

1 **Impact of chimera-less long reads on metagenomics of human gut viromes treated with**
2 **multiple displacement amplification**

3

4 Yuya Kiguchi^{1, 2, 3}, Suguru Nishijima^{1, 2, 4}, Naveen Kumar³, Masahira Hattori^{1, 3*}, Wataru
5 Suda^{3*}

6

7 ¹Graduate School of Advanced Science and Engineering, Waseda University, Tokyo 169-8555,
8 Japan.

9 ²Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of
10 Advanced Industrial Science and Technology, Tokyo 169-8555, Japan.

11 ³Laboratory for Microbiome Sciences, RIKEN Center for Integrative Medical Sciences,
12 Yokohama 230-0045, Japan.

13 ⁴Structural and Computational Biology Unit, European Molecular Biology Laboratory,
14 Heidelberg, Germany.

15

16 *Correspondence: wataru.suda@riken.jp; hattori@edu.k.u-tokyou.ac.jp

17

18

19

20

21

22

23

24

25

26 **Abstract**

27 **Background:** The ecological and biological features of the indigenous phage community
28 (virome) in the human gut microbiome are poorly understood, possibly due to many
29 fragmented contigs and fewer complete genomes based on conventional short-read
30 metagenomics. Long-read sequencing technologies have attracted attention as an
31 alternative approach to reconstruct long and accurate contigs from microbial
32 communities. However, the impact of long-read metagenomics on human gut virome
33 analysis has not been well evaluated.

34

35 **Results:** Here we present chimera-less PacBio long-read metagenomics of multiple
36 displacement amplification (MDA)-treated human gut virome DNA. The method
37 included the development of a novel bioinformatics tool, SACRA (Split Amplified
38 Chimeric Read Algorithm), which efficiently detects and splits numerous chimeric reads
39 in PacBio reads from the MDA-treated virome samples. SACRA treatment of PacBio
40 reads from five samples markedly reduced the average chimera ratio from 72 to 1.5%,
41 generating chimera-less PacBio reads with an average read-length of 1.8 kb. *De novo*
42 assembly of the chimera-less long reads generated contigs with an average N50 length of
43 11.1 kb, whereas those of MiSeq short reads from the same samples were 0.7 kb,
44 dramatically improving contig extension. Alignment of both contig sets generated 378
45 high-quality merged contigs (MCs) composed of the minimum scaffolds of 434 MiSeq and
46 637 PacBio contigs, respectively, and also identified numerous MiSeq short fragmented
47 contigs ≤ 500 bp additionally aligned to MCs, which possibly originated from a small
48 fraction of MiSeq chimeric reads. The alignment also revealed that fragmentations of the
49 scaffolded MiSeq contigs were caused primarily by genomic complexity of the community,
50 including local repeats, hypervariable regions, and highly conserved sequences in and

51 between the phage genomes. We identified 142 complete and near-complete phage
52 genomes including 108 novel genomes, varying from 5 to 185 kb in length, the majority
53 of which were predicted to be *Microviridae* phages including several variants with
54 homologous but distinct genomes, which were fragmented in MiSeq contigs.

55

56 **Conclusions:** Long-read metagenomics coupled with SACRA provides an improved
57 method to reconstruct accurate and extended phage genomes from MDA-treated virome
58 samples of the human gut, and potentially from other environmental virome samples.

59

60 **Keywords:** virome, bacteriophage, gut, metagenomics, long-read, chimera, multiple
61 displacement amplification

62

63 **Background**

64 Indigenous gut bacteriophages (phages) comprise the major component of the human gut
65 microbiome [1–3]. The estimated ratio of phages to bacterial cells in the feces varies from 1:1
66 to 1:10, suggesting that phages exist at numbers comparable to bacterial cells [4]. Furthermore,
67 gut phages have a primary role in functions of the gut microbiome by altering the community
68 structure and metabolism through the infection of host bacteria [5, 6]. Several papers also
69 reported associations between gut phages and microbiome dysbiosis, linking it to diseases such
70 as inflammatory bowel diseases and diabetes [7–12] and implying the impact of gut phages on
71 the human physiological state.

72 In analyses of the gut phage community (virome) [1–3, 13, 14], the first step is the
73 enrichment of virus-like particles (VLPs) from vast numbers of bacterial cells in fecal samples.
74 VLPs are then proteolytically lysed to generate viral DNA, which is further subjected to
75 multiple displacement amplification (MDA) using DNA polymerases with strand displacement

76 activity to obtain the increased DNA quantity sufficient for sequencing. More importantly,
77 MDA enables the detection of single-stranded phage genomes in the sequence data [15].

78 Although virome metagenomics is a powerful approach to comprehensively identify
79 phage genomes in a community [1–3, 14], the assembly generates many fragmented and fewer
80 completed contigs [13, 14, 16]. Contig fragmentation is problematic in virome studies, making
81 it difficult to precisely conduct taxonomic assignments, diversity and abundance estimations,
82 and host prediction based on the genomic data [17–19]. This fragmentation is possibly caused
83 by the genomic complexity in the viral communities [17, 20], which is often encountered in
84 metagenomics with short reads <300 bp [21]. To overcome such shortcomings in short-read
85 metagenomics, several studies have reported improvements of overall contig extension in
86 metagenomics of the microbial communities with long reads >10 kb using single-molecule
87 sequencing technologies [22–26]. However, long-read metagenomics of the human gut virome
88 has not been evaluated. Here, we provide a method for efficient reconstruction of accurate and
89 extended contigs including completed phage genomes by long-read metagenomics of the
90 MDA-treated human gut virome, for which we developed a novel bioinformatics tool, SACRA
91 (Split Amplified Chimeric Read Algorithm). SACRA was crucial for the detection and splitting
92 of numerous chimeric reads from the MDA-treated DNA to obtain chimera-less long reads
93 with high efficiency.

94

95 **Results**

96 **Metagenomic sequencing of MDA-treated viral DNA**

97 We prepared viral DNA samples from virus-like particles (VLPs) enriched from the feces of
98 five healthy Japanese adults, and the samples containing the spike-in lambda phage (λ) DNA
99 were then subjected to MDA using EquiPhi29 polymerase [27] (Methods) to obtain sufficient
100 amounts of DNA for sequencing (**Table S1**). We sequenced the MDA-treated DNA with the

101 PacBio and MiSeq platforms and removed reads mapped to human, fungus, and the PacBio
102 internal control sequences (**Table S2**). The filter-passed PacBio reads were further treated with
103 canu [28] to obtain error-corrected PacBio reads (ECLs) (**Table S3**). The ECLs and canu-
104 untreated reads had an average sequence similarity of 96.7% and 90.6% with high-quality
105 contigs from the filter-passed MiSeq short reads (FMSs) obtained from the same MDA-treated
106 samples, respectively, improving the accuracy of ECLs (**Suppl. Fig. 1**).

107

108 **Chimeric reads in metagenomic sequencing of the MDA-treated virome samples**

109 We estimated the average chimera ratio to 72% for ECLs and 2.2% for FMSs, respectively, by
110 aligning them to the spike-in λ genome with $\geq 95\%$ identity, indicating a considerably high
111 chimera rate in ECLs, as observed previously [29] (**Fig. 1**). However, normalization of the
112 chimera ratio by read-length suggested similar chimera occurrences between them (**Suppl. Fig.**
113 **2**). We applied Pacasus, a bioinformatics tool for correcting PacBio chimeric reads containing
114 palindromic sequence [29], to ECLs but observed an inefficient reduction in the average
115 chimera ratio to only 34% for the ECLs aligned to the λ genome (**Fig. 1**), suggesting the
116 presence of Pacasus-insensitive chimeric reads in the ECLs.

