



Published in final edited form as:

Arch Virol. 2020 January ; 165(1): 127–135. doi:10.1007/s00705-019-04466-9.

In-depth serum virome analysis in patients with acute liver failure with indeterminate etiology

Yi Ren^{1,4}, Yanjuan Xu¹, William M. Lee³, Adrian M Di Bisceglie^{1,2,*}, Xiaofeng Fan^{1,2,*}

¹Division of Gastroenterology & Hepatology, Department of Internal Medicine, Saint Louis University School of Medicine, St. Louis, Missouri 63104, USA

²Saint Louis University Liver Center, Saint Louis University School of Medicine, St. Louis, Missouri 63104, USA

³Division of Digestive and Liver Diseases, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA

⁴Wuhan Pulmonary Hospital, Wuhan 430030, Hubei, China

Abstract

In clinical virome research, whole-genome/transcriptome amplification is required when starting material is limited. An improved method, named “template-dependent multiple displacement amplification” (tdMDA), has recently been developed in our lab (Wang et al. (2017) *BioTechniques* 63:21–25). In combination with Illumina sequencing and bioinformatics pipelines, its application in virome sequencing was explored using a serum sample from a patient with chronic hepatitis C virus (HCV) infection. In comparison to an amplification-free procedure, virome sequencing via tdMDA showed a 9.47-fold enrichment for HCV-mapped reads and, accordingly, an increase in HCV genome coverage from 28.5% to 70.1%. Eight serum samples from acute patients liver failure (ALF) with or without known etiology were then used for virome sequencing with an average depth at 94,913x. Both similarity-based (mapping, NCBI BLASTn, BLASTp, and profile hidden Markov model analysis) and similarity-independent methods (machine-learning algorithms) identified viruses from multiple families, including *Herpesviridae*, *Picornaviridae*, *Myoviridae*, and *Anelloviridae*. However, their commensal nature and cross-detection ruled out an etiological interpretation. Together with a lack of detection of novel viruses in a comprehensive analysis at a resolution of single reads, these data indicate that viral agents might be rare in ALF cases with indeterminate etiology.

*Correspondence: Division of Gastroenterology & Hepatology, Department of Internal Medicine, Saint Louis University School of Medicine, St. Louis, Missouri 63104, USA. Tel: +1 314 977-7833; Fax: +1 314 577-8125, xiaofeng.fan@health.slu.edu (X.F.) or adrian.dibisceglie@health.slu.edu (A.M.D.).

Author contributions

Conceived and designed the experiments: XF, WML, AMD; performed the experiments: YR, YX; data interpretation: XF, AMD; analyzed the data: XF; wrote the paper: XF.

Competing interest statement

The authors have no conflict of interest to declare with respect to this manuscript.

Publisher's Disclaimer: This Author Accepted Manuscript is a PDF file of an unedited peer-reviewed manuscript that has been accepted for publication but has not been copyedited or corrected. The official version of record that is published in the journal is kept up to date and so may therefore differ from this version.

Keywords

acute liver failure; virome; multiple displacement amplification; etiology

Introduction

The human virome, defined as the collection of all viruses at a given anatomical site, is an emerging topic in biomedical research [1]. An essential step in human virome research is to taxonomize and/or discover known and novel viruses. Methodologically, this can be achieved by using next-generation sequencing (NGS), a so-called metagenomics or metatranscriptome approach [2]. An unsolved issue in viral metagenomics, however, is its low sensitivity due to the overwhelming number of sequencing reads from the human genome in NGS data output (reviewed in reference 3). A standard pipeline of library construction in NGS is often associated with the loss of low-abundance DNA/cDNA species [4], which is a challenge when working with a human specimen that is limited in quantity. An amplification step is thus required prior to library preparation. For this purpose, phi29-DNA-polymerase-based multiple-strand displacement amplification (MDA) is frequently employed for whole-genome or whole-transcriptome amplification due to its high fidelity and low amplification bias [5]. A notable concern about MDA is the generation of artifacts from high concentrations of primers, which sometimes account for up to 70% of NGS data [6, 7]. By designing random pentamer primers with their 5' ends blocked by a C18 spacer, we have recently found that these kinds of artifacts can be eliminated completely from an MDA that works in a template-dependent manner (tdMDA) [8]. In the current study, we have explored the utility of this technique for studying the virome in human serum, which is the most common type of clinical specimen and is notorious for its low concentration of circulating nucleic acids and highly degraded nature [9]. We first evaluated the efficiency of tdMDA using a serum sample from a patient with chronic hepatitis C virus (HCV) infection and then applied the method to patients with acute liver failure (ALF) with unknown etiology.

Materials and methods

Patient samples

A serum sample from our sample repository was included in the study under Saint Louis University Institutional Review Board protocol 10592. This serum sample, available in a large volume, was collected from a patient with chronic HCV infection who had an HCV viral load at 4.3×10^6 copies/mL (approximately equivalent to 4.73×10^6 IU/mL) as quantified by Roche Amplicor HCV Monitor (v2.0). Using HCV as a model virus, the sample was used to estimate the efficiency of our experimental protocol for virome sequencing.

A total of eight serum samples, 0.5 mL each, were obtained from the Acute Liver Failure Study Group (ALFSG), an NIDDK-sponsored ongoing clinical trial ([ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT01101100) identifier:).

