

ORIGINAL ARTICLE

Insertions and deletions in the hypervariable region of the hepatitis E virus genome in individuals with acute and chronic infection

Paula Biedermann^{1,2} | Patrycja Klink¹ | Maximilian K. Nocke³  | Christian-Patrick Papp^{1,2} | Dominik Harms¹ | Marianne Kebelmann¹ | Andrea Thürmer⁴ | Mira Choi⁵ | Britta Altmann¹ | Daniel Todt^{3,6}  | Jörg Hofmann^{2,7} | Claus-Thomas Bock^{1,8} 

¹Division of Viral Gastroenteritis and Hepatitis Pathogens and Enteroviruses, Department of Infectious Diseases, Robert Koch Institute, Berlin, Germany

²German Centre for Infection Research, Institute of Virology, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin Institute of Health, Berlin, Germany

³Department of Molecular and Medical Virology, Ruhr University Bochum, Bochum, Germany

⁴Genome Sequencing, Methodology and Research Infrastructure, Robert Koch Institute, Berlin, Germany

⁵Department of Nephrology and Intensive Medical Care, Charité Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

⁶European Virus Bioinformatics Center (EVBC), Jena, Germany

⁷Labor Berlin, Charité-Vivantes GmbH, Berlin, Germany

⁸Institute of Tropical Medicine, University of Tuebingen, Tuebingen, Germany

Correspondence

Claus-Thomas Bock, Department of Infectious Diseases, Robert Koch Institute, 13353 Berlin, Germany.
Email: bockc@rki.de

Funding information

German Federal Ministry of Health (BMG), Grant/Award Number: ZMVI1-2518FSB705; German Bundestag by the Federal Government; German Research Association DFG, Grant/Award Number: 448974291; German Federal Ministry of Education and Research (BMBF), Grant/Award Number: 01KI2106; Claussen-Simon-Stiftung; Claussen-Simon Foundation; CSF; Fazit-Stiftung 'Promotions Stipendium'; Investitionsbank Berlin, Germany, Grant/Award Number: 10169028; EFRE, Grant/Award Number: 10169096

Handling Editor: Alessio Aghemo

Abstract

Background and Aims: Hepatitis E virus is a major cause of acute hepatitis worldwide and can progress to chronicity in immunocompromised individuals. Various virus–host recombination events have been reported in the hypervariable region of the hepatitis E virus genome, but the patterns of assembly and selection remain unclear.

Methods: To gain further insight into viral evolution, we assessed the presence of low abundance variants in 16 samples from individuals with acute or chronic infection using a targeted next-generation sequencing approach.

Results: In seven samples, different variants with insertions and/or deletions were identified. Among them, eight insertions originating either from human genes or from the hepatitis E virus genome. Five different deletions could be identified. The amino acid composition of sequences with insertions showed a higher frequency of lysine and a lower abundance of proline, and additionally acetylation and ubiquitination sites were more frequent than in hepatitis E virus wild-type sequences.

Abbreviations: ALAT, alanine aminotransferase; ASAT, aspartate aminotransferase; γ -GT, γ -glutamyltransferase; cDNA, complementary DNA; DNA, deoxyribonucleic acid; GSTA1, glutathione S-transferase alpha 1; HEV, hepatitis E virus; HVR, hypervariable region; Indel, insertion and/or deletion; KTR, kidney transplant recipient; NGS, next-generation sequencing; NHL, non-Hodgkin lymphoma; 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; RTX, rituximab.

Daniel Todt, Jörg Hofmann, Claus-Thomas Bock contributed equally to this work and thus shared last authorship.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Liver International* published by John Wiley & Sons Ltd.

Conclusions: These findings suggest that the nucleotide composition of insertions and sites for post-translational modification may contribute to recombination events. Although the impact of low-level hepatitis E virus variants is uncertain, our results highlight the importance of a highly sensitive next-generation sequencing approach to capture the full diversity of hypervariable region.

KEYWORDS

gene insertion, hepatitis E virus, hypervariable region, next-generation sequencing, recombination, sequence deletion

1 | INTRODUCTION

Hepatitis E virus (HEV) is one of the leading causes of viral hepatitis worldwide and responsible for large outbreaks and epidemics with an acute course of disease in resource-poor countries due to waterborne/faecal-oral transmission of the virus.¹ In industrialized countries, sporadic cases with acute and chronic cases occur mainly due to zoonotic transmission.² The single-stranded positive sense RNA-virus belongs to the genus *Paslahepevirus*, species *Paslahepevirus balayani* in the family of Hepeviridae and is divided into 8 genotypes of which 4 (HEV-1–HEV-4) are known to be pathogenic in humans.^{3,4} In addition, one case was reported in which a person became chronically infected with HEV-7 after consuming camel meat and milk.⁵ HEV genotype 3 (HEV-3) is the predominant genotype causing HEV infection in industrialized countries.⁶ Over the past decade, an increasing number of autochthonous cases with HEV-3 have been identified, which can lead to a chronic course of infection.⁷ Chronic infections primarily occur in immunocompromised individuals, such as organ transplant recipients, patients with haematological disorders, and patients with HIV/AIDS.⁸ In immunocompromised individuals, treatment options are scarce with an increased risk of progression to liver cirrhosis.⁹ Therefore, chronic HEV infections are an increasingly important clinically and public health relevant issue.

The exact mechanisms leading to chronic infections are still unknown. Host factors resulting in alteration of the immune system may play an important role.⁸ The T-cell response has been identified as a key variable in the body's capability to control the virus.^{10–12} In addition, viral characteristics may contribute to chronification. The hypervariable region (HVR) shows a highly divergent nucleotide composition compared with other regions of the HEV genome and is known to influence virus adaption in vitro and in vivo.^{13,14} In recent studies, insertions in the HEV HVR have been shown to enhance replication ability in vitro.^{15–19} Furthermore, virus–host recombinants have been detected in samples of chronically infected individuals, suggesting a potential role of HVR diversity in the development of persistence.^{14–20} Recently, insertions have also been detected in an acutely infected individual.²¹ However, the impact of these recombination events in the HVR on the chronic course of the infection remains unclear.

Key Points

Hepatitis E Virus infections lead to acute or chronic courses. In chronically infected patients, fragments of human mRNA or of the own viral genome inserted into the hypervariable region have previously been reported. In this study, we found previously unknown genetic rearrangements and investigated their impact on protein sequences and their characteristics.