117

118 **Structure of Pacasus-insensitive chimeric reads**

119 From analyses of Pacasus-insensitive chimeric reads aligned to the λ genome, we found that
120 those were intragenomic non-palindromic chimeric reads (intra-NPCRs) composed of two
121 different lambda sequences flanked with inverted (intra-NPCRI) and tandem rearrangements
122 (intra-NPCRT), which were further classified into two types based on the position in the
123 chimeric reads, and intergenomic non-palindromic chimeric reads (inter-NPCRs) composed of
124 the lambda and unrelated sequences (**Suppl. Fig. 3a**). The average ratio was 85.1% for intra-
125 NPCRs, designated as the intra-chimera rate (see below), and 14.9% for inter-NPCRs based on

126 the λ genome. In addition, chimeric position [30,31] and frequency on the λ genome were
127 highly similar between the technical replicates with an average Pearson's correlation of 0.82,
128 and the average frequency was estimated to be one per \sim 2.8 kb (**Table S4** and **Suppl. Fig. 3b**).

129

130 **Development of SACRA to correct PacBio chimeric reads**

131 We developed SACRA (Split Amplified Chimeric Read Algorithm) for correcting both NPCRs
132 and Pacasus-sensitive chimeric reads. To detect the chimeric reads, we first obtained aligned
133 read clusters (ARCs) by the pairwise local alignment of all ECLs in each sample. The generated
134 ARCs included those with both CARs (continuously aligned reads spanning the chimeric
135 position) and PARs (partially aligned reads with \geq 50-bp unaligned sequences from the
136 chimeric position), those with only PARs, and those with only CARs, accounting for 83.6%,
137 7.3%, and 0.4% of all ECLs on average, respectively (**Table S5**). ARCs with both CARs and
138 PARs could be further divided into chimeric ARCs composed of the true chimeric CARs and
139 non-chimeric PARs, and non-chimeric ARCs composed of non-chimeric CARs and the true
140 chimeric PARs (**Fig. 2a**). From ARCs on the λ genome, the average ratio of chimeric and non-
141 chimeric ARCs was estimated to be \sim 15% and \sim 85%, respectively (**Suppl. Fig. 4a**), suggesting
142 the requirement for a way to distinguish these two ARCs without references such as the λ
143 genome for the actual samples; otherwise, non-chimeric ARCs might be excessively split by
144 SACRA.

145 To this end, we considered the ratio of PARs/CARs (PC ratio) as a parameter to
146 distinguish the two ARCs, because PC ratios tended to be higher for chimeric ARCs than non-
147 chimeric ARCs on the λ genome (**Suppl. Fig. 4b**). We examined the relationship between the
148 PC ratio and the efficiency of the correct assignment of chimeric (sensitivity) and non-chimeric
149 ARCs (specificity) and determined the optimized minimum PC (mPC) ratio that maximized
150 the sensitivity and specificity to 0.08, 0.08, 0.05, 0.05, and 0.11 for each sample, with which

151 both specificity and sensitivity were >95% on the λ genome (**Fig. 2b**). Accordingly, ARCs with
152 greater than the optimized mPC ratio and ARCs with only PARs were subjected to SACRA to
153 split them as chimeric ARCs, whereas ARCs with less than the optimized mPC ratio and ARCs
154 with only CARs were not (**Fig. 3**).

155

156 **Efficiency of SACRA for correcting PacBio chimeric reads**

157 SACRA treatment with the optimized mPC ratio dramatically reduced the average chimera
158 ratio from ~72% to ~1.5%, similar to that of FMSs, on the λ genome in the five samples,
159 including corrections of >99% of the Pacasus-sensitive reads (**Fig. 1, Suppl. Figs. 5a and b**).
160 A comparison of ECLs, Pacasus-treated ECLs (PALs), and SACRA-treated ECLs (SALs) with
161 similar sequence quantities (**Table S3**) revealed that the average read-length was 5.5 kb for
162 ECLs, 3.1 kb for PALs, and 1.8 kb for SALs, and the overall read-length distribution of PALs
163 and SALs shifted to shorter than that of ECLs, due to the generation of shorter reads than ECLs
164 by splitting chimeric reads with Pacasus and SACRA (**Suppl. Fig. 5c**). The total bases of PALs
165 and SALs (≥ 1 kb) were similar to those of ECLs (≥ 1 kb) with ratios (PALs/ECLs and
166 SALs/ECLs) of >0.91 (**Suppl. Fig. 5d**), indicating no substantial reduction in the total bases
167 with Pacasus and SACRA treatments.

168

169 ***De novo* assembly of FMSs, PALs, and SALs**

170 We performed *de novo* assembly of FMSs, PALs, and SALs (Methods, **Table S6**), generating
171 a total of 80,716 contigs for FMSs, 8,593 for PALs, and 1,707 for SALs (**Table S7**). The
172 average N50 was 0.7 kb for FMS contigs, 7.2 kb for PAL contigs, and 11.1 kb for SAL contigs,
173 and >99% of the FMS contigs had a length <5 kb, whereas ~27% of the SAL contigs had a
174 length ≥ 5 kb (**Figs. 4a and 4b**).

175 The single complete contigs of the λ genome were reconstructed from all samples in
176 the SAL assembly, only from one sample in the FMS assembly, and not from all samples in
177 the PAL assembly (**Suppl. Fig. 6**), which might be partly due to a higher chimera ratio in PALs
178 and shorter reads in FMSs than SALs. In addition, no complete λ genome was recovered from
179 the four samples in the assembly of SALs from SACRA treatment with a PC ratio ≥ 0 , probably
180 due to excessive splitting of non-chimeric ARCs (**Suppl. Fig. 6**).

181

182 **Alignment of FMS and SAL contigs**

183 The alignment of FMS and SAL contigs with $\geq 95\%$ identity generated a total of 701 MCs with
184 a minimum scaffold of 1,338 FMS and 1,410 SAL contigs, accounting for 1.9 FMS and 2.0
185 SAL contigs per MC on average, respectively (**Fig. 5a** and **Table S8**). In addition to the
186 scaffolded FMS contigs, we also found 40,455 short FMS contigs (mostly ≤ 500 bp) aligned to
187 the MCs (**Fig. 5a**), most of which disappeared in the assembly of FMSs excluding chimeric
188 reads for the spike-in λ genome with subtle changes in assembly outcomes (**Tables S8** and **S9**,
189 **Discussion**). The remaining 26,788 FMS and 297 SAL contigs unaligned between them were
190 specific to each. Quantification of the average abundance of MCs and FMS- and SAL-specific
191 contigs by mapping FMSs revealed 104 FMSs/kb for FMS-specific contigs, 8,423 for MCs,
192 and 0.9 for SAL-specific contigs (**Table S10**), suggesting that the generated MCs included
193 relatively highly abundant genomes. We then removed all SAL-specific contigs and 323 MCs
194 with low read-depth from further analysis because of the low sequence accuracy. The average
195 length of 378 high read-depth MCs increased to 20.1 kb, fairly longer than that of all FMS (591
196 bp) and all SAL contigs (5.9 kb) (**Fig. 5b**). We finally obtained 324 non-redundant MCs by
197 clustering the 378 MCs with $\geq 99\%$ identity and coverage.

198 We found four types of contig fragmentations in alignments of the scaffolded FMS and
199 SAL contigs (**Table S11** and **Suppl. Fig. 7**). Type 1 fragmentations were observed for FMS

200 contigs at the local repeat region in SAL contigs. Type 2 fragmentations were observed for
201 FMS contigs aligned to multiple SAL contigs homologous but distinct, and included contigs
202 with mosaic sequences consisting of the highly similar sequences between the homologous
203 SAL contigs. Type 3 fragmentations of FMS contigs occurred mostly near the hypervariable
204 loci containing many single nucleotide polymorphic sites [20] in SAL contigs. Type 4
205 fragmentations were observed only in SAL contigs, mostly at the relatively low read-depth
206 region in FMS contigs. Among the FMS fragmentations, type 2 fragmented contigs were most
207 frequent (**Fig. 5c** and **Table S11**).

