Viromes of one year old infants reveal the impact of birth mode on microbiome diversity

Angela McCann [1,2], Feargal J Ryan [1,2], Stephen R Stockdale [1,2,3], Marion Dalmasso [1], Tony Blake [1,2], C. Anthony Ryan [1,4], Catherine Stanton [1,2], Susan Mills [1,2], Paul R Ross [1,2,3], Colin Hill [Corresp. 1,2]

[1] APC Microbiome Institute, Cork, Cork, Ireland

[2] School of Microbiology, University College Cork, Cork, Cork, Ireland

[3] Teagasc Food Research Centre, Fermoy, Cork, Ireland

[4] Department of Neonatology, Cork University Maternity Hospital, Cork, Cork, Ireland

Corresponding Author: Colin Hill
Email address: c.hill@ucc.ie

Establishing a diverse gut microbiota after birth is being increasingly recognised as important for preventing illnesses later in life. It is well established that bacterial diversity rapidly increases post-partum; however, few studies have examined the infant gut virome/phageome during this developmental period. We performed a metagenomic analysis of 20 infant faecal viromes at 1 year of age to determine whether spontaneous vaginal delivery (SVD) or caesarean section (CS) influenced viral composition. We find that birth mode results in distinctly different viral communities, with SVD infants having greater viral and bacteriophage diversity. We demonstrate that CrAssphage is acquired early in life, both in this cohort and two others, although no difference in birth mode is detected. A previous study has shown that bacterial OTU's (operational taxonomic units) identified in the same infants could not discriminate between birth mode at 12 months of age. Therefore, our results indicate that vertical transmission of viral communities from mother to child may play a role in shaping the early life microbiome, and that birth mode should be considered when studying the early life gut virome.

1   **Viromes of one year old infants reveal the impact of birth mode on microbiome diversity**

2   Angela McCann*[1,3], Feargal J. Ryan*[1,3], Stephen R. Stockdale*[1,2,3], Marion Dalmasso[1,3,†], Tony

3   Blake[1,3], C. Anthony Ryan[1,4],Catherine Stanton[1,3], Susan Mills[1,3], Paul R. Ross[1,2,3], Colin Hill[1,3,‡]

4   [1]APC Microbiome Institute, University College Cork, Co. Cork, Ireland

5   [2]Teagasc Food Research Centre, Moorepark, Fermoy, Co. Cork, Ireland

6   [3]School of Microbiology, University College Cork, Cork, Ireland

7   [4]Department of Neonatology, Cork University Maternity Hospital, Co. Cork, Ireland

8   †Present address: Normandie Univ, UNICAEN, ABTE, 14000 Caen, France

9   *These authors contributed equally to this work.

10   Corresponding author: Colin Hill

11   c.hill@ucc.ie

19 **Abstract**

20   Establishing a diverse gut microbiota after birth is being increasingly recognised as

21 important for preventing illnesses later in life. It is well established that bacterial diversity rapidly

22 increases post-partum; however, few studies have examined the infant gut virome/phageome

23 during this developmental period. We performed a deep-sequencing metagenomic analysis of 20

24 infant faecal viromes at 1 year of age to determine whether spontaneous vaginal delivery (SVD)

25 or caesarean section (CS) influenced viral composition. We find that birth mode results in

26 distinctly different viral communities, with SVD infants having greater viral and bacteriophage

27 diversity. We demonstrate that CrAssphage is acquired early in life, both in this cohort and two

28 previously published studies of the infant virome, although no difference in birth mode is detected.

29 A previous study has shown that bacterial OTU's (operational taxonomic units) identified in the

30 same infants could not discriminate between birth mode at 12 months of age. In conclusion, our

31 results indicate that vertical transmission of viral communities from mother to child may play a

32 role in shaping the early life microbiome, and that birth mode should be considered when studying

33 the early life gut virome.

34

35 **Introduction**

36   The human gut microbiota is a diverse community densely populated with bacteria,

37 archaea, protists, fungi, and viruses. Studies focused on gut bacteria suggest that healthy

38 individuals are characterised by high species diversity (Heiman & Greenway, 2016), with

39 compositional alterations and decreased diversity linked to conditions such as obesity, diabetes

40 and inflammatory bowel disease (IBD) (Imhann et al 2016, Karlsson et al 2013, Ley et al 2005).

41   Gut microbiota colonization in infants is a critical process, characterised by initial low bacterial

42   diversity which increases with time such that by 1 year of age the microbiota converges towards

43   that of an adult and fully resembles an adult microbiota by 2-5 years of age (Rodriguez et al 2015).

44   Several factors have been shown to influence an infant's microbiota, from birth mode to antibiotic

45   usage, diet, geographical location, lifestyle and age (Milani et al 2017, Rodriguez et al 2015).

46   Indeed, Hill *et al.* confirmed that delivery mode and gestational age significantly influence

47   bacterial composition in the infant gut during the first 24 weeks of life (Hill et al 2017).

48   The gut virome is an area of growing interest with relation to the microbiota (Breitbart et

49   al 2003, Minot et al 2013) and gut virome alterations have been recorded between healthy and

50   diseased states; for instance, an increase in the taxonomic richness of *Caudovirales* has been

51   associated with IBD (Norman et al 2015). However, there is a significant knowledge gap about

52   healthy human viral populations, with large portions of the sequence data from human virome

53   metagenomic studies representing uncharacterised viruses either not present in current databases

54   or described using in silico methods without host information or taxonomic assignment

55   (Krishnamurthy and Wang 2017). Reyes *et al.* (2010) demonstrated intrapersonal virome variation

56   between adult twins over a one year period was low, while interpersonal variation was high (Reyes

57   et al 2010). Manrique *et al* (2016) demonstrated the presence of a healthy human phageome, which

58   is a collection a bacteriophage which are present in a large portion of healthy individuals and

59   hypothesized that this community plays a key role in the structure of the human microbiome and

60   by extension human health. Research on the gut virome in infancy and early life (0 – 3 years) thus

61   far has exclusively focused on longitudinal studies in twin pairs (Lim et al 2015, Reyes et al 2015)

62   or has been based on a single infant (Breitbart et al. 2008). Reyes *et al.* (2015) identified different

63   viral assembly stages of the gut microbiota in 20 healthy infant twin pairs (0-3 years) and revealed

64 that this program of assembly is impaired in twins discordant for severe acute malnutrition (Reyes

65 et al 2015). Lim *et al.* (2015) conducted a longitudinal study of the virome and bacterial

66 microbiome in 4 twin pairs, from birth to 2 years, and revealed that the expansion of the bacterial

67 microbiome with age was accompanied by a contraction and shift in the bacteriophage composition

68 (Lim et al 2015). The work conducted by Lim *et al*. was unsuited to examining the impact of birth

69 mode as only a single twin pair in that study was born by standard vaginal delivery. Reyes *et al*

70 did not report on the birth mode of the infants in their study; however, as many of the twin pairs

71 in that study were discordant for forms of malnutrition and received a dietary intervention as a

72 result, measuring the impact of birth mode in those infants would be complicated by confounding

73 factors. Thus, to date, no investigation has had a study design allowing for direct investigation of

74 the impact of birth mode on diversity and composition of the virome in early life but birth mode

75 has been proposed as a putative modulator of microbiome diversity (Milani et al 2017). Therefore,

76 we performed a deep-sequencing metagenomic examination of faecal DNA viromes of 20 infants

77 at 1 year of age and investigated whether spontaneous vaginal delivery (SVD) or caesarean section

78 (CS) influenced gut virome composition and diversity.

79

80 **Materials & Methods**

81 **Selection of faecal samples**

82 Faecal samples for all infants were collected as part of the INFANTMET (Hill et al 2017) study.

83 Infant guardians were approached for written consent between February 2012 and May 2014 in

84 Cork University Maternity Hospital, with ethical approval provided by the Cork University

85 Hospital Research Ethics Committee. Ethical approval reference: ECM (w) 07/02/2012. In order

86 to control for potential variants, faecal samples were randomly chosen from those available that

87 best met the following criteria: (1) an equal number of Spontaneous Vaginal Delivered (SVD) and

88 emergency Caesarean-Section (CS) infants, (2) gestational-term matched infants, (3) age matched

89 infants, and (4) a balanced number of breastfed versus bottle-fed infants from SVD and CS

90 available samples. In addition, technical criteria included the availability of >1g of starting faecal

91 material. As a result, 20 infant faecal samples were selected; 10 were from SVD infants and 10

92 were from CS infants. Of the 10 infant faecal samples per cohort, 7 and 8 of the CS and SVD

93 delivered infants, respectively, were breastfed. For details related to samples chosen in this study,

94 see Supplementary Table 1.

95 **Preparation of faecal viral suspensions**

96 Viruses were separated from faecal solids using the following method. Faeces (0.5 g) was

97 suspended in 10 ml of SM buffer (50 mM Tris-HCl; 100 mM NaCl; 8.5 mM MgSO$_4$; pH 7.5).

98 Samples were homogenised by vortexing for 5 min, before centrifuging twice at 4,075 x g for 10

99 min at 4 ºC in a swing-bucket centrifuge to remove large particulates and bacterial cells. Faecal

100 viral suspensions were filtered twice through a 0.45 µm pore diameter filter and processed

101 immediately for DNA extraction.

102 **Extraction of viral DNA**

103 Preparation of viral suspensions and DNA extractions were optimised for the small faecal samples

104 collected from infant adsorbent nappies. To viral suspensions, NaCl (final conc. 0.5 M) and 10 %

105 (w/v) polyethylene glycol (PEG; average molecular weight 8000) were dissolved before samples

106 were chilled on ice for 3 hrs. Viruses were then precipitated from solution in a 4 ºC pre-chilled

107 centrifuge at 4, 075 x g in a swing-bucket centrifuge for 20 min. The viral-PEG pellet was

108 suspended in 400 µL of SM buffer, viruses were then separated from the PEG by treating the

109 samples with an equal volume of chloroform, vortexing for 30 sec, and centrifuging at 2,500 x g.

