Ultradeep Pyrosequencing of Hepatitis C Virus to Define Evolutionary Phenotypes

Brendan A. Palmer[1], Zoya Dimitrova[2], Pavel Skums[2], Orla Crosbie[3],
Elizabeth Kenny-Walsh[3] and Liam J. Fanning[1, *]

[1]Molecular Virology Diagnostic & Research Laboratory, Department of Medicine, University College Cork, Cork, Ireland; [2]Division of Viral Hepatitis, Centers of Disease Control and Prevention, Atlanta, Georgia, USA; [3]Department of Gastroenterology, Cork University Hospital, Cork, Ireland
*For correspondence: l.fanning@ucc.ie

[Abstract] Analysis of hypervariable regions (HVR) using pyrosequencing techniques is hampered by the ability of error correction algorithms to account for the heterogeneity of the variants present. Analysis of between-sample fluctuations to virome sub-populations, and detection of low frequency variants, are unreliable through the application of arbitrary frequency cut offs. Cumulatively this leads to an underestimation of genetic diversity. In the following technique we describe the analysis of Hepatitis C virus (HCV) HVR1 which includes the E1/E2 glycoprotein gene junction. This procedure describes the evolution of HCV in a treatment naïve environment, from 10 samples collected over 10 years, using ultradeep pyrosequencing (UDPS) performed on the Roche GS FLX titanium platform (Palmer *et al.*, 2014). Initial clonal analysis of serum samples was used to inform downstream error correction algorithms that allowed for a greater sequence depth to be reached. PCR amplification of this region has been tested for HCV genotypes 1, 2, 3 and 4.

Keywords: Ultradeep pyrosequencing, Virus, Quasispecies, Hypervariability

[Background] Analysis of UDPS datasets derived from virus amplicons frequently relies on software tools that are not optimized for amplicon analysis, assume random incorporation of sequencing mutations and are focused on finding true sequences rather than false variants. These difficulties are further complicated by the presence of hypervariable regions present in RNA virus genomes. Many studies utilizing UDPS look to overcome these issues by applying arbitrary frequency cut offs to the data, resulting in the loss of minor variants. Here, a temporally matched clonal dataset, together with an error correction methodology designed to overcome the problems outlined, facilitated the retention of valuable sequence information.

Materials and Reagents

1. 1.5 ml tube (SARSTEDT, catalog number: 72.690.001)
2. 200 µl MicroAmp® PCR tube (Thermo Fisher Scientific, Applied Biosystems™, catalog number: N8010840)
3. Clean stainless steel blade

4. One Shot® TOP10 Competent Cells (Thermo Fisher Scientific, Invitrogen™, catalog number: C404003)
5. QIAamp® Viral RNA mini kit (QIAGEN, catalog number: 52904)
6. Random primer (Promega, catalog number: C1181)
7. Deoxynucleoside triphosphate (dNTP's, 100 mM) set, PCR grade (Roche Molecular Systems, catalog number: 11969064001)
8. AMV reverse transcriptase (Promega, catalog number: M5101)
9. RNasin® Ribonuclease inhibitor (Promega, catalog number: N2511)
10. Outer-forward primer: 5'- ATGGCATGGGATATGAT -3' (10 pmol/µl, Eurofins)
11. Outer-reverse primer: 5'- AAGGCCGTCCTGTTGA -3' (10 pmol/µl, Eurofins)
12. Inner-forward primer: 5'- GCATGGGATATGATGATGAA -3' (10 pmol/µl, Eurofins)
13. Inner-reverse primer: 5'- GTCCTGTTGATGTGCCA -3' (10 pmol/µl, Eurofins)
14. Pwo DNA polymerase (5 U/µl,) including 10x reaction buffer (- $MgSO_4$) and $MgSO_4$ stock solution (25 mM) (Roche Molecular Systems, catalog number: 11644955001)
15. dH_2O (Sigma-Aldrich, catalog number: W4502)
16. Sybr safe DNA gel stain (Thermo Fisher Scientific, Invitrogen™, catalog number: S33102)
17. Agarose (Sigma-Aldrich, catalog number: A9539)
18. GeneRuler 100 bp Plus DNA ladder (Thermo Fisher Scientific, Thermo Scientific™, catalog number: SM0323)
19. Gel extraction kit (QIAGEN, catalog number: 28704)
20. CloneJet PCR Cloning Kit (Thermo Fisher Scientific, Thermo Scientific™, catalog number: K1231)
21. GeneJet Plasmid Miniprep Kit (Thermo Fisher Scientific, Thermo Scientific™, catalog number: K0503)
22. Trizma® base (Sigma-Aldrich, catalog number: T1503)
23. Acetic acid glacial (BDH Laboratory Supplies, catalog number: 10001CU)
24. Ethylenediaminetetraacetic acid solution 0.5 M (EDTA) (Sigma-Aldrich, catalog number: 03690)
25. 1x TAE (see Recipes)

