scientific reports

Check for updates

OPEN

Development of a direct whole genome sequencing for hepatitis A virus from serum and analysis of genetic characteristics

Daseul Yeo[1], Soontag Jung[1], Danbi Yoon[1], Seongwon Hwang[1], Dong Jae Lim[1], Songfeng Jin[1], Jinho Choi[2], Ki Ho Hong[3] & Changsun Choi[1]✉

Hepatitis A virus (HAV) is transmitted via the fecal–oral route, including through person-to-person contact and consumption of contaminated food. In 2019, 17,638 cases of HAV infection were reported in South Korea. The use of whole-genome sequencing (WGS) for rapid and accurate epidemiological investigation of HAV remains challenging due to the low viral titers and the lack of a cell culture system. Here, we developed a multiplex PCR (mPCR)-based next generation sequencing (NGS) method for direct WGS of HAV from serum using the Illumina platform. Overlapping primers were designed to generate amplicons covering the entire HAV genome, enabling successful sequencing from samples with viral titers as low as 3.0 log_{10} copies/μL. The method achieved a mapping rate of 98.95% and 98.32% genome coverage. All nine HAV strains analyzed were classified as genotype IA and showed > 99% sequence homology. Phylogenetic analysis of the whole genome, structural, and non-structural regions demonstrated clustering with HAV strains previously isolated in Japan. Correlation coefficients between WGS and individual gene regions approached 1.0, confirming the method's reliability for molecular epidemiology. This mPCR-based NGS approach provides a robust tool for HAV surveillance and facilitates high-resolution genomic analysis from non-culturable clinical samples.

Keywords Hepatitis A virus, Illumina platform, Whole-genome sequencing, Phylogeny, Dendrogram analysis

The hepatitis A virus (HAV), also called hepatovirus A, is a single-strand positive-strand RNA virus belonging to the *Picornaviridae* family. HAV has been the most common cause of viral hepatitis worldwide. HAV is a member of the *Hepatovirus* genus and includes the species *Hepatovirus ahepa–ishrewi*. Among them, *Hepatovirus ahepa* infects to humans and non-human primates[1]. It is transmitted through the fecal–oral route and can be classified into three main genotypes (I, II, and III) with two sub-genotypes (A and B). HAV consists of ~ 7.5 kb genome with the 5′-UTR, structural proteins (VP4, VP2, VP3, VP1, 2A (pX)) and non-structural proteins (2BC and 3A–D) and a 3′-UTR followed by a poly (A) tail[2].

The World Health Organization estimated a total of 159 million cases of acute HAV infections worldwide in 2019, leading to 39,000 deaths and 2.3 million disability-adjusted life years[3]. The United States (US) Centers for Disease Control and Prevention reported that between 2016 and 2024, 37 states experienced hepatitis A outbreaks, resulting in 44,926 cases, 27,457 hospitalizations, and 424 deaths[4]. In South Korea, more than 1,000 cases of hepatitis A have been reported annually from 2021 to 2023, with an average of 4,321 cases per year[5]. Notably, in 2019, the country faced a serious hepatitis A epidemic, with 17,598 cases reported and an incidence rate of 33.95 per 100,000[6].

HAV has a prolonged incubation period of 14–28 days and is primarily transmitted through fecal–oral routes. Causes include foodborne, waterborne, and close physical contact[3]. Person-to-person transmission occurs as the unvaccinated population increases, with factors like oral-anal sex, drug use, and travel from low-endemic countries to high-endemic areas[7–11].

[1]Department of Food and Nutrition, School of Food Science and Technology, College of Biotechnology and Natural Resources, Chung-Ang University, 4726 Seodongdaero, Daeduck-myeon, Anseong-si, Gyeonggi-do 17546, Republic of Korea. [2]Sanigen, 16 Heungan-daero, Dongan-gu, Anyang-si, Gyeonggi-do 14059, Republic of Korea. [3]Department of Laboratory Medicine, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea. ✉email: cchoi@cau.ac.kr

PCR amplification of the VP1/2A junction region is widely utilized for the detection and genotyping of HAV. Genotyping based on the VP1 region classifies HAV into three genotypes (I–III), whereas analysis of the VP1–P2A junction allows for classification into four genotypes (I–III and VII)[12]. Discrepancies in genotyping across different target regions have introduced ambiguities in viral sequence comparison and hindered comprehensive genomic analysis[13]. To overcome these limitations, next generation sequencing (NGS) has emerged as a powerful approach, enabling high-resolution genome characterization, particularly for culturable viruses such as SARS-CoV-2, Zika virus, and Chikungunya virus[14,15].

NGS-based whole genome sequencing (WGS) for HAV is a valuable tool for identifying viral strains and conducting genomic characterization. By employing multiplex PCR (mPCR)-based NGS, whole genome sequences of HAV were directly obtained from hepatitis A patient serum. To confirm the accuracy of the WGS method developed in this study, the mapping rate, coverage sequence quality data, and the tanglegram were used. The identified HAV strain characteristics were analyzed with a phylogenetic tree and SimPlot.

Results

WGS and data quality analysis of HAV from patient sera

Twelve serum samples (cau230022–cau230033) obtained from patients with confirmed HAV infection, as determined by anti-HAV IgM seropositivity, were subjected to analysis. The maximum, median, and minimum anti-HAV-IgM values were 14.9, 6.9, and 1.8 mg/dL, respectively (Table 1). Using quantitative reverse transcription PCR (RT-qPCR), HAV viremia was detected in all serum samples except cau230024, cau230028, and cau230033. The cau230030 had the highest HAV viral load at 4.16 $\log_{10}$ copies/μL, while the cau230032 had the lowest viral load at 1.23 $\log_{10}$ copies/μL. The median was 2.69 $\log_{10}$ copies/μL in cau230026.

WGS analysis was performed on nine serum samples with detected HAV viral load. Each sample was sequenced up to five times until the minimum sequencing depth exceeded 10×. The mapping rates ranged from 64.23% to 98.95%, with seven samples showing high mapping rates exceeding 90% (Table 1). The lowest mapping rates were observed in cau230025 (64.23%). Coverage rates were consistently high across all samples, ranging from 88.05% to 99.10% (Fig. 1). Notably, samples with higher viral loads (> 3 $\log_{10}$ copies/μL), such as cau230022 (3.29), cau230027 (3.55), cau230029 (3.50), and cau230030 (4.16) demonstrated both high mapping rates (> 98%) and coverage rates (> 97%). However, cau230032, despite having the lowest viral load (1.23 $\log_{10}$ copies/μL), maintained a relatively high coverage rate of 96.78%, although its mapping rate was lower (73.54%).

HAV whole-genome characteristic analysis

The sequence identity matrix between different samples is shown in Fig. S1. The HAV strains identified in this study showed high sequence similarity, with nucleotide sequence identities ranging from 98 to 99% (Fig. S1a). Additionally, the amino acid sequence identity was consistently 99% across all sequences. Comparison with LC373510.1, the reference strain from Japan, demonstrated 99% sequence homology at the nucleotide level in serum HAV strains (Fig. S1b). The amino acid sequences of serum HAV strains also exhibited 99% homology with IA genotypes of the reference strain.