Since 1998, the ALFSG registry has enrolled and collected bio-samples from patients meeting the standard definition of acute liver failure (coagulopathy with international normalized ratio (INR) ≥ 1.5 and any degree of hepatic encephalopathy) at 23 transplant centers across North America. Informed consent was obtained from next of kin in all instances, since the patient by definition had an altered mental status. All samples were collected on day 1 of study enrollment from ALF with indeterminate etiology ($n = 5$) or known etiology (acetaminophen-related acute liver failure) ($n = 3$) and promptly frozen at -80°C until use. The etiology in each case was determined initially by the site investigator but was supplemented by later review by a causality committee composed of senior hepatologists with access to additional data in many instances. The designation “indeterminate” was based on this additional review [10]. Five indeterminate ALF cases were deemed suitable for further analysis. All samples were coded and shipped to our lab at Saint Louis University, and the study protocol was reviewed and approved by the Saint Louis University Institutional Review Board (assurance no: FWA00005304).

RNA extraction, reverse transcription (RT), tdMDA, and Illumina sequencing

Total RNA was extracted from 140 μL of serum and eluted into 60 μL of Tris buffer (pH 8.5) using a QIAamp Viral RNA Mini Kit (QIAGEN, Valencia, CA). Despite the name of the kit, it actually extracts both cell-free DNA and RNA larger than 200 bp, as indicated by the manufacturer. Total RNA was used for reverse transcription (RT)-tdMDA as described previously [8]. In brief, 10.6 μL of extracted RNA was mixed with 9.4 μL RT matrix consisting of 1x SuperScript III buffer, 10 mM DTT, 80 μM exonuclease-resistant random pentamer primers with their 5' ends blocked by a C18 spacer [8], 2 mM dNTPs (Epicentre), 20 U (0.5 μL) of RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen), and 200 U of SuperScript III reverse transcriptase (Life Technologies). The reaction was incubated at 37°C for 30 minutes and 50°C for 30 minutes and inactivated at 70°C for 15 minutes. An aliquot of 4 μL of RT reaction was used for tdMDA in a 40- μL volume consisting of 1x phi29 DNA polymerase buffer, 1 mM dNTPs, 80 μM random pentamer primers (the same as used for RT), and 20 units of phi29 DNA polymerase (New England Biolabs, Ipswich, MA). The reaction was incubated at 28°C for 14 hours and then terminated by heating at 65°C for 15 minutes. After the purification using a QIAamp DNA Mini Kit (QIAGEN), 1 ng of RT-tdMDA product at a concentration of 0.4 ng/ μL was used for library construction with a Nextera XT DNA Sample Preparation Kit (Illumina, San Diego, CA) and sequencing on the Illumina MiSeq platform (1 \times 250-bp single reads and mid-output) at MOgene (St. Louis, MO). A negative control in which water was used instead of total RNA in RT-tdMDA was also included for Illumina sequencing.

In addition to RT-tdMDA-based Illumina sequencing, the HCV sample was processed for direct sequencing without an amplification step. In this procedure, the RT reaction (first cDNA strand synthesis), was performed as described above using total RNA extracted from 4 mL of serum. All RT reactions were then combined and purified using a QIAquick Nucleotide Removal Kit (QIAGEN), followed by second-strand synthesis using the reagents from a SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen). After purification, approximately 0.9 μg of double-stranded cDNA was obtained. An aliquot of 1 ng of cDNA was used for library preparation and sequencing as described above.

Estimation of the efficiency of serum virome sequencing via tdMDA

Raw sequence reads in fastq format from the HCV patient were first filtered in PRINSEQ (v0.20) for read quality control to include read length ≥ 70 bp, mean read quality score ≥ 25 , low complexity with DUST score ≤ 7 , ambiguous bases $\leq 1\%$, and all duplicates [11]. Using Bowtie 2 mapper [12], quality reads from human sequences (The National Center for Biotechnology Information [NCBI] GRCh38 build) [13], NCBI microbial reference sequences for bacteria, archaea, fungi and protist (downloaded on September 11, 2018) [14], and microbial reference genome sequences from the Human Microbiome Project (HMP) were subtracted sequentially [15]. HCV-specific reads were then extracted using Bowtie 2 by indexing 184 HCV reference sequences downloaded from the HCV database [16]. Using a two-step strategy developed in our previous study [17], a patient-specific HCV consensus sequence was generated, and this served as the reference for read mapping. The percentage of HCV-specific reads and the HCV genome coverage were compared between the experimental protocols with or without tdMDA amplification.