110 Clarified viral preparations were treated with 20 U of DNase I and 10 U of RNase I (final

111 concentrations; Ambion) for 1 hr at 37 ^{o}C, after the addition of 40 µL of 10x Nuclease Buffer (50

112 mM $CaCl_2$; 10 mM $MgCl_2$). Nucleases were inactivated at 70 ^{o}C for 10 min before samples were

113 treated with 20 µL of 10 % SDS and 2 µL of freshly prepared 20 mg/ml Proteinase K for 20 min

114 at 56 ^{o}C. Remaining intact viruses were lysed by the addition of 100 µL of Phage Lysis Buffer (4.5

115 M guanidine thiocyanate; 45 mM sodium citrate; 250 mM sodium lauroyl sarcosinate; 562.5 mM

116 β-mercaptoethanol; pH 7.0) with incubation at 65 ^{o}C for 10 min. Viral DNA was purified by two

117 treatments with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) and passing the

118 resulting purified DNA through a QIAGEN Blood and Tissue Purification Kit and eluting samples

119 in 50 µL of TE Buffer.

120 **Viral DNA amplification, library preparation and sequencing**

121 Infant faecal viral DNA concentrations were equalised before amplification for sequencing using

122 an Illustra GenomiPhi V2 kit (GE Healthcare). Amplifications of purified viral DNA were

123 performed in triplicate on all samples as described by the manufacturer. Subsequently, an equal

124 volume of each amplification and an equal volume of the original viral DNA purification were

125 pooled together for paired-end Nextera XT library preparation (Illumina) as described by the

126 manufacturer. Metagenomic sequencing of stool filtrates was performed using the Illumina MiSeq

127 (Illumina Inc. U.S.A) by generating 300 bp paired-end read libraries following the manufacturer's

128 instructions.

129 **Torque Teno Virus qPCR detection**

130   Voided infant faeces were suspended in 1:20 (w/v) SM buffer, centrifuged twice at 4,075 x g for

131   10 min at 4 °C, before filtering twice through 0.45 μm pore diameter filters. A 200 μL aliquot of

132   the faecal viral-enriched suspension was lysed using a QIAGEN Blood and Tissue Purification Kit

133   following the manufacturer's recommendations with elution in 50 μL of TE Buffer. The

134   concentration of double stranded viral DNA was calculated using a Qubit 3.0 Flurometer (Life

135   Technologies) using a Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher). Subsequently,

136   all dsDNA concentrations were normalised to 0.05 ng/μl. The choice of primers and conditions

137   for detecting pan-human associated TTV by qPCR were as described by Ssemadaali *et al.* (2016)

138   , using SensiFAST SYBR No-ROX mastermix and a LightCycler 480 thermocycler. A two-fold

139   serial dilution of the purified TTV PCR product was also included in the qPCR for standard curve

140   analysis.

141

142   **Analysis of virome sequencing data**

143   **Metagenomic analysis**

144   The quality of the raw reads was visualized with FastQC v0.11.3. Nextera adapters were removed

145   with Cutadapt v1.9.1 (Martin 2011) followed by read trimming and filtering with Trimmomatic

146   v0.36 (Bolger et al 2014) to ensure a minimum length of 60, maximum length of 150, and a sliding

147   window that cuts a read once the average quality in a window size of 4 falls below a Phred score

148   of 30. Levels of Bacterial contamination were estimated by classifying reads with SortMeRNA

149   v2.0 (Kopylova et al 2012) against the SILVA database and by aligning reads against the cpn60db

150   (Hill et al 2004) with bowtie2 in end-to-end alignment mode (Langmead and Salzberg 2012).

151   Reads were then assembled with the metaSPAdes assembler (Nurk et al 2017).

152 Virome sequence reads were classified into known viral orders and families using the Kaiju

153 metagenomic classifier (Menzel et al 2016) and the NCBI non-redundant protein database (NCBI

154 Resource Coordinators, 2018). The number of *Torque Teno virus* (TTV) homologues was counted

155 by predicting genes from all contigs with Prodigal (Hyatt et al 2010), and then by BLASTp search

156 against ORF1 from known TTV genomes (Hsiao et al 2016). Prototypical crAssphage was

157 downloaded from GenBank using accession number NC_024711.1 (Dutilh et al 2014). Assembled

158 contigs, in this study and from the Reyes *et al.* dataset, with similarity to crAssphage were detected

159 by BLAST homology (Reyes *et al.*, 2015). Complete, or near complete, crAss-like genomes (96-

160 99kb) were compared using the 'Pyani' program (https://github.com/widdowquinn/pyani),

161 implementing the ANIm method with a 500bp window size. The pyani percentage identity

162 comparison calculations were exported to R and graphed using the gplot 'heatmap.2' package.

163 GenBank files of the 6 crAss-like phages were generated and used to visualise whole genome

164 comparisons by EasyFig v2.2.2 (Sullivan et al 2011), using a minimum Blast length of 50bp and

165 identity of 30bp.

166 **Statistical analyses**

167 16S OTU tables from the INFANTMET (Hill et al 2017) cohorts were obtained and used in this

168 study to examine the connection between the virome and the bacteriome. In order to account for

169 partially assembled viruses, abundances were correlated and those with a Spearman correlation of

170 greater than 0.9 were grouped into a single feature. All statistical analyses were performed in R

171 v3.3.0 (Team 2000). Alpha diversity metrics including Chao1 richness and Shannon index were

172 computed with PhyloSeq v1.16.2 (McMurdie and Holmes 2013) and plotted with ggplot2 v2.2.1

173 (Wickham 2016). Between-group differences in alpha diversity were tested with a Mann-Whitney

174 test (also known as a two sample Wilcoxon test). Unweighted Bray-Curtis distance was used as

175   input for a Principle Coordinates Analysis (PCoA) as performed by the pcoa function in the ape

176   package v4.1. Adonis tests were performed using the vegan package v2.4.3 (Oksanen et al 2007)

177   in to test community level differences. Differential abundance analyses for both virome and 16S

178   rRNA datasets was carried out with DESeq2 (Love et al 2014) based upon the previous reporting

179   that it has increased sensitivity on datasets with less than 20 samples per group (Weiss et al 2017).

180

Results & Discussion

182   Sequencing resulted in a mean of 924,917 paired end reads per sample, which dropped to

183   697,558 following strict quality control, making this the deepest sequenced infant virome dataset

184   to date (Lim et al 2015, Reyes et al 2015). Paired end sequence reads were classified against the

185   nr database from NCBI using Kaiju (Menzel et al 2016) which translates reads into 6 possible read

186   frames for classification based on amino acid homology (Fig 1a). As with previous published

187   findings, a large portion of sequence reads from the viromes could not be classified to any known

188   viral taxonomic group (Norman et al 2015, Reyes et al 2015), with a mean 46.59 % unclassified

189   reads per sample in this cohort (Fig 1a). For viruses which were classifiable, it was only possible

190   to do so at higher taxonomic ranks. The most abundant viruses detected were *Caudovirales*,

191   *Microviridae* and *Anelloviridae* (Fig 1a), in agreement with previously published published

192   findings (Lim et al 2015). The number of sequence reads classifiable as *Anelloviridae*, a family of

193   single stranded DNA vertebrate viruses, showed a large difference between birth modes (Fig 1a,

194   Wilcox test, p = 0.02). The *Anelloviridae* and specifically their type species, Torque Teno Virus

195   (TTV), have been characterised by a very high prevalence in humans worldwide, although their

196   host interaction remains poorly understood (Spandole et al 2015). Previous research of the infant

197   virome to date have described the *Anelloviridae* as important but variable members of the gut

198   virome in the first years of life. Lim *et al* reported *Anelloviridae* peak in abundance between 6 and

199   12 months of age and that infants harbour multiple *Anelloviridae* species (Lim et al 2015). Reyes

200   *et al* reported that the abundance of the *Anelloviridae* decreases after 15 months of age and that

201   members of this family were able to discriminate twin pairs discordant for malnutrition (Reyes et

202   al 2015). Further investigation of *Anelloviridae* in this cohort found that the richness of TTV was

203   significantly increased in infants delivered by vaginal birth (Fig 1b), but not by breastfeeding status

204   (Fig S1). Similar observations of vertical transmission of TTV have previously been reported,

205   although it has been unclear whether this transmission occurs in the birth canal or in the post-

206   partum period through mother-infant contact such as breast feeding (Tyschik et al 2017).

207   Transmission of TTV to infants could occur at any point during their development through

208   environmental exposure, contact with other infants or parental contact. Given the ubiquity of

209   *Anelloviridae* throughout humans (Spandole et al 2015) it seems likely that transmission can

210   happen through multiple routes and forms of contact but based on the difference observed here it

211   would suggest that vertical transmission is one such route. As multiple displacement amplification

212   is known to distort the abundance of ssDNA viruses such as TTV (Roux et al 2016), we sought to

213   verify these results using quantitative PCR on the unamplified DNA from the infant faecal

214   samples. However, due to limited sample material this was possible for only 50% of the samples

215   (Table S2). The abundance of detected TTV DNA was found to be significantly higher in the SVD

216   cohort over the CS group (Wilcox test, $p = 0.048$) with no difference detected by breastfeeding

217   status (Wilcox test, $p = 0.49$).

218   CrAssphage is a highly abundant constituent of the human gut microbiome (Dutilh et al

219   2015) and has previously been suggested as not present in the early life microbiome (Lim et al

220 2015). However, we recovered several complete crAssphage genomes from infants, both in this

221 cohort and from virome assemblies examining the gut virome of Malawian infants (Fig 1c,

222 Supplementary figure 3) (Reyes *et al.* 2015). These crAssphage genomes showed high levels of

223 nucleotide homology and synteny to the prototypical crAssphage genome as originally described

224 by Dutilh *et al* with average nucleotide identity between all six crass genomes here between 95%

225 and 97% (Fig 1c, Fig S2). The impact of this highly abundant and prevalent bacteriophage on the

226 stability of the gut microbiota is thus far unknown, but crAssphage was recently described as just

227 one member of a previously unknown but expansive bacteriophage family (Yutin et al 2017).

228 CrAssphage is thought to predate on bacteria within the phylum Bacteroidetes (Dutilh et al 2014,

229 Yutin et al 2017), which is a constituent of the infant microbiome from as early as one week after

230 birth (Hill et al 2017).

