1 **Thousands of previously unknown phages discovered in whole-community human gut**
2 **metagenomes**

3 Sean Benler[1], Natalya Yutin[1], Dmitry Antipov[2], Mikhail Raykov[2], Sergey Shmakov[1], Ayal B.
4 Gussow[1], Pavel Pevzner[2,3], Eugene V. Koonin[1]

5   1. National Center for Biotechnology Information, National Library of Medicine, Bethesda,
6 Maryland 20894, USA
7   2. Center for Algorithmic Biotechnology, Institute for Translational Biomedicine, St.
8 Petersburg State University, 199004 St. Petersburg, Russia
9   3. Department of Computer Science and Engineering, University of California, San Diego, La
10 Jolla, CA 92093, USA.

11 *Correspondence : koonin@ncbi.nlm.nih.gov

12

13 Sean Benler – sean.benler@nih.gov

14 Natalya Yutin - yutin@ncbi.nlm.nih.gov

15 Dmitry Antipov - dmitrij.antipov@gmail.com

16 Mikhail Raykov – Mike.Rayko@gmail.com

17 Sergey Shmakov – sergey.shmakov@nih.gov
18 Ayal Gussow – ayal.gussow@nih.gov

19 Pavel Pevzner – ppevzner@ucsd.edu

20 Eugene Koonin - koonin@ncbi.nlm.nih.gov

21

Abstract

Background:

Double-stranded DNA bacteriophages (dsDNA phages) play pivotal roles in structuring human gut microbiomes; yet, the gut phageome is far from being fully characterized, and additional groups of phages, including highly abundant ones, continue to be discovered by metagenome mining. A multilevel framework for taxonomic classification of viruses was recently adopted, facilitating the classification of phages into evolutionary informative taxonomic units based on hallmark genes. Together with advanced approaches for sequence assembly and powerful methods of sequence analysis, this revised framework offers the opportunity to discover and classify unknown phage taxa in the human gut.

Results:

A search of human gut metagenomes for circular contigs encoding phage hallmark genes resulted in the identification of 3,738 apparently complete phage genomes that represent 451 putative genera. Several of these phage genera are only distantly related to previously identified phages and are likely to found new families. Two of the candidate families, "Flandersviridae" and "Quimbyviridae", include some of the most common and abundant members of the human gut virome that infect *Bacteroides, Parabacteroides* and *Prevotella*. The third proposed family, "Gratiaviridae", consists of less abundant phages that are distantly related to the families *Autographiviridae*, *Drexlerviridae* and *Chaseviridae*. Analysis of CRISPR spacers indicates that phages of all three putative families infect bacteria of the phylum Bacteroidetes. Comparative genomic analysis of the three candidate phage families revealed features without precedent in phage genomes. Some "Quimbyviridae" phages possess Diversity-Generating Retroelements (DGRs) that generate hypervariable target genes nested within defense-related genes, whereas

46   the previously known targets of phage-encoded DGRs are structural genes. Several

47   "Flandersviridae" phages encode enzymes of the isoprenoid pathway, a lipid biosynthesis

48   pathway that so far has not been known to be manipulated by phages. The "Gratiaviridae"

49   phages encode a HipA-family protein kinase and glycosyltransferase, suggesting these phages

50   modify the host cell wall, preventing superinfection by other phages. Hundreds of phages in

51   these three and other families are shown to encode catalases and iron-sequestering enzymes that

52   can be predicted to enhance cellular tolerance to reactive oxygen species.

53   **Conclusions:**

54   Analysis of phage genomes identified in whole-community human gut metagenomes resulted

55   in the delineation of at least three new candidate families of *Caudovirales* and revealed diverse

56   putative mechanisms underlying phage-host interactions in the human gut. Addition of these

57   phylogenetically classified, diverse and distinct phages to public databases will facilitate

58   taxonomic decomposition and functional characterization of human gut viromes.

59

60

Background

The bulk of the human-associated virome resides in the distal gastrointestinal tract and is composed of tailed double-stranded (ds) DNA bacteriophages (dsDNA phages) [1-3] that, in the recent virus megataxonomy, are classified as the class *Caudoviricetes* under the phylum *Uroviricota* [4]. The ternary interactions between phages, bacteria and their human hosts are being elucidated at an increasing pace through experiments on model systems and sequencing of the uncultured community of viruses (virome) [5-9]. Comparisons of the human gut virome within and between individuals unveils remarkable longitudinal stability and high diversity of resident phages [2, 10, 11]. Although the human gut offers a rich source of phage genomic diversity, the virome so far has been explored to a much lesser extent than the whole community (metagenome), composed of viruses, Bacteria and Archaea. The rapid growth of the public whole-community metagenomic data offers the opportunity to identify numerous novel phage genomes lurking in metagenomes.

Depending on the terminal genomic arrangement, many complete phage genomes assemble into a 'circular' contig (i.e., a contig with direct terminal repeats) [12]. Thus, circularity can be used as one feature to identify putative complete phage genomes in viromes and metagenomes. However, the comparatively small size of dsDNA phage genomes (~50 kb, on average) [13] and the estimated low virus-to-microbe ratio in the gut (1 : 10) [1] jointly translate into a relatively small amount of phage DNA present in whole community metagenomic libraries [14]. Moreover, similar-sized plasmids also assemble into circular contigs [15]. A recently developed computational method aims to address this problem by focusing specifically on the assembly of circular phage genomes and their automatic discrimination from plasmids based on gene content [16]. The genetic repertoire shared between plasmids and phages, for example, the

84 *parABS* partitioning system encoded by both *Escherichia coli* phage P1 and plasmids [17], can

85 obfuscate their automatic annotation-based discrimination and necessitate manual curation.

86 Despite these challenges, there is a pressing need to reduce the amount of viral "dark matter" in

87 the human gut by identifying and classifying phages for reference-based analyses [18, 19].

88   The global organization of the virosphere was recently captured in a comprehensive,

89 unified framework using protein domains encoded by viral hallmark genes to infer evolutionary

90 connections between major groups of viruses [4] and subsequently approved by the International

91 Committee on the Taxonomy of Viruses (ICTV) as the comprehensive, multi-rank taxonomy of

92 viruses. In particular, dsDNA viruses possess either the HK97 fold or the double jelly-roll fold in

93 their major capsid proteins, along with distinct ATPases involved in capsid maturation, and thus

94 appear to have independent origins, justifying their separation into two realms (the highest virus

95 taxon rank) [4]. Tailed dsDNA phages, with their HK97 major capsid proteins, comprise the

96 order *Caudovirales* within the class *Caudoviricetes*, under the phylum *Uroviricota* (that also

97 include the distantly related herpesviruses of animals) and are further classified into 9 families.

98 With the now formally recognized ability to classify viruses from sequence data alone [20],

99 phylogenomic analysis of uncultured phage genomes can delineate novel taxa.

100   Here, we describe 3,738 completely assembled phage genomes discovered by analysis of

101 5,742 whole-community human gut metagenomes. Using abundance, taxonomy and genomic

102 composition as criteria to select genomes for further scrutiny, three groups of phages, all

103 infecting bacteria of the phylum Bacteroidetes comprising potential new families, were analyzed

104 in detail. All these candidate families, named "Quimbyviridae", "Flandersviridae", and

105 "Gratiaviridae" consist of phages infecting bacteria of the phylum Bacteroidetes, and the first

106 two are widely distributed and abundant in human gut viromes. The phages in these families and

5

107 others yet to be classified encode enzymes that are involved in the response of cells to oxidative

108 stress, implicating phages in the tolerance of anaerobes to oxygen. Furthermore, comparative

109 genomic analysis exposed genetic cassettes that are unique to some genera in each family and

110 thus appear to be relatively recent acquisitions involved in phage-host interactions. Addition of

111 all the phage genomes identified here to public databases will substantially expand the known

112 phage diversity and augment taxonomic classification of the human gut virome.

113

Methods

114 **Methods**

115 **Identification of phage genomes in human gut metagenomes**

116 5,742 whole-community metagenome assemblies generated from human fecal samples were

117 downloaded from the NCBI Assembly database (accessed 8/2019). To limit the search space to

118 likely complete genomes, 95,663 'circular' contigs (50-200 bp direct overlap at contig ends)

119 were extracted from these assemblies. Next, 304 phage-specific protein alignments from the

120 CDD database [21] and 117 custom alignments (Yutin et al., *in press*) were converted to Hidden

121 Markov Models (HMMs) using hmmpress (v. 3.2.1). Proteins in the 95,663 contigs were

122 predicted by Prodigal (v. 2.6.3) [22] in the metagenomic mode and searched against the set of

123 304 phage-specific HMMs using hmmsearch, with the relaxed e-value cutoff of < 0.05. Contigs

124 with at least one hit (n = 4,907) were selected for a second round of searches after correcting for

125 re-assigned codons, as follows. All contigs were searched for the presence of tRNAs using

126 tRNA-scan-SE (v. 2.0) [23]. In 212 contigs, an amber stop codon-suppressor tRNA was

127 identified. ORFs were re-predicted for these contigs with the amber stop codon re-assigned to

128 glutamine, given that this reassignment is most commonly observed in human gut phages [24,

129   25]. The re-translated contigs were added back to the database and all contigs were subjected to a

130   second profile search with a stricter e-value cutoff (< 0.01). Contigs were classified as phages

131   when exceeding 3 kbp in length and possessing at least one ORF that matched a capsid, portal or

132   large terminase subunit protein profile below the e-value threshold. The phage classifications

133   were cross-checked with Seeker [26] and ViralVerify [16]. In cases where both tools classified a

134   contig as non-phage, the protein annotations were examined manually, revealing four contigs of

135   ambiguous identity that were discarded.

136

137   **Collection of phage genomes in GenBank**

138   Taxonomic accession codes corresponding to all prokaryotic viruses were collected from

139   the NCBI Taxonomy database and used to extract sequences longer than 3 kbp from the non-

140   redundant nucleotide database (accessed 09/2019). The protein predictions for each genome

141   sequence were retrieved using the 'efetch' functionality in the entrez direct command line tools

142   [27]. Genomic sequences lacking protein predictions were discarded.

143

144   **Dereplication and annotation of phage genomes**

145   The collections of GenBank and human gut phage genomes were each dereplicated at 95%

146   average nucleotide identity across 80% of the genome length using dRep (v. 2.6.2) [28] and its

147   associated dependencies, Mash [29] and FastANI [30], with all other settings left as default. The

148   proteins from these contigs were collected and clustered at 95% amino acid identity across 33%

149   of the protein length using mmclust [31]. The representative protein sequences were combined

150   into a single BLAST database and compared against the multiple sequence alignments (MSAs)

151 in the CDD database [21] with PSI-BLAST [32] at an evalue cutoff of 0.01. If the representative

152 protein sequence produced a significant result, the representative and all constituent members of

153 the protein cluster were annotated using the best hit.

154

Phylogenetic reconstruction

156 Alignments of the large terminase subunit (TerL), capsid or portal protein were

157 constructed as previously described [33]. Marker proteins from the metagenomic phages were

158 combined with markers from GenBank phages into a single database and initially clustered to

159 50% amino acid identity using mmclust [31]. The clusters were aligned using MUSCLE [34];

160 cluster alignments were then compared to each other using HHsearch (v. 3.0) [35]. The cluster-

161 cluster similarity scores were converted to distances as $-\ln(S_{A,B}/\min(S_{A,A},S_{B,B}))$, where $S_{A,B}$ is

162 similarity between the profiles A and B, then, an unweighted pair group method with arithmetic

163 mean (UPGMA) dendrogram was constructed using the estimated cluster distances. Tips of the

164 tree (depth <1.5) were used to guide the pairwise alignment of the clusters at the tree leaves with

165 HHalign, creating larger protein clusters. The resulting alignments were filtered to remove sites

166 with more than 50% gaps and a homogeneity lower than 0.1 [36]. The filtered alignment was

167 used to construct an approximate maximum-likelihood tree using FastTree [37], with the

168 Whelan-Goldman models of amino acid evolution and gamma-distributed site rates. Examination

169 of the trees identified 353 nearly identical PhiX-174 sequences that were removed from

170 subsequent analyses as a contamination from a sequencing reagent.

Phage genome analysis

172   A gene-sharing network of phage genomes was constructed using Vcontact2 (v. 0.9.19)

173   [38], with default search settings against the database of dereplicated GenBank phage genomes.

174   The results were imported into Cytoscape (v. 3.8) [39] for visualization.

175   The ORFs for selected groups of phages (see the main text) were additionally annotated

176   through HHblits searches against the Uniprot database clustered to 30% identity and the PDB

177   database clustered to 70% identity (available at

178   http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/, accessed 02/2020)

179   [40]. Genomes encoding a predicted reverse transcriptase (RT) were examined for the presence

180   of repeats corresponding to a diversity-generating retroelement using DGRScan with default

181   settings [41]. To identify repeats outside of the 10 kb RT-centered window (the default window

182   of DGRScan), the template repeats were used as BLASTn queries against the encoding genome

183   with the following parameters: -dust no -perc_identity 75 -qcov_hsp_perc 50 -ungapped -

184   word_size 4.

185

186   **Fractional abundance of phage genomes in metagenomes**

187   Dereplicated phage genomes from the NCBI Genbank database were combined with the

188   dereplicated gut phages into a single database and indexed for read recruitment using Bowtie2

189   [42]. A collection of 1,256 human gut viromes were downloaded from the NCBI SRA using the

190   SRA-toolkit (v. 2.10) and quality filtered with fastp (v. 0.20.1) [43]. The quality-filtered virome

191   reads were recruited to the phage database using Bowtie2 with default parameters, except for the

192   following additions "--no-unal --maxins 1000000". The length-normalized fractional abundance

193   of each phage genome in each virome was calculated as described previously [1].

Host prediction from CRISPR-spacer matches

194

A database of CRISPR spacers was compiled from previous surveys of CRISPR-Cas

195

systems [44, 45]. Each spacer was used as a BLASTN [46] query against the phage genomes,

196

using word size of 8 and low complexity filtering disabled. A phage-host prediction was inferred

197

if the spacer was 95% identical over 95% of its length to a phage sequence.

198

Prediction of anti-CRISPR proteins

199

Identification of anti-CRISPR proteins (Acrs) was carried as out as previously described

200

[47]. Briefly, each protein was assigned a score by the Acr prediction model ranging between 0

201

and 1, where a higher score corresponds to a higher likelihood of the protein being an Acr. The

202

proteins were then clustered at 50% amino acid identity and considered a candidate Acr if they

203

satisfied the following criteria: 1) received a mean score of 0.9 or above, 2) are present in a

204

directon of 5 or fewer genes, 3) at least one of the directons encodes an HTH domain-containing

205

protein and 4) the cluster does not produce a hit with an HHpred probability greater than 0.9 to

206

any PDB or CDD database sequence [21].

207

Results

208

Identification of novel phage genomes from whole-community human gut metagenomes

209

The collection of 5,742 whole-community assembled metagenomes was searched for the

210

presence of complete phage genomes. To limit the search space to putatively closed genomes,

211

only 'circular' contigs with direct repeats (50 – 200 bp) at their termini (n = 95,663) were

212

searched for open reading frames (ORFs) matching a known phage marker profile (i.e, the

213

terminase large subunit, major capsid protein, or portal protein). In total, 3,738 contigs encode at

214

least one ORF that passed the e-value and length cutoff criteria (Methods) (**Additional file 1**).

215

216 Dereplication at approximately 95% mean nucleotide identity reduced the number of phage

217 marker-matching contigs to 1,886 (**Additional file 2**). A subset of 664 contigs encoded all three

218 markers, 531 encoded two of the three and the remaining 691 possessed a single detectable

219 marker (**Additional file 3**). The putative phage contigs had a median length of 44.9 kb which is

220 consistent with the recent estimates of the median genome size of dsDNA phages [13]. To

221 exclude any contaminating contigs (e.g., a plasmid harboring an integrated phage), each was

222 assessed with ViralVerify [16] and Seeker [26], two bioinformatic tools trained to discriminate

223 phage genomes from other sequences. These tools classified almost all the selected contigs as

224 phages with varying levels of confidence except for several cases that, upon manual

225 examination, were found to represent false negative classifications by these tools (**Additional**

226 **file 2**). Although we cannot rule out the possibility that some non-phage contigs were retained

227 erroneously, the results collectively suggest that the set of circular marker-matching contigs

228 predominantly consists of complete phage genomes.

229 To determine the host ranges of the phages, a database of CRISPR spacers from

230 prokaryotic genomes was used to query the metagenomic phages for potential matches. In total,

231 553 (29%) of the dereplicated phage genomes were found to be targeted by at least one CRISPR-

232 Cas system allowing host prediction (**Additional file 4**). The most common predicted hosts were

233 *Firmicutes* (321 phages), followed by *Bacteroidetes* (143), *Actinobacteria* (43), *Proteobacteria*

234 (41) and *Verrucomicrobia* (4).

235 Many phages have been found to encode Anti-CRISPR proteins (Acrs) to parry CRISPR-

236 Cas defenses [48-51]. Given their function in counter-defense, Acrs evolve rapidly and show

237 limited sequence similarity to experimentally characterized Acrs, making inference challenging

238 [48]. However, a machine-learning based method has been recently developed that utilizes

11

239   genomic context to identify candidate Acrs [47]. Application of this method showed that 41

240   phages, 16 of which were found to be targeted by a CRISPR-Cas system of their inferred host,

241   encoded at least one candidate Acr, (**Additional file 5**). The highest-scoring Acrs belong to four

242   phages that are targeted by *Bifidobacterium* CRISPR-Cas systems. All four phages are $\geq$ 97%

243   identical over $\geq$ 90% of their length at the nucleotide level to uncharacterized prophages in

244   cultured *Bifidobacterium* isolates **(Additional file 5)**, confirming their host-tropism assignment

245   via CRISPR spacer-protospacer matches. In these phages, two candidate Acr-encoding genes lie

246   between the large terminase subunit and integrase (**Additional file 6**). The localization of the

247   Acr-encoding genes suggests they are expressed not only upon initial entry into the host cell and

248   during lysogeny [52], but also upon transition to the lytic program to prevent cleavage of

249   progeny phage genomes by CRISPR-Cas, as demonstrated experimentally in *Listeria*-infecting

250   phages [53]. Transcription of Acrs is typically regulated by HTH domain-containing proteins

251   termed Acr-associated proteins (Acas) [54]. Indeed, in the *Bifidobacterium* phages identified

252   here, a short HTH domain-encoding ORF is located immediately downstream of the Acrs and

253   can be predicted to regulate the expression of these two genes throughout the phage lifecycle.

254   While these uncharacterized *Bifidobacterium* phages possess the characteristic features of Acr

255   loci, the great majority of the phages identified in this work did not harbor any detectable Acrs

256   yet were targeted by CRISPR-Cas (**Additional file 4**). Some of these phages might encode

257   distinct Acrs undetectable by the method we used that was trained on a collection of previously

258   characterized Acrs, whereas others might employ alternative anti-CRISPR strategies.

259

260   **Taxonomic decomposition of the gut phages identifies previously unknown putative**

261   **families**

262   Phylogenetic trees were constructed for the large terminase subunit (TerL), major capsid

263   protein (MCP) and portal protein encoded in each phage genome using an iterative approach to

264   construct the underlying alignments [33]. The trees were constructed alongside reference

265   proteins derived from phage genomes extracted from the NCBI GenBank database. Reflecting

266   the set of protein profiles employed to identify the phage contigs, 1,480 (78%) genomes were

267   assigned to the phylum *Uroviricota*, 360 (19%) to the *Phixviricota* and 46 (2%) to the *Loebvirae*

268   (**Additional file 3**). The phylum *Phixviricota* includes *Escherichia coli* phage phiX174 that is

269   used as a sequencing reagent; however, the 360 Phixviricota phages detected in this analysis do

270   not include any sequences closely related to phiX174 (see Methods). The remaining analyses

271   focus on the taxonomic decomposition of the phages that belong to *Uroviricota*, given that these

272   contigs represent by far the largest fraction of recovered genomes.

273   The phylum *Uroviricota* is organized into a single class (*Caudoviricetes*) and order

274   (*Caudovirales*), but a new order encompassing the crAss-like phages, a common and apparently

275   most abundant group of phages in the human gut virome, is being proposed [55]. Our profiles

276   recovered 141 phage genomes (dereplicated from 601 total genomes) that displayed phylogenetic

277   relationship with the crAss-like phages and are the subject of a separate study (Yutin *et al.*, in

278   press). The *Caudovirales* are presently organized into 9 families, but 3 of these (*Myoviridae,*

279   *Podoviridae, Siphoviridae*) are expansive and demonstrably polyphyletic [56-58] and were thus

280   not used for family-level taxonomic assignment although the remaining 6 families represent only

281   a small fraction of the phages available in GenBank. The phylogenetic tree of TerL encoded by

282   *Caudovirales* phages in our set reveals only 34 gut phages belong to one of these 6 ICTV-

283   accepted families (**Figure 1** and **Additional file 3**). The remainder of these unclassified phages

284   are likely to found new families presently composed entirely of uncultured phages or belong to

285   families with a cultured representative that have yet to be defined under the new multi-rank

286   taxonomy of viruses.

287

288   **Selection of candidate families for comparative genomics**

289   The taxonomic analysis based on phage hallmark proteins demonstrates that few phages

290   in the human gut belong to a currently accepted ICTV-family. To prioritize candidate families

291   for in-depth analysis, we next complemented the hallmark gene-based taxonomic analysis with

292   whole-genome comparisons and abundance calculations of each phage relative to GenBank

293   phages.

294   A gene-sharing network was constructed with the phages recovered from metagenomes

295   and those deposited in GenBank. Edges are drawn between two viral genomes, represented as

296   nodes, based on the number of ORFs that share significant sequence similarity [38]. Most of the

297   metagenome-recovered phages bore multiple connections within the network to GenBank

298   phages, in agreement with the manual curation of these contigs as genuine phage genomes

299   (**Figure 1B**). However, two large groups of phages (tentatively labelled "Flandersviridae" and

300   "Gratiaviridae") were weakly connected to the larger network, reflecting disparate genome

301   content. The divergence of the gene content of these phages from those of previously known

302   phages and their distinct position in phylogenetic trees (**Figure 1A** and see below) indicate that

303   they represent novel genera, and likely, new families.

304   To quantify the fractional abundance [1] of each phage in the human gut viral

305   community, reads from a collection of 1,265 human gut viromes were competitively mapped

306   against a database containing the metagenome-recovered and GenBank phages. The majority of

14

307   the genomes do not recruit any reads ("detection") from more than 2% of the viromes (Q1-Q3, 0-

308   2% of viromes) (**Figure 1C**-**D**), consistent with the previously reported individuality of the

309   human gut viromes [2, 7, 11]. A notable exception are the crAss-like phages [59] that recruit at

310   least one read from about one-third of the viromes (Q1 - Q3, 9 - 28%), in agreement with

311   previous reports of their cosmopolitan distribution [60, 61]. One uncharacterized *Caudovirales*

312   genome was frequently observed in the collection of human gut viromes (54%, **Figure 1C**),

313   suggesting that this phage is also cosmopolitan. To rule out the possibility that the observed

314   frequency stemmed from non-specific read mapping to one or a few loci, rather than the

315   complete genome, the coverage of sequencing reads across the genome (accession

316   OMAC01000147.1) was examined. The broad coverage of this genome in the viromes confirms

317   that its frequent detection is not an artifact, although several loci present in the reference

318   sequence were absent in the viromes (**Additional file 7**). The exceptional detection of this

319   uncharacterized phage (hereafter referred to as Quimbyvirus, after the character Mayor Quimby

320   from the *Simpsons*) in the human gut viral community warrants its detailed examination.

321   Thus, three groups of phages were selected for in-depth analysis based on their distinct

322   positions in the phylogenetic trees of the marker genes (all three groups), combined with

323   divergent gene contents ("Flandersviridae" and "Gratiaviridae") and high abundance in the

324   human gut viral community ("Flandersviridae" and "Quimbyviridae"). A comparative genomic

325   analysis of each candidate family is presented below, case-by-case.

326

327   **"Quimbyviridae" phages are abundant, hypervariable phages infecting *Bacteroides***

328   In the TerL phylogenetic tree, Quimbyvirus belongs to a group of phages whose closest

329   characterized relatives include the *Vequintavirinae* and *Ounavirinae* subfamilies, under the now

330   defunct *Myoviridae* family. To elucidate the taxonomic affiliation of Quimbyvirus, genomes

331   from adjacent branches were examined (**Figure 2**). The median genome length of Quimby-like

332   phages is 75.2 kb, close to the genome size of a branch basal to the Quimby-like branch, "group

333   4986" (72 kb), but smaller than the genomes of other phages in adjacent branches, *Ounavirinae*

334   (88 kb) and *Vequintavirinae* (145 kb). Despite the similarity in genome size, phylogenetic

335   reconstruction of the portal protein and MCP separate the Quimby-like phages from group 4986

336   (**Additional file 8**). Moreover, most Quimby-like phages encode a DnaG-family primase and

337   DnaB-family helicase that are both absent in group 4986. However, in one branch of Quimby-

338   like phages, the primase was lost from the replication module. The genomes of this branch

339   encode a protein adjacent to the DnaB-family helicase with significant structural similarity to the

340   winged helix-turn-helix domain of RepA (HHpred probability, 96.5) (**Figure 2**). RepA-family

341   proteins mediate replication of plasmids by interacting with host DnaG primases [62], suggesting

342   that the RepA-like protein coopts the host primase during replication, triggering the loss of the

343   phage-encoded *dnaG* in this lineage. Consistent with a RepA-mediated episomal replication

344   strategy, no integrase is identifiable in the genomes on this branch yet the phages encode

345   numerous antirepressors, proteins involved in the lysis-lysogeny decision of temperate phages

346   [63, 64]. The rest of the Quimby-like phages harbor a full-length integrase, indicating that these

347   phages integrate into their host cell genome (**Figure**). Based on the topologies of the TerL,

348   portal, MCP and DnaG trees, we propose that Quimby-like phages represent a novel taxonomic

349   group at the family rank (henceforth, the "Quimbyviridae"). The potential differences in

16

350   replication strategies (episomal vs. integrated) combined with the topologies of the phylogenetic

351   trees of marker proteins suggest that "Quimbyviridae" splits into two distinct subfamilies.

352   The Quimbyvirus genome aligns with a cryptic prophage of the bacterium *Bacteroides*

353   *dorei* (CP011531.1), with 95% nucleotide sequence identity across 92% of its length, indicating

354   that *B. dorei*, a common constituent of human gut microbiomes [65], carries a prophage closely

355   related to Quimbyvirus. Inspection of the alignment shows that Quimbyvirus site-specifically

356   integrates into the tRNA-Asp gene of *B. dorei*, a typical site of prophage integration [66]. The

357   hosts of the other "Quimbyviridae" phages, determined through CRISPR-spacer analysis, include

358   the *Prevotella*, *Bacteroides* and *Parabacteroides* genera within the phylum *Bacteroidetes* and the

359   *Lachnospiraceae* within the phylum *Firmicutes*. In contrast, the hosts of group 4986 do not

360   include any *Bacteroidetes*. The differences in the inferred host ranges support separating group

361   4986 from 'Quimbyviridae" phages and suggests that group 4986 might represent a novel family,

362   but these genomes were not investigated further.

363   Some of the "Quimbyviridae" phages harbor diversity-generating retroelements (DGRs),

364   a cassette of genes that selectively mutate a short locus, known as the variable repeat, that is part

365   of a C-type lectin or an immunoglobulin-like domain [67, 68]. Targeted mutation of these

366   domains yields proteins with altered binding affinities and specificities [69]. The DGR cassette

367   in *Bordetella* phage BPP-1 of the genus *Rauchvirus* is the only experimentally studied DGR

368   system in a phage, where diversification of the C-type lectin domain-containing tail fiber gene

369   enables adsorption to different host cell receptors [70]. In Quimbyvirus, the RT component of the

370   DGR is encoded by overlapping ORFs in all three frames (ORFs 52-54), suggesting that the

371   active RT is produced by two programmed frameshifts. Although overlapping ORFs and

372   programmed frameshifts have been identified in in many compact tailed phage genomes [71-74],

17

373 DGR RTs have thus far only been predicted to be encoded by a single ORF. To discern if the

374 frameshifts render the RT inactive, the variable repeats were examined for adenine-specific

375 substitutions, a hallmark of DGR-mediated variation [68]. The two variable repeats reside in

376 ORF 47 and 80 of the Quimbyvirus genome which both encode proteins containing C-type lectin

377 domains, the canonical target of DGRs [67] (**Figure 2**). Alignment of the variable repeats with

378 their cognate template repeats from nearly identical Quimbyvirus genomes ($\geq$ 95 % average

379 nucleotide identity) allowed the detection of 22 adenine sites in the variable repeat exhibiting

380 substitutions whereas all other bases were nearly perfectly conserved (**Additional file 9**).

381 Collectively, these results suggest that the frameshifted RT possesses the selective infidelity that

382 characterizes DGR-mediated hypervariation.

383   The first variable repeat resides in the C-terminus of ORF 51 that is located downstream

384 of the tail fiber genes, suggesting that this gene codes for a structural component of the virion,

385 similar to the hypervariable tail fiber of phage BPP-1 [70, 75]. The second DGR target locus is in

386 ORF 84 that is distal to the phage structural gene module and is expressed from the opposite

387 DNA strand, suggestive of a non-structural protein. The genomic neighborhood of ORF 80

388 includes genes coding for a nuclease, four methyltransferases and a tRNA ligase within 7 kb.

389 The nuclease shows significant sequence and structural similarity to *E. coli mutY* (HHpred

390 probability, 97.3, **Additional file 10**), a DNA glycosylase involved in base excision repair. The

391 methyltransferases are most similar to adenine- and cytosine-modifying enzymes (HHpred

392 probability 100 and 99.9, respectively, **Additional file 10**) that likely prevent cleavage by host

393 restriction endonucleases. Similarly, the tRNA ligase might repair tRNAs cleaved by host

394 anticodon endonucleases [76]. Overall, the adjacency of ORF 84 with defense- and

395 counterdefense-related genes implies that this hypervariable phage protein plays a role in the

18

396   phage-host conflicts; however, the exact functions of the DGR and hypervariable target proteins

397   during the life cycle of "Quimbyviridae" phages remain to be investigated.

398

399   **"Flandersviridae" phages are common and abundant in whole-community metagenomes**

400   Analysis of the phylogenetic trees of TerL identified a deep branch of 29 gut phages

401   (dereplicated from 196 total genomes) that joins the family *Ackermannviridae* (**Figure 3A**).

402   Annotation of the ORFs encoded by the 29 representative contigs demonstrated that the genomes

403   are colinear, confirming that they belong to a cohesive group (**Figure 3B**). The cohesiveness of

404   this group was confirmed by the gene-sharing network, where these genomes form a coherent

405   cluster that has few connections to the larger network (**Figure 1B**), reflecting distant (if any)

406   similarity between most of the proteins encoded by these phages and proteins of phages in

407   GenBank. The median genome size of the phages in this group is 85.2 kb, compared to 157.7 kb

408   among the *Ackermannviridae* phages. There is a conserved module of structural genes that

409   encode the MCP, portal, sheath and baseplate proteins, TerL and the virion maturation

410   proteinase. The presence of a contractile tail sheath indicates that these viruses possess

411   contractile tails similar to those in the family *Ackermannviridae*, in agreement with the TerL

412   phylogeny. Several of the genes within the structural block contain immunoglobulin-like or C-

413   type lectin domains (e.g., BACON and GH5, respectively) which are predicted to play a role in

414   adhesion of the virion to bacterial cells or host-associated mucosal glycans [77-80]. Downstream

415   of the structural block is a module of genes involved in DNA replication that includes a DnaB-

416   family helicase, DnaG-family primase and DNA polymerase I (PolA). The *polA* gene is widely

417   distributed among dsDNA phages and therefore serves as a useful marker for delineating the

418   diversity of phage replication modules [81]. Phylogenetic reconstruction of both *polA* and *dnaG*

19

419   encoded by these phages confirmed their monophyly (**Additional file 11**). Following the

420   replication module is an approximately 20 kb long locus containing ORFs that showed no

421   detectable similarity to functionally characterized proteins. Two of the phages harbor matches to

422   CRISPR spacers encoded by *Bacteroides* and *Parabacteroides* spp., indicating these bacteria

423   serve as hosts. Based on the large terminase and *polA* phylogeny, colinearity of their genomes

424   and differences from known phages in both genome size and content, we propose that these

425   *Bacteroides*-infecting phages represent a novel taxonomic group, probably, with a family rank,

426   hereafter, named "Flandersviridae" (after the region where some of the metagenomes were

427   sampled).

428   Although all members of the "Flandersviridae" are syntenic, some contain an insertion of

429   two adjacent genes encoding nucleotidyltransferase superfamily enzymes within the DNA

430   replication module. One enzyme belongs to the *ispD* family that is involved in the biosynthesis

431   of isoprenoids [82, 83], and the other is a *licD* family enzyme that is responsible for the addition

432   of phosphorylcholine to teichoic acids present in bacterial cell walls [84] (**Figure 3**). To our

433   knowledge, neither of these enzymes has been reported in phages previously. Given that only

434   some members of the "Flandersviridae" possess these genes, they are unlikely to perform

435   essential functions in phage reproduction, and instead could be implicated in phage-host

436   interactions. The *licD* family enzyme might modify teichoic acids to prevent superinfection by

437   other phages, given that these polysaccharides serve as receptors for some phages to adsorb to

438   the host cells [85]. The role of *ispD* is less clear because *ispD* family enzymes catalyze one step

439   in the biosynthesis of isopentenyl pyrophosphate, a building block for a large variety of diverse

440   isoprenoids [86]. Phages manipulate host metabolic networks including central carbon

441   metabolism, nucleotide metabolism and translation [87]; the discovery of *ispD* present in the

442   "Flandersviridae" phage genomes might add to this list the isoprenoid biosynthetic pathway.

443   Complete "Flandersviridae" phage genomes were recovered from 249 whole-community

444   human gut metagenomic assemblies. Their frequent assembly into closed contigs suggests that

445   these phages might persist in their host cells as extrachromosomal circular DNA molecules,

446   similar to phage P1 [88]. However, neither genes involved in DNA partitioning nor lysis-

447   lysogeny switches are readily identifiable in the "Flandersviridae" genomes. Thus, this group of

448   phages might be obligately lytic although discerning the lifestyle of a phage from the genome

449   sequence alone is challenging [89]. Regardless of their lifestyle, the frequent recovery of these

450   phages from whole-community metagenomes implies that they are common members of the

451   human gut virome. Indeed, the "Flandersviridae" phages reach similar detection frequency as the

452   crAss-like phages (**Figure 1D**) although there are fewer Flanders-like phages in the database.

453   Like the "Quimbyviridae", the even coverage of sequencing reads across one "Flandersviridae"

454   genome (accession OLOC0100071.1) confirms its detection is not artifactual (**Additional file**

455   **12**). The high fractional abundance and detection of Flanders-like phages in viromes generally

456   agrees with their frequent assembly from whole community metagenomes although they were

457   not the most abundant (see Discussion). Overall, Flanders-like phages represent a previously

458   undetected phage group that is widely distributed in human gut viromes.

459

460   **"Gratiaviridae", a putative novel family of phages infecting *Bacteroides***

461   A deeply branching cluster of 18 genomes (dereplicated from 45 total) is basal to the

462   families Autographiviridae, Drexlerviridae and Chaseviridae on the TerL phylogenetic tree

463   (**Figure 4A)**. Although not commonly present in gut viromes (**Figure 1D**), the deep relationship

464   between these contigs and established phage families prompts in depth genome analysis of these

465   putative phages. All 18 genomes encode a DnaG-family primase and a DnaE-family polymerase,

466   and phylogenetic reconstruction for these genes demonstrates monophyly of these phages, the

467   sole exception being the *dnaE* gene of bacteriophage phiST, a marine *Cellulophaga*-infecting

468   phage that belongs to the polyphyletic, currently defunct Siphoviridae family [90] (**Additional**

469   **file 13**). The *dnaG* and *dnaE* genes are nested within a module of other replication-associated

470   genes that include superfamily I and II helicases, SbcCD exonucleases and a RecA family

471   ATPase (**Figure 4B**). The structural module is composed of genes that encode an MCP, capsid

472   maturation protease, portal protein, baseplate proteins and a contractile tail sheath protein.

473   Although these genomes are not strictly colinear as observed for the "Flandersviridae" phages,

474   the overall similarity of the proteins encoded by these phages is apparent in the gene-sharing

475   network where they form a coherent cluster that shares some edges with the crAss-like phages

476   (**Figure 1B**). Similar to crAss-like phages, the predicted hosts suggested by CRISPR-spacer

477   matches are the *Bacteroides* and *Parabacteroides* genera (**Additional file 4**). Taken together, the

478   phylogenetic and genomic organization of these phages indicate that they represent a new family,

479   provisionally named "Gratiaviridae" (after the pioneering phage biologist Dr. Andre Gratia).

480   In addition to structural and replication proteins, "Gratiaviridae" phages encode several

481   enzymes of the ferritin-like diiron-carboxylate superfamily. The ferritin-like enzymes encoded

482   by these phages belong to two families, namely, DNA protecting proteins (DPS) and manganese-

483   catalases. Manganese-catalases have not been documented in phage genomes and DPS-like

484   enzymes have only been observed in seven *Lactobacillus*-infecting phages [91]. Both enzymes

485   are involved in the tolerance of anaerobes to oxidative stress. Catalases detoxify hydrogen

22

486 peroxide to oxygen and water, enhancing survival of anaerobic *Bacteroides* in the presence of

487 oxygen [92]. DPS enzymes catalyze a reaction between oxygen and free iron to yield insoluble

488 iron oxide, lowering the concentration of both intracellular oxygen and free iron levels that

489 would otherwise react with hydrogen peroxide and produce a hydroxyl radical, the most toxic

490 reactive oxygen species [93, 94]. "Gratiaviridae" phages might deploy catalase- and DPS-like

491 enzymes during infection to enhance the tolerance of their strictly anaerobic *Bacteroides* hosts to

492 oxidative damage. Notably, these enzymes were not restricted to the "Gratiaviridae" but could be

493 identified in 196 (manganese catalase) and 36 (Dps) other phage genomes, including the

494 "Flandersviridae". The frequent identification of these enzymes in gut phage genomes

495 underscores the importance of intracellular iron and reactive oxygen species concentration for

496 productive infections in an anaerobic environment.

497   Five of the "Gratiaviridae" phages encode a protein containing a serine/threonine protein

498 kinase domain with distant but significant sequence similarity to HipA family kinases (HHpred

499 probability 99, **Additional file 10**). Whereas HipA family kinases are present in numerous,

500 phylogenetically distinct bacterial genomes as the toxin component of a distinct variety of type II

501 toxin-antitoxin systems [95, 96], there are only two characterized examples of protein kinases

502 encoded by phages. The protein kinase of T7-like phages phosphorylates RNA polymerase and

503 RNAse III early during infection as part of the takeover of the host cell transcriptional and

504 translational machinery [97-99]. In contrast, the protein kinase of *E. coli* phage 933W is

505 expressed during lysogeny and mediates abortive infection upon superinfection of the host cell

506 by phage HK97 [100]. The HipA-like kinase is unlikely to function early during infection like

507 the kinase of T7-like phages because, in all five "Gratiaviridae" phages, the kinase is encoded

508 between the portal protein and MCP genes, which are expressed late during infection in

23

509   numerous cultured phages [101, 102]. Instead, the kinase might confer immunity to heterotypic

510   phage infection, analogous to the kinase encoded by 933W [100]. In support of an immunity-

511   related role, an AAA-family ATPase and a glycosyltransferase are encoded immediately

512   upstream of the kinase in all five phage genomes (**Figure 4**). Glycosyltransferases are encoded

513   within capsular polysaccharide biosynthetic loci [103] and phase variation of the capsular

514   polysaccharides confers immunity from phages that rely on these molecules for adsorption [104].

515   The specific roles of the HipA-family kinase, ATPase and glycosyltransferase are unknown but,

516   collectively, these enzymes might modify host cell capsules, granting temporary immunity to

517   heterotypic phage infection while the morphogenesis of "Gratiaviridae" progeny virions

518   completes.

519

Discussion

521   A search of human gut metagenomes identified 3,738 putative complete phage genomes.

522   In an attempt to recover complete phage genomes, this analysis restricts the search space to

523   metagenomic contigs with direct terminal repeats which are present at the termini of some phage

524   genomes that consequently form circular assemblies [12]. Phages with different replication and

525   DNA packaging strategies, such as members of the phylum *Preplasmiviricota* or *Escherichia*

526   phage Mu [12], that lack direct repeats do not yield circular assemblies and thus were not

527   detected here. As a result, the set of phage genomes recovered by this strategy is both biased and

528   an underestimate. The results are also skewed towards smaller genomes that are more likely to

529   assemble into a single contig although, in one metagenome, a 294 kb phage genome was

530   identified (**Additional file 1**). Phylogenetic and comparative genomic analyses suggest that this

531   set of contigs includes many previously unnoticed lineages of phages, some, most likely, at the

532   family rank.

533   Two groups of phages were selected for in-depth analysis based on their frequent

534   recovery in metagenomes and viromes. Complete genomes of "Flandersviridae" and

535   "Quimbyviridae" phages were assembled in 249 and 20 whole-community metagenomes,

536   respectively. Yet, Quimbyvirus was more frequently detected in the viromes than any

537   "Flandersviridae" phage (**Figure 1D**). The discrepancy can be attributed to several factors,

538   including sampling bias, the greater number of "Flandersviridae" genomes in the reference

539   database "diluting" the number of mapped reads per genome, or the presence of variable loci

540   (e.g., the variable repeats of DGRs) that break contig assemblies [105]. Regardless, both groups

541   encompass abundant members of the human gut virome. Predictably, the hosts of these phages

542   include *Bacteroides spp.*, which are some of the most dominant bacterial taxa of the human gut

543   [106] and serve as hosts for other common human gut phages [61, 107]. Much of the

544   uncharacterized "dark matter" in these phage genomes is likely to be dedicated to preventing

545   superinfection of the *Bacteroides* host cells by such phages and to counter the host defenses.

546   Although in general defense systems in *Bacteroidetes* remain poorly characterized, most of the

547   bacteria possess active CRISPR-Cas systems, and numerous CRISPR spacers targeting the

548   phages analyzed here were detected. This implies that many if not most of the phages infecting

549   *Bacteroidetes* would encode Acrs. However, the currently available prediction method that was

550   trained on the sequences of previously identified Acrs detected putative Acrs only in a small

551   minority of these phages. The remaining phages of *Bacteroidetes* might encode distinct Acrs or

552   employ alternative anti-CRISPR strategies.

553   Several phage genera possess DGRs, including Quimbyvirus. Metagenomic surveys have

554   shown that DGRs are enriched in the viruses that inhabit gastrointestinal environments [105,

555   108]. Combined with the isolation of DGR-carrying phages from human gut bacteria [109, 110],

556   these observations reflect a prominent role of hypervariability underlying phage-host interactions

557   in the gastrointestinal environment. Notably, Quimbyvirus and another DGR-carrying phage

558   (Hankyphage, BK010646.1), lysogenize the same *Bacteroides* species and both phages are

559   frequently detected in human gut viromes [110]. The commonalities aside, the Quimbyvirus

560   DGR RT is encoded by three overlapping reading frames and targets two proteins, one in the

561   structural module and one in a defense-related island. DGRs have been associated with putative

562   defense and signaling systems in cyanobacterial and gammaproteobacterial genomes [111, 112],

563   but beyond the presence of the C-type lectin fold, the hypervariable proteins possess few other

564   recognizable domains that obfuscate their precise roles.

565   The third group analyzed in this study, the "Gratiaviridae", is not abundant but occupies a

566   deep position on the TerL tree relative to the *Autographiviridae*, *Chaseviridae* and

567   *Drexlerviridae* families. Analysis of the "Gratiaviridae" genomes will facilitate the future

568   organization of these families into higher taxonomic ranks, potentially, at the order level.

569   Furthermore, analysis of the "Gratiaviridae" genomes demonstrated the presence of catalase- and

570   DPS-family enzymes that arbitrate cellular responses to oxidative stress [113]. Oxygen

571   concentrations vary along the length of the gastrointestinal tract, where the concentration is

572   lower in the distal vs. proximal gut [114]. Oxygen also diffuses from tissues radially into the

573   lumen [115] and, in combination with other factors, these gradients affect the structure and

574   composition of the gastrointestinal microbiota [116]. The acquisition of oxygen detoxifying-

575   enzymes by the "Gratiaviridae" and other gut phages signals a need to supplement their host

576   cell's tolerance to oxidative damage during infection, which might be especially important for

577   cells that reside near the tissue surface where oxygen exposure is higher.

578   A unique feature of some "Gratiaviridae" phages is a HipA-family protein kinase. The T7-

579   like phages (within the *Autographiviridae* family) and *Escherichia* phage 933W (currently

580   unclassified at the family level) encode PKC-family protein kinases that function during host cell

581   takeover and abortive infection, respectively [97, 100]. A third, CotH-family protein kinase

582   domain is occasionally observed in phage genomes where it is fused to a hypervariable C-type

583   lectin domain [67, 108], but these proteins are currently unstudied. The "Gratiaviridae" family

584   phage recruited of a fourth family of protein kinases that, together with the phage encode

585   glysosyltransferase, might modify the host cell envelope, contributing to the prevention of

586   superinfection.

Conclusions

588   In summary, comparative genomic analysis of the phages described here, along with the

589   complementary analysis of crAss-like phages (Yutin *et. al,* in press), substantially increases the

590   characterized diversity of phages, primarily, those infecting Bacteroidetes bacteria, which are

591   major components of the human gut microbiome. These findings also expand the repertoire of

592   phage gene functions, notably, by adding the isoprenoid metabolic pathway, catalase-like

593   enzymes, HipA family protein kinases and hypervariable genes implicated in defense, and open

594   multiple directions for experimental study.

Declarations

596   **Ethics approval and consent to participate** – Not applicable

597   **Consent for publication** – Not applicable

598   **Availability of data and materials** – All phage genomes are available in the NCBI GenBank

599   database using the accession numbers listed in **Additional file 1**. The genomes, protein

600   predictions and annotations are also provided at
601   ftp://ftp.ncbi.nih.gov/pub/benlersm/gut_phages_2020/. Underlying alignments of marker protein
602   sequences used to generate phylogenetic trees are provided in **Additional file 14.** Raw sequence
603   data was downloaded from the NCBI SRA database using the accession numbers listed in
604   **Additional file 15.**

605   **Competing interests** - The authors declare that they have no competing interests.

609   **Authors contributions** –NY, MR and DA analyzed the gut metagenomes; SS and AG executed
610   the CRISPR analyses; SB analyzed the phage genomes and wrote the manuscript; PP and EVK
611   conceived the project and wrote the manuscript. All authors read and approved the final
612   manuscript.
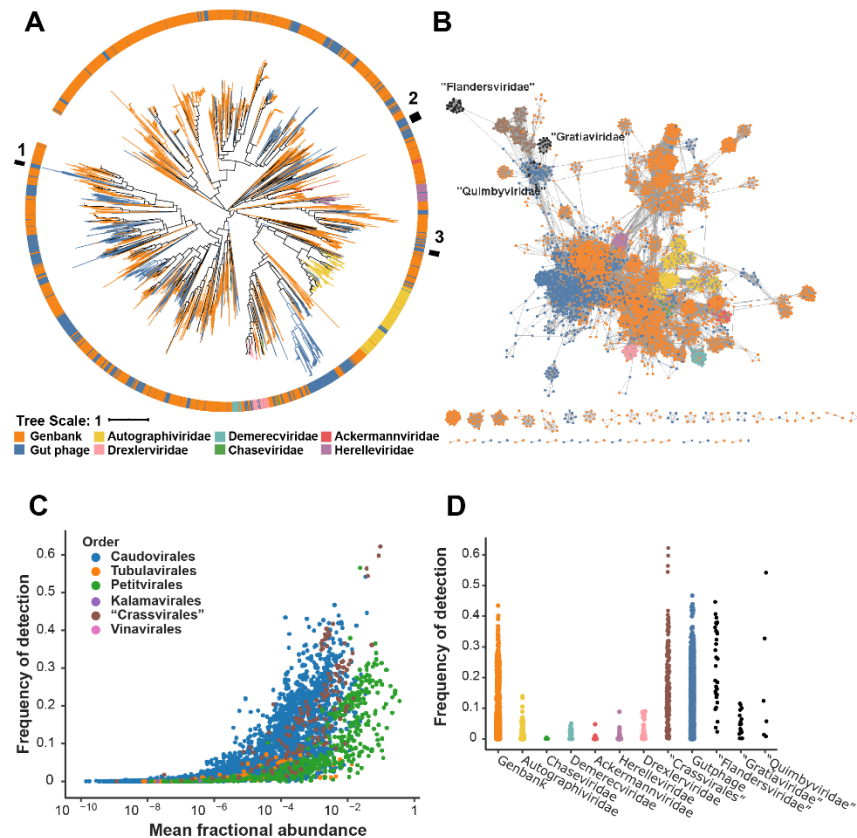
617

618

Figures

620

Figure 1: Three candidate families of Caudovirales phages discovered in human gut metagenomes. (A) Phylogenetic tree of the large terminase subunit encoded by Caudovirales phage genomes in GenBank (n = 3,931) and in gut metagenomes (n = 1,298). Branches are colored according to the current ICTV families, except for the Myoviridae, Podoviridae or Siphoviridae, which are in orange. The outermost ring indicates the location of candidate families proposed in this study: 1, "Quimbyviridae" phages; 2, "Flandersviridae" phages; 3, "Gratiaviridae" phages (see main text). (**B**) Gene sharing network of the Urovicota phages. Phage genomes identified in human gut metagenomes (blue nodes) were compared to phages in the GenBank database (colored as in Figure 1, with the addition of the crAss-like phages in brown and the new Caudovirales families proposed in this study in black). (**C**) Abundance of phages across human gut viromes. The x-axis depicts the fractional abundance of a given phage averaged across all viromes (n = 1,258); the y-axis is the fraction of viromes that a given phage recruits at least one read. Each phage genome (n = 7,888 total) is colored at the taxonomic level of Order (**C**) or Family (Uroviricota families only) (**D**). The raw data used to generate panel (A)

29

637   are provided in **Additional file 14** and data for panels (C) and (D) are provided in **Additional**
638   **file 15**.

639

640   *Figure*

641

642   **Figure 2:** Phylogenetic tree of the large terminase subunit and genome maps of Quimby-like
643   phages. (A) Individual genome maps of Quimby-like phages and ICTV classified phages are
644   shown to the right of each branch. The ORFs are colored according to function: large terminase
645   subunit (red), structural components (blue), DNA replication and repair (orange), lysogeny
646   (pink), general function (green) and unknown (grey). **(B)** Expansion of four Quimby-like phages
647   and a single gut phage genome from an adjacent branch ("group 4986"). The diversity-

30

648   generating retroelement and hypervariable ORFs are highlighted with a dashed box and asterisk.
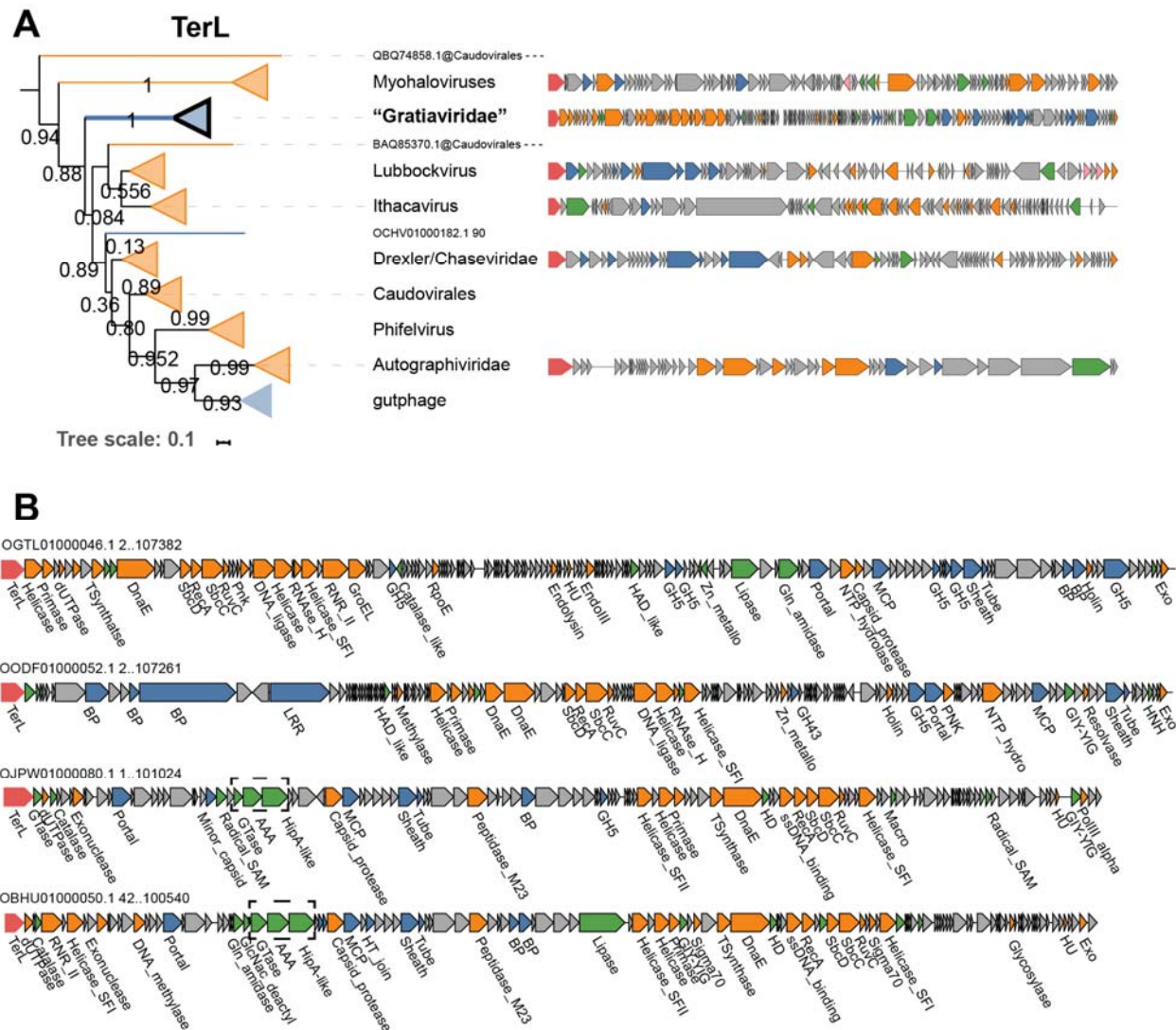649   The nucleotide scales differ between individual genome maps in both panels.

650

651
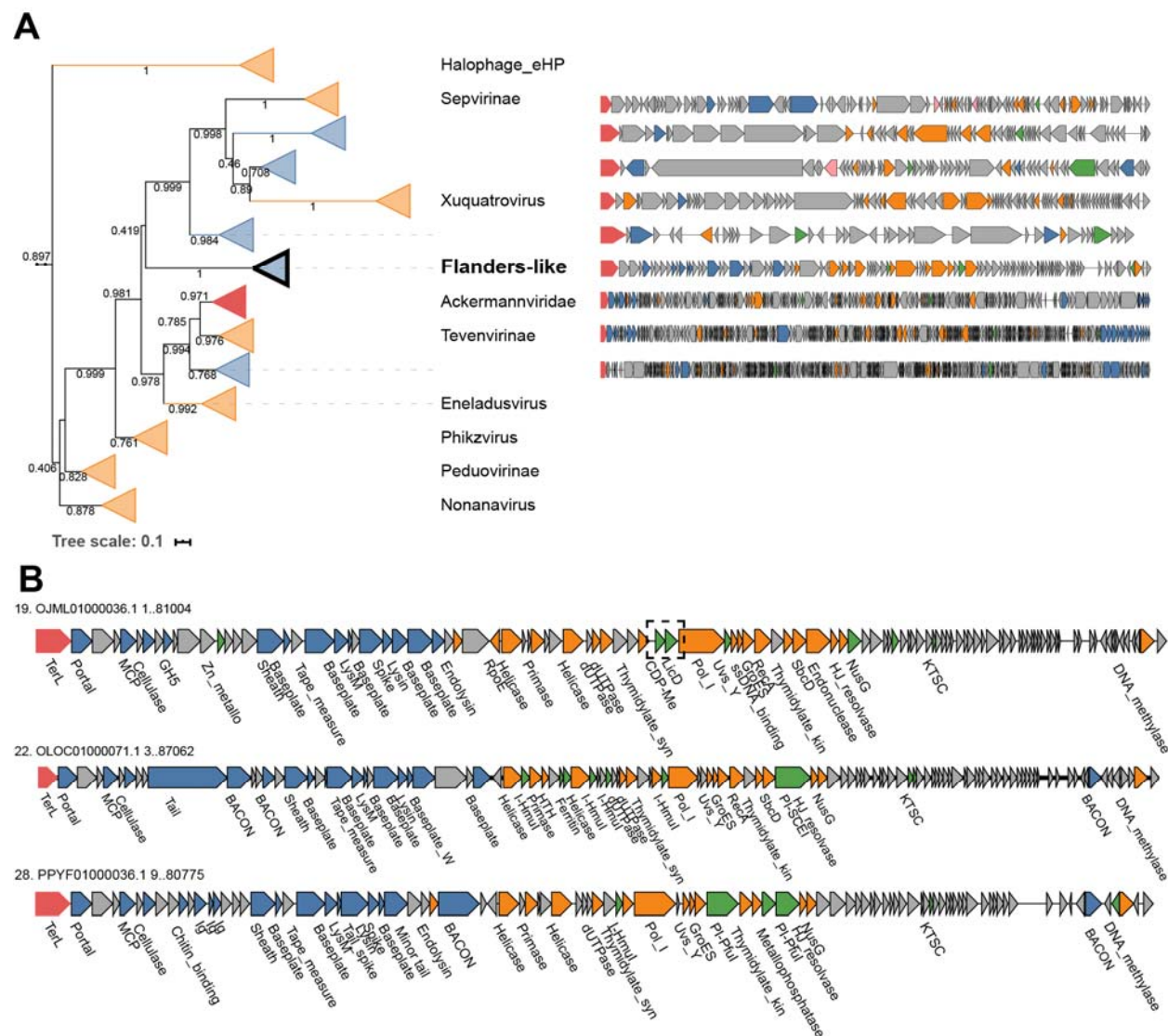
652   **Figure 3**: Phylogenetic tree of the large terminase subunit and complete genome maps for
653   "Flandersviridae". **(A)** Genome maps of members of the "Flandersviridae and selected ICTV-
654   classified phages were constructed and colored as in Figure 3. **(B)** Genome maps of three genera
655   from the "Flandersviridae" family. The dashed box highlights the insertion of licD- and ispD-
656   family enzymes in the replication module of one "Flandersviridae" phage.

657

658

659

660

661

Figure 4: Phylogenetic tree of the large terminase subunit and genome maps of the "Gratiaviridae" phages. (A) Genome maps of ICTV-classified phages were constructed and colored as in Figure 3. (B) Genome maps of four genera from the Gratiaviridae family. The dashed box highlights a HipA-family kinase domain-containing protein, AAA-family ATPase and glycosyltransferase (see main text).

667

668

Additional files

Additional file 1 – Clustering information for 3,378 gut phages identified in the study. Phage genomes were dereplicated at 95% identity over 80% of the contig length. The GenBank accession codes for each phage and the representative sequence is provided in columns 1 and 3, respectively (.csv).

Additional file 2 – Taxonomic information for 1,886 representative phage genomes. Each taxonomically classified phage genome was scored by ViralVerify, Seeker and vcontact2 and the assignments are provided in columns 9-10, 11 and 12-13, respectively (.csv).

Additional file 3 - Distribution of the marker profiles identified on each phage genome and pie chart representation of the gut phage taxonomy. (**A**) Histogram of marker proteins detected on each phage genome recovered from human gut metagenomes. Abbreviations are as follows: M, major capsid protein; P, portal, T, terminase large subunit. (**B**) Taxonomic assignments of the dereplicated contigs (n = 1,886), with the outermost ring corresponding to ICTV families (.pdf).

Additional file 4 – Host ranges inferred from CRISPR-spacer matches. The nucleotide coordinates of the protospacer are provided in columns 2 and 3. The sequence of the CRISPR spacer and protospacer are provided in columns 4 and 5. The taxonomic information of the host is listed in the subsequent columns (.csv).

Additional file 5 – Candidate anti-CRISPR (Acr) proteins encoded on each phage genome. Proteins predicted to be an Acr with high confidence (score > 0.9) and meeting all other heuristic criteria are listed along with their inferred host. For the *Bifidobacteria* phages (see the main text), the nucleotide coordinates of closely related prophages integrated in their host genome are provided in columns 6-8 (.csv)

Additional file 6 - Genome maps of predicted anti-CRISPR proteins (Acrs) in uncharacterized *Bifidobacteria* phages. Open reading frames are colored according to function: large terminase subunit (red), structural components (blue), replication (orange), integrase (pink), general function (green) and unknown (grey). The candidate Acrs are indicated with a dashed box (.pdf)

Additional file 7 - Coverage heatmap of a "Quimbyviridae" genome across human gut viromes. The coverage of the most abundant "Quimbyviridae" phage genome (accession OMAC01000147.1) is plotted as a heatmap, scaled from 0 – 100x per 100 bp window (.pdf).

Additional file 8 - Phylogenetic tree of the MCP, primase and portal proteins encoded by "Quimbyviridae" phages (.pdf).

Additional file 9 - Alignment of the template and variable repeats from the Quimbyvirus DGRs. The template repeat (TR) from Quimbyvirus is the first listed sequence and is followed by the variable repeats from either ORF80 (VR1) or ORF 47 (VR2) encoded in phage genomes that are nearly identical to Quimbyvirus ($\geq$ 95 % average nucleotide identity). A total of 21 adenine

33

residues (green) in the template repeat exhibit a least one substitution in a corresponding variable repeat (.pdf).

Additional file 10 - HHPred alignments of four Quimbyviridae proteins and one Gratiaviridae protein with their top-scoring templates, including a replication initiator protein, a cytosine-specific methyltransferase, an adenine-specific methyltransferase, a MutY nuclease and HipA kinase (.docx).

Additional file 11 - Phylogenetic tree of the polA and dnaG genes in Flanders-like phages. Branches composed of GenBank phages are colored in orange and branches of gut metagenomic phages in blue. Branches with sequences labelled as bacteria in the GenBank database, likely representing cryptic prophages, are colored in grey (.pdf).

Additional file 12 - Coverage heatmap of a "Flandersviridae" genome across human gut viromes. The coverage of the most abundant "Flandersviridae" phage genome (accession OLOC0100071.1) is plotted as a heatmap, scaled from 0 – 100x fold coverage per 100 bp window (.pdf).

Additional file 13 - Phylogenetic tree of the *dnaG* and *dnaE* genes in Gratiaviridae phages (.pdf).

Additional file 14 – Fasta-formatted alignments and Newick-formatted files underlying all trees generated in the study (.zip).

Additional file 15 – Abundance matrix of phage genomes in human gut viromes. The first column lists the GenBank accessions of each reference phage genome and the first row lists the SRA accessions for each virome. Values represent the length-normalized fractional abundance of each phage in each virome (see methods) (.zip).

References

1.   Cobián Güemes AG, Youle M, Cantú VA, Felts B, Nulton J, Rohwer F: **Viruses as Winners in the Game of Life**. *Annual Review of Virology* 2016, **3**(1):197-214.

2.   Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI: **Viruses in the faecal microbiota of monozygotic twins and their mothers**. *Nature* 2010, **466**(7304):334-338.

3.   Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F: **Metagenomic analyses of an uncultured viral community from human feces**. *Journal of bacteriology* 2003, **185**(20):6220-6223.

4.   Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH: **Global Organization and Proposed Megataxonomy of the Virus World**. *Microbiology and Molecular Biology Reviews* 2020, **84**(2):e00061-00019.

5.   Barr JJ, Auro R, Furlan M, Whiteson KL, Erb ML, Pogliano J, Stotland A, Wolkowicz R, Cutting AS, Doran KS *et al*: **Bacteriophage adhering to mucus provide a non–host-derived immunity**. *Proceedings of the National Academy of Sciences* 2013, **110**(26):10771-10776.

6.   Roach DR, Leung CY, Henry M, Morello E, Singh D, Di Santo JP, Weitz JS, Debarbieux L: **Synergy between the Host Immune System and Bacteriophage Is Essential for Successful Phage Therapy against an Acute Respiratory Pathogen**. *Cell Host & Microbe* 2017, **22**(1):38-47.e34.

7.   Moreno-Gallego JL, Chou S-P, Di Rienzi SC, Goodrich JK, Spector TD, Bell JT, Youngblut ND, Hewson I, Reyes A, Ley RE: **Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins**. *Cell Host & Microbe* 2019, **25**(2):261-272.e265.

8.   Gogokhia L, Buhrke K, Bell R, Hoffman B, Brown DG, Hanke-Gogokhia C, Ajami NJ, Wong MC, Ghazaryan A, Valentine JF *et al*: **Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis**. *Cell Host & Microbe* 2019, **25**(2):285-299.e288.

9.   Waller AS, Yamada T, Kristensen DM, Kultima JR, Sunagawa S, Koonin EV, Bork P: **Classification and quantification of bacteriophage taxa in human gut metagenomes**. *The ISME Journal* 2014, **8**(7):1391-1402.

10.   Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD: **Rapid evolution of the human gut virome**. *Proceedings of the National Academy of Sciences* 2013, **110**(30):12450-12455.

11.   Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, McDonnell SA, Khokhlova EV, Draper LA, Forde A *et al*: **The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific**. *Cell Host & Microbe* 2019, **26**(4):527-541.e525.

12.   Casjens SR, Gilcrease EB: **Determining DNA Packaging Strategy by Analysis of the Termini of the Chromosomes in Tailed-Bacteriophage Virions**. In: *Bacteriophages Methods in Molecular Biology*. Edited by Clokie M.R., A.M. K, vol. 502: Humana Press; 2009.

13.   Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A, Castelle CJ, Olm MR, Bouma-Gregson K, Amano Y *et al*: **Clades of huge phages from across Earth's ecosystems**. *Nature* 2020, **578**(7795):425-431.

14.   Shkoporov AN, Ryan FