High performance enrichment-based genome sequencing to support the investigation of hepatitis A virus outbreaks

Authors

Sara E. Zufan[1,2], Karolina Mercoulia[1,3], Jason Kwong[4,5], Louise M. Judd[1,6], Benjamin P. Howden[1,2,3], Torsten Seemann[1,2,3], Timothy P. Stinear[1,2]

Affiliation(s)

[1]Department of Microbiology and Immunology, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia

[2]The Centre for Pathogen Genomics, The University of Melbourne, Melbourne, VIC, Australia

[3]Microbiological Diagnostic Unit Public Health Laboratory, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia

[4]Department of Infectious Diseases, Austin Health, Heidelberg, VIC, Australia

[5]Department of Infectious Diseases, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia

[6]Doherty Applied Microbial Genomics, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia

Corresponding author

Sara E. Zufan, s.zufan@unimelb.edu.au

Keywords

Public health genomics, culture-independent diagnostics, whole genome sequencing, food-borne surveillance, molecular epidemiology

Abstract

Hepatitis A virus (HAV) infections are an increasing public health concern in low-endemicity regions due to outbreaks from foodborne infections and sustained transmission among vulnerable groups, including persons experiencing homelessness, those who inject drugs, and men who have sex with men (MSM), which is further compounded by aging, unvaccinated populations. DNA sequence characterisation of HAV for source tracking is performed by comparing small subgenomic regions of the virus. While this approach has been successful when robust epidemiological data are available, poor genetic resolution can lead to conflation of outbreaks with sporadic cases. HAV outbreak investigations would greatly benefit from the additional phylogenetic resolution obtained by whole virus genome sequence comparisons. However, HAV genomic approaches can be difficult because of challenges in isolating the virus, low sensitivity of direct metagenomic sequencing in complex sample matrices like various foods such as fruits, vegetables and molluscs, and difficulty designing highly multiplexed PCR primers across diverse HAV genotypes. Here, we introduce a proof-of-concept pan-HAV oligonucleotide hybrid capture enrichment assay from serum and frozen berry specimens that yields complete and near-complete HAV genomes from as few as four input HAV genome copies. We used this method to recover HAV genomes from human serum specimens with high Cτ values (34·7—42·7), with high assay performance for all six human HAV sub-genotypes, both contemporary and historical. Our approach provides a highly sensitive and streamlined workflow for HAV WGS from diverse sample types, that can be the basis for harmonised and high-resolution molecular epidemiology during HAV outbreak surveillance.

Importance

This proof-of-concept study introduces a hybrid capture oligo panel for whole genome sequencing (WGS) of all six human pathogenic hepatitis A virus (HAV) sub-genotypes, exhibiting a higher sensitivity than some conventional genotyping assays. The ability of hybrid capture to enrich multiple targets allows for a single, streamlined workflow, thus facilitating the potential harmonization of molecular surveillance of HAV with other enteric viruses. Even challenging sample matrices can be

accommodated, making it suitable for broad implementation in clinical and public health laboratories. This innovative approach has significant implications for enhancing multijurisdictional outbreak investigations, as well as our understanding of the global diversity and transmission dynamics of HAV.

57 # Introduction

58 Hepatitis A is a vaccine-preventable disease caused by hepatitis A virus (HAV), a nonenveloped
59 virus, with a small (7.5kb) positive-sense RNA genome in the family *Picornaviridae*. HAV is
60 transmitted faecal-orally, by consuming contaminated food or drink, as well as person-to-person
61 between close contacts.[1] It is the leading cause of acute viral hepatitis, symptoms can be prolonged
62 but rarely result in fulminant hepatitis. Endemicity can be classified as high, intermediate, low, or
63 very low and is generally associated with socioeconomic indicators such as hygienic and sanitary
64 conditions.[2] Incidence of HAV infections declined in many low-endemicity countries after the
65 implementation of routine childhood immunization by the early 2000s.[3,4] As a result, immunity in
66 adult populations has declined significantly, thereby increasing the number of susceptible individuals
67 who are more likely to experience and report symptoms of hepatitis.[5] Epidemiology in these regions
68 then shifted from cyclical community wide outbreaks associated with childcare facilities or
69 households with children, to sporadic outbreaks stemming from travellers or consumption of
70 contaminated food from high- and intermediate-endemicity regions.[2,6–9] Another epidemiological
71 shift began in 2016 when several North American and European countries observed a sharp increase
72 in cases driven by person-to-person transmission, particularly in vulnerable populations including
73 persons reporting injection drug use, experiencing homelessness, and men who have sex with men
74 (MSM).[10–19] The combination of infection in older adults and persons with one or more comorbidities
75 has resulted in an uptick in hospitalisations and death related to HAV infection.[5,13,20,21]

76 Human pathogenic HAV can be characterised into three genotypes (I, II, II), each with two
77 subgenotypes (A, B). Distribution of subgenotypes varies geographically, although IA accounts for
78 more than half of all reported cases.[22] Combined with epidemiological data, subgenotyping has been
79 shown to improve outbreak characterisation and response by differentiating between sporadic and
80 outbreak cases, determining the source of infection, and tracking chains of transmission.[17,18] Several
81 genome regions are used for HAV genotyping. The most common regions in HAV query databases
82 are found in or around a 168 nt fragment of the VP1/2A junction.[22–24] However, the accuracy of
83 phylogenetic relationships inferred using these small regions has been put into question by several

84 studies that found significant variations in genetic relatedness when comparing different subgenomic
85 regions.[25–27]

86 WGS comparisons have become increasingly employed for viral surveillance and outbreak detection
87 as variation across the entire virus genome can increase the spatial and temporal phylogenetic
88 resolution of molecular epidemiology.[28] Indeed, Vaughn et al. found that subgenotyping could yield
89 misleading transmission interpretations when applied to cases with unknown epidemiological
90 associations, demonstrating that WGS was able to differentiate between sporadic and outbreak cases
91 with identical VP1/2A sequences.[27] This additional resolution could be particularly useful for guiding
92 response to growing person-to-person transmission. For example, a recent analysis of a large outbreak
93 in New York (USA) found a cluster of identical VP1/2A sequences from neighbouring jurisdictions.
94 WGS revealed 3-41 single nucleotide variants (SNVs) between cases, where the case with the greatest
95 nucleotide difference had travelled to Florida, suggesting the infection was acquired outside of the
96 local outbreak.[29]

97 Despite the potential benefits, WGS of HAV in a public health context has been challenging for
98 several reasons. First, HAV grows poorly in cell culture, making isolation infeasible for routine
99 culture-based WGS. Second, despite moderate genetic diversity compared to other RNA viruses,
100 enough heterogeneity across the genome exists to make it difficult to design primers for multiplex
101 amplicon sequencing.[30,31] Recently, two primer schemes for HAV WGS have been designed targeting
102 IA and/or IB, where an effort to incorporate additional subgenotypes reduced primer efficiency.[29,32]
103 Both studies had difficulty reliably recovering whole genomes from serum samples above 30 Cτ.

104 Alternatively, hybridisation capture is a highly flexible option for culture-agnostic WGS. Unlike
105 PCR-based amplification, hybrid capture reactions can contain hundreds of thousands of
106 oligonucleotide baits (probes) without interactions and are unaffected by amplification inhibitors in
107 complex sample matrices. Furthermore, conservative tiling probe design can improve the ability to
108 capture novel sequences as well as increase sensitivity in degraded samples. These characteristics
109 make it an ideal solution for routine WGS in public health laboratories for pathogens like HAV that
110 are genetically diverse and present in a variety of complex sample matrices. The flexibility of probe

111 design can also be leveraged to incorporate additional enteric pathogens, which can streamline library
112 preparation and increase the sequence throughput of routine diagnostic or surveillance workflows.