Equipment

1. PCR thermal cycler (Thermo Fisher Scientific, Applied Biosystems™, model: Applied Biosystems® 2720)
2. BioPhotometer (Eppendorf, http://arboretum.harvard.edu/wp-content/uploads/Biophotometer-manual.pdf)
3. Water bath (JULABO, model: SW22)
4. Orbital shaker incubator (Grant, model: ES-80)
5. Ultraviolet transilluminator (UVP, model: TMW-20)

Software

1. SFFFile tools (Roche Molecular Systems)
2. k-mer error correction (KEC) and empirical threshold (ET) (Skums *et al.*, 2012)
3. MEGA 6.0 (Tamura *et al.*, 2013)

Procedure

A. RNA extraction and cDNA generation

1. Whole patient serum, surplus to diagnostic testing requirements and with a mean viral titer of 6 HCV RNA $\log_{10}$ IU/ml, was used as the starting material.
2. RNA was extracted from 140 µl of serum using QIAamp® Viral RNA mini kit according to the manufacturer's instructions into 1.5 ml RNase free tubes and a final volume of 50 µl.
3. 11 µl of extracted viral RNA was mixed with 1 µl (0.5 µg) random primer.
4. Samples were incubated at 75 °C for 10 min.
5. To this was added a master mix which contained 2 µl (80 mM) dNTP mix, 1 µl (10 U) AMV reverse transcriptase, 1 µl (40 U) RNasin, 4 µl AMV reaction buffer.
6. cDNA generation took place at 42 °C for 60 min, followed by 94 °C for 3 min.
7. Samples were kept at 4 °C until required.

B. Nested PCR to amplify the HCV E1/E2 gene junction

1. Prepare the primary PCR master mix to a final volume of 45 µl:

Outer-forward primer:	1.5 µl
Outer-reverse primer:	1.5 µl
10x reaction buffer (- $MgSO_4$):	5 µl
dNTP mix:	1 µl
$MgSO_4$ stock solution:	3 µl
Pwo:	0.5 µl
PCR grade water:	32.5 µl

2. 5 µl of cDNA is then added to the master mix.
3. 1° PCR cycle parameters:
   a. Initial denaturation: 3 min at 94 °C
   b. Cycle conditions (repeat for 35 cycles):
      Denaturation: 15 sec at 94 °C
      Annealing: 30 sec at 51 °C
      Extension: 30 sec at 72 °C
   c. Final extension: 7 min at 72 °C
4. Keep sample at 4 °C until required.
5. Prepare master mix for secondary PCR to a final volume of 46 µl:

Inner-forward primer:	1.5 µl
Inner-reverse primer:	1.5 µl
10x reaction buffer (- $MgSO_4$):	5 µl
dNTP mix:	1 µl
$MgSO_4$ stock solution:	2 µl
Pwo:	0.5 µl
PCR grade water:	34.5 µl

6. 4 µl of primary PCR sample is then added to the master mix.
7. 2° PCR cycle parameters:
   a. Initial denaturation: 3 min at 94 °C
   b. Cycle conditions (repeat for 35 cycles):
      Denaturation: 15 sec at 94 °C
      Annealing: 30 sec at 53 °C
      Extension: 30 sec at 72 °C
   c. Final extension: 7 min at 72 °C
8. Samples were kept at 4 °C until required.
9. To ensure that the initial amount of the template was not limiting, 1:100 dilution of the viral RNA was prepared which, when used as the starting template for nested PCR as described, should yield an amplicon visualized by gel electrophoresis for each sample.

C. Preparation of samples for pyrosequencing
1. Two 2% TAE agarose gels were poured, one containing Sybr safe DNA gel stain and one without.
2. Once set, the gels were split in two, with one half of the gel containing the gel stain joined with the second gel without gel stain.
3. The 50 µl amplicon sample was split in two (10 µl and 40 µl) and resolved on the above gel. The 10 µl sample was stained using Sybr safe, while the 40 µl sample was not stained and went forward for downstream procedures. The resultant amplicon in this instance was 320 bp (Figure 1).
4. The region of the gel containing the unstained band (40 µl sample) was cut out using a clean stainless steel blade using the stained 10 µl sample as a positioning guide and transferred to a clean 1.5 ml tube.
5. The amplicon was gel extracted using a gel extraction kit according to the manufacturer's instructions.
6. Extracted amplicons were quantified using a BioPhotometer.
7. Samples were prepared in equimolar concentrations and diluted to a final concentration of 1 x 10^7 molecules/ml.
8. Pyrosequencing was outsourced to Roche 454 Life Sciences (Brandford, CT, USA).

