Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 626083, 10 pages
http://dx.doi.org/10.1155/2013/626083

Research Article

Ultradeep Pyrosequencing of Hepatitis C Virus Hypervariable Region 1 in Quasispecies Analysis

Kamila Caraballo Cortés,[1,2] Osvaldo Zagordi,[3] Tomasz Laskus,[1] Rafał Płoski,[4] Iwona Bukowska-Ośko,[1] Agnieszka Pawełczyk,[1] Hanna Berak,[5] and Marek Radkowski[1]

[1] *Department of Immunopathology of Infectious and Parasitic Diseases, Medical University of Warsaw, 3c Pawińskiego Street, 02-106 Warsaw, Poland*
[2] *Postgraduate School of Molecular Medicine, Żwirki i Wigury 61 Street, 02-091 Warsaw, Poland*
[3] *Institute of Medical Virology, University of Zurich, Winterthurerstrasse, 190 8057 Zurich, Switzerland*
[4] *Department of Medical Genetics, Medical University of Warsaw, 3c Pawińskiego Street, 02-106 Warsaw, Poland*
[5] *Hospital for Infectious Diseases, 37 Wolska Street, 01-201 Warsaw, Poland*

Correspondence should be addressed to Kamila Caraballo Cortés; kcaraballo@wum.edu.pl

Received 22 October 2012; Accepted 12 February 2013

Academic Editor: Ozgur Cogulu

Genetic variability of hepatitis C virus (HCV) determines pathogenesis of infection, including viral persistence and resistance to treatment. The aim of the present study was to characterize HCV genetic heterogeneity within a hypervariable region 1 (HVR1) of a chronically infected patient by ultradeep 454 sequencing strategy. Three independent sequencing error correction methods were applied. First correction method (Method I) implemented cut-off for genetic variants present in less than 1%. In the second method (Method II), a condition to call a variant was bidirectional coverage of sequencing reads. Third method (Method III) used *Short Read Assembly into Haplotypes* (ShoRAH) program. After the application of these three different algorithms, HVR1 population consisted of 8, 40, and 186 genetic haplotypes. The most sensitive method was ShoRAH, allowing to reconstruct haplotypes constituting as little as 0.013% of the population. The most abundant genetic variant constituted only 10.5%. Seventeen haplotypes were present in a frequency above 1%, and there was wide dispersion of the population into very sparse haplotypes. Our results indicate that HCV HVR1 heterogeneity and *quasispecies* population structure may be reconstructed by ultradeep sequencing. However, credible analysis requires proper reconstruction methods, which would distinguish sequencing error from real variability *in vivo*.

1. Introduction

Genetic variability is a characteristic feature of hepatitis C virus (HCV), due to an absence of error correction mechanisms of the viral RNA-dependent RNA polymerase, fast replication, and recombination events [1–3]. As a consequence, HCV displays high intrahost population diversity, forming a pool of closely related but distinct genetic variants (*quasispecies*) [1]. The viral genetic variability is not evenly distributed through the entire genome; the highest variable regions include HVR1, HVR2, and HVR3 of the envelope E2 protein [4]. It is believed that HCV variability has significant clinical implications, since it may result in the generation of immune escape mutants, which may contribute to chronic infection and treatment resistance [5].

The detailed study of the minor variants within the *quasispecies* population is hampered by the absence of sensitive sequencing strategies which would allow for the detection of low-frequency genomes. The traditional method for studying viral *quasispecies* is based on Sanger sequencing of bacterially cloned viral sequences. However, this strategy requires extensive cloning to achieve the desired sensitivity for minor variants detection, a process, that is, costly and time consuming. Another limitation of the Sanger method is its difficulty in sequencing GC-rich regions.

Other studies employed single strand conformational polymorphism (SSCP), an electrophoretic method shown to detect variants constituting as little as 3% of the viral population [6]. However, SSCP it is not informative of the

nature of genetic changes or genetic distance between variants and therefore could not be used for some applications such as investigation of the drug resistance. In addition, in a mixture of heterogenous sequences, certain bands may overlap, underrating viral complexity.

With next-generation sequencing (NGS) platforms, it is now possible to investigate viral *quasispecies* at much greater detail. Their high throughput allows for generation of millions of reads in a single sequencing run, facilitating in-depth sequencing.

NGS can detect variants at low frequencies, which would go undetected by standard sequencing methods [7]. Nevertheless, in order to make reliable reconstruction of the viral *quasispecies* from the noisy, incomplete data obtained by NGS, a proper data analysis is required [8, 9].

In the present study we used ultradeep pyrosequencing (454/Roche) to characterize the complexity and heterogeneity of hypervariable region 1 (HVR1) in a patient persistently infected with HCV genotype 1b. This region was chosen as its protein product is under constant selection pressure of the host immune responses, especially of cytotoxic T cells and neutralizing antibodies [10, 11]. We sequenced this short region at very high coverage, aiming at detecting a large number of minority variants. We took into account sequencing errors in order to have a reliable reconstruction of the viral *quasispecies* on this region.

Reports taking advantage of deep sequencing to investigate HCV genetic diversity for clinical and epidemiological studies are currently available [12–15]. Likewise, there are also works reporting and comparing bioinformatic approaches to infer the viral population from clinical samples, mostly HIV [9, 16–19]. Our study contributes progress in the evaluation of reconstruction methods and extends it for HCV *quasispecies* phenomenon investigation.

2. Patient and Methods

2.1. Sample. A serum sample from a 66-year-old treatment-naive female patient with genotype 1b chronic HCV infection was used. The serum HCV viral load was 1.54×10^6 IU/mL. The patient provided informed consent and the study was approved by the Institutional Bioethical Committee.

2.2. HVR1 Amplification. Viral RNA was extracted from 250 μL of serum by a modified guanidinium thiocyanate-phenol/chlorophorm method using a commercially available Trizol reagent (Invitrogen) and suspended in 20 μL of water. Five μL of the solution containing RNA was subjected to reverse transcription at 37°C for 30 minutes using AccuScript High Fidelity Reverse Transcriptase (Stratagene). HVR1 sequences were amplified in a two-step PCR using FastStart High Fidelity Taq DNA Polymerase (Roche) as described previously [20]. Primers used for reverse transcription (E2 AS) and first round HCV HVR1 amplification (E2 S) were as follows: 5′-CATTGCAGTTCAGGGCCGTGCTA-3′ and 5′-GGTGCTCACTGGGGAGTCCT-3′. Primers for the second round PCR (E2 NS and E2 NAS) were as follows: 5′-CGT ATC GCC TCC CTC GCG CCA TCAG TCC ATG GTG GGG AAC TGG GC-3′ and 5′-CTA TGC GCC TTG CCA GCC CGC TCAG TGC CAA CTG CCA TTG GTG TT-3′. The latter contained tags recognized by GS Junior Sequencing System (underlined).

2.3. SSCP Analysis of HVR1 Quasispecies. Second round PCR product was purified using Wizard SV Genomic DNA Purification System (Promega) and resuspended in 20 μL of water. Next, 2–5 μL of purified PCR product was subjected to SSCP assay as described previously [21]. Complexity of a population was reflected by the number of distinct bands.