113 Here, we present a hybridisation capture panel for undertaking WGS designed to be highly sensitive
114 across all HAV human subgenotypes. We demonstrate the sensitivity of the capture panel to recover
115 WGS from food (berry) and human serum samples spiked with serially diluted HAV RNA ranging
116 from 27·9 copies/µL to below the limit of detection (LOD) (<0·2 copies/µL), corresponding to cycle
117 threshold (Cτ) values 33·8 to >42·0. We also use the panel to recover complete and near complete
118 (93·3%) genomes from clinical serum samples of infected patients, with and without subgenotyping
119 data available. Finally, we use *in silico* prediction to extrapolate the efficiency of the panel to capture
120 subgenotypes not evaluated in this study.

121 Materials and Methods

122 Pan-HAV probe design

123 Complete genomes of human pathogenic HAV were downloaded from GenBank on 25 Jan 2021 (N =
124 164). Genomes were manually selected for temporal and geographic representation of each
125 subgenotype (N = 44). Probes were designed by IDT as previously described using a 120-mer 1X
126 tiling approach.[33]

127 Optimisation experiment

128 Sample preparation
129

130 Food samples were prepared with frozen mixed berries according to International Organization for
131 Standardization report ISO 15216-1:2017 with two modifications.[34] First, RNA extraction was
132 performed on a QIASymphony with a custom protocol using the DSP Virus/Pathogen Kit with an
133 input volume of 800µL and elution volume of 60µL. The second modification was the use of an
134 Applied Biosystems 7500 Fast DX system for detection and quantification. “Low”, “mid”, and “high”
135 virus-spiked berry specimen were prepared based on ISO 15216-2:-2019 validation studies.[35] For
136 serum samples, ultraconcentrated spike-in material was prepared by resuspending one LENTICULE

137 disc (RM000HAV) in 222µL PBS, followed by 10^{-1} serial dilutions in human sera. RNA extractions
138 were again performed using a QiaSymphony using the DSP Virus/Pathogen Kit with an input volume
139 of 200µL and elution volume of 60 µL. Quantification of target copies per microliter of sample RNA
140 was performed as described in ISO 15216-1:2017. HAV LENTICULE discs were used for both the
141 quantification standards and as the spike-in material.

142 The volume of RNA extracts was equally divided to create four experimental variables: pre-
143 enrichment, pre-enrichment with concentration, post-enrichment, and post-enrichment with
144 concentration. A subset of samples were concentrated two-fold using the Zymo Clean and
145 Concentrate -5 kit including on-column DNAse treatment with a final elution volume of 10µL.

146 **Constructing sequencing libraries**
147

148 Sequencing libraries were prepared by Illumina RNA Prep with Enrichment with 8.5µL RNA input,
149 according to the manufacturer's protocol. Total RNA for all samples was below the recommended 10
150 ng. Briefly, RNA is synthesized to cDNA, followed by tagmentation library preparation. The indexed
151 library is then hybridized with the custom probe panel, captured, and, finally, PCR amplified. Pre-
152 enrichment controls were removed from the workflow after purification of the indexed library.
153 Enriched sample libraries were normalised to 200ng and pooled for three-plex hybridisation overnight
154 (16 hours) using the previously described pan-HAV probes (100uM). Final sequencing libraries were
155 normalised to 20nM prior to pooling for sequencing on an Illumina NextSeq2000 using a P1 2 x 150
156 bp reagent kit.

157 **Analysis of pre- and post-enrichment libraries**
158

159 Reads were adapter trimmed using cutadapt (v.2.6) with parameters -a CTGTCTCTTATACACATCT
160 -A CTGTCTCTTATACACATCT for paired-end trimming of universal dual indices.[36] Taxonomic
161 classification of trimmed reads was performed with Kraken 2 (v.2.1.2) using the viral RefSeq
162 database indexed on 05 May 2021 (available from https://benlangmead.github.io/aws-indexes/k2).[37]
163 Trimmed reads were mapped to the spike-in material reference sequence (GenBank accession
164 M59809.1) using minimap2 (v.2.18-r1015) with parameters -ax sr.[38,39] PCR duplicates were removed

165 from the mapped reads using samtools *markdup* (v.1.17).[40] Mapping and quality metrics were
166 obtained from samtools *samstats* and qualimap *bamqc* (v.2.2.2a), and collated with multiQC
167 (v.1.14).[41,42]

168 Application to clinical specimen
169

170 Patient samples

171 Archived serum samples, collected in Victoria, AUS between February and May 2023, from two
172 infected patients were referred to the Microbiological Diagnostic Unit Public Health Laboratory for
173 HAV WGS. A single patient had two samples taken one month apart (AUSMDU00085964,
174 AUSMDU00085966). AUSMDU00085964 was referred with subgenotype documented as IIIA, but
175 no subgenotype data was provided for the second referral (AUSMDU00085965). Ethical approval
176 was granted from the University of Melbourne Human Research Ethics Committee (HREC no. 2021-
177 22158-20072-2).

178 Sample preparation and sequencing
179
180 RNA extraction and quantification were performed as described earlier for serum samples. Prior to
181 library preparation, RNA extracts were vacuum concentrated 3·7-fold for 26 minutes at ambient
182 temperature. Library preparation was performed as described earlier in duplicate with ~7.5µL input
183 RNA. Total RNA was below the limit of quantification by Qubit HS RNA assay. Libraries were
184 normalised to 4nM prior to pooling for sequencing on an Illumina iSeq 100 using an i1 2 x 150bp PE
185 reagent v2 kit.

186 Analysis of clinical sequences

187 Reads were quality controlled as described earlier. To determine the subgenotype of
188 AUSMDU00085965, deduplicated reads were filtered to discard those that were taxonomically
189 unclassified. Filtered reads were *de novo* assembled using SPAdes (v.3.15.5) with the rnaviral
190 parameter enabled.[43] The longest contig was queried against a local nucleotide (nt) database of HAV
191 sequences from GenBank and HAVNet using BLAST (v.2.14.0) with an e-value threshold of 10^{-4}.[44,45]
192 The highest scoring matches belonged to subgenotype IIIA. Deduplicated reads for all samples were

193 then mapped to the most recently available complete IIIA genome (MN062167.1) collected in 2019
194 from California, USA.[17] Variants were called using samtools *mpileup* and iVar *consensus* (v.1.0) with
195 parameters -m 10 -q 30 -t 0.9.[46] Mapping and quality statistics were obtained as before. Duplicate
196 sample reads were concatenated to create high quality consensus genomes and were deposited to
197 GenBank.