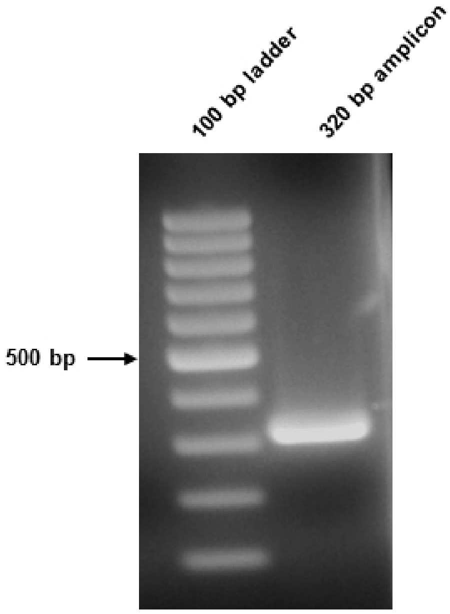

Figure 1. Amplicon visualization. Successful amplification of the 320 bp amplicon was confirmed following agarose gel electrophoresis. 10 µl of the 2° PCR sample was loaded.

D. Clonal analysis

1. Purified amplicons were cloned using CloneJet PCR Cloning Kit and transformed into One Shot® TOP10 Competent Cells using the manufacturer's instructions using a molar ratio of 3:1 insert to vector.
2. 20 clones per sample were generated.
3. Plasmids were purified using GeneJet Plasmid Miniprep Kit as per manufacturer's instructions.
4. Sequencing of E1/E2 inserts was performed by Eurofins.
5. All trace files were inspected to exclude sequences where double peaks or regions of ambiguous sequence were present.

E. Data handling and error correction

1. The raw sff data files were managed using SFFFile tools.
2. Low-quality reads and reads shorter than 90% of the expected amplicon lengths were removed.
3. Phylogenetic separation of the clonal data using a general time-reversible model with gamma-distributed and invariant sites (GTR+G+I) using MEGA 6.0 (Tamura *et al.*, 2013).
4. Main branches with bootstrap values (of 1,000 resamplings) > 85 were categorised as (sub-)lineages (Palmer *et al.*, 2014).
5. Two 24-bp motifs, that defined the HVR1 amino acid profile of each (sub-)lineage, were subsequently applied to the sequence analysis pipeline. The first 15-bp of the motif span the conserved 3'-end of E1. The remaining 9-bp include the first three amino acids of the HVR1 at the 5'-end of E2.

6. The overall number of motifs used reflected the observed changes in the dominant HVR1 over time. For each (sub-)lineage, two motif reference sequences were deemed sufficient.
7. To increase the sensitivity of the sequencing error correction algorithms (KEC-ET), the UDPS data was partitioned according to the presence of corresponding motifs.
8. In order to ensure the quality of the analyzed data and the absence of PCR and sequencing chimeras, reads that had more than a 3 bp difference from the best-matching sequence from this motif set were removed.
9. KEC consists of the three stages
    a. In stage 1, the set of k-mers (substring of fixed length k) of reads from the processed data set is calculated and the distribution of frequencies of k-mers is analyzed. The error threshold is calculated as the minimal frequency of k-mers separating two different distributions.
    b. In stage 2, k-mers with frequencies lower than the error threshold are considered erroneous and are used to identify and correct the errors. The corrections are based on an analysis of different factors, including the length of a segment of consecutive erroneous k-mers, the sequences of nucleotides at the end of that segment, and the frequencies of the similar correct k-mers. The procedure of error correction is repeated iteratively i times.
    c. In stage 3, the reads containing k-mers that were not corrected in stage 2 are discarded.
10. The following parameters of KEC were used: $k = 25$ and $i = 3$.

Data analysis

A more complete description of the data handling and error correction procedure can be found in the original article, http://jvi.asm.org/content/88/23/13709.short (Palmer *et al.*, 2014).

Notes

All serum samples were genotyped and quantified by the Molecular Virology Diagnostic & Research Laboratory at Cork University Hospital, Cork, Ireland. https://www.ucc.ie/en/meddept/people/liam-fanning/mvdrl/

Recipes

1. 1x TAE
   4.84 g Tris base
   1.15 ml acetic acid glacial
   2 ml 0.5 M EDTA
   Add dH_2O to 1 L

References

1. Palmer, B. A., Dimitrova, Z., Skums, P., Crosbie, O., Kenny-Walsh, E. and Fanning, L. J. (2014). Analysis of the evolution and structure of a complex intrahost viral population in chronic hepatitis C virus mapped by ultradeep pyrosequencing. *J Virol* 88(23): 13709-13721.
2. Skums, P., Dimitrova, Z., Campo, D. S., Vaughan, G., Rossi, L., Forbi, J. C., Yokosawa, J., Zelikovsky, A. and Khudyakov, Y. (2012). Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics* 13 Suppl 10: S6.
3. Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 30(12): 2725-2729.