

1	Norovirus whole genome sequencing by SureSelect target enrichment: a robust and
2	sensitive method
3	
4	Julianne R Brown, <sup>a,b</sup> # Sunando Roy, <sup>c</sup> Christopher Ruis, <sup>c</sup> Erika Yara Romero, <sup>c</sup> Divya Shah, <sup>a,b</sup>
5	Rachel Williams, <sup>c</sup> Judy Breuer <sup>a, c</sup>
6	
7	Microbiology, Virology and Infection Control, Great Ormond Street Hospital for Children
8	NHS Foundation Trust, London, UK <sup>a</sup> ; NIHR Biomedical Research Centre at Great Ormond
9	Street Hospital for Children NHS Foundation Trust and University College London, UK <sup>b</sup> ;
10	Division of Infection and Immunity, University College London, UK <sup>c</sup>
11	
12	Running title: Norovirus whole genome sequencing by target enrichment
13	

14 #Address correspondence to Julianne R brown, julianne.brown@nhs.net

# 15 Abstract

Norovirus full genome sequencing is challenging due to sequence heterogeneity between
genomes. Previous methods have relied on PCR amplification, which is problematic due to
primer design, and RNASeq which non-specifically sequences all RNA in a stool specimen,
including host and bacterial RNA.

Target enrichment uses a panel of custom-designed 120-mer RNA baits which are 20 21 complementary to all publicly available norovirus sequences, with multiple baits targeting 22 each position of the genome, thus overcoming the challenge of primer design. Norovirus 23 genomes are enriched from stool RNA extracts to minimise sequencing non-target RNA. 24 SureSelect target enrichment and Illumina sequencing was used to sequence full genomes 25 from 507 norovirus positive stool samples with RT-qPCR Ct values 10-43. Sequencing on an Illumina MiSeq in batches of 48 generated on average 81% on-target-reads per sample and 26 27 100% genome coverage with >12,000-fold read depth. Samples included genotypes GI.1, 28 GI.2, GI.3, GI.6, GI.7, GII.1, GII.2, GII.3, GII.4, GII.5, GII.6, GII.7, GII.13, GII.14 and GII.17. Once outliers are accounted for, we generate over 80% genome coverage for all positive samples, 29 regardless of Ct value. 30

31 164 samples were tested in parallel with conventional PCR genotyping of the capsid shell 32 domain. 164/164 samples were successfully sequenced, compared to 158/164 that were 33 amplified by PCR. Four of the samples that failed capsid PCR had low titres, suggesting 34 target enrichment is more sensitive than gel-based PCR. Two samples failed PCR due to 35 primer mismatches; target enrichment uses multiple baits targeting each position, thus 36 accommodating sequence heterogeneity between norovirus genomes.

#### 37 Introduction

Norovirus is a leading cause of outbreaks of acute gastroenteritis (1, 2) with an estimated 38 prevalence of 20% in cases of acute gastroenteritis in developed countries (3) and a high 39 40 financial burden in healthcare settings associated with ward and hospital closures (4). In 41 countries where rotavirus vaccine has been introduced, norovirus is now the leading cause 42 of medically-attended gastroenteritis in children (5, 6). Norovirus has a 7.5kb single stranded RNA genome, organised into 3 open reading frames; 43 ORF1, ORF2 and ORF3. ORF1 encodes a non-structural polyprotein which is cleaved post-44 45 translationally and includes the RNA-dependent RNA polymerase. ORF2 encodes the major 46 structural capsid protein, which is divided into shell (S) and protruding (P) domains. The P domain has two subdomains, P1 and P2. P2 is the most exposed antigenic site and contains 47 immunogenic epitopes; consequently it has the greatest sequence variation. ORF3 codes for 48

a minor capsid protein.

50

51 Comparison of viral genetic sequences allows linking of previously unrecognised 52 transmission events or exclusion of cases from an outbreak. Traditionally, norovirus genotyping has involved polymerase chain reaction (PCR) amplification and 53 54 capillary sequencing of partial regions of the polymerase and capsid sequences, followed by additional sequencing of the P2 region for outbreak investigations. This is a labour intensive 55 process requiring several rounds of PCR and sequencing, each requiring genogroup or 56 57 genotype specific primers and only yields partial genome sequences at the end. Moreover, 58 whilst the P2 domain can identify linked outbreak events with 64–73% specificity (assuming

bootstrap support >70 or <70, respectively), the full capsid sequence can identify linked</li>
outbreak events with 100% specificity (7) and thus is more informative.

Whole genome sequencing simplifies investigation of norovirus molecular epidemiology by
generating all the regions of interest in one step, thus allowing identification of the
genotype, variant type and full capsid sequence; negating the need for sequential PCR and
sequencing reactions. However, unlike bacteria, which can be isolated in pure culture,
norovirus culture is difficult (8). Moreover, as norovirus replicates within the host cell, viral
nucleic extracts are contaminated by host DNA, and if obtained from clinical specimens, by
DNA and RNA from enteric bacteria.

68 To date norovirus sequencing from clinical material has been achieved by two methods: sequencing of overlapping PCR fragments (9-12) and direct sequencing of total RNA (13-16). 69 70 The former generates pure viral template, which improves the quality of sequence, but 71 requires multiple PCR amplifications. The latter necessitates great depth of sequencing to generate the target norovirus genome. Here we describe the application of a third method, 72 73 SureSelect target enrichment (Agilent), which has been successfully used to generate full 74 pathogen genomes for hard to culture bacteria as well as DNA and RNA viruses directly from 75 clinical samples (17-19). Norovirus genomes are enriched directly from stool RNA extracts 76 using a panel of custom-designed 120-mer RNA baits which are complementary to all publicly available norovirus sequences, with multiple baits targeting each position of the 77 genome. This approach overcomes the problems of primer design in PCR and of non-target 78 79 sequencing in RNASeq.

