In Silico Biology 11 (2011/2012) 183–192
DOI 10.3233/ISB-2012-0453
IOS Press

Evaluation of viral heterogeneity using next-generation sequencing, end-point limiting-dilution and mass spectrometry

Z. Dimitrova[a,*], D.S. Campo[a], S. Ramachandran[a], G. Vaughan[a], L. Ganova-Raeva[a], Y. Lin[a], J.C. Forbi[a], G. Xia[a], P. Skums[a], B. Pearlman[b] and Y. Khudyakov[a]
[a]*Laboratory Branch, Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, GA, USA*
[b]*Center for Hepatitis C, Atlanta Medical Center, Atlanta, GA, USA*

Received 10 January 2012
Revised 20 June 2012
Accepted 6 July 2012

Abstract. Hepatitis C Virus sequence studies mainly focus on the viral amplicon containing the Hypervariable region 1 (HVR1) to obtain a sample of sequences from which several population genetics parameters can be calculated. Recent advances in sequencing methods allow for analyzing an unprecedented number of viral variants from infected patients and present a novel opportunity for understanding viral evolution, drug resistance and immune escape. In the present paper, we compared three recent technologies for amplicon analysis: (i) Next-Generation Sequencing; (ii) Clonal sequencing using End-point Limiting-dilution for isolation of individual sequence variants followed by Real-Time PCR and sequencing; and (iii) Mass spectrometry of base-specific cleavage reactions of a target sequence. These three technologies were used to assess intra-host diversity and inter-host genetic relatedness in HVR1 amplicons obtained from 38 patients (subgenotypes 1a and 1b). Assessments of intra-host diversity varied greatly between sequence-based and mass-spectrometry-based data. However, assessments of inter-host variability by all three technologies were equally accurate in identification of genetic relatedness among viral strains. These results support the application of all three technologies for molecular epidemiology and population genetics studies. Mass spectrometry is especially promising given its high throughput, low cost and comparable results with sequence-based methods.

1. Introduction

Hepatitis C virus (HCV) infection is a major cause of liver disease in the world. It is estimated that ~130 million people are infected with HCV globally [1]. HCV is a heterogeneous single-stranded (+) RNA virus that belongs to the Flaviviridae family. The HCV genome contains one large open reading frame that encodes a polyprotein which can be processed into ten mature proteins [23]. HCV causes chronic infection in 60–85% of infected adults. There is no vaccine against HCV and current anti-viral therapy is effective in only 40%–79% of chronically infected patients [14].

The HCV neutralizing epitope was mapped in hypervariable region 1 (HVR1) located at amino acid (aa) positions 384–410 in the structural protein E2. Sequence variation in HVR1 correlates with neutralization escape and is associated with viral persistence during chronic infection [7,13,16,22,24,31,33]. The HCV sequence studies mainly focus on the viral amplicon containing HVR1 to obtain a sample of sequences from which several population genetics parameters can be calculated [9,15,20,25,28]. Recent advances in sequencing methods allow for analyzing an unprecedented number of viral variants from infected patients and present a novel

*Corresponding author: Z. Dimitrova, Laboratory Branch, Division of Viral Hepatitis, Centers for Disease Control and Prevention, 1600 Clifton Rd, MS A-33, Atlanta, GA 30300, USA. Tel.: +1 404 639 2342; Fax: +1 404 639 1563; E-mail: izd7@cdc.gov.

opportunity for understanding viral evolution, drug resistance and immune escape [4,25,32,34]. Three strategies have been extensively used to analyze viral amplicons:

i. End-Point Limiting-Dilution (EPLD) [26] involves isolation of individual coexisting sequence variants from serum specimens using a limiting-dilution protocol, real-time PCR and sequencing. This method provides a cost-effective alternative to conventional cloning but, compared to the other methods considered in this article, is the most time-consuming and expensive.

ii. Next-Generation Sequencing (NGS) [18] results in a large number of viral variants. However, the increase in the amount of data increases the probability of observing erroneous reads. For pyrosequencing conducted using 454/Roche GS FLX, the mean error rate is 1.07% and error-free haplotypes represent from 10.09% to 67.57% of all reads, depending on the read length [11]. Originally, the emphasis was on obtaining the consensus sequence, provided that the depth of coverage easily allowed for retrieving the main true sequence and its most common polymorphisms irrespective of the suboptimal quality of numerous individual reads. However, analysis of viral amplicons is usually applied to biological tasks requiring in-depth characterization of viral populations and entails examination of individual error-free reads rather than consensus sequences. The main disadvantage of this strategy is, therefore, the extensive post-processing needed to correct errors and discriminate between artefacts and actual sequences.

iii. The mass spectrometry (MS) approach evaluated here is a high-performance comparative-sequencing strategy based on matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) analyses of complete base-specific cleavage reactions of a target RNA obtained from PCR fragments [30]. This strategy is the most cost-effective and can be used to process a great amount of samples in very little time. The approach is accurate and reproducible, can be applied to amplicons of up to 800 nucleotides in length, and the resulting mass pattern is suitable for identification of nucleotide polymorphisms representing >10% of the amplicon population. Although the technology is able to reliably obtain the actual nucleotide sequence of a single clone, in the case of viral amplicons there is a mixture of variants. Therefore, mass data represent a composite pattern rather than sequences of individual variants. Although this pattern can be used to accurately assess major viral genetic types [10], it cannot be used to assess many genetic parameters such as synonymous and non-synonymous changes, protein sequence, presence of stop codons, etc. and cannot be used to directly compare with sequences obtained previously. However, mass patterns contain very rich genetic information on the intra-host viral populations and are suitable for molecular epidemiological characterization of viral strains.