2.4. Ultradeep Pyrosequencing. Pyrosequencing was carried out according to the manufacturer's protocol for amplicons using GS Junior System (454/Roche). In order to lower contamination with short sequences (i.e., primer residues), HVR1 product of the second round PCR was purified from agarose gel by QIAquick Gel Extraction Kit (Qiagen). The extracted product was measured fluorometrically using Quant-iT PicoGreen dsDNA Assay Kit (Molecular Probes), and the amount of DNA equivalent to 3×10^7 copies was subjected to emulsion PCR using GS Junior Titanium emPCR Kit (Lib-A). Pyrosequencing was performed according to the amplicon processing procedure for 100 cycles (recommended for amplicons up to 250 bp).

2.5. Data Analysis. Reads that did not match primer sequences or had undetermined bases (Ns) were excluded from further analysis. Retained sequences of 179 bp were visualized using GS Amplicon Variant Analyzer (Roche). Subsequently, primer sequences were trimmed from the target sequence and reads of 138 bp were aligned to the reference sequence for genotype 1b HCV (GenBank accession number AJ406073) and translated to amino acid sequences by (*Molecular Evolutionary Genetics Analysis)* MEGA, version 5.0 (http://www.megasoftware.net/) [22]. Phylogenetic analyses were conducted in MEGA5 using the Maximum Likelihood method based on the Tamura-Nei model [23] using MEGA 5.0 software. Genetic parameters such as genetic diversity and sequence polymorphisms within sequences were 5 assessed by DNA SP version (http://www.ub.edu/dnasp/). The program diri_sampler from the ShoRAH software was used to correct sequencing errors and infer haplotypes. Given the high number of reads obtained in the sequencing, the dataset was split equally in two, and the obtained sets were analyzed independently. Error correction included mismatches as well as insertions and deletions.

3. Results

3.1. Amplification and Sequencing Errors. As our experiment used RT-PCR-amplified material, we attempted to assess the error rate in the consecutive experimental steps taking into account error rates of employed enzymes. For reverse transcription, AccuScript High Fidelity Reverse Transcriptase (Stratagene) was used, which displays three times higher fidelity than commonly used MMLV reverse transcriptase [24]. The estimated AccuScript RT error rate is 2×10^{-5}

Figure 1: The SSCP image of HVR1 amplified from the serum of HCV-infected patient.

(manufacturers data). For PCR amplification, we used FastStart High Fidelity Taq DNA Polymerase (Roche), which has estimated error rate of 2×10^{-6} (three times lower than Taq DNA polymerase) [25]. Finally, the pyrosequencing error rate is estimated to be 1.07%, including mismatches (0.088%), insertions (0.541%), deletions (0.359%), and ambiguous base calls (0.085%) [26].

Studying clonal samples, or control samples where a set of clones are mixed in predetermined proportions are important to evaluate the error rate of the sequencing process and the performance of the haplotype reconstruction methods. Since these have already been reported elsewhere [16, 26], it seems not requisite to perform these experiments for every new study of the viral *quasispecies*.

3.2. Heterogeneity of HCV HVR1 Viral Variants Assessed by SCCP Analysis. Based on gel analysis, at least nine SSCP bands of HVR1 were observed (Figure 1). The frequency was not uniform across variants, as could be seen by the different intensities of the bands.

3.2.1. Heterogeneity of HCV HVR1 Assessed by Ultradeep Sequencing. To check the applicability of ultradeep pyrosequencing for HCV HVR1 heterogeneity analysis, the amplified product was sequenced by GS Junior System (454/Roche). Based on the data of GS Amplicon Variant Analyzer, the total number of sequenced nucleotides was 1.37×10^8. The system read 76 332 individual sequences, among them 73 236 (95.9%) (28 098 forward and 45 138 reverse) were aligned to the reference sequence AJ406073 of genotype 1b HCV. The GS Amplicon Variant Analyzer software detected 15 917 haplotypes. Mean coverage of each variant (expressed by the number of identical reads) was 4.6. The most abundant haplotype coverage was 4540 reads. The rarest haplotypes comprised single sequence reads (74,6% of detected haplotypes). Our results are summarized in Table 1.

Table 1: HVR1 HCV characteristics obtained by pyrosequencing using GS Junior System (454/Roche).

Number of sequenced nucleotides	1.37×10^8
Number of individual sequences that passed the quality control*	76 332
Number of individual sequences aligned to reference genome	73 236
Mean coverage per sequence	4.6
Identified haplotypes	15 917

*No undetermined bases, 100% match with primer sequences.

3.2.2. Error Correction in Haplotype Reconstruction. In order to reflect the HVR1 HCV population *in vivo* as accurately as possible, we explored the effect of different strategies to take the sequencing error rate into account. In a very conservative approach (Method I), we only considered variants detected at a frequency higher than 1%. This amounts to discard most variants, even if, given the high coverage, they appear in hundreds of reads. With this strategy we only retained 8 haplotypes.

A second strategy (Method II) consisted in requiring bidirectional coverage, that is, in only retaining variants supported by at least one forward and one reverse read. This method identified 40 HVR1 variants.

In the third approach we used the program diri_sampler from the software suite ShoRAH [16]. In this analysis, inference of the viral *quasispecies* is done in probabilistic manner using a Bayesian approach. It does not rely on the input of an error rate, rather, it estimates it from the sequencing data. Reads are clustered together and the consensus sequence of each cluster represents the original haplotype. Together with the frequency of each variant, the program enables assessments of the posterior probability of each haplotype, a confidence value for their existence. The number of diverse reads sequenced was higher than what the program can handle on a desktop computer with 4 GB of RAM. In order to face this limitation, we split the reads equally in two subsets, and performed haplotype reconstruction independently. Only haplotypes with confidence value >95% were retained. As an additional measure of reliability, only haplotypes supported by at least 5 reads were included. Since we are dealing with a coding sequence, frameshift inducing insertions/deletions were resolved correcting to the most common nucleotide for that position. As a result of two independent computations on raw data halves, two populations (A and B), consisting of 333 and 315 haplotypes respectively, were obtained. Their frequencies varied from 10.54% and 10.44% (the most abundant variants in population A and B, resp.) down to 0.013% and 0.014% (the least abundant variants in population A and B, resp.). 186 haplotypes were common to both populations and their frequencies were all above 0.02%. Seventeen haplotypes were present with a frequency >1%, constituting in total 58.6% of the entire population.

3.2.3. Characteristics of Inferred HVR1 Populations. After application of error correction methods, such parameters as

Table 2: The impact of haplotype reconstruction method on the variability parameters of HVR1.