80 Materials and Methods

81 Samples

507 norovirus positive stool samples from 382 patients in four UK healthcare centres were 82 processed for whole genome sequencing. Samples included genotypes GI.1, GI.2, GI.3, GI.6, 83 GI.7, GII.1, GII.2, GII.3, GII.4, GII.5, GII.6, GII.7, GII.13, GII.14 and GII.17, as detailed in Table 84 1. The presence of norovirus was verified in all samples using a multiplex norovirus GI and 85 86 GII-specific one-step reverse-transcription real-time PCR (RT-qPCR); the primer and probe 87 sequences and cycling conditions have been previously described (manuscript submitted to 88 J Clin Virol). For 78/507 samples provided by one of the centres, the presence of norovirus RNA was not verified in the re-extracted residual specimen; for these samples the RT-qPCR 89 Ct value corresponds to the original extract used as part of diagnostic service. The RT-qPCR 90 cycle threshold (Ct) value is used in this study as a semi-quantitative indicator of viral titre. 91

All specimens were residual diagnostic specimens obtained from patients with confirmed
norovirus infections. Specimens were submitted to the UCL Infection DNA Bank for use in
this study. All samples were supplied to the study in an anonymised form; the use of these
specimens for research was approved by the NRES Committee London – Fulham (REC
reference: 12/LO/1089). All stool samples were stored at -80°C in between diagnostic
testing and RNA extraction for full genome sequencing.

164 stool samples were genotyped using capsid PCR and Sanger sequencing in parallel to
SureSelect target enrichment whole genome sequencing. PCR primer sequences and cycling
conditions for genotyping have been described previously (manuscript submitted to *J Clin Virol*). Briefly, GI or GII-specific primers were used to amplify a 597 or 468 nt region of the
norovirus capsid shell domain, respectively; amplicons were capillary sequenced in the

forward and reverse direction. Generated sequences were submitted to the Norovirusgenotyping tool to identify the capsid genotype (20).

105 RNA extraction

106 RNA was purified from 200  $\mu$ l of a clarified 10% w/v stool suspension using the Qiagen EZ1 107 virus mini kit or Qiasymphony DSP Virus/Pathogen kit with a 90  $\mu$ l elution volume. All 108 purified RNA was stored at -80°C prior to cDNA synthesis.

109 *cDNA synthesis* 

RNA extracts were concentrated to 11 µl using a vacuum centrifuge at 65°C prior to first 110 111 strand cDNA synthesis. First strand cDNA was synthesised using random primers and SuperScript III (SS III, Life Technologies) as per manufacturer's instructions. Briefly, 1 µl of 112 113 10mM (each) dNTP mix and 1  $\mu$ l of 3  $\mu$ g/ml random primers were incubated with 11  $\mu$ l RNA 114 for five minutes at 65 °C to anneal primers to RNA template, followed by incubation on ice for 1 minute. RNA–primer templates were mixed with 4  $\mu$ l 5x first strand buffer, 1  $\mu$ l 0.1M 115 116 DTT, 1 µl RNase OUT and 1 µl SS III at 25 °C for 5 minutes followed by cDNA synthesis at 50 117 °C for 1 hour and enzyme inactivation at 70 °C for 15 minutes. Second strand cDNA was synthesised using Second Strand cDNA Synthesis kit (NEB) as per manufacturer's instruction. 118 Briefly, 20  $\mu$ l first strand cDNA was incubated with 48  $\mu$ l water, 8  $\mu$ l 10x 2<sup>nd</sup> strand buffer 119 and 4 µl 2<sup>nd</sup> strand enzyme mix at 16 °C for 2.5 hours. Double stranded cDNA was purified 120 and concentrated with Genomic DNA Clean and Concentrator (Zymo Research), as per 121 122 manufacturer's instructions, with a 30  $\mu$ l elution volume and quantified with Qubit dsDNA high sensitivity (HS) kit (Invitrogen). 123

124 SureSelect Target Enrichment: RNA baits design

Overlapping 120-mer RNA baits complementary to and spanning the length of 622 norovirus 125 126 partial or complete genomes from Genbank were designed using an in-house PERL script. Briefly, a 120 nucleotide sliding window is scanned along each reference genome at 127 intervals of 10 nucleotides. If the 120-mer is sufficiently different to other 120-mer 128 129 sequences in the baitset (as assessed by BLAT (21)), it is retained in the baitset; otherwise 130 that 120-mer is discarded. In this way, the baitset spans the diversity in all of the included 131 reference genomes. The baitset is available upon request. The reference genomes included 132 samples from polymerase genotypes GI.P1, GI.P2, GI.P3, GI.P4, GI.P6, GI.P8, GI.Pb, GI.Pc, GI.Pd, GI.Pf, GII.P1, GII.P2, GII.P3, GII.P4, GII.P5, GII.P6, GII.P7, GII.P8, GII.P11, GII.P12, 133 GII.P15, GII.P16, GII.P17, GII.P18, GII.P21, GII.P22, GII.Pc, GII.Pe, GII.Pg, GII.Pp, GIII, GIV, GV 134 135 and GVI and capsid genotypes GI.1, GI.2, GI.3, GI.4, GI.5, GI.6, GI.8, GII.2, GII.3, GII.4, GII.5, GII.6, GII.7, GII.8, GII.10, GII.11, GII.12, GII.13, GII.14, GII.15, GII.16, GII.17, GII.18, GII.21, 136 137 GII.22, GIII, GIV, GV and GVI. The GII.4 reference genomes included samples from all major 138 GII.4 strains: CHDC1970s, Bristol 1993, Camberwell 1994, US95/96, Farmington Hills 2002, Lanzhou 2002, Asia 2003, Hunter 2004, Yerseke 2006a, Den Haag 2006b, Osaka 2007, 139 140 Apeldoorn 2007, New Orleans 2009 and Sydney 2012. The custom designed norovirus bait library was uploaded to Agilent SureDesign and synthesised by Agilent Biotechnologies. 141 142

143 SureSelect Target Enrichment: Library preparation, hybridisation and enrichment

144 Norovirus cDNA samples were quantified and carrier G147 Human Genomic DNA: male

145 (Promega) was added if necessary to obtain a total of 200ng.

146 All DNA samples were mechanically sheared for 150 seconds using a Covaris E210 focused-

147 ultrasonicator (duty cycle 5%, PIP 175 and 200 cycles per burst) to yield a fragment size of

approximately 270 bp. End-repair, non-templated addition of 3' –A adapter ligation,

149	hybridisation, enrichment PCR and all post-reaction clean-up steps were performed
150	according to the SureSelect Illumina Paired-End Sequencing Library XT protocol. All
151	recommended quality steps were performed between steps.
152	Negative controls
153	All RNA extraction batches included a negative extract control, consisting of sterile Qiagen
154	Buffer ASL extracted with the Qiagen EZ1 virus mini kit alongside stool samples. All negative
155	extracts were tested by norovirus-specific real-time RT-PCR to verify the absence of
156	contaminating RNA.
157	To determine the level of contaminating norovirus RNA in the sequencing pipeline, two
158	negative extracts were processed for sequencing.
159	Illumina sequencing
160	Samples were multiplexed with 48 samples per run. Paired end sequencing was done on an
161	Illumina MiSeq sequencing platform with the 500 cycle v2 Reagent Kit. Base calling and
162	sample demultiplexing were generated as standard on the MiSeq producing paired FASTQ
163	files for each sample.
164	Sequence assembly
165	All assemblies were done in CLC genomics workbench v8, as summarised in Figure 1. All
166	reads were quality trimmed and adapter sequences removed. Trimmed reads were mapped
167	to a curated reference list consisting of all norovirus complete genome and complete gene
168	sequences in Genbank as of $14/07/2015$ (n = 688). All paired reads mapping to the reference
169	list (filtered reads) were taken forward to <i>de novo</i> assembly using workbench default
170	parameters and a minimum contig length of 200 nucleotides. Contigs generated from the de

171 novo assembly were aligned to a single Genbank reference sequence of the relevant

172 genotype to check the orientation of the contig and, where multiple contig sequences were

173 generated, the position of each contig relevant to the reference. Multiple contig sequences

were joined based on overlapping nucleotide sequences or with a manually inserted gap. All

trimmed reads (pre-filtering) were mapped to the full length contig sequence generated

176 from the *de novo* assembly to generate a final consensus sequence. Areas of low coverage

177 (<10) were assigned the ambiguity symbol N.

#### 178 Simulated mixed infection

179 To assess whether a reliable consensus sequence can be generated from a mixed infection,

the reads generated from two single infections (one GII.3 and one GII.4) were merged into a

181 single assembly pipeline. The consensus sequences generated from the single infection

182 (original) and the mixed (simulated) infection were aligned to identify the number of

183 differences between the two consensus sequences.

184 Statistical analysis

185 All statistical analysis was performed in SPSS v23 using two-tailed tests at the 5%

186 significance level.

187 The difference in % on-target-reads (% OTR), read depth and % genome coverage between

188 norovirus genotypes and in PCR Ct value between Pass/Sub-optimal/Failed samples was

189 tested by Kruskal-Wallis ANOVA, with pairwise multiple comparison of significant results

and P values adjusted for multiple comparisons.

191 The relationship between PCR Ct value and % OTR, read depth and % genome coverage was

192 assessed by Spearman's correlation.

A simple linear regression model (independent variable, PCR Ct value; dependant variable, logit transformed %genome coverage) was fitted to generate prediction intervals for % genome coverage from the PCR Ct value. % genome coverage was transformed using the formula  $tr\_genome = \frac{\%genome \ coverage \times (N-1)+0.5}{N}$  to ensure there are no proportions of 0 or 1 and then transformed again using the logit function

198 (*logit transformed %genome coverage* =  $log\left(\frac{tr\_genome}{1-tr\_genome}\right)$  where log is the natural 199 logarithm with base *e*. Outliers (highlighted in Figure A3) were excluded from regression 200 analysis.

201 Results

# 202 Overall sequencing outcomes

Since the aim was to generate full genome sequences, we defined the cut-off for sequencing success as >90% coverage of the full norovirus genome with >100-fold mean read depth to ensure a robust consensus sequence. Samples that met only one of these criterions were categorised as "sub-optimal", and those which did not meet either criteria were considered a "fail".

208 Of 507 samples across all sampled genotypes, 453 (89%) passed; i.e. had >90% genome

209 coverage and >100-fold read depth (Table 1, Figure 2, Figure A1). However in total, 93% of

samples had a genome coverage of >90% at any depth. A median of 81.22% of the total

sequencing reads generated for each sample mapped to the norovirus genome, referred to

- as the % on-target-reads (% OTR). On average, 100% of the full genome was covered (%
- genome coverage) with median read depth of 12,227-fold (Table 1).

There was no significant difference in % OTR (P = 0.127), mean read depth (P = 0.398) or %

215 genome coverage (P = 0.203) between norovirus genotypes (Figure 3 (a–c)).

A significant correlation was found between % OTR and read depth (R = 0.757, P < 0.001,

Figure A2) and between PCR Ct value and (i) % OTR (R = -0.536, P < 0.001), (ii) read depth (R

218 = −0.468, P <0.001) and (iii) % genome coverage (R = −0.223, P <0.001) (Figure 3 (d−f)). It

219 follows that there is a significant difference in PCR Ct value between samples that passed

compared to those that were sub-optimal (P <0.001) or failed (P <0.001) with median Ct

values of 22, 32 and 32, respectively (Figure 4). There is an inverse relationship between Ct

value and viral load (22); thus samples with a smaller Ct value (higher viral titre) resulted in

higher %OTR, read depth and genome coverage.

## 224 Predicted genome coverage

The estimated linear regression model is y = 7.432 - 0.059x where the dependent variable y is the logit of transformed genome coverage proportion and the independent

variable *x* is the PCR Ct value (n = 477,  $R^2 = 0.058$ , P < 0.001).

228

229 Prediction intervals generated using the linear regression model predict that stool samples

with a norovirus RT-qPCR Ct value <40 will generate 92–100% of the full genome sequence,

with 95% certainty (Figure 5).

232 Failed samples

The outliers in Figure 3(f) are dominated by samples from two sequencing runs (#30 and 31;

Figure A3), which were known to have had processing problems during cDNA preparation.

Six of the 16 samples with Ct <30 and genome coverage <80% had sufficient residual</li>
specimen to be repeated; all of these passed on repeat.

237 Three samples (highlighted in Figure A3, detailed in Table A1) generated unexpectedly low % genome coverage (49–73%) given the RT-qPCR Ct values (22–29) but were not part of 238 239 sequencing runs 30 or 31. Sequences from all three samples were fragmented throughout 240 ORF 1, with ORF3 and ORF2 downstream from the capsid protruding domains, P1 and P2 241 absent (Figure A4). In all three cases, the % OTR (0.01, 2.53 and 6.76%) and average read 242 depth (1-, 120- and 137-fold) was low for ORF 1 despite apparently good Ct values. Coverage of ORF 1 and the 5' end of ORF 2 was sufficient to confirm two samples as GII.4 and one as 243 244 GII.5 using the norovirus genotyping tool; we have shown good sequencing outcomes for both genotypes in other samples (Table 1). It is not possible to exclude the possibility of a 245 246 novel recombinant strain, with recombination at the P1/P2 junction in ORF2, and subsequent failure due to missing complementary baits in the enrichment; however if this 247 were the case we would expect to see good coverage of the enriched region, in this case 248 249 ORF 1, which we do not. Moreover all three samples had been re-extracted at referring 250 centres and the Ct value supplied was obtained from PCRs carried out on the original 251 diagnostic extracts. This, combined with the low coverage of ORF 1, suggests that extraction failure at the local hospital may explain the unexpected sequencing failure. It has not been 252 possible to test either possibility, since none of the original sample remains. 253

254 Low titre samples

Seven samples generated full genome sequences despite low viral titres (PCR Ct ≥36). To
determine whether these samples had misleadingly late Ct values due to a mismatch in the
RT-qPCR primer target region, the seven genome sequences were aligned to the RT-qPCR

primer and probe sequences used to generate the Ct value. There were no mismatches in
the primer or probe sites (Figure A5), suggesting they are genuinely low titre samples and
confirming the sensitivity of the method for low titre samples.

261 Comparison to capsid genotyping

96% (158/164) and 100% (164/164) of samples processed in parallel were successfully
genotyped by PCR with Sanger sequencing and by our method, respectively (Table A2). For
the 158 samples typed by both methods, there was 100% agreement in the respective
genotypes. Of the 6 samples that failed capsid typing by PCR, four were GII.4, one GII.7 and
one GI.3 (Table A3).

Two of the failed samples, with Ct values 20 and 27, had mismatches at the genotyping
primer sites (Figure A6) which accounts for genotyping failure in these instances.

269 The remaining four of the six samples that failed genotyping had Ct values >30 (range 31–

270 37), which suggests the genotyping PCR is less sensitive than sequencing by target

271 enrichment.

272 Contamination

Two "negative extract" samples, consisting of Buffer ASL that was treated in the same way as, and alongside, stool samples, were negative for norovirus RNA by RT–qPCR. Nonetheless target enrichment and sequencing generated 16–36% OTR with 3–81-fold read depth. The genome coverage for each sample was only 9 and 12%, with reads fragmented across the genome (Figure A7 and Figure A8). The mapped regions do not correspond to PCR amplicon sites.

#### 279 Mixed infections

280 Three (3/507) samples were identified as having sequences from more than one genotype

during the assembly pipeline (Table A4). For two of the samples, the mixed infections were

- evident during the "mapping to reference list" step of the *de novo* pipeline (Figure 1), in
- 283 which reads mapped to reference sequences corresponding to multiple norovirus
- 284 genotypes, as per Table A4. For the third sample, mixed infection was evident during the
- 285 "align contigs to single reference of appropriate genotype" step, in which a full length contig
- mapped to the reference sequence at ORF1 but not at ORF 2 and ORF 3.
- 287 Comparison of the consensus sequences generated from a single infection and from a

simulated mixed infection showed 178–332 single nucleotide polymorphisms (SNPs) and

289 95.53–97.61% sequence identity between the consensus sequences from the single and

290 mixed datasets (Table A5).