In the present work, these three strategies were used to assess diversity of intra-host HCV populations and genetic relatedness among HCV strains.

2. Results

2.1. Intra-host diversity

2.1.1. EPLD and NGS comparison

In this section, we compare sequences obtained using EPLD and NGS (Table 1). In general, the diversity of samples obtained with EPLD and NGS was similar. The number of unique sequences obtained with NGS was not significantly different from EPLD (PTMP, $p = 0.0664$) (Fig. 1A). The nucleotide diversity of samples obtained with NGS was not significantly different from EPLD (PTMP, $p = 0.0775$) (Fig. 1B). The average frequency of the major intra-host variant in NGS (62.04%) was significantly higher (PTMP, $p = 0.0017$) than average frequency of the major in EPLD (54.60%) (Fig. 1C).

The number of unique sequences found in both NGS and EPLD was rather low (Fig. 2A), with an average of 3.29 sequences per sample (S.E.M. = 0.36). However, when the frequencies of these sequences were considered, the level of agreement between the two methods was high, as measured by the overlap probability (Fig. 2B). Interestingly, we found that an important factor determining the level of agreement was the actual nucleotide diversity of the sample; samples with low EPLD nucleotide diversity showed higher overlap with NGS than high-diversity samples (Fig. 2C) ($r = -0.63$; $p = 0.0001$).

We also investigated whether the NGS error-correction algorithm was removing some of the expected sequences from the raw file. In order to establish this, we searched for all the EPLD sequences of a given sample in the raw NGS file. Although we found that the number of shared unique sequences in the raw file was higher than after error-correction (Average = 5.63; S.E.M. = 0.79), the overlap between both methods was still low. These results suggest that differences in stochastic sampling of variants

Table 1
Assessment of viral heterogeneity using EPLD, NGS and MS. n_seq, number of sequences; Unique, number of unique sequences; n_shar, number of unique sequences shared between NGS and EPLD; P_Overlap, probability of overlap; Dist_major, the Hamming distance between the major sequences found with EPLD and NGS (see methods for details in the measurement of overlap and diversity)

	n_seq		Unique		NGS and EPLD Overlap			Diversity		
Sample	NGS	EPLD	NGS	EPLD	n_shar	P_Overlap	Dist_major	NGS	EPLD	MS
1	3064	46	22	18	4	0.4264	0	0.0674	0.0306	4.4741
2	4133	13	12	5	3	0.6754	0	0.0058	0.0184	4.3561
3	4049	25	13	13	6	0.4216	0	0.0077	0.0118	4.2796
4	1342	43	16	31	2	0.0753	0	0.0693	0.0264	4.4141
5	1297	40	17	30	4	0.1535	0	0.012	0.0259	4.3836
6	1480	20	19	17	0	0	27	0.0102	0.0802	4.4419
7	2619	33	22	21	6	0.4101	2	0.0329	0.0373	4.4277
8	5197	68	21	27	3	0.4523	0	0.0042	0.0078	4.3337
9	4037	44	33	5	2	0.4976	0	0.0082	0.0069	4.2982
10	5411	32	11	10	3	0.7299	0	0.0043	0.0076	4.3689
11	5808	35	16	9	1	0.6621	0	0.004	0.0048	4.2955
12	3081	44	25	18	5	0.5772	0	0.0047	0.0055	4.3026
13	3866	38	28	19	6	0.2924	0	0.0115	0.0147	4.3482
14	2238	30	19	22	6	0.2573	0	0.0269	0.0284	4.3621
15	5393	48	2	7	1	0.8679	0	0.0038	0.0073	4.3488
16	2975	48	5	9	1	0.7751	0	0.0039	0.0059	4.359
17	2756	42	15	12	2	0.5781	0	0.0043	0.0059	4.3846
18	2395	48	11	16	2	0.6022	0	0.013	0.0174	4.3353
19	1211	54	14	32	6	0.2048	30	0.0619	0.0438	4.6323
20	1223	48	10	22	6	0.2995	0	0.007	0.009	4.4234
21	2691	48	17	15	4	0.5137	0	0.0074	0.0084	4.3807
22	4708	48	5	8	1	0.1577	1	0.0039	0.0051	4.4006
23	3829	48	7	15	4	0.7073	1	0.0055	0.0068	4.4504
24	3900	52	12	6	2	0.7574	0	0.0043	0.0047	4.3697
25	1909	48	10	7	1	0.673	0	0.0731	0.0046	4.3603
26	2543	38	3	7	1	0.7965	0	0.0038	0.0041	4.3983
27	1233	47	8	7	2	0.7334	0	0.0041	0.0053	4.5738
28	4290	48	16	19	7	0.6339	0	0.0076	0.0093	4.4268
29	3728	47	9	16	8	0.8209	0	0.006	0.0074	4.3405
30	5511	47	15	25	5	0.025	1	0.0095	0.01	4.4108
31	5209	48	2	13	1	0.7199	0	0.0038	0.0047	4.3577
32	2777	49	8	20	5	0.5189	10	0.026	0.0227	4.3262
33	4577	48	7	13	4	0.645	1	0.0057	0.0079	4.304
34	663	48	9	8	1	0.7846	0	0.0045	0.004	4.3811
35	3432	47	11	19	7	0.5825	0	0.0111	0.0119	4.3603
36	3957	47	3	11	1	0.7677	0	0.0038	0.0054	4.4322
37	3006	48	5	6	1	0.8589	0	0.0038	0.004	4.3593
38	2148	48	9	7	1	0.7984	0	0.0039	0.0053	4.368