198 Two sets of context sequences were selected for phylogenetic analysis, which were obtained by
199 querying the individual whole genomes (*WG-query*) and the extracted VP1/2A region (*VP1/2A-query*)
200 sequences against the local nt HAV database to retrieve the 50 highest scoring matches for each
201 patient. These highest matching sequences to each patient were combined into a single WG-query or
202 VP1/2A-query multifasta file and duplicate sequences were removed. Three alignments were created
203 using MAFFT (v.7.520): WGS consensus sequences with WG-query sequences, VP1/2A consensus
204 sequences with VP1/2A-query sequences, and WGS consensus sequences with VP1/2A-query
205 sequences.[47] Maximum-likelihood phylogenetic tree construction was performed with IQtree
206 (v.2.2.2.3) using the best-fit model and 1000 bootstrap replicates.[48]

207 *In silico* probe efficiency assessment

208 Complete and near complete (>95%; >7051nts) genomes of hepatitis A virus from human hosts were
209 downloaded from GenBank on 24 April 2023 (N=239). Sequences used to create the probe panel were
210 removed, with the exception of the single available IIB genome (AY644670.1), the remaining
211 sequences were randomly subsampled to account for oversampling from outbreak studies by selecting
212 no more than three sequences per country and year for each subgenotype using augur *filter grep*
213 (v.21.1.0) (N=86) plus the spike-in control sequence.[49] ProbeTools (v.0.1.9) modules *capture* and
214 *stats* were used to align the pan-HAV probe sequences against the selected target sequences to assess
215 probe coverage and depth.[50] ProbeTools *getlowcov* was used to find low coverage regions, defined as
216 120-kmers with zero probe coverage at more than 40 consecutive sites. A Mann-Whitney U test was
217 performed to evaluate whether the low probe coverage regions identified in the control sequence had
218 significantly lower read depth compared to the rest of the genome in enriched samples using scipy
219 (v.1.18.1).[51]

220 **Data Availability**

221 Complete genome sequences from AUSMDU00085964 and AUSMDU00085965 were deposited in
222 GenBank (Accessions: OR261026, OR261027). Additional data and supplemental materials are
223 available from https://doi.org/10.26188/23648025.

224 Results

225 Pan-HAV probe design

226 A total of 425, 120-mer biotinylated oligonucleotide probes were selected to achieve a minimum 1X
227 tiling across the representative target genomes (Supplemental Tables 1-2).

228 Optimisation experiment

229 Frozen berries and serum spiked with "high", "mid", and "low" viral genome titres of HAV RNA
230 corresponded to Cτs 35·1, 40·1, and Undetermined (<LOD) for berries, and 33·6, 36·6, and 41·0 for
231 serum, respectively (Table 1). In berries, between 0—5 unique on-target reads were detected in the
232 unenriched samples, yielding no genome coverage with a minimum read depth of 10x. In serum,
233 between 69—346 unique on-target reads were detected in unenriched samples, yielding genome
234 coverage ranging from 0—1.9% with a minimum read depth of 10.

235 An increase in read depth and genome coverage was observed in all enriched samples. After
236 enrichment of berry libraries, mean-fold enrichment increased the read depth by 315·2x in high-titre
237 samples with concentration treatment and 21·4x without; by 3·5x in mid-titre samples with
238 concentration and 7·3x without; and by 3·6x in low-titre samples with concentration and 1·3x without
239 (Figure 1A). Up to 74%, 2·3%, and <0·1% of the genome was recovered from high-, mid-, and low
240 titres at a minimum read depth of 10, respectively. Target reads were present at ≥ 1 depth up to
241 99·5%, 81·8%, and 11·7% of the genome positions in high-, mid-, and low-titre berry samples,
242 respectively (Figure 1B). After enrichment of serum libraries, mean-fold enrichment increased the
243 read depth by 368·9x in high-titre samples with concentration and 662·3x without; by 109·5x in mid-
244 titre samples with concentration and 23·0x without; and by 3·1x in low-titre-samples with

concentration and 3·0x without (Figure 1A). Up to 100%, 80·8%, and 10·3% of the genome was recovered from high-, mid-, and low titres using a minimum read depth of 10, respectively. Target reads were present at ≥ 1 depth up to 100%, 100%, and 91·4% of positions in high-, mid-, and low-titre serum samples, respectively (Figure 1B).

Patient samples

One patient sampled twice over a one-month period had a Cτ value of 35·8 for the first (AUSMDU00085964) and 42·7 for the second (AUSMDU00085966). AUSMDU00085965 had a Cτ value of 34·7 (Table 1). Genotyping was available for AUSMDU00085964, thus presumed for AUSMDU00085966, reported as IIIA. The genotype for AUSMDU00085965 was determined to be IIIA using the longest *de novo* assembled contig (3482nt) queried against the local HAV database. Enriched libraries of AUSMDU00085964, AUSMDU00085966, and AUSMDU00085965 were comprised of up to 12·9%, 17·2%, and 0·9% HAV classified reads (Figure 2A), yielding a mean depth up to 2455·3X, 24·3X, and 3446·4X unique target reads (Table 1), respectively. The complete genome (100%) was recovered for AUSMDU00085964 and both AUSMDU00085965 replicates, while 89·5% and 93·3% were recovered for AUSMDU00085966 replicates using a minimum read depth of 10. Target reads were present at ≥ 99·9% of genome positions in all samples (Figure 2B).

Comparing the nucleotide differences in the WG and VP1/2A region (503 nt) between the complete sequences of both patients, there are 80 (0·011 substitutions/site) and 6 SNVs (0·012 substitutions/site), respectively. All 51 of the context sequences obtained by querying the WG sequences were >7kb, while 71·2% of the 52 context sequences obtained by querying the VP1/2A sequences were <1kb. Phylogenetic analysis of the WG sequences with WG-query context samples places the patient samples in the most recent clade descending from a clade of strains collected from India in the early 2000s (Figure 3A). Patient samples cluster together in the tree of VP1/2A sequences with VP1/2A-query context samples amongst contemporary sequences, primarily collected in European countries (Figure3B). In the tree comparing WG sequences with VP1/2A-query context samples, patient samples are similarly found in a contemporary European clade but are distributed further apart (Figure 3C).

272 ### *In silico* prediction

273 A table of accession numbers and metadata for sequence data used for *in silico* prediction are
274 available in Supplemental Table 3. On average, 98·6% of the queried target positions were covered by
275 one or more probes (median 99·2, IQR 98·8—100). Subgenotypes I—IIIA and IIB are covered by one
276 or more probes at >99·1% positions, while IB and IIIB have 97·7 and 92·8% coverage, respectively
277 (Figure 4). Amongst strains with more than one target sequence, an average of 1·1 low probe
278 coverage regions were identified. The single IIB and IIIB strains had zero and 5 low probe coverage
279 regions, respectively. Three low probe coverage sites were predicted in the control sequence between
280 positions 1690—1809, 2633—2752, and 6730—6849, none of which had significantly lower read
281 depth compared to the rest of the genome.

282 ## Discussion

283 We present a hybrid capture oligo panel for WGS of all six human pathogenic HAV subgenotypes
284 from serum and food samples, demonstrating a highly sensitive and generalized HAV WGS workflow
285 for outbreak surveillance. This approach is a more cost-effective alternative to HAV WGS by
286 metagenomic sequencing, can be used with degraded or challenging sample matrices, and can be
287 employed without *a priori* genotyping.

288 ### Optimisation experiment

289 Enrichment of the complete HAV genome from berry specimen was observed as low as 83 input
290 copies, and near complete (78·3%) from 4 input copies, providing the most cost-effect approach of its
291 kind to date. Previously, deep metagenomic sequencing had recovered a near complete HAV from
292 frozen raspberries, generating 5M reads and yielding a mean read depth of 60.[52] Lower HAV read
293 depth observed in berry specimen compared to serum may be attributed to lower genome input copies
294 and an over fragmented library. To address the former, future optimisation should increase the input
295 volume into the concentration workflow to both remove residual inhibitors and increase input viral
296 RNA. Regarding the latter, the vigorous purification process necessary for removing inhibitors from
297 food samples resulted in nearly all sequencing libraries having average read lengths <120

298 (Supplemental Figure 1). Previous work has reported reduced hybridisation efficiency when the mean
299 insert size of the library was shorter than probe length.[53] Another possibility is residual inhibitors
300 present in the berry RNA extracts could negatively affect cDNA synthesis and tagmentation prior to
301 hybridisation, which could explain why mean-fold enrichment of cleaned and concentrated berry
302 samples was significantly greater than the untreated berry samples at high-titre. Whereas the inverse
303 was observed in lower-titres, and serum, suggesting sample loss in the spin-column outweighs the
304 benefits of purification (Figure 1A). It is recommended that future optimisation includes concentration
305 from higher input volumes to increase viral RNA input, and further determination of titres suitable for
306 column-based purification for food or other complex matrices. Additional troubleshooting is needed
307 to minimize over fragmentation during library tagmentation.