Correction method	I Cut-off 1%	II Bidirectional coverage	III ShoRAH
Number of haplotypes	8	40	186
Number of nucleotide substitutions within HVR1	51	59	70
Percentage of mutated amino acid positions within HVR1 (%)	55.6	55.6	74.1
Genetic distance	3.874	0.065	0.110
Genetic diversity	0.923	0.998	0.984

percentage of mutated amino acid positions, genetic distance, genetic diversity as well as number of substitutions were calculated (Table 2). The highest genetic distance characterized population reconstructed by cut-off method (3.874) followed by ShoRAH method (0.110) and bi-directional coverage method (0.065), whereas genetic diversities were similar for all populations (0.923, 0.998 and 0.984 for method I, II and III, resp.). The highest number of nucleotide substitutions was detected in ShoRAH-reconstructed population (overall 70). 47 (67%) of them were present in genetic variants constituting more than 1% of the entire population.

HVR1 populations were also compared on amino acid level (Figure 2). Within 27 amino acid stretch of HVR1, only 15 (55.5%) positions were polymorphic after application of methods I and II, and 20 (74.1%) after ShoRAH computations. Based on ShoRAH computation results, the most variable was the fourth HVR1 position, where 11 amino acid substitutions were detected when compared to reference sequence (V/D, V/M, V/T, V/L, V/R, V/A, V/E, V/G, V/N, V/I, V/Q).

Viral populations were also analyzed phylogenetically. As shown in Figure 3, the general topology of three populations was similar. However, the tree topology based on ShoRAH computation was the most extensive.

4. Discussion

Pyrosequencing is a relatively novel technique which may help to decipher complex viral populations in terms of their diversity and structure. To date, it was successfully used in human immunodeficiency virus (HIV) research to identify minor drug resistant variants, analyze variable regions of heavy and light chains of neutralizing antibodies against HIV, as well as to determine HIV tropism, analyze superinfections and assess diversity of genital microbiota in HIV-infected women [27–31]. Ultradeep sequencing strategies also offers a new approach in HCV research. However, application of this method requires that several issues are taken into account. The foremost of these is the generation of mutations during reverse transcription and amplification reactions, due to enzyme errors [32]. Reverse transcriptase is the most error-prone, as it lacks a proofreading activity. For instance, error rate of common reverse transcriptases used *in vitro* to synthesize cDNA is at least 10^{-4} [24], and errors that occurred during this step are propagated during the subsequent PCR amplification. In the present study, in order to minimize errors, high fidelity enzymes were used in amplification reactions preceding sequencing (AccuScript High Fidelity Reverse Transcriptase and FastStart High Fidelity Taq DNA Polymerase). Nevertheless, the resulting hypothetical error rate of amplification is estimated to be lower than the sequencing error rate itself. The sequencing step introduces various types of errors related to the pyrosequencing chemistry and detection technology. The major contributor to errors is the ambiguity of homopolymer length, which results from the difficulty to resolve intensity of luminescence when a homopolymer is encountered. Moreover, insufficient flushing may lead to single base insertions. Overall, it was estimated that the mean error rate of pyrosequencing (defined as the number of errors such as miscalled bases or inserted or deleted bases divided by the total number of sequenced bases) was 1.07% [26]. This value may be considered as the experimentally confirmed resolution of the method. For the above reasons, the raw data obtained from sequencing should be additionally processed in order to remove low-quality reads and reads containing errors.

Three different error correction methods were applied to the raw sequencing data, which resulted in three HVR1 populations, differing in complexity and heterogeneity. The most sensitive was ShoRAH program reconstruction, which allowed to obtain the broadest spectrum of HVR1 sequences. This method has already been shown to reliably detect variants down to about 0.1% [33]. In this study we detected variants down to 0.02%, confirmed in two independent computations. The cut-off method, in which variants present in less than 1% of the population were discarded, was the least sensitive, as it allowed to detect only 8 haplotypes. Similar cut-off was applied in analysis of pyrosequencing reads of *pol/gag* of HIV population as well as *PePHD* E2 of HCV [14, 34]. It was reported that this method may result in inadequate haplotype reconstruction of low precision, low recall, or both, depending on the cut-off value. Too low cut-off value may result in low precision (fraction of true haplotypes among all called haplotypes) and conversely, high cut off may significantly lower recall (fraction of called haplotypes among all true haplotypes). For instance, based on the analysis of *gag/pol* HIV genes, it was shown that the cut-off of 50 read observations resulted in 80% precision but only 40% recall [33]. Application of the bi-directional coverage correction method II allowed us to determine the presence of 40 haplotypes, but such verification is laborious and raises concern regarding the acceptance of haplotypes characterized by high disproportion in forward and reverse strand counts. Among these forty sequences with bidirectional coverage, we could identify twenty that matched exactly one of the

```
                                                      HVR1
#AJ406073                     KVLIVMLLFA GVDGGHTHVT GGAAAHAAYG FRHLFSPGSA QKLQLV
#CON_93_amplicon_Amp_1        .......... ....E-.RTI ...E.R..S. LSG..TR..Y ..I...
#CON_11189_amplicon_Amp_1     .......... ....E-.RTI ..-E.R..S. LSG..TR..Y ..I...
#CON_7239_amplicon_Amp_1      .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_7283_amplicon_Amp_1      .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_40_amplicon_Amp_1        .......... ....E-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_7319_amplicon_Amp_1      .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_7250_amplicon_Amp_1      .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_11507_amplicon_Amp_1     .......... ....E-..TI ..-V.RTTS. LSG..RA..H ..I...
```