Virome analysis in ALF patients

After the read quality control and subtraction as described above, the remaining reads were subjected to a stepwise procedure for complete annotation. First, reads from eight ALF cases and the control were mapped onto NCBI viral reference sequences (9,687 complete viral genomes downloaded on September 11, 2018) [14], followed by read assignment to the mapped viral genome using SAMtools [18], a procedure called “viral categorization” [19]. Second, upon an additional subtraction from NCBI viral reference sequences and the negative control, reads from each sample were assembled *de novo* using the short-read assembler SPAdes [20]. The resulting contigs were combined with unassembled reads (singletons) to generate a sequence dataset that was re-labelled with a sample identifier using PRINSEQ [11]. All eight sequence datasets were then pooled together and compressed by similarity at 90% with CD-HIT [21], followed by the removal of sequences with a low complexity (DUST score >7) in PRINSEQ [11]. These sequences were annotated using a similarity-based strategy, first by NCBI BLASTn comparison against the NCBI collection of nucleotide acid sequences (database “nt”) with a conserved e-value setting of 1×10^{-5} . Sequences with no BLASTn hits were translated in all six reading frames using a custom script [22]. Amino acid sequences were searched using BLASTp against the NCBI non-redundant protein database (“nr”), also with an e-value of 1×10^{-5} . Amino acid sequences without BLASTp hits were used for a remote homology search using profile hidden Markov model (HMM) analysis in HMMER (v3.2.1) [23]. In this way, a search with default settings were performed consecutively with three rounds of profile-HMM built from NCBI viral RefSeq except for phage (vFam) [24], prokaryotic virus orthologous groups (pVOGs) (phages) [25], and the collection of protein families (Pfam, 17,929 entries in version 32) [26]. Those sequences without HMMER hits were examined for possible viruses using a machine-learning method implemented in VirFinder using the model of “VF.modEPV_k8.rda”, which was trained using 5,800 eukaryotic virus sequences collected from NCBI [27].

Finally, sequences suspected to be of viral origin were examined using the following criteria: supported by >25 reads from the original data, lacking cross-mapping in one of three ALF

cases with known etiology and the negative control, and not a tandem repeat [28]. Only sequences that met the criteria were considered candidate viral sequences. In addition, given the potential genome similarity between bacteria and viruses, the criteria were also applied to sequences with bacterial hits. One of the bacterial sequences that met the criteria was selected for PCR to determine whether it was present in the serum or was a contaminant from the experimental pipeline.

Confirmation of a bacterium-like sequence by PCR and Sanger sequencing

For a selected bacterium-like sequence that was present in one ALF case with unknown etiology, RT-PCR was conducted with primers designed from its predicted coding region. In brief, 10 μ L of total RNA was used for 20- μ L RT consisting of 1x SuperScript III reverse transcriptase buffer, 10 mM DTT, 1 μ M reverse primer R1 (5' - TCG GCA ACA ACA AGA CCA TC - 3'), 1 mM dNTPs (Invitrogen), 16 U of RNasin Ribonuclease Inhibitor (Promega), and 200 U of SuperScript III reverse transcriptase (Invitrogen). The reaction was incubated at 50°C for 45 min and inactivated by heating at 70°C for 15 min. The first round of PCR was done with 5 μ L of RT product in a 50- μ L reaction including 1x Q5 polymerase buffer, 0.8 mM dNTPs, primers R1 and F1, each at 0.4 μ M (F1, 5' - GCT CTC ATC TTA CCC GTC CC - 3'), and 1 U of Q5 DNA polymerase (New England Biolabs). Cycle parameters were programmed as 94°C for 1 min connected by the first 5 cycles of 94°C for 1 min, 60°C for 1 min and 72°C for 1 min, linked by 25 cycles in which the annealing temperature was reduced to 50°C (touchdown protocol), followed by a final 7-min incubation at 72°C. A 2- μ L aliquot of the first-round PCR product was used for the second round of PCR with primers F2 (5' - AAA GGT GGA GAG AGT TGG CG - 3') and R2 (5' - GCA GGA ATG CAG AAG CGA C - 3') using the same cycle settings. The product was gel-purified and subjected to direct sequencing as described previously [29].

Data availability

Raw sequence data in fastq format from the HCV patient, eight ALF cases, and the negative control were deposited in the NCBI Sequence Read Archive (SRA) under BioProject ID: PRJNA527118. A bacterium-like sequence identified in an experimental reagent in the current study was also deposited in the GenBank database under the accession number [MK659570](#).

Results

Enhanced HCV detection via tdMDA-mediated serum virome sequencing

To minimize the influence of low-complexity reads from homogeneous regions on read count, human and bacterial sequences were first removed from the serum virome sequencing data, which were then mapped onto the HCV reference sequences. Consequently, there were 457 and 56 HCV-mapped reads, respectively, from the virome sequencing with and without an amplification step by tdMDA. The tdMDA-based approach gave a 9.47-fold enrichment with regard to HCV detection (457/1,597,044 vs. 56/1,853,336) (Fig. 1). The read coverage of the HCV genome was also increased from 28.5% to 70.1% upon the inclusion of an amplification step (Fig. 1). Finally, both library preparation methods showed a similar read mapping pattern that was centered on a 3,000-bp region in the HCV nonstructural region

from NS3 to NS5a (Fig. 1). Therefore, tdMDA did not result in an apparent bias compared to direct library construction of double-strand cDNA from serum.

Detection of known viruses in ALF patients by read mapping

The experimental protocol for the HCV patient was applied to eight ALF cases with (acetaminophen overdose for cases 6, 7, and 8) or without known etiology (cases 1, 2, 3, 4, and 5). Because tdMDA eliminates primer-related artifacts but not the contamination from the reagents, the negative control that showed a positive amplification was also included for Illumina sequencing [8]. After read quality control, the average number of reads for each ALF case was $7,641,992 \pm 2,553,282$, with a read length at 248.4 ± 1.83 nt. Given that the typical product size for RT-tdMDA is 20 kb [8], these data were transformed into a sequencing depth of 94,913x, which is approximately 4.4 times deeper when compared to the assignment for the HCV patient. Using an empirical read number cutoff of 5, the mapping detected a total of 43 viruses that were present in at least one ALF patient. These viruses belonged to the families *Herpesviridae*, *Picornaviridae*, *Myoviridae* (bacteriophage), and *Anelloviridae* (Fig. 2). Not surprisingly, torque teno virus (TTV) was a major virus detected in all of the ALF cases except for cases 2, 4, and 6. Cases 5 and 8 had a large number of mapping reads assigned to TTV. All of the viruses from the four families were detected in all eight ALF cases, suggesting a lack of etiological association.