The nine identified strains formed a monophyletic clade within genotype IA, as shown in Fig. 2. In the WGS phylogenetic tree, the serum HAV strains clustered with subgenotype IA strains, ranging from cau230025 to LC049340 from Mongolia. All serum HAV strains formed a genetically homogeneous lineage with LC373510 from Japan. The phylogenetic analysis revealed three nodes: (i) a single-strain node containing cau230025, (ii) a node grouping cau230029 and cau230030, and (iii) a larger node comprising cau230022, cau230027, cau230023, cau230026, cau230031, and cau230032. Specifically, cau230025 occupied a separate node, with nucleotide

Sample ID	GenBank number	Anti-HAV IgM[1]	Ct value	HAV viral load ($\log_{10}$ copies/μL)	Total read	Mapped read segment	Mapping rate (%)[2]	Coverage rate (%)[2]	Minimum depth
cau230022	PP389036	9.6	29.81	3.29	23,636,370	23,348,852	98.78%	98.87	11×
cau230023	PP389037	5.1	36.35	1.54	21,270,065	20,508,740	96.42%	97.13	12×
cau230024	NA	6.9	NA	NA	NA	NA	NA	NA	NA
cau230025	PP389038	14.9	35.05	1.88	24,063,686	15,457,271	64.23%	92.24	13×
cau230026	PP389039	5.1	32.03	2.69	13,618,195	13,475,437	98.95%	98.32	22×
cau230027	PP389040	9.6	28.82	3.55	17,985,303	17,754,777	98.72%	98.29	174×
cau230028	NA	6.9	NA	NA	NA	NA	NA	NA	NA
cau230029	PP389041	10	29.03	3.50	22,785,081	22,491,430	98.71%	99.10	95×
cau230030	PP389042	4.9	26.56	4.16	20,043,855	19,829,322	98.93%	97.38	11×
cau230031	PP389043	9.4	36.53	1.49	21,899,707	19,892,628	90.84%	88.05	47×
cau230032	PP389044	10.8	37.50	1.23	24,063,532	17,695,927	73.54%	96.78	60×
cau230033	NA	1.8	NA	NA	NA	NA	NA	NA	NA

Table 1. Comprehensive serological and genomic analysis of hepatitis A virus in clinical serum samples via multiplex PCR-based next generation sequencing. [1]A positive outcome for HAV IgM was indicated by a signal/cutoff value of 1.2 or greater. [2]The HAJFF-Kan12 strain (accession number: AB819870) was utilized for the virus coverage and reads mapping reference sequence. NA: Not available.

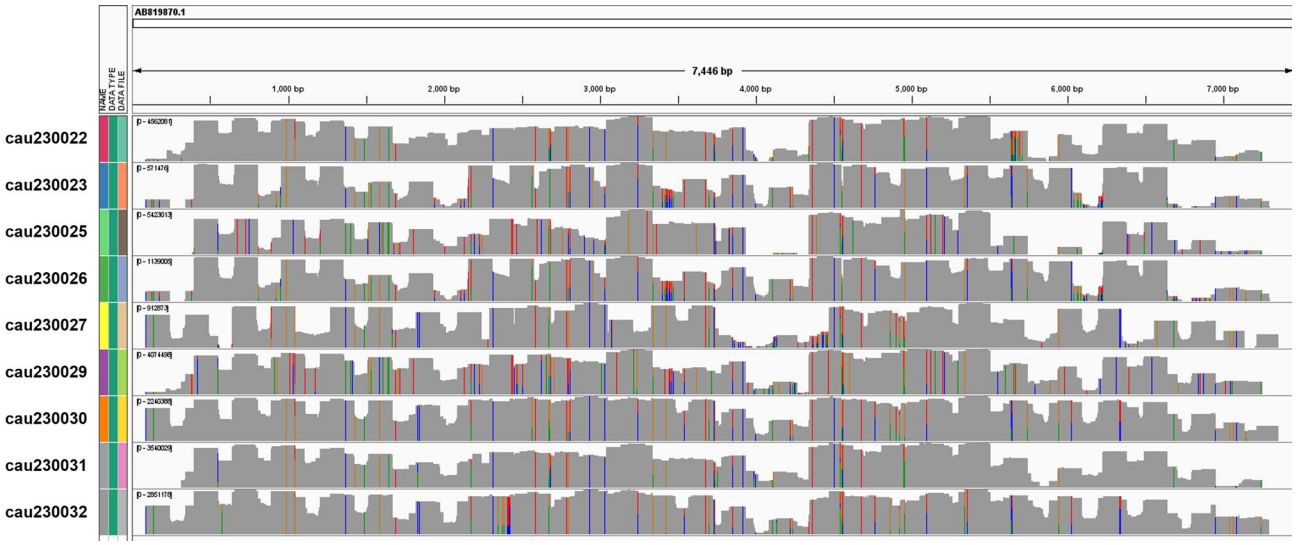

Fig. 1. Whole genome sequencing analysis of hepatitis A virus from serum samples in integrative genomics viewer. Gray bars indicate reads that perfectly match the reference genome. Base mismatches with the reference sequence are highlighted in color when they differ by more than 20% from the reference sequence (green: A, red: T, blue: C, orange: G).

sequence identities of 98.2% cau230029 and 98.3% cau230031 and amino acid sequence identities of 99.09% and 99.14%, respectively.

A phylogenetic tree based on the HAV capsid region (nucleotides 741–3674) is shown in Fig. 3a. The results reveal a topology consistent with the WGS phylogenetic structure. The HAV strains isolated in this study, which clustered within genotype IA, demonstrate a close genetic relationship to the reference strain LC373510. Within one major clade (from cau230025 to LC049340), the tree bifurcated into three nodes comprising cau230025–cau230030, cau230026–LC373510, and LC049340. Sequence identity analysis revealed high genetic similarity, with 99.88% identity between isolates cau230029 and cau230030. The genetic distance among cau230025 variants ranged from 97.58% to 97.70%. Additional phylogenetic analysis of the non-structural region (nucleotides 3675–7415, spanning proteins 2B to 3D) demonstrated concordant tree topology with both the WGS and capsid region phylogenies (Fig. 3b).

Tanglegram analysis for comparing different phylogenies

We compared with individual HAV genes phylogeny and whole-genome phylogeny. The capsid region VP4, VP3, VP2, VP1, and 2A showed phylogenetic clades with 155.56% (126/81), 76.54% (62/81), 93.83% (76/81), 74.07% (60/81), and 71.60% (58/81) interactions, respectively. The non-structural region of the phylogenetic tree of 2B, 2C, 3A, 3B, 3C, and 3D regions indicated conflicts of 88.89% (72/81), 88.89% (72/81), 106.17% (86/81), 111.11% (90/81), 88.89% (72/81), and 83.95% (68/81) respectively. The capsid region from VP4 to 2A showed 66.68% (54/81). The non-structural regions from 2B to 3D showed 37.04% (30/81).

The correlation between each HAV gene and the WGS dendrogram is shown in Fig. 4. Compared to the whole-genome sequences, phylogenetic clades were observed in the VP4, VP2, VP3, VP1, and 2A regions, with correlation coefficients of 0.70, 0.98, 0.99, 0.99, and 0.99, respectively. When comparing WGS with the dendrogram, all 2B and 2C regions exhibited a correlation of 0.99. The 3A, 3B, 3C, and 3D regions displayed dendrogram intersections of 0.96, 0.95, 0.99, and 0.99, respectively. All HAV gene regions, except the VP4 region, demonstrated high similarity values of 0.95 or above. The detailed data is provided in Table S1.

SimPlot analysis of HAV

SimPlot analysis revealed no evidence of recombination events among the examined sequences. Similarity comparison was performed between serum HAV isolates (PP389036–PP389044) and representative strains from HAV genotypes (IA, IB, IIA, IIB, IIIA, and IIIB), as shown in Fig. 5a. The analysis indicated that genotype IA exhibited the highest similarity score of 0.9 with the serum HAV sequences. A comparative analysis was further conducted between the identified HAV sequences (PP389036–PP389044) and a selection of HAV IA reference strains (AB623053, AB839692, KP177964, LC373510, LC416594, MN832786, and MN062166), as presented in Fig. 5b.