(a)

```
                                                      HVR1
#AJ406073                     KVLIVMLLFA GVDGGHTHVT GGAAAHAAYG FRHLFSPGSA QKLQLV
#CON_236_amplicon_Amp_1       .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_1037_amplicon_Amp_1      .......... ....E-.RTI ...E.R..S. LSG..TR..Y ..I...
#CON_1051_amplicon_Amp_1      .......... ....E-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_1080_amplicon_Amp_1      .......... ....E-..TI ...V.RTTS. LSS..TA..Q ..I...
#CON_1111_amplicon_Amp_1      .......... ....E-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_1120_amplicon_Amp_1      .......... ....E-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_1174_amplicon_Amp_1      .......... ....E-..TI ...V.RTTS. .SS..TA..Q ..I...
#CON_1179_amplicon_Amp_1      .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_1241_amplicon_Amp_1      .......... ....E-.RTI ...E.R..S. LSG..TR..Y ..I...
#CON_125_amplicon_Amp_1       .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_126_amplicon_Amp_1       .......... ....E-..TI ...V.RTTS. .SG..RA..Q ..I...
#CON_1294_amplicon_Amp_1      .......... ....K-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_1620_amplicon_Amp_1      .......... ....E-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_1710_amplicon_Amp_1      .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_1727_amplicon_Amp_1      .......... ....E-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_181_amplicon_Amp_1       .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_1832_amplicon_Amp_1      .......... ....E-.RTI ...E.R..S. LSG..TR..Y ..I...
#CON_184_amplicon_Amp_1       .......... ....E-..TI ...V.RTTS. LSG..RA..H ..I...
#CON_223_amplicon_Amp_1       .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_240_amplicon_Amp_1       .......... ....E-..TI ...V.RTTS. LSG..RA..H ..I...
#CON_243_amplicon_Amp_1       .......... ....E-..TI ...V.RTTS. LSG..RA..H ..I...
#CON_245_amplicon_Amp_1       .......... ....E-..TI ...V.RTTS. LSG..TA..Q ..I...
#CON_2667_amplicon_Amp_1      .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_3101_amplicon_Amp_1      .......... ....T-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_39_amplicon_Amp_1        .......... ....E-..TI ...V.RTTS. .SG..RA..H ..I...
#CON_397_amplicon_Amp_1       .......... ....E-..TI ...V.RTTS. LSG..RA..Q ..I...
#CON_40_amplicon_Amp_1        .......... ....E-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_307_amplicon_Amp_1       .......... ....E-..TI ...V..G.R. LSS..TA..Q ..I...
#CON_48_amplicon_Amp_1        .......... ....E-..TI ...V.RTTS. .SG..RA..H ..I...
#CON_491_amplicon_Amp_1       .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_493_amplicon_Amp_1       .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_508_amplicon_Amp_1       .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_54_amplicon_Amp_1        .......... ....E-..TI ...V.RTTS. LSG..RA..H ..I...
#CON_548_amplicon_Amp_1       .......... ....E-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_688_amplicon_Amp_1       .......... ....E-.RTI ..SV..G.R. LSS..TA..Q ..I...
#CON_692_amplicon_Amp_1       .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_724_amplicon_Amp_1       .......... ....E-..TI ...V.RTTS. LSG..RA..H ..I...
#CON_788_amplicon_Amp_1       .......... ....R-..TI ..SV..G.R. LSS..TA..Q ..I...
#CON_883_amplicon_Amp_1       .......... ....E-..TI ...V.RTTS. LSG..TR..Y ..I...
#CON_947_amplicon_Amp_1       .......... ....E-.RTI ...E.R..S. LSG..TR..Y ..I...
```

(b)

```
                                                      HVR1
#AJ406073                     KVLIVMLLFA GVDGGHTHVT GGAAAHAAYG FRHLFSPGSA QKLQLV
#seq_1_frequency_10.496963    .......... .....DS.DR .-.E.R..S. LSG..TR..Y ..I...
#seq_2_frequency_8.547829     .......... .....DS.MI ...E.R..S. LSG..TR..Y ..I...
#seq_3_frequency_7.419855     .......... ..E.-G..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_4_frequency_6.384569     .......... ..E.-G..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_5_frequency_4.296772     .......... ....-E..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_6_frequency_3.850357     .......... ..E.-G..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_7_frequency_3.171748     .......... ..E.-G..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_8_frequency_2.787145     .......... .....DP.DR .-.V.RTTS. LSG..RA..H ..I...
#seq_9_frequency_2.250011     .......... ...-.E..TI ...V.RTTS. LSG..RA..H ..I...
#seq_10_frequency_1.504237    .......... ..E.-G..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_11_frequency_1.251363    .......... ..E.-G..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_12_frequency_1.187379    .......... ...-.E..TI ...V.RTTS. .SG..RA..H ..I...
#seq_13_frequency_1.178896    .......... .....DS.MI ...E.R..S. LSG..TR..Y ..I...
#seq_14_frequency_1.083473    .......... .....DS.DR .-.E.R..S. LSG..TR..Y ..I...
#seq_15_frequency_1.075477    .......... ....-E..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_16_frequency_1.038557    .......... .....DP.DR .-.V.RTTS. .SG..RA..H ..I...
#seq_17_frequency_1.000225    .......... .....DS.MI ...E.R..S. LSG..TR..Y ..I...
#seq_18_frequency_0.926960    .......... .....DP.DR .-.V..G.R. LSS..TA..Q ..I...
#seq_19_frequency_0.882146    .......... ..EW.DS.DR .-.E.R..S. LSG..TR..Y ..I...
#seq_20_frequency_0.863415    .......... ....-E..TI ...V..G.R. LSS..TA..Q ..I...
#seq_21_frequency_0.803007    .......... .....DS.DR .-.V.RTTS. LSG..RA..H ..I...
#seq_22_frequency_0.777475    .......... ..E.-G..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_23_frequency_0.760876    .......... .....DS.DR .-.E.R..S. LSG..TR..Y ..I...
#seq_24_frequency_0.755084    .......... ..E.-G..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_25_frequency_0.701852    .......... ....-E..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_26_frequency_0.672432    .......... .....DP.DR .-.V.RTTS. LSG..TR..Y ..I...
#seq_27_frequency_0.655984    .......... .....DP.DR .-.V.RTTS. LSG..RA..H ..I...
#seq_28_frequency_0.654307    .......... ..E.-G..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_29_frequency_0.651253    .......... ...-.E..TI ...V.RTTS. LSG..RA..H ..I...
#seq_30_frequency_0.650095    .......... ...-.E..TI ...V.RTTS. LSG..TR..Y ..I...
#seq_31_frequency_0.625705    .......... .....DP.DR .-.E.R..S. LSG..TR..Y ..I...
#seq_32_frequency_0.612851    .......... ....-E..TI ...E.R..S. LSG..TR..Y ..I...
#seq_33_frequency_0.602512    .......... ...-.E..TI ...V.RTTS. .SG..RA..Q ..I...
```

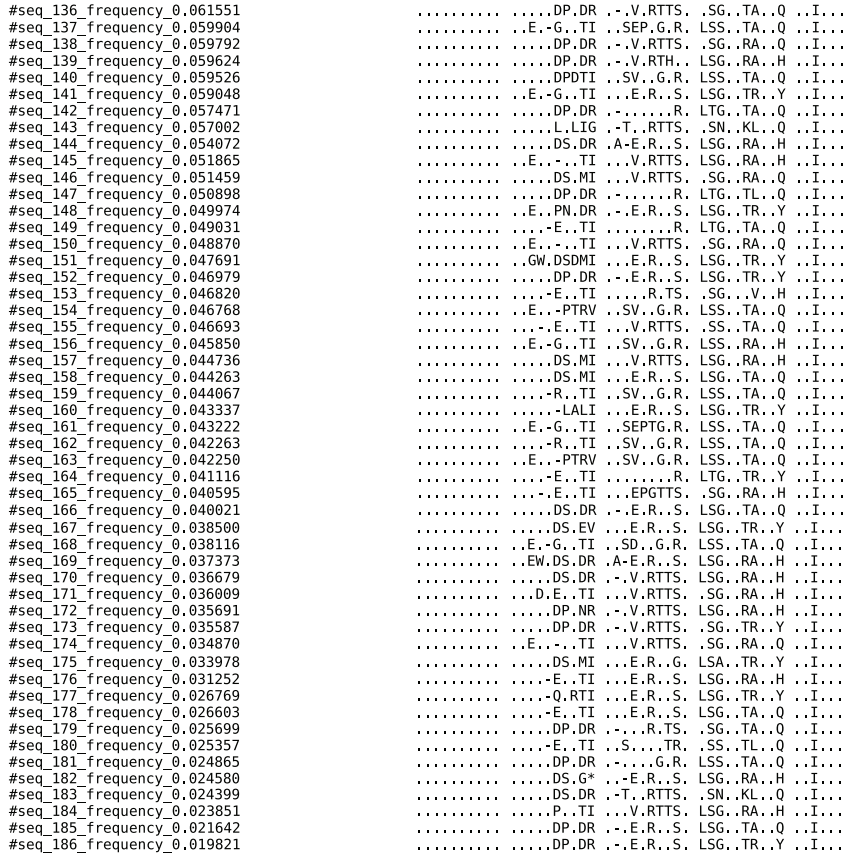
(c)

Figure 2: Continued.

```
#seq_34_frequency_0.566364              ..... ...-.E..TI ...V.RTTS. LSG..RA..H ..I...
#seq_35_frequency_0.539622              ..... .....DP.DR .-.V.RTTS. LSG..RA..Q ..I...
#seq_36_frequency_0.537195              ..... ...-.E..TI ...V.RTTS. .SG..RA..H ..I...
#seq_37_frequency_0.521421              ..... ...-.E..TI ...V.RTTS. LSG..RA..Q ..I...
#seq_38_frequency_0.511224              ..... .....DS.MI ..SV..G.R. LSS..TA..Q ..I...
#seq_39_frequency_0.506563              ..... .....DS.MI ...E.R..S. LSG..TR..Y ..I...
#seq_40_frequency_0.489579              ..... .....DS.DR .-.E.R..S. LSG..TR..Y ..I...
#seq_41_frequency_0.477541              ..... ...-.E..TI ...V.RTTS. LSG..RA..H ..I...
#seq_42_frequency_0.475100              ..... ....-E..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_43_frequency_0.416991              ..... .....DP.DR .-SV..G.R. LSS..TA..Q ..I...
#seq_44_frequency_0.403574              ..... .....DS.DR .-.E.R..G. LSA..TR..Y ..I...
#seq_45_frequency_0.398722              ..... .....DS.DR ...E.R..S. LSG..TR..Y ..I...
#seq_46_frequency_0.386373              ..... ....-E..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_47_frequency_0.375498              ..... .....DP.DR .-.V.RTTS. LSG..RA..H ..I...
#seq_48_frequency_0.368952              ..... ...-.E..TI ...V.RTTS. LSG..TA..Q ..I...
#seq_49_frequency_0.353843              ..... ...-.E..TI ...V.RTTS. .SS..RA..H ..I...
#seq_50_frequency_0.350268              ..... ..E..PN.DR .-SV..G.R. LSS..TA..Q ..I...
#seq_51_frequency_0.346676              ..... ..E..PN.DR .-SV..G.R. LSS..TA..Q ..I...
#seq_52_frequency_0.336838              ..... .....DP.DR .-.V.RTTS. LSG..RA..H ..I...
#seq_53_frequency_0.332207              ..... ...-.E..TI ...V.RTTS. LSG..RA..H ..I...
#seq_54_frequency_0.330468              ..... .....DP.DR .-.V.RTTS. .SG..RA..Q ..I...
#seq_55_frequency_0.325899              ..... .....DS.DR .-.E.R..S. LSG..TA..Q ..I...
#seq_56_frequency_0.324684              ..... .....DP.DR .-.V.RTTS. .SS..RA..H ..I...
#seq_57_frequency_0.323406              ..... ....-E..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_58_frequency_0.322620              ..... .....DS.MS R..E.R..S. LSG..TR..Y ..I...
#seq_59_frequency_0.291277              ..... .....DS.MI ...E.R..S. LSG..TA..Q ..I...
#seq_60_frequency_0.290212              ..... .....DS.DR .A-E.R..S. LSG..RA..H ..I...
#seq_61_frequency_0.265723              ..... .....DS.MI ...E.R..G. LSA..TR..Y ..I...
#seq_62_frequency_0.256780              ..... .....DP.DR .-.V.RTTS. LSG..RA..H ..I...
#seq_63_frequency_0.243733              ..... .....DS.MI ...E.R..S. LSG..RA..H ..I...
#seq_64_frequency_0.232532              ..... .....DP.DR .-.V.RTTS. .SG..TA..Q ..I...
#seq_65_frequency_0.232330              ..... ...-.E..TI ...V.RTTS. .SS..TA..Q ..I...
#seq_66_frequency_0.231031              ..... ..E..PN.DR .-SV..G.R. LSS..TA..Q ..I...
#seq_67_frequency_0.226299              ..... .....P..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_68_frequency_0.226212              ..... ...-.E..TI ...V.RTTS. LSG..RA..H ..I...
#seq_69_frequency_0.226001              ..... ...-.E..TI ...V.RTTS. LSS..TA..Q ..I...
#seq_70_frequency_0.214432              ..... .....P..TI ..S....TR. .SS..TL..Q ..I...
#seq_71_frequency_0.198007              ..... ...-.E..TI ...V.RTTS. .SG..TA..Q ..I...
#seq_72_frequency_0.189106              ..... ..E..PN.DR .-SV..G.R. LSS..TA..Q ..I...
#seq_73_frequency_0.171091              ..... .....DP.DR .-.V.RTTS. .SS..TA..Q ..I...
#seq_74_frequency_0.161615              ..... .....DP.DR .-.V.RTTS. LSG..RA..H ..I...
#seq_75_frequency_0.158776              ..... ....-E.RAI ...E.R..S. LSG..TR..Y ..I...
#seq_76_frequency_0.156317              ..... ....-K..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_77_frequency_0.148586              ..... .....DS.DR ..RE.R..S. LSG..TR..Y ..I...
#seq_78_frequency_0.147589              ..... ..E..-LTLI ..SV..G.R. LSS..TA..Q ..I...
#seq_79_frequency_0.144746              ..... .....DP.DR .-.V.RTTS. LSS..TA..Q ..I...
#seq_80_frequency_0.143410              ..... ....-K..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_81_frequency_0.141619              ..... .....DS.DR .-.V.RTTS. LSG..RA..H ..I...
#seq_82_frequency_0.140555              ..... .....R..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_83_frequency_0.138271              ..... .....DP.DR .-.V..G.R. LSS..TA..Q ..I...
#seq_84_frequency_0.132354              ..... ..E.-G..TI ..SEP.G.R. LSS..TA..Q ..I...
#seq_85_frequency_0.127754              ..... .....DP.DR .-.V.RTTS. LSG..TA..Q ..I...
#seq_86_frequency_0.123660              ..... ...-.E..TI ...V.RTH.. LSG..RA..H ..I...
#seq_87_frequency_0.121928              ..... ....-G..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_88_frequency_0.120898              ..... .....DS.MI ..SV..G.R. LSS..TA..Q ..I...
#seq_89_frequency_0.118086              ..... .....DS.EV ...E.R..S. LSG..TR..Y ..I...
#seq_90_frequency_0.117558              ..... ..E..-..TI ...V.RTTS. .SG..RA..H ..I...
#seq_91_frequency_0.117161              ..... .....R..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_92_frequency_0.112927              ..... .....P..TI ..S....TR. .SS..TA..Q ..I...
#seq_93_frequency_0.112631              ..... ....-.E..TI ...V.RTTS. LSG..RA..H ..I...
#seq_94_frequency_0.112472              ..... .....DS.MI ...V.RTTS. LSG..RA..H ..I...
#seq_95_frequency_0.112397              ..... ....-Q.RTI ...E.R..S. LSG..TR..Y ..I...
#seq_96_frequency_0.111663              ..... .....DP.DR ..RV.RTTS. .SG..RA..H ..I...
#seq_97_frequency_0.107630              ..... ..E..-..TI ...V.RTTS. LSG..RA..H ..I...
#seq_98_frequency_0.106809              ..... ....-.E..TI ...V.RTTS. S.G..RA..H ..I...
#seq_99_frequency_0.103442              ..... ..E..PN.DR .-.V.RTTS. LSG..RA..H ..I...
#seq_100_frequency_0.100934             ..... ...-.E..TI ...V.RTTS. .SG..RA..Q ..I...
#seq_101_frequency_0.100065             ..... ..E..-PTRV ..SV..G.R. LSS..TA..Q ..I...
#seq_102_frequency_0.099957             ..... ..E..-PDTI ..SV..G.R. LSS..TA..Q ..I...
#seq_103_frequency_0.099954             ..... ...-.E..TI ...V.RTTS. LSG..TR..Y ..I...
#seq_104_frequency_0.097549             ..... ..E..PN.DR .-.V.RTTS. LSG..RA..H ..I...
#seq_105_frequency_0.095057             ..... ..EW.DS.DR .-.E.R..S. LSG..TR..Y ..I...
#seq_106_frequency_0.095048             ..... ....-G..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_107_frequency_0.094927             ..... .....DS.DR .-.E..G.R. LSS..TA..Q ..I...
#seq_108_frequency_0.093505             ..... .....DP.DR .-SV..G.R. LSS..TA..Q ..I...
#seq_109_frequency_0.090986             ..... .....DP.DR .-.V.RTTS. LSG..TR..Y ..I...
#seq_110_frequency_0.090194             ..... ....-E..TI ...E.R..S. LSG..TR..Y ..I...
#seq_111_frequency_0.089734             ..... ..E.-G..TI ..SV..G.R. LSS..TR..Y ..I...
#seq_112_frequency_0.089089             ..... .....DP.DR .-.E.R..S. LSG..TR..Y ..I...
#seq_113_frequency_0.088302             ..... .....DS.MI ...E.R.G*. LSG..TR..Y ..I...
#seq_114_frequency_0.081532             ..... ....-K..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_115_frequency_0.081455             ..... ..E..-PTQI ..SV..G.R. LSS..TA..Q ..I...
#seq_116_frequency_0.079521             ..... ..GR.P..TI ...V.RTTS. .SG..RA..H ..I...
#seq_117_frequency_0.079391             ..... ....-E..TI ..SV..G.R. LSS..TA..Q ..I...
#seq_118_frequency_0.078571             ..... ...-.E..TI ..R*.RTTS. LSG..RA..H ..I...
#seq_119_frequency_0.076269             ..... .....DS.G* ..PE.R..S. LSG..TR..Y ..I...
#seq_120_frequency_0.074612             ..... .....DP.DR .-.V.RTTS. S.G..RA..H ..I...
#seq_121_frequency_0.074134             ..... .....DP.EV ..SV..G.R. LSS..TA..Q ..I...
#seq_122_frequency_0.074079             ..... ...-.E..TI ..R*PRTTS. LSG..RA..H ..I...
#seq_123_frequency_0.073492             ..... ..E.-G..TI ..SE..G.R. LSS..TA..Q ..I...
#seq_124_frequency_0.072251             ..... .....DS.G* ...E.R..S. LSG..TR..Y ..I...
#seq_125_frequency_0.071245             ..... ..E.-G..TI ..SE..G.R. LSS..TA..Q ..I...
#seq_126_frequency_0.070688             ..... ..E.-G..TI ..SE..G.R. LSS..TA..Q ..I...
#seq_127_frequency_0.070584             ..... .....DP.DR .-...R.TS. .SG...V..H ..I...
#seq_128_frequency_0.070491             ..... .....DS.DR .-.E.R..G. LSA..TR..Y ..I...
#seq_129_frequency_0.069469             ..... ....-E..TI ........R. LTG..TL..Q ..I...
#seq_130_frequency_0.065577             ..... ..E.-G..TI ..SV..G.R. LSS..TR..Y ..I...
#seq_131_frequency_0.065258             ..... ...-.E..TI ..RV.RTTS. LSG..RA..H ..I...
#seq_132_frequency_0.064382             ..... .....DS.MI ...E..G.R. LSS..TA..Q ..I...
#seq_133_frequency_0.064154             ..... ...-.E..TI ...V.RTTS. .SG..RA..H ..I...
#seq_134_frequency_0.063206             ..... ...-.E..TI ...EPRTTS. LSG..RA..H ..I...
#seq_135_frequency_0.061966             ..... ....-R..TI ..SV..G.R. LSS..TA..Q ..I...
```