rather than conservative cleaning were responsible for this low overlap. It must be noted that the frequency of these missed sequences in the raw file was extremely low, with an average frequency of 0.07% (S.E.M. = 0.01). Given that the low frequency of these missed variants is below the NGS sequencing error rate, their correct identification is unlikely.

In order to visualize the nucleotide differences between sequences obtained with these two techniques, we generated an MJN for selected samples. Figure 3A shows a sample with the highest agreement between NGS and EPLD. Figure 3B shows the sample with the highest diversity difference between NGS and EPLD. Diversity assessed with EPLD was higher than assessed with NGS in this sample. Figure 3C shows another sample with the high diversity difference between NGS and EPLD. In this case, diversity was higher when assessed by NGS than EPLD.

2.1.2. MS sample diversity

We explored several ways to translate the MS pattern of a sample into a measure of its underlying HVR1 diversity. This task is complex because the number of peaks is related to the number of different k-mers in the mixture, but this number has two different sources: (i) k-mer heterogeneity along the sequence; and (ii) new k-mers

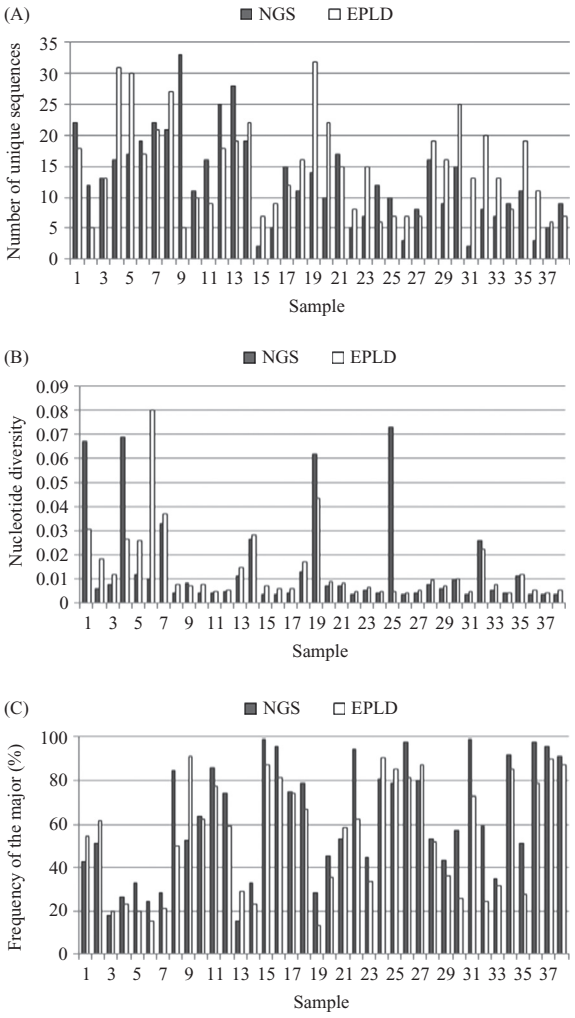

Fig. 1. Intra-host diversity by EPLD and NGS. (A) Number of unique sequences. (B) Nucleotide diversity. (C) Frequency of the major variant.

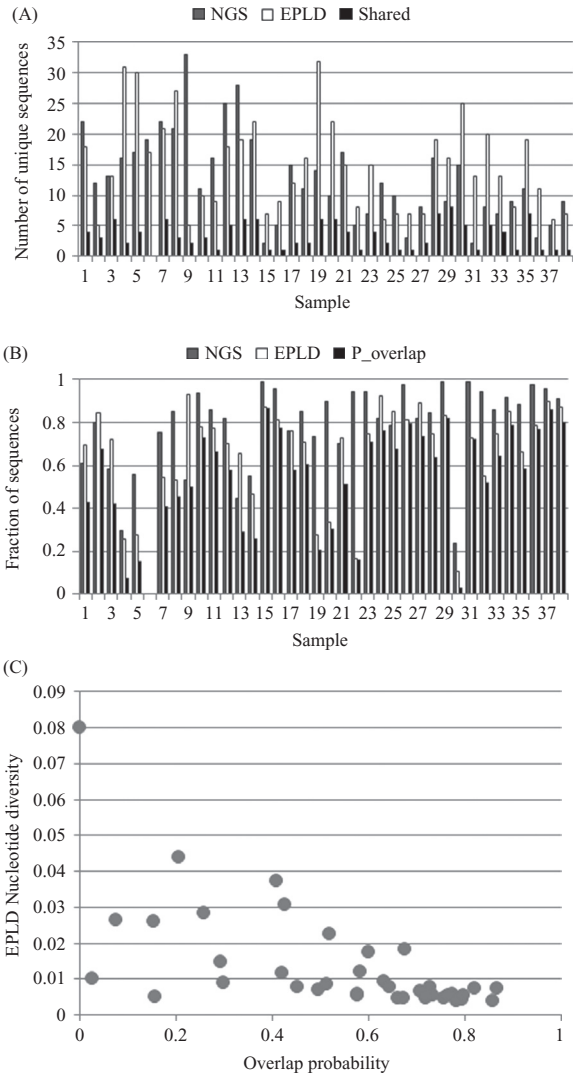

Fig. 2. Overlap between EPLD and NGS. (A) Number of shared sequences. (B) Fraction of sequences in the overlap. (C) Scatterplot of nucleotide diversity and overlap probability.