Discussion

To address discrepancies in genotyping caused by variations across target regions, this study introduces a novel application of mPCR based NGS for WGS of HAV directly from clinical serum samples with low viral titers. Unlike previous approaches that relied on high-titer, culturable strains or partial genome amplification, this method enables the direct sequencing of non-culturable HAV strains with viral loads as low as 3.0 $\log_{10}$ copies/μL.

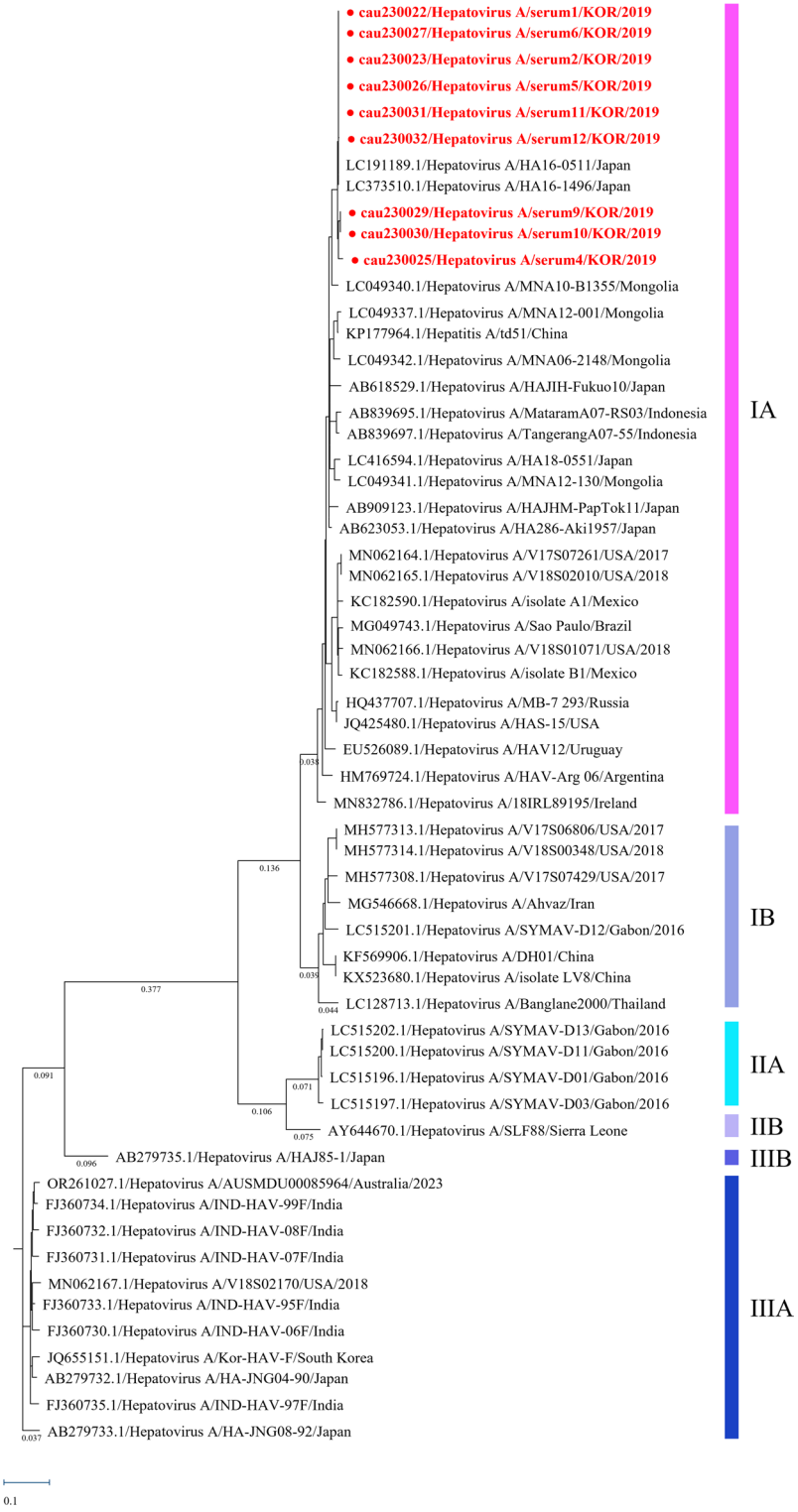

Fig. 2. Phylogenetic tree of hepatitis A virus (HAV) whole genome sequence. Red circles indicate serum HAV strains identified in this research. The phylogenetic tree was analyzed using the maximum likelihood method. The numbers on the branches represent the genetic distance between strains.

Comparative analysis between whole-genome and partial-region sequencing demonstrated consistent genotyping results and accurate identification of strain-specific characteristics across both capsid and non-structural regions. Correlation analysis among functional regions approached unity, with the VP1/2A junction—commonly used for HAV genotyping—showing particularly high concordance (correlation coefficient: 0.99). Although two samples (cau230025 and cau230032) exhibited relatively lower mapping rates (64.23% and 73.54%,