(c)

Figure 2: Continued.



(c)

Figure 2: Amino acid sequences of HVR1 populations inferred after the application of three different error correction methods. (a) Cut-off method >1% (I), (b) bidirectional coverage (II), and (c) ShoRAH computation (III). Top sequence corresponds to reference sequence AJ406073 for genotype 1b HCV. Dots indicate consensus positions. Dashes indicate positions not present in the sequence. Asterisks indicate stop codons.

sequences obtained with ShoRAH. This is probably a very precise dataset, although, for example, some sequences might have a very biased forward/reverse read ratio and yet be included in it. Undoubtedly, using the strand information while reconstructing haplotypes is a strategy worth pursuing. A promising approach seems to be a proper statistical treatment of the strand bias, implemented together with the error correction of ShoRAH (McElroy, unpublished data).

In our study, using two independent ShoRAH computations, 186 haplotypes were reconstructed. In contrast, in the study of Bull et al. [13], 100 E2 variants were detected by ShoRAH. However, patients in that study were in an early phase of HCV infection and the study was performed along the whole genome, at much lower coverage, and thus HVR1 complexity could have been lower [13]. Based on ShoRAH reconstruction, we found that the most frequent HVR1 variants constitute a relatively small percentage of the entire population. Thus, the most abundant variant constituted 10.5%, and only 17 haplotypes were present in proportions higher than 1%. These data suggest that during the chronic phase of infection the *quasispecies* population is highly dispersed into minor variants, with no predominant sequences present. Similar haplotype frequency distribution was observed in foot-and-mouth virus population reconstructed by next generation genome sequencing [35]. In the only other HCV study investigating this issue, two to five variants were detected in frequency higher than 2.5%, whereas we detected eight such sequences. However, as already mentioned, viral samples in that study were drawn in the acute phase of infection [13]. The highest number of substitutions (70) within HVR1 was detected in population reconstructed by ShoRAH. Importantly, 47 (67%) of these were detected in variants constituting more than 1%. In contrast, during the acute phase of infection, less than 50% of substitutions were detected in variants present in more than 1% [13]. This suggests that, during the acute phase of infection, rare variants contribute more to the population diversity than during the chronic infection.