Analysis of unmapped reads from ALF patients

After subtractive mapping to NCBI reference sequences, microbial genomes from HMP, and sequences from the negative control, unmapped reads occupied a small portion of the data: 36,856 (0.58%), 16,990 (0.25%), 4,105 (0.06%), 12,029 (0.14%), 21,117 (0.28%), 172,538 (1.23%), and 28,752 (0.42%) reads for cases 1, 2, 3, 4, 6, 7, and 8, respectively. An exception was case 5, in which 3,430,454 reads (56.99%) were unmapped. Case-based *de novo* assembly of these unmapped reads generated 18,436 contigs. Together with unmapped reads to the contigs (i.e., singletons), a total of 25,457 sequences were collected from all eight ALF cases. After the removal of short (<100 nt) and low-complexity sequences (DUST score >7 in PRINSEQ), these sequences were further compressed by 90% similarity in CD-HIT. Eventually, a total of 10,923 sequences were subjected to similarity-based annotation. First, BLASTn analysis resulted in 3,733 hits, mostly associated with the human genome or viruses of the family *Anelloviridae* (Supplementary Table 1). The remaining 7,190 sequences were translated in all six reading frames into 9,089 coding fragments ranging from 12 to 671 amino acids, which successively had 1,929 unique matches in the “nr” database in a BLASTp search (Supplementary Table 2), 229 hits in vFam, 119 hits in pVOGs, and 305 hits in Pfam by HMMER. vFam, which includes sequences from all known eukaryotic viruses, is a valuable resource for virus discovery. Among 64 vFams detected by HMMER search, the vFams from members of the family *Anelloviridae* received 62% of the 229 hits (Fig. 3). None of remaining sequences met the criteria for candidate viral sequences, suggesting their commensal nature, like TTV or the contaminants from the environment or experimental pipeline. Of the final 5,525 sequences, VirFinder identified only 35 candidate viral sequences with *p*-values less than 0.039, which was the empirical cutoff determined by analysis of 11,000 non-overlapping genome fragments (300–1,000 bp) derived from known hepatotropic viruses, including hepatitis A virus, hepatitis B virus,

HCV, hepatitis D virus, and hepatitis E virus (available from the author upon request). Thirteen sequences were actually tandem repeats ranging in size from 8 to 213 nt. Another 22 sequences did not qualify as a candidate viral sequence using the defined criteria.

A bacterium-like sequence from RNaseOUT Recombinant Ribonuclease Inhibitor

Approximately 9% of sequences with bacterial hits in BLASTp met the defined criteria. One of these sequences, alf1_3, shared ~50% sequence identity with the CBS domain-containing protein from a *Candidatus Melainabacteria* bacterium. This sequence matched a large number of reads (467 reads) and appeared to be unique for ALF case 1. Reads extracted with this sequence were re-assembled by SPAdes and an additional assembler, Newbler [32], both of which gave the same contig sequence with a size of 1,662 nt. RT-PCR showed its appearance in multiple ALF cases, but the results were less reproducible. This phenomenon, sometime called a PCR ghost, might result from the use of a very small amount of template. The sequence was thus suspected to be a contaminant from the reagents. When using the reagents as a template, nested PCR confirmed its presence in RNaseOUT Recombinant Ribonuclease Inhibitor (Fig. 4). Therefore, this sequence was amplified by chance in RT-tdMDA of ALF case 1.

Discussion

In the current study, the feasibility of using tdMDA in serum virome sequencing was first evaluated using a serum sample from a patient with chronic HCV infection. In comparison to amplification-free library construction, tdMDA enriched HCV detection by 9.42-fold through serum virome sequencing (Fig. 1). It is known that the phi29 DNA polymerase used in tdMDA favors large or circular templates [5]. As a consequence, small DNA/cDNA fragments that are the most abundant in circulation might be amplified inefficiently by tdMDA. However, TTV was also detected in this patient. In spite of its circular genome structure [33], tdMDA achieved only 1.23-fold enrichment (980 vs. 796 reads/million reads).

In an analysis of our previous data generated using tdMDA [8], HCV-mapped reads, normalized per million total reads, could vary by up to one log (about 500–5,000 reads) among patients with similar viral titers. We speculate that the enrichment efficiency of tdMDA may depend on multiple factors, such as the complexity, composition, and amount of the template. In our previous work using standard MDA, HCV-mapped reads consistently accounted for fewer than 50 reads/million reads [19]. Taken together, while the efficiency may fluctuate from sample to sample, tdMDA generally outperforms MDA in terms of viral sequence enrichment, perhaps due to the elimination or suppression of primer-mediated artifacts [8]. It should also be noted that our protocol does not involve any sample pretreatment that might damage the intrinsic virome [34]. In addition, tdMDA generates an amplicon with the size around 20 kb that is large enough for direct library preparation without a need for additional steps such as concatemerization. Therefore, tdMDA provides a simple but efficient approach for serum virome sequencing. The method presented here is especially useful when the serum is available in a limited volume, such as the ALF samples in the present study, of which only 0.5 mL had been provided for research purposes.