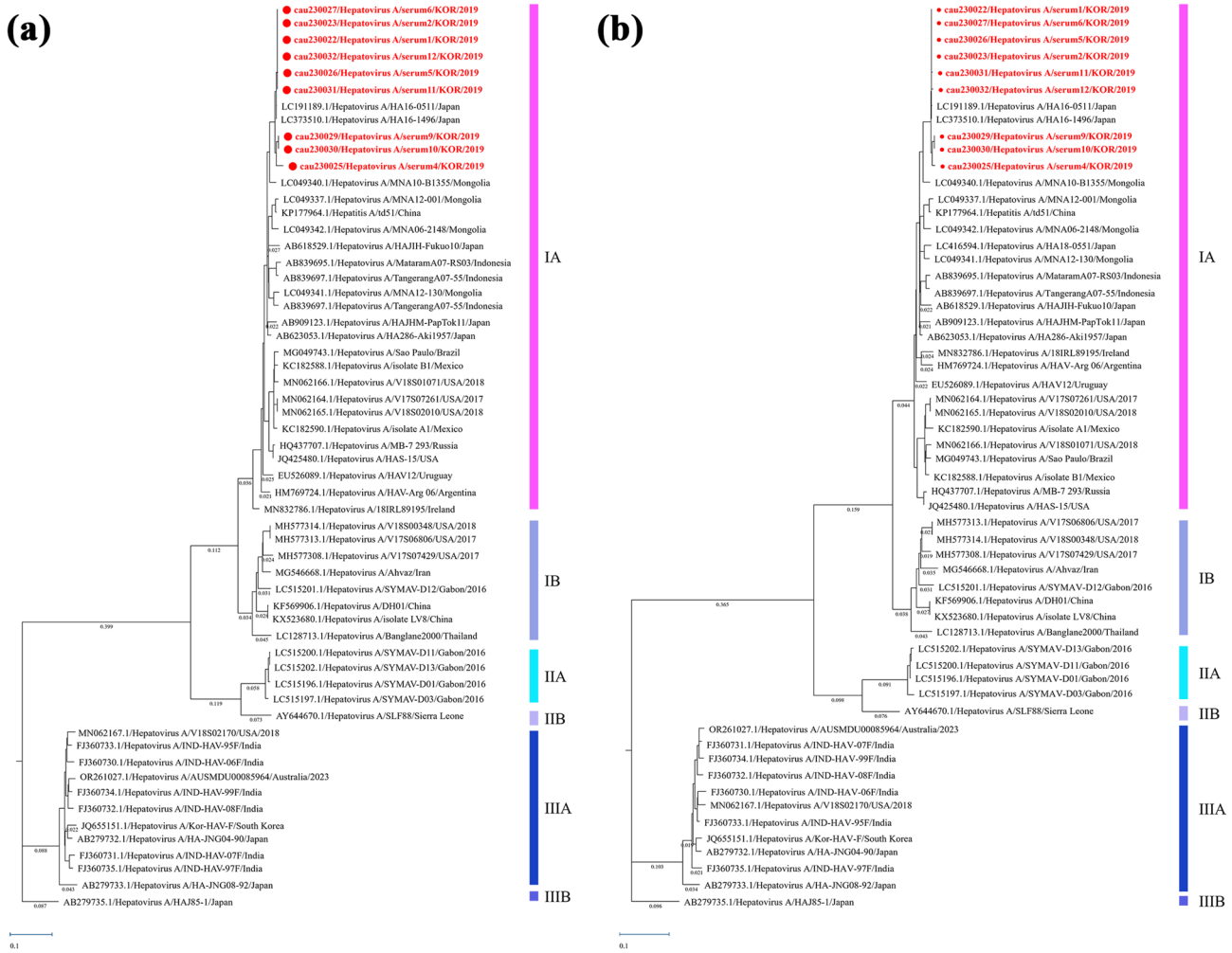

Fig. 3. Phylogenetic trees of hepatitis A virus (HAV) (**a**) capsid region, and (**b**) nonstructural region. Red circles indicate serum HAV strains identified in this study. The phylogenetic tree was analyzed using the maximum likelihood method. The numbers on the branches represent the genetic distances between strains.

respectively), both achieved genome coverage exceeding 95%, underscoring the robustness of the developed method. These results highlight the superior effectiveness of the mPCR-based NGS approach for WGS of HAV.

The established WGS method successfully sequenced HAV genome at viral concentrations as low as 3.0 $\log_{10}$ copies/μL. In contrast, previous PCR-free studies using nanopore sequencing were limited to culturable strains such as pHM175, requiring viral concentrations above 5 $\log_{10}$ copies/μL[12]. Other studies employing mPCR and Sanger sequencing targeted low-titer clinical samples, but only achieved mapping rates of approximately 19.9% at Ct values of 33[15]. In comparison, this study achieved a mapping rate of 98.95% and 98.32% coverage in a sample with a similar Ct value (cau230026, Ct values of 32.03).

Phylogenetic analysis revealed high sequence homogeneity among all examined HAV strains. The largest sequence identity gap among serum samples was observed between cau230025 and cau230026 (97.96%), while the highest similarity (99.90%) was between cau230022 and cau230027. Serum strains clustered closely with Japanese reference strains LC191189 and LC373510, which were isolated from hepatitis A patients and shared 99% identity[16,17]. HAV genotype IA strains identified in this study formed a regional cluster with strains previously reported in South Korea, Japan, Mongolia, and Indonesia. In contrast, other IA genotype clades from Mexico, US, and Brazil were genetically distinct, suggesting strong regional clustering. While most epidemiological studies to date have relied on seroprevalence data[2], this mPCR-based WGS approach can overcome limitations associated with long incubation period of HAV and region-specific sequencing practices[1,2].

Correlation studies between WGS and partial gene segments confirmed that the developed method is suitable for accurate genotyping and molecular epidemiology. Diagnostic tools typically rely on highly conserved 5'UTR regions for detection, while the VP1-2A region is often used for genotyping[2]. The mPCR-based NGS method demonstrated a high correlation (0.99) in VP1 and 2A regions. Compared to a previous NGS method that showed divergence in 76.3% of strains in the VP1/2A region[18], our approach is more consistent and reliable. Although the VP4 region showed relatively low correlation (0.70), this is likely due to its short sequence length (63 bp)[18], which reduces its phylogenetic resolution[19].

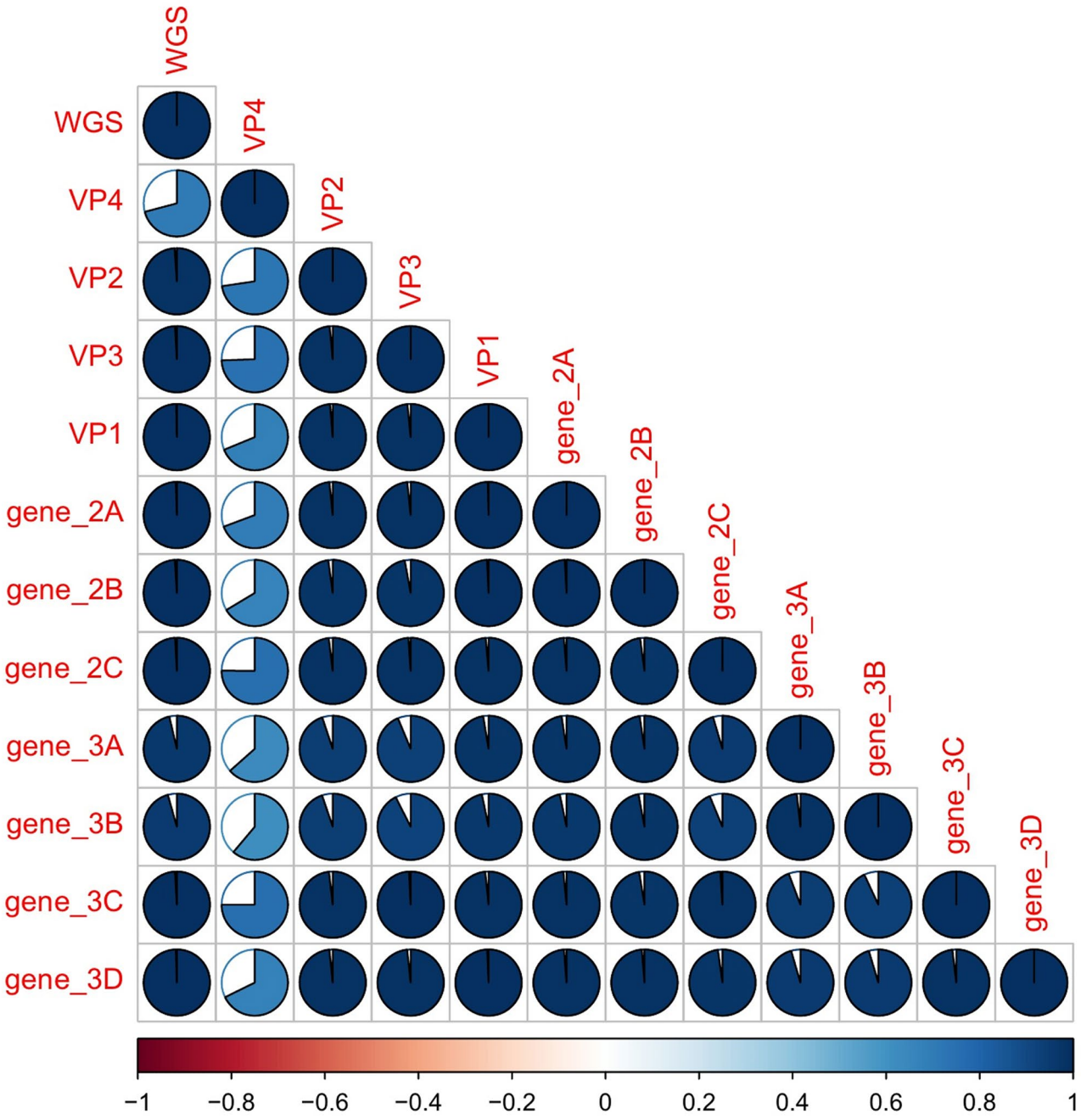

Fig. 4. The correlation plot between whole genome sequence phylogeny and partial gene phylogeny. The matrix displays pairwise correlations using pie charts, where the filled portion represents the magnitude of correlation, and the color indicates the direction and strength of the relationship.