The aim of this study was to analyse and characterize the HVR variants of individuals with acute and chronic HEV infection to provide detailed evidence of the involvement of HVR variants in the chronification of HEV infection. The application of a targeted NGS approach allows the detection of low-frequency viral variants that may carry insertions that would not be detectable with Sanger sequencing. In addition, the insertions detected in the sequences of the analysed samples were characterized in terms of amino acid composition and post-translational modifications to provide further evidence on the potential role of HVR in the chronification of HEV infection.

2 | MATERIALS AND METHODS

2.1 | Patient samples

Sixteen blood samples including 15 EDTA-plasma samples and one serum sample from HEV-infected individuals were analysed. The selection was based on availability, availability of required volume, and viral load needed for PCR and sequencing. Baseline characteristics of the study cohort and individual patient characteristics were summarized in [Tables 1](#) and [2](#). The samples were obtained from routine diagnostics and were pseudonymized before further investigation. Sample 19-0195 was a follow-up sample of patient 1 described recently by Papp et al.²² Nine samples were from acute and seven from chronically HEV-infected individuals ([Table 2](#)). Chronic HEV infection is defined as the persistence of HEV RNA for more than 3 months.²³

TABLE 1 Baseline characteristics of the study cohort

	Reference range	Total (n = 16)	Acute HEV (n = 9)	Chronic HEV (n = 7)
Age (years)		57.5 (18–90)	59 (18–90)	56 (24–74)
Sex (ratio m:f)		1:8	9:0	1:2.5
ASAT (U/L)	<50 U/L	56 (27–814)	55 (27–814)	57 (45–152)
ALAT (U/L)	<41 U/L	103.5 (30–1201)	119 (52–1201)	74 (30–134)
Bilirubin (mg/dl)	<1 mg/dl	0.535 (0.09–4.02)	0.58 (0.3–4.02)	0.35 (0.09–3.84)
γ-GT (U/L)	8–61 U/L	111 (60–1401)	109 (60–811)	113 (76–1401)
Viral load (IU/ml)		1.19E+06 (3.24E+04–3.46E+07)	1.58E+06 (3.24E+04–2.92E+07)	6.39E+05 (3.68E+04– 3.46E+07)

Note: Data in median and range.

Abbreviations: m, male; f, female; ASAT, aspartate aminotransferase; ALAT, alanine aminotransferase; γ-GT, γ-glutamyltransferase; HEV, hepatitis E viral infection.

2.2 | Ethics statement

This study was approved by the local ethics committee (approval number 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.

2.3 | RNA extraction and quantification

RNA was extracted from 140 μl EDTA-plasma or serum using the QIAamp Viral RNA mini Kit (Qiagen) and the QIAcube (Qiagen) according to the manufacturer's instructions. For quantification, a real-time RT-PCR was performed as described previously.²⁴

2.4 | Genotyping and phylogenetic analysis

For genotyping and subsequent phylogenetic analysis, a nested RT-PCR assay modified from the protocol described by Wang et al²⁴ targeting a 307 nt fragment in the ORF 1 was performed. In detail, first-round RT-PCR was performed using the QIAGEN OneStep RT-PCR kit (Qiagen) and a modified primer pair (38_forward: GAGGC YATGGTSGAGAARG and 39_reverse: GCCATGTTCCAGACRGT RTTCC). The PCR was performed under the following conditions: 15 min at 95°C followed by 40 cycles consisting of 10 s at 94°C, 30 s at 50°C and extension at 72°C for 50 s and a subsequent final elongation at 72°C for 2 min. For nested PCR, a modified primer pair (37_forward: GGTTCGCGCTATTGARAARG and 27_reverse: TCRCAGAGTGYTTCTTCC) and HotStarTaq Master Mix (Qiagen, Hilden, Germany) were used under the following conditions: 15 min at 95°C, followed by 30 cycles at 94°C for 10 s, 49°C for 30 s, 72°C for 20 s and a final elongation step of 2 min at 72°C. Due to the short fragment length and lack of further NGS sequencing, no proof-reading polymerase was deemed necessary for this amplification step. The fragment was primer and quality trimmed to 256 nt. The HEV geno- and subtypes were assigned by phylogenetic

analysis using genotype 3 reference sequences according to the latest proposal of the ICTV for Hepeviridae and the most closely related sequence of each sample from GenBank according to the Max Score value.²⁵ To generate a cladogram, all genotype 3 reference sequences described earlier and genotype 3 sequences from the HEV-GLUE Sequence Database were downloaded.^{25,26} A multi-sequence alignment (MSA) of all sequences was built using Clustal-Omega version 1.2.4.²⁷ Based on this MSA, all loaded sequences were cut to the same genomic section as the patient sequences. A second MSA was created by QIAGEN CLC Genomics Workbench 22.0 (<https://digitalinsights.qiagen.com/>) providing the basis for the cladogram, which was computed and visualized using the neighbour-joining construction method from the same tool.

2.5 | Amplification and sequencing of the HVR

RNA was reverse transcribed using the SuperScript IV Reverse Transcriptase Synthesis System (Invitrogen) according to the manufacturer's recommendations with a prolonged incubation step of 20 min at 50°C and 15 min at 55°C using oligo(dT)s and random hexamers. The nested PCR and amplicon preparation was performed according to Papp et al²² with the following modifications: The Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Fisher Scientific) was used and the first PCR was performed under the following conditions: 30 s at 98°C and 35 cycles of 5 s at 98°C, 30 s at 62°C, 20 s at 72°C, followed by 2 min at 72°C. Conditions of the second PCR were: 30 s at 98°C, 13 cycles of 5 s at 98°C, 30 s at 62°C and 10 s at 72°C, 15 cycles of 5 s at 98°C and 30 s at 72°C, followed by 2 min at 72°C.