308 In serum samples, the complete genome was recovered from as few as 65 input copies, and near-
309 complete (91·4%) from 4 input copies. These results are similarly sensitive compared to amplicon-
310 based sequencing for HAV WGS from serum but offers a simplified solution for routine sequencing
311 in public health laboratories. For instance, the IA scheme presented by Lee et al. yielded a maximum
312 of 94·7% genome coverage in specimen with Cτs ranging from 27·6 to 33·3. Cleary et al. designed a
313 scheme for IA-B where nearly half of the samples with Cτ values between 16-40 yielded >=95·0%
314 genome coverage using a combination of multiplex amplicon and Sanger sequencing.[29] Previous
315 iterations of this scheme included IIA resulting in lower amplification efficiency due to the degree of
316 genetic variation between the genotypes. Unlike primer schemes for amplicon-based sequencing,
317 hybrid capture allows for the enrichment of all six subgenotypes using a pan-HAV probe panel.
318 Moreover, hybrid capture can tolerate thousands of probes, meaning additional targets can be added to
319 the current panel to facilitate robust pathogen surveillance.

320 Patient samples

321 After observing potential target loss from column-based concentration in the optimisation experiment,
322 vacuum concentration was implemented with the clinical specimen as well as using a greater RNA
323 extract input volume (30µL) prior to enrichment. As a result, mean read depth and genome coverage
324 from patient samples increased compared to contrived samples with similar Cτ values. At a minimum

325 read depth of ten, complete genomes were recovered from AUSMDU00085964 and
326 AUSMDU00085965, and near complete from AUSMDU00085966 (93·3%). However, 99·9% of the
327 positions in AUSMDU00085966 were covered by one or more reads, suggesting the possibility of
328 recovering a complete genome with greater input volume or sequencing depth. Indeed, combining
329 reads from technical duplicates yields 97·7% coverage and 46·7 mean read depth. Regardless, such
330 recovery at this Cτ is remarkable given the difficulty of detecting subgenotype IIIA. Recently, a
331 reverse transcription-PCR (rRT-PCR) HAV subgenotyping assay found the LOD for IIIA to be 10
332 RNA copies per reaction, equivalent to 34·2 and 33·9 Cτ for singleplex and triplex reactions,
333 respectively.[54] Whereas a subgenotype-specific optimised reverse transcription quantitative real-time
334 PCR (RT-qPCR) assay found the limit of detection for IIIA to be 5000 copies per reaction, equivalent
335 to Cτ near 41·3.[55] This may explain why subgenotyping data was not available for
336 AUSMDU00085965, or typing had not been requested.

337 It has been demonstrated that the whole genome can provide greater geographic and temporal
338 resolution for virus phylogenetics, generally.[28] For HAV, specifically, WGS has been shown to
339 increase phylogenetic resolution and provide additional context in the absence of epidemiological
340 data.[27,29] Even though substitutions per site were slightly higher in the VP1/2A region compared to the
341 whole genome sequence (0·012 versus 0·011), phylogenetic analysis comparing whole genomes to
342 the top VP1/2A hits provided additional resolution due to the sequences in the alignment including
343 additional regions around the VP1/2A junction (Figure 3B-C). The previous difficulty of WGS HAV
344 is reflected in the sequence database where few complete genomes are available, thus resulting in high
345 similarity to sequences more than 15 years old (Figure 3A). Until HAV genomes are better
346 represented in these databases, it is recommended that subgenomic regions are used to select closely
347 related sequences for phylogenetic analysis of the whole genome.

348 *In silico* prediction

349 We observed high capture efficiency in two of the six subgenotypes *in vitro*. Similarly, *in silico*
350 prediction shows broad specificity across all human subgenotypes with one or more probes covered at
351 >92% genome positions (Figure 2). Analysis of low probe coverage found an average of 1·1 120-

352 kmer regions, representing less than 2% of the HAV genome. While low probe coverage regions were
353 identified *in silico*, variables such as random fragmentation and hybridisation temperature can
354 improve efficiency *in vitro*. In fact, we did not observe significantly lower read depth in regions of the
355 enriched contrived samples where low probe coverage was predicted in the control sequence.

356 Limitations

357 This study serves as a proof-of-concept for the feasibility of a highly sensitive culture-agnostic
358 method for HAV WGS, however, additional work is needed to further optimise and validate the
359 approach for implementation in public health. For example, more complex sample types of public
360 health relevance remain untested, such as stool, wastewater, or shellfish. Promising results have been
361 found from ultra-deep sequencing of wastewater enriched with a broad viral discovery probe panel,
362 yielding three partial HAV genomes (1482—6729nts).[56,57] Cost may also be a limiting factor for
363 wider implementation. In the present study, the price per sample was $157AUD, approximately three-
364 times the cost of highly multiplexed amplicon sequencing.[58] However, increased multiplexing as well
365 as additional optimisation to increase target read proportions could lower per-sample costs.
366 Furthermore, we anticipate the value of this approach to be the greatest for supporting food-borne
367 surveillance, where PCR-based amplification is limited by low, degraded viral input and the presence
368 of inhibitors.[59,60] Finally, overnight hybridisation was used to accommodate hands-on staffing
369 schedules. Implementation of liquid handing automation could bring the turnaround time from
370 extraction to sequencer to under 24 hours.

371 # Conclusion

372 This study presents, for the first time, a solitary approach for achieving highly sensitive WGS of HAV
373 from clinical and food samples, while proving relatively cost-effective. WGS of HAV can support
374 outbreak investigations by offering enhanced phylogenetic resolution and simultaneously facilitate the
375 refinement and optimisation of molecular diagnostic assays through the deposition of genome data in
376 public databases. In conclusion, flexible oligo design and hybridisation-based enrichment make this a
377 promising workflow for streamlining the surveillance of enteric viruses, thereby furthering the
378 genomic capacities of public health laboratories.

379 # Acknowledgements

380 This work was supported by the National Health and Medical Research Council, Australia Partnership
381 Grant (GNT1149991) and the Australian Government Medical Research Future Fund, Genomics
382 Health Futures Mission Flagships – Pathogen Genomics Grant (FSPGN000045). SEZ is supported by
383 the Melbourne Research Scholarship from the University of Melbourne.

384 We are grateful to Dr Andrew Gador-Whyte and the Department of Microbiology at Austin Health for
385 their assistance. We thank the Victorian Infectious Disease Reference Laboratory for their valuable
386 guidance in the conceptualization of this project. We would also like to show our gratitude to Dr.
387 Ketan Patel for insightful comments on the manuscript.