SimPlot analysis further confirmed that strain LC373510 had the highest similarity to the serum isolates. This result was consistent across nucleotide, amino acid, and phylogenetic analyses (Fig. S1b, 3). Among HAV genes, VP2 displayed the most variability within genotype IA, likely due to its structural function. No evidence of HAV genomic recombination was observed in this study. This observation may be related to the measure of natural selection pressure at the molecular level. Previous studies have shown that HAV has one of the lowest ratio of non-synonymous to synonymous substitutions (dN/dS) ratios among RNA viruses (0.014 ± 0.002)[20], supporting its genetic stability. The high similarity among our samples may reflect their similar temporal and geographical origins.

Compared to existing methods, the mPCR-based NGS platform developed in this study demonstrates superior performance at lower detection thresholds and enables high-efficiency genome recovery. This capability is particularly valuable for early outbreak detection and the identification of mild or asymptomatic infections. Nevertheless, this study is limited by its focus on genotype IA and the absence of detailed clinical metadata.

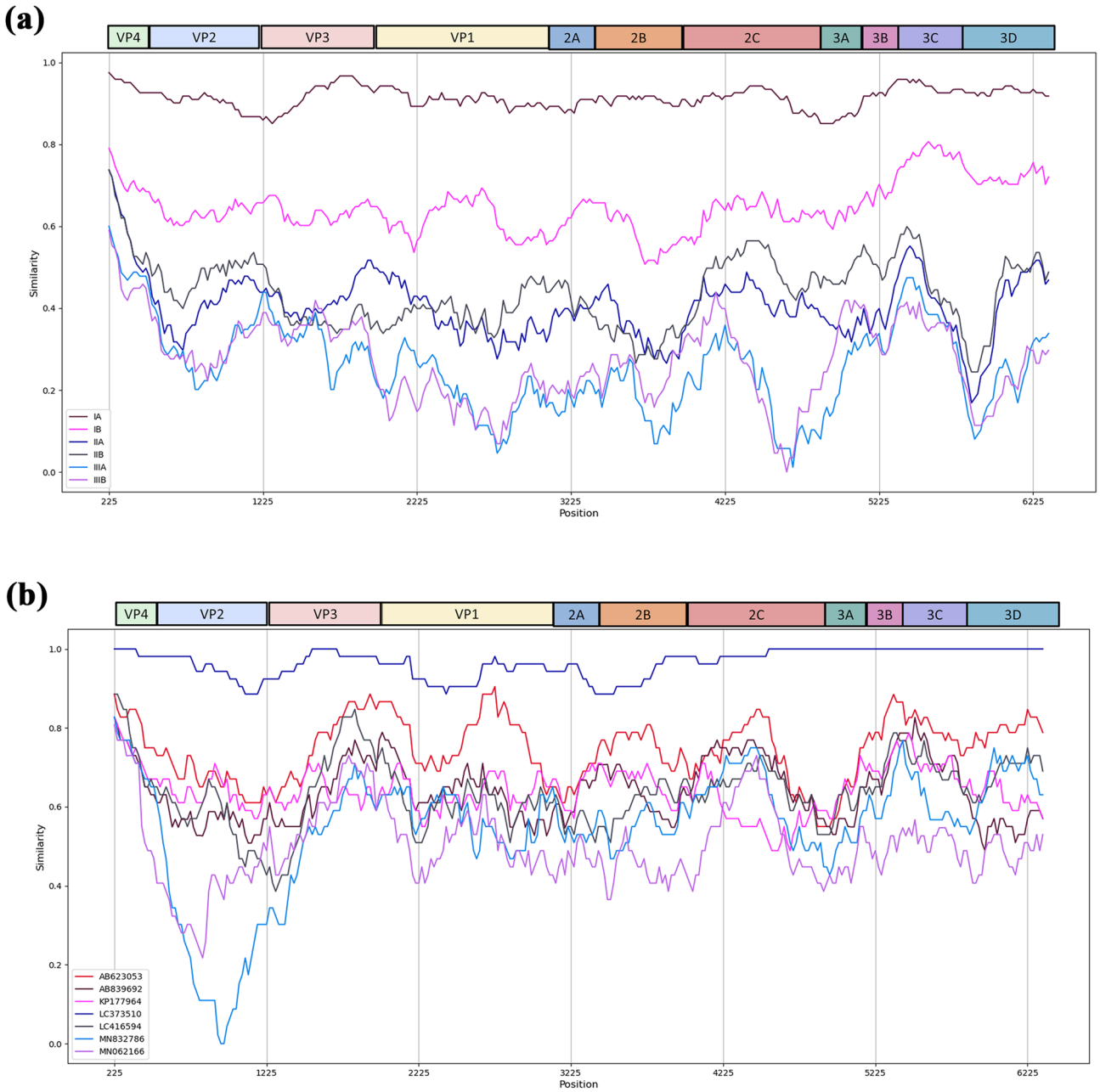

Fig. 5. Comparative genomic analysis of hepatitis A virus (HAV) strains. (**a**) SimPlot analysis between serum HAV stains and IA, IB, IIA, IIB, IIIA, and IIIB genotype strains and (**b**) comparison of serum HAV strains and IA genotype strains. Analysis was performed using SimPlot + + software with a Jukes-Cantor model and window size of 450 bp with steps of 20 nucleotides.

Future studies should aim to incorporate more diverse genotypes and integrate clinical information. Additionally, applying this method to other specimen types, such as urine or stool, could further expand the utility of HAV genomic surveillance.

In summary, this study presents a robust mPCR-based NGS method for WGS of HAV directly from serum samples. The method achieved high genome coverage, even in low-titer samples, and identified nine HAV-IA strains with strong inter-strain similarity. Phylogenetic analyses confirmed distinct geographical clustering patterns, and gene-level correlation analyses validated the method's reliability for molecular epidemiology. The developed platform offers an advanced tool for enhancing the speed and accuracy of HAV outbreak investigations and public health surveillance.

Materials and methods

Ethical approval

The study was approved by the Bioethics Committee of Chung-Ang University (Approval No. 1041078–202,007-BR-179–01) and conducted in accordance with the principles of the Declaration of Helsinki. Serum samples were collected in compliance with ethical standards, and informed consent was obtained from all participants. HAV infection was confirmed in twelve patients through serological detection of anti-HAV IgM antibodies. All serum specimens were stored at – 80 °C until further analysis.

Reverse transcription PCR (RT-PCR) and quantitative PCR (qPCR)

Following the manufacture protocol, total RNA was extracted using an RNeasy mini kit (Qiagen, Hilden, Germany). The nucleic acid was stored at – 80 °C until analysis. cDNA was synthesized from 1.0 µg of total RNA using the RevertAid H Minus First Strand cDNA Synthesis Kit (Thermo Fisher Scientific Inc, Cincinnati, OH, USA), which contains 50 µM oligo dT primers and 50 µM random hexamer primers[21].

HAV genomic quantification was performed via qPCR, targeting the 5'UTR region using previously validated primer sequences[22]. The qPCR was conducted using a CFX96™ Real-Time PCR system (Bio-Rad, Hercules, CA, USA) with Premix Ex Taq™ (2X) (Takara, Shiga, Japan). Quantitative standard curves were generated using genomic DNA from reference HAV strain (VR-3257SD, ATCC, Manassas, VA, USA). Viral copy numbers were determined through linear regression analysis, achieving a correlation coefficient (R^2) of 0.99.

Experimental design

We describe a method for direct WGS for HAV from serum samples. The method proceeds in four stages: (i) mPCR primer pool design, (ii) multiplex PCR, (iii) sequencing on Illumina platform and (iv) bioinformatics analysis and quality control (QC) (Fig. 6).

Multiplex PCR primer pool design

The mPCR primers were designed using custom-designed primers developed through Primal Scheme (http://primal.zibraproject.org)[15]. Six reference genomes were used as templates: HAJFF-Kan12 (accession no. AB819870.1), HM-175 wild-type strain (M14707), SYMAV-D13/Gabon/2016 (LC515202), SLF88 (AY644670), Kor-HAV-F (JQ655151), and HAJ85-1 (AB279735). As illustrated in Fig. 6a, two primer sets—Set A and Set B—were designed to span the entire HAV genome. Each set contained 18 forward and 18 reverse primers (0.075 µM each). The pooled primers from each set were used in parallel amplification reactions to ensure comprehensive genome coverage (Table 2)[15].

Multiplex PCR for illumina sequencing

Prepared mPCR primer pools were used for NGS in Fig. 6b. The reaction mixture of 25 µL contained 12.5 µL of 2 × Platinum SuperFi II PCR Master mix (Thermo Fisher Scientific Inc) 50 µM of primer set A (including pool 1 and 2) or set B (including pool 3 and 4), 6.20 µL of RNase-free water, and 2 µL of cDNA template. The mPCR was performed at 98 ℃ for 30 s, followed by 40 cycles at 98 ℃ for 15 s, 60 ℃ for 5 min, and a final cycle at 72 ℃ for 1 min. Using Qubit 4.0 (Life Technologies, Carlsbad, CA, US), over 10 ng/µL (total amount 500 ng) PCR product was used for Illumina sequencing.

Sequencing on illumina platform

The DNA library was constructed using the TruSeq Nano DNA Library Prep Kit (Illumina, San Diego, CA, US), according to the manufacturer's instructions. After the genome pooling process, the quality of the library was assessed using a bioanalyzer equipped with an Agilent DNA 1000 Kit (Agilent Technologies, Carlsbad, CA, US). Subsequently, NGS was conducted using a 150 bp paired-end NextSeq 2000 sequencer (Illumina)[23].

Bioinformatics analysis and quality control

The quality of the reads was checked using FastQC (v0.11.9). Using Trimmomatic (v0.39), adapters attached to reads were removed, and reads with an average accuracy of less than 99% (Q20) per 4 bp and reads with a total length of less than 200 bp were removed. Bwa (v0.7.17) maps read to reference sequences. Using Samtools (v1.10), we obtained the final HAV genome sequence of the present invention[24]. HAV genome mapping rates were calculated as 'mapped reads/total reads × 100' using a reference sequence HAJFF-Kan 12 strain. HAV genome coverage was calculated as 'sequence coverage/7.5 kp × 100'. The coverage rate was visualized using the Integrative Genomics Viewer (IGV)[25].

Sequence identity and phylogenetic analysis of HAV

The nucleotide and amino acid sequence identity were conducted by DNASTAR Lasergene MegAlign Pro (DNASTAR Inc., WI, US). The Expasy Translation tool (https://web.expasy.org/translate/) and Blast translated the amino acid sequences. The multiple-sequence alignments were conducted using MAFFT with the G-INS-I algorithm. The 28 sequences percent identity was visualized by R v4.5.1. The phylogenetic tree was constructed based on the multiple sequence alignment by the Maximum Likelihood method. The Maximum Likelihood method was performed with 1000 rapid bootstrapping replicates using the RAxML-NG (ver. 1.1.0) with the GTR+F+I+G4 nucleotide model. Phylogenetic clustering patterns were analyzed to assess the genetic relationships between the sequences. The tree topology was further validated using bootstrap values to ensure statistical reliability[24].

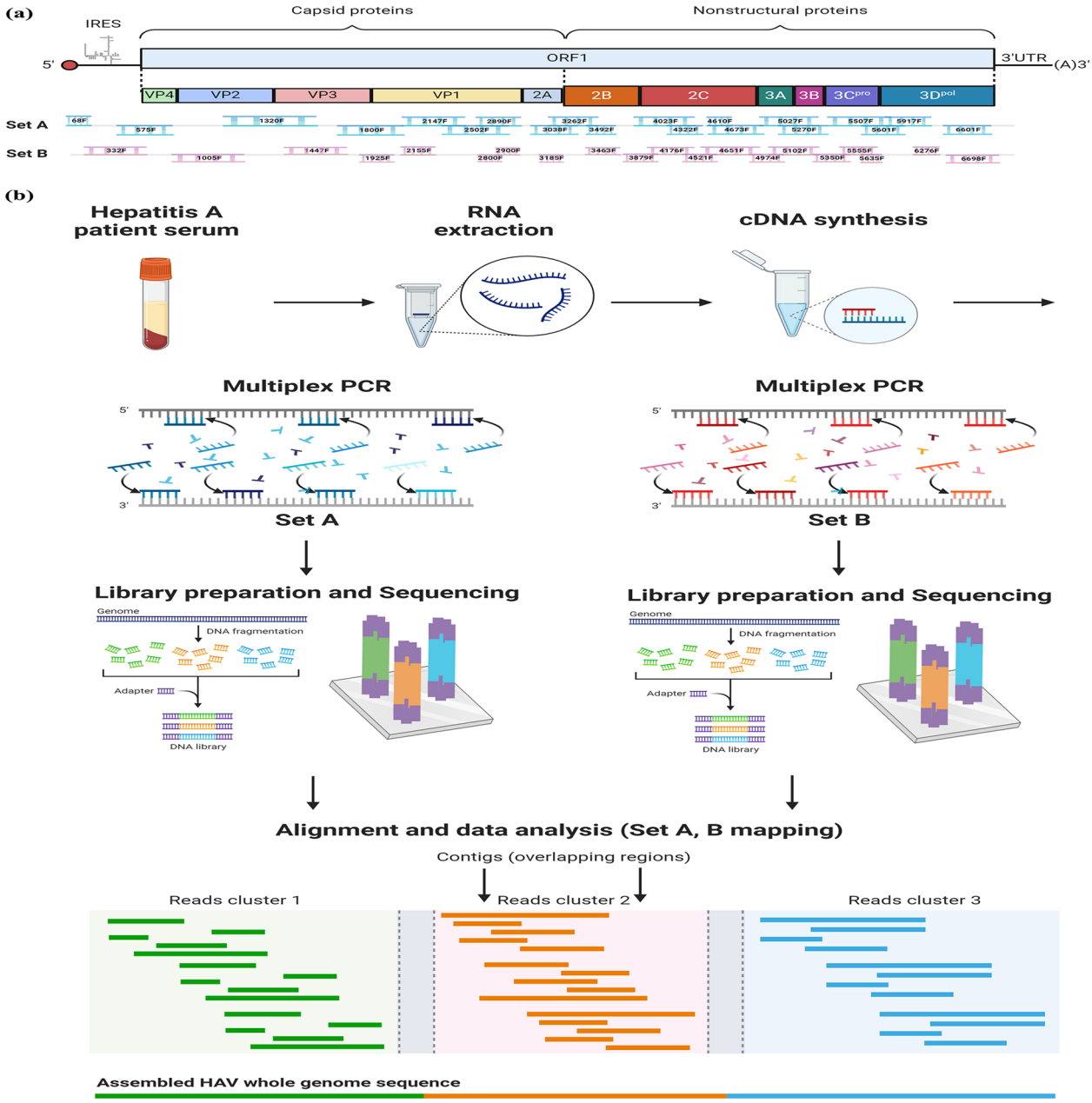

Fig. 6. Experimental design for hepatitis A virus (HAV) whole genome sequencing (WGS). (**a**) Schematic showing expected amplicon products for each set of HAV WGS. (**b**) Workflow of multiplex PCR-based NGS method for HAV WGS.