2.6 | Next-generation sequencing of HVR

Next-Generation sequencing was performed on the MiSeq Sequencer (Illumina, Inc.) with paired-end technology and 300 bp read length as previously described.²² In addition, sequences were confirmed by Sanger sequencing performed as described recently.²²

TABLE 2 Patient characteristics

Sample ID	Sex	Age ^a	AST ^b (U/L)	ALT ^b (U/L)	Bilirubin ^b (mg/dl)	γ-GT ^b (U/L)	Days since first positive PCR	Status of infection	Reason for impaired immune response	Viral load (IU/ml) (sample type)
19-0142	M	33	46	109	0.98	60	27	Acute	KTR	1.58E+06 (plasma)
19-0143	M	41	77	160	0.55	811	0	Acute		1.47E+06 (plasma)
19-0144	W	56	49	44	0.52	102	1679	Chronic	KTR	5.72E+05 (plasma)
19-0150	M	63	57	106	0.35	118	187	Chronic	KTR	3.68E+04 (plasma)
19-0151	M	66	45	46	0.26	113	359	Chronic	KTR	9.07E+05 (plasma)
19-0157	M	54	48	30	0.46	167	265	Chronic	KTR	1.48E+07 (plasma)
19-0162	M	18	27	52	0.55	61	6	Acute	RTX-patient	8.50E+06 (plasma)
19-0164	M	56	814	1201	0.61	495	0	Acute		2.92E+07 (plasma)
19-0166	M	90	n.a.	n.a.	n.a.	n.a.	0	Acute		3.18E+05 (plasma)
19-0171	M	74	60	134	0.09	76	686	Chronic	KTR	5.31E+04 (plasma)
19-0174	M	62	54	119	0.58	109	69	Acute	KTR	2.14E+06 (serum)
19-0194	M	90	208	265	4.02	152	0	Acute		3.24E+04 (plasma)
19-0195	M	24	152	74	3.84	1401	404	Chronic	RTX-patient, Bruton's disease ^c	3.46E+07 (plasma)
19-0196	W	50	68	101	0.35	93	489	Chronic	IgG-deficiency, T-lymphopaenia	6.39E+05 (plasma)
19-0201	M	59	n.a.	n.a.	n.a.	n.a.	8	Acute	NHL	3.13E+06 (plasma)
19--0204	M	64	55	63	0.3	63	0	Acute	Immuno-suppression	4.09E+05 (plasma)

Abbreviations: KTR, kidney transplant recipient; RTX, rituximab; NHL, non-Hodgkin lymphoma; n.a., not available.

^aAge – age in years at sample date.

^bFor reference ranges see Table 1.

^cAgammaglobulinemia.

2.7 | Bioinformatics

The different amplicon sequences with their different indel-patterns, each representing at least one viral haplospecies, can be handled as molecular 'operational taxonomic units' (OTU). Using the tool UPARSE, amplicon sequences were clustered into OTU's based on similarity.²⁸ Similar OTUs were grouped by length and their indel-pattern and within each group, the OTU with the largest cluster size was picked as group representative. Subsequently, we offered all representative OTUs as mapping reference in a python-based pipeline, which originally was designed for the molecular surveillance of HIV and includes a merging step using FLASH (version 1.2.11), trimming with Trimmomatic (version 0.36) and using the BWA mapper (version 0.7.15).^{29–32} Briefly, paired sequences were merged using FLASH with a minimum overlap of 15 and a maximum overlap of 300bp and a max-mismatch density of 0.25. Merged reads as well as unmerged reads were than quality trimmed by Trimmomatic with the options sliding window (8:20) and minimum sequence length of 50 base pairs. Subsequently, all survived reads were mapped against a reference file containing all representative OTUs of one sample plus the wild-type HVR sequence, using the mem-algorithm of BWA with a mismatch-penalty of 2. The resulting alignments were analysed by their codon frequencies and a consensus sequence for each offered reference was generated based on the most frequent codon-variants. Samtools (version 1.11) was used to analyse the distribution of reads between

the offered reference sequences.³³ Sequences with at least 100 reads and a minimum share of 0.5% reads were included. For the following analyses, the sequences were assigned into three different groups: sequences without indels (insertions and/or deletions), with deletions, and with insertions. Sequences containing both an insertion and a deletion were assigned to the group 'sequences with insertions'. The programming language Julia version 1.7.3 was utilized for data processing and pipeline organization, while R version 4.2.0 was used for visualization.^{34,35} For visualization, the HEV-3c reference strain wbGER27 (Acc. No. FJ705359.1) as proposed by Smith et al²⁵ was used.

2.8 | Analysis of regulation sites and of amino acid characteristics

To analyse the amino acid composition and for the creation of a graph, the biostrings package from the 'bioconductor software project' based on the programming language R was used.^{35,36} The following tools were used for the in silico determination of putative post-translational modifications: prediction of acetylation on internal lysines: GPS-PAIL with medium threshold and all putative acetylation (<http://pail.biocuckoo.org/>)³⁷; prediction of ubiquitination: BDM-PUB with a balanced cut-off (<http://bdmpub.biocuckoo.org/>); prediction of phosphorylation: (<http://www.cbs.dtu.dk/services/NetPhos/>)³⁸; prediction of glycosylation: (<http://www.cbs.dtu.dk/services/NetPhos/>)³⁸.

dtu.dk/services/NetNGlyc/)³⁹; prediction of methylation: (<http://www.jci-bioinfo.cn/iMethyl-PseAAC>).^{40,41} For the analysis of amino acid characteristics, the “Emboss Pepstats Tool” was used (https://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats/).⁴²

2.9 | Statistical analysis

Quantitative variables were compared by employing the Wilcoxon rank sum test. The analysis was performed using the programming language R.³⁵ $p < .05$ were considered statistically significant.

2.10 | Nucleotide sequence accession numbers

Sequences generated using NGS have been deposited in the NCBI GenBank under the accession numbers OL469119–OL469148.

3 | RESULTS

3.1 | Characteristics of the study cohort

Samples of 16 HEV-positive individuals were analysed, of which seven were chronically infected and nine had an acute infection. Fourteen individuals were male and two were female. The median age was 56 and 59 in the chronically and acutely infected group with an age range

between 24 and 74 years and 18–90 years, respectively (Table 1). The enzymes aspartate aminotransferase (AST), alanine aminotransferase (ALT), γ -glutamyltransferase (γ -GT) were moderately elevated with no significant differences between acutely or chronically infected individuals. Furthermore, there was no significant difference regarding the viral load between the group of acutely infected and the group of chronically infected patients (Wilcoxon rank sum test; data not shown) (Table 1). The duration of infection among the chronically infected patients ranged from 6 to 55 months (Table 2). Out of seven chronically infected individuals, five were kidney transplant recipients and two suffered from immunodeficiencies (Table 2). Two patients from the group of acutely infected patients developed a chronic infection (19-0174 and 19-0204). All sequences were assigned to HEV genotype 3. Fifteen sequences clustered within HEV subtype 3c and one sequence assigned to HEV subtype 3e (Figure 1).