388

389

Tables and Figures

					Unenriched		Enriched	
Sample	**Matrix**	**Cτ**	**Copy no. (copies/µL)**	**Concentrated**	**Mean RD**	**Coverage**	**Mean RD**	**Coverage**
BerryLow	Berry	Undet.	Undet.	N	0.0	0.0	0.2	0.0
				Y†	0.0	0.0	0.2	0.0
BerryMid		40.2	0.45	N	0.0	0.0	2.6	1.1
				Y†	0.0	0.0	2.7	2.3
BerryHigh		35.1	11.2	N	0.0	0.0	12.0	60.5
				Y†	0.1	0.0	14.8	74.1
SerumLow	Serum	41.0	0.28	N	3.1	1.9	1.2	1.0
				Y†	2.5	0.9	12.5	10.3
SerumMid		36.6	4.4	N	2.1	0.0	19.2	77.7
				Y†	2.0	0.0	28.0	80.8
SerumHigh		33.6	27.9	N	1.2	0.0	324.7	100.0
				Y†	2.3	0.6	160.7	99.6
AUSMDU00085964-1	Serum	35.8	7.1	Y‡	-	-	2455.3	100.0
AUSMDU00085964-2				Y‡	*	*	*	*
AUSMDU00085966-1		42.7	0.09	Y‡	-	-	24.3	93.3
AUSMDU00085966-2				Y‡	-	-	23.0	89.5
AUSMDU00085965-1		34.7	14.07	Y‡	-	-	3446.4	100.0
AUSMDU00085965-2				Y‡	-	-	2387.9	100.0

Table 1. Summary of sample variables and genome coverage results. Undetermined (Undet.) Cτ values occur when the target is not present or is present below the limit of detection of the assay used. Copy number is the genome copies/µL quantified from the RNA extract as determined by the ISO 15216-1 HAV quantification real-time RT-PCR protocol. Mean read depth (RD) is calculated as the average number of deduplicated reads covering one position in the

396 genome. Coverage reported here is the percentage of genome positions with a depth of 10 or more reads. A summary of positive- and negative controls can be

397 found in Supplemental Table 4.

398 [†]Zymo Clean and Concentrator -5 kit.

399 [‡]Vacuum concentration.

400 [*]No indices belonging to AUSMDU00085964-2 were identified in the sequenced reads, suggesting an error occurred during library preparation.

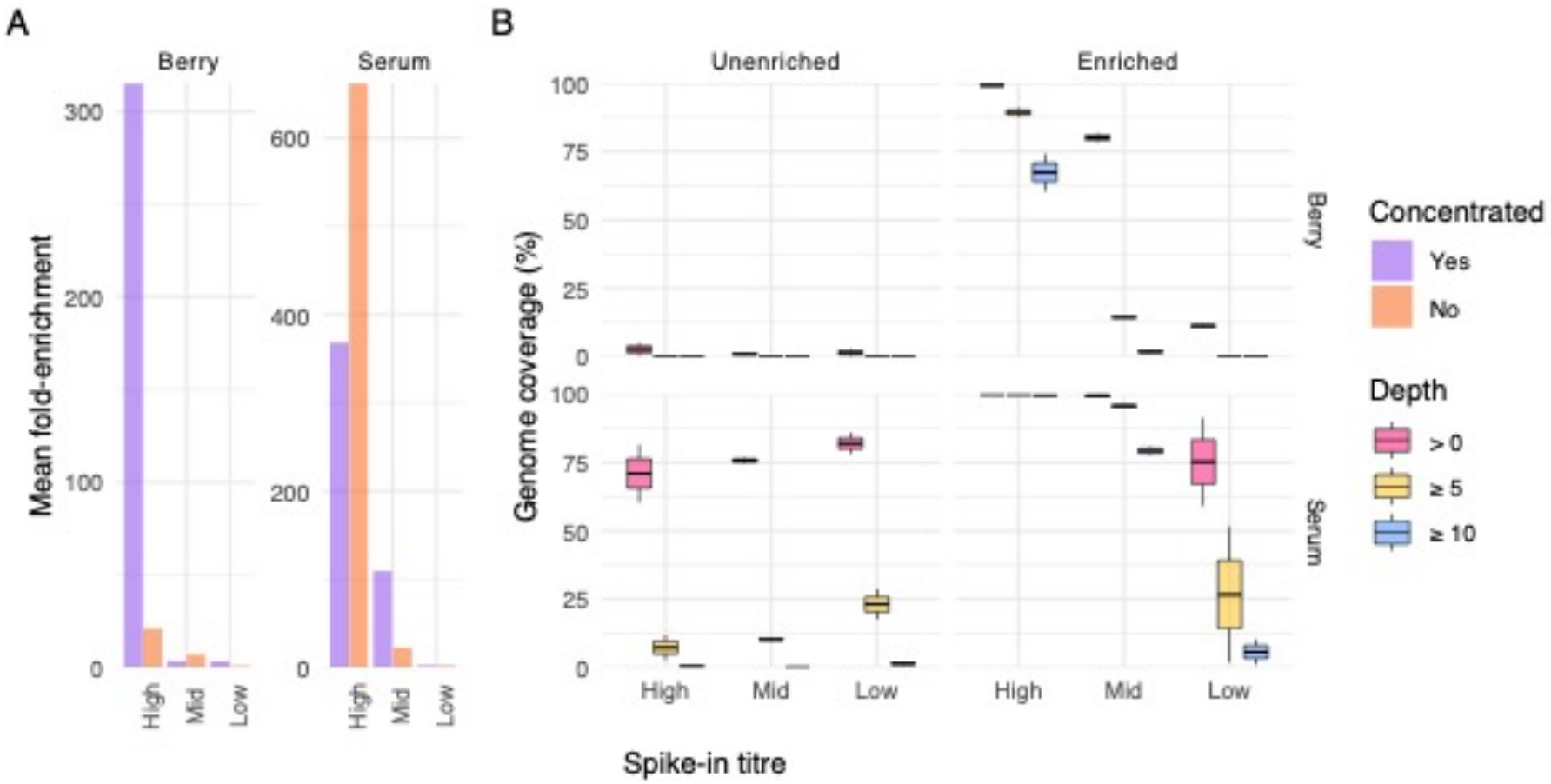


Figure 1. A) Mean-fold enrichment between paired pre- and post-enrichment samples. To avoid infinite values or overestimation using a small pseudocount, a depth of 1 was assigned to positions in pre-enriched samples where the corresponding position in the post-enriched sample was greater than 0 under the biological assumption that one or more reads must have been present for enrichment to have occurred. B) Genome coverage by spike-in titre levels at read depth thresholds of >0, ≥5, and ≥10.

