Aus dem Institut für Virologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Identification of Ribavirin-Induced Mutations in Patient-Derived Hepatitis E Virus

Identifizierung von Ribavirin-induzierten Mutationen in von Patienten stammenden Hepatitis-E-Viren

zur Erlangung des akademischen Grades
Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Christian Patrick Papp

Datum der Promotion: 29.11.2024

Table of contents

List of tables

List of abbreviations

CHE – Chronic Hepatitis E

cDNA – Complementary DNA

DNA – Desoxyribonucleic Acid

EASL – European Association for the Study of the Liver

GSTA1 – Glutathione S-Transferase Alpha 1

Hel – RNA Helicase

HEV – Hepatitis E Virus

HIV – Human Immunodeficiency Virus

HVR – Hypervariable Region

lrPCR – long-range PCR

MeT – Methyltransferase

NGS – Next-Generation Sequencing

ONT – Oxford Nanopore Technology

ORF – Open Reading Frame

PCR – Polymerase Chain Reaction

qPCR – Quantitative PCR

RdRp – RNA-dependent RNA-Polymerase

RNA – Ribonucleic Acid

RPL18 – Ribosomal Protein L18

RT – Reverse Transcription

SMS – Single-molecule Sequencing

SNP – Single Nucleotide Polymorphism

SOT – Solid Organ Transplant

UTR – Untranslated Region

zOTU – Zero-radius Operational Taxonomic Unit

Abstract

The Hepatitis E Virus (HEV) is an emerging yet underdiagnosed pathogen that is increasingly diagnosed in high-income countries. HEV infection has a predominantly subclinical manifestation; however, HEV genotype 3 can chronify in immunocompromised individuals. Currently, there is no approved antiviral therapy for HEV infection and chronic hepatitis E in particular. Nevertheless, ribavirin has been established as a treatment option for individuals with chronic HEV infection. Mutations in the RNA-dependent RNA-polymerase region (RdRp) associated with therapy failure as well as insertions in the hypervariable region (HVR) are increasingly detected in individuals with chronic HEV infection. Accordingly, for a more detailed characterization of the viral population, new approaches were developed that allowed both full-length and targeted sequencing of HEV-3 genomic regions using novel technologies such as next-generation and third-generation sequencing. Due to the full-length amplification using one pair of primers combined with single-molecule sequencing, haplotyping was facilitated, leading to the first-time detection of an HEV-variant containing two insertions simultaneously. In addition, targeted ultra-deep sequencing has demonstrated utility in the clinical setting by elucidating the effects of standard or novel antiviral treatment options on the occurrence of mutations in the RdRp. The targeted sequencing approach for the HVR facilitated the detection of multiple variants containing different insertions in the same sample and showed that insertions in the HEV genome may occur more frequently than previously assumed. Nevertheless, this work contributes critically to HEV research by means of providing new sequencing methods for HEV and demonstrating utility for assessing therapeutic outcomes in patients with chronic HEV infections. This molecular approach is anticipated to expand our current knowledge of the HEV genome and pathogenesis in chronic hepatitis E.

Zusammenfassung

Das Hepatitis-E-Virus (HEV) ist ein neu aufkommender aber unterdiagnostizierter Erreger, der in Ländern mit hohem Einkommen zunehmend diagnostiziert wird. Die HEV-Infektion verläuft überwiegend subklinisch, wobei der HEV-Genotyp 3 bei immungeschwächten Personen zu einer Chronifizierung führen kann. Derzeit gibt es keine zugelassene antivirale Therapie für HEV-Infektionen und insbesondere der chronischen Hepatitis E; Ribavirin hat sich jedoch als Behandlungsoption für Personen mit chronischer HEV-Infektion etabliert. Mutationen in der RNA-abhängigen RNA-Polymerase-Region (RdRp), die mit Therapieversagen assoziiert sind, sowie Insertionen in der hypervariablen Region (HVR) werden zunehmend bei Personen mit chronischer HEV-Infektion nachgewiesen. Dementsprechend wurden für eine detailliertere Charakterisierung der Viruspopulation neue Ansätze entwickelt, die sowohl das komplette Genom als auch die gezielte Sequenzierung von HEV-3-Genomregionen mit Hilfe moderner Technologien wie Next-Generation- und Third-Generation-Sequencing ermöglichen. Durch die Amplifikation des kompletten Genoms mit einem Primerpaar in Kombination mit der Single-Molecule-Sequencing wurde die Haplotypisierung ermöglicht, was zum ersten Nachweis einer HEV-Variante führte, die zwei Insertionen gleichzeitig enthält. Darüber hinaus hat die gezielte ultra-tiefe Sequenzierung ihren Nutzen im klinischen Umfeld bewiesen, indem sie die Auswirkungen von Standard- oder neuen antiviralen Behandlungsoptionen auf das Auftreten von Mutationen im RdRp aufklärt. Der gezielte Sequenzierungsansatz für das HVR erleichterte den Nachweis mehrerer Varianten mit unterschiedlichen Insertionen in derselben Probe und zeigte, dass Insertionen im HEV-Genom häufiger auftreten können als bisher angenommen. Die vorliegende Arbeit leistet einen wichtigen Beitrag zur HEV-Forschung, indem sie neue Sequenzierungsmethoden für HEV beschreibt und deren Nutzen für die Beurteilung des therapeutischen Erfolgs bei Patienten mit chronischen HEV-Infektionen aufzeigt. Es ist zu erwarten, dass dieser molekulare Ansatz unser derzeitiges Wissen über das HEV-Genom und die Pathogenese der chronischen Hepatitis E erweitern wird.

1. Introduction

1.1 General introduction

Worldwide, the Hepatitis E Virus (HEV) is responsible for an estimated 20 million infections each year and the WHO estimates that HEV caused approximately 44 thousand deaths in 2015 (1). Transmission occurs mainly by contaminated water causing large outbreaks; however, zoonotic transmission by ingesting infected undercooked animal products is also a relevant source of infection (1). The infection is usually self-limiting, although, in immunocompromised individuals, there is a higher risk of progression to chronicity, including solid organ transplant (SOT) recipients, HIV-infected individuals, and patients with malignancies (2). A systematic review and meta-analysis from 2020 investigated the global burden of HEV infection in pregnancy and showed that HEV infection during pregnancy was associated with maternal deaths, low birth weight, small for gestational age, preterm birth, stillbirth, and intrauterine deaths (3).

1.2 Structure of HEV and epidemiology

HEV is a ca. 7.2 kb single-stranded positive-sense hepatotropic RNA virus that consists of three open reading frames (ORF1, ORF2, and ORF3), 5′ and 3′ untranslated regions (UTRs), and a poly(A) tract at the 3′ end. ORF1 encodes the enzymes methyltransferase (MeT), RNA helicase (Hel), and RNA-dependent RNA polymerase (RdRp), the ORF2 encodes the capsid protein, and the ORF3 encodes a multifunctional phosphoprotein important for cellular signaling and particle secretion (4). The genotypes HEV-1 to HEV-8 belong to *Orthohepevirus A* of which at least HEV-1 to HEV-4 are relevant in human infections. HEV causes epidemics and substantial outbreaks, especially in tropical and subtropical countries, affecting tens of thousands of individuals (1, 5), while approximately two billion people live in areas endemic for HEV (6). While HEV-3 and HEV-4 are distributed throughout the world, HEV-1 is common in Asia, HEV-2 in Africa and Mexico, and HEV-7 in the Middle East (5, 7). HEV-1 and HEV-2 exclusively infect humans, whereas HEV-3 and HEV-4 have a wider host range, including swine, wild boar, rabbits, and deer species (8). HEV-7 can infect both camels and humans (7). In low-income countries, the transmission is mainly waterborne and caused by HEV-1 and HEV-2 (1, 9). However, in

the last two decades, the number of HEV infections documented in high-income countries has risen exponentially. There, HEV-3 and HEV-4 are responsible for foodborne infections with mainly pigs and wild boars as main hosts and are transmitted to humans by consumption of undercooked meat (10). Depending on the genotype, there are significant differences not only in distribution and transmission patterns, but also in clinical progression. HEV-1 and HEV-2 cause severe acute hepatitis without chronification (11), whereas infections with HEV-3, HEV-4 and, HEV-7 do not only cause acute hepatitis but can also lead to chronic hepatitis in the immunocompromised (7, 12). Infection with HEV-1 is associated with severe forms of liver disease and complications in pregnant women (3, 11). These data indicate the significance of HEV variability in pathogenesis and clinical course.

1.3 HEV infection and treatment

As mentioned above, symptomatic infections can cause both acute and chronic hepatitis, including fulminant liver failure and extrahepatic symptoms such as Guillain-Barré Syndrome, acute pancreatitis, glomerulonephritis, or neuralgia amyotrophy (13). Acute hepatitis E is defined as a self-limiting disease that lasts up to 8 weeks with unspecific symptoms (9). HEV RNA is detectable in serum less than one month after onset of symptoms, while it is detectable in stool over a longer period of time (11).

Chronic hepatitis E infection is diagnosed when HEV-RNA or anti-HEV IgM persists for more than three months with elevated alanine aminotransferase (ALT) levels. The main factor that influences development of a chronic HEV infection is the host immune response. Due to suppressed immunity, viral replication persists, resulting in chronic infection. There is no targeted antiviral therapy for HEV infections available yet, hence, increased immune system fitness, in combination with off-label antiviral treatment, should be pursued early in persistent HEV infection, as recommended by current EASL guidelines (14). In many cases the use of ribavirin alone is preferred due to the increased risk of rejection in organ transplant recipients after reduction in immunosuppression or administration of peginterferon alfa (15). However, ribavirin is shown to increase heterogeneity in the viral population, especially in ORF1 in patients who did not achieve sustained virological response (SVR) (16). If ribavirin treatment fails, mutations may occur that enhance viral fitness (16-18). One such example is the G1634R mutation in the C-terminal region of the HEV-3 RdRp, which significantly increases viral replication efficiency and infectivity

compared to the wild-type (19). A clinical observation supports the effect of the 1634R mutant, by showing that plasma HEV-RNA levels are significantly increased in patients in whom this mutation was detected (18). These findings suggest that G1634R may cause antiviral resistance by enhancing HEV replication. Approximately 10% of patients with chronic HEV infection do not respond to the treatment approaches mentioned above, with an increased risk of rapid progression into cirrhosis and liver failure (14). Thus, novel treatment options are urgently needed.

1.4 Viral evolution

Evolutionary and population studies of HEV show that ancestors of the virus infecting animals might have adapted, and thus, have become infectious for humans (20). Furthermore, as already described, there are significant differences in clinical course and severity between different genotypes (21, 22) which indicates that variability influences pathogenesis in HEV infection. Structural variations in HEV have been repeatedly described in samples from chronically infected patients, most frequently as insertions in the hypervariable region (HVR) (23-25). The origin of inserted sequences is mostly in HEV; however, sequences from the human genome have also been described that have been associated with adaptability in cell culture and increased viral replication potential (26-29). Although variability in the HEV genome is increased by the RdRp, which has no proof-reading function during the transcription process, the exact mechanisms leading to insertions are not fully understood. In addition, antiviral drugs and host immune responses are increasing HEV variability through selection pressure (28). HVR shows sequence heterogeneities and length variations between HEV strains and genotypes and can withstand deletions and insertions (30). A 282 bp-insertion with a fragment derived from the RdRp in the HVR of an isolate from a patient with chronic hepatitis E is associated with increased replication (17). Therefore, variance in HVR may be associated with HEV pathogenesis and clinical outcome. Furthermore, greater heterogeneity of the viral population during the acute phase of infection is associated with HEV persistence (28, 31, 32).
In contrast, the RdRp domain is a conserved genomic region that exhibits high homology to RdRps of other RNA viruses such as hepatitis C virus (33). Several HEV mutations such as Y1320H, K1383N, D1384G, K1398R, V1479I, Y1587F, and G1634R were identified in patient-derived HEV isolates and associated with HEV replication capacity *in vitro* (16, 17, 19). Non-synonymous substitutions in the ORF2 causing a change in

epitope structure may lead to escape from the immune response and to chronicity (16). Consequently, there is an urgent need for in-depth analyses of the viral population with detection and tracking of new variants in patients with a chronic HEV infection.

1.5 Next-Generation Sequencing

Public health laboratories are increasingly integrating modern sequencing techniques for surveillance and investigation of foodborne disease, tuberculosis, SARS-CoV-2, HIV, among others (34). Complete genome and next-generation sequencing (NGS) methods were proven to be superior and are being established for more accurate typing and evolutionary analysis. While traditional pathogen testing procedures require clinicians to employ tests that target specific pathogens, NGS has the potential to detect all types of microorganisms in a sample. Thus, clinicians can benefit from the use of NGS in the field of diagnostic microbiology. Recent advances have led to a significant reduction in time and cost of NGS, and the COVID-19 pandemic has pushed some of the applications of NGS into widespread usage, delivering key findings such as information regarding the spread of SARS-CoV-2 within and across countries (35-38), detection and characterization of different viral clades around the world (38), and mutational rates of the SARS-CoV-2 genome (38). In the case of HEV, deep-sequencing demonstrated that ribavirin therapy is associated with an increased HEV heterogeneity (16) and revealed that the G1634R mutant was already present as a minor population before therapy onset and went on to be the dominant strain after 11 months of ribavirin therapy. This demonstrates that NGS technology plays an important role in the research of infectious diseases and may be used diagnostically to identify patients at risk for treatment failure early who may require alternative treatments. Whole-genome sequences are currently needed to support correct subtyping of the HEV. Therefore, a full-length sequencing method adds a major epidemiological value to the HEV field.

1.6 Research question

Ribavirin is used widely as an off-label drug in the treatment of chronic HEV infections. In a substantial number of cases, therapy with ribavirin fails, and there are no other treatment alternatives for clinicians yet. To reconstruct the mechanisms leading to therapy failure, there is an urgent need for accurate and sensitive solutions for detecting mutations

and recombination events in the HEV genome driving viral evolution. One solution may be modern sequencing technologies, such as NGS and single-molecule sequencing. Thus, this study's purpose was to develop a novel method to allow the use of NGS and single-molecule sequencing in samples from chronic HEV patients not achieving SVR after ribavirin therapy and to evaluate if detection of variants that impact ribavirin treatment outcome is thus improved.

2. Methods

2.1 Samples

The evaluation of the newly established molecular methods presented in this work was performed using a total of twenty-one samples from ten patients with chronic hepatitis E infection (CHE). All patients with CHE were immunosuppressed (Table 1). From one patient, five samples collected over a period of six months were available (Table 2). All samples were pseudonymised before usage.

Table 1. Samples and clinical characteristics (adapted from *Table 1* by Papp et al. 2022)

Patient	Sample Name	Sample Type	Cause of Immunodeficiency	Collection Date	Virus Concentration (Copies/ml)
1 [39]	17-0421	Plasma	Atypical hemolytic-uremic syndrome, severe immunoglobulinaemia	10/2017	$4.1x10^8$
1 [39]	17-0420	Stool		10/2017	$1.17x10^9$
1 [39]	17-0534	Plasma		11/2017	$2x10^5$
1 [39]	18-0002	Plasma		01/2018	$6.5x10^7$
1 [39]	18-0056	Plasma		04/2018	$8.5x10^7$
2 [39]	16-0005	Serum	Kidney and pancreas transplantation	01/2016	$4.0x10^5$
2 [39]	17-0371	Plasma		07/2017	$4.4x10^6$
3	16-0016	Serum	Kidney transplantation	01/2016	$5.8x10^4$
3 [39]	17-0535	Plasma		12/2017	$1.6x10^5$
4	17-0426	Plasma	Selective IgG deficiency, T-lymphopenia with helper cells < 200	10/2017	$6x10^6$
4 [39]	18-0058	Plasma		04/2018	$1.3x10^6$
5	18-0063	Plasma	Peripheric T-cell lymphoma	12/2017	$1.8x10^6$
5	18-0064	Stool		12/2017	$1.6x10^7$
5 [39]	18-0066	Plasma		01/2018	$1.2x10^6$
6 [39]	18-0068	Plasma	Kidney transplantation	04/2018	$8.1x10^6$

7	16-0007.1	Serum	Lung transplantation	02/2016	$1.5x10^7$
7	16-0007	Serum		04/2016	$1.6x10^7$
8	16-0023	Serum	Kidney transplantation	08/2015	$5.6x10^6$
8	16-0010	Serum		02/2016	$5.4x10^5$
9	16-0026	Serum	Kidney transplantation	04/2016	$1.1x10^6$
10	16-0283	Stool	Kidney transplantation	09/2016	$7.8x10^8$

[39] Adapted from Papp et al., Table 1, Sci Rep. 2022;12(1):1720.

Table 2. Overview of the sequencing methods used for the samples from patient 1 (own representation)

Sample name	Whole genome Sanger	Whole genome Next-Generation Sequencing	Whole genome Oxford Nanopore Technology	Amplicon Next-Generation Sequencing HVR	Amplicon Next-Generation Sequencing RdRp	Amplicon Next-Generation Sequencing – HVR and RdRp pool
17-0420				x	x	x
17-0421	x	x	x	x	x	
17-0534					x	
18-0002					x	
18-0056	x	x	x			

2.2 Sample preparation and long-range PCR

RNA extraction, cDNA synthesis, and lrPCR were performed as described by Papp et al. (39) In brief, RNA was extracted using QIAcube (QIAGEN, Hilden Germany) and the QIAamp Viral RNA Mini kit (QIAGEN, Hilden, Germany) according to the manufacturer's instruction. Quantitative PCRs and genotyping were performed as previously described (40, 41). cDNA synthesis was performed with the SuperScript IV First-Strand Synthesis System (Invitrogen, Thermo Fisher Scientific, Carlsbad, CA, USA). Oligo d(T) primers and 11 µL template RNA were used as described in the manufacturer's protocol, except for the synthesis step, which was modified and carried out at 60 °C for 20 min. The success of the whole HEV genome synthesis was verified using four genome-wide PCRs. For the lrPCR Kapa HiFi Readymix (Sigma-Aldrich, Merck KGaA, Darmstadt, Germany) was used with an optimized program for GC-rich templates for all PCRs. Primers and PCR protocols are listed in the Supplementary Tables S1 and S2 - Papp et al. (39) An agarose gel of 1% was used for electrophoresis and the bands were visualized using the BioDocAnalyze software (Biometra GmbH, Göttingen, Germany).

2.3 Long-range PCR and Illumina sequencing

Detailed description of sequencing is presented by Papp et al. (39) Concisely, a library was prepared for all lrPCR products with the Nextera XT library kit (Illumina, San Diego)followed by a 2 × 250 bp sequencing run on the HiSeq 2500 platform (Illumina Inc., San Diego, USA). The generated files after conversion using bcl2fastq v 1.8.4 software (Illumina Inc., San Diego, USA) and demultiplexing were imported into Geneious version 11.1.5 (Biomatters Ltd, Auckland, NZ) for analysis including adapter and quality trimming using BBDuk Adapter/Quality Trimming Version 37.64 by Brian Bushnell with a minimum Phred score quality of 30. Mapping onto reference sequence wbGER27_RAS (FJ705359.1) was performed using Geneious mapper with medium sensitivity and five iterations. The consensus sequence was generated containing the most common bases.

2.4 Oxford Nanopore sequencing of the long-range PCR products

Single-molecule sequencing was performed using Oxford Nanopore Technologies (ONT) MinION. The detailed method is described by Papp et al. (39) Briefly, to multiplex samples on the high accuracy $1D^2$ flow cells (Oxford Nanopore Technologies, Oxford, UK), the manufacturer's native barcoding protocol (XP-NBD104) was merged with the Ligation $1D^2$ (SQK-LSK308) manufacturer's protocol and barcoding of all samples was done according to the ONT 1D Native barcoding genomic DNA protocol. This was followed by dA-tailing. The products were pooled and sequenced according to the $1D^2$ sequencing protocol. The reads were processed and demultiplexed using ONT Guppy basecaller (Oxford Nanopore Technologies, Oxford, UK) and the resulting fastq files were mapped against a reference (GenBank ID: FJ705359.1) using Geneious mapper set at medium sensitivity in Geneious version 11.1.5 (Biomatters Ltd, Auckland, NZ). The sequences that contain two insertions simultaneously were validated by mapping the Illumina reads against the ONT reads containing the insertions.

2.5 Amplicon-based NGS

As described in detail by Papp et al., target specific primers with flow cell adapter sequence at the 5' end were designed (39). The amplicons generated with these primers that contained the region of interest were sequenced on Illumina MiSeq using the MiSeq Reagent Kit v3 (Illumina, San Diego, USA) resulting in 2 × 300 bp reads. The size of the amplicons is optimized to allow two paired reads to overlap and cover the whole sequence of the amplicon. Thus, merging two reads of a pair generates a single end-to-end-sequence covering potential insertions with high accuracy. Regarding cDNA synthesis, in contrast to the lrPCR protocol, for the amplicon NGS, random hexamers were used.

2.6 Amplicon sequencing and reads processing

The PCR products were purified using magnetic beads purification with the MagSi-NGS Prep Plus kit (magtivio B.V., The Netherlands) followed by DNA concentration was measured by Qubit fluorometer using the double-stranded DNA High Sensitivity Assay kit (Thermo Fisher Scientific). Libraries were generated using the Nextera XT library kit (Illumina, San Diego). Indexed libraries were pooled and sequenced on the Illumina MiSeq

using 2 × 300 bp reads. The generated files containing paired-end reads were paired, trimmed, and mapped for each sample as described by Papp et al. (39)

2.7 Single-nucleotide polymorphisms detection

The processed and mapped reads were used for single-nucleotide polymorphism (SNP) detection with the Geneious “Find Variations/SNPs” tool. The minimum frequency was 0.005 and only regions with a coverage of over 500 were analyzed as described by Papp et al. (39)

2.8 Detection of true biological variants

The rationale behind the calculation of the cut-off for biological variants is described in Papp et al. (39) In brief, besides the use of a high-fidelity RT, a maximum of 11 µL template RNA was used for cDNA synthesis to maximize the amount of cDNA synthesized. A Phred quality score of 30 was used for quality trimming of the reads. The plausibility of the mutations was assessed by comparing the values of different sequencing approaches used to sequence the same sample or with the values detected in follow-up samples. Variants were considered if they had a frequency above 0.5%.

2.9 Proportion of polymorphisms and type of selection

Proportion of polymorphisms were calculated by dividing the number of SNPs over 0.5% detected in a specific region by the number of nucleotides in that region. Type of selection was determined by calculating the ratio between nonsynonymous (dN) and synonymous (dS) substitutions per site (ratio dN/dS). Insertions were excluded from the calculation.

2.10 Detection of insertions

Zero-radius operational taxonomic units (zOTUs) from the HVR-amplicon reads were generated by the UNOISE3-Suite(42). zOTU sequences which represent consensus sequences of clustered highly similar reads were imported into Geneious and mapped to

the reference sequence wbGER27_RAS (FJ705359.1) as described in Papp et al. (39) HVR insertion origine was determined using the BLASTn search engine (https://blast.ncbi.nlm.nih.gov).

2.11 Correlation between insertions and mutations

As described in Papp et al., the long-read sequences were grouped based on their insertions (39). For each insertion group, the frequencies of the G1634R, Y1587F, V1479I, and K1383N mutations were determined separately and compared with the frequencies in the ungrouped dataset. A higher frequency of a mutation in one of the groups compared to the ungrouped frequency would indicate a link between mutation and insertion.

2.12 Accession numbers

The sequences published in Papp et al. can be found in the NCBI GenBank under the following accession numbers: MW837243 to MW837255 (39).

2.13 Ethical approval

This study was approved by the Ethics Committee of Charité Universitätsmedizin Berlin (approval number EA1/367/16). All participating subjects gave their written informed consent and all samples were de-identified for this study. The experiments in this study were conducted in conformity with the relevant guidelines and regulatory requirements.

3. Results

To detect mutations and recombination events that may lead to enhanced viral fitness, and thus, answer the research question of this study, the first step was the development of new methods that facilitate the use of NGS and single-molecule sequencing. In brief, two sequencing approaches were developed, one near full-length approach based on a long-range PCR that amplifies the near full-length genome of the HEV in a 7kb amplicon and another amplicon-based NGS approach that allowed the targeted sequencing of the RdRp and HVR by generating Illumina reads that cover the whole amplicon sequence by one pair of reads (39). The lrPCR products can be used either for single-molecule sequencing, granting an overview of each haplotype or for NGS with high accuracy and allowing low-abundance variant detection. Furthermore, a protocol was developed that allowed multiplex sequencing of the lrPCR products using $1D^2$ ONT technology, enabling deep-sequencing using single-molecule sequencing (39). Here, it was shown that the virus can cumulate insertions in the HVR through the detection of five reads representing the full-length HEV genome that revealed two insertions in the same HEV-molecule (39). To our knowledge, this is the first time such a variant was detected. This detection was possible not only due to single-molecule sequencing technology but also due to the presented preparation protocol that allowed ultra-deep multiplex sequencing of six samples. Moreover, due to the single-molecule sequencing approach, a potential correlation between the RdRp mutations and the HVR insertions could be excluded. Furthermore, due to the full-length sequencing based on Illumina technology, the detection of SNPs with a frequency threshold as low as 0.5% could be detected throughout the whole genome, the lowest in published literature. On the other hand, the amplicon-based sequencing method, although limited by amplicon size, showed to be the most sensitive for insertion detection and more appropriate for targeted screening of samples. Moreover, multiplexing different regions of the genome increases data output only for the cost of two PCRs.

In a further yet unpublished study, twenty-one samples from ten patients with chronic HEV infection were sequenced with the amplicon HVR method to screen the available samples for insertions. The amplicon generation and sequencing were successful for all twenty-one samples and fourteen samples from nine patients had insertions in the HVR. All insertions are listed in Table 3. In four samples, more than one insertion could be detected. Most insertions were fragments from human genes. Interestingly, an insertion

that was detected in one sample was previously described by Lhomme et al. (27) It is an insertion from the human genome that has been described in a patient with chronic HEV infection, however, in a different HEV-3 subtype, namely HEV-3f, whereas the insertion presented in this work was detected in a subtype 3c. The insertion is 126 bp long, of which the whole inserted sequence originated in the sequence coding for the human tyrosine aminotransferase gene (TAT). Compared to the sequence described by Lhomme et al., this sequence was in-frame, and two parts of the insertion are inserted in reverse order (Table 3). Interestingly, the AHNAK insertion described in Papp et al. (39) had a similar insertion pattern. A different sample pair, collected from a patient from whom the strain 47832c was originally isolated (26), was screened with the amplicon-based method. The 47832c strain was shown to efficiently replicate in A549/D3 cells and has been used widely in published studies (43, 44). Sequencing the HVR of these two follow-up samples revealed that the insertion persisted over at least four years in the patient (unpublished data). The insertion originates from duplications of an adjacent part of its HVR and of a part of its RdRp region.

Another interesting finding was the insertion with sequence deriving from human coagulation factor V (F5), which was detected in the sample 18-0066 (Table 3). The variant containing this insertion was the most prevalent strain, although the plasma (18-0063) and stool (18-0064) samples collected simultaneously ten days earlier than sample 18-0066 showed no insertion (unpublished data). Furthermore, a second insertion from the human AHNAK gene was detected in one sample (Table 3). However, this insertion has its origin in a tandem repeat region of the gene and has no similarity to the AHNAK sequence previously described (39).

As for the amplicon-based NGS of the RdRp, the presented method delivered important findings in the clinical context. There, it demonstrated that mutations such as G1634R and K1383N were detectable, albeit in very low abundance, before treatment with ribavirin, and an increasing proportion of these variants under ongoing therapy (45). Furthermore, the same method was used for assessing the effectiveness of combination therapy of sofosbuvir and ribavirin on RdRp in a multi-visceral transplanted patient with chronic HEV infection. This was the first reported case of a multi-visceral transplanted patient with chronic HEV infection who failed to clear the HEV infection under combination therapy with sofosbuvir and ribavirin. The method presented here shows that a stepwise accumulation of ribavirin-associated mutations leads to treatment failure (46).

Furthermore, in this work, the dynamics of described RdRp mutations was investigated using NGS. For this purpose, five samples from a patient under ribavirin therapy collected within six months were sequenced. Mutations G1634R, Y1587F, V1479I, and K1383N were examined. Notably, already after the second time-point, all indicated mutations increased. The Y1587F and K1383N mutations even increased from values below the cut-off at the first-time point to about 90% and 98% of the reads at the third time-point. At time-point four, the mutations G1634R and V1479I increased even further, reaching 50% and 8.5%, respectively (39). This study shows that although the lrPCR and consequent whole-genome NGS and single-molecule sequencing is an elegant approach that facilitates whole-genome sequencing including variant detection and haplotype characterization, the amplicon-based targeted deep-sequencing should be the method of choice in screening HEV samples for insertions or mutations. These new methods are not just suitable for detecting recombination events that occur with very low frequency in the viral population but also demonstrated utility in detecting resistant strains in a clinical setting when antiviral therapeutic success requires objectification.