SSCP analysis, which has the sensitivity limit of 3%, revealed the presence of 9 bands. If the same cut-off value were applied in ShoRAH-based reconstruction, 7 haplotypes would have been identified. This fact suggests that the SSCP sensitivity might be higher than expected.

However, it must be stressed that the absolute SSCP band number may not reflect the haplotype number as it represents

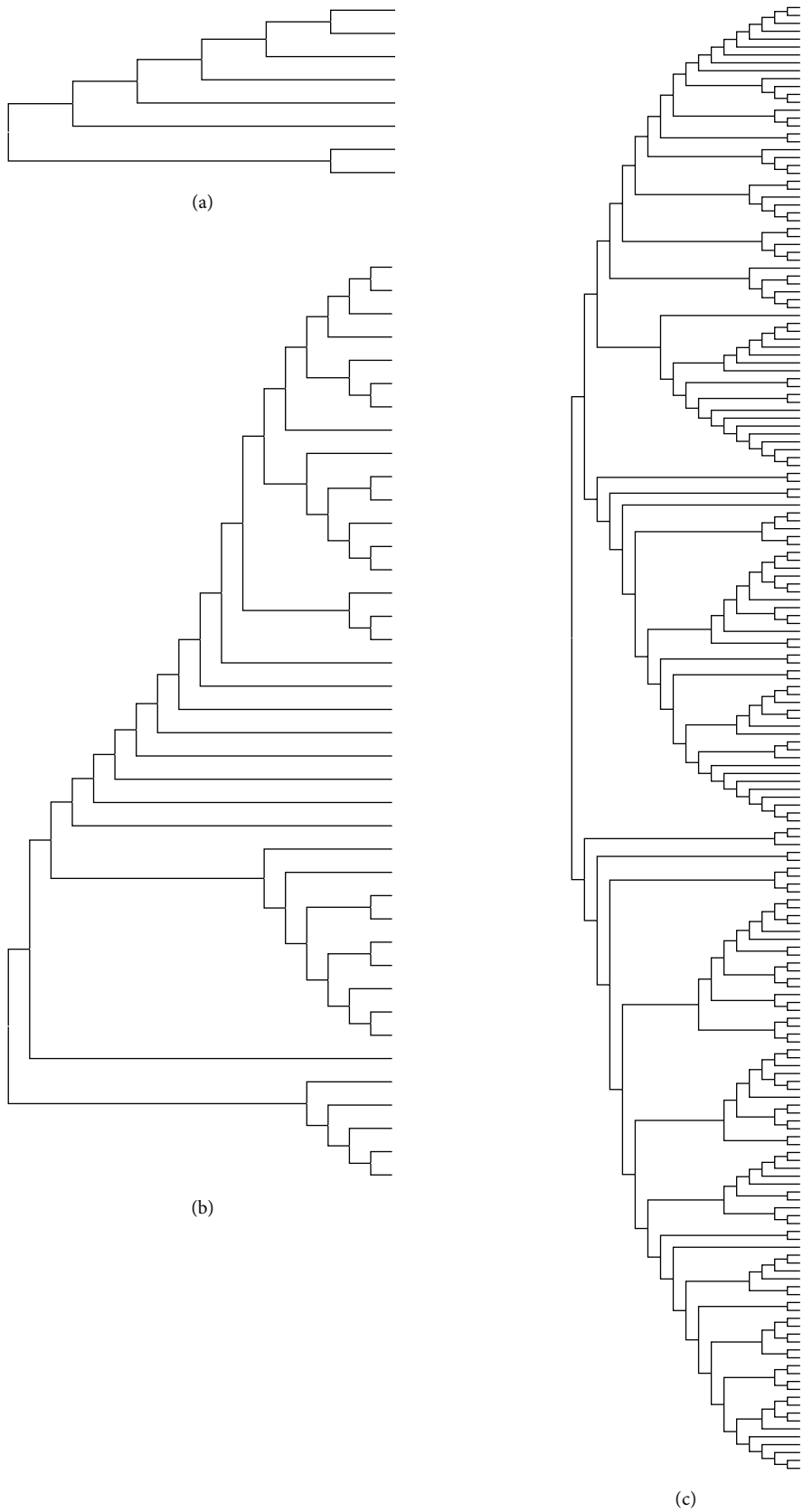


Figure 3: Molecular phylogenetic analysis of HVR1 populations inferred after the application of three different error correction methods. (a) Cut-off method >1% (I), (b) bidirectional coverage method (II), and (c) ShoRAH algorithm (III). The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model [23]. Evolutionary analyses were conducted using MEGA 5.0 [22].

resolved single DNA strands. Importantly, only sequencing provides information about the nucleotide sequence.

5. Conclusions

The newly available pyrosequencing technique opens a new approach to the analysis of complex viral genomes as it allows for detection of rare molecular variants. Better understanding of population genetics of complex viral populations seems crucial for understanding *quasispecies* phenomenon, viral evolution, and drug resistance.

In the evaluation presented here, we used ShoRAH to obtain the broadest spectrum of HVR1 variants while trying to preserve their reliability. The use of different sequencing platforms, the optimization of library preparation, and data analysis will further improve the reconstruction of viral *quasispecies*.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgment

This work was supported by Projects from Polish National Science Centre number NN 401 64 67 40 and 1 M24/PM12/12.

References

[1] M. Martell, J. I. Esteban, J. Quer et al., "Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution," *Journal of Virology*, vol. 66, no. 5, pp. 3225–3229, 1992.

[2] F. Kurbanov, Y. Tanaka, D. Avazova et al., "Detection of hepatitis C virus natural recombinant RF1_2k/1b strain among intravenous drug users in Uzbekistan," *Hepatology Research*, vol. 38, no. 5, pp. 457–464, 2008.

[3] P. Moreno, M. Alvarez, L. Lápez et al., "Evidence of recombination in Hepatitis C Virus populations infecting a hemophiliac patient," *Virology Journal*, vol. 6, article 203, 2009.

[4] J. M. Cuevas, M. Torres-Puente, N. Jiménez-Hernández et al., "Refined analysis of genetic variability parameters in hepatitis C virus and the ability to predict antiviral treatment response," *Journal of Viral Hepatitis*, vol. 15, no. 8, pp. 578–590, 2008.

[5] E. A. Duarte, I. S. Novella, S. C. Weaver et al., "RNA virus quasispecies: significance for viral disease and epidemiology," *Infectious Agents and Disease*, vol. 3, no. 4, pp. 201–214, 1994.