## 291 Turn-around times and costs

292 The turn-around times associated with full genome sequencing by SureSelect target

293 enrichment is 6 days; three days longer than genotyping (RNA-dependent RNA polymerase

and capsid regions) by PCR and Sanger sequencing with an extra associated cost of £54

when reagents are purchased in bulk (Table 2).

## 296 Discussion

- 297 Target enrichment is a highly effective method for sequencing norovirus full genomes across
- 298 genotypes with a high read depth averaging over 12,000-fold and complete or almost
- complete genomes in 89% of samples. We report median genome coverage of 100% across

all sequenced samples and, once outliers are accounted for, over 80% genome coverage
 regardless of the viral titre.

However, despite good molecular practice, low level contamination does occur. Since 302 negative extracts were RT-qPCR negative but target enrichment yielded reads that map to 303 304 the norovirus genome, we suspect the source of contamination is the automated equipment 305 used for target enrichment and sequencing library preparation. In the context of norovirus-306 positive specimens, the contamination is low; reads are fragmented and only map to 9-12% 307 of the genome with <100-fold read depth, which is significantly below the observed median 308 % genome coverage and read-depth seen for norovirus-positive samples (100% and >12,000-fold, respectively) and below the 95% prediction intervals for % genome coverage 309 310 (92–100% for a sample with Ct <40). These findings support our acceptance criteria for 311 downstream analysis, which is >100-fold read depth and >90% genome coverage. Where a complete genome sequence is not critical for downstream analysis, based on the 95% 312 prediction intervals, >60% genome coverage would be acceptable if read depth is >100-fold. 313 314 However due to the potential for low level contamination, specimens for which norovirus 315 RNA is not detectable by real-time PCR should not be sequenced.

Previous reports have described whole norovirus genome sequencing with overlapping PCR
amplicons or using RNASeq, the findings of which are summarised in Table A6. PCR- based
methods yield high read depth; however, due to sequence heterogeneity between
genotypes, primers generally need to be genotype specific (9). Broad-range primers have
been reported by Cotton *et al.* (10) nonetheless this approach retains a limited success rate;
full genome sequences were amplified from a comparable proportion of samples of GII.13
(83% versus 100% in this study), GII.6 (88% vs. 95%) and GII.4 (92% vs. 89% or 93%

irrespective of read depth). However PCR fared worse, recovering fewer full genomes from 323 324 GI (20% vs. 100% in this study), GII.2 (40% vs. 88%), GII.3 (77% vs. 87% or 90% irrespective of read depth) and GII.7 (0% vs. 90%). Norovirus whole genome sequencing from a single 7.5 325 kb amplicon has also been described and used to generate 25 full genome sequences (23) 326 327 however the authors do not report the success rate using this approach; it is generally very difficult to amplify fragments of such a size. Conversely here we report complete or nearly-328 complete genome sequences in 93% of processed samples. In target enrichment, baits are 329 330 designed using all publically available norovirus sequences, across all GI and GII genotypes; unlike PCR which uses a single primer at each target site, multiple baits are designed to 331 cover each position in the genome thus accounting for sequence variation between 332 333 norovirus genomes. This allows un-biased sequencing across known genotypes in a single reaction. A disadvantage of the method is that it may fail to generate sequences for a newly 334 335 emerging genotype where the existing baits are a poor match.

336 Whole transcriptome sequencing, or RNASeq, involves sequencing the total RNA or mRNA 337 content of a stool specimen. The advantage of RNASeq is that there is no requirement for 338 PCR primers therefore it is completely unbiased. Although all whole genomes by RNASeq 339 reported to date are predominantly GII.4, it is theoretically possible to sequence all 340 genotypes with equal success as evidenced by Bavelaar et al who successfully sequenced five non-GII.4 genomes (16). The data generated by RNASeq is sufficient to generate almost 341 complete norovirus genome sequences; 40–100% of reported samples achieved >90% 342 343 genome coverage (13-16) (summarised in Table A6). However the median % OTR across all 344 reported samples is only 2–3% using a MiSeq or HiSeq (13, 15) and 28% using an Ion Torrent 345 PGM (16), compared to 81% OTR by SureSelect target enrichment. The high proportion of

non-target data using RNASeq makes the technique uneconomical and, critically, results in
low read depth; on average only 9–259-fold using a MiSeq or HiSeq (13-15) and 1,309 using
an Ion Torrent PGM (16). Conversely, the median read depth using target enrichment is over
12,000-fold which allows large sample batches to be sequenced on a single MiSeq run and
downstream analysis of minority variants.

351 Our *de novo* assembly pipeline identified mixed genotype infections in three samples. 352 However with as many as 332 SNPs between the consensus sequences generated from a 353 single and simulated mixed infection, we suggest that a reliable consensus sequence cannot 354 be generated using this assembly pipeline. This is due to mis-mapping of reads in relatively conserved regions, as evidenced by the majority of SNPs being found in ORF1 (163/178 and 355 356 284/332 in the GII.3 and GII.4 consensus sequences respectively). Thus whilst this pipeline 357 can identify infections with a mixture of genotypes, an alternative approach is required for assembly and generating the consensus sequence, possibly involving the use of minority 358 variants and haplotype reconstruction. 359

360 We have shown target enrichment to be superior to PCR capsid amplification for 361 genotyping; all samples (164/164) that were processed in parallel successfully generated genome sequences by target enrichment, whereas 96% (158/164) were successfully 362 363 amplified by capsid typing PCR. Four of the six samples that failed capsid genotyping but were sequenced by target enrichment had low norovirus titres (based on PCR Ct values), 364 which suggests target enrichment is more sensitive than the conventional genotyping 365 366 methods. The remaining two failed samples had primer mismatches that account for 367 amplification failure. Target enrichment overcomes the limitations of primer design by

allowing multiple baits with different sequences to target each region of the genome, thus
 accounting for sequence heterogeneity in a way that PCR primers cannot.