208

209 **Identification of complete and near-complete contigs**

210 From the 324 non-redundant MCs, we identified 159 MCs including 17 FMS-specific contigs
211 as complete contigs generating circular contigs (CCs), a hallmark of the full-length genome
212 [22]. For determination of the near-complete linear contigs (ncLCs), we developed and used
213 two parameters to assess the completeness of contigs (Discussion). One was the intra-chimera
214 rate of SAL contigs, which was significantly reduced from ~85% to ~50% as the contig length
215 was shortened to 80%, 50%, and 20% of the complete λ genome (100%) (**Suppl. Fig. 8**). We
216 set the cut-off value of the intra-chimera rate to 80% to screen the LCs/MCs (≥ 10 kb) and
217 obtained 95 LCs with an intra-chimera rate $\geq 80\%$. Of them, 13 were identified as ncLCs with
218 the second parameter that the total length of either shorter SAL or FMS contigs aligned to the
219 MC was $\geq 97\%$ of that of the MC, which was consistently observed for all complete CCs
220 composed of both SAL and FMS contigs. Overall, we identified a total of 172 complete and
221 near-complete contigs in the five samples (**Tables S12** and **S13**).

222

223 **Characterization of complete and near-complete contigs**

224 The 172 contigs were characterized by phage classification assessments such as VirSorter [32],
225 gene annotation by pVOG [33], and a similarity search with the publicly available databases.
226 The results showed that 142 contigs (129 CCs and 13 ncLCs) were classified as phage with the
227 VirSorter category 1, 2, or 3 and with at least one pVOG (**Table S12**). Other questionable
228 contigs negative for the VirSorter assessment or with no pVOG, including nine matched with
229 known plasmids, were excluded from further analysis (**Table S13**).

230 Of the 142 phage contigs, 108 (101 CCs and seven ncLCs) were likely novel phage
231 genomes because of a lack of similarity with the known genomes, whereas 34 (28 CCs and six
232 ncLCs) had high sequence similarities of $\geq 88.9\%$ and almost identical lengths with the known
233 full-length phage genomes. Of the 34 contigs, seven had similarity with phages of the family
234 *Microviridae* and 11 had similarity with those of the crAss-like phages, but the known phages
235 matched with other 16 contigs lacked taxonomic information (**Table S12**). The pVOG-based
236 family-level taxonomy prediction of the novel and taxon-unknown phage contigs revealed that
237 77 contigs were possibly assigned to *Microviridae*, 27 to *Siphoviridae*, 12 to *Myoviridae*, six
238 to *Podoviridae*, one to *Inoviridae*, and one unassigned (**Table S14**). In total, we identified six
239 families including 84 *Microviridae*, 27 *Siphoviridae*, 12 *Myoviridae*, 11 crAss-like, six
240 *Podoviridae*, and one *Inoviridae* in the samples (**Suppl. Fig. 9**).

241 We identified a total of 4,675 open reading frames (ORFs) as putative genes in the 142
242 phage contigs including 1,077 ORFs matched with pVOGs (**Table S12**). The average number
243 of ORFs on contigs was 32.9 per phage contig, 8.8 per SAL contig, and 1.5 per FMS contig,
244 respectively. The median ORF length in the 142 phage contigs was 150 amino acids (aa), which
245 was longer than 76 aa in FMS contigs, 110 aa in SAL contigs, and 132 aa in the reference
246 phage genomes (**Fig. 6**).

247

248 **Discussion**

249 The present data revealed numerous chimeric reads in PacBio reads from the MDA-treated
250 human gut virome samples (**Fig. 1**). The high chimera rate was also observed in PacBio reads
251 from MDA-treated DNA of non-metagenomic samples, in which Pacasus improved the
252 mapping rate by reducing the chimeric reads [29]. Nevertheless, Pacasus inefficiently reduced
253 the chimeric ratio from ~72% to only ~34% in our PacBio reads, which substantially hampered
254 reconstruction of the complete spike-in λ genome (**Fig. 1** and **Suppl. Fig. 6**). We found that
255 Pacasus-insensitive NPCRs were composed of two different sequences, differing from the
256 Pacasus-sensitive reads with palindromic sequences (**Suppl. Fig. 3a**). The mechanism for
257 NPCR formation by MDA might involve the template switching of DNA extension by
258 polymerases to annealable sequences spatially close but sequentially distant [30]. Because the
259 MDA-mediated chimera formation was highly reproducible in terms of the position and
260 frequency in the genomes (**Suppl. Fig. 3b**), it might not be largely influenced by other
261 coexisting genomic DNA under the present MDA conditions, which might also be supported
262 by the higher rate of intra-NPCRs than inter-NPCRs (**Suppl. Fig. 8**). As the estimated chimera
263 frequency was similar between the samples (**Suppl. Fig. 2, Table S4**), the higher chimera ratio
264 in PacBio than FMS reads is explained by its longer read-length than that of FMSs.

265 SACRA was developed to correct both Pacasus-sensitive and NPCRs with high
266 efficiency. SACRA included simple computational steps, pairwise local alignment of PacBio
267 reads to construct ARCs, PC ratio calculation, PC ratio-based selection, and splitting of
268 chimeric ARCs (**Fig. 3**), in which the PC ratio was most crucial. In fact, SACRA treatment
269 with the optimized mPC ratio, which maximized the sensitivity and specificity for
270 distinguishing the two ARCs (**Fig. 2**), dramatically reduced the average chimera ratio to ~1.5%
271 (**Fig. 1** and **Table S3**), recovering the complete λ genome in all the samples, whereas no
272 complete λ genome was recovered from SALs with an mPC ratio ≥ 0 in four samples (**Suppl.**
273 **Fig. 6**). Of note, the overall SAL assembly was improved by SACRA treatment with the

274 optimized mPC ratio in each sample (**Fig. 4**), suggesting that the optimized mPC ratio
275 determined from reads aligned to the spike-in λ genome is effective for reads from other phage
276 genomes in the samples as well.

277 *De novo* assembly of chimera-less SALs markedly increased the N50 and contigs with
278 a length >5 kb, whereas the contig number was much less than that of FMSs (**Fig. 4** and **Table**
279 **S7**), largely due to differences in read numbers between them, where were ~ 7 -fold higher in
280 FMSs than SALs in this study (**Table S6**). For the alignment of FMS and SAL contigs, we
281 found numerous short FMS contigs (mostly ≤ 500 bp) aligned to the MCs, in addition to the
282 minimum scaffolded FMS contigs (**Table S8**). Assembly of the FMSs excluding chimeric
283 reads on the λ genome effectively suppressed the generation of these short contigs (**Table S9**),
284 suggesting that the short contigs originated from a small fraction of FMS chimeric reads, which
285 could not be efficiently removed prior to assembly in the actual samples without the reference
286 genomes.

287 Typical FMS contig fragmentations were caused by local repeats (type 1) and
288 hypervariable regions (type 3) [20] in individual phage genomes and by the highly conserved
289 regions between multiple homologous phage genomes in the community (type 2) (**Suppl. Fig.**
290 **7** and **Table S11**), suggesting that FMS contig fragmentation largely depended on the
291 community's genomic complexity, and the most frequent type 2 fragmentation for FMS contigs
292 was difficult to identify without SAL contigs. However, SAL contig fragmentations (type 4)
293 occurred mostly at the low read-depth regions in FMS contigs, which might be caused by the
294 biased MDA [34]. Although type 4 fragmentation was more frequent than other types (**Table**
295 **S11**), this was probably specific and limited to this study due to much fewer SALs than FMSs,
296 as described, as well as undetected SAL contigs corresponding to the relatively low abundance
297 of FMS-specific contigs (**Table S6**). In other words, SAL contig fragmentations will be

298 decreased by increasing the SALs per sample, whereas FMS contig fragmentations could be
299 marginally improved even when the reads increase.

300 It is challenging to identify complete LCs without TDRs because they have no
301 characteristic structural feature of the full-length genome, unlike the linear genomes with TDRs
302 [22]. In this study, we therefore defined ncLCs without TDRs as those containing SAL contigs
303 with a $\geq 80\%$ intra-chimera rate (**Suppl. Fig. 8**) and with a $\geq 97\%$ length coverage of either
304 shorter SAL or FMS contigs in the MC. Using these two parameters, we conservatively
305 identified 13 ncLCs/MCs, of which, six were aligned to the known full-length phage genomes
306 with high similarity (**Table S12**). Because all of the six ncLCs aligned to the known phage
307 genomes were captured by the intra-chimera rate parameter, the intra-chimera rate of SAL
308 contigs could be used to assess the completeness of non-TDR linear genomes in MDA-treated
309 virome samples.

310 The 142 complete and near-complete phage contigs identified in this study were
311 characterized by the majority (71 novel and 13 known phages) of relatively small genomes
312 (from 4.8 to 7 kb in length) belonging to the *Microviridae* family with ssDNA genomes [35]
313 (**Suppl. Fig. 9** and **Table S12**). Of the 84 complete *Microviridae* genomes, 18 (13 novel) were
314 recovered as type 2 fragmented contigs including mosaic sequences in the FMS assembly
315 (**Table S11** and **Suppl. Fig. 7**), suggesting the existence of some *Microviridae* phages with
316 homologous but distinct genomes in the human gut, which might be hardly reconstructed from
317 the conventional short reads. However, the abundance of these small circular ssDNA phages
318 in the gut cannot be precisely evaluated from data of the MDA-treated samples because they
319 were preferentially amplified by MDA as described previously [15]. Only 13 CCs, all of which
320 were novel *Microviridae* phages, were shared with multiple samples in this study (**Table S12**),
321 suggesting the high inter-individual variability in the gut phages, except for some phages
322 prevalent in the human gut.

323 Of the 129 complete CCs, some might have linear genomes with TDRs like
324 crAssphages, because the conversion of linear to circular DNA might also occur in MDA with
325 a mechanism similar to template switching in PCR-extension between the two unconnected
326 TDRs [22]. Metagenomic sequencing of non-MDA virome DNA samples will allow to
327 determine whether the contigs have circular or linear genomes with TDRs and to quantify the
328 abundance of phage genomes including ssDNA in the human gut [15,22]. The increased
329 number and length of ORFs in SAL contigs was remarkably (**Fig. 6**), providing more
330 information on gene organization and gene products in the phage genome than the conventional
331 short-read metagenomics, which will facilitate the accurate prediction of phage taxonomy
332 based on genomic signatures [36–38].

333

334 **Conclusion**

335 We present the chimera-less long-read metagenomics of MDA-treated human gut virome DNA
336 including the development of a novel bioinformatics tool, SACRA, to correct chimeric reads
337 in the PacBio reads. *De novo* assembly of the chimera-less long reads generated extended and
338 accurate contigs, facilitating identification of the complete phage genomes and the scaffolding
339 of fragmented contigs in the short-read metagenomics. Thus, the long-read metagenomics
340 coupled with SACRA is of great use to reconstruct high-quality complete phage genomes from
341 MDA-treated virome samples, and potentially from other environmental virome samples.

342

343 **Methods**

344 **Subjects and fecal sample collection**

345 Five unrelated healthy Japanese volunteers were recruited at RIKEN. None of the subjects were
346 treated with antibiotics during the collection of fecal samples. Freshly collected fecal samples
347 were transported at 4 °C to the laboratory in a plastic bag containing disposable oxygen-

348 absorbing and carbon dioxide-generating agents. In the laboratory, the fecal samples were
349 immediately frozen using liquid nitrogen and stored at $-80\text{ }^{\circ}\text{C}$ until use [39].

350

351 **Preparation of VLPs and viral DNA**

352 Viral DNA was prepared from feces according to the literature with some modifications [14].

353 Frozen feces (1.0 g) was suspended in 7 mL SM buffer (with 0.01% gelatin) by vortexing. The

354 suspension was filtrated with 100- μm pore membrane filters (Merck Millipore) and the filtrate

355 was centrifuged at $5,000 \times g$ for 10 min at $4\text{ }^{\circ}\text{C}$ to pellet the debris. The supernatant was further

356 filtrated with 5.0- μm and 0.45- μm PVDF pore membrane filters (Merck Millipore, Steriflip).

357 An equal volume of polyethylene glycol solution (20% PEG-6000-2.5M NaCl) was added to

358 the filtrate and the well-mixed solution was stored overnight at $4\text{ }^{\circ}\text{C}$. The solution was

359 centrifuged at $20,000 \times g$ for 45 min at $4\text{ }^{\circ}\text{C}$ to collect VLP pellets. Lysozyme (10. mg/sample,

360 SIGMA ALDRICH) was added in the pellet suspended in 1 ml SM buffer, and the solution

361 was incubated for 60 min at $37\text{ }^{\circ}\text{C}$ with gentle shaking. The lysate was incubated with 10 U

362 DNase (NIPPON GENE), 10 U TURBO DNase (Thermo Fisher Scientific), 20 U Baseline-

363 ZERO DNase (Epicentre), and 500 U Benzonase (SIGMA ALDRICH) in DNase buffer

364 (TURBO DNase, $1\times$ concentration) for 2 h at $37\text{ }^{\circ}\text{C}$ with gentle shaking. EDTA (final conc.

365 20 mM) was added, and the sample was heated for 15 min at $70\text{ }^{\circ}\text{C}$ to inactivate DNases. VLPs

366 were then lysed with Proteinase K (SIGMA ALDRICH; 0.5 mg/reaction) in the presence of

367 SDS (final conc. 0.1%) at $55\text{ }^{\circ}\text{C}$ for 30 min with gentle shaking. The lysate was mixed with an

368 equal volume of phenol/chloroform/isoamyl alcohol (Life Technologies Japan, Ltd) and

369 centrifuged at $9,000 \times g$ for 10 min at room temperature. DNA was precipitated by adding

370 sodium acetate (final conc. 0.3M) and an equal volume of isopropanol to the aqueous phase

371 and pelleted by centrifugation at $12,000 \times g$ for 15 min at $4\text{ }^{\circ}\text{C}$. The pellet was rinsed with 75%

372 ethanol and dissolved in TE10 buffer (10 mM tris-HCl 10 mM EDTA). RNase (NIPPON

373 GENE; 37.5 µg/sample) was added to the solution and incubated for 30 min at 37 °C. An equal
374 volume of the polyethylene glycol solution (20% PEG6000-2.5 M NaCl) was added and the
375 solution was kept on ice for at least 20 min. RNA-free DNA was pelleted by centrifugation at
376 12,000 × g for 10 min at 4 °C and rinsed with 75% ethanol twice. The DNA was dried and
377 dissolved in TE buffer (10 mM tris-HCl 1 mM EDTA). The DNA concentration was measured
378 with the Qubit HS DNA Assay Kit in a Qubit 2.0 fluorometer (Thermo Fisher Scientific).

379

380 **Multiple displacement amplification and sequencing**

381 Lambda phage genomic DNA, supplied by PacBio (Part number 001-119-535), was added to
382 viral DNA as spike-in DNA (1/20 of viral DNA amount). MDA was performed for 2.2–10.4
383 ng of viral DNA for 3 h at 45 °C using EquiPhi29 (Thermo Fisher Scientific) [27]. The
384 amplified DNA was purified using a 0.65× volume ratio of AMPure XP (Beckman Coulter).