Table 3. BLASTn results of HVR insertions (own representation)

Amplicon-based NGS					
Sample	**Source of Insertion**	**Position in HEV[a] (bp)**	**Length (bp)**	**BLASTn Results Position on Insertion : Position in Source (bp)**	
16-0005 and 17-0371	HEV – RdRp/HVR	2,336	186	34-104 : 4,949-5,019 Identity with FJ705359.1: 88.7% Mismatches: 8	105-159 : 2,254-2,308 Identity with FJ705359.1: 85.5% Mismatches: 8
16-0016 and	HEV - HVR	2,329	99	Identity with FJ705359.1: 96.7% Mismatches: 4	
17-0535	HEV - HVR	2,329	99	Identity with FJ705359.1: 91.8% Mismatches: 10	
17-0426 and 18-0058	Homo sapiens AHNAK nucleoprotein (AHNAK)	2,246	204	1-105 : 10,011-10,115 Identity with NG_051822.1: 99% Mismatches: 1	103-204 : 10,188-10,289 Identity with NG_051822.1: 99% Mismatches: 1
17-0426	HEV – RdRp/HVR	2,361	132	1-37 : 4,885-4,921 Identity with FJ705359.1: 97% Mismatches: 1	56-132 : 2,284-2,360 Identity with FJ705359.1: 90% Mismatches: 8

17-0462 and 18-0058	Homo sapiens glutathione S-transferase alpha 1 (GSTA1)	2,353	138	1-138 : 563-700 Identity with NM_145740.5: 100% Mismatches: 0	
18-066	Homo sapiens coagulation factor V (F5)	2,238	207	1-207 : 5,705-5,911 Identity with AH005274.2: 99% Mismatches: 2	
16-0007 and 16-0007.1	Homo sapiens utrophin (UTRN)	2,241	168	1-168 : 484-660 Identity with NM_007124.3: 92% Mismatches: 15	
16-0010	Homo sapiens tyrosine aminotransferase (TAT)	2,329	126	54-125 : 80-151 Identity with NM_000353.3: 96% Mismatches: 3	1-60 : 153-212 Identity with NM_000353.3: 95% Mismatches: 3
16-0026	Homo sapiens guanylyl cyclase domain containing 1 (GUCD1)	2,349	96	2-62 : 2,792-2,852 Identity with NR_104286.2: 97% Mismatches: 2	57-94 : 2,754-2,791 Identity with NR_104286.2: 100% Mismatches: 0
16-0283	Homo sapiens calmin (CLMN)	2,319	150	1-106 : 622-727	106-150 : 125-169

				Identity with NM_024734.4: 98% Mismatches: 2	Identity with NM_024734.4: 100% Mismatches: 0
16-0283	Homo sapiens spen family transcriptional repressor (SPEN)	2,229	84	2-84 : 7,298-7,380 Identity with NM_015001.3: 98% Mismatches: 2	
16-0283	Homo sapiens heterogeneous nuclear ribonucleoprotein D (HNRNPD)	2,502	102	1-102 : 400-501 Identity with NM_031370.3: 99% Mismatches: 1	
17-0420 and 17-0421[b]	HEV – RdRp/HVR	2,360	162	1–88 : 4,544-4,631 Identity with FJ705359.1: 95.5% Mismatches: 4	88-162 : 2,285-2,359 Identity with FJ705359.1: 88% Mismatches: 9
17-0420 and 17-0421[b]	Homo sapiens AHNAK nucleoprotein (AHNAK)	2,227	138	1-30 : 27,578-27,607 Identity with NG_051822.1: 100% Mismatches: 0	27-138 : 27,466-27,577 Identity with NG_051822.1: 98.2% Mismatches: 2

17-0420 and 17-0421[b]	Homo sapiens ribosomal protein L18 (RPL18-2), variant 1	2,273	150	4-150 : 160-306 Identity with L11566.1: 98% Mismatches: 3	
17-0420 and 17-0421[b]	Homo sapiens ribosomal protein L18 (RPL18-2), variant 2	2,246	105	1-105 : 202-306 Identity with L11566.1: 98.1% Mismatches: 2 deletion of HEV sequence from 2,246 to 2,275	
17-0420 and 17-0421[b]	HEV – RdRp/HVR	2,360	162	1-88 : 4,544-4,631 Identity with FJ705359.1: 95.5% Mismatches: 4	88-162 : 2,285-2,359 Identity with FJ705359.1: 88% Mismatches: 9
17-0420 and 17-0421[b]	HEV – HVR	2,393	60	1-60 : 2,333-2,392 Identity with FJ705359.1: 96.7% Mismatches: 2	

[a]Using wbGER27 as reference (accession No: FJ705359.1)

[b]Papp et al. (39)

4. Discussion

4.1 Short summary of results

The key results of this work are the new sequencing approaches that include a unique long-range PCR protocol for full-length genome amplification, an efficient multiplex single-molecule ultra-deep sequencing protocol, and an amplicon-based NGS protocol for targeted ultra-deep sequencing. These new methods in turn enabled the discovery of unique findings, which include the numerous insertions detected in the HVR, the first description of a variant with cumulated insertions, and the dynamics of the RdRp mutations associated with treatment failure in patients undergoing antiviral ribavirin therapy. Thus, the hypothesis of this work was verified, especially in regards to NGS and single-molecule sequencing being more adequate than standard Sanger sequencing-based methods in detecting variants that impact ribavirin treatment failure in patients with chronic HEV infection. Therefore, this study's findings improve the detection potential for mutations and recombination events that could lead to enhanced viral fitness and/or therapy failure in HEV infections.

4.2 Interpretation of results

The development of a long-range PCR protocol for full-length HEV genome amplification was difficult and time-consuming, yet its necessity was highlighted by the lack of such protocols for HEV in the published literature. Long-range PCR refers to the amplification of DNA targets over 5 kb which normally cannot be amplified using routine PCR methods or reagents. The key factor that facilitated successful amplification of the near-full length HEV genome was the higher temperature during cDNA synthesis and elongation step of the PCR as well as longer primers. Although the manufacturer's protocol was used and no changes in salt concentration of the PCR mix were made, optimization steps showed that the lrPCR was most efficient at an elongation temperature of 74 °C, higher than most PCR protocols. Furthermore, six different primer combinations were tested to achieve the optimal PCR. The protocol presented here was developed due to the lack of full-length PCR protocols with a single pair of primers available in the literature for HEV. Single-molecule sequencing facilitates sequencing of the whole genome in one piece, thus full-

length amplification without genome fragmentation allows for exact haplotyping of the HEV population. The only protocol in the literature for single-molecule sequencing of the HEV genome from patients with HEV infection uses an amplification approach in two overlapping fragments (47). There, the full-length sequences of 114 HEV samples were generated using single-molecule sequencing with consequent quasispecies analysis (48). However, haplotype reconstruction from a set of reads representing only fragments of the genome is theoretical and one of the most challenging problems in bioinformatics today. Also, it excludes the possibility of assessing the correlation between events in distant parts of the genome. Similar to the method presented in this work, a protocol for a near full-length long-range PCR was published; however, this was developed for an avian HEV strain (49). The authors of this study propose the use of whole genome sequencing of pathogens directly from clinical samples to reduce the risk of genomic modifications due to cell culture passaging. Furthermore, HEV cannot be propagated easily in cell culture, making whole genome sequencing directly from clinical samples the most suitable option for the characterisation of viral isolates.

As for single-molecule sequencing, due to its higher accuracy, the ONT $1D^2$ flow cell was chosen. However, for this type of flow cell, there was no protocol for multiplexing samples. Therefore, a multiplexing protocol using the multiplexing kit designed for the 1D sequencing kit was developed. This led to multiplex sequencing of six samples, generating a substantial output. The number of reads per sample with sizes between 6.5 – 8 kb ranged from 11,000 to more than 68,000 reads. As shown in Papp et al., in the dataset of one sample, five full-length HEV reads were detected that contained two insertions simultaneously (39). This variant was confirmed using the Illumina NGS dataset by mapping the Illumina reads onto the ONT read. Thus, Illumina reads were detected that contained sequences of both insertions simultaneously (39). Unfortunately, Lhomme et al. did not publish the sequencing statistics and depth (48); however, due to the low count of variants found, the depth was presumably low, and thus, only high-abundance variants were detected. A significant challenge regarding the single-molecule reads was the analysis of the dataset generated. The alignment of tens of thousands of reads with the length of the HEV genome that were generated per sample exceeded the computational power of the bioinformatics department at the Robert Koch Institute. Illumina reads were available for one sample and could be used for validation of ONT reads, and therefore, only the ONT reads from this particular sample were analyzed in detail. The ONT reads from the re-

maining samples will be analyzed in more detail in the future with more streamlined analysis pipelines. Nevertheless, third-generation sequencing is less suitable for the detection of polymorphisms due to its higher error rate compared to NGS (50).

The amplicon-based sequencing was the most efficient method regarding sequencing preparation and dataset processing. Five different insertions could be detected in one sample using this method, whereas using whole genome sequencing only two insertions were detected. The divergent results regarding variant calling in whole-genome and targeted NGS datasets can be explained by sampling biases, different cDNA synthesis methods, and possible strand selection of PCR primers. As previously mentioned, the detection of recombined variants is less likely successful using other sequencing approaches, primarily due to their relatively lower proportion in the viral population. Furthermore, the targeted NGS approach detected insertions in thirteen out of twenty-one samples from patients with chronic HEV. Nine out of eleven patients had, in at least one sample, HEV variants with at least an insertion in the HVR (unpublished data). Lhomme et al. described insertions in 7/114 patients of which all seven were immunocompromised. Interestingly, three recombination events were detected in patients with acute HEV infection (48). Nevertheless, Lhomme et al. used single-molecule sequencing, and as described above, probably only the variants with the highest abundance were detected. Interestingly, an insertion previously described by Lhomme et al. was discovered in one of the samples presented in this work, namely the TAT insertion from the human genome that has been described in a patient with chronic HEV infection (27). The TAT insertion found in sample 16-0010 (Table 3) is 126 bp long and originates completely in the human genome (unpublished data). Compared to the sequence described earlier, the newly found sequence is in-frame, rearranged and inserted in inversed order. Interestingly, the AHNAK insertion described in this work had a similar insertion pattern (39). Furthermore, two samples were sequenced that were collected from the same patient from which the 47832c strain was isolated, and it was found that the variant with an insertion that has been shown to increase viral replication in cell culture (26, 43), persisted over four years in this patient (unpublished data). Thus, both *in vitro* and *in vivo* data suggest an increased viral fitness of this variant.

In Papp et al. (39), the proportions of known minority variants detected using different sequencing approaches in one plasma and one stool sample collected from one patient at the same time-point were compared. In both samples, G1634R mutation was detected at a frequency as low as 1.1%. Notably, this mutation was detectable already in the first

sample in low abundance and was being selected in the course of infection (39). This consolidates the findings from other studies showing the presence of G1634R as a minority variant using NGS in an early stage of infection and before initiation of ribavirin (16, 45). The amplicon-based method presented here was also used successfully to assess the effects of ribavirin and sofosbuvir combination therapy on RdRp mutations in a multivisceral transplant patient, where it showed that G1634R increased in frequency while combination therapy failed (46). Thus, this method has demonstrated utility in assessing the dynamics of mutations in patients with CHE, while facilitating the pooling of amplicons from different regions, such as RdRp and HVR, thereby maximizing the data obtained and simultaneously reducing costs. Due to its highly conserved nature, the sequences from the RdRp region were used to validate the variability in the HVR reads. HVR showed twice as many polymorphic loci as the RdRp region, and the ratio between non-synonymous and synonymous SNPs indicated positive selection in the HVR whereas the RdRp experienced negative selection. The HVRs in both stool and plasma samples were positively selected, which indicates that changes in the amino acid sequence in this region may contribute to viral replication and fitness. Previous studies have shown that positive selection in the HVR is characteristic of the zoonotic HEV genotypes HEV-3 and HEV-4 (20).

4.3 Embedding the results into the current state of research

EASL guidelines recommend reduction of immunosuppression as well as ribavirin treatment to achieve viral freedom in patients with chronic HEV infection (14). However, ribavirin therapy increases the mutation rate in HEV, and some mutations associated with severe forms of infection and potentially antiviral resistances occur during treatment. Most mutations associated with treatment failure mostly due to higher replication competence occur in the RdRp region, which makes this region of clinical relevance (16, 17).
In HEV, NGS facilitated an in-depth analysis of the viral population, showing ribavirin therapy is associated with an increased HEV heterogeneity (16). Using the novel amplicon-based NGS method, the dynamics of five mutations associated with therapy failure or increased replication potential was analyzed, and it was shown that their proportion increased substantially throughout ribavirin therapy. Similarly, separate studies showed that using this novel sequencing approach the mutations K1383N and G1634R were de-

tected in low proportion even before the initiation of antiviral treatment (45) and in increased proportion after failure of combination therapy of sofosbuvir and ribavirin (46). The latter study was the first report of a patient with chronic HEV infection who failed to achieve SVR under combination therapy with sofosbuvir and ribavirin. There, the amplicon-based deep sequencing method showed a stepwise accumulation of ribavirin-associated mutations leading to treatment failure. These results suggest that mutations associated with ribavirin treatment failure are present in the viral population before therapy onset and are being selected during ribavirin therapy. Thus, this novel approach should be considered in CHE patients who are at increased risk for selection of variants with higher pathogenicity.

Insertions in the ORF1 that increase HEV heterogeneity have been repeatedly described and are influencing pathogenesis and transmission patterns (27). This work contributes with a large number of insertions detected in the HEV using novel sequencing approaches, which may indicate that recombination events occur more frequently than previously thought, and it shows that the method of choice for detecting HVR insertions is the amplicon-based approach. However, single-molecule sequencing by ONT detected a viral variant with two insertions that accumulated on the same strand. This was the first report of an HEV variant with mixed insertions. Furthermore, using the single-molecule sequencing reads a correlation between mutations in the RdRp and recombination events in the HVR could be ruled out. Thus, these are probably unrelated mechanisms that affect viral diversity and adaptation in different ways. Some HEV variants containing an HVR insertion have been shown to efficiently replicate in cell culture (26, 43, 51), including insertions from the human genome such as a 174 bp insertion of the human RPS17 gene (23, 52). In a study where the HVR from 27 immunocompromised patients with HEV persistence, infections were characterized, and recombination occurred in three HEV strains (27). Insertions of human origin, such as inter-α-trypsin inhibitor (ITI), and tyrosine aminotransferase (TAT), were shown to enhance HEV replication, likely by providing new potential regulatory sites (27). Interestingly, using the amplicon-based NGS method, the same TAT sequence was detected in a different patient (16-0010, Table 3) using the methods presented herein, compounding the evidence of a higher selective advantage of the variant with this particular insertion. Furthermore, an insertion was detected in the samples 16-0005 and 17-0371 (Table 3) that persisted over four years in the samples from a patient from which the strain 47832c was originally isolated. This strain was shown to efficiently replicate in cell culture and has been used widely in basic and applied studies

(26, 43, 51). The insertion contains duplications of an adjacent part of its HVR and a part of its RdRp region. In a recent paper, several mutants containing this insertion were tested in cell culture and showed that removal of the insertion stopped the replication process. Furthermore, a frameshift of the inserted sequence rendered the virus non-infectious, whereas a mutant with synonymous codons in the insertion replicated similarly to the wild type. Therefore, the authors concluded that it is not the RNA sequence but the translated amino acid sequence of the insertion that is responsible for the effects of an insertion (51). These *in vivo* and *in vitro* findings demonstrate the increased viral fitness of the variant containing this insertion. In a different recent study, sixteen samples from individuals with acute or chronic HEV infection were analyzed by Biedermann et al. using the targeted NGS approach presented in this work (53). This method allowed detection of insertions, deletions, or both in the HVR of seven samples. Interestingly, more than one recombinant variant was found in each sample. One of the insertions described was detected in a sample from a patient with acute infection who later developed chronic infection. Using sequences discovered with this novel sequencing approach, Biedermann et al. demonstrated that HVR sequences with insertions or deletions had significantly more acetylation sites as well as an increase in presumed methylation sites. These results also support the hypothesis described by Scholz et al. mentioned above that amino acid sequence rather than nucleotide sequence is important for the persistence of HEV infection. Together, these findings suggest that diversity in HVR contributes to chronification. However, the precise effects of insertions on the chronic course of infection, as well as viral genome stability and replication remains unclear. Therefore, there is an obvious need for robust methods that facilitate the use of modern technologies, such as NGS and single-molecule sequencing, for understanding the pathophysiology of HEV and the mechanisms that lead to persistence and adaptation in the host. NGS technology has already demonstrated and established its importance in research and diagnosis of pathogens, most recently in the SARS-CoV-2 pandemic, by tracking variants and monitoring vaccine response, and thus, fighting the spread of the virus. The identification of new SARS-CoV-2 strains in different regions of the world has resulted in the wider use of NGS technology in diagnostic laboratories (36). In HEV, the amplicon-based NGS method detected up to five insertions in one sample (Table 3), while in a different study, the novel method presented in this work demonstrated its utility by detecting insertions, deletions, or both in samples from acute and chronic HEV infections as described above (53). It is therefore evident that this approach is best suited for screening HEV samples with high accuracy

while facilitating multiplexing. However, this approach is also of great value for studying SNP dynamics associated with antiviral therapy failure. Its use in monitoring mutations to evaluate standard and novel therapeutic options has been highlighted by Schulz et al. (46)

4.4 Strengths and weaknesses of the study

The main strength of this work is that both amplicon-based and full-length genome sequencing involving lrPCR are robust methods that are reproducible and were able to capture HEV variability, which in essence, is the driving force in viral evolution. Furthermore, the amplicon-based method was shown to be of significance in assessing the effects of antiviral therapy on the HEV, and thus, permits a shift to personalized treatments by generating high-throughput data, which in turn, enables tailored therapy strategies in patients with chronic HEV infections.

However, it is worth noting that all samples from patients with chronic HEV infections presented in this work were collected from patients who received ribavirin therapy. At the time of sample collection and analysis, every patient received early ribavirin therapy as soon as the HEV infection was considered chronic. Therefore, a comparison with control samples from treatment-naïve patients with chronic HEV infection is lacking. However, it has been previously shown that ribavirin increases the error rate of the RdRp and some mutations have been previously associated with ribavirin therapy. Thus, the focus was on these already described mutations. Therefore, the exact aetiology of other mutations cannot be determined in this study. However, the presented advanced sequencing methods can be used to elucidate this question in further studies.

A limitation of the lrPCR is the lower efficiency in samples with low viral loads compared to shorter PCR target sequences. Samples with low viral loads however are seldom of great interest for single-molecule sequencing. Also noteworthy is the incomplete coverage of the coding region of the lrPCR and the somewhat limited number of samples used for validation and testing. Also, this study was performed only with HEV subtype 3c which is the most prevalent subtype in Germany. Most likely, the approaches described herein will be applicable also to other subtypes with small adaptations. Regarding sequencing, limitations of the presented methods are addressed in Papp et al. (39). In brief, amplicon-based sequencing is limited by the length of the insertion. If the insertion exceeds a certain length and the PCR product containing the insertion exceeds the length of two paired-

end reads, some insertions may be missed during the mapping process. Also, designing amplicons that match the criteria for amplicon-based sequencing can be challenging. Exemplarily, PCR products that exceeded the optimal length led to a decrease in coverage in the middle sequences of the RdRp amplicons. Long-read sequencing technologies, on the other hand, have relatively high error rates, which impedes accurate variant analysis. Therefore, to overcome this limitation, simultaneous sequencing was performed with both Illumina and ONT technologies. Finally, it must be noted that no studies were undertaken to investigate the impact of insertions on HEV replication, genome stability, or pathogenesis, which is a limitation of this study. Here, the focus was on detecting and validating insertions using advanced sequencing techniques. The mechanisms underlying recombination with fragments of human or viral sequences remain unclear.

4.5 Implications for practice and future research

As demonstrated above, the presented methods are of great value for the HEV field with implications in research, diagnostic and therapy. The lrPCR is an elegant method and a powerful tool that facilitates whole-genome sequences, which are currently needed for more accurate subtyping of the HEV. Therefore, a full-length sequencing method adds a major epidemiological value to the HEV field. Furthermore, single-molecule sequencing offers a closer representation of the viral haplotypes and an overview of the HEV genome and rearrangement events. The amplicon-based sequencing is a versatile method with multiple applications due to the advantages outlined in this work. As presented above, the method can be used to assess viral adaptation in the context of existing or novel antiviral therapies, and thus, it can support personalized medicine. Assessment of viral evolution or screening for SNPs and/or recombination events, while allowing multiplexing and reducing NGS costs makes this approach powerful. The great number of insertions described will help in further analysis of the HEV physiology and pathology. In clinical practice, systematic analysis of HEV variants using NGS should be considered to predict therapy outcomes and the progression of liver diseases.

5. Conclusion

HEV can cause chronic infections in immunocompromised individuals that can result in cirrhosis and liver failure. EASL guidelines recommend antiviral therapy with ribavirin for chronic HEV infections; however, ribavirin was shown to increase viral diversity, resulting in therapy failure. The presented methods herein can capture viral diversity accurately in samples from patients with chronic HEV infection, allowing clinicians and scientists to monitor variants associated with therapy failure and to detect new variants that may lead to the persistence of the infection. Structural variations and substitutions in the HEV genome may affect replication and virus-host interaction, and thus, may be linked with pathogenesis. Further studies will help determine the potential contribution of HEV variants in HEV pathogenesis and their clinical relevance. In this regard, the use of NGS and single-molecule sequencing will play a key role.

Reference list

1. WHO. Fact sheet: Hepatitis E [Internet], Access Date: January 24, 2023. https://www.who.int/news-room/fact-sheets/detail/hepatitis-e.

2. Hoerning A, Hegen B, Wingen AM, Cetiner M, Lainka E, Kathemann S, Fiedler M, Timm J, Wenzel JJ, Hoyer PF, Gerner P. Prevalence of hepatitis E virus infection in pediatric solid organ transplant recipients--a single-center experience. Pediatr Transplant. 2012;16(7):742-7.

3. Bigna JJ, Modiyinji AF, Nansseu JR, Amougou MA, Nola M, Kenmoe S, Temfack E, Njouom R. Burden of hepatitis E virus infection in pregnancy and maternofoetal outcomes: a systematic review and meta-analysis. BMC Pregnancy Childbirth. 2020;20(1):426.

4. Parvez MK, Al-Dosari MS. Evidence of MAPK-JNK1/2 activation by hepatitis E virus ORF3 protein in cultured hepatoma cells. Cytotechnology. 2015;67(3):545-50.

5. Kamar N, Bendall R, Legrand-Abravanel F, Xia NS, Ijaz S, Izopet J, Dalton HR. Hepatitis E. Lancet. 2012;379(9835):2477-88.

6. Pérez-Gracia MT, Mateos Lindemann ML, Caridad Montalvo Villalba M. Hepatitis E: current status. Rev Med Virol. 2013;23(6):384-98.

7. Lee GH, Tan BH, Teo EC, Lim SG, Dan YY, Wee A, Aw PP, Zhu Y, Hibberd ML, Tan CK, Purdy MA, Teo CG. Chronic Infection With Camelid Hepatitis E Virus in a Liver Transplant Recipient Who Regularly Consumes Camel Meat and Milk. Gastroenterology. 2016;150(2):355-7.e3.

8. Kamar N, Dalton HR, Abravanel F, Izopet J. Hepatitis E virus infection. Clin Microbiol Rev. 2014;27(1):116-38.

9. Wedemeyer H, Pischke S, Manns MP. Pathogenesis and treatment of hepatitis e virus infection. Gastroenterology. 2012;142(6):1388-97.e1.

10. Spahr C, Knauf-Witzens T, Vahlenkamp T, Ulrich RG, Johne R. Hepatitis E virus and related viruses in wild, domestic and zoo animals: A review. Zoonoses Public Health. 2018;65(1):11-29.

11. Krain LJ, Nelson KE, Labrique AB. Host immune status and response to hepatitis E virus infection. Clin Microbiol Rev. 2014;27(1):139-65.

12. Geng Y, Wang Y. Epidemiology of Hepatitis E. Adv Exp Med Biol. 2016;948:39-59.

13. Bazerbachi F, Haffar S, Garg SK, Lake JR. Extra-hepatic manifestations associated with hepatitis E virus infection: a comprehensive review of the literature. Gastroenterol Rep (Oxf). 2016;4(1):1-15.

14. easloffice@easloffice.eu EAftSotLEa, Liver EAftSot. EASL Clinical Practice Guidelines on hepatitis E virus infection. J Hepatol. 2018;68(6):1256-71.

15. Behrendt P, Steinmann E, Manns MP, Wedemeyer H. The impact of hepatitis E in the liver transplant setting. J Hepatol. 2014;61(6):1418-29.

16. Todt D, Gisa A, Radonic A, Nitsche A, Behrendt P, Suneetha PV, Pischke S, Bremer B, Brown RJ, Manns MP, Cornberg M, Bock CT, Steinmann E, Wedemeyer H. In vivo evidence for ribavirin-induced mutagenesis of the hepatitis E virus genome. Gut. 2016;65(10):1733-43.

17. Debing Y, Ramière C, Dallmeier K, Piorkowski G, Trabaud MA, Lebossé F, Scholtès C, Roche M, Legras-Lachuer C, de Lamballerie X, André P, Neyts J. Hepatitis E virus mutations associated with ribavirin treatment failure result in altered viral fitness and ribavirin sensitivity. J Hepatol. 2016;65(3):499-508.

18. Lhomme S, Kamar N, Nicot F, Ducos J, Bismuth M, Garrigue V, Petitjean-Lecherbonnier J, Ollivier I, Alessandri-Gradt E, Goria O, Barth H, Perrin P, Saune K, Dubois M, Carcenac R, Lefebvre C, Jeanne N, Abravanel F, Izopet J. Mutation in the Hepatitis E Virus Polymerase and Outcome of Ribavirin Therapy. Antimicrob Agents Chemother. 2015;60(3):1608-14.

19. Debing Y, Gisa A, Dallmeier K, Pischke S, Bremer B, Manns M, Wedemeyer H, Suneetha PV, Neyts J. A mutation in the hepatitis E virus RNA polymerase promotes its replication and associates with ribavirin treatment failure in organ transplant recipients. Gastroenterology. 2014;147(5):1008-11.e7; quiz e15-6.

20. Purdy MA, Khudyakov YE. Evolutionary history and population dynamics of hepatitis E virus. PLoS One. 2010;5(12):e14376.

21. Mizuo H, Yazaki Y, Sugawara K, Tsuda F, Takahashi M, Nishizawa T, Okamoto H. Possible risk factors for the transmission of hepatitis E virus and for the severe form of hepatitis E acquired locally in Hokkaido, Japan. J Med Virol. 2005;76(3):341-9.

22. Aggarwal R, Jameel S. Hepatitis E. Hepatology. 2011;54(6):2218-26.

23. Shukla P, Nguyen HT, Torian U, Engle RE, Faulk K, Dalton HR, Bendall RP, Keane FE, Purcell RH, Emerson SU. Cross-species infections of cultured cells by hepatitis E virus and discovery of an infectious virus-host recombinant. Proc Natl Acad Sci U S A. 2011;108(6):2438-43.

24. Wang H, Zhang W, Ni B, Shen H, Song Y, Wang X, Shao S, Hua X, Cui L. Recombination analysis reveals a double recombination event in hepatitis E virus. Virol J. 2010;7:129.

25. Smith DB, Simmonds P, Jameel S, Emerson SU, Harrison TJ, Meng XJ, Okamoto H, Van der Poel WH, Purdy MA, Group ICoToVHS. Consensus proposals for classification of the family Hepeviridae. J Gen Virol. 2014;95(Pt 10):2223-32.

26. Johne R, Reetz J, Ulrich RG, Machnowska P, Sachsenröder J, Nickel P, Hofmann J. An ORF1-rearranged hepatitis E virus derived from a chronically infected patient efficiently replicates in cell culture. J Viral Hepat. 2014;21(6):447-56.

27. Lhomme S, Abravanel F, Dubois M, Sandres-Saune K, Mansuy JM, Rostaing L, Kamar N, Izopet J. Characterization of the polyproline region of the hepatitis E virus in immunocompromised patients. J Virol. 2014;88(20):12017-25.

28. Lhomme S, Garrouste C, Kamar N, Saune K, Abravanel F, Mansuy JM, Dubois M, Rostaing L, Izopet J. Influence of polyproline region and macro domain genetic heterogeneity on HEV persistence in immunocompromised patients. J Infect Dis. 2014;209(2):300-3.

29. Kenney SP, Meng XJ. The lysine residues within the human ribosomal protein S17 sequence naturally inserted into the viral nonstructural protein of a unique strain of hepatitis E virus are important for enhanced virus replication. J Virol. 2015;89(7):3793-803.

30. Pudupakam RS, Huang YW, Opriessnig T, Halbur PG, Pierson FW, Meng XJ. Deletions of the hypervariable region (HVR) in open reading frame 1 of hepatitis E virus do not abolish virus infectivity: evidence for attenuation of HVR deletion mutants in vivo. J Virol. 2009;83(1):384-95.

31. Lhomme S, Abravanel F, Dubois M, Sandres-Saune K, Rostaing L, Kamar N, Izopet J. Hepatitis E virus quasispecies and the outcome of acute hepatitis E in solid-organ transplant patients. J Virol. 2012;86(18):10006-14.

32. Suneetha PV, Pischke S, Schlaphoff V, Grabowski J, Fytili P, Gronert A, Bremer B, Markova A, Jaroszewicz J, Bara C, Manns MP, Cornberg M, Wedemeyer H. Hepatitis E virus (HEV)-specific T-cell responses are associated with control of HEV infection. Hepatology. 2012;55(3):695-708.

33. Agrawal S, Gupta D, Panda SK. The 3' end of hepatitis E virus (HEV) genome binds specifically to the viral RNA-dependent RNA polymerase (RdRp). Virology. 2001;282(1):87-101.

34. CDC. Advanced Molecular Detection. https://www.cdc.gov/amd/: U.S. Centers for Disease Control and Prevention, Access Date: October 01, 2022.

35. Yadav PD, Potdar VA, Choudhary ML, Nyayanit DA, Agrawal M, Jadhav SM, Majumdar TD, Shete-Aich A, Basu A, Abraham P, Cherian SS. Full-genome sequences of the first two SARS-CoV-2 viruses from India. Indian J Med Res. 2020;151(2 & 3):200-9.

36. John G, Sahajpal NS, Mondal AK, Ananth S, Williams C, Chaubey A, Rojiani AM, Kolhe R. Next-Generation Sequencing (NGS) in COVID-19: A Tool for SARS-CoV-2 Diagnosis, Monitoring New Strains and Phylodynamic Modeling in Molecular Epidemiology. Curr Issues Mol Biol. 2021;43(2):845-67.

37. Lorusso A, Calistri P, Mercante MT, Monaco F, Portanti O, Marcacci M, Cammà C, Rinaldi A, Mangone I, Di Pasquale A, Iommarini M, Mattucci M, Fazii P, Tarquini P, Mariani R, Grimaldi A, Morelli D, Migliorati G, Savini G, Borrello S, D'Alterio N. A "One-Health" approach for diagnosis and molecular characterization of SARS-CoV-2 in Italy. One Health. 2020;10:100135.

38. Wang R, Hozumi Y, Yin C, Wei GW. Decoding SARS-CoV-2 Transmission and Evolution and Ramifications for COVID-19 Diagnosis, Vaccine, and Medicine. J Chem Inf Model. 2020;60(12):5853-65.

39. Papp CP, Biedermann P, Harms D, Wang B, Kebelmann M, Choi M, Helmuth J, Corman VM, Thürmer A, Altmann B, Klink P, Hofmann J, Bock CT. Advanced sequencing approaches detected insertions of viral and human origin in the viral genome of chronic hepatitis E virus patients. Sci Rep. 2022;12(1):1720.

40. Wang B, Harms D, Papp CP, Niendorf S, Jacobsen S, Lütgehetmann M, Pischke S, Wedermeyer H, Hofmann J, Bock CT. Comprehensive Molecular Approach for Characterization of Hepatitis E Virus Genotype 3 Variants. J Clin Microbiol. 2018;56(5).

41. Jothikumar N, Cromeans TL, Robertson BH, Meng XJ, Hill VR. A broadly reactive one-step real-time RT-PCR assay for rapid and sensitive detection of hepatitis E virus. J Virol Methods. 2006;131(1):65-71.

42. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. 2016:081257.

43. Johne R, Trojnar E, Filter M, Hofmann J. Thermal Stability of Hepatitis E Virus as Estimated by a Cell Culture Method. Appl Environ Microbiol. 2016;82(14):4225-31.

44. Johne R, Wolff A, Gadicherla AK, Filter M, Schlüter O. Stability of hepatitis E virus at high hydrostatic pressure processing. Int J Food Microbiol. 2021;339:109013.