370 Unlike classical genotyping, which requires sequential PCR and sequencing reactions yielding 371 only fragments of the genome in return, full genome sequences can, in a single reaction, provide us with the RNA polymerase and capsid sequences, which are important for 372 373 genotyping, and in addition can identify recombination and reveal minority variants in the intra-host viral population. The cost of targeted enrichment whole genome sequencing is 374 375 around £50 more expensive than PCR genotyping of the capsid and polymerase genes. 376 However, whole genome sequencing using overlapping amplicons is comparable in cost to 377 enrichment methods. Turnaround time for the target enrichment is 6 days compared to 378 three days for capsid and polymerase genotyping. The semi-automated target enrichment 379 hands-on-time is 4 hours more than conventional genotyping and comparable to RNASeq. A current drawback is the need for batch processing of samples to achieve the costs savings. 380 This is feasible for a regional sequencing service or a named study, but might be difficult for 381 382 a diagnostic laboratory. Further developments to shorten hybridization and sequencing 383 times and to enable random access processing would address these drawbacks.

The advancement of sequencing techniques, from PCR with capillary sequencing to target enrichment with deep sequencing, facilitates the use of norovirus full genomes in clinical practice. In conjunction with growing expertise, lower costs and faster turn-around times, full genomes can be sequenced for under £100 in less than a week; this makes full genome sequencing a reality not just in academic settings but for informing public health practice in real time.

390

## 391 Acknowledgements

392	The authors thank the Great Ormond	Street Hospital Virology departme	ent, Royal Free
-----	------------------------------------	-----------------------------------	-----------------

393 Hospital Virology department, Norfolk and Norwich University Hospital Microbiology

394 department and Public Health England Virus Reference Department for supplying specimens

395 for sequencing.

396 The authors declare no conflicts of interest.

397

408

## 398 Funding Information

399 This work was funded by the PATHSEEK FP7 EU grant. PATHSEEK is funded by the European 400 Union's Seventh Programme for research, technological development and demonstration 401 under grant agreement no. 304875. JRBrown is funded by a National Institute for Health 402 Research (NIHR) doctoral fellowship (NIHR-HCS-D12-03-15). JRBrown and DS are supported 403 by the NIHR Biomedical Research Centre (BRC) at Great Ormond Street Hospital for Children 404 NHS Foundation Trust and University College London (UCL). JBreuer receives funding from 405 the NIHR UCL/UCLH BRC. We acknowledge the infrastructure support from the UCL 406 Pathogen Genomics Unit (PGU), the NIHR UCL/UCLH BRC and the UCL MRC CMMV. The funders had no role in study design, data collection and interpretation, or the decision 407

to submit the work for publication.

# 409 References

- de Wit MA, Koopmans MP, Kortbeek LM, Wannet WJ, Vinje J, van Leusden F, Bartelds AI,
   van Duynhoven YT. 2001. Sensor, a population-based cohort study on gastroenteritis in the
   Netherlands: incidence and etiology. Am J Epidemiol 154:666-674.
- Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, Griffin PM.
   2011. Foodborne illness acquired in the United States--major pathogens. Emerg Infect Dis
   17:7-15.
- Ahmed SM, Hall AJ, Robinson AE, Verhoef L, Premkumar P, Parashar UD, Koopmans M,
   Lopman BA. 2014. Global prevalence of norovirus in cases of gastroenteritis: a systematic
   review and meta-analysis. Lancet Infect Dis 14:725-730.
- Lopman BA, Reacher MH, Vipond IB, Hill D, Perry C, Halladay T, Brown DW, Edmunds WJ,
   Sarangi J. 2004. Epidemiology and cost of nosocomial gastroenteritis, Avon, England, 2002 2003. Emerg Infect Dis 10:1827-1834.
- 422 5. Koo HL, Neill FH, Estes MK, Munoz FM, Cameron A, Dupont HL, Atmar RL. 2013.
  423 Noroviruses: The Most Common Pediatric Viral Enteric Pathogen at a Large University
  424 Hospital After Introduction of Rotavirus Vaccination. J Pediatric Infect Dis Soc 2:57-60.
- Payne DC, Vinje J, Szilagyi PG, Edwards KM, Staat MA, Weinberg GA, Hall CB, Chappell J,
   Bernstein DI, Curns AT, Wikswo M, Shirley SH, Hall AJ, Lopman B, Parashar UD. 2013.
   Norovirus and medically attended gastroenteritis in U.S. children. N Engl J Med 368:1121 1130.
- Verhoef L, Williams KP, Kroneman A, Sobral B, van Pelt W, Koopmans M. 2012. Selection of
  a phylogenetically informative region of the norovirus genome for outbreak linkage. Virus
  Genes 44:8-18.
- Jones MK, Watanabe M, Zhu S, Graves CL, Keyes LR, Grau KR, Gonzalez-Hernandez MB,
   Iovine NM, Wobus CE, Vinje J, Tibbetts SA, Wallet SM, Karst SM. 2014. Enteric bacteria
   promote human and mouse norovirus infection of B cells. Science 346:755-759.
- 435 9. Kundu S, Lockwood J, Depledge DP, Chaudhry Y, Aston A, Rao K, Hartley JC, Goodfellow I,
  436 Breuer J. 2013. Next-generation whole genome sequencing identifies the direction of
  437 norovirus transmission in linked patients. Clin Infect Dis 57:407-414.
- 438 10. Cotten M, Petrova V, Phan MV, Rabaa MA, Watson SJ, Ong SH, Kellam P, Baker S. 2014.
  439 Deep sequencing of norovirus genomes defines evolutionary patterns in an urban tropical 440 setting. J Virol 88:11056-11069.
- Won YJ, Park JW, Han SH, Cho HG, Kang LH, Lee SG, Ryu SR, Paik SY. 2013. Full-genomic
  analysis of a human norovirus recombinant GII.12/13 novel strain isolated from South Korea.
  PLoS One 8:e85063.
- 444 12. Chhabra P, Walimbe AM, Chitambar SD. 2010. Complete genome characterization of
  445 Genogroup II norovirus strains from India: Evidence of recombination in ORF2/3 overlap.
  446 Infect Genet Evol 10:1101-1109.
- 13. Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, Yamashita A, Goto N, Takahashi K,
  Yasunaga T, Ikuta K, Mizutani T, Okamoto Y, Tagami M, Morita R, Maeda N, Kawai J,
  Hayashizaki Y, Nagai Y, Horii T, Iida T, Nakaya T. 2009. Direct metagenomic detection of
  viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing
  approach. PLoS One 4:e4219.
- 452 14. Wong TH, Dearlove BL, Hedge J, Giess AP, Piazza P, Trebes A, Paul J, Smit E, Smith EG,
  453 Sutton JK, Wilcox MH, Dingle KE, Peto TE, Crook DW, Wilson DJ, Wyllie DH. 2013. Whole
  454 genome sequencing and de novo assembly identifies Sydney-like variant noroviruses and
  455 recombinants during the winter 2012/2013 outbreak in England. Virol J 10:335.
- 456 15. Batty EM, Wong TH, Trebes A, Argoud K, Attar M, Buck D, Ip CL, Golubchik T, Cule M,
  457 Bowden R, Manganis C, Klenerman P, Barnes E, Walker AS, Wyllie DH, Wilson DJ, Dingle