385 For MiSeq shotgun metagenomic sequencing, Illumina libraries were constructed from
386 10 ng of MDA-treated DNA using KAPA HyperPrep Kits (KAPA Biosystems) with 13 cycles
387 of amplification. The concentration and size distribution of the libraries was determined with
388 a Qubit 2.0 and Bioanalyzer, respectively. KAPA HyperPrep libraries were subjected to 300-
389 bp paired-end sequencing with the MiSeq platform. After trimming terminal 50-bp sequences,
390 any 5' and 3' low quality bases (<20 QV) bases were also trimmed. Reads with a mean QV <20
391 or those with <100 bp were removed. The quality filtering of reads was performed using
392 prinseq-lite.pl [40]. Filter-passed MiSeq reads were mapped to the human genome (hg19),
393 RefSeq fungi (release 92), the phiX genome to remove them using Bowtie2 (version 2.3.2)
394 [41].

395 For PacBio Sequel shotgun metagenomic sequencing, SMRTbell libraries were
396 constructed from 1 µg of MDA-treated DNA according to the manufacturer's protocol
397 (Procedure & Checklist – Preparing SMRTbell Libraries using PacBio Barcoded Adapters for

398 Multiplex SMRT Sequencing, PN 101-069-200-02). In brief, MDA-treated DNA was sheared
399 by g-TUBE to ~10-kb fragments, which were then concentrated using a 0.45× volume ratio of
400 AMPure PB (Pacific Biosciences). Sheared DNA was treated as follows: 1) Exo VII digestion,
401 2) DNA damage repairing, 3) end repairing, 4) ligation of SMRTbell barcoded adaptor, 5)
402 pooling libraries, 6) Exo III and VII digestion, and 7) polymerase binding using Sequel binding
403 Kit 3.0. SMRTbell libraries were sequenced on the SMRT cell 1M v3 using diffusion loading
404 and a 600-min movie time. PacBio Sequel subreads were mapped to the human genome (hg19)
405 and a sequel internal control sequence using minimap2 (map-pb option)[42] with ≥80%
406 identity and ≥80% alignment coverage to remove them.

407

408 **Error correction of PacBio reads**

409 PacBio raw reads were subjected to error-correction by canu (v1.8) using -correct
410 genomeSize=10m minReadLength=500 minOverlapLength=250 corOutCoverage=10000
411 corMinCoverage=0 corMhapSensitivity=high -fast parameters [28]. For quantification of the
412 sequence accuracy, PacBio raw and error-corrected reads were aligned to the short-read contigs
413 generated from the assembly of MiSeq short reads by MEGAHIT using LAST (v 963) [43]
414 with ≥80% identity and ≥80% length coverage using -a 0 -A 10 -b 15 -B 7 and -a 8 -A 16 -b
415 12 -B 5 parameters, respectively, and sequence similarity of the alignments was calculated. To
416 determine optimal parameters for LAST alignment, PacBio raw and error-corrected reads were
417 aligned to the lambda phage (λ) genome using minimap2 (map-pb option) with ≥80% identity
418 and length coverage, respectively. Then, the existence and extension costs of gaps and
419 insertions were determined with LAST-TRAIN [44] from PacBio reads aligned to the λ
420 genome. The reference index was made using lastdb with -uNEAR and -R01 options. The
421 alignment with the highest similarity was selected as the top-hit alignment for the PacBio reads
422 aligned to multiple regions. The error-corrected PacBio reads were subjected to Pacasus

423 (v1.1.1) [29] to correct chimeric reads containing palindromic sequences with default
424 parameters.

425

426 **Detection and quantification of PacBio and MiSeq chimeric reads on the spike-in lambda**
427 **phage genome**

428 Filter-passed MiSeq short reads (FMSs) were aligned to the λ genome using BLASTN
429 (BLAST+ 2.6.0) with -e value 1.0e-5 -task blastn-short options. After removing alignments
430 with <95% identity and <50-bp length, chimeric reads were detected in alignments with <90%
431 length coverage of query reads and were removed prior to the assembly of FMSs without them.
432 Error-corrected PacBio reads (ECLs), Pacasus-treated ECLs (PALs), and SACRA-treated
433 ECLs (SALs) ≥ 500 bp were aligned to the λ genome with LAST using -a 8 -A 16 -b 12 -B 5
434 parameters with $\geq 95\%$ identity and ≥ 50 -bp length. Mapped reads with sequences partially
435 aligned to different regions of the λ genome were assigned as intragenomic chimeric reads, and
436 those with one sequence aligned to the λ genome and another to λ -unrelated genomes were
437 assigned as intergenomic chimeric reads.

438

439 **Calculation of PARs/CARs (PC ratio) of aligned read clusters (ARCs) mapped to the**
440 **lambda phage genome**

441 All ECLs were aligned to each other with LAST using -a 0 -A 10 -b 15 -B 7 parameters with
442 $\geq 75\%$ identity and ≥ 100 -bp alignment length to obtain ARCs in each sample. The reference
443 index of LAST was made using lastdb with -uNEAR and -R01 options. For the ARCs mapped
444 to the λ genome, the PC ratio was calculated by counting the number of PARs and CARs in
445 ARCs.

446

447 **Relationship between the PC ratio and the specificity and sensitivity of chimera detection**
448 **based on the spike-in lambda phage genome**

449 The relationship between PC ratios and the sensitivity and specificity of the assignments of
450 chimeric and non-chimeric ARCs mapped to the λ genome was examined. The sensitivity was
451 defined as $100 \times TP / (TP + FN)$, where TP and FN indicate the number of ARCs assigned as
452 chimeric ARCs and those unassigned as chimeric ACRs among the chimeric ARCs,
453 respectively. The specificity was defined by $100 \times TN / (FP + TN)$, where TN and FP indicate
454 the number of non-chimeric ARCs assigned as non-chimeric ARCs and those unassigned as
455 non-chimeric ARCs among the non-chimeric ARCs, respectively.

456

457 **Determination of chimeric position**

458 The chimeric junctions containing possible chimeric positions were mostly <25 bp in length in
459 our samples and contained alignments of multiple PARs and CARs with a few base mismatches
460 as described previously [30,31]. The chimeric position shared by multiple PARs was
461 determined as the chimeric position. When several PARs with subtly different chimeric
462 positions were present, the chimeric position of PARs with the highest read depth was selected.

463

464 **Whole process of SACRA treatment of PacBio reads**

465 First, ECLs were aligned to each other using LAST to generate ARCs. Second, the PC ratio
466 was calculated to divide ARCs with both PARs and CARs into those with greater than the
467 optimized mPC ratio and less than the optimized mPC ratio. Third, ARCs with greater than the
468 optimized mPC ratio and those with only PARs were subjected to SACRA to split them at the
469 chimeric position. Finally, SACRA-split reads, unsplit reads from ARCs with less than the
470 optimized mPC ratio and with only CARs, and singletons were combined and assembled
471 (Figure 3).

472

473 ***De novo* assembly of PacBio and MiSeq reads**

474 PALs and SALs were assembled using *canu* (v1.8) [28] with `-trim-assemble`
475 `minReadLength=1000 minOverlapLength=1000` options. All generated contigs were polished
476 by filter-passed MiSeq reads with *Pilon* [45]. Filter-passed MiSeq reads were assembled by
477 *MEGAHIT* (v1.1.4) [46] with default parameters. Assembly statistics were obtained from all
478 assembled contigs using *seqkit* [47].

479

480 **Dot-plot of lambda phage genome contigs from the assembly of FMSs, PALs, and SALs**

481 Contigs generated from the assembly of FMSs, PALs, and SALs were aligned to the λ genome
482 with *minimap2* using the `asm20` parameter. The dot-plot of the *minimap2* alignment was
483 visualized with *D-GENIES* [48].