3.2 | NGS amplicon sequencing of HVR

Sequences without indels were present in all samples. In addition, insertions and deletions were detected in the HVR domain of seven samples (Table S1). All indel-patterns found were in frame. In total, eight different insertions were identified in the HVR region of four samples, of which three were derived from individuals in a chronic state of infection (19-0144, 19-0195, 19-0196) and one from an acutely infected patient (19-0174) (Figure 2). All quasispecies with an indel contained more than one indel variant (Figure 2 and Table S1). The insertions

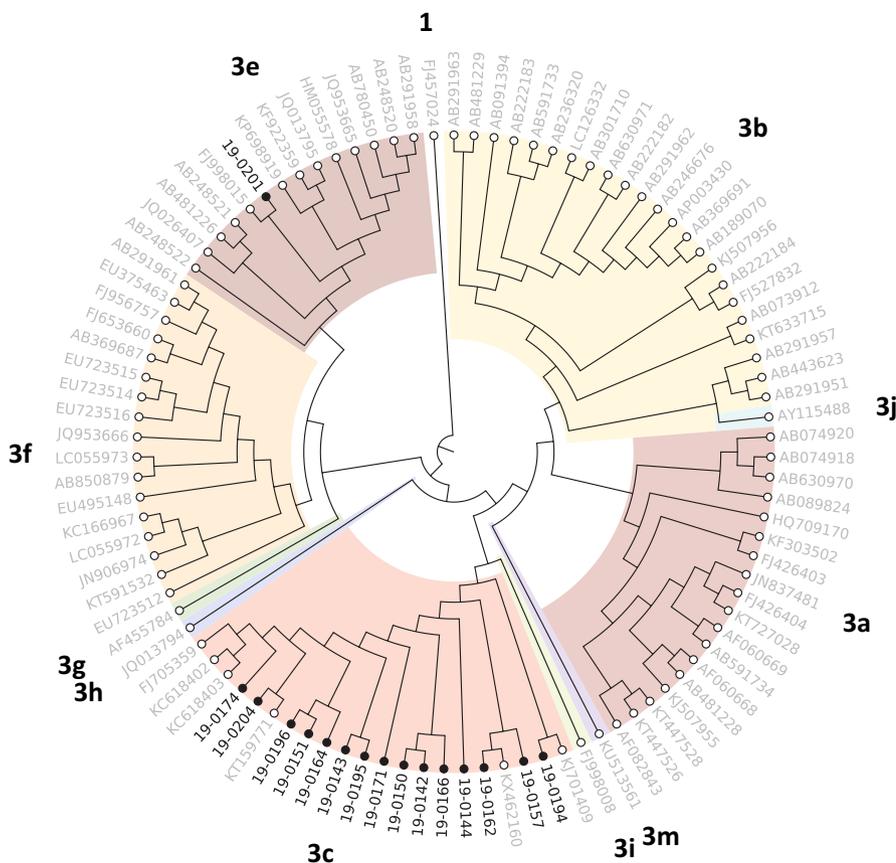


FIGURE 1 Cladogram of the patient-derived HEV sequences based on partial ORF1 sequences (256 nt). Subgenotyping results were confirmed with HEVnet (<https://www.rivm.nl/mpf/typingtool/hev/>). The HEV genotype 1 sequence FJ457024 was set as an outgroup. HEV sequences from this study are marked in black font (e.g. 19-0194) and were compared with HEV reference sequences representing HEV-3 (depiction: HEV subtype, accession number, e.g., 3e FJ998015).

ranged from 96 to 162 nt in size and were present with frequencies between 0.8% and 95.9% (Table S1). Among these eight insertions, five were derived from the HEV genome itself, either from the HVR or the RNA-dependent RNA polymerase (RdRp) region (19-0144, 19-0174, 19-0195, 19-0196; Figure 2 and Table S1). The remaining three insertions were derived from human genes (19-0195, 19-0196; Figure 2 and Table S1). Of these, one insertion showed the highest similarity with the human gene for the nucleoprotein Ahnak (19-0195). Ahnak is located on chromosome 11 and possibly plays a role in tumour metastasis and calcium regulation in cardiomyocytes and acts as a scaffold protein.⁴³ The insertion consisted of two Ahnak fragments (Figure 2 and Table S1). The second insertion in the HVR sequence of the sample 19-0195 was derived from the ribosomal protein L18 (RPL18). The remaining human-derived insertion in the HVR sequence of sample 19-0196 showed the highest identity to Glutathione S-transferase A1 (GSTA1), a gene that is abundantly expressed within the liver (Figure 2 and Table S1).⁴⁴ All insertions deriving from the HEV genome itself were made up of at least two fragments with one of them being a duplication of the adjacent region (Figure 2). Three insertions consisted of one HVR and one RdRp fragment. Two insertions were a combination of two different HVR fragments.

Deletions were observed in the HEV viral quasispecies of five samples (19-0142, 19-0151, 19-0157, 19-0174, 19-0196; Figure 2 and Table S1). Two samples within this group were collected from acutely infected patients (19-0142, 19-0174) and three from chronically infected patients (19-0151, 19-0157, 19-0196). The deletions ranged from 6 to 45 nt in size and were present in frequencies between 0.5% and 87.6% (Table S1). In the HVR sequence of sample 19-0174, a combination of two deletions consisting of a 7 and an 8 nt deletions was detected (Figure 2 and Table S1). Two variants were detected which featured both a deletion and an insertion in one strain (19-0174, 19-0196; Figure 2 and Table S1).

In addition, one 90 nt insertion not meeting our inclusion criteria (frequency: 0.4%) was identified in the HEV viral quasispecies of sample 19-0174 which was derived from the human gene NCKAP1 and was in frame (data not shown). The NCKAP1 Gene is located on chromosome 2 and is part of the WAVE complex that regulates lamellipodia formation.⁴⁵

3.3 | Position and fragment composition of insertions

In total, the position of the identified insertions varied, but all were located between nucleotide position (nt) nt 2227 and nt 2400

(numbering according to the HEV reference strain FJ705359.1). No pattern in regard to the nucleotides and amino acid composition before and after the insertion was identified.