Applying tdMDA-based virome sequencing, we did not find any viruses, either known or novel, that could be associated with ALF etiology. Given a high prevalence of TTV in the general population [35], it is not surprising that TTV was detected in 5 of 8 ALF cases. In particular, TTV was identified as a major virus at various levels using similarity-based methods, including read mapping, BLASTn, BLASTp, and HMMER analysis. For ALF case 5, there was a significantly large number of unmapped reads (56.99%) in which approximately 55% of the reads could be assigned to TTV by re-mapping of TTV-specific contigs (Supplementary Tables 1 and 2, Fig. 3). This suggests a huge intra-patient genetic diversity. However, co-circulation of multiple TTV genotypes has been observed in blood donors [36, Fan et al., unpublished data]. Therefore, high genetic diversity is unlikely to be an etiological factor in ALF patients.

The rate of unknown etiology in ALF patients varies among countries [37]. In the United States, acetaminophen overdose is a major etiological factor, and no more than 15% of all ALF cases are of unknown etiology [38]. After integrative analyses, including more-sensitive lab tests and a thorough clinical re-evaluation, however, the percentage of ALF cases with unknown etiology was reduced to 5.5% [10]. Moreover, in a recent study from the ALFSG, acetaminophen was undetectable in serum in more than 50% patients deemed to have had an acetaminophen overdose [39]. Therefore, the incidence of true ALF cases with indeterminate etiology should be well below 5.5%. The lack of detection of viruses in the current study suggests that viral etiology might be rare among these patients. Our results are consistent with those of a previous study using serum samples from the ALFSG, although unmapped reads were not analyzed [40]. It should be noted that our method is limited in its sensitivity for viral detection. Using HCV as a model virus, 457 HCV-specific reads were recovered out of 1,597,044 quality reads. The average read output per ALF case was 7,641,992 reads after quality control. Using this method, a virus like HCV is only detectable if it has a titer of 9.8×10^3 copies/mL when using an empirical cutoff of 5 reads. To compensate for this limitation in viral detection, two strategies were applied in the current study. First, all eight ALF samples were analyzed at ultra (saturated) sequencing depths, as indicated by an average rate of read duplicates at $28.3 \pm 7.8\%$. The second strategy was bioinformatics analysis at the level of single reads. Both strategies together increase the chance of detection of known or unknown viruses and thereby enhance the sensitivity of the entire pipeline. Finally, analysis of the serum virome is a relatively inefficient way of detecting a virus that is transmitted in a non-parenteral manner. Therefore, viral agents cannot be excluded completely as a cause of ALF with indeterminate etiology, although the probability is low. A more-thorough virome analysis of other types of specimens, such as liver tissue, is needed to reach a firm conclusion regarding a viral etiology in ALF patients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the US National Institutes of Health (NIH) grants AI117128 (X.F.) and AI139835 (X.F.) and a seed grant from the Saint Louis University Liver Center (X.F.). The ALFSG was supported by U01

DK58369. Special acknowledgment is given to all the patients, families, coordinators and PI's that participated in this network, 1998–2019.

References

1. Handley SA (2016) The virome: a missing component of biological interaction networks in health and disease. *Genome Med* 8:32 10.1186/s13073-016-0287-y [PubMed: 27037032]
2. Mokili JL, Rohwer F, Dutilh BE (2012) Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* 2:63–77. 10.1016/j.coviro.2011.12.004 [PubMed: 22440968]
3. Houldcroft CJ, Beale MA, Breuer J (2017) Clinical and biological insights from viral genome sequencing. *Nat Rev Microbiol* 15:183–192. 10.1038/nrmicro.2016.182 [PubMed: 28090077]
4. Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W, Fodor SP (2014) Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proc Natl Acad Sci USA* 111:1891–1896. 10.1073/pnas.1323732111 [PubMed: 24449890]
5. Nelson JR (2014) Random-primed, Phi29 DNA polymerase-based whole genome amplification. *Curr Protoc Mol Biol* 105: Unit 15.13. 10.1002/0471142727.mb1513s105 [PubMed: 24510438]
6. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 10:R32 10.1186/gb-2009-10-3-r32 [PubMed: 19327155]
7. Pan X, Durrett RE, Zhu H, Tanaka Y, Li Y, Zi X, Marjani SL, Euskirchen G, Ma C, Lamotte RH, Park IH, Snyder MP, Mason CE, Weissman SM (2013) Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proc Natl Acad Sci USA* 110:594–599. 10.1073/pnas.1217322109 [PubMed: 23267071]
8. Wang W, Ren Y, Lu Y, Xu Y, Crosby SD, Di Bisceglie AM, Fan X. Template-dependent multiple displacement amplification for profiling human circulating RNA (2017) *Biotechniques* 63:21–27. 10.2144/000114566 [PubMed: 28701144]
9. Butt AN, Swaminathan R (2008) Overview of circulating nucleic acids in plasma/serum. *Ann N Y Acad Sci* 1137:236–242. 10.1196/annals.1448.002 [PubMed: 18837954]
10. Ganger DR, Rule J, Rakela J, Bass N, Reuben A, Stravitz RT, Sussman N, Larson AM, James L, Chiu C, Lee WM; Acute Liver Failure Study Group (2018) Acute liver failure of indeterminate etiology: a comprehensive systematic approach by an expert committee to establish causality. *Am J Gastroenterol* 113:1319–1328. 10.1038/s41395-018-0160-2 [PubMed: 29946176]
11. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27: 863–864. 10.1093/bioinformatics/btr026 [PubMed: 21278185]
12. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. 10.1038/nmeth.1923 [PubMed: 22388286]
13. Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y (2017) Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* 109:83–90. 10.1016/j.ygeno.2017.01.005 [PubMed: 28131802]
14. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue):D61–65. 10.1093/nar/gkl842 [PubMed: 17130148]
15. Lloyd-Price J, Mahurkar A, Rahnavard G, et al. (2017) Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550:61–66. 10.1038/nature23889 [PubMed: 28953883]
16. Kuiken C, Yusim K, Boykin L, Richardson R (2005) The Los Alamos HCV Sequence Database. *Bioinformatics* 21:379–384. 10.1093/bioinformatics/bth485 [PubMed: 15377502]
17. Wang W, Zhang X, Xu Y, Weinstock GM, Di Bisceglie AM, Fan X (2014) High-resolution quantification of hepatitis C virus genome-wide mutation load and its correlation with the outcome of peginterferon-alpha2a and ribavirin combination therapy. *PLoS One* 9:e100131 10.1371/journal.pone.0100131 [PubMed: 24950220]

18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. 10.1093/bioinformatics/btp352 [PubMed: 19505943]
19. Wang W, Zhang X, Xu Y, Di Bisceglie AM, Fan X (2013) Viral categorization and discovery in human circulation by transcriptome sequencing. *Biochem Biophys Res Commun* 436:525–529. 10.1016/j.bbrc.2013.05.139 [PubMed: 23764402]
20. Bankevich A, Nurk S, Antipov D, et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. 10.1089/cmb.2012.0021 [PubMed: 22506599]
21. Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* 11:187 10.1186/1471-2105-11-187 [PubMed: 20388221]
22. Novichkov PS, Ratnere I, Wolf YI, Koonin EV, Dubchak I (2009) ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. *Nucleic Acids Res* 37(Database issue):D448–454. 10.1093/nar/gkn684 [PubMed: 18845571]
23. Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure. *BMC Bioinformatics* 11:431 10.1186/1471-2105-11-431 [PubMed: 20718988]
24. Skewes-Cox P, Sharpton TJ, Pollard KS, DeRisi JL (2014) Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS One* 9:e105067 10.1371/journal.pone.0105067 [PubMed: 25140992]
25. Grazziotin AL, Koonin EV, Kristensen DM (2017) Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res* 45(Database issue): D491–498. 10.1093/nar/gkw975 [PubMed: 27789703]
26. Finn RD, Coghill P, Eberhardt RY, et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44(D1):D279–D285. 10.1093/nar/gkv1344 [PubMed: 26673716]
27. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5:69 10.1186/s40168-017-0283-5 [PubMed: 28683828]
28. Benson G. Tandem repeats finder: a program to analyze DNA sequences (1999) *Nucleic Acids Res* 27:573–580. 10.1093/nar/27.2.573 [PubMed: 9862982]
29. Lu Y, Xu Y, Di Bisceglie AM, Fan X (2013) Comprehensive cloning of patient-derived 9022-bp amplicons of hepatitis C virus. *J Virol Methods* 191:105–112. 10.1016/j.jviromet.2013.04.010 [PubMed: 23602804]
30. Carver T, Harris SR, Otto TD, Berriman M, Parkhill J, McQuillan JA (2013) BamView: visualizing and interpretation of next-generation sequencing read alignments. *Brief Bioinformatics* 14:203–212. 10.1093/bib/bbr073 [PubMed: 22253280]
31. Zhao S, Guo Y, Sheng Q, Shyr Y (2014) Heatmap3: an improved heatmap package with more powerful and convenient features. *BMC Bioinformatics* 15(Suppl 10): 16 10.1186/1471-2105-15-S10-P16 [PubMed: 24428894]
32. Reinhardt JA, Baltrus DA, Nishimura MT, Jeck WR, Jones CD, Dangl JL (2009) De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Res* 19:294–305. 10.1101/gr.083311.108 [PubMed: 19015323]
33. Biagini P. Classification of TTV and related viruses (anelloviruses) (2009). *Curr Top Microbiol Immunol* 331:21–33. [PubMed: 19230555]
34. Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J (2001) A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc Natl Acad Sci USA* 98:11609–11614. 10.1073/pnas.211424698 [PubMed: 11562506]
35. Spandole S, Cimponeriu D, Berca LM, Mih escu G (2015) Human anelloviruses: an update of molecular, epidemiological and clinical aspects. *Arch Virol* 160:893–908. 10.1007/s00705-015-2363-9 [PubMed: 25680568]

36. Niel C, Saback FL, Lampe E (2000) Coinfection with multiple TT virus strains belonging to different genotypes is a common event in healthy Brazilian adults. *J Clin Microbiol* 38:1926–30. [PubMed: 10790123]
37. Lee WM (2008) Etiologies of acute liver failure. *Semin Liver Dis.* 28:142–152. 10.1055/s-2008-1073114 [PubMed: 18452114]
38. Lee WM, Squires RH Jr, Nyberg SL, Doo E, Hoofnagle JH (2008) Acute liver failure: Summary of a workshop. *Hepatology* 47:1401–1415. 10.1002/hep.22177 [PubMed: 18318440]
39. Leventhal TM, Gottfried M, Olson JC, Subramanian RM, Hameed B, Lee WM; Acute Liver Failure Study Group (2019) Acetaminophen is undetectable in plasma from more than half of patients believed to have acute liver failure due to overdose. *Clin Gastroenterol Hepatol.* Epub ahead of print pii: S1542–3565(19)30087–4. 10.1016/j.cgh.2019.01.040
40. Somasekar S, Lee D, Rule J, Naccache SN, Stone M, Busch MP, Sanders C, Lee WM, Chiu CY (2017) Viral surveillance in serum samples from patients with acute liver failure by metagenomic next-generation sequencing. *Clin Infect Dis* 65:1477–1485. 10.1093/cid/cix596 [PubMed: 29020199]

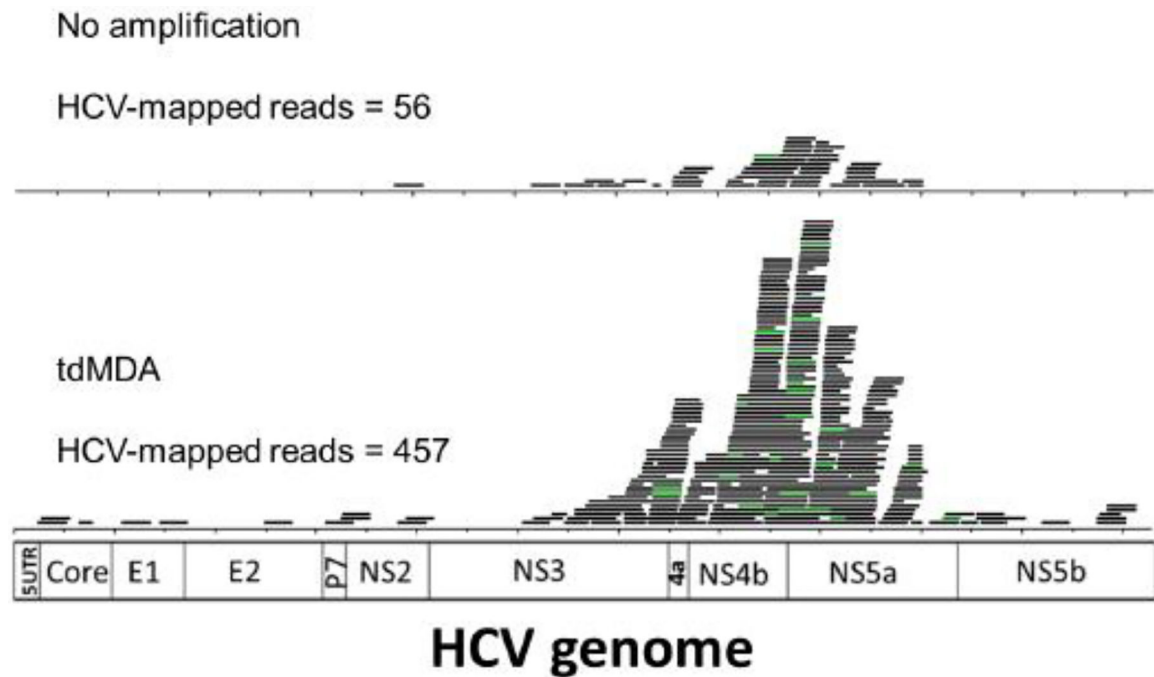


Fig. 1. HCV detection by virome sequencing. The alignment of HCV-mapped reads from Illumina sequencing with (bottom) or without amplification step (top) in BamView [30]. Reads with matching start and end positions were collapsed into one line and are shown in green. The HCV genome structure is shown at the bottom.