484

485 **Alignment of FMS and SAL contigs and identification of complete and near-complete**
486 **contigs**

487 FMS and SAL contigs were aligned using *NUCmer* (v4.0.0) with $\geq 95\%$ identity in each sample
488 to obtain redundant merged contigs (MCs) with the minimum scaffolded contigs. MCs with a
489 low read-depth of < 100 mapped FMSs/kb were removed because of the low sequence accuracy.
490 FMS short contigs (≤ 500 bp) additionally aligned to the MCs were also removed from the
491 analysis. Remaining high read-depth MCs were further clustered with $\geq 99\%$ identity and length
492 coverage to obtain non-redundant MCs. The alignment results were visualized using *AliTV*
493 with the `nogapped` option of *lastz* [49]. Merged and FMS-specific contigs generating circular
494 contigs (CCs) were identified as complete contigs by the *BLAST* detection of terminal direct
495 repeats (TDRs) with $\geq 95\%$ identity and ≥ 50 -bp alignment length.

496 Chimeric paired reads generated from Pacasus-insensitive chimeric reads (NPCRs)
497 split by SACRA were mapped to contigs using minimap2 with the map-pb option and $\geq 80\%$
498 identity and coverage. The paired reads mapped to the same contig were assigned as intra-
499 NPCRs, and those mapped to different contigs were assigned as inter-NPCRs. The intra-
500 chimera rate of a contig was obtained by dividing the number of intra-NPCRs by the total
501 number of NPCRs in the contig. The fragmented λ contigs were generated by randomly
502 dividing the complete λ sequence into 10 subsequences with 20%, 50%, and 80% coverage of
503 the complete length.

504

505 **Mapping of MiSeq and PacBio reads to contigs**

506 FMSs were used for quantification of the relative abundance of contigs by mapping them to
507 contigs using Bowtie2 with $\geq 95\%$ identity in each sample, and the relative abundance was
508 calculated by dividing the number of mapped reads by the contig size. Mapping SALs to
509 contigs was individually conducted using minimap2 with the map-pb option [42] and $\geq 95\%$
510 identity and $\geq 85\%$ length coverage. Of the alignments containing PacBio reads that aligned to
511 multiple contigs, the alignment with the highest identity was selected as the top hit alignment.
512 The mapping results were visualized with IGV (v2.7.2) [44].

513

514 **Analysis of complete and near-complete contigs**

515 Virsorter (v1.0.3) with virome db and virome decontamination options in the CyVerse
516 environment [32] was used for phage assessment. Open reading frames (ORFs) were predicted
517 by prodigal with ≥ 20 amino acids [51], and a similarity search was conducted against pVOG
518 [33] using Diamond BLASTP [52] with an e-value $< 1e-5$ threshold and --sensitive option. A.
519 similarity search of the contigs against the virus and plasmid database was conducted using
520 NUCmer with $\geq 85\%$ identity and length coverage. In this study, we integrated and used

521 publicly available phage sequences from RefSeq phage genomes (release 85), complete and
522 high-quality draft viral genomes in IMG/VR v2.0 [16], crAss-like phage genomes [53],
523 *Microviridae* genomes [35], and the full-length phage genomes reconstructed from human
524 fecal samples [22]. The gene lengths of reference phage genomes were obtained from the
525 RefSeq phage genomes. The plasmid sequences were also obtained from RefSeq plasmid
526 (February 2020), PLSDB (version 2020_03_04) [54] and those from human fecal samples [22].
527 Taxonomic assignments of novel complete and near-complete phage contigs were performed
528 by a similarity search of pVOGs against those in the known phage genomes with the voting
529 system [38], in which the taxon with the highest number of matched pVOGs was assigned as
530 that of the contig. When the contig had the same number of pVOGs that hit multiple different
531 taxa, the taxon with the highest similarity was assigned as that of the contig.

532

533 **Declarations**

534 **Ethics approval and consent to participate**

535 This study was approved by the research ethics committee of RIKEN and Waseda University,
536 and written consent was obtained from all subjects.

537

538 **Consent for publication**

539 Written informed consent for publication was obtained from all subjects.

540

541 **Availability of data and materials**

542 The sequencing data of MiSeq and PacBio Sequel and circular and near-complete linear contigs
543 are available from the DNA Data Bank of Japan (DDBJ; Accession No: DRA009211). SACRA
544 used in this study is available at <https://github.com/hattori-lab/SACRA>.

545

546 **Competing interests**

547 The authors declare that they have no competing interests.

548

549 **Funding**

550 This study was supported in part by the scholarships for young doctoral students from Waseda
551 University to Y.K., the Leading Advanced Projects for Medical Innovation (LEAP) and
552 Advanced Research and Development Programs (CREST) from AMED to M.H., and RIKEN
553 Integrated Symbiology (to W.S.).

554

555 **Authors' contributions**

556 Y.K. designed the study. W.S and M.H supervised the study. Y.K. performed biological
557 experiments including virome DNA sample preparation. Y.K. and W.S. performed Illumina
558 and PacBio sequencing. Y.K., N.K., S.N., and W.S. performed computational analysis. Y.K.,
559 M.H., and W.S. wrote the manuscript, which was approved by all authors.

560

561 **Acknowledgements**

562 We thank M. Tanokura, K. Kaida, C. Shindo, K. Honda, K. Komiya, and Y. Hattori for
563 technical supports for DNA preparation, sequencing, and computational analysis.

564

565 **References**

- 566 1. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, et al. Viruses in the
567 faecal microbiota of monozygotic twins and their mothers. *Nature*. 2010;466:334–8.
- 568 2. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, et al. Metagenomic
569 Analyses of an Uncultured Viral Community from Human Feces. *J Bacteriol*.
570 2003;185:6220–6223.

- 571 3. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, et al. The
572 Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe*.
573 2019;26:527–41.
- 574 4. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI. Going viral: next-
575 generation sequencing applied to phage populations in the human gut. *Nat Rev Microbiol*.
576 2012;10:607–17.
- 577 5. Reyes A, Wu M, McNulty NP, Rohwer FL, Gordon JI. Gnotobiotic mouse model of
578 phage-bacterial host dynamics in the human gut. *Proc Natl Acad Sci USA*. 2013;110:20236–
579 20241.
- 580 6. Hsu BB, Gibson TE, Yeliseyev V, Liu Q, Lyon L, Bry L, et al. Dynamic Modulation of the
581 Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model. *Cell Host Microbe*.
582 2019;25:803–14.
- 583 7. Norman JM, Handley SA, Baldrige MT, Droit L, Liu CY, Keller BC, et al. Disease-
584 specific alterations in the enteric virome in inflammatory bowel disease. *Cell*. 2015;160:447–
585 60.
- 586 8. Nakatsu G, Zhou H, Wu WKK, Wong SH, Coker OO, Dai Z, et al. Alterations in Enteric
587 Virome Are Associated With Colorectal Cancer and Survival Outcomes. *Gastroenterology*.
588 2018;155:529–41.
- 589 9. Zhao G, Vatanen T, Droit L, Park A, Kostic AD, Poon TW, et al. Intestinal virome
590 changes precede autoimmunity in type I diabetes-susceptible children. *Proc Natl Acad Sci*
591 *USA*. 2017;114:E6166–E6175.
- 592 10. Ma Y, You X, Mai G, Tokuyasu T, Liu C. A human gut phage catalog correlates the gut
593 phageome with type 2 diabetes. *Microbiome*. 2018;6:24.
- 594 11. Zuo T, Lu XJ, Zhang Y, Cheung CP, Lam S, Zhang F, et al. Gut mucosal virome
595 alterations in ulcerative colitis. *Gut*. 2019;68:1169–79.

- 596 12. Clooney AG, Sutton TDS, Shkoporov AN, Plevy SE, Ross RP, Hill C, et al. Whole-
597 Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell*
598 *Host Microbe*. 2019;20:2083–8.
- 599 13. Manrique P, Bolduc B, Walk ST, van der Oost J, de Vos WM, Young MJ. Healthy
600 human gut phageome. *Proc Natl Acad Sci USA*. 2016;113:10400–10405.
- 601 14. Shkoporov AN, Ryan FJ, Draper LA, Forde A, Stockdale SR, Daly KM, et al.
602 Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome*.
603 2018;6:68.
- 604 15. Roux S, Solonenko NE, Dang VT, Poulos BT, Schwenck SM, Goldsmith DB, et al.
605 Towards quantitative viromics for both double-stranded and single-stranded DNA viruses.
606 *PeerJ*. 2016;4:e2777.
- 607 16. Paez-Espino D, Roux S, Chen I-MA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR
608 v.2.0: an integrated data management and analysis system for cultivated and environmental
609 viral genomes. *Nucleic Acids Res*. 2018;47:D678–D686.
- 610 17. Roux S, Emerson JB, Eloë-Fadrosch EA, Sullivan MB. Benchmarking viromics: An in
611 silico evaluation of metagenome-enabled estimates of viral community composition and
612 diversity. *PeerJ*. 2017;5:e3817.
- 613 18. Galiez C, Siebert M, Enault F, Vincent J, Söding J. WIsH: who is the host? Predicting
614 prokaryotic hosts from metagenomic phage contigs. *Bioinformatics*. 2017;33:3113–4.
- 615 19. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d2* oligonucleotide
616 frequency dissimilarity measure improves prediction of hosts from metagenomically-derived
617 viral sequences. *Nucleic Acids Res*. 2017;45:39–53.
- 618 20. Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD. Hypervariable loci in the human
619 gut virome. *Proc Natl Acad Sci USA*. 2012;109:3962–6.
- 620 21. Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, et al.

621 Metagenomic assembly through the lens of validation: recent advances in assessing and
622 improving the quality of genomes assembled from metagenomes. *Brief Bioinform.*
623 2017;20:1140–1150.

624 22. Suzuki Y, Nishijima S, Furuta Y, Yoshimura J, Suda W, Oshima K, et al. Long-read
625 metagenomic exploration of extrachromosomal mobile genetic elements in the human gut.
626 *Microbiome.* 2019;7:119.

627 23. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, et al.
628 Long-read viral metagenomics enables capture of abundant and microdiverse viral
629 populations and their niche-defining genomic islands. *PeerJ.* 2019;7:e6800.

630 24. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang X-S, Davis-Richardson A, et al.
631 Metagenomic binning and association of plasmids with bacterial host genomes using DNA
632 methylation. *Nat Biotechnol.* 2017;36:61–69.

633 25. Bertrand D, Shaw J, Kalathiappan M, Ng AHQ, Muthiah S, Li C, et al. Hybrid
634 metagenomic assembly enables high-resolution analysis of resistance determinants and
635 mobile elements in human microbiomes. *Nat Biotechnol.* 2019;37:937–944.

636 26. Moss EL, Maghini DG, Bhatt AS. Complete , closed bacterial genomes from
637 microbiomes using nanopore sequencing. *Nat Biotechnol.* 2020;

638 27. Stepanauskas R, Fergusson EA, Brown J, Poulton NJ, Tupper B, Labonté JM, et al.
639 Improved genome recovery and integrated cell-size analyses of individual uncultured
640 microbial cells and viral particles. *Nat Commun.* 2017;8:84.

641 28. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable
642 and accurate long-read assembly via adaptive k-mer weighting and repeat separation.
643 *Genome Res.* 2016;27:722–736.

644 29. Warris S, Schijlen E, van de Geest H, Vegesna R, Hesselink T, Te Lintel Hekkert B, et al.
645 Correcting palindromes in long reads after whole-genome amplification. *BMC Genomics.*

646 2018;19:798.

647 30. Lasken RS, Stockwell TB. Mechanism of chimera formation during the Multiple
648 Displacement Amplification reaction. *BMC Biotechnol.* 2007;7:19.

649 31. Tu J, Guo J, Li J, Gao S, Yao B, Lu Z. Systematic characteristic exploration of the
650 chimeras generated in multiple displacement amplification through next generation
651 sequencing data reanalysis. *PLoS One.* 2015;10:e0139857.

652 32. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from
653 microbial genomic data. *PeerJ.* 2015;3:e985.

654 33. Graziotin AL, Koonin E V., Kristensen DM. Prokaryotic Virus Orthologous Groups
655 (pVOGs): A resource for comparative genomics and protein family annotation. *Nucleic
656 Acids Res.* 2017;45:D491–8.

657 34. Parras-moltó M, Rodríguez-galet A, Suárez-rodríguez P, López-bueno A. Evaluation of
658 bias induced by viral enrichment and random amplification protocols in metagenomic
659 surveys of saliva DNA viruses. *Microbiome.* 2018;6:119.

660 35. Roux S, Krupovic M, Poulet A, Debroas D, Enault F. Evolution and diversity of the
661 microviridae viral family through a collection of 81 new complete genomes assembled from
662 virome reads. *PLoS One.* 2012;7:e40418.

663 36. Rohwer F, Edwards R. The phage proteomic tree: A genome-based taxonomy for phage.
664 *J Bacteriol.* 2002;184:4529–35.

665 37. Jang H Bin, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al.
666 Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing
667 networks. *Nat Biotechnol.* 2019;37:632–639.

668 38. Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. Rapid evolution of the
669 human gut virome. *Proc Natl Acad Sci U S A.* 2013;110:12450–5.

670 39. Nishijima S, Suda W, Oshima K, Kim S-W, Hirose Y, Morita H, et al. The gut

671 microbiome of healthy Japanese and its microbial and functional uniqueness. *DNA Res.*
672 2016;23:125–133.

673 40. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets.
674 *Bioinformatics.* 2011;27:863–4.

675 41. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.*
676 2012;9:357–9.

677 42. Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics.*
678 2018;34:3094–100.

679 43. Wan R, Sato K, Kielbasa SM, Frith MC, Horton P. Adaptive seeds tame genomic
680 sequence comparison. *Genome Res.* 2011;21:487–93.

681 44. Hamada M, Ono Y, Asai K, Frith MC, Hancock J. Training alignment parameters for
682 arbitrary sequencers with LAST-TRAIN. *Bioinformatics.* 2017;33:926–8.

683 45. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An
684 integrated tool for comprehensive microbial variant detection and genome assembly
685 improvement. *PLoS One.* 2014;9.

686 46. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node
687 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
688 *Bioinformatics.* 2015;31:1674–6.

689 47. Shen W, Le S, Li Y, Hu F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q
690 file manipulation. *PLoS One.* 2016;11:1–10.

691 48. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient
692 and simple way. *PeerJ.* 2018;6:e4958.

693 49. Ankenbrand MJ, Hohlfield S, Hackl T, Förster F. AliTV-interactive visualization of whole
694 genome comparisons. *PeerJ Comput Sci.* 2017;3:e116.

695 50. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al.

696 Integrative genomics viewer. *Nat Biotechnol.* 2011;29:24–6.

697 51. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal:
698 Prokaryotic gene recognition and translation initiation site identification. *BMC*
699 *Bioinformatics.* 2010;11:119.

700 52. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND.
701 *Nat Methods.* 2014;12:59–60.

702 53. Guerin E, Shkoporov A, Stockdale S, Clooney AG, Ryan FJ, Draper LA, et al. Biology
703 and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell*
704 *Host Microbe.* 2018;24:653–64.

705 54. Galata V, Fehlmann T, Backes C, Keller A. PLSDB: A resource of complete bacterial
706 plasmids. *Nucleic Acids Res.* 2019;47:D195–202.

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721 **Figure legends**

722

723 **Figure 1 | The average chimera ratio based on the spike-in lambda phage genome.**

724 The average chimera ratio of PacBio and MiSeq reads in metagenomic sequencing of the
725 multiple displacement amplification (MDA)-treated human gut virome DNA is shown. FMSs,
726 ECLs, PALs, and SALs indicate filter-passed MiSeq short reads, error-corrected PacBio long
727 reads, Pacasus-treated ECLs, and SACRA-treated ECLs, respectively. Error bars represent the
728 standard deviation (SD).