created by nucleotide variations at a given position. Our measure of MS diversity is based on the number of peaks and therefore confounds these two sources. However, genetic diversity is only associated with the second source. The correlation between MS diversity and sequence-based nucleotide diversity assessed by NGS and EPLD was 0.3730 ($p = 0.0211$) and 0.3816 ($p = 0.0181$), correspondingly.

2.2. Inter-host distances

Molecular epidemiological investigations are frequently based on estimation of genetic relatedness among viral strains. We evaluated performance of the three technologies in measuring distances between samples. The Mantel tests showed very high correlations among distance matrices: NGS and EPLD ($r = 0.9991$; $p = 0.0001$), NGS and MS ($r = 0.9395$; $p = 0.0001$), EPLD and MS ($r = 0.9414$; $p = 0.0001$). Figure 4 shows the scatterplots of inter-sample distances. There is a clear high agreement between NGS and EPLD estimates of genetic identity among HCV strains (Fig. 4A). However, inter-sample distances obtained using MS for genetically distant HCV strains were generally lower and more variable than distances estimated using NGS and EPLD

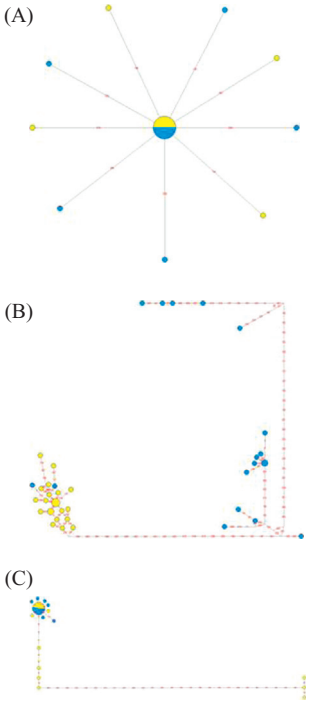

Fig. 3. MJN of EPLD and NGS sequences from selected samples. Yellow represents NGS sequences and blue represents EPLD sequences. Node size is proportional to the frequency of the sequences in each method. Nucleotide changes are shown in red. (A) Sample 37; (B) Sample 6; (C) Sample 25.

(Fig. 4B and 4C). This difference is probably associated with the binary representation of MS patterns implemented here and k-mer structure of the MS data obtained from short amplicons.

In order to visualize the relatedness among samples, we also built an MDS using data obtained by each method (Fig. 5). The MDS mapping showed correct subgenotype separation by all methods. There is a substantial concordance between distance estimates obtained by NGS and EPLD. Consistent with the scatterplot of distances shown in Fig. 4, the MDS distances between HCV strains from subgenotype 1a and 1b are smaller for MS than for NGS and EPLD. A set of samples were epidemiologically related (samples 21 to 38) and we expected them to form a cluster of closely related variants. The similarity tree based on sequence data is shown of Fig. 6A, showing that all samples formed a tight cluster for both NGS and EPLD. All these samples are also part of a single cluster in the similarity tree based on MS data (Fig. 6B).

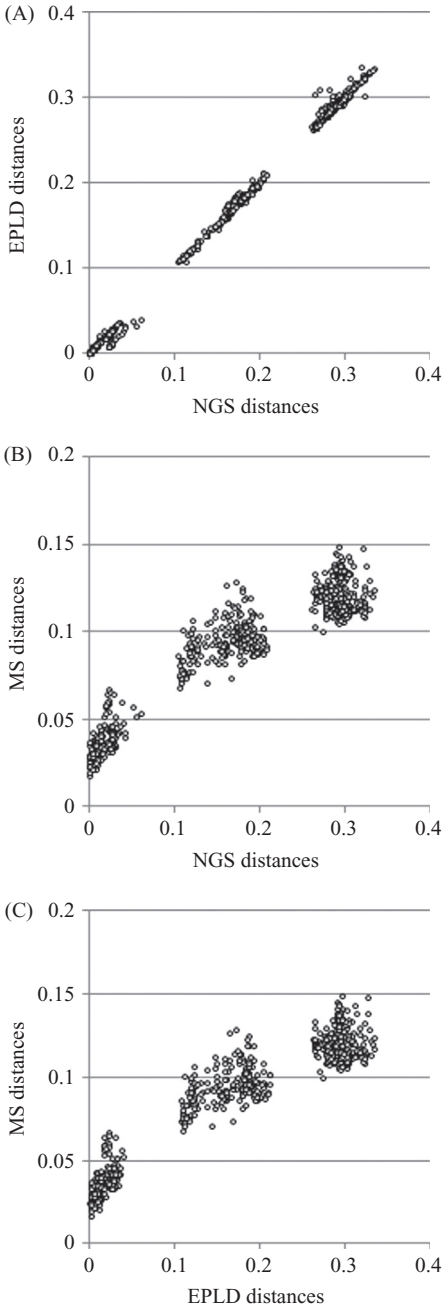

Fig. 4. Scatterplot of inter-host distances. (A) NGS and EPLD. (B) NGS and MS. (C) EPLD and MS.