3.4 | Amino acid composition of the HVR

For the amino acid analysis, the sequences were divided into three different groups: (1) sequences without indels, (2) sequences with insertions, and (3) sequences with deletions. Furthermore, the ratio of specific amino acids within each group (group (1)–(3)) was analysed (Figures 3 and 4). All the analysed sequences were rich in alanine (A), proline (P), and serine (S). The sequences with insertions contained significantly more lysine (K) (5.5% vs. 2.2%; $p < .0001$) and glycine (G) (4% vs. 2.5%; p -value: .013) compared with the groups without insertions (groups 1 and 3). On the other hand, these sequences (group 2) contained significantly less proline (P) (23.3% vs 28.5%; p -value: .00064), glutamate (E) (4.2% vs. 4.7%; p -value: .040) and valine (V) (5.6% vs. 6.8%; p -value: .032).

Sequences with deletions consisted of significantly more arginine (R) (8.4% vs. 6.2%; p -value: .0062), asparagine (N) (1.0% vs. 0.3%; p -value: .017), histidine (H) (1.3% vs. 1.2%; p -value: .012), and lysine (K) (3.0% vs. 2.2%; p -value: .048) compared with sequences without indels (group 1).

Sequences with insertions displayed significantly more polar (p -value: .029) and basic amino acids ($p < .0001$) and less small (p -value: .0023) and non-polar amino acids (p -value: .029) compared with sequences without indels (Figure 4). Sequences with deletions consisted of more basic (p -value: .014) and less acidic amino acids (p -value: .0065) (Figure 4). Further details are given in Table S2.

3.5 | Characteristics of sequences

Sequences with insertions provided significantly more putative sites for acetylation (p -value: .010), methylation (lysine) (p -value: .0018), and ubiquitination (p -value: .010) compared with sequences without indels (Figure 4). In detail, sequences with insertions provided between 3.2 and 8.6 putative sites for ubiquitination and acetylation per 100 amino acids in comparison to 1.2 and 2.4 putative sites in sequences without indels. The sequences without indels carried between one and two putative post-translational modification sites for acetylation and ubiquitination. The number of putative post-translational modifications was high in the sequences with

FIGURE 2 Amino acid (aa) sequence alignment of HEV HVR genomic regions. Depiction of insertions and deletions in the HEV HVR region of samples from HEV-infected individuals. The pseudonym of the patient is indicated above each aa sequence alignment (e.g., 19-0144). The upper sequence depicts the HEV reference strain wbGER27 (FJ705359.1) and numbering is according to this HEV reference strain. A hyphen indicates a deletion reference strain. The green bars indicate the human or HEV sequence insertion in the HEV HVR region. (A) displays sequences with deletions only, while (B) shows all sequences with insertions. Below the HVR variant name, the percentage of reads is displayed. For more detailed information, view Table S1. HEV sequence data of the patient samples are available at NCBI GenBank database (Acc. No. OL469119–OL469148).



(A)

wbGER27

2153 2183 2213 2243 2273 2303 2333 2363 2393
 FSSDFSPPEAAVAAPAATPGLRHPTPPVSDVWVLPSPSEEFQVDTAPTTPAPEPAQPSSSAGPKAPVRKPPTPPSPRTRRL

19-0142_2

0.7 % of reads

.....L.....K.....A..T...T...NP.....LM.....

19-0151_1_1

3.2 % of reads

.....TV...IS.....A.....VRS..PV.....M.....

19-0157_6

0.5 % of reads

.....A....A..SP.....N...R....L..V--A.....VL.P..VA.RV....AA..P.....

19-0174_1

87.6 % of reads

.....V...A.....A.GI.--RLLKNFRLTQ--P.....P.P.....L.....

(B)

19-0144_23

26.0 % of reads

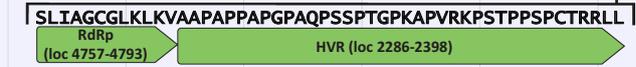
.....A.....H.....IS.....K.....A.....F.P.....S.....S.....



19-0144_25

64.5 % of reads

.....A.....H.....I...S.....A.....S.....A.....



19-0174_2

4.9 % of reads

.....V...A...C...A.GI.....L.A...L.A..P.....P.P.....L.....S.....



19-0174_2_a

1.0 % of reads

.....V...A...A.GI.--RLLKNFRLTQ--P.....P.P.....L.....S.....



19-0195_1

95.9 % of reads

.....S.....R.....S.....S.....P.PV...T.....A.....



19-0195_2

0.8 % of reads

.....SS...IR.....T...L.....P.....T.....A..P.....



19-0195_5

1.5 % of reads

.....ASS...R...S.....S.....P.....T.....



19-0196_1

62.8 % of reads

.....V.....S.....SA..IR.....A...S.....P.....L.S.....



19-0196_2

21.7 % of reads

.....V.....SH...S...I.....A...S.....P.....L.S.....



19-0196_3

9.5 % of reads

.....VV.....SH...SA..I.....L.....AL..S.L.....P.....S.L.S.....



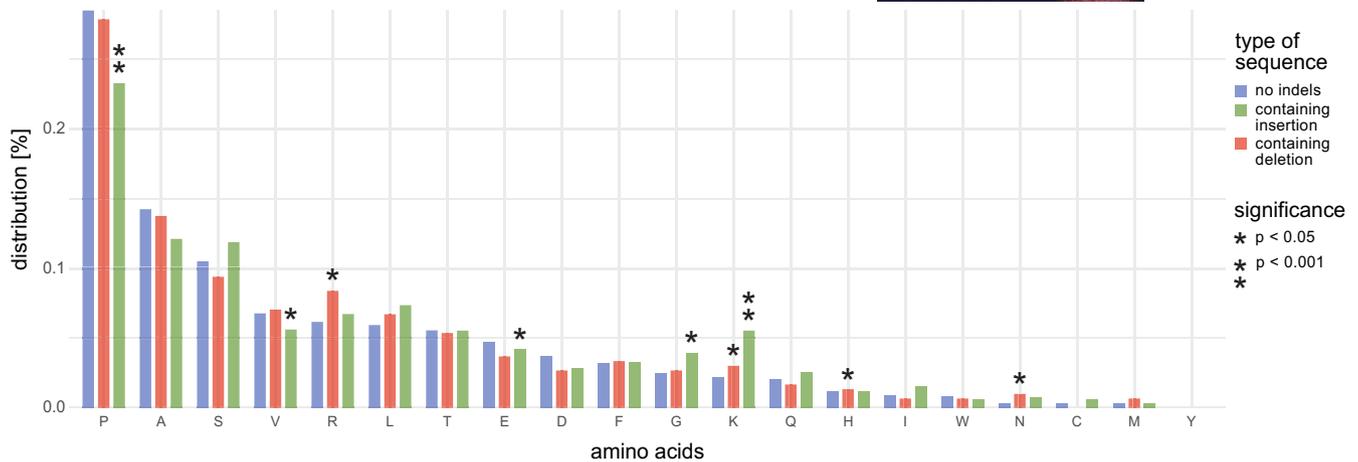


FIGURE 3 Amino acid composition of the HVR sequences of the three groups: Sequences without indels (in blue), with insertions (in green), and with deletions (in red). The mean percentage composition of amino acids is indicated on the left (y-axis). The single letter code was used to represent the amino acids (x-axis). The amino acids are ranked from most common to least common according to the group of sequences without indels. Asterisks indicate significant statistical differences between sequences without indels compared with sequences with insertions or sequences with deletions. $p < .05^*$ and $< .001^{**}$ indicate statistical significance.

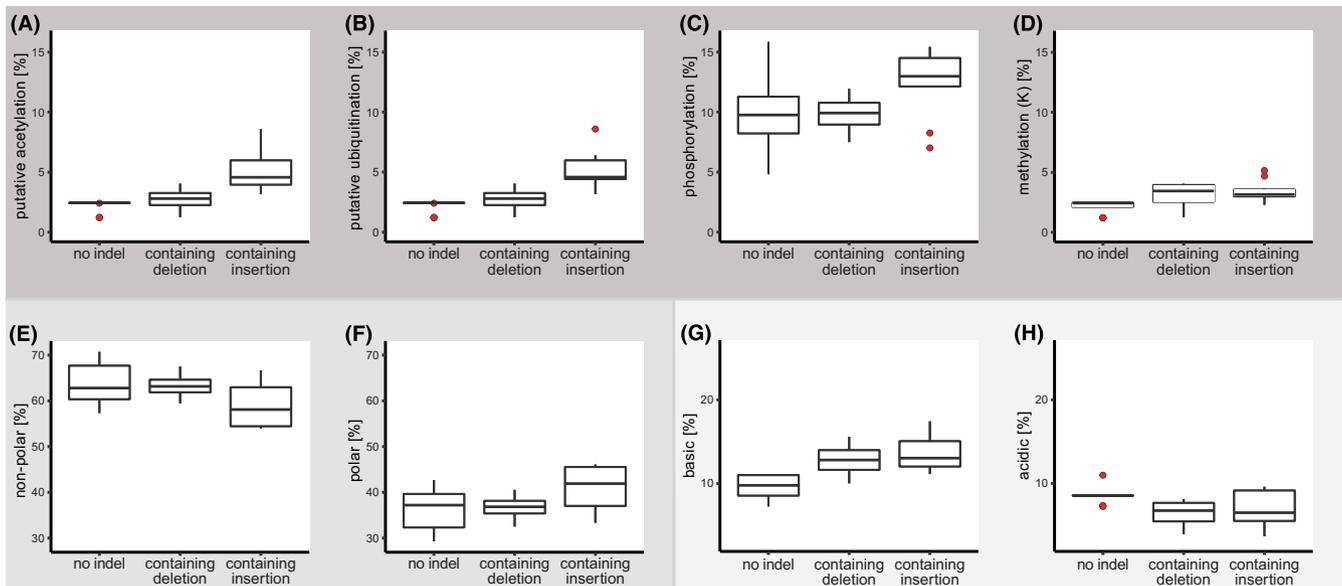


FIGURE 4 Comparison of potential post-translational modifications (dark grey background, A–D) and amino acid characteristics, that is, polarity (medium grey background, E, F) as well as ratio of acidic and basic amino acids (light background, G, H) in sequences without indels, sequences with deletions and sequences with insertions. Boxplots range from minimum to maximum, excluding outliers, which are represented by red dots. Second and third quartiles are displayed as a box divided by the median. The data shown are a subset of [Table S2](#) (Supplement), which points out differences between the three sequence groups.

the Ahnak insertion (19-0195_2) which carried the most sites with 11 putative acetylation and ubiquitination sites. Furthermore, the sequences with the GSTA1 insertion (19-0196_1, 19-0196_2) provided the second most ubiquitination and acetylation sites with eight putative post-translational modification sites. Sequences with deletions provided significantly more sites for putative acetylation (p -value: .048), ubiquitination (p -value: .048) and methylation (lysine) (p -value: .037) post-translational modifications with absolute numbers of putative post-translational modification sites ranging between one and three per sequence (Figure 4).

4 | DISCUSSION

Over the past decade, the awareness and assessment of the risk potential of the HEV infection has changed. This is partly due to the recent observation that autochthonous HEV infections in industrialized countries can cause severe chronic hepatitis in immunocompromised individuals, and partly due to new sensitive and specific molecular HEV detection methods that allow certain detection of the virus. However, the knowledge of the viral life cycle, molecular virology, and pathogenesis of HEV infection is still incomplete.

Recent studies revealed that recombination events in the HEV HVR can enhance replication ability *in vitro* and can influence virus adaptation *in vitro* and *in vivo*.^{13–19}

To gain further insights into HEV HVR recombination events, we analysed the sequences of the HVR obtained from 16 samples of individuals with acute or chronic HEV infection using a targeted NGS approach. This amplicon-based NGS approach also allows the detection of low-frequency variants in the viral population that would go unnoticed with Sanger sequencing.²² In this study cohort, all but two of the individuals were men, and the mean age was 57.5 years. This observation is not unexpected, as patients infected with HEV are known to be predominantly middle aged to elderly men.⁴⁶ In contrast to other findings where higher elevations of ALT were observed in acutely HEV-infected patients compared with chronically infected patients, we detected no significant difference (Tables 1 and 2).¹ In the phylogenetic analysis, all samples but one were assigned to HEV subgenotype 3c (Figure 1), the most common HEV subgenotype in Germany.⁴⁷

By applying the targeted NGS strategy, insertions, deletions, and a mixture of both were identified in the HVR quasiespecies of seven samples (Figure 2). Eight different insertions in the sequences of four patient samples were identified, all of which showing more than one recombinant variant in their viral population (Figure 2 and Table S1). The insertions derived either from the HVR and/or the RdRp region of the HEV genome or from human genes such as Anhak, GSTA1, and RPL18 (Table S1). An insertion deriving from GSTA1 gene has not been described before; however, Anhak and RPL18 sequences have already been reported recently²²; the RPL18 fragment identified in our study was 3 nt shorter. Viral-host recombinants in the HVR region including duplications of the HVR and insertions of the RdRp have already been detected in chronically HEV-infected individuals and recently in one acutely infected individual.^{14–17,19–21} In this study, three of the samples with insertions in the HVR sequences were obtained from chronically HEV-infected individuals (19-0144, 19-0195, 19-0196) and one from an acutely HEV-infected individual (19-0174). The sample from the acutely infected individual was drawn 69 days after the first positive PCR and the patient became chronically infected afterwards. Further investigations are necessary to assess whether screening for insertions using a sensitive NGS approach could become a powerful tool in a personalized medicine approach to identify individuals at risk for developing a chronic infection.