45. Gerhardt F, Maier M, Liebert UG, Platzbecker U, Wang SY, Papp CP, Bock CT, Berg T, van Bömmel F. Early Detection of Hepatitis E Virus Ribavirin Resistance Using Next-Generation Sequencing. Antimicrob Agents Chemother. 2019;64(1).

46. Schulz M, Papp CP, Bock CT, Hofmann J, Gerlach UA, Maurer MM, Eurich D, Mueller T. Combination therapy of sofosbuvir and ribavirin fails to clear chronic hepatitis E infection in a multivisceral transplanted patient. J Hepatol. 2019;71(1):225-7.

47. Nicot F, Jeanne N, Roulet A, Lefebvre C, Carcenac R, Manno M, Dubois M, Kamar N, Lhomme S, Abravanel F, Izopet J. Diversity of hepatitis E virus genotype 3. Rev Med Virol. 2018;28(5):e1987.

48. Lhomme S, Nicot F, Jeanne N, Dimeglio C, Roulet A, Lefebvre C, Carcenac R, Manno M, Dubois M, Peron JM, Alric L, Kamar N, Abravanel F, Izopet J. Insertions and Duplications in the Polyproline Region of the Hepatitis E Virus. Front Microbiol. 2020;11:1.

49. Asif K, O'Rourke D, Sabir AJ, Shil P, Noormohammadi AH, Marenda MS. Characterisation of the whole genome sequence of an avian hepatitis E virus directly from clinical specimens reveals possible recombination events between European and USA strains. Infect Genet Evol. 2021;96:105095.

50. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 2015;3:1-8.

51. Scholz J, Falkenhagen A, Johne R. The Translated Amino Acid Sequence of an Insertion in the Hepatitis E Virus Strain 47832c Genome, But Not the RNA Sequence, Is Essential for Efficient Cell Culture Replication. Viruses. 2021;13(5).

52. Kenney SP, Meng XJ. Identification and fine mapping of nuclear and nucleolar localization signals within the human ribosomal protein S17. PLoS One. 2015;10(4):e0124396.

53. Biedermann P, Klink P, Nocke MK, Papp CP, Harms D, Kebelmann M, Thurmer A, Choi M, Altmann B, Todt D, Hofmann J, Bock CT. Insertions and deletions in the hypervariable region of the hepatitis E virus genome in individuals with acute and chronic infection. Liver Int. 2023.

Statutory Declaration

“I, Christian Patrick Papp, by personally signing this document in lieu of an oath, hereby affirm that I prepared the submitted dissertation on the topic - Identification of Ribavirin-Induced Mutations in Patient-Derived Hepatitis E Virus (Identifizierung von Ribavirin-induzierten Mutationen in von Patienten stammenden Hepatitis-E-Viren), independently and without the support of third parties, and that I used no other sources and aids than those stated.

All parts which are based on the publications or presentations of other authors, either in letter or in spirit, are specified as such in accordance with the citing guidelines. The sections on methodology (in particular regarding practical work, laboratory regulations, statistical processing) and results (in particular regarding figures, charts and tables) are exclusively my responsibility.

Furthermore, I declare that I have correctly marked all of the data, the analyses, and the conclusions generated from data obtained in collaboration with other persons, and that I have correctly marked my own contribution and the contributions of other persons (cf. declaration of contribution). I have correctly marked all texts or parts of texts that were generated in collaboration with other persons.

My contributions to any publications to this dissertation correspond to those stated in the below joint declaration made together with the supervisor. All publications created within the scope of the dissertation comply with the guidelines of the ICMJE (International Committee of Medical Journal Editors; http://www.icmje.org) on authorship. In addition, I declare that I shall comply with the regulations of Charité – Universitätsmedizin Berlin on ensuring good scientific practice.

I declare that I have not yet submitted this dissertation in identical or similar form to another Faculty.

The significance of this statutory declaration and the consequences of a false statutory declaration under criminal law (Sections 156, 161 of the German Criminal Code) are known to me.”

Date Signature

Declaration of your own contribution to the publications

Publication: C.-Patrick Papp, Paula Biedermann, Dominik Harms, Bo Wang, Marianne Kebelmann, Mira Choi, Johannes Helmuth, Victor M. Corman, Andrea Thürmer, Britta Altmann, Patrycja Klink, Jörg Hofmann, C.-Thomas Bock,
Advanced sequencing approaches detected insertions of viral and human origin in the viral genome of chronic hepatitis E virus patients,
Scientific Reports, 2022

Hereby, I declare the following methods were conceptualized, designed, and performed by me, alone:

- Sequencing approaches based on the long-range PCR and amplicon generation.
- Primer designs and all PCR protocols.
- Optimization of the long-range PCR and the nested PCRs used for amplicon-based sequencing.
- RNA extraction; cDNA synthesis; amplification, including gel electrophoresis; and purification of the samples analyzed.
- Optimization of the cDNA synthesis method.
- Sample preparation for NGS and single-molecule sequencing including DNA quantification and equimolar amplicon pooling.
- $1D^2$ multiplexing protocol for ONT sequencing (Andrea Thürmer had an equal contribution).
- Processing of the raw NGS and ONT datasets and data analysis using Genious 11.1.5 (Biomatters Ltd, Auckland, NZ), including adapter and quality trimming, mapping to the reference, de novo assembly, consensus sequence generation,

determining parameters for variant calling, plausibility assessment of minority mutations, analysis of insertions, calculation of polymorphisms and type of selection, performing BLASTn searches and determining insertion characteristics, and determining the correlation between insertions and mutations.

- Drafting and final editing of the manuscript including the design of all tables and figures.

_____

Signature of doctoral candidate

Printing copy of the publication

scientific reports

Check for updates

OPEN

Advanced sequencing approaches detected insertions of viral and human origin in the viral genome of chronic hepatitis E virus patients

C.-Patrick Papp[1,2], Paula Biedermann[1,2], Dominik Harms[1], Bo Wang[1,3], Marianne Kebelmann[1,2], Mira Choi[4], Johannes Helmuth[5], Victor M. Corman[2], Andrea Thürmer[6], Britta Altmann[1], Patrycja Klink[1], Jörg Hofmann[2,5,8] & C.-Thomas Bock[1,7,8]✉

The awareness of hepatitis E virus (HEV) increased significantly in the last decade due to its unexpectedly high prevalence in high-income countries. There, infections with HEV-genotype 3 (HEV3) are predominant which can progress to chronicity in immunocompromised individuals. Persistent infection and antiviral therapy can select HEV-3 variants; however, the spectrum and occurrence of HEV-3 variants is underreported. To gain in-depth insights into the viral population and to perform detailed characterization of viral genomes, we used a new approach combining long-range PCR with next-generation and third-generation sequencing which allowed near full-length sequencing of HEV-3 genomes. Furthermore, we developed a targeted ultra-deep sequencing approach to assess the dynamics of clinically relevant mutations in the RdRp-region and to detect insertions in the HVRdomain in the HEV genomes. Using this new approach, we not only identified several insertions of human (AHNAK, RPL18) and viral origin (RdRp-derived) in the HVR-region isolated from an exemplary sample but detected a variant containing two different insertions simultaneously (AHNAK- and RdRpderived). This finding is the first HEV-variant recognized as such showing various insertions in the HVR-domain. Thus, this molecular approach will add incrementally to our current knowledge of the HEV-genome organization and pathogenesis in chronic hepatitis E.

Hepatitis E virus (HEV) is one of the main causes of acute viral hepatitis worldwide. It is a hepatotropic, approximately 7.2 kb single-stranded positive-sense RNA virus whose genome contains three open reading frames (ORF1, ORF2, and ORF3), 5′ and 3′ untranslated regions (UTRs), and a poly(A) tract at the 3′ end. The genotypes HEV-1 to HEV-8 belong to the genus *Orthohepevirus A* of which at least HEV-1 to HEV-4 can infect

[1] Division of Viral Gastroenteritis and Hepatitis Pathogens and Enteroviruses, Department of Infectious Diseases, Robert Koch Institute, Berlin, Germany. [2]Institute of Virology, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität Zu Berlin, Berlin Institute of Health, German Centre for Infection Research, Berlin, Germany. [3]College of Veterinary Medicine, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. [4]Department of Nephrology and Intensive Medical Care, Charité Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität Zu Berlin, Berlin, Germany. [5]Charité-Vivantes GmbH, Labor Berlin, Berlin, Germany. [6]Genome Sequencing, Methodology and Research Infrastructure, Robert Koch Institute, Berlin, Germany. [7]Institute of Tropical Medicine, University of Tübingen, Tubingen, Germany. [8]These authors jointly supervised this work: Jörg Hofmann and C.-Thomas Bock. ✉email: BockC@rki.de

humans. HEV-1 and HEV-2 are responsible for large waterborne outbreaks in low-income countries whereas HEV-3 and HEV-4 cause foodborne infections and are autochthonous in high-income countries[1,2]. The latter two are zoonotic strains with mainly pigs and wild boars as putative main hosts and are transmitted to humans by the ingestion of raw or undercooked meat [3–5]. It has been shown that HEV-3 has a higher variability compared to other HEV genotypes and has been correlated with a higher host range[6]. Recombination events in this genotype have been repeatedly described, most frequently as insertions in the hypervariable region (HVR, also denoted as

Patient	Sample no	Sample name	Sample type	Cause of immunodeficiency	Collection date	Virus Concentration (copies/ml)
1	1a	17-0421–*sample 1*	Plasma	Atypical hemolytic-uremic syndrome, severe immunoglobulinaemia	10/2017	4.1×10^8
1	1b	17–0420	Stool		10/2017	1.17×10^9
1	2	17–0534	Plasma		11/2017	2×10^5
1	3	18–0002	Plasma		01/2018	6.5×10^7
1	4	18-0056–*sample 2*	Plasma		04/2018	8.5×10^7
2	5	17–0371	Plasma	Kidney and pancreas transplantation	07/2017	4.4×10^6
3	6	17–0535	Plasma	Kidney transplantation	12/2017	1.6×10^5
4	7	18–0058	Plasma	Selective IgG deficiency, T-lymphopenia with helper cells < 200	04/2018	1.3×10^6
5	8	18–0066	Plasma	Peripheric T-cell lymphoma	01/2018	1.2×10^6
6	9	18–0068	Plasma	Kidney transplantation	04/2018	8.1×10^6

Table 1. Samples from patients with chronic HEV infection.

polyproline region, PPR)[7–11]. The inserted sequences mostly originated in HEV, however, sequences originating in the human genome have also been identified[7,12–14]. Recombinants and other HEV variants were associated with ribavirin treatment failure, adaptability in cell culture, and increased viral replication potential [10,12,13,15,16]. Besides recombination events, some single-nucleotide polymorphisms (SNP) such as G1634R were linked with RBV treatment failure[17]. Further SNPs identified in the RNA-dependent-RNA-polymerase (RdRp) coding sequence in patients treated with RBV were Y1320H, K1383N, K1398R, V1479I, Y1587F, G1634K[18]. Similar to other RNA viruses, HEV builds a so-called mutant cloud, which represents an intra-host heterogeneous population, with the advantage of rapid adaptation to environmental conditions[19]. Population-based Sanger sequencing has been shown to miss minor variants with a frequency below 20%[20]. Therefore, to capture the heterogeneity of the HEV quasispecies, including recombination events and SNP that occur with very low frequencies, we developed new sequencing approaches for HEV genotype 3 based on the amplification of the near full-length genome of HEV by long-range PCR (lrPCR) followed by subsequent next-generation sequencing (NGS) and third-generation sequencing. These methods allow the identification of insertions and SNPs in the HEV genome. Furthermore, single-molecule sequencing using Oxford Nanopore Technologies (ONT) enables the analysis of potential correlation in the occurrence of these events, as mutations and insertions can be detected on either the same or different DNA strands. A multiplex third-generation sequencing protocol (Oxford Nanopore Technologies, 1 D^2 flow cell) was also established as an additional optimization step to perform parallel sequencing of multiple samples. Amplicon-based NGS was performed to detect multiple insertions that coexist in the viral population and to determine the dynamic of SNPs in the RdRp region of the HEV genome.

Material and methods

Chronic hepatitis E patients and sample collection. To evaluate the newly established molecular methods, a total of nine plasma samples and one faecal sample from patients with chronic hepatitis E infection (CHE), defined as persistence of HEV RNA for longer than three months [21], were used in this study. All patients were immunosuppressed (Table 1). Five samples were follow-up samples from one patient (Table 1). The samples 17-0421 and 18-0056 were used to compare the sequencing methods and for the sake of clarity will be referred to as *sample 1* and *sample 2,* respectively. All samples were obtained from routine diagnostics and were pseudonymised before usage. The HEV genotype for all samples was determined as described previously [22].

RNA extraction, cDNA synthesis, and long-range PCR (lrPCR). RNA extraction was performed using QIAcube (QIAGEN, Hilden Germany) with the QIAamp Viral RNA Mini kit (QIAGEN, Hilden, Germany) according to the manufacturer's instruction. The RNA obtained from each sample was stored at − 80 °C until use. Quantitative PCRs for HEV were performed as previously described[22,23]. Isolated RNA was used for cDNA synthesis with the SuperScript IV First-Strand Synthesis System (Invitrogen, Thermo Fisher Scientific, Carlsbad, CA, USA). The synthesis was performed with Oligo d(T) primers and 11 µL template RNA as described in the manufacturer's protocol with minor modification of the cDNA synthesis step at 60 °C for 20 min. To verify the successful synthesis of the whole HEV genome, four genome-wide PCRs were performed.

Subsequently, the cDNA was used for the near full-length lrPCR. The lrPCR was performed using the Kapa HiFi Readymix (Sigma-Aldrich, Merck KGaA, Darmstadt, Germany) with an optimised program for GC-rich templates for both, first and heminested PCR. Primers and PCR protocols are listed in the Supplementary Tables S1 and S2. The PCR products were visualized on a 1% agarose gel using the BioDocAnalyze software (Biometra GmbH, Göttingen, Germany).

Illumina sequencing of the lrPCR products and reads processing. Magnetic beads purification of the lrPCR products was performed with the MagSi-NGS Prep Plus kit (magtivio B.V., The Netherlands) according to the manufacturer's instructions. The purified PCR products were measured by Qubit fluorometer using the double-stranded DNA High Sensitivity Assay Kit (Thermo Fisher Scientific, Carlsbad, CA, USA). A wholegenome (WG) library was prepared for all PCR products using the Nextera XT library kit (Illumina, San Diego), following the manufacturer`s instructions. WG libraries were sequenced in a 2 × 250 bp sequencing run on an Illumina HiSeq 2500 platform (Illumina Inc., San Diego, USA).

Illumina raw sequencing data were converted using bcl2fastq v 1.8.4 conversion software (Illumina Inc., San Diego, USA) and demultiplexed according to their multiplex identifier. For each sample, two fastq files were generated representing the paired-end reads. These two files were imported into Geneious version 11.1.5 (Biomatters Ltd, Auckland, NZ) and automatically paired. Adapter and quality trimming was performed using BBDuk Adapter/Quality Trimming Version 37.64 by Brian Bushnell implemented in Geneious. Minimum quality was set to 30. Trimmed reads were mapped to the reference sequence wbGER27_RAS (FJ705359.1) using Geneious mapper with medium sensitivity and performing five iterations. The consensus sequence containing the most common bases was extracted.

Oxford Nanopore sequencing of the lrPCR products and reads processing. Third-generation sequencing for long-read sequencing was performed using Oxford Nanopore Technologies (ONT) MinION. To multiplex samples on the high accuracy $1D^2$ flow cells (Oxford Nanopore Technologies, Oxford, UK), the manufacturer's native barcoding protocol (XP-NBD104) recommended for Ligation-Libraries (SQK-LSK109) was combined with the Ligation 1 D^2 (SQK-LSK308) manufacturer's protocol. Barcoding of all samples was performed according to the ONT 1D Native barcoding genomic DNA protocol. After barcoding, dA-tailing of the barcoded DNA was performed followed by magnetic beads clean-up and the ONT 1 D^2 adapter ligation step from the $1D^2$ sequencing of the genomic DNA protocol. After 1 D^2 adapter ligation, elution and DNA concentration measurements using a Qubit fluorometer (Thermo Fisher Scientific) were performed for each sample. The prepared samples were pooled to a final mix with a volume of 50 µL and a concentration of approximately 10 ng/ µL. The prepared sample pool was used for sequencing according to the 1 D^2 sequencing protocol.

The raw reads were processed and demultiplexed using ONT Guppy basecaller (Oxford Nanopore Technologies, Oxford, UK). Post-analysis of the resulting fastq files was done with Geneious version 11.1.5 (Biomatters Ltd, Auckland, NZ). Due to the high amount of data generated by this sequencing approach, the lack of NGS data for ONT read correction for samples 6–10, and the limited computational power available, only the reads for *sample 1* and *2* were analysed. Therefore, sequences with lengths between 6.5 and 8 kb were extracted. In the size selected sequence list, each insertion found was annotated. The function "extract annotation" was used to extract the entire sequences based on their annotation. To validate the ONT reads that contain two insertions simultaneously, the trimmed and error-corrected Illumina reads were mapped against the MinION sequence from *sample 2* which contained two insertions simultaneously. Mapping was performed using Geneious mapper at medium sensitivity, executing five iterations. The consensus sequence containing the most common bases was extracted. In addition, the size selected long-reads from *sample 2* were mapped against a reference (GenBank ID: FJ705359.1) using Geneious mapper set at medium sensitivity. The consensus sequence containing the most common bases was extracted.

Amplicon-based NGS. Target specific primers containing flow cell adapters at the 5' end were designed (Supplementary Table S2). These primers were used to generate amplicons containing the region of interest which were subsequently sequenced on Illumina MiSeq using the MiSeq Reagent Kit v3 (Illumina, San Diego, USA) generating 2 × 300 bp reads. The amplicons are between 350 and 500 bp in size allowing two paired reads to overlap. Merging the two reads of a pair results in a single end-to-end-sequence covering potential insertions with high accuracy.

RNA was isolated as described above and used for cDNA synthesis with the SuperScript IV First-Strand Synthesis System (Invitrogen, Thermo Fisher Scientific, Carlsbad, CA, USA) using random hexamers and 11 µL template RNA as described in the manufacturer's protocol. In contrast to the cDNA synthesis performed for lrPCR, for the amplicon NGS, we used random hexamers increasing the cDNA yield compared to Oligo d(T) primers [24].

Hypervariable region amplicons. For the HVR amplicons, a fragment of approximately 800 bp in size was amplified. The resulting PCR products were used for nested PCR with primers containing overhang adapters for the Illumina flow cell (Supplementary Table S2). Depending on the length of insertion, the size of the generated amplicons varies. Therefore, two sets of primers aiming to generate the optimal amplicon size for

NGS were designed. For amplification in the first and nested PCRs the Kapa HiFi Readymix (Sigma-Aldrich, Merck KGaA, Darmstadt, Germany) was used. The success of amplification and the amplicon size was verified by 1.5% agarose gel electrophoresis using the BioDocAnalyze software (Biometra GmbH, Göttingen, Germany).

RNA-dependent RNA-polymerase amplicons. A 2 kb target sequence of the complete HEV polymerase region was amplified by TaKaRa Ex Taq DNA Polymerase (Takara Bio Inc., Japan) according to the manufacturer's instructions using the primers listed in the Supplementary Table S2 with the PCR conditions described in the Supplementary Table S1. The PCR product was used as template for three approximately 550 bp overlapping PCR amplicons that cover the complete coding RdRp domain. Target specific primers with Illumina MiSeq overhang adapters were designed. Amplification was performed using the Takara Ex Taq DNA polymerase (Takara Bio Inc., Japan) with the PCR conditions described in the Supplementary Table S1. The success of amplification was verified by 1.5% agarose gel electrophoresis using the BioDocAnalyze software (Biometra GmbH, Göttingen, Germany).

Amplicon preparation, sequencing, and reads processing. Magnetic beads purification of the PCR products was performed with the MagSi-NGS Prep Plus kit (magtivio B.V., The Netherlands) according to the manufacturer's instructions. Purified PCR products were measured by Qubit fluorometer using the doublestranded DNA High Sensitivity Assay kit (Thermo Fisher Scientific). To compare the quality between a pooled and a separate sequencing approach, for sample 1b two different pools were prepared: one containing only the three RdRp amplicons, the second containing the RdRp and the HVR amplicons. These two pools were sequenced simultaneously using the Illumina MiSeq technology.

Amplicon libraries were generated using the Nextera XT library kit (Illumina, San Diego). The index-PCR was performed according to the manufacturer's protocol. Indexed libraries were pooled and sequenced on the Illumina MiSeq using 2 × 300 bp reads. The read files containing paired-end reads provided for each sample were paired, trimmed, and mapped as described above.

Single-nucleotide polymorphisms detection. Single-nucleotide polymorphisms (SNP) were detected using the Geneious "Find Variations/SNPs" tool. The minimum variant frequency was set to 0.005, maximum variant *P*-value to 10^{-6} and minimum strand-bias *P*-value at 10^{-5} when exceeding 65% bias. Regions below a coverage of 500 were excluded from variant calling.

Cut-off for biological variants. cDNA synthesis, PCR, and the sequencing process are sources of error that need to be considered when sequencing RNA viruses. Therefore, we used the SuperScript IV high-fidelity RT which has a misincorporation frequency of 1.8×10^{-4}[25]. *Sample 1* and *2* had high viral loads of 10^8 and 10^6 copies/ml, respectively (Table 1), which translates into a higher yield for the cDNA synthesis. Furthermore, to maximise the amount of cDNA synthesised and consequently used for lrPCR, a maximum of 11 µL template RNA was used for cDNA synthesis. The cut-off for biological variants was set at 0.5%.

Regarding sequencing errors, by trimming all bases with a Phred quality score lower than 30, the expected error rate is five times lower than our cut-off. The plausibility of the minority mutations detected was assessed by comparing their values with the values detected when other sequencing approaches were used or with the values detected in other follow-up samples.

Calculation of polymorphisms and type of selection. In order to calculate the proportions of the polymorphism, the number of SNPs over 0.5% detected in a specific region was divided by the number of nucleotides in that region. To determine the type of selection in a specific region of the HEV genome, we used the ratio between nonsynonymous (dN) and synonymous (dS) substitutions per site (ratio dN/dS). Insertions in the HVR were excluded from the calculation.

Detection of insertions. Due to the high variability of the HVR and to the targeted manner of the sequencing approach, the UNOISE3-Suite was used to get zero-radius operational taxonomic units (zOTUs) from the HVR-amplicon reads[26]. Thus, clustering of highly similar reads and generation of representative consensus sequences as zOTUs was performed. zOTU sequences were imported into Geneious and mapped to the reference sequence wbGER27_RAS (FJ705359.1). The origin of the HVR insertions was determined using the BLASTn search engine (https:// blast. ncbi. nlm. nih.go v).

Correlation between insertions and mutations. The long-read ONT sequences from *sample 1* and *2* were grouped based on their insertions. Accordingly, there were the *AHNAK insertion*, the *HEV-derived insertion,* and the *no insertion* sequence groups. The frequencies of the RdRp mutations G1634R, Y1587F, V1479I, and K1383N that were detected in the analysed samples were determined for each group and compared. A correlation between a certain mutation and an insertion would result in a higher frequency of this mutation in one of the groups. However, given a Phred score of 10 for our ONT reads, the frequencies of the mutations were interpreted as estimates.

Accession numbers. HEV sequences described in this article have been submitted to NCBI GenBank under the accession numbers: MW837243 to MW837255 (Supplementary Table S3).

Ethical approval. The ethics committee of the Charité Universitätsmedizin Berlin approved the study (approval number No. EA1/367/16) and written informed consent was obtained from all participating individuals. Patient samples were de-identified for this study. All experiments were performed in accordance with relevant guidelines and regulations.

Results

HEV samples. To evaluate the newly established molecular methods, a total of nine plasma samples and one faecal sample from patients with clinically confirmed chronic hepatitis E (CHE) were used in this study. All these CHE infections were observed in immunosuppressed patients. (Table 1). The samples were collected between October 2017 and April 2018 with viral loads ranging between 10^5 and 10^9 copies/ml (Table 1). Seven samples from CHE patients, including *sample 1* and *2* (Table 1), were used to test the lrPCR method and subsequently the multiplex sequencing method using $1D^2$ flow cell from ONT. Four samples (1a, 1b, 2, and 3, Table 1) collected from one CHE patient representing three time-points of the CHE infection (follow-up samples) were used to assess the amplicon-based NGS methods. In addition, the lrPCR products from *sample 1* and *2* were further sequenced with the Illumina WG method and additionally by Sanger sequencing (data not shown), that confirmed the mutations with high frequency and the most frequent and prevalent insertions.

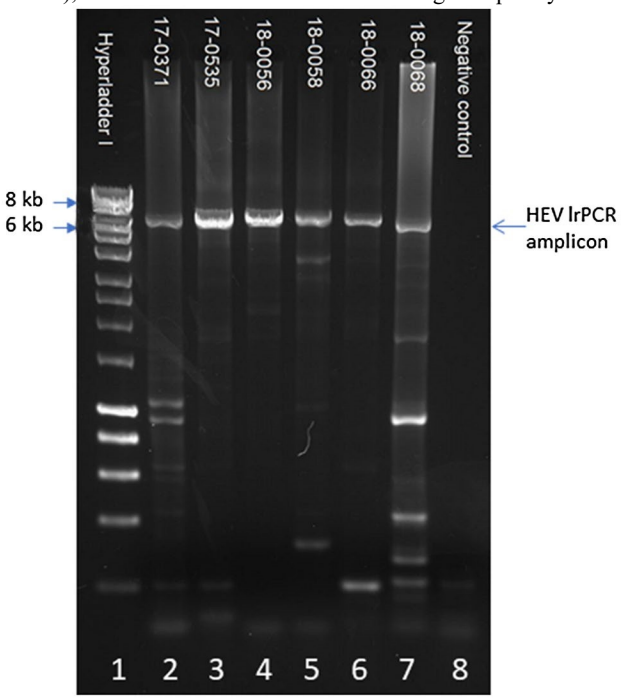

Figure 1. Agarose gel electrophoresis of HEV lrPCRs. Lanes 2–7 show lrPCR results of patient samples 17-0371 to 18-0068 with signals of correct size between 6 and 8 kb. Lane 1: size marker (HyperLadder™ 1 kb, Bioline)(for the creation of the gel image the software BioDocAnalyze, Version 2.67.5.0, www. biome tra. com was used).

Insertion	Accession No	Sequence (5'-3')
AHNAK[b]–WG-NGS	MW837253	TGA CAT AAC AGG TCC AAA AGT TGA TAT TAA TAT CGA AGG CAA GTC AAA GAA ATC TCG TTT TAA GCT TCC CAA ATT TAA TTT TTC GGG CTC TAA AGT TCA GAC ACT TGA AGT GGA TGT CAA AGG TAA AAA ACC AGA AAT
HEV–RdRp[a]—WG-NGS	MW837246	CGG TCG GCC TGG ATC TTA CAG GCG CCG AAG GTG TCT CTT AAG GGT TTT TGG AAG AAG CAT TCT GGT GAG CCT GGT ACC CTC CTT TCG ACA GCA TCC GCC TCC CCT GCC CCT GAG CCC GCT CAA CCA CCT GGC TCC GCT GGG CCA AAG ACT CCT GTG CGT AAG
AHNAK[b]–Amplicon-NGS	MW837244	TGA CAT AAC AGG TCC AAA AGT TGA TAT TAA CAT CGA AGG CAA GTC AAA GAA ATT TCG TTT TAA GCT TCC CAA ATT TAA TTT TTC GGG CTC TAA AGT TCA GAC ACC TGA AGT GGA TGT CAA AGG TAA AAA GCC AGA TAT

RPL18[c] (1)–Amplicon-NGS	MW837245	TCT AAG AGG TTG TTT ATG AGT CGC ACC AAC CGG CCG CCT CTG TCC CTT TCC CGG ATG ATC CGG AAG ATG AAG CTC CCT GGC CGG GGA AAC AAG ACG GCC GTG GCT GTG GGG ACC ATA ACT GAT GAT GTG CGG GTT CAG GAG GTA CCC AAA
RPL18[c] (2)–Amplicon-NGS	MW837248	CTT TCC CGG ATG ATC CGG AAG ATG AGG CTT CCT GGC CGG GAA AAC AAG ACG GCC GTG GCT GTG GGG ACC ATA ACT GAT GAT GTG CGG GTT CAG GAG GTA CCC AAA
HEV–RdRp[a]–Amplicon-NGS	MW837246	CGG TCG GCC TGG ATC TTA CAG GCG CCG AAG GTG TCT CTT AAG GGT TTT TGG AAG AAG CAT TCT GGT GAG CCT GGT ACC CTC CTT TCG ACA GCA TCC GCC TCC CCT GCC CCT GAG CCC GCT CAA CCA CCT GGC TCC GCT GGG CCA AAG ACT CCT GTG CGT AAG
HEV—HVR duplication[a]– Amplicon-NGS	MW837247	GCT GGG CCA AAG ACT CCC GTG CGT AAG CCG CCA ACG CCA CCA CCC CCG CGC ACC CGC CGC

Table 2. HVR insertions. [a] Using wbGER27 as reference (accession No: FJ705359.1); [b]homo sapiens AHNAK nucleoprotein (human neuroblast differentiation-associated protein (desmoyokin), accession No.: NG_051822.1); [c]homo sapiens ribosomal protein L18 (RPL18, accession No.: L11566.1); HVR = hypervariable region; NGS = next generation sequencing; WG-NGS = whole genome next generation sequencing.

Long-range PCR and Illumina sequencing. The near full-length HEV genome amplification was successful for all samples (Fig. 1; an unprocessed image of Fig. 1 is presented in the Supplementary Information). To gain a deeper insight into the diversity of viral variants, the lrPCR products from two samples (*sample 1* and *sample 2*) collected from the same patient at a six-month interval were sequenced with Illumina technology. More than two million reads were generated for each sample. After processing the raw data, 99.9% of the reads successfully mapped onto the reference with a mean coverage of approximately 60 thousand and minimum coverage of approximately 12 thousand (Supplementary Table S4). Two insertions in the HVR were detected using Illumina NGS in both samples analysed. One insertion was identified as a fragment of the human AHNAK gene, and one as a fragment from the RdRp region of the HEV genome. The exact sequences and their characteristics are presented in Tables 2 and 3.

Source of Insertion	Position in HEV[a] (bp)	Length (bp)	BLASTn Results Position on Insertion: Position in Source (bp)	
Whole-genome NGS				
AHNAK[b]	2,227	138	Segment 1: 1–30:27,578–27,607 Identity with NG_051822.1: 100% Mismatches: 0	Segment 2: 27–138:27,466–27,577 Identity with NG_051822.1: 97.3% Mismatches: 3
HEV–RdRp[a]	2,360	162	Segment 1: 1–88:4,544–4,631 Identity with FJ705359.1: 95.5% Mismatches: 4	Segment 2: 88–162:2,285–2,359 Identity with FJ705359.1: 88% Mismatches: 9
Amplicon-based NGS				
AHNAK[b]	2,227	138	Segment 1: 1–30:27,578–27,607 Identity with NG_051822.1: 100% Mismatches: 0	Segment 2: 27–138:27,466–27,577 Identity with NG_051822.1: 98.2% Mismatches: 2
RPL18[c], variant 1	2,273	150	4–150:160–306 Identity with L11566.1: 98% Mismatches: 3	
RPL18-2[c], variant 2	2,246	105	1–105:202–306 Identity with L11566.1: 98.1% Mismatches: 2 deletion of HEV sequence from 2,246 to 2,275	
HEV–RdRp[a]	2,360	162	Segment 1: 1–88:4,544–4,631 Identity with FJ705359.1: 95.5% Mismatches: 4	Segment 2: 88–162:2,285–2,359 Identity with FJ705359.1: 88% Mismatches: 9
HEV–HVR[a]	2,393	60	1–60:2,333–2,392 Identity with FJ705359.1: 96.7% Mismatches: 2	

Table 3. BLASTn results of the insertions. [a] Using wbGER27 as reference (accession No: FJ705359.1); [b]homo sapiens AHNAK nucleoprotein (accession No.: NG_051822.1); [c]homo sapiens ribosomal protein L18 (accession No.: L11566.1).

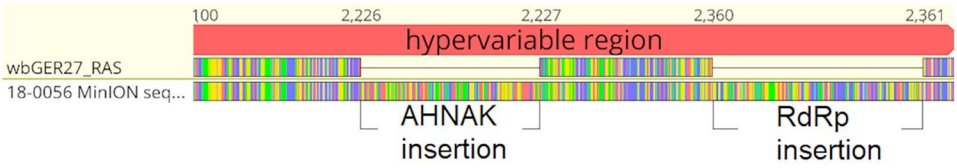

Figure 2. Cumulated AHNAK- and RdRp-derived Insertions in the HVR. The red bar represents the sequence annotation. Reference sequence is wbGER27_RAS (FJ705359.1) and below the sequence containing two insertions (AHNAK and RdRp) detected in sample 18-0056 (*sample 2*) using long-read sequencing. Visualization of data by Geneious version 11.1.5 (Biomatters Ltd, Auckland, NZ).

Long-read sequencing with Oxford Nanopore technology. The $1D^2$ flow cell generated a mean yield of approximately 285,000 reads per sample. The number of reads per sample after size selection ranged from 11,000 to more than 68,000 reads (Supplementary Table S5). When mapped to the HEV reference the reads compared well to the length of the PCR products covering 95% of the HEV genome. Thus, near fulllength sequences of six samples were generated using long-read sequencing with the MinION device. The two insertions identified by Illumina NGS were also detected when analysing the MinION reads of *sample 1* and *2*. This confirmed the coexistence of both variants at an early time-point of HEV infection. Interestingly, we were able to detect in *sample 2* five ONT reads showing both the AHNAK-derived and the RdRp-derived insertions, separated by a 132 bp HEV specific connecting sequence (Fig. 2). To validate the detection of the HVR variant with simultaneous AHNAK- and RdRp-derived insertions also in the Illumina reads, the Illumina reads from *sample* 2 were mapped against ONT reads. Thus, Illumina reads were detected that contained the 3′-end of the AHNAK insertion, the whole HEV specific connecting sequence, and the 5′end of the RdRp insertion. Moreover, approximately 0.1% of the Illumina reads covering this specific region of the HVR contained parts of both AHNAK- and RdRp-derived sequences simultaneously. This confirms the double-insertion variant showing that HEV can cumulate insertions.

Targeted NGS. For the targeted NSG approach we used *sample 1* which was a plasma (sample 1a) and a stool sample (sample 1b) from the same patient and same time-point (Table 1). The read qualities and coverages of the HVR amplicons after mapping to the reference wbGER27_RAS (FJ705359.1) are shown in Table S6. Here,

Sequencing method	Insertion	Sample	Reads containing insertion/total reads	Percentage of reads containing insertion
Whole-genome NGS	AHNAK[b]	*Sample 1*	36,170/42,760	~ 85%
		Sample 2	4,770/23,410	~ 20%
	HEV–RdRp[a]	*Sample 1*	1,360/42,760	~ 3%
		Sample 2	5,590 /23,410	~ 24%
Amplicon-based NGS	AHNAK[b]	*Sample 1*	620/15,980	~ 4%
		Sample 1b	31,840/135,350	~ 23%
	RPL18[c], variant 1	*Sample 1*	5,210/15,980	~ 33%
		Sample 1b	47,220/135,350	~ 35%
	RPL18-2[c], variant 2	*Sample 1*	0	0%
		Sample 1b	1,580/135,350	~ 1%
	HEV–RdRp[a]	*Sample 1*	20/15,980	< 1%
		Sample 1b	1,280/135,350	~ 1%
	HEV–HVR[a]	*Sample 1*	0	0%
		Sample 1b	814/135,350	< 1%

Table 4. Proportions of reads supporting insertions. [a] Using wbGER27 as reference (accession No: FJ705359.1); [b]homo sapiens AHNAK nucleoprotein (accession No.: NG_051822.1); [c]homo sapiens ribosomal protein L18 (accession No.: L11566.1).

Sample no	Time-point	Weeks after 1st Time-point	G1634R	Y1587F	V1479I	K1383N
1a (*sample 1*)	1—blood	0	1.1	< cut-off	< cut-off	< cut-off
1b	1—stool	0	1.1	< cut-off	< cut-off	< cut-off
2	2	6	< cut-off	< cut-off	< cut-off	< cut-off
3	3	12	6.5	90	2	98
4 (*sample 2*)	4	25	50	85	8.5	100

Table 5. RdRp mutations in percent.

we detected further insertions in addition to the ones found by whole-genome sequencing (Tables 2, 3, and 4). The insertions are all in-frame and range between 21 and 54 amino acids in length. A more detailed analysis showed that one insertion derives from the AHNAK nucleoprotein sequence from the human genome, two were from the human ribosomal protein L18 sequences, one has its origin in the RdRp of the HEV genome, and one is a duplication of 21 amino acids of the HVR. The sequences of the two human RPL18 insertions are identical, however, one sequence lacks the first 15 amino acids and is preceded by a deletion of the HEV sequence. Notably, the truncated RPL18-derived insertion was detected only in the stool sample (Table 4). Using amplicon-based NGS sequencing a total of five insertions in the stool and three insertions in the plasma sample could be detected compared to Illumina WG or MinION sequencing where only the AHNAK and RdRp derived insertions could be identified. To confirm the high extent of variability detected in the HVR we additionally sequenced the more conserved RdRp region from the same samples using the amplicon-based NGS method. Since mutations in the RdRp region could be associated with therapy failure, a targeted deep-sequencing approach for this region is useful for mutational analysis. Read quality and coverage of the RdRp amplicons after mapping against wbGER27_RAS (FJ705359.1) are shown in Table S7. The RdRp reads showed no structural changes besides polymorphisms. The variability of sample 1a and 1b of both HVR and RdRp amplicons was compared. The number of polymorphic loci for both HVR and RdRp showed a ratio approximately twice as high for the HVR (1a: 0.37 and 1b: 0.41) compared to the RdRp (1a: 0.22 and 1b: 0.17). Furthermore, the ratio between non-synonymous SNPs and synonymous SNPs was 1.5 and 1.2 for HVR compared to 0.5 and 0.23 for the RdRp, respectively. This finding indicated that the HVR is under positive selection whereas the RdRp is under negative selection. Notably, sample 1b had a nine times higher mean coverage than 1a for the RdRp amplicons and a seven times higher mean coverage for the HVR amplicon (Table S7). In addition, all insertions described above could be detected also in the pooled HVR and RdRp amplicons. The read quality and coverage of the pooled amplicons after mapping against wbGER27_RAS (FJ705359.1) are shown in Table S7.

Dynamics of SNPs. The RdRp region of five samples (1a, 1b, 2, 3, and 4, Table 1) from one patient under ribavirin therapy collected within six months was analysed. The frequency of mutations in the RdRp region found in these samples that are associated with therapy failure or higher replication, namely K1383N, V1479I, Y1587F, and G1634R are shown in Table 5. Notably, between the second and third time-points all specified mutations increased, especially Y1587F and K1383N. The mutations Y1587F and K1383N increased from values lower than the cut-off to approximately 90% and 98% of reads at time-point three. At time-point four, the percentages of G1634R and V1479I increased further, whereas Y1587F and K1383N maintained consistent. When determining the frequencies of the RdRp mutations for each insertion group using the ONT reads there was no difference in the distribution of these mutations compared to the initial analysis where all sequences were assessed together. Therefore, there is most likely no correlation between insertion in the HVR and mutations in the RdRp.

Discussion

To gain insights into the mechanisms leading to pathophysiology, persistence, and adaptation of HEV in their hosts and to improve clinical outcome as well as management of HEV infections, the need for robust methods that facilitate the use of cutting-edge technologies is obvious. Amplifying the whole-genome by lrPCR allows sequencing of viral strains as single molecules using third-generation sequencing. Whole-genome sequencing is closer to physiology compared to smaller amplicons and fragments while it reduces the number of PCRs needed, the preparation costs, and saves time. Using lrPCR products for NGS we obtained a high coverage throughout the whole HEV genome which makes this method suitable for intra-populational analysis. Increasing the amount of target cDNA and validating mutations using data from follow-up samples a cut-off value for variant detection as low as 0.5% can be applied. Exemplarily for this study, proportions of known minority variants, such as G1634R, were compared in *sample 1* (plasma sample) and sample 1b (stool sample) which were collected at the same time-point and sequenced using different approaches. Consequently, we were able to validate and compare the number of mutations in the HEV genome by different methods despite the low cut-

off value of 0.5%. However, regarding the detection of long insertions, single-molecule sequencing by ONT was superior to whole genome NGS with the Illumina technology as the long-reads of the ONT captured a viral variant containing AHNAK- and RdRp-derived insertions which cumulated on the same strand. This is the first report of such a mixed HVR viral variant. Furthermore, a correlation between mutations in the RdRp and recombination events in the HVR could be excluded. This indicated that mutation selection and recombination events are probably unlinked mechanisms that drive viral diversity and adaptation through different molecular mechanisms highlighting the advantage of long-read sequencing technologies. Nevertheless, it is worth mentioning that third-generation sequencing has a higher error rate than NGS, which makes this method less suitable for the detection of polymorphisms[27].

With regard to the detection of insertions, the amplicon-based sequencing proved to be the method of choice. The amplicon-based sequencing showed the best performance by detecting the most insertions if compared to whole-genome sequencing; however, due to the 2 × 300 bp reads and thus smaller amplicon size, this method failed to detect the HEV variant with two insertions that occur on the same strand with a total length of 432 bp which could be detected only by third-generation sequencing. Both NGS methods presented allowed for a very detailed analysis; however, only the AHNAK and RdRp derived insertions were detected when sequencing was carried out using lrPCR. This may be due to the cDNA synthesis which was performed using Oligo d(T)s for the lrPCR whereas for the HVR amplicon random hexamers were used which tend to achieve higher cDNA yield[24]. Sampling bias, different cDNA synthesis methods, and a potential strand selection of the PCR primers could explain the differences in the proportion of the insertions when comparing the whole-genome NGS and the amplicon-based NGS datasets. Therefore, the frequencies given for each insertion were listed for completeness and should be interpreted accordingly. While standard sample preparation for NGS includes tagmentation which can lead to sequencing of host DNA and thus artifacts that could be misinterpreted as insertions, the ampliconbased sequencing method was performed without tagmentation. This approach aimed to generate PCR products and subsequently reads that detected both HEV specific sequence of the HVR region with possible insertion. The PCR products were generated with target-specific primers designed to contain the flow cell adapters. In addition, the generated reads were clustered and consensus sequences of highly similar reads were generated. Therefore, to exclude contamination by host DNA which may also have been sequenced, only insertions were validated where the reads simultaneously contained the insertion and HEV-specific sequences. Moreover, the AHNAK insertion was confirmed by long-read sequencing, indicating that human sequences were integrated into the viral genome. Insertions originating in the human genome have previously been described in HEV and some of these HEV variants have been shown to efficiently replicate in the human hepatocyte cell line HepG2/C3A[7,28,29]. Insertions in the HEV genome and particularly in the HVR region have already been detected in samples of CHE individuals and additionally, although very rarely, in samples from individuals with acute HEV infection, suggesting a potential role of HVR diversity in chronification [11,12,30]. However, the impact of these recombination events in the HVR region on the chronic course of infection, the stability of viral genome and viral replication remains unclear. Using the amplicon-based method five different insertions in the HVR could be detected in one sample. Two insertions derived from the human RPL18 gene showing identical sequences; however, one of these lacked 15 amino acids from the RPL18 insertion and nine amino acids from the HEV sequences that precedes the insertion. A possible explanation for this truncated RPL18 insertion is that HEV highjacked the RPL18 gene after which a deletion took place. This finding is in accordance with a recent report where 114 HEV strains have been sequenced using PacBio platforms and insertions from human genes or duplications of the HEV genome were detected in seven patients. However, in one sample six nucleotides truncated human gene fragment RNA18SP5 was detected using single-molecule sequencing[11]. Besides the possibility of sequencing error, Lhomme et al. also discussed the possibility of different biological variants that have been detected separately by different sequencing methods [11]. As noted above, some recombination variants are likely missed using different sequencing approaches due to their low quantity in the viral population. However, using the amplicon-based ultra-deep sequencing method of the HVR, we were able to detect not just one but five different insertions in the same sample. It is therefore evident that amplicon-based ultra-deep sequencing is best suited for the screening of the HVR of many samples allowing multiplexing, high accuracy, and ultra-deep sequencing.

In contrast to the HVR, the RdRp region is highly conserved and could be used to validate HVR sequencing. The RdRp region is clinically relevant due to the variants that can occur by the selection of mutations associated with higher replication competence and potential therapy failure[10,17,18]. The HVRs in samples 1a (blood) and 1b (stool) were under positive selection which suggests that amino acid changes in this region may play an important role for the virus replication and fitness. Positive selection in the HVR has been described previously and it seems to be a characteristic of the zoonotic HEV genotypes HEV-3 and HEV-4[6]. Considering the quality and validity of our data, a cut-off of 0.5% for variant detection was implemented focusing on the dynamics of the mutations. In both, sample 1a (plasma) and 1b (stool), the mutation G1634R was detected at 1.1% confirming the robustness of the method. Notably, the G1634R mutation was a minority mutation that was detectable already at the first time-point at a low level and was being selected in the course of infection. Using the amplicon sequencing method for sample analyses of other patients the mutations K1383N and G1634R could be detected in low frequencies even before the initiation of antiviral treatment[31] or in increased frequencies after sofosbuvir and ribavirin combination therapy failure [32]. The RdRp mutations presented in Table 5 are known to be selected during ribavirin therapy as previously reported[10,18]. Interestingly, in the samples analysed here, Y1587F and K1383N mutations appeared within six weeks between the second and third time-point and increased

subsequently to 85 and 100%, respectively. Thus, the amplicon-based deep sequencing method has proven useful in capturing the dynamics of mutations in patients with CHE. Notably, this method allows pooling of amplicons from RdRp and HVR amplification maximising the data gained and reducing the costs.

Nevertheless, the presented methods have their limitations. One limitation of the lrPCR is the lower efficiency compared to shorter PCR amplicons. Samples with low viral load might be difficult or even impossible to amplify. Further limitations of the lrPCR were determined by the incomplete coverage of the coding region and by the low number of samples used for validation and testing. A further limitation was that the HEV subtype of all samples analysed here was HEV-3c, the most prevalent subtype in Germany. However, we expect that the here described approaches will be applicable also for other subtypes; although, marginal adaptation might be necessary. Regarding sequencing, limitations of the presented methods are more specific to the sequencing technologies used. Illumina sequencing requires fragmentation of the PCR products if these products exceed the length of two paired end reads while some insertions can be missed. On the other hand, if amplicon NGS is possible designing the right amplicon size can be challenging. For instance, we saw a drop of coverage in the middle-sequences of all RdRp amplicons due to the weak overlap of the paired end reads. Amplicons should be therefore of optimal length to present an advantage over, e.g., the tagmentation method. Long-read sequencing technologies, such as ONT, have relatively high error rates which is detrimental for variant analysis. Thus, we overcame this limitation by performing simultaneous sequencing with both Illumina and ONT technologies. Furthermore, we have not undertaken any studies to show the effect of insertions on HEV replication and genome stability. Underlying mechanisms relevant to viral gene duplication of human gene recombination with the HEV genome is not clear. This may be considered as limitation of our study. In this study, we focused only on identifying and validating insertions in the HVR region to gain more detailed insights into the complexity and variation of HVR insertions by using advanced sequencing techniques.

In summary, we have shown here that lrPCR is an elegant method that facilitates whole-genome sequencing with the Illumina and ONT technologies; however, the amplicon-based deep-sequencing of the HVR and RdRp region is the method of choice in screening HEV samples for insertions or mutations. Amplicon-based deepsequencing has the accuracy of the Illumina technology, allows multiplexing, and reduces sequencing costs. This approach should be considered especially in CHE patients where sustained virologic response cannot be achieved facing an elevated risk for the selection of viral variants with possible increased pathogenicity.

Data availability

All data needed to evaluate the conclusions in the paper are present in the manuscript and/or the Supplementary Materials. Further data will be made available to interested researchers by reasonable request to the first and/or the corresponding author.

Received: 8 June 2021; Accepted: 17 January 2022
Published online: 02 February 2022

References

1. Wedemeyer, H., Pischke, S. & Manns, M. P. Pathogenesis and treatment of hepatitis e virus infection. *Gastroenterology* **142**, 13881397.e1381. https://doi .org /10 . 1053/j.gastro .2012 .02 . 014 (2012).
2. Kamar, N. *et al. Lancet* **379**, 2477–2488. https://doi . org/10 .1016 / S0140- 6736(11)61849- 7 (2012).
3. Dalton, H. R. *et al.* Autochthonous hepatitis E in Southwest England: natural history, complications and seasonal variation, and hepatitis E virus IgG seroprevalence in blood donors, the elderly and patients with chronic liver disease. *Eur J Gastroenterol Hepatol* **20**, 784–790. https:// doi. org/ 10. 1097/ MEG. 0b013 e3282f5195 a (2008).
4. Said, B. *et al.* Pork products associated with human infection caused by an emerging phylotype of hepatitis E virus in England and Wales. *Epidemiol Infect* **145**, 2417–2423. https:// doi. org/ 10.1017 / S09502688 17001 3 88 (2017).
5. Spahr, C., Knauf-Witzens, T., Vahlenkamp, T., Ulrich, R. G. & Johne, R. Hepatitis E virus and related viruses in wild, domestic and zoo animals: a review. *Zoonoses Public Health* **65**, 11–29. https://doi .org / 10.1111 /zph .1240 5 (2018).
6. Purdy, M. A. & Khudyakov, Y. E. Evolutionary history and population dynamics of hepatitis E virus. *PLoS ONE* **5**, e14376. https:// doi. org/ 10. 1371/ journ al. pone.0014 3 76 (2010).
7. Shukla, P. *et al.* Cross-species infections of cultured cells by hepatitis E virus and discovery of an infectious virus-host recombinant. *Proc. Natl. Acad. Sci. U. S .A* **108**, 2438–2443. https:// doi. org/10 . 1073/pnas . 1018878 10 8 (2011).
8. Wang, H. *et al.* Recombination analysis reveals a double recombination event in hepatitis E virus. *Virol J.* **7**, 129. https://doi . org/ 10. 1186/1743 -422X-7 -12 9 (2010).
9. Smith, D. B., Simmonds, P., Jameel, S., Emerson, S. U., Harrison, T. J., Meng, X. J., Okamoto, H., Van der Poel, W. H., Purdy, M. A. & Group, I. C. o. T. o. V. H. S. Consensus proposals for classification of the family Hepeviridae. *J Gen Virol* **95**, 2223–2232. https:// doi.org /10 .1099 /vir.0 . 068429-0 (2014).
10. Debing, Y. *et al.* Hepatitis E virus mutations associated with ribavirin treatment failure result in altered viral fitness and ribavirin sensitivity. *J Hepatol* **65**, 499–508. https:// doi. org/ 10. 1016/j. jhep. 2016. 05. 002 (2016).
11. Lhomme, S. *et al.* insertions and duplications in the polyproline region of the hepatitis E virus. *Front. Microbiol.* **11**, 1. https://doi . org/10 .3389 /fmicb .2020 . 00001 (2020).
12. Lhomme, S. *et al.* Characterization of the polyproline region of the hepatitis E virus in immunocompromised patients. *J. Virol.* **88**, 12017–12025. https:// doi. org/ 10. 1128/ JVI. 01625- 14 (2014).
13. Kenney, S. P. & Meng, X. J. The lysine residues within the human ribosomal protein S17 sequence naturally inserted into the viral nonstructural protein of a unique strain of hepatitis E virus are important for enhanced virus replication. *J. Virol.* **89**, 3793–3803. https:// doi.org /10 . 1128/ JVI.03582 -1 4 (2015).
14. Kenney, S. P. & Meng, X. J. Identification and fine mapping of nuclear and nucleolar localization signals within the human ribosomal protein S17. *PLoS ONE* **10**, e0124396. https://doi .org /10 .1371 / journ al.pone . 01243 96 (2015).

15. Johne, R. *et al.* An ORF1-rearranged hepatitis E virus derived from a chronically infected patient efficiently replicates in cell culture. *J. Viral Hepat.* **21**, 447–456. https:// doi. org/10 .1111 /jvh .1215 7 (2014).
16. Lhomme, S. *et al.* Influence of polyproline region and macro domain genetic heterogeneity on HEV persistence in immunocompromised patients. *J Infect Dis* **209**, 300–303. https:// doi. org/10 . 1093/ infdis/ jit438 (2014).
17. Debing, Y., Gisa, A., Dallmeier, K., Pischke, S., Bremer, B., Manns, M., Wedemeyer, H., Suneetha, P. V. & Neyts, J. A mutation in the hepatitis E virus RNA polymerase promotes its replication and associates with ribavirin treatment failure in organ transplant recipients. *Gastroenterology* **147**, 1008–1011.e1007; quiz e1015–1006. https://doi .org /10 . 1053/j. gastro. 2014.08 . 040 (2014).
18. Todt, D. *et al.* In vivo evidence for ribavirin-induced mutagenesis of the hepatitis E virus genome. *Gut* **65**, 1733–1743. https://doi . org/ 10. 1136/ gutjnl- 2015- 311000 (2016).
19. Lauring, A. S. & Andino, R. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog* **6**, e1001005. https://doi .org / 10. 1371/ journ al.ppat . 10010 05 (2010).
20. Mohamed, S. *et al.* Comparison of ultra-deep versus Sanger sequencing detection of minority mutations on the HIV-1 drug resistance interpretations after virological failure. *AIDS* **28**, 1315–1324. https:// doi. org/ 10. 1097/ QAD.0000 0 0000000026 7 (2014).
21. EASL Clinical Practice Guidelines on hepatitis E virus infection. *J Hepatol* **68**, 1256–1271. https://doi .org /10 .1016 /j.jhep .2018 .03 . 005 (2018).
22. Wang, B., Harms, D., Papp, C. P., Niendorf, S., Jacobsen, S., Lütgehetmann, M., Pischke, S., Wedermeyer, H., Hofmann, J. & Bock, C. T. Comprehensive Molecular Approach for Characterization of Hepatitis E Virus Genotype 3 Variants. *J Clin Microbiol* **56**. doi:https://doi .org /10 .1128 /JCM .01686 -1 7 (2018).
23. Jothikumar, N., Cromeans, T. L., Robertson, B. H., Meng, X. J. & Hill, V. R. A broadly reactive one-step real-time RT-PCR assay for rapid and sensitive detection of hepatitis E virus. *J Virol Methods* **131**, 65–71. https://doi .org /10 . 1016/j. jviro met. 2005.07 .00 4 (2006).
24. Zucha, D., Androvic, P., Kubista, M. & Valihrach, L. Performance Comparison of Reverse Transcriptases for Single-Cell Studies. *Clin Chem.* https://doi .org /10 .1373 /clin c hem. 2019. 307835 (2019).
25. Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 183–195. https:// doi. org/ 10. 1261/ rna. 063479. 117 (2018).
26. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*, 081257. https:// doi. org/ 10.1101 / 081257 (2016).
27. Laver, T. *et al.* Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif* **3**, 1–8. https:// doi.org /10 . 1016/j. bdq. 2015. 02.00 1 (2015).
28. Shukla, P. *et al.* Adaptation of a genotype 3 hepatitis E virus to efficient growth in cell culture depends on an inserted human gene segment acquired by recombination. *J Virol* **86**, 5697–5707. https://doi .org / 10. 1128/ JVI. 00146-1 2 (2012).
29. Nguyen, H. T. *et al.* A naturally occurring human/hepatitis E recombinant virus predominates in serum but not in faeces of a chronic hepatitis E patient and has a growth advantage in cell culture. *J Gen Virol* **93**, 526–530. https:// doi. org/ 10. 1099/ vir.0. 037259-0 (2012).
30. Munoz-Chimeno, M., Cenalmor, A., Garcia-Lugo, M. A., Hernandez, M., Rodriguez-Lazaro, D. & Avellon, A. Proline-rich hypervariable region of hepatitis e virus: Arranging the disorder. *Microorganisms* **8**. https:// doi. org/ 10. 3390/ micro organ isms8 091417 (2020).
31. Gerhardt, F., Maier, M., Liebert, U. G., Platzbecker, U., Wang, S. Y., Papp, C. P., Bock, C. T., Berg, T. & van Bömmel, F. Early Detection of Hepatitis E Virus Ribavirin Resistance Using Next-Generation Sequencing. *Antimicrob Agents Chemother* **64**. doi:https:// doi. org/ 10. 1128/AAC . 01525- 19 (2019).
32. Schulz, M. *et al.* Combination therapy of sofosbuvir and ribavirin fails to clear chronic hepatitis E infection in a multivisceral transplanted patient. *J Hepatol* **71**, 225–227. https:// doi. org/ 10. 1016/j. jhep. 2019. 03. 029 (2019).

Acknowledgements

We are grateful to Steffen Zander (RKI) and the staff members of the laboratory of the FG15 at RKI for their excellent technical assistance.

Author contributions

C.P.P., D.H., B.W., J.HO., C.T.B.: Conceptualisation; C.P.P., D.H., B.W., A.T., C.T.B., J.HO.: methodology design;
C.P.P., V.C., A.T., J.HE.: Data acquisition and processing; C.P.P.: Investigation; C.P.P., P.B., M.K., B.A., J.HO.: Data analysis, interpretation, and project assistance; C.T.B., M.C., J.HO.: Resources; C.P.P.: Writing original draft preparation; C.P.P., P.K., C.T.B., J.HO.: Visualisation, drafting; D.H., B.W., A.T., B.A., V.C., P.K., M.C., J.H.O., C.T.B: Reviewing, editing; J.H.O., C.T.B.: Project supervision, sample acquisition, final approval; C.T.B.: Funding acquisition. All authors read, revised, and approved the final version of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This research was funded by grants from the German Federal Ministry of Health (BMG) with regard to a decision of the German Bundestag by the Federal Government (CHED-project grant No: ZMVI1-2518FSB705). D.H. is supported by the Claussen-Simon-Stiftung (Claussen-Simon Foundation; CSF) “Dissertation Plus” program, Germany, and the Fazit-Stiftung “Promotions Stipendium”. B.W. is supported by the China Scholarship Council (CSC), Beijing, China. B.A. is supported by ProFIT grant of the Investitionsbank Berlin (IBB, ProFIT No. 10169028, Berlin, Germany). The authors declare no conflict of interest. The funders BMG, CSF, Fazit, CSC, and ProFit had no role in the design of the study, in the collection, analyses or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-05706-w.

Correspondence and requests for materials should be addressed to C.-T.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Curriculum Vitae

For data protection reasons, my curriculum vitae will not be published in the electronic version of my work.

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Publication list

1. Biedermann P, Klink P, Nocke MK, Papp CP, Harms D, Kebelmann M, Thurmer A, Choi M, Altmann B, Todt D, Hofmann J, Bock CT. Insertions and deletions in the hypervariable region of the hepatitis E virus genome in individuals with acute and chronic infection. Liver Int. 2023.

2. Papp CP, Biedermann P, Harms D, Wang B, Kebelmann M, Choi M, Helmuth J, Corman VM, Thürmer A, Altmann B, Klink P, Hofmann J, Bock CT. Advanced sequencing approaches detected insertions of viral and human origin in the viral genome of chronic hepatitis E virus patients. Sci Rep. 2022;12(1):1720.

3. Harms D, Choi M, Allers K, Wang B, Pietsch H, Papp CP, Hanisch L, Kurreck J, Hofmann J, Bock CT. Specific circulating microRNAs during hepatitis E infection can serve as indicator for chronic hepatitis E. Sci Rep. 2020;10(1):5337.

4. Gerhardt F, Maier M, Liebert UG, Platzbecker U, Wang SY, Papp CP, Bock CT, Berg T, van Bömmel F. Early Detection of Hepatitis E Virus Ribavirin Resistance Using Next-Generation Sequencing. Antimicrob Agents Chemother. 2019;64(1).

5. Schulz M, Papp CP, Bock CT, Hofmann J, Gerlach UA, Maurer MM, Eurich D, Mueller T. Combination therapy of sofosbuvir and ribavirin fails to clear chronic hepatitis E infection in a multivisceral transplanted patient. J Hepatol. 2019;71(1):225-7.

6. Wang B, Harms D, Papp CP, Niendorf S, Jacobsen S, Lütgehetmann M, Pischke S, Wedermeyer H, Hofmann J, Bock CT. Comprehensive Molecular Approach for Characterization of Hepatitis E Virus Genotype 3 Variants. J Clin Microbiol. 2018;56(5).

7. Harms D, Wang B, Papp CP, Bock CT. Capturing virus evolution by proteomic bioinformatics: Hunting for characteristic mutations in the hepatitis E virus genome. Virulence. 2018;9(1):13-6.

Acknowledgments

I would like to express my deepest gratitude to Prof. Dr. C.- Thomas Bock and Prof. Dr. Jörg Hofmann for their invaluable support, patience, and feedback.

I am also extremely grateful to Dr. Andrea Thürmer, who generously provided methodological expertise and guidance.

Special thanks to Tony Lesmeister, first and foremost, for his continuous and unconditional moral support as well as for his invaluable help especially in editing and proofreading.

Many thanks to Dominik Greb, not only for his patience and moral support, but also for giving me valuable advice and suggestions.

Finally, I would like to mention my family, especially my parents and my sister, who made great sacrifices so that I could pursue my goals, and who have always supported me morally and unconditionally.