3. Discussion

The data obtained in this study show that the MS assessment of HCV intra-host diversity differs from assessments obtained using sequences. This finding

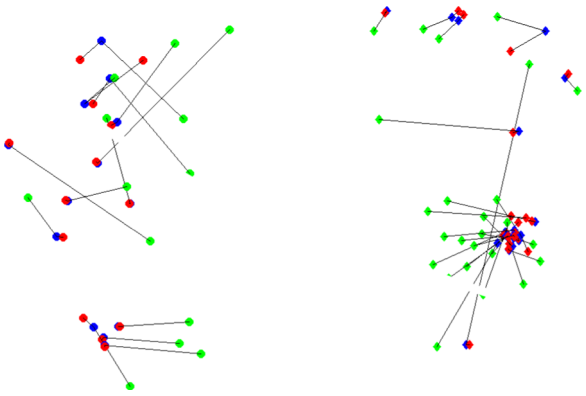

Fig. 5. Two-dimensional representation of the distance matrices by MDS. The 2D-coordinates of MS and NGS data were rotated to maximize agreement to the EPLD coordinates. Subgenotype 1a strains are shown as circles and subgenotype 1b strains as rhombi. NGS samples are shown in red, EPLD samples in blue and MS in green. Black lines join results of the three technologies that correspond to the same sample.

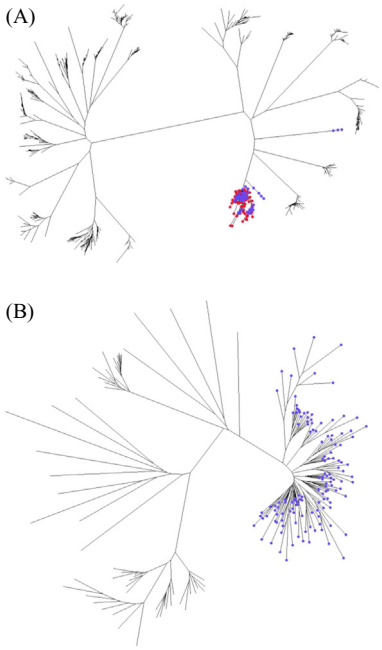

Fig. 6. Cluster of related samples in a similarity tree. (A) NGS (blue) and EPLD (red). (B) MS.

reflects the essential difference in the data structure produced by MS compared to EPLD and NGS and suggests the development of novel computational strategies for measuring genetic parameters of viral populations from MS patterns.

Although NGS produces a significantly greater number of raw reads than EPLD, which usually produces dozens of sequences, there is almost no difference in diversity of unique variants assessed by both technologies after the error correction of NGS data. The agreement between the methods in detection of unique sequence variants was low but improved considerably after taking frequencies into consideration. We found that an important factor determining the level of agreement between the two methods was the actual nucleotide diversity of the sample, suggesting that stochastic sampling is responsible for the differences found between the two methods. Finally, assessments of genetic distances of HCV strains by these two technologies are virtually identical.

4. Conclusion

Molecular epidemiological investigations are frequently based on the degree of genetic relatedness among the viral strains harbored by different patients. As shown in this study, all three technologies provide equally accurate assessment of genetic distances for HCV strains. However, NGS and MS are significantly more cost- effective, less time-consuming and more amenable to tracking viral variants and detecting transmission events in the course of epidemiological investigation.

5. Materials and methods

5.1. Sample description

HCV strains (n = 38) belonged to subgenotype 1a (n = 13) and subgenotype 1b (n = 25). Ethical Review and Informed Consent approval was granted by the institutional review boards at Atlanta Medical Center and Centers for Disease Control and Prevention.

Total nucleic acids from the specimens were extracted from serum by using the Roche MagNA Pure LC instrument and the MagNA Pure LC Total Nucleic Acid Isolation Kit (Roche Diagnostics, Mannheim, Germany), and eluted with 50 ul of buffer according to manufacturer's instructions. PCR quantification was determined by COBAS® AmpliPrep/COBAS® TaqMan® HCV Test (Roche Diagnostics, Mannheim, Germany) by Reference Lab of DVH, genotyping was done by VERSANT®HCV Genotype 2.0 Assay (LiPA) (Innogenetic NV, Gent, Belgium). RNA was precipitated and reverse-transcribed using both random and specific primers as previously described [26]. The junction E1/E2 region (309 nucleotides), which contains the HVR1

region, was amplified using the nested PCR protocol described in [26].

5.2. EPLD experimental conditions

The EPLD protocol allows for sensitive detection of viral variants [26]. The average number of clones amplified was 43.5 (Standard Error of the mean, S.E.M. = 1.59). The amplifications were carried out as described in [26]. Sequencing reactions were performed using the BigDye v3.1 chemistry sequencing kit (Applied Biosystems, Foster City, CA), and products were sequenced using an automated sequencer (3130xl Genetic Analyzer, Applied Biosystems). The sequence files were aligned with Muscle [8] and clipped to a size of 264 nucleotides.