Analysis of the amino acid composition of the sequences identified in this study revealed significant differences between HEV HVR variants with and without insertions as previously reported (Figure 3).^{14,21} The sequences with insertions consisted of significantly more lysine and glycine and less proline and valine which is consistent with the results reported by Lhomme et al.²¹ However, especially the decrease in proline and the increase in lysine are remarkable. Proline is the key characteristic of this genome region.⁴⁸ Lysine, on the other hand, offers broad possibilities for post-translational modifications, and thus increased lysine residues indicate the possible significance in regulatory processes. In our observation,

sequences with insertions also contained significantly more positively charged amino acids (Figure 4). In Lhomme et al.²¹ human insertions increased the proportion of positively charged amino acids and HEV duplications decreased the proportion of negatively charged amino acids. In Munoz et al.,²⁰ human insertions equally increased the proportion of positive amino acids, but HEV duplications increased the proportion of negative amino acids. However, we did not subdivide into human insertions and HEV duplications for the calculations in this study. Overall, these observations are of particular importance as Scholz et al.⁴⁹ showed in reverse genetics experiments that the amino acid sequences not the nucleotide sequences or the length of the inserted fragment seemed to be important for the deregulation of viral replication.

The analysis of post-translational modifications revealed that HVR sequences with insertions provided significantly more sites for post-translational acetylation than sequences without indels. Furthermore, an increase in the putative sites for methylation (lysine) was observed in this study which has not been described before. Comparably with other reports, the increase in putative acetylation and ubiquitination sites in HVR sequences with insertions could be confirmed, but an increase in phosphorylation or in glycosylation sites was not observed, contrary to previous reports.^{20,21} The significance of the increased number of post-translational modification sites remains still unclear, however, as no further functional analyses were performed in our study.

In the analysed sequences, deletions in the HVR sequences of five patient samples could be identified. In our examination, sequences with deletions also provided significantly more putative sites for acetylation, ubiquitination, and methylation compared with sequences without indels. Deletions in the HEV genome have only been described by Nyugen et al so far and most recently by Munoz et al.^{19,20} While the deletions detected by Munoz et al were only up to 30nt in length, Nguyen et al observed large deletions in the HVR of a sample from a chronically HEV-infected patient. With a range from 42nt to more than 700nt, the size of the detected deletions was much larger compared with the deletions observed in this study. However, using our NGS approach, deletions of this dimension (up to 700nt) could not have been detected. Equally larger insertions could possibly go unnoticed due to the limited amplicon size with a 2×300bp sequencing approach. Another limitation of this study is the limited sample size. To gain a deeper insight into the prevalence and structure of indels in the HVR region larger cohorts should be studied.

5 | CONCLUSIONS

The data presented here supports the hypothesis that the amino acid composition with its increased proportion of lysine residues and putative post-translational modification sites within the HVR region is significant for the sustainability and breakthrough of indels in the persistence of HEV infection. Further studies are needed to evaluate the significance of an insertion in the acute stage of disease and later progression to chronic infection and whether this could be a

screening parameter to identify individuals at risk for a prolonged and potentially more severe course of HEV infection as part of a personalized medicine approach in the future.

ACKNOWLEDGEMENTS

We are grateful to Steffen Zander (RKI) and the staff members of the laboratory of the FG15 and MF2 at RKI for their excellent technical assistance. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

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: ZMV11-2518FSB705). D.T. was supported by a grant from the German Research Association DFG (Project number 448974291) and by the German Federal Ministry of Education and Research (BMBF) [project: VirBio, grant number: 01KI2106]. D.H. is supported by the Claussen-Simon-Stiftung (Claussen-Simon Foundation; CSF) 'Dissertation Plus' program, Germany, and the Fazit-Stiftung 'Promotions Stipendium'. B.A. is supported by ProFIT grant of the Investitionsbank Berlin, Germany (IBB, ProFIT No. 10169028 co-funded by EFRE no. 10169096). The funders BMG, DFG, BMBF, CSF, Fazit, 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.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICS APPROVAL AND PATIENT CONSENT STATEMENT

This study was approved by the local ethics committee (approval number 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.