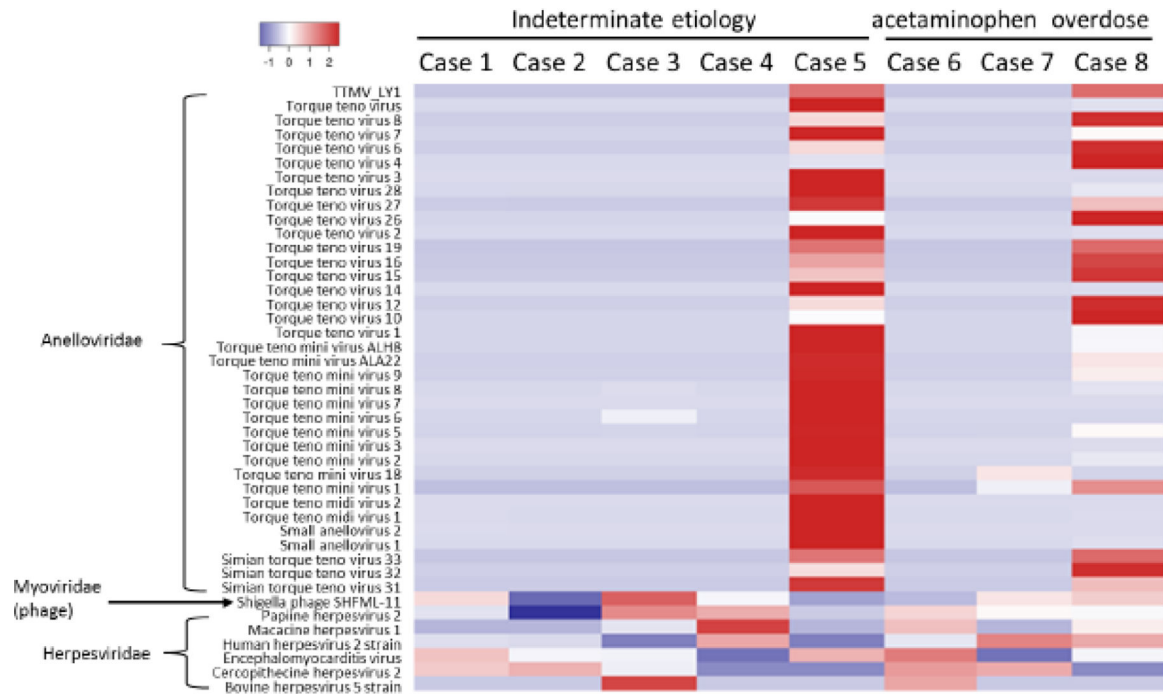


Fig. 2. Detection of known viruses by read mapping. Original read numbers assigned to each virus were normalized based on the average data output for eight ALF cases (7,641,992) and are presented as a heat map using Heatmap3 [31]. The viruses and their corresponding families are indicated on the left.

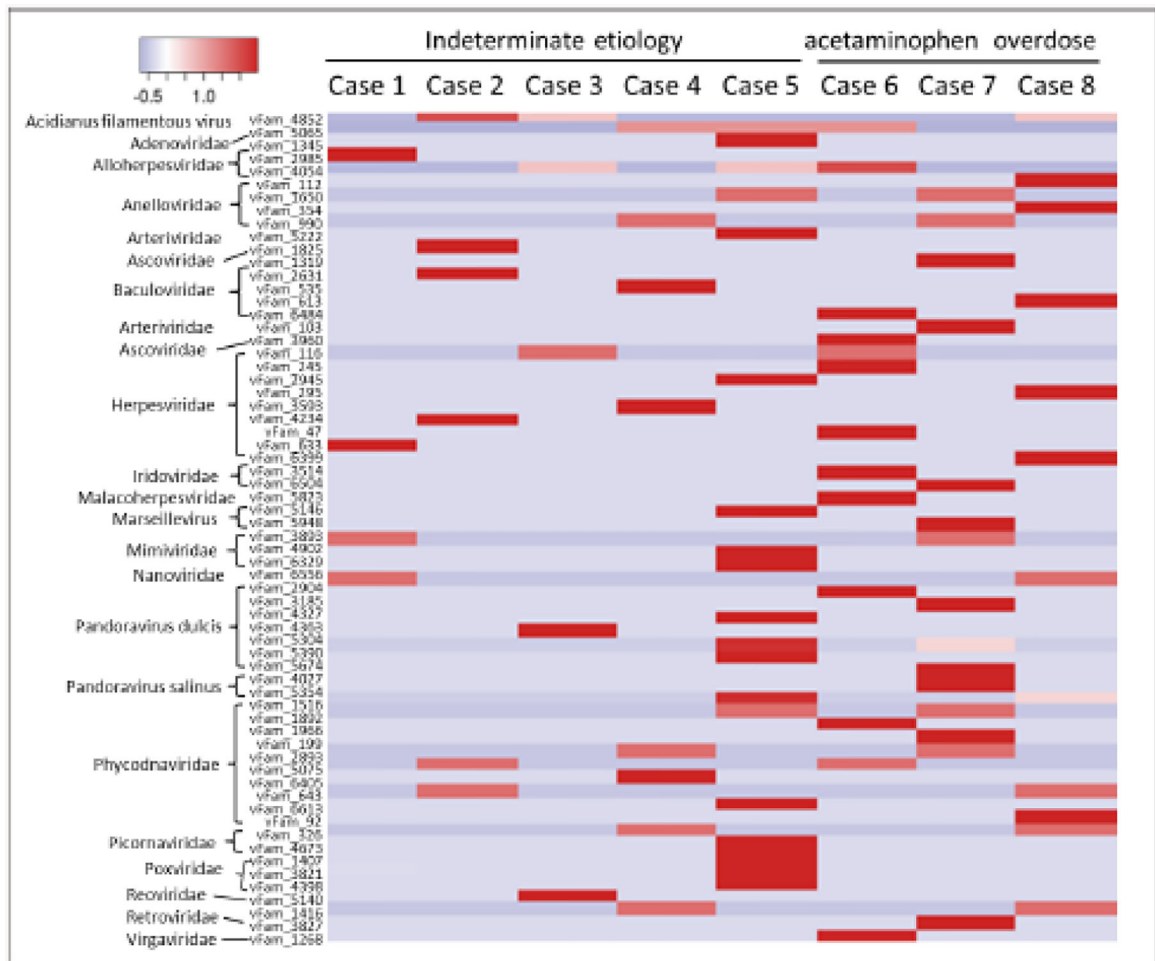


Fig. 3. A list of vFams hits by HMMER search. The detection was performed using amino acid sequences translated from contigs and singletons from individual ALF cases. It was therefore a qualitative rather than quantitative examination. The viral families from which the profile HMM (vFam) was built are indicated on the left.