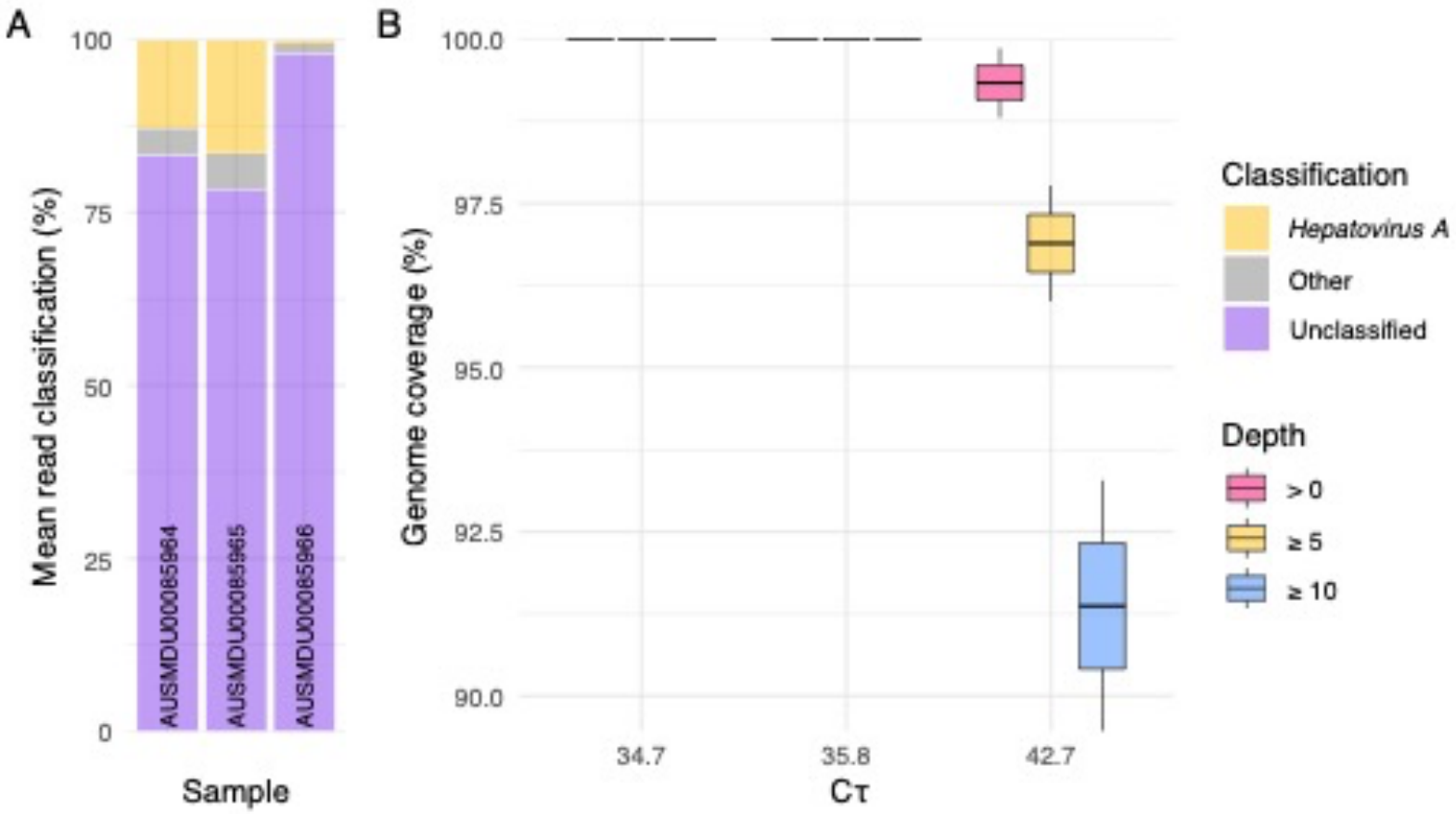


Figure 2. HAV read statistics of enriched clinical specimen. A) Taxonomic classification of sample reads averaged over technical duplicates belonging to *Hepatovirus A*, non-HAV viral taxonomies, and unclassified. B) Box and whisker plot of genome coverage by sample Cτ at read depth thresholds of >0, ≥5, and ≥10.

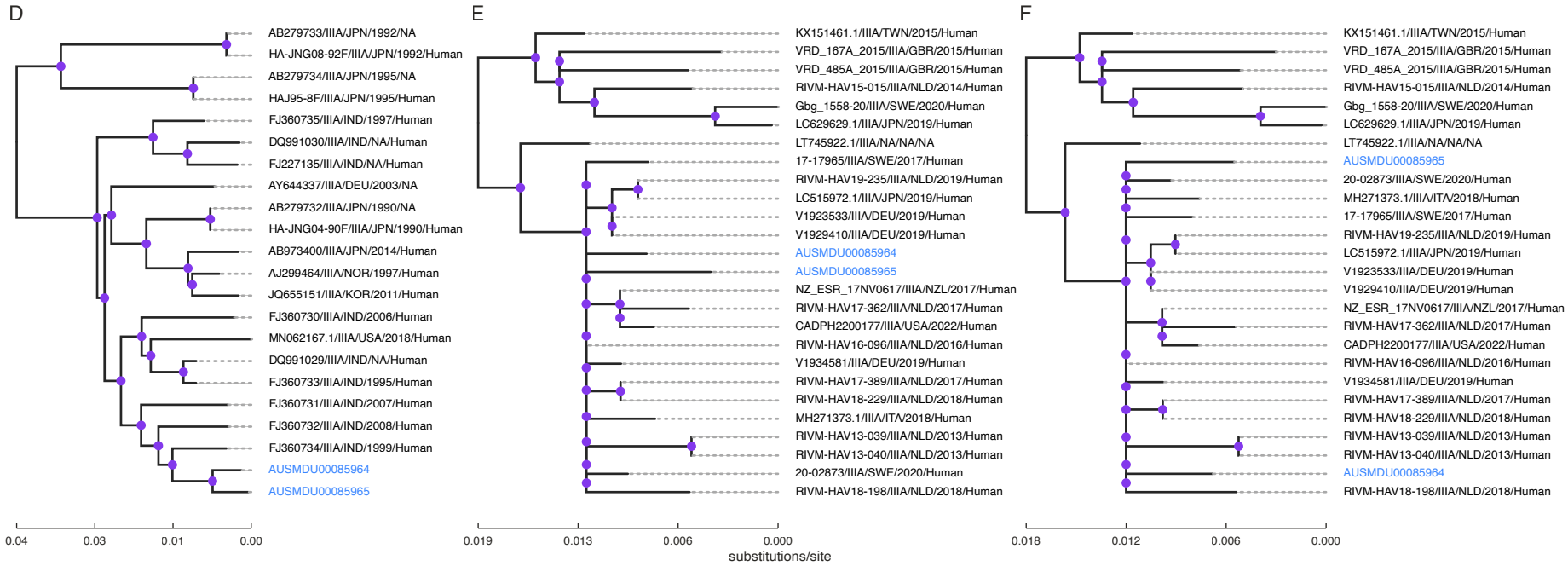


Figure 3. Maximum likelihood trees comparing WGS- and VP1/2A-query results. Ancestral nodes were traversed to clearly visualise patient samples (blue). Nodes with high bootstrap support (≥80%) are coloured in purple. Trees were visualised using toytree (v.2.0.5).[62] A) WGS consensus sequences in the context of WG-query sequences (best-fit model = TIM2+F+I+R3). B) VP1/2A consensus sequences in the context of VP1/2A-query sequences (best-fit model = TIM2+F+I+G4). C) WGS consensus sequences in the context of VP1/2A-query sequences (best-fit model = TIM2+F+I+G4).

417

418

419 **Figure 4.** *In silico* prediction of probe efficiency, averaged by subgenotype and stratified by probe

420 coverage depth. Percentage annotation indicates the mean* predicted percent genome coverage at

421 respective probe depths.

422 *Subgenotypes IIB and IIIB each have a single genome available for comparison (AB279735.1 and

423 AY644670.1, respectively).

Supplemental tables and figures

Additional data and supplemental materials are available from https://doi.org/10.26188/23648025.