5.3. NGS experimental conditions

The amplicons generated during first-round PCR were used as templates for a nested PCR using hybrid primers composed of NGS primer adaptors, multiple identifiers and specific sequences complementary to the HCV genome. This allowed for multiplexing and downstream NGS procedure. Resulting amplicons were quantified using the Picogreen kit (Invitrogen, Carlsbad, CA). Integrity of each fragment was evaluated using Bioanalyzer 2100 (Agilent, Santa Clara, CA). PCR products were pooled and subjected to NGS using the GS FLX Titanium Series Amplicon kit in a 454/Roche GS FLX instrument. The initial reads were processed by matching to the corresponding identifier. Low quality reads were removed using the GS Run Processor v2.3 (Roche, 2010). The NGS files were post-processed with our Empirical Threshold error correction algorithm [29], which shows very high accuracy in finding true haplotypes, removing false haplotypes and estimating the frequency of true ones. The Empirical Threshold algorithm includes a calibration step using sequence reads from single-clone HVR1 samples, estimating an empirical frequency threshold for indels and haplotypes, and also correcting homopolymer errors using sequence alignment [29].

5.4. MS experimental conditions

5.4.1. HVR1 amplicon sampling, transcription and cleavage

To obtain a profile of HVR1 heterogeneity we sampled the first-round PCR products 8-12 times without limiting dilution to serve as a template for second-round reaction. The second-round products were then subjected to the MassCLEAVE™ protocol. The mass data were collected on a Compact Analyzer (Sequenom, San Diego, USA). Samples were processed in parallel in 384-well microtiter plates. All PCR set up, SAP, post-PCR base-specific cleavage reactions (MassCLEAVE, Sequenom, San Diego, USA) and post-cleavage treatments were performed using the automated liquid handler Biomek 3000 (Beckman Coulter, Fullerton, CA). All products were then printed on 384-SpectroCHIPs (Sequenom, San Diego, USA) by a Nanodispenser (Samsung) and analyzed by MALDI-TOF MS as described in [30].

5.4.2. MS pattern analysis

The data collected by MALDI-TOF MS for each specimen represent peak patterns of masses of all nucleic acid fragments of a particular amplicon generated by four separate base-specific RNaseA cleavage reactions. Mass-peak lists were exported from iSEQ and results from different plates were merged according to the existing algorithm provided by Sequenom (San Diego, USA). The process results in one output file that contains the mass size and intensity data across all experiments assigned into discrete bins.

5.5. Data analysis

5.5.1. Comparison between NGS and EPLD

We quantified the number of unique sequences shared between NGS and EPLD. This quantification can only be done after the following processing: (i) pairwise alignment of each sequence in EPLD against all other sequences in NGS; (ii) For each sequence in EPLD, choose its closest sequence in NGS; (iii) Clip the gapped ends of the longest read; (v) Calculating hamming distance and establishing identity. We also computed the probability that a sequence found by NGS will also be found by EPLD (sum of the shared sequences' frequencies in NGS, multiplied by the sum of the shared sequences' frequencies in EPLD).

5.5.2. Diversity

The extent of sequence heterogeneity in each sample was examined by unbiased estimates of nucleotide diversity, calculated according to Nei and Li [21]. For MS, each pattern was transformed into a binary vector where the presence or absence of a peak was identified

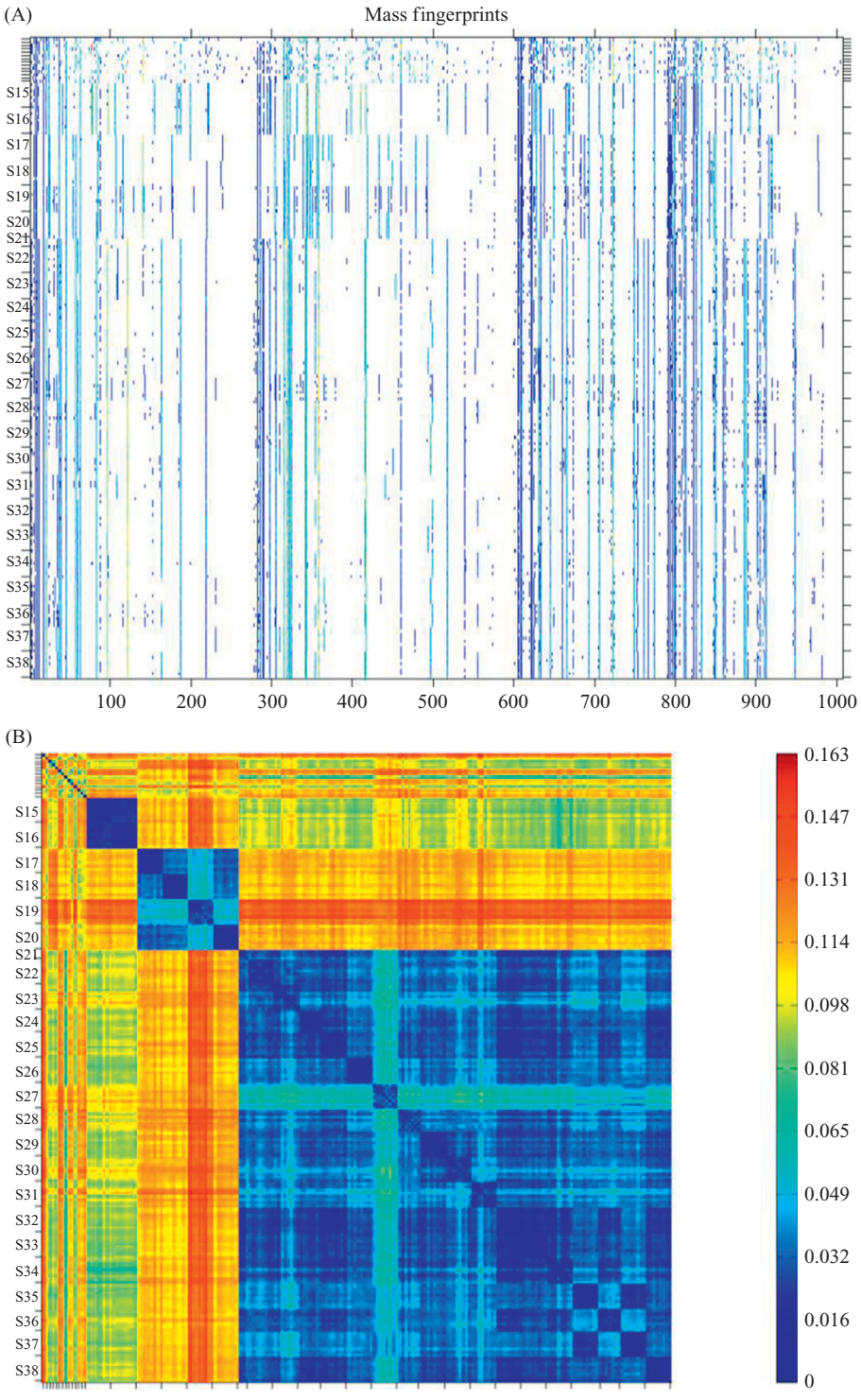

Fig. 7. MS Patterns for 38 samples. (A) List of peaks. (B) Color map of the matrix of distances among MS patterns. Distances are shown in a scale from low (blue) to high (red).

with 1 or 0, correspondingly; then, the natural logarithm of the number of peaks was calculated.

5.5.3. Permutation test for matched pairs (PTMP)

We tested the differences in diversity between methods by means of a PTMP [5,19]. PTMP is a permutation version of the paired t-test, a non-parametric test for testing the hypothesis of no difference between two groups of paired samples. Permutation tests allow for deriving the exact probabilities associated with a test statistic, rather than approximating them from common probability distributions, such as t, F and Chi square [6]. The population distribution is frequently unknown, and assuming a normal distribution (an underlying assumption in parametric testing) is inappropriate for many biological datasets, which often are skewed, discontinuous, and multi-modal. We used the PTMP implemented in BLOSSOM [5] to evaluate 100,000 permutations.

5.5.4. Distances among samples

For each method, the distance between HCV strains was measured. In the case of EPLD and NGS, the average Hamming distance between haplotypes of two samples was calculated and weighted by the frequency of each haplotype. In the case of MS, each peak pattern was transformed into a binary vector as described above (Fig. 7A) and then the Hamming distance between pairs of patterns was calculated (Fig. 7B).

5.5.5. Mantel test

The agreement between distance matrices was measured using Mantel test [17]. In this test, the null hypothesis is that distances in a matrix A are independent of the distances, for the same objects, in another matrix B. The significance is assessed by a randomization procedure (n = 10,000), in which the original value of the statistic is compared with the distribution found by randomly reallocating the order of the elements in one of the matrices [3].

5.5.6. Median-Joining network (MJN)

We applied MJN to visualize nucleotide differences between the sequences obtained with NGS and EPLD from the same sample. The MJN method begins by computing the minimum spanning trees (a graph that connects all the sequences with the minimum necessary total branch length), following which all the constructed graphs are combined within a single (reticulate) network. MJN were performed using NETWORK 4.0 [2].

5.5.7. Multidimensional scaling (MDS)

We applied MDS to visualize the distances among samples. The metric stress was calculated to evaluate how closely a particular configuration in the MDS plot reproduces the observed distance matrix. The MDS of the three methods were rotated to maximize agreement by means of Procrustes analysis [12]. All calculations and statistical analyses were performed with Matlab (Mathworks, Natick, MA).

5.5.8. Neighbor-joining tree

We applied the Neighbor-joining method [27] to calculate similarity trees based on hamming distances among samples.

Author contributions

ZD, DSC and YK designed the study. BP contributed samples and materials. SR, GV, LGR, YL and JCF performed laboratory analysis. ZD, DSC, PS and GX performed data analysis. ZD, DSC and YK wrote the manuscript.

References

[1] Alter, M., *Epidemiology of hepatitis C virus infection*. World J Gastroenterol, 2007. **13**(17): p. 2436–2441.

[2] Bandelt, H., P. Forster, and A. Rohl, *Median-Joining Networks For Inferring Intraspecific Phylogenies*. Mol Biol Evol, 1999. **16**(1): p. 37–48.

[3] Bonnet, E. and d. Van, Peer, Y., *ZT: A Sofware Tool for Simple and Partial Mantel Tests*. Journal of Statistical Software, 2002. **7**(10): p. 1–12.

[4] Bull, R.A., et al., *Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection*. PLoS Pathog, 2011. **7**(9): p. e1002243.

[5] Cade, B. and J. Richards, *User Manual For BLOSSOM Statistical Software*. Midcontinent Ecological Science Center US Geological Survey Fort Collins, Colorado, 2001.

[6] Cai, L., *Multi-response Permutation Procedure as An Alternative to the Analysis of Variance: An SPSS Implementation*. Department of Psychology, University of North Carolina, 2004.

[7] Eckels, D., et al., *Identification of antigenic escape variants in an immunodominant epitope of hepatitis C virus*. Int Immunol, 1999. **11**(4): p. 577–583.

[8] Edgar, R., *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Nucleic Acids Res, 2004. **32**(5): p. 1792–1797.

[9] Farci, P., et al., *The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies*. Science, 2000. **288**(5464): p. 339–344.

[10] Ganova-Raeva, L., et al., *Robust hepatitis B virus genotyping by mass spectrometry*. J Clin Microbiol, 2010. **48**(11): p. 4161–4168.

[11] Gilles, A., et al., *Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing*. BMC Genomics., 2011. **12**(1): p. 245.
[12] Gower, J. and G. Dijksterhuis, *Procrustes Problems*. 2004, New York: Oxford University Press 248.
[13] Isaguliants, M., *Hepatitis C virus clearance: the enigma of failure despite an impeccable survival strategy*. Curr Pharm Biotechnol, 2003. **4**(3): p. 169–183.
[14] Jacobson, I.M., et al., *Telaprevir for previously untreated chronic hepatitis C virus infection*. N Engl J Med, 2011. **364**(25): p. 2405–2416.
[15] Liu, L., et al., *Acceleration of hepatitis C virus envelope evolution in humans is consistent with progressive humoral immune selection during the transition from acute to chronic infection*. J Virol, 2010. **84**(10): p. 5067–5077.
[16] Lopez-Labrador, F., et al., *Genetic variability of hepatitis C virus NS3 protein in human leukocyte antigen-A2 liver transplant recipients with recurrent hepatitis C*. Liver Transpl, 2004. **10**(2): p. 217–227.
[17] Mantel, N. and R. Valand, *A technique of nonparametric multivariate analysis*. Biometrics, 1970. **26**: p. 547–558.
[18] Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors*. Nature, 2005. **437**(7057): p. 376–380.
[19] Mielke, P. and K. Berry, *Permutation methods: A distance function approach*. 2001, New York: Springer-Verlag.
[20] Mondelli, M.U., et al., *Antibody responses to hepatitis C virus hypervariable region 1: evidence for cross-reactivity and immune-mediated sequence variation*. Hepatology, 1999. **30**(2): p. 537–545.
[21] Nei, M. and W. Li, *Mathematical model for studying genetic variation in terms of restriction endonucleases*. Proc Natl Acad Sci U S A., 1979. **76**: p. 5269–5273.
[22] Pavio, N. and M. Lai, *The Hepatitis C Virus Persistence: How To Evade The Immune System?* J Biosci, 2003. **3**: p. 287–304.
[23] Penin, F., et al., *Structural biology of hepatitis C virus*. Hepatology., 2004. **39**(1): p. 5–19.
[24] Puntoriero, G., et al., *Towards a solution for hepatitis C virus hypervariability: mimotopes of the hypervariable region 1 can induce antibodies cross-reacting with a large number of viral variants*. EMBO J, 1998. **17**(13): p. 3521–3533.
[25] Ramachandran, S., et al., *Temporal Variations in the Hepatitis C Virus Intra-Host Population During Chronic Infection*. J virol, 2011.
[26] Ramachandran, S., et al., *End-point limiting-dilution real-time PCR assay for evaluation of hepatitis C virus quasispecies in serum: performance under optimal and suboptimal conditions*. J Virol Methods, 2008. **151**(2): p. 217–224.
[27] Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees*. Mol Biol Evol, 1987. **4**: p. 406–425.
[28] Sheridan, I., et al., *High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression*. J Virol, 2004. **78**(7): p. 3447–3454.
[29] Skums, P., et al., *Efficient Error Correction of High-throughput Viral Sequencing*. BMC Bioinformatics, 2011.
[30] Stanssens, P., et al., *High-throughput MALDI-TOF discovery of genomic sequence polymorphisms*. Genome Res, 2004. **14**(1): p. 126–133.
[31] Van Doorn, L., et al., *Sequence evolution of the hypervariable region in the putative envelope region E2/NS1 of hepatitis C virus is correlated with specific humoral immune responses*. J Virol, 1995. **69**(2): p. 773–778.
[32] Wang, G., et al., *Hepatitis C virus transmission bottlenecks analyzed by deep sequencing*. J Virol, 2010. **84**(12): p. 6218–6228.
[33] Wang, H., et al., *Sequence variation in the gene encoding the nonstructural 3 protein of hepatitis C virus: evidence for immune selection*. J Mol Evol, 2002. **54**(4): p. 465–73.
[34] Zagordi, O., et al., *Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies*. Nucleic Acids Research, 2010. **38**(21): p. 7400–7409.