ORCID

Maximilian K. Nocke  <https://orcid.org/0000-0003-4632-525X>

Daniel Todt  <https://orcid.org/0000-0002-3564-1014>

Claus-Thomas Bock  <https://orcid.org/0000-0002-2773-486X>

REFERENCES

- Kamar N, Bendall R, Legrand-Abravanel F, et al. Hepatitis E. *Lancet*. 2012;379(9835):2477-2488.
- Webb GW, Dalton HR. Hepatitis E: an underestimated emerging threat. *Ther Adv Infect Dis*. 2019;6:2049936119837162.
- Purdy MA, Drexler JF, Meng XJ, et al. ICTV virus taxonomy profile: hepeviridae 2022. *J Gen Virol*. 2022;103(9):1-2. doi:10.1099/jgv.0.001778
- Walker PJ, Siddell SG, Lefkowitz EJ, et al. Recent changes to virus taxonomy ratified by the international committee on taxonomy of viruses (2022). *Arch Virol*. 2022;167:2429-2440.
- Lee G-H, Tan B-H, Teo EC-Y, et al. Chronic infection with camelid hepatitis E virus in a liver transplant recipient who regularly consumes camel meat and Milk. *Gastroenterology*. 2016;150(2):355-357.e3.
- Izopet J, Tremoux P, Marion O, et al. Hepatitis E virus infections in Europe. *J Clin Virol*. 2019;120:20-26.
- Kamar N, Selves J, Mansuy J-M, et al. Hepatitis E virus and chronic hepatitis in organ-transplant recipients. *N Engl J Med*. 2008;358(8):811-817.
- Kamar N, Izopet J, Dalton HR. Chronic hepatitis e virus infection and treatment. *J Clin Exp Hepatol*. 2013;3(2):134-140.
- Kamar N, Garrouste C, Haagsma EB, et al. Factors associated with chronic hepatitis in patients with hepatitis E virus infection who have received solid organ transplants. *Gastroenterology*. 2011;140(5):1481-1489.
- Narayanan S, Abutaleb A, Sherman KE, Kottillil S. Clinical features and determinants of chronicity in hepatitis E virus infection. *J Viral Hep*. 2019;26(4):414-421.
- Suneetha PV, Pischke S, Schlaphoff V, et al. Hepatitis E virus (HEV)-specific T-cell responses are associated with control of HEV infection. *Hepatology*. 2012;55(3):695-708.
- Lhomme S, Abravanel F, Dubois M, et al. Hepatitis E virus quasispecies and the outcome of acute hepatitis E in solid-organ transplant patients. *J Virol*. 2012;86(18):10006-10014.
- Purdy MA, Lara J, Khudyakov YE. The hepatitis E virus Polyproline region is involved in viral adaptation. *PLoS One*. 2012;7(4):e35974.
- Lhomme S, Abravanel F, Dubois M, et al. Characterization of the polyproline region of the hepatitis E virus in immunocompromised patients. *J Virol*. 2014;88(20):12017-12025.
- Shukla P, Nguyen HT, Faulk K, 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*. 2012;86(10):5697-5707.
- Johne R, Reetz J, Ulrich RG, et al. An ORF1-rearranged hepatitis E virus derived from a chronically infected patient efficiently replicates in cell culture. *J Viral Hep*. 2014;21(6):447-456.
- Shukla P, Nguyen HT, Torian U, et al. Cross-species infections of cultured cells by hepatitis E virus and discovery of an infectious virus-host recombinant. *Proc Natl Acad Sci USA*. 2011;108(6):2438-2443.
- Kenney SP, Meng X-J. The lysine residues within the human ribosomal protein S17 sequence naturally inserted into the viral non-structural protein of a unique strain of hepatitis E virus are important for enhanced virus replication. *J Virol*. 2015;89(7):3793-3803.
- Nguyen HT, Torian U, Faulk K, 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*. 2012;93(3):526-530.
- Muñoz-Chimeno M, Cenalmor A, Garcia-Lugo MA, Hernandez M, Rodriguez-Lazaro D, Avellon A. Proline-rich hypervariable region of hepatitis E virus: arranging the disorder. *Microorganisms*. 2020;8(9):1417.
- Lhomme S, Nicot F, Jeanne N, et al. Insertions and duplications in the polyproline region of the hepatitis E virus. *Front Microbiol*. 2020;11:1.
- Papp CP, Biedermann P, Harms D, et al. 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.
- European Association for the Study of the Liver. Electronic address: easloffice@easloffice.eu; European Association for the Study of the Liver. EASL clinical practice guidelines on hepatitis E virus infection. *J Hepatol*. 2018;68(6):1256-1271.
- Wang B, Harms D, Papp CP, et al. Comprehensive molecular approach for characterization of hepatitis E virus genotype 3 variants. *J Clin Microbiol*. 2018;56(5):e01686-e01617.
- Smith DB, Izopet J, Nicot F, et al. Update: proposed reference sequences for subtypes of hepatitis E virus (species Orthohepevirus A). *J Gen Virol*. 2020;101(7):692-698.

26. HEV-GLUE 2019. Accessed May 12, 2022. <http://hev.glue.cvr.ac.uk/#/home>.
27. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 2011;7:539.
28. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods.* 2013;10(10):996-998.
29. Machnowska P, Hauser A, Meixenberger K, et al. Decreased emergence of HIV-1 drug resistance mutations in a cohort of Ugandan women initiating option B+ for PMTCT. *PLoS One.* 2017;12(5):e0178297.
30. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics.* 2011;27(21):2957-2963.
31. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114-2120.
32. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics.* 2010;26(5):589-595.
33. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011;27(21):2987-2993.
34. Bezanson J, Edelman A, Karpinski S, Shah VB. Julia: a fresh approach to numerical computing. *SIAM Rev.* 2017;59(1):65-98.
35. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing; 2021.
36. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.62.0. <https://bioconductor.org/packages/Biostrings2021>
37. Deng W, Wang C, Zhang Y, et al. GPS-PAIL: prediction of lysine acetyltransferase-specific modification sites from protein sequences. *Sci Rep.* 2016;6:39787.
38. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol.* 1999;294(5):1351-1362.
39. Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics.* 2004;4(6):1633-1649.
40. Qiu W-R, Xiao X, Lin W-Z, Chou K-C. iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. *J Biomol Struct Dyn.* 2015;33(8):1731-1742.
41. Wang P, Xiao X, Chou K-C. NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. *PLoS One.* 2011;6(8):e23505.
42. Madeira F, Ym P, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 2019;47(W1):W636-W641.
43. Davis TA, Loos B, Engelbrecht A-M. AHNAK: the giant jack of all trades. *Cell Signal.* 2014;26(12):2683-2693.
44. Sheehan D, Meade G, Foley VM, Dowd CA. Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. *Biochem J.* 2001;360(Pt 1):1-16.
45. Bladt F, Aippersbach E, Gelkop S, et al. The murine Nck SH2/SH3 adaptors are important for the development of mesoderm-derived embryonic structures and for regulating the cellular Actin network. *Mol Cell Biol.* 2003;23(13):4586-4597.
46. Dalton HR, Bendall RP, Rashid M, et al. Host risk factors and autochthonous hepatitis E infection. *Eur J Gastroenterol Hepatol.* 2011;23(12):1200-1205.
47. Adlhoch C, Avellon A, Baylis SA, et al. Hepatitis E virus: assessment of the epidemiological situation in humans in Europe, 2014/15. *J Clin Virol.* 2016;82:9-16.
48. Koonin EV, Gorbalenya AE, Purdy MA, Rozanov MN, Reyes GR, Bradley DW. Computer-assisted assignment of functional domains in the nonstructural polyprotein of hepatitis E virus: delineation of an additional group of positive-strand RNA plant and animal viruses. *Proc Natl Acad Sci USA.* 1992;89(17):8259-8263.
49. Scholz J, Falkenhagen A, John 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):762.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Biedermann P, Klink P, Nocke MK, et al. Insertions and deletions in the hypervariable region of the hepatitis E virus genome in individuals with acute and chronic infection. *Liver Int.* 2023;43:794-804. doi:[10.1111/liv.15517](https://doi.org/10.1111/liv.15517)