458		KE, Peto TE, Crook DW, Piazza P. 2013. A modified RNA-Seq approach for whole genome
459		sequencing of RNA viruses from faecal and blood samples. PLoS One 8:e66129.
460	16.	Bavelaar HH, Rahamat-Langendoen J, Niesters HG, Zoll J, Melchers WJ. 2015. Whole
461		genome sequencing of fecal samples as a tool for the diagnosis and genetic characterization
462		of norovirus. J Clin Virol <b>72:</b> 122-125.
463	17.	Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZ, Depledge DP,
464		Nikolayevskyy V, Broda A, Stone MJ, Christiansen MT, Williams R, McAndrew MB, Tutill H,
465		Brown J, Melzer M, Rosmarin C, McHugh TD, Shorten RJ, Drobniewski F, Speight G, Breuer
466		J. 2015. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly
467		from Clinical Samples. J Clin Microbiol 53:2230-2237.
468	18.	Christiansen MT, Brown AC, Kundu S, Tutill HJ, Williams R, Brown JR, Holdstock J, Holland
469		MJ, Stevenson S, Dave J, Tong CY, Einer-Jensen K, Depledge DP, Breuer J. 2014. Whole-
470		genome enrichment and sequencing of Chlamydia trachomatis directly from clinical samples.
471		BMC Infect Dis <b>14:</b> 591.
472	19.	Depledge DP, Palser AL, Watson SJ, Lai IY, Gray ER, Grant P, Kanda RK, Leproust E, Kellam
473		P, Breuer J. 2011. Specific capture and whole-genome sequencing of viruses from clinical
474		samples. PLoS One <b>6:</b> e27805.
475	20.	Kroneman A, Vennema H, Deforche K, v d Avoort H, Penaranda S, Oberste MS, Vinje J,
476		Koopmans M. 2011. An automated genotyping tool for enteroviruses and noroviruses. J Clin
477		Virol <b>51:</b> 121-125.
478	21.	Kent WJ. 2002. BLATthe BLAST-like alignment tool. Genome Res 12:656-664.
479	22.	Brown JR, Gilmour K, Breuer J. 2016. Norovirus Infections Occur in B-Cell-Deficient Patients.
480		Clin Infect Dis <b>62:</b> 1136-1138.
481	23.	Eden JS, Tanaka MM, Boni MF, Rawlinson WD, White PA. 2013. Recombination within the
482		pandemic norovirus GII.4 lineage. J Virol <b>87:</b> 6270-6282.

	Number of samples sequenced	Number samples Pass (%)	Number samples Sub- optimal (%)	Number samples Fail (%)	Median % OTR (min–max)	Median read depth (min–max)	Median % genome coverage (min– max)	Median Ct values (min– max)
GI.1	2	2 (100%)	0 (0%)	0 (0%)	63.05 (43.85-82.25)	11,194 (7,239–15,149)	100 (100–100)	31 (30–32)
GI.2	4	4 (100%)	0 (0%)	0 (0%)	77.60 (2.17-94.70)	11,464 (379–21,843)	100 (99–100)	29 (24–33)
GI.3	15	15 (100%)	0 (0%)	0 (0%)	74.13 (1.08-93.25)	13,157 (246–27,569)	100 (90–100)	27 (17–35)
GI.6	1	1 (100%)	0 (0%)	0 (0%)	86.56 (n/a)	8,642 (n/a)	100 (n/a)	29 (n/a)
GI.7	1	1 (100%)	0 (0%)	0 (0%)	83.88 (n/a)	18,414 (n/a)	100 (n/a)	21 (n/a)
Gl.ut	2	1 (50%)	0 (0%)	1 (50%)	40.34 (9.50-71.18)	7,000 (42–13,957)	91 (83–100)	29 (23–35)
GII.1	3	3 (100%)	0 (0%)	0 (0%)	95.61 (20.06-97.04)	11,990 (4,365–16,506)	100 (99–100)	15 (14–31)
GII.13	1	1 (100%)	0 (0%)	0 (0%)	77.44 (n/a)	10,043 (n/a)	100 (n/a)	21 (n/a)
GII.14	6	6 (100%)	0 (0%)	0 (0%)	53.31 (4.20-81.60)	10,238 (1,081–15,215)	100 (100–100)	27 (21–32)
GII.17	2	2 (100%)	0 (0%)	0 (0%)	63.30 (40.27-86.33)	13,204 (8,598–17,811)	100 (100–100)	24 (21–27)
GII.2	24	21 (88%)	0 (0%)	3 (12.5%)	57.60 (0.60-99.47)	4,717 (7–23,889)	100 (64–100)	24 (18–32)
GII.3	105	91 (87%)	3 (2.9%)	11 (10.5%)	85.00 (0.02-99.36)	16,034 (7–38,843)	100 (3–100)	21 (10–38)
GII.4	281	250 (89%)	12 (4.3%)	19 (6.8%)	83.75 (0.02-99.63)	12,465 (1–46,996)	100 (5–100)	22 (12–43)
GII.5	6	5 (83%)	0 (0%)	1 (16.7%)	70.21 (0.04-97.13)	16,468 (1–29,488)	100 (49–100)	19 (16–23)
GII.6	40	38 (95%)	0 (0%)	2 (5%)	70.32 (0.45-98.23)	9,356 (3–31,643)	100 (22–100)	21 (13–33)
GII.7	10	9 (90%)	0 (0%)	1 (10%)	53.14 (2.72-83.88)	12,779 (2,106–26,914)	100 (96–100)	25 (22–30)
Gll.ut	4	3 (75%)	1 (25%)	0 (0%)	49.02 (0.59-92.61)	11,356 (98–23,588)	100 (94–100)	25 (19–35)
NegEx	2	0 (0%)	0 (0%)	2 (100%)	26.30 (16.18-36.42)	42 (3–81)	11 (9–12)	Not detected
TOTAL Total	509	453 (89%)	16 (3%)	40 (8%)	81.22 (0.02–99.63)	12,227 (1–46,996)	100 (3–100)	22 (10–43)
excl. Run 30 & 31	413	381 (92%)	16 (4%)	16 (4%)	84.45 (0.02–99.63)	14,341 (1–46,996)	100 (13–100)	22 (10–40)

484 **Table 1.** Metrics of norovirus whole genome sequencing for all samples (TOTAL) and for each genotype.

485 Pass, >90% genome coverage and >100-fold read depth; Sub-optimal, >90% genome coverage or >100-fold read depth; Fail, <90% genome

486 coverage and <100-fold read depth; n/a, range not applicable due to single sample; % OTR, percent on target reads; Ct, real-time PCR cycle

487 threshold; GI.ut, genogroup I untypable; GII.ut, genogroup II untypable; NegEx, negative control

488 **Table 2**. Turn-around times and costs associated with norovirus genotyping by PCR and

489 Sanger sequencing compared to SureSelect target enrichment full genome sequencing.

Genotyping method	Hands on time	Total turn- around time	Reagent costs per sample
PCR and Sanger sequencing*	7 hrs.	3 days	£32
Full genome sequencing by SureSelect target enrichment	11 hrs. 30 mins	6 days	£86–£93**

491 polymerase (RdRp), capsid shell domain and capsid P2 domain, including one round of

492 nested PCR, assuming RdRp and capsid shell domain targets are amplified and sequenced

493 simultaneously

490

<sup>494</sup> \*\* Cost based on batches of 96 or 48 samples and sequencing on an Illumina MiSeq.

495	Figure	legends
-----	--------	---------

496 **Figure 1**. Schematic of norovirus full gnome assembly pipeline

497

- 498 **Figure 2.** Number of samples sequenced according to norovirus genotype, classified by
- 499 sequencing outcome. Pass, >90% genome coverage and >100-fold read depth; Sub-optimal,
- 500 >90% genome coverage or >100-fold read depth; Fail, , <90% genome coverage and <100-

501 fold read depth. Genotype refers to capsid genotype only.

502

503 **Figure 3.** Norovirus full genome sequencing outcome metrics according to (a–c) norovirus

504 genotype and (d–f) RT-qPCR Ct value. Red lines indicate median value

505

**Figure 4**. RT-qPCR Ct value of all samples, excluding Run 30 and 31, (n = 413) sequenced by

507 SureSelect. Pass, >90% genome coverage and >100-fold read depth; Sub-optimal, >90%

508 genome coverage or >100-fold read depth; Fail, , <90% genome coverage and <100-fold

509 read depth

510

511 Figure 5. Observed and predicted % genome coverage values with 95% prediction intervals,

excluding outliers identified in Figure A3. Fitted linear regression model: y = 7.432 - 7.432

513 0.059x where the dependent variable y is the logit transformed genome coverage

proportion and the independent variable x is the PCR Ct value (n = 477).