729

730 **Figure 2 | Development of SACRA correction for PacBio chimeric reads.**

731 **a**, Illustration of chimeric and non-chimeric alignment read clusters (ARCs). The aligned and
732 unaligned sequences to chimeric and non-chimeric ARCs are shown by solid and dashed lines,
733 respectively. Reads composed of red and blue lines are chimeric, and those only with red or
734 blue lines are non-chimeric. Continuously aligned reads (CARs) spanning the chimeric
735 positions and partially aligned reads (PARs) are shown. Vertical dashed lines indicate the
736 candidate chimeric positions based on PARs. **b**, Relationship between the specificity and
737 sensitivity for detection of non-chimeric and chimeric ARCs and the minimum PC ratio in each
738 sample.

739

740 **Figure 3 | Whole process of SACRA treatment of PacBio reads.**

741 **STEP1.** Construction of aligned read clusters (ARCs) from error-corrected PacBio reads by
742 pairwise local alignment with the LAST aligner. The generated ARCs include ARCs-1 with
743 both CARs and PARs, ARCs-2 only with PARs, and ARCs-3 only with CARs, and singleton
744 reads (singletons) unaligned with any other read. **STEP2.** The PC ratio (PARs/CARs) of all
745 ARCs-1 is calculated to determine the optimized minimum PC (mPC) ratio, with which ARCs-

746 1 are divided into those \geq mPC and $<$ mPC ratio. **STEP3.** ARCs-1 \geq mPC ratio and ARCs-2 are
747 split by SACRA. Finally, all the split and unsplit reads are combined for assembly.

748

749 **Figure 4| Statistics of *de novo* assembly of FMSs, PALs, and SALs.**

750 **a**, N50 contig length in assembly of FMSs, PALs, and SALs. FMSs, PALs, and SALs indicate
751 filter-passed MiSeq short reads, Pacasus-treated ECLs, and SACRA-treated ECLs,
752 respectively. Boxes represent the inter-quartile range (IQR) and lines inside the box indicate
753 the median. \times represents the average. Whiskers show the $1.5\times$ IQR. Points represent outliers.

754 **b**, Ratio of contig-length distribution in assembly of FMSs, PALs, and SALs in the five samples.

755

756 **Figure 5 | Comparison of FMS, SAL, and merged contigs (MCs).**

757 **a**, The average number of the minimum scaffolded FMS and SAL contigs, and short
758 fragmented FMS contigs aligned to the MCs in the five samples. FMSs and SALs indicate
759 filter-passed MiSeq short reads and SACRA-treated ECLs, respectively. **b**, The average contig
760 length of FMS, SAL, and MC. Error bars represent the SD. **c**, Proportion of fragmentation
761 types of FMS contigs.

762

763 **Figure 6 | Statistics of ORFs in FMS, SAL, and 142 complete and near-complete phage**
764 **contigs.**

765 **a**, Distribution of the number of open reading frames (ORFs; ≥ 20 aa) per SAL, FMS, and
766 merged contig in the five samples. FMSs and SALs indicate filter-passed MiSeq short reads
767 and SACRA-treated ECLs, respectively. **b**, Comparison of ORF length (aa) in complete and
768 near-complete contigs (CCs & nLCs) in the five samples and the reference complete phage
769 genomes (left two data), as well as between FMS and SAL contigs in each sample. Boxes
770 represent the inter-quartile range (IQR). Lines inside the box indicate the median. Whiskers

771 show the 1.5× IQR. For visualization, outliers are not shown.

772

773 **Supplementary Figure 1 | Sequence similarity of the filter-passed and error-corrected**
774 **PacBio reads with corresponding Miseq contigs.**

775 **a**, Boxes represent the inter-quartile range (IQR) and lines inside the box indicate the median.

776 × represents the average. Whiskers show the 1.5× IQR. Points represent outliers.

777

778 **Supplementary Figure 2 | Normalized chimera ratio by read-length.**

779 **a**, Average chimera ratio normalized by read-length in the spike-in lambda phage genome.

780 Error bars represent the SD.

781

782 **Supplementary Figure 3 | Characterization of Pacasus-insensitive chimeric reads**
783 **(NPCRs).**

784 **a**, Structure of non-palindromic intragenomic (intra-NPCRs) and intergenomic chimeric reads

785 (inter-NPCRs) on the spike-in lambda phage genome. **b**, An example comparison of the

786 position (x-axis) and frequency (y-axis) of chimeric positions every 20 bp on the lambda phage

787 genome between technical replicate 1 and 2 of sample S1.

788

789 **Supplementary Figure 4 | Characteristics of chimeric and non-chimeric ARCs and ARC**
790 **PC ratio on spike-in lambda phage genome.**

791 **a**, Ratio of chimeric and non-chimeric ARCs in each sample. **b**, PC ratio of ARCs in each

792 sample. Boxes represent the inter-quartile range (IQR) and lines inside the box indicate the

793 median. Whiskers show the 1.5× IQR. Points represent outliers.

794

795 **Supplementary Figure 5 | Statistics of FMSs, ECLs, PALs, and SALs.**

796 **a**, Chimera ratio (%) of FMSs, ECLs, PALs, and SALs on the spike-in lambda phage genome
797 in each sample. **b**, The ratio (%) of the corrected Pacasus-sensitive chimeric reads based on
798 SACRA in each sample. **c**, An example of the read-length distribution of ECLs, PALs, and
799 SALs in sample S3. **d**, Average ratio of the total bases of PALs and SALs (≥ 1 kb) to that of
800 ECLs (≥ 1 kb). Error bars represent SD.

801

802 **Supplementary Figure 6 | Comparison of the spike-in lambda phage genomes**
803 **reconstructed from FMSs, PALs, and SALs.**

804 Dot-plots indicate the similarity at $\geq 95\%$ identity between the reconstructed lambda phage
805 genomes. The x-axis indicates the consensus sequence of lambda phage genomes reconstructed
806 from SALs in the five samples, almost identical to the reference, and the y-axis indicates the
807 sequences reconstructed from FMSs, PALs, and two SALs with an optimized mPC ratio and
808 mPC ratio of zero, respectively. Horizontal lines in dot-plots show the gaps in the reconstructed
809 contigs, and the total contig number (#contigs) is shown on the top-left of each dot-plot.

810

811 **Supplementary Figure 7 | Examples for FMS and SAL contig fragmentations.**

812 Four types of contig fragmentations are exemplified. Upper red and blue bars show the forward
813 and reverse sequences of SALs and FMSs mapped to the contigs, respectively. **a**, Type 1
814 fragmentation in FMS contigs. Dot-plot shows self-alignment of the MC, and a local repeat
815 region is boxed. **b**, Type 2 fragmentation in the FMS contig. Five different fragmented FMS
816 contigs including mosaic sequences caused by two homologous genomes are shown. The
817 degree of similarity in lastz alignments is shown on the right. **c**, Type 3 fragmentation in FMS
818 contigs. The middle vertical lines in grey indicate changes in the read-depth of FMSs across
819 the MC, in which base variations are colored. **d**, Type 4 fragmentation in SAL contigs. The
820 middle vertical lines in grey indicate changes in the read-depth of FMSs across the MC, in

821 which a region with the extreme low read-depth is marked.

822

823 **Supplementary Figure 8 | Comparison of the intra-chimera rate of lambda contigs.**

824 **a**, The intra-chimera (intra-NPCRs) rate of fragmented lambda contigs with completeness from

825 20% to 100% of the full-length genome. Boxes represent the inter-quartile range (IQR) and

826 lines inside the box indicate the median. Whiskers show the 1.5× IQR. Points represent outliers.

827 **p*-value <0.01 NS, no significance; Wilcoxon rank sum test with Bonferroni correction.

828

829 **Supplementary Figure 9 | Family-level taxonomic assignment of 142 complete and near-**

830 **complete phage contigs.**

831