[6] T. Laskus, J. Wilkinson, J. F. Gallegos-Orozco et al., "Analysis of hepatitis C virus quasispecies transmission and evolution in patients infected through blood transfusion," *Gastroenterology*, vol. 127, no. 3, pp. 764–776, 2004.

[7] L. Barzon, E. Lavezzo, V. Militello, S. Toppo, and G. Palù, "Applications of next-generation sequencing technologies to diagnostic virology," *International Journal of Molecular Sciences*, vol. 12, no. 11, pp. 7861–7884, 2011.

[8] N. Beerenwinkel, H. F. Gunthard, V. Roth, and K. J. Metzner, "Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data," *Frontiers in Microbiology*, vol. 3, article 329, 2012.

[9] N. Beerenwinkel, "Ultra-deep sequencing for the analysis of viral populations," *Current Opinion in Virology*, vol. 1, no. 5, pp. 413–418, 2011.

[10] S. Guglietta, A. R. Garbuglia, V. Pacciani et al., "Positive selection of cytotoxic T lymphocytes escape variants during acute hepatitis C virus infection," *European Journal of Immunology*, vol. 35, no. 9, pp. 2627–2637, 2005.

[11] C. Di Lorenzo, A. G. Angus, and A. H. Patel, "Hepatitis C virus evasion mechanisms from neutralizing antibodies," *Viruses*, vol. 3, no. 11, pp. 2280–2300, 2011.

[12] A. Escobar-Gutierrez, M. Vazquez-Pichardo, M. Cruz-Rivera et al., "Identification of hepatitis C virus transmission using a next-generation sequencing approach," *Journal of Clinical Microbiology*, vol. 50, no. 4, pp. 1461–1463, 2012.

[13] R. A. Bull, F. Luciani, K. McElroy et al., "Sequential bottlenecks drive viral evolution in early acute hepatitis c virus infection," *PLoS Pathogens*, vol. 7, no. 9, Article ID e1002243, 2011.

[14] F. Bolcic, M. Sede, F. Moretti et al., "Analysis of the PKR-eIF2alpha phosphorylation homology domain (PePHD) of hepatitis C virus genotype 1 in HIV-coinfected patients by ultra-deep pyrosequencing and its relationship to responses to pegylated interferon-ribavirin treatment," *Archives of Virology*, vol. 157, no. 4, pp. 703–711, 2012.

[15] S. Fonseca-Coronado, A. Escobar-Gutierrez, K. Ruiz-Tovar et al., "Specific detection of naturally occurring hepatitis C virus mutants with resistance to telaprevir and boceprevir (protease inhibitors) among treatment-naive infected individuals," *Journal of Clinical Microbiology*, vol. 50, no. 2, pp. 281–287, 2012.

[16] O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel, "ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data," *BMC Bioinformatics*, vol. 12, article 119, 2011.

[17] I. Astrovskaya, B. Tork, S. Mangul et al., "Inferring viral quasispecies spectra from 454 pyrosequencing reads," *BMC Bioinformatics*, vol. 12, supplement 6, 2011.

[18] C. Quince, A. Lanzen, R. J. Davenport, and P. J. Turnbaugh, "Removing noise from pyrosequenced amplicons," *BMC Bioinformatics*, vol. 12, article 38, 2011.

[19] P. Skums, Z. Dimitrova, D. S. Campo et al., "Efficient error correction for next-generation sequencing of viral amplicons," *BMC Bioinformatics*, vol. 13, Supplement 10, p. S6, 2012.

[20] M. Gerotto, F. Dal Pero, S. Loffreda et al., "A 385 insertion in the hypervariable region 1 of hepatitis C virus E2 envelope protein is found in some patients with mixed cryoglobulinemia type 2," *Blood*, vol. 98, no. 9, pp. 2657–2663, 2001.

[21] T. Laskus, M. Radkowski, L. F. Wang, M. Nowicki, and J. Rakela, "Uneven distribution of hepatitis C virus quasispecies in tissues from subjects with end-stage liver disease: confounding effect of viral adsorption and mounting evidence for the presence of low-level extrahepatic replication," *Journal of Virology*, vol. 74, no. 2, pp. 1014–1017, 2000.

[22] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[23] K. Tamura and M. Nei, "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees," *Molecular Biology and Evolution*, vol. 10, no. 3, pp. 512–526, 1993.

[24] I. Malet, M. Belnard, H. Agut, and A. Cahour, "From RNA to quasispecies: a DNA polymerase with proofreading activity is

highly recommended for accurate assessment of viral diversity," *Journal of Virological Methods*, vol. 109, no. 2, pp. 161–170, 2003.

[25] J. Cline, J. C. Braman, and H. H. Hogrefe, "PCR fidelity of Pfu DNA polymerase and other thermostable DNA polymerases," *Nucleic Acids Research*, vol. 24, no. 18, pp. 3546–3551, 1996.

[26] A. Gilles, E. Meglécz, N. Pech, S. Ferreira, T. Malausa, and J. F. Martin, "Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing," *BMC Genomics*, vol. 12, article 245, 2011.

[27] B. B. Simen, J. F. Simons, K. H. Hullsiek et al., "Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes," *Journal of Infectious Diseases*, vol. 199, no. 5, pp. 693–701, 2009.

[28] X. Wu, T. Zhou, J. Zhu et al., "Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing," *Science*, vol. 333, no. 6049, pp. 1593–1602, 2011.

[29] A. Gonzalez-Serna, R. A. McGovern, P. R. Harrigan et al., "Correlation of the virological response to short-term maraviroc monotherapy with standard and deep-sequencing-based genotypic tropism prediction methods," *Antimicrobial Agents and Chemotherapy*, vol. 56, no. 3, pp. 1202–1207, 2012.

[30] A. D. Redd, A. Collinson-Streng, C. Martens et al., "Identification of HIV superinfection in seroconcordant couples in Rakai, Uganda, by use of next-generation deep sequencing," *Journal of Clinical Microbiology*, vol. 49, no. 8, pp. 2859–2867, 2011.

[31] G. T. Spear, M. Sikaroodi, M. R. Zariffard, A. L. Landay, A. L. French, and P. M. Gillevet, "Comparison of the diversity of the vaginal microbiota in HIV-infected and HIV-uninfected women with or without bacterial vaginosis," *Journal of Infectious Diseases*, vol. 198, no. 8, pp. 1131–1140, 2008.

[32] I. Vandenbroucke, H. Van Marck, P. Verhasselt et al., "Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications," *Biotechniques*, vol. 51, no. 3, pp. 167–177, 2011.

[33] O. Zagordi, R. Klein, M. Däumer, and N. Beerenwinkel, "Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies," *Nucleic Acids Research*, vol. 38, no. 21, pp. 7400–7409, 2010.

[34] N. Eriksson, L. Pachter, Y. Mitsuya et al., "Viral population estimation using pyrosequencing," *PLoS Computational Biology*, vol. 4, no. 4, Article ID e1000074, 2008.

[35] C. F. Wright, M. J. Morelli, G. Thébaud et al., "Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing," *Journal of Virology*, vol. 85, no. 5, pp. 2266–2275, 2011.