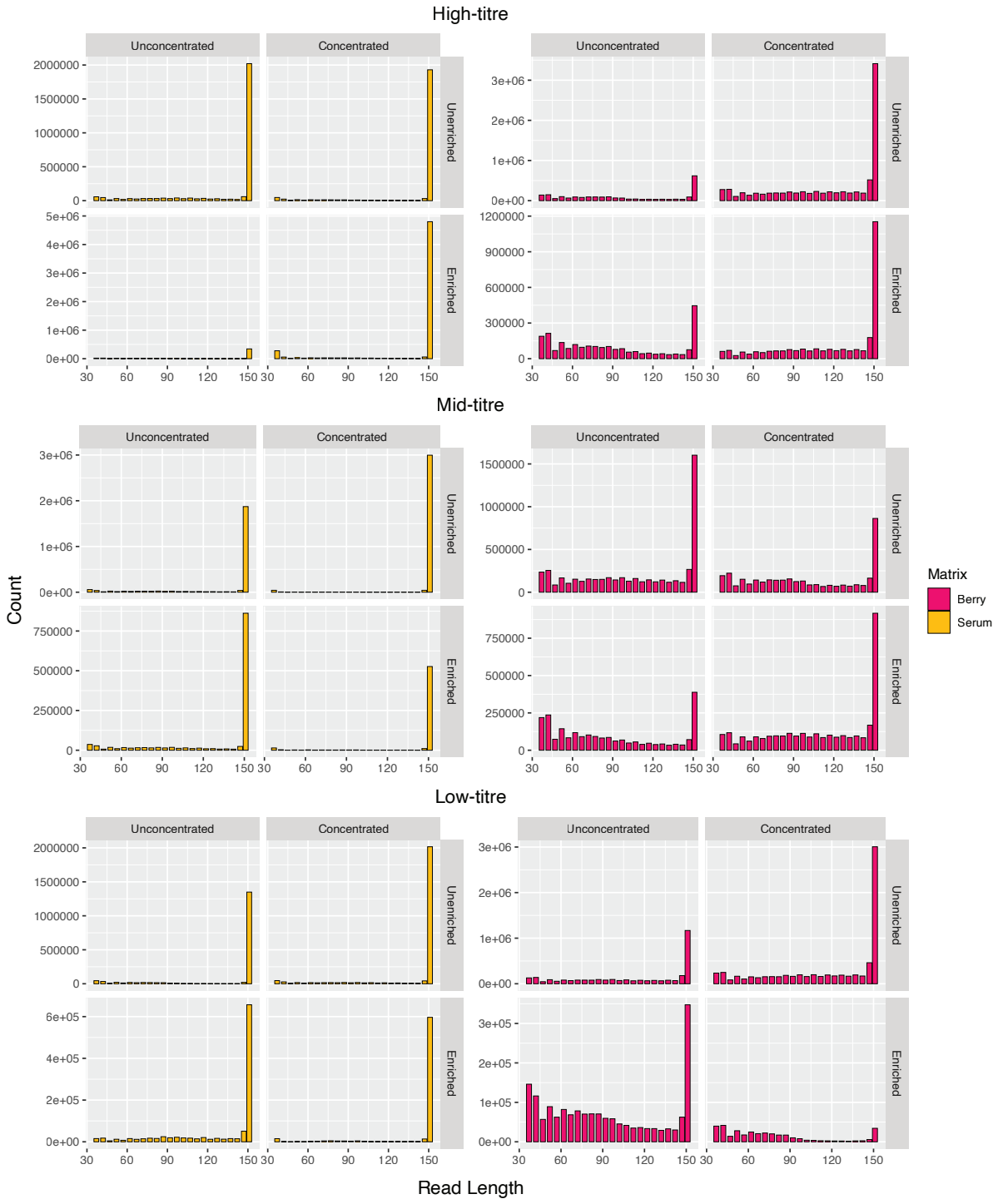
S Table 1. Resultant pan-HAV oligos designed for capture of all six human subgenotypes of HAV.

S Table 2. GenBank accession numbers for sequences used to design pan-HAV oligos and associated subgenotype and collection metadata. Date of sample collection, where available, is formatted as DD-MM-YYYY.

S Table 3. GenBank accession numbers for sequences used for *in silico* prediction of the pan-HAV oligo design efficiency, including associated subgenotype and collection metadata. Date of sample collection, where available, is formatted as DD-MM-YYYY.

S Table 4. Summary of positive and negative controls from the optimisation experiment and resultant genome coverage results. Undetermined (Undet.) Cτ values occur when the target is not present or is present below the limit of detection of the assay used. Copy number is the genome copies/µL quantified from the RNA extract as determined by the ISO 15216-1 HAV quantification real-time RT-PCR protocol. Mean RD is calculated as the average number of deduplicated reads covering one position in the genome. Coverage reported here is the percentage of genome positions with a depth of 10 or more reads.

- Indicates no reads from the corresponding index were present in the sequence data.



S Figure 1. Read length distribution of spike-in sample libraries measured before quality control.

444 # References

445 1. Koff, R. S. Hepatitis A. *Lancet* **351**, 1643–1649 (1998).

446 2. Shapiro, C. N. & Margolis, H. S. Worldwide epidemiology of hepatitis A virus infection. *J.*
447 *Hepatol.* **18 Suppl 2**, S11-4 (1993).

448 3. Centers for Disease Control and Prevention. Prevention of hepatitis A through active or passive
449 immunization: recommendations of the Advisory Commitee on Immunization Practices (ACIP).
450 *MMWR Recomm. Rep.* (1999).

451 4. Advisory Committee on Immunization Practices (ACIP), Fiore, A. E., Wasley, A. & Bell, B. P.
452 Prevention of hepatitis A through active or passive immunization: recommendations of the
453 Advisory Committee on Immunization Practices (ACIP). *MMWR Recomm. Rep.* **55**, 1–23
454 (2006).

455 5. Klevens, R. M., Denniston, M. M., Jiles-Chapman, R. B. & Murphy, T. V. Decreasing immunity
456 to hepatitis A virus infection among US adults: Findings from the National Health and Nutrition
457 Examination Survey (NHANES), 1999-2012. *Vaccine* **33**, 6192–6198 (2015).

458 6. Bell, B. P. *et al.* The diverse patterns of hepatitis A epidemiology in the United States-
459 implications for vaccination strategies. *J. Infect. Dis.* **178**, 1579–1584 (1998).

460 7. Staes, C. J. *et al.* Sources of infection among persons with acute hepatitis A and no identified risk
461 factors during a sustained community-wide outbreak. *Pediatrics* **106**, E54 (2000).

462 8. Mutsch, M., Spicher, V. M., Gut, C. & Steffen, R. Hepatitis A virus infections in travelers, 1988-
463 2004. *Clin. Infect. Dis.* **42**, 490–497 (2006).

464 9. Klevens, R. M. *et al.* The evolving epidemiology of hepatitis a in the United States: incidence
465 and molecular epidemiology from population-based surveillance, 2005-2007. *Arch. Intern. Med.*
466 **170**, 1811–1818 (2010).

467 10. Beebeejaun, K. *et al.* Outbreak of hepatitis A associated with men who have sex with men
468 (MSM), England, July 2016 to January 2017. *Euro Surveill.* **22**, (2017).

469 11. Freidl, G. S. *et al.* Hepatitis A outbreak among men who have sex with men (MSM)
470 predominantly linked with the EuroPride, the Netherlands, July 2016 to February 2017. *Euro*
471 *Surveill.* **22**, (2017).
472 12. Latash, J. *et al.* Notes from the Field: Increase in Reported Hepatitis A Infections Among Men
473 Who Have Sex with Men - New York City, January-August 2017. *MMWR Morb. Mortal. Wkly.*
474 *Rep.* **66**, 999–1000 (2017).
475 13. Foster, M. *et al.* Hepatitis A Virus Outbreaks Associated with Drug Use and Homelessness -
476 California, Kentucky, Michigan, and Utah, 2017. *MMWR Morb. Mortal. Wkly. Rep.* **67**, 1208–
477 1210 (2018).
478 14. Friesema, I. H. *et al.* Spillover of a hepatitis A outbreak among men who have sex with men
479 (MSM) to the general population, the Netherlands, 2017. *Euro Surveill.* **23**, (2018).
480 15. Ndumbi, P. *et al.* Hepatitis A outbreak disproportionately affecting men who have sex with men
481 (MSM) in the European Union and European Economic Area, June 2016 to May 2017. *Euro*
482 *Surveill.* **23**, (2018).
483 16. Foster, M. A. *et al.* Increase in Hepatitis A Virus Infections - United States, 2013-2018. *MMWR*
484 *Morb. Mortal. Wkly. Rep.* **68**, 413–415 (2019).
485 17. Probert, W. S., Gonzalez, C., Espinosa, A. & Hacker, J. K. Molecular Genotyping of Hepatitis A
486 Virus, California, USA, 2017-2018. *Emerg. Infect. Dis.* **25**, 1594–1596 (2019).
487 18. Haddix, M. *et al.* Use of Molecular Epidemiology to Inform Response to a Hepatitis A Outbreak
488 - Los Angeles County, California, October 2018-April 2019. *MMWR Morb. Mortal. Wkly. Rep.*
489 **69**, 820–824 (2020).
490 19. Foster, M. A. *et al.* Widespread Hepatitis A Outbreaks Associated with Person-to-Person
491 Transmission - United States, 2016-2020. *MMWR Morb. Mortal. Wkly. Rep.* **71**, 1229–1234
492 (2022).
493 20. Batdorf, S. J. *et al.* Estimated Medicaid Costs Associated with Hepatitis A During an Outbreak -
494 West Virginia, 2018-2019. *MMWR Morb. Mortal. Wkly. Rep.* **70**, 269–272 (2021).

495 21. Helmick, M. J., Morrow, C. B., White, J. H. & Bordwine, P. Widespread Community
496 Transmission of Hepatitis A Virus Following an Outbreak at a Local Restaurant - Virginia,
497 September 2021-September 2022. *MMWR Morb. Mortal. Wkly. Rep.* **72**, 362–365 (2023).
498 22. Kroneman, A. *et al.* Usability of the international HAVNet hepatitis A virus database for
499 geographical annotation, backtracing and outbreak detection. *Euro Surveill.* **23**, (2018).
500 23. Robertson, B. H. *et al.* Genetic relatedness of hepatitis A virus strains recovered from different
501 geographical regions. *J. Gen. Virol.* **73 (Pt 6)**, 1365–1377 (1992).
502 24. Nainan, O. V. *et al.* Hepatitis a molecular epidemiology in the United States, 1996-1997: sources
503 of infection and implications of vaccination policy. *J. Infect. Dis.* **191**, 957–963 (2005).
504 25. Costa-Mattioli, M. *et al.* Genetic variability of hepatitis A virus. *J. Gen. Virol.* **84**, 3191–3201
505 (2003).
506 26. Costa-Mattioli, M. *et al.* Molecular evolution of hepatitis A virus: a new classification based on
507 the complete VP1 protein. *J. Virol.* **76**, 9516–9525 (2002).
508 27. Vaughan, G. *et al.* Genetic relatedness among hepatitis A virus strains associated with food-
509 borne outbreaks. *PLoS One* **8**, e74546 (2013).
510 28. Dudas, G. & Bedford, T. The ability of single genes vs full genomes to resolve time and space in
511 outbreak analysis. *BMC Evol. Biol.* **19**, 232 (2019).
512 29. Cleary, N. G., Bryant, P. W., Lamson, D. M., Newman, A. P. & St George, K. A custom
513 hepatitis A virus assay for whole-genome sequencing. *J. Virol. Methods* **312**, 114649 (2023).
514 30. Moratorio, G. *et al.* Bayesian coalescent inference of hepatitis A virus populations: evolutionary
515 rates and patterns. *J. Gen. Virol.* **88**, 3039–3042 (2007).
516 31. Vaughan, G. *et al.* Hepatitis A virus: host interactions, molecular epidemiology and evolution.
517 *Infect. Genet. Evol.* **21**, 227–243 (2014).
518 32. Lee, G.-Y. *et al.* Genotyping and Molecular Diagnosis of Hepatitis A Virus in Human Clinical
519 Samples Using Multiplex PCR-Based Next-Generation Sequencing. *Microorganisms* **10**, (2022).
520 33. Bonsall, D. *et al.* ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-
521 genome sequencing of HCV and other highly diverse pathogens. *F1000Res.* **4**, 1062 (2015).

522 34. International Organization for Standardization. *Microbiology of the Food Chain–Horizontal*
523 *Method for Determination of Hepatitis A Virus and Norovirus Using Real-time RT-PCR- Part1:*
524 *Method for Quantification*. (2017).
525 35. International Organization for Standardization. *Microbiology of the Food Chain-Horizontal*
526 *Method for Determination of Hepatitis A Virus and Norovirus Using Real-time RT-PCR- Part2:*
527 *Method for Detection*. (2019).
528 36. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
529 *EMBnet J.* **17**, 10 (2011).
530 37. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome*
531 *Biol.* **20**, 257 (2019).
532 38. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100
533 (2018).
534 39. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574
535 (2021).
536 40. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
537 41. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality
538 control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
539 42. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for
540 multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
541 43. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes DE
542 Novo Assembler. *Curr. Protoc. Bioinformatics* **70**, e102 (2020).
543 44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search
544 tool. *J. Mol. Biol.* **215**, 403–410 (1990).
545 45. HAVNet. *National Institute for Public Health and the Environment (RIVM)* www.havnet.nl.
546 46. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring
547 intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).
548 47. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
549 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

550 48. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective
551 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–
552 274 (2015).
553 49. Huddleston, J. *et al.* Augur: a bioinformatics toolkit for phylogenetic analyses of human
554 pathogens. *J. Open Source Softw.* **6**, 2906 (2021).
555 50. Kuchinski, K. S., Duan, J., Himsworth, C., Hsiao, W. & Prystajecky, N. A. ProbeTools:
556 designing hybridization probes for targeted genomic sequencing of diverse and hypervariable
557 viral taxa. *BMC Genomics* **23**, 579 (2022).
558 51. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat.*
559 *Methods* **17**, 261–272 (2020).
560 52. Chen, H., Wang, W., Wang, S. & Hu, Y. Near-Complete Genome Sequence of a Hepatitis A
561 Subgenotype IB Virus Isolated from Frozen Raspberries. *Microbiol Resour Announc* **8**, (2019).
562 53. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel
563 targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
564 54. Probert, W. S. & Hacker, J. K. New Subgenotyping and Consensus Real-Time Reverse
565 Transcription-PCR Assays for Hepatitis A Outbreak Surveillance. *J. Clin. Microbiol.* **57**, (2019).
566 55. Coudray-Meunier, C. *et al.* Hepatitis A virus subgenotyping based on RT-qPCR assays. *BMC*
567 *Microbiol.* **14**, 296 (2014).
568 56. Briese, T. *et al.* Virome Capture Sequencing Enables Sensitive Viral Diagnosis and
569 Comprehensive Virome Analysis. *MBio* **6**, e01491-15 (2015).
570 57. Schaeffer, J., Desdouits, M., Besnard, A. & Le Guyader, F. S. Looking into sewage: how far can
571 metagenomics help to detect human enteric viruses? *Front. Microbiol.* **14**, 1161674 (2023).
572 58. Gohl, D. M. *et al.* A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2.
573 *BMC Genomics* **21**, 863 (2020).
574 59. Lee, D.-Y., Cho, S.-R., Chae, S.-J., Choi, W. & Han, M.-G. Evaluation of a test method to detect
575 hepatitis A virus in salted shellfish. *J. Food Saf.* **41**, (2021).

576 60. Fitzpatrick, A. H. *et al.* Amplicon-Based High-Throughput Sequencing Method for Genotypic
577 Characterization of Norovirus in Oysters. *Appl. Environ. Microbiol.* e0216522 (2023)
578 doi:10.1128/aem.02165-22.
579 61. Eaton, D. A. R. Toytree: A minimalist tree visualization and manipulation library for Python.
580 *Methods Ecol. Evol.* **11**, 187–191 (2020).
581