231 In order to assess the full diversity of the DNA virome, further analysis was based on the

232 abundance of contigs assembled with MetaSPAdes. Estimates of 16S rRNA and 60 kDa chaperone

233 protein (cpn60), two commonly used bacterial phylogenetic markers, and comparison to shotgun

234 metagenomic samples from the Human Microbiome Project indicated that all samples contained

235 low levels of bacterial contamination (Fig S3 and Table S3). However in order to avoid any

236 bacterial sequence being considered for analysis only contigs which passed the VirSorter virome

237 "decontamination" mode (Roux et al 2015), contained genes that corresponded to at least one

238 known Prokaryotic Virus Orthologous Groups (pVOGs) (Grazziotin et al 2016) or showed

239 nucleotide homology to a known virus in the nt database (Coordinators 2016) were used for further

240 analysis. This resulted in a total of 2028 assembled contigs (Table S4) being taken forward for

241 analysis, out of a possible 5629, which recruited a median of 64.075% of reads per sample. There

242 was no difference detected in the percentage of reads recruited between birth mode groups (Wilcox

243   test, p = 0.57). Close to half of these assemblies (925, 45.1%) bore no homology at any length to

244   anything in the nt database with an E-value cutoff of 1e-5, highlighting the lack of viral

245   representation in current databases (Krishnamurthy and Wang 2017). The largest assembled

246   sequence included in the analysis a 146kb which had a best hit of 126 bases at 88.1 percent identity

247   to a region of Lachnoclostridium sp. YL32 annotated as a transfer RNA. Of the 2028 included

248   contigs only 36 were detected as circular by VirSorter (Table S4). Alpha and beta diversity

249   analyses identified significant differences between infant viromes by birth mode (Figure 2a,c, Fig

250   S4a). Differences in bacterial diversity at 1 year of age was not observed with the 16S rRNA

251   sequencing data (Figure 2b,d, Fig S4b). The lack of taxonomic resolution with the 16S rRNA gene

252   possibly masks diversity differences at the species, or strain level which may only be observable

253   through shotgun metagenomic sequencing (Yarza et al 2014).

254   No single virus, or viral taxon, was identified as being universally absent in CS and

255   universally present in SVD. However, DESeq2 did identify 32 contigs differentially abundant by

256   birth mode, including TTV and several contigs bearing high levels of nucleotide homology to

257   *Bifidobacteria* temperate phages including those from *Bifidobacterium longum* subsp. *infantis* and

258   subsp. *longum* (Table S5) being increased in infants born by SVD. This may be reflective of

259   differential colonisation of *Bifidobacterium* by birth mode, an observation which is supported by

260   16S rRNA sequence based studies (Hill et al 2017). Only 5 of the differentially abundant contigs

261   were significantly increased in CS relative to SVD, none of which showed high enough levels of

262   homology to reliably infer their taxonomy or host (Table S4 & Table S5).

263

264   **Conclusion**

265   Birth mode has been established to impact the microbiome but the exact mechanism or

266   duration of the impact has yet to be established. Here we observe a strong correlation between

267   birth mode and diversity of the gut virome at one year of age. This may indicate that vertical

268   transmission of viral communities may help shape the early life microbiome. In theory, the ability

269   of the virome to predate on bacterial hosts could increase bacterial diversity, and thus assist overall

270   community fitness. However, before causation can be established this phenomena will need to be

271   characterized both in animal models and in larger human cohorts incorporating longitudinal

272   sample collection. Future studies of gut virome composition and diversity in the first years of life

273   should also consider birth mode as a potential confounding factor.

274   **Data availability**

275   The raw sequence data has been deposited in the NCBI Sequence Read Archive under the

276   accession number SRP106048. Accession numbers for each individual subject are available in

277   Table S1. FastQC reports, R code, assembled sequences, BLASTn output against the nt database,

278   taxonomic assignments 16S sequences and count tables for both virome and 16S rRNA data were

279   deposited in FigShare and are available at https://figshare.com/s/3020d1009f59db059732.

280

281

282   **References**

283   Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence
284   data. *Bioinformatics* **30:** 2114-2120.

285

286   Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F (2003).
287   Metagenomic analyses of an uncultured viral community from human feces. Journal of
288   bacteriology 185: 6220-6223.

289

Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, Felts B, Mahaffy JM, Mueller J, Nulton J, Rayhawk S, Rodriguez-Brito B, Salamon P, Rohwer F (2008). Viral diversity and dynamics in an infant gut. Research in Microbiology 159: 367-373.

293

Coordinators NR (2016). Database resources of the national center for biotechnology information. Nucleic acids research 44: D7.

296

Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, Felts B, Dinsdale EA, Mokili JL, Edwards RA (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications* **5:** 4498.

301

Grazziotin AL, Koonin EV, Kristensen DM (2016). Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic acids research: gkw975.

305

Heiman ML, Greenway FL (2016). A healthy gastrointestinal microbiome is dependent on dietary diversity. Molecular metabolism 5: 317-320.

308

Hill CJ, Lynch DB, Murphy K, Ulaszewska M, Jeffery IB, O'Shea CA, Watkins C, Dempsey E, Mattivi F, Tuohy K, Ross RP, Ryan CA, O' Toole PW, Stanton C (2017). Evolution of gut microbiota composition from birth to 24 weeks in the INFANTMET Cohort. *Microbiome* **5:** 4.

312

Hill JE, Penny SL, Crowell KG, Goh SH, Hemmingsen SM (2004). cpnDB: a chaperonin sequence database. Genome research 14: 1669-1675.

315

Hsiao K-L, Wang L-Y, Lin C-L, Liu H-F (2016). New Phylogenetic Groups of Torque Teno Virus Identified in Eastern Taiwan Indigenes. *PloS one* **11:** e0149901.

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **11:** 119.

Imhann F, Vich Vila A, Bonder MJ, Fu J, Gevers D, Visschedijk MC, Spekhorst LM, Alberts R, Franke L, van Dullemen HM, Ter Steege RWF, Huttenhower C, Dijkstra G1, Xavier RJ, Festen EAM, Wijmenga C, Zhernakova A, Weersma RK (2016). Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut*.

324

325   Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, Fagerberg B, Nielsen J, Bäckhed
326   F (2013). Gut metagenome in European women with normal, impaired and diabetic glucose
327   control. *Nature* **498:** 99-103.

328

329   Kopylova E, Noé L, Touzet H (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs
330   in metatranscriptomic data. *Bioinformatics* **28:** 3211-3217.

331

332   Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI (2005). Obesity alters
333   gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of*
334   *America* **102:** 11070-11075.

335

336   Lim ES, Zhou Y, Zhao G, Bauer IK, Droit L, Ndao IM, Warner BB, Tarr PI, Wang D, Holtz LR
337   (2015). Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nature*
338   *Medicine* **21:** 1228.

339

340   Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for
341   RNA-seq data with DESeq2. *Genome biology* **15:** 550.

342

343   Martin M (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads.
344   *EMBnet journal* **17:** pp. 10-12.

345

346   McMurdie PJ, Holmes S (2013). phyloseq: an R package for reproducible interactive analysis and
347   graphics of microbiome census data. *PloS one* **8:** e61217.

348

349   Menzel P, Ng KL, Krogh A (2016). Fast and sensitive taxonomic classification for metagenomics
350   with Kaiju. *Nature communications* **7**.

351

352   Milani C, Duranti S, Bottacini F, Casey E, Turroni F, Mahony J, Belzer C, Delgado Palacio S,
353   Arboleya Montes S, Mancabelli L, Lugli GA, Rodriguez JM, Bode L, de Vos W, Gueimonde M,
354   Margolles A, van Sinderen D, Ventura M (2017). The First Microbial Colonizers of the Human
355   Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. *Microbiology*
356   *and Molecular Biology Reviews* **81**.

357

358   Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD (2013). Rapid evolution of the
359   human gut virome. *Proc Natl Acad Sci U S A* **110:** 12450-12455.

360

361   NCBI Resource Coordinators (2018). Database resources of the national center for biotechnology
362   information. *Nucleic acids research* **46:** D8-D13.

363

364   Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL,
365   Zhao G, Fleshner P, Stappenbeck TS, McGovern DP, Keshavarzian A, Mutlu EA, Sauk J, Gevers
366   D, Xavier RJ, Wang D, Parkes M, Virgin HW (2015). Disease-specific Alterations in the Enteric
367   Virome in Inflammatory Bowel Disease. *Cell* **160:** 447-460.

368

369   Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017). metaSPAdes: a new versatile
370   metagenomic assembler. *Genome research* **27:** 824-834.

371

372   Oksanen J, Blanchet FG, Kindt R, Legendre P, O'Hara RG, Simpson GL, Solymos P, Stevens
373   MHH, Wagner H (2007). The vegan package. *Community ecology package* Available at
374   https://CRAN.R-project.org/package=vegan (accessed 09 April 2018)

375

376   Reyes A, Blanton LV, Cao S, Zhao G, Manary M, Trehan I, Smith MI, Wang D, Virgin HW,
377   Rohwer F, Gordon JI (2015). Gut DNA viromes of Malawian twins discordant for severe acute
378   malnutrition. *Proc Natl Acad Sci U S A* **112:** 11941-11946.

379

380   Rodriguez JM, Murphy K, Stanton C, Ross RP, Kober OI, Juge N, Avershina E, Rudi K, Narbad
381   A, Jenmalm MC, Marchesi JR, Collado MC (2015). The composition of the gut microbiota
382   throughout life, with an emphasis on early life. *Microbial ecology in health and disease* **26:** 26050.

383

384   Roux S, Enault F, Hurwitz BL, Sullivan MB (2015). VirSorter: mining viral signal from microbial
385   genomic data. PeerJ 3: e985.

386

387   Roux S, Solonenko NE, Dang VT, Poulos BT, Schwenck SM, Goldsmith DB, Coleman ML,
388   Breitbart M, Sullivan MB (2016). Towards quantitative viromics for both double-stranded and
389   single-stranded DNA viruses. PeerJ 4: e2777.

390

391   Spandole S, Cimponeriu D, Berca LM, Mihaescu G (2015). Human anelloviruses: an update of
392   molecular, epidemiological and clinical aspects. Archives of virology 160: 893-908.

393

394 Ssemadaali MA, Effertz K, Singh P, Kolyvushko O, Ramamoorthy S (2016). Identification of
395 heterologous Torque Teno Viruses in humans and swine. Scientific reports 6: 26655.

396

397 Sullivan MJ, Petty NK, Beatson SA (2011). Easyfig: a genome comparison visualizer.
398 *Bioinformatics* **27:** 1009-1010.

399

400 R core team (2016). R: A language and environment for statistical computing. *Vienna, Austria: R*
401 *foundation for statistical computing*. Available at http://www.R-project.org/ (accessed 09 April
402 2018)

403
404 Tyschik EA, Shcherbakova SM, Ibragimov RR, Rebrikov DV (2017). Transplacental transmission
405 of torque teno virus. *Virology Journal* **14**.

406

407 Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR,
408 Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R (2017). Normalization and microbial
409 differential abundance strategies depend upon data characteristics. *Microbiome* **5:** 27.

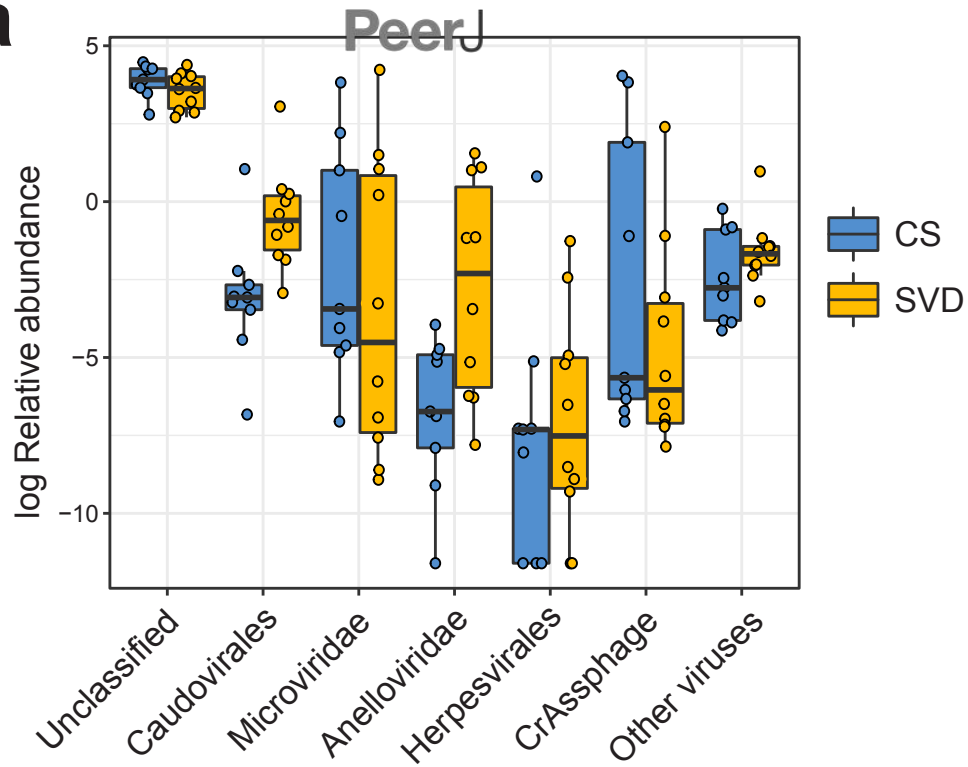
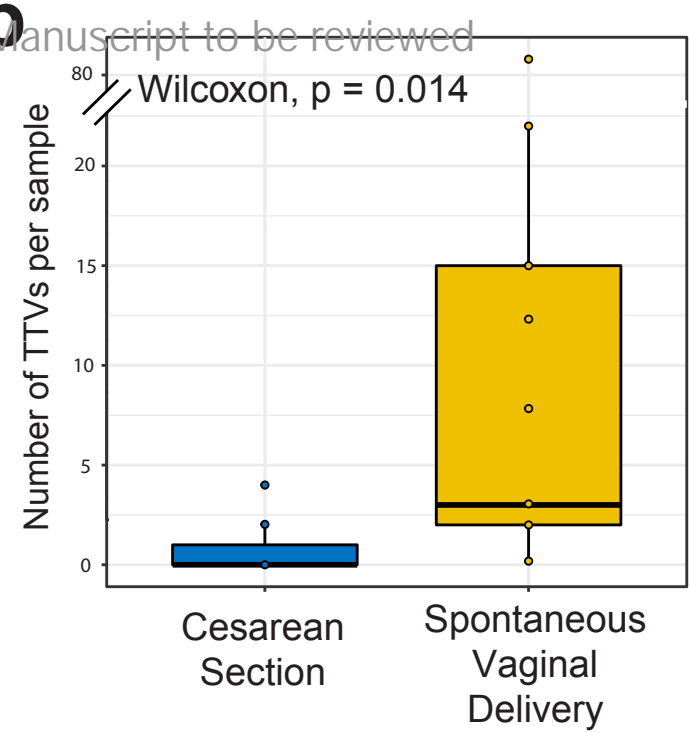
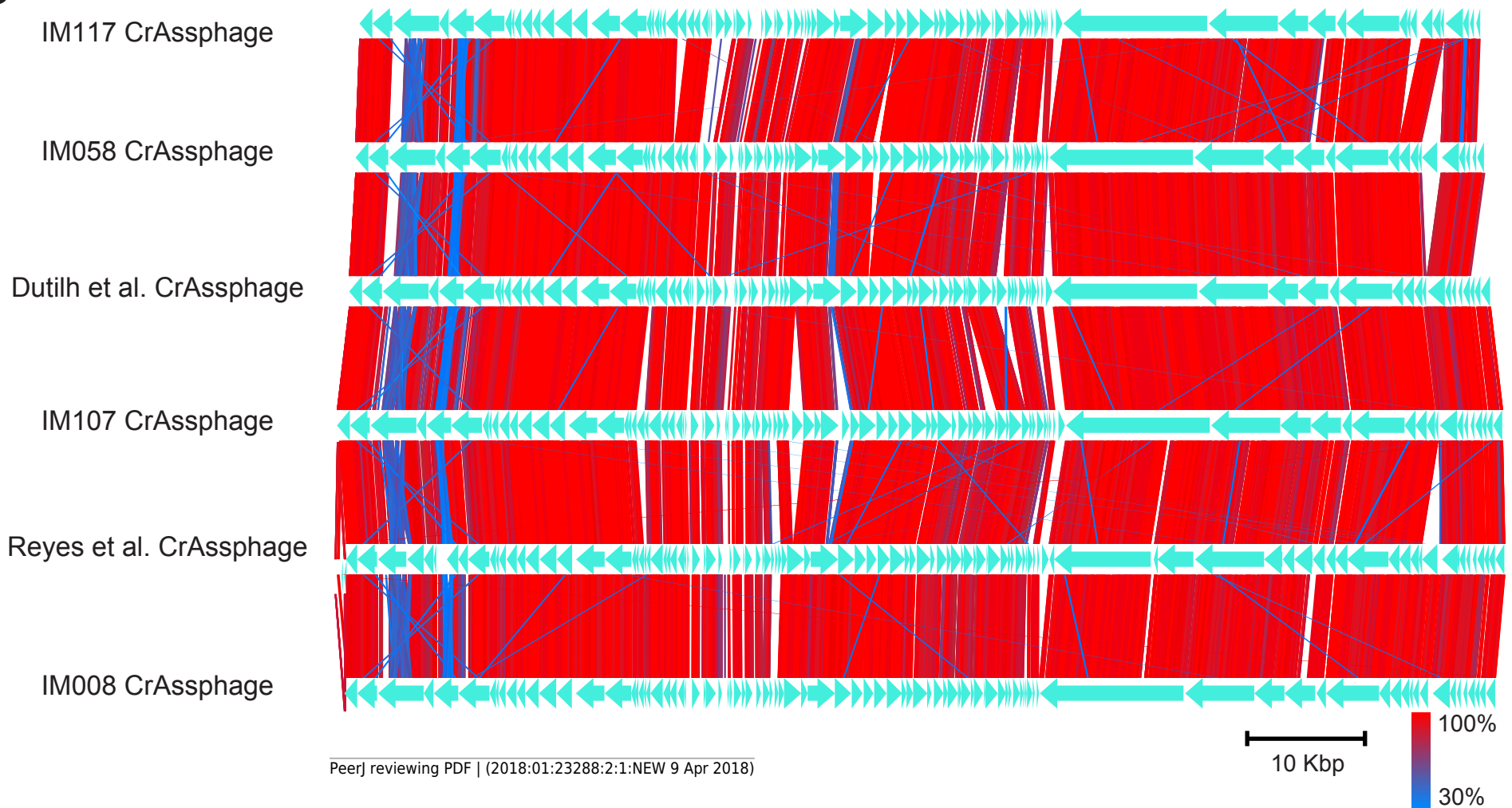
410

411

412 Wickham H (2016). *ggplot2: elegant graphics for data analysis*. Springer.
413
414
415 Yarza P, Yilmaz P, Pruesse E, Glockner FO, Ludwig W, Schleifer KH, Whitman WB5, Euzéby
416 J6, Amann R2, Rosselló-Móra R (2014). Uniting the classification of cultured and uncultured
417 bacteria and archaea using 16S rRNA gene sequences. *Nature reviews Microbiology* **12:** 635-645.

418
419 Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, Koonin EV*l* (2017).
420 Discovery of an expansive bacteriophage family that includes the most abundant viruses from the
421 human gut. *Nature Microbiology*.

Classification and abundance of known viral groups in the INFANTMET cohort.

(A): Log relative abundance of classifiable viral groups by the Kaiju amino acid classifier against the NR protein database. (B) Boxplot of the number of detectable homologues of Torque Teno Virus (TTV) ORF1 in each sample by birth mode. (C) Visualized alignment of multiple CrAssphage genomes of infant origin.

Alpha and beta diversity measures for virome and 16S rRNA sequence data in the INFANTMET cohort.

PCoAs of unweighted Bray-Curtis distances for the (A) virome and (B) 16S rRNA sequence datasets, respectively. Boxplots of Shannon diversity in the (C) virome and (D) 16S rRNA sequence datasets, respectively.

DNA Virome

16S rRNA

a

Adonis, p = 0.002

PC2: 10.38 variation

PC1: 16.08 variation

b

Adonis, p = 0.503

PC2: 9.66 variation

PC1: 14.49 variation

CS

SVD

c

Wilcoxon, p = 0.0022

Shannon

Cesarean
Section

Spontaneous
Vaginal
Delivery

d

Wilcoxon, p = 0.81

Cesarean
Section

Spontaneous
Vaginal
Delivery