Dendrogram analysis comparing phylogenies of whole-genome and partial genome regions

The phylogenetic correlation between whole-genome sequences and partial genomic regions of HAV strains was evaluated using dendrogram analysis. The HAV genome was divided into eleven functional regions, including VP4, VP2, VP3, VP1, 2A, 2B, 2C, 3A, 3B, 3C, and 3D, based on established genomic annotations[26]. For each region, a separate phylogenetic tree was constructed using the Maximum Likelihood method with appropriate substitution models. Comparative analyses between the whole-genome and each partial region phylogenetic tree were performed using the 'dendextend' package in R v4.5.1. The dend.list function with the 'complete' method was applied to calculate pairwise tree correlations based on the Robinson–Foulds distance metric[27].

HAV sequence similarity analysis and recombination detection

SimPlot++ (v3.5.1; https://sray.med.som.jhmi.edu/SCRoftware/SimPlot/) was used to analyze nucleotide sequence similarity between HAV strains isolated from serum samples and reference sequences representing genotypes IA, IB, IIA, IIB, IIIA, and IIIB[28]. Multiple sequence alignment was performed using the MAFFT algorithm to ensure accurate alignment across the full genome. The similarity plot was generated using a sliding

Primer name	Multiple PCR set	Pool	Sequence (5'→3')	Size	Primer name	Multiple PCR set	Pool	Sequence (5'→3')	Size
68F	A	1	TGC TTG TAA ATA TTA ATT CCT GCA GGT	475	332F	B	3	ATA GGG TAA CAG CGG CGG ATA T	526
542R	A	2	ATT CCC TCA ATG CAT CCA CTG G		857R	B	4	ACC AGT CAC TGC AGT CCT ATC A	
575F	A	1	AGG TAC TCA GGG GCA TTT AGG T	450	1005F	B	3	TGG CTC ACT ACA CAT GCT CTT T	401
1024R	A	2	CAA CCC TTG AAC AGC AAA CTG C		1405R	B	4	TGA AGT GTA AGC TGA AGT TCC TGT	
1320F	A	1	AAA GAT CCA CAA TAC CCA GTT TGG	399	1447F	B	3	GAT CAG GAA GAT TGG AAA TCT GAT CC	531
1718R	A	2	TGG GTC AAC TGG AAT AAC TTT GAT CT		1977R	B	4	TCA AGG TTG ATT GCA CTC CTG T	
1800F	A	1	GTT TTG TTC CTG GCA ATG AGC T	488	1925F	B	3	CTG ATG TGG ATG GTA TGG CCT G	206
2287R	A	2	GTT GTT ATG CCG ACT TGG GGA T		2130R	B	4	TCC CAA GCC ATC CAA TGC TC	
2147F	A	1	GCA AGC ACC TGT GGG AGC TA	394	2155F	B	3	CAG GTT GGA GAT GAT TCW GGA GGT	422
2540R	A	2	AGC TGA CTC CTT YTC YAC CCA AG		2576R	B	4	GGT CCT CTA TAY AAC TGA AAC AGG T	
2502F	A	1	ACC TTC AAT TCA AAC AAT AAA GAG TAC ACA	409	2800F	B	3	TCA GAT TAG ATT GCC ATG GTA TTC TTA TT	141
2910R	A	2	TTG CAA TCT GAA TAG AAA CCA ATC CA		2940R	B	4	GCA TTT GAG TTC AAT GGA GCT CT	
2890F	A	1	CTG GTT TCT ATT CAG ATT GCA AAT TAC A	208	2900F	B	3	GGA GGA CAG AAG ATT TGA GAG TCA	281
3097R	A	2	CTG TCC TCC TCT GAT CTG GGA T		3180R	B	4	CCA AGA AAA TTT CAT TAT TTC ATG CTC CT	
3038F	A	1	AGT CAT ATA GAA TGC AGG AAG CCA	385	3185F	B	3	AGA GTT GTC AAA TGA AGT GCT TCC	201
3422R	A	2	TGT GGG AAA TTC ACT TTG GAC CA		3385R	B	4	GCA TCC ATC TCA AGA GTC CAC A	
3262F	A	1	ATG GAT GCT GGG GTT CTT ACT G	459	3463F	B	3	GAT TTG TGT TTC TTG CTG CAT TGG	518
3720R	A	2	AGC ACA GTG TTT ATA ATT TCA ACA GTC A		3980R	B	4	CCA TCA TTC TGG AGT CCA TTT GC	
3492F	A	1	TGG TCT AAA GTG AAT TTT CCA CAT GG	280	3879F	B	3	TGA ATT ATG CAG ATA TTG GTT GTT CAG T	294
3771R	A	2	CAA CAG TCA CAG AAT GAT GAA ACC C		4172R	B	4	ACC TGA GCC ATT GAA TGA ACA GT	
4023F	A	1	GAG AAA GCY MTT GAR GAA GCC GAT	455	4322F	B	3	AGA TGT GAG CCA GTT GTT TGC T	276
4477R	A	2	TCT GAC CAR TCC TCA TCT GTT GTG T		4597R	B	4	GTT GTG TTT TGG CCA ATA TCA TCA AT	
4521F	A	1	AGA GTA CAC ATT TCC AAT AAC TCT GTC	290	4610F	B	3	GTC AGA TTT TTG TCA ATT AGT GTC AGG A	140
4810R	A	2	GAC TCC TTT TCT ACC CAA GGC G		4749R	B	4	GGA TTT TTG AAA AAT GAA GCA GGT TTA AC	
4651F	A	1	GGC CAA AAC YCA ACA GAT GAR GAT	422	4673F	B	3	AAG GAA GCA ATT GAY CGY AGR CT	383
5072R	A	2	AYT CCC TGW GAC CAC AAC TCC A		5055R	B	4	GCT CCC ACA GCA ACC CAC TT	
4974F	A	1	AGT GAA TTC ATG GAG TTG TGG TCT	200	5027F	B	3	CAG TTG GCA TTC TTG GAG TGC T	513
5173R	A	2	AAC CAT CCT CCC ACA AGC AC		5539R	B	4	TGA AAC CCC ACA TCC AAA GAT TGA	
5102F	A	1	TGT TAC TAA TCA CAA GTG GGT TGC T	207	5270F	B	3	TGG GAG TGA AAG ATG ATT GGC T	401
5308R	A	2	GCT ATT TCC AAA GTT GAC TGA GAC T		5670R	B	4	TTC ACC TTT TCC TCT CCA TGC C	
5350F	A	1	AGA AAA ATG GAT GTG TGA GAT GGG	202	5555F	B	3	TGT TAA TTT CTG AGG GCC CAC T	458
5551R	A	2	AGA ACA ACA TCT TGA AAC CCC AC		6012R	B	4	GCT TTA GAA AAG GGC ATA GCT GC	
5917F	A	1	TGT GGT CTC CAA AAC GCT TTT T	365	6276F	B	3	TCT CCT GGG TTT CCT TAT GTC CA	413
6281R	A	2	ACA GTA TTG AAT AAG ATT CTC TGA GCC		6688R	B	4	GCG GAA AAA TCT AAA TCA AGA CCA AC	
6601F	A	1	TGG CAT AGA TCC TGA TAG ACA GTG	397	6698F	B	3	AGT CTT TTC CAG AGA TGT TCA AAT TGA	600
6997R	A	2	ACT ATC AAA ACA TCA TCT CCA TAA CAG AG		7297R	B	4	CCA AAT TGT CTT TTC TGA AAT TGC AGG	

Table 2. Primers for multiplex PCR-based NGS of hepatitis A virus. F: forward primer, R: reverse primer.

window approach with a window size of 450 base pairs and a step size of 20 base pairs. The Jukes-Cantor model was applied to calculate pairwise nucleotide distances across the aligned sequences[28].

Data availability

The accession numbers can be found in the article in Table 1. The datasets generated and analyzed during the current study are available in GenBank, the accession numbers PP389036–PP389044.

Received: 6 April 2025; Accepted: 30 June 2025
Published online: 11 July 2025

References

1. He, M., He, C.Q. & Ding, N. Z. Human viruses: An ever-increasing list. *Virology*, 110445 (2025).
2. Van Damme, P. et al. Hepatitis A virus infection. *Nat. Rev. Dis. Primers.* **9**, 51 (2023).
3. World Health Organization (WHO), Weekly Epidemiological Record, 2022, vol. 97, 40 [full issue]. Weekly Epidemiological Record 97, 493–512, https://iris.who.int/handle/10665/363396 (2022).
4. Zulli, A., Chan, E. M. & Boehm, A. B. Detection of Hepatovirus A (HAV) in wastewater indicates widespread national distribution and association with socioeconomic indicators of vulnerability. *mSphere* **9**, e00645-e1624 (2024).

5. Kim, Y. W., Chae, S. H. & Yang, J. S. Surveillance of Hepatitis A virus and its outbreak status in the Republic of Korea 2019 to 2022. *Public Health Weekly Rep.* **17**, 404–417. https://doi.org/10.56786/PHWR.2024.17.10.2 (2024).
6. Hyun, J. H., Yoon, J. Y. & Lee, S. H. A case-control study of acute hepatitis A in South Korea, 2019. *Osong Public Health Res. Perspect.* **13**, 352 (2022).
7. Beauté, J. et al. Travel-associated hepatitis A in Europe, 2009 to 2015. *Eurosurveillance* **23**, 1700583 (2018).
8. Foster, M. et al. Hepatitis A virus outbreaks associated with drug use and homelessness—California, Kentucky, Michigan, and Utah, 2017. *Morb. Mortal. Wkly Rep.* **67**, 1208 (2018).
9. Lee, Y. L. et al. Less severe but prolonged course of acute Hepatitis A in human immunodeficiency virus (HIV)–infected patients compared with HIV-uninfected patients during an outbreak: A multicenter observational study. *Clin. Infect. Dis.* **67**, 1595–1602 (2018).
10. Ndumbi, P. et al. Hepatitis A outbreak disproportionately affecting men who have sex with men (MSM) in the European Union and European Economic Area, June 2016 to May 2017. *Eurosurveillance* **23**, 1700641 (2018).
11. Nelson, R. Hepatitis A outbreak in the USA. *Lancet. Infect. Dis* **18**, 33–34 (2018).
12. Costa-Mattioli, M. et al. Molecular evolution of hepatitis A virus: a new classification based on the complete VP1 protein. *J. Virol.* **76**, 9516–9525 (2002).
13. Batista, F. M. et al. Whole genome sequencing of hepatitis A virus using a PCR-free single-molecule nanopore sequencing approach. *Front. Microbiol.* **11**, 874 (2020).
14. Kousathanas, A. et al. Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature* **607**, 97–103 (2022).
15. Quick, J. et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).
16. Miura, Y. et al. Hepatitis A virus genotype IA-infected patient with marked elevation of aspartate aminotransferase levels. *Clin. J. Gastroenterol.* **10**, 52–56 (2017).
17. Saito, A. et al. Guillain-Barré syndrome associated with acute hepatitis AA case report and literature review. *Rinsho Shinkeigaku=Clin. Neurol.* **58**, 574–577 (2018).
18. Lee, J. et al. Whole-genome sequencing and genetic diversity of severe fever with thrombocytopenia syndrome virus using multiplex PCR-based nanopore sequencing, Republic of Korea. *PLoS Negl. Trop. Dis.* **16**, e0010763 (2022).
19. Hole, G. Eight things you need to know about interpreting correlations. *Recuperado de* http://users.sussex. ac. uk/~ grahamh/RM 1web/Eight% 20things% 20you% 20need% 20to% 20know% 20about% 20interpreting% 20correlations. pdf (2014).
20. Lin, J. J., Bhattacharjee, M. J., Yu, C. P., Tseng, Y. Y. & Li, W.-H. Many human RNA viruses show extraordinarily stringent selective constraints on protein evolution. *Proc. Natl. Acad. Sci.* **116**, 19009–19018 (2019).
21. Martín-Alonso, S., Frutos-Beltrán, E. & Menéndez-Arias, L. Reverse transcriptase: from transcriptomics to genome editing. *Trends Biotechnol.* **39**, 194–210 (2021).
22. Jothikumar, N., Cromeans, T. L., Sobsey, M. D. & Robertson, B. H. Development and evaluation of a broadly reactive TaqMan assay for rapid detection of Hepatitis A virus. *Appl. Environ. Microbiol.* **71**, 3359–3363. https://doi.org/10.1128/AEM.71.6.3359-3363.2005 (2005).
23. Gladkikh, A. et al. Comparative analysis of library preparation approaches for SARS-CoV-2 genome sequencing on the illumina MiSeq platform. *Int. J. Mol. Sci.* **24**, 2374 (2023).
24. Cho, A., Tikhonenkov, D. V., Lax, G., Prokina, K. I. & Keeling, P. J. Phylogenomic position of genetically diverse phagotrophic stramenopile flagellates in the sediment-associated MAST-6 lineage and a potentially halotolerant placididean. *Mol. Phylogenet. Evol.* **190**, 107964 (2024).
25. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Giotechnol.* **29**, 24–26 (2011).
26. Li, W. & Tahiri, N. Host-virus cophylogenetic trajectories: investigating molecular relationships between coronaviruses and bat hosts. *Viruses* **16**, 1133 (2024).
27. Galili, T. dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720. https://doi.org/10.1093/bioinformatics/btv428 (2015).
28. Kvisgaard, L. K. et al. A recombination between two Type 1 Porcine Reproductive and Respiratory Syndrome Virus (PRRSV-1) vaccine strains has caused severe outbreaks in Danish pigs. *Transbound. Emerg. Dis.* **67**, 1786–1796 (2020).

Acknowledgements

This research was supported by the Chung-Ang University Graduate Research Scholarship (Academic scholarship for College of Biotechnology and Natural Resources) in 2024. This research was supported by a grant [grant number: 20162MFDS033] from the Ministry of Food and Drug Safety of the Republic of Korea for 2020–2021. Figure 6 was created using BioRender.com. Available at: https://BioRender.com/xqu0p79.

Author contributions

Daseul Yeo.: Data Curation, Methodology, Validation, Visualization, Writing – Original Soontag Jung.: Data Curation, Formal Analysis, Investigation, Methodology, Visualization Danbi Yoon.: Methodology Seongwon Hwang.: Investigation Dong Jae Lim.: Investigation Songfeng Jin.: Writing – Review & Editing Jinho Choi.: Formal Analysis, Project Administration Ki Ho Hong.: Resources Changsun Choi.: Conceptualization, Funding Acquisition, Project Administration, Supervision. The authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

The establishment of these serum samples was approved by the Chung-Ang University Bioethics Committee (approval number: 1041078-202007-BR-179-01), and consent was obtained for the sample collection. Therefore, the ethical approval statement and the need for informed consent were waived for this manuscript.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-09812-3.

Correspondence and requests for materials should be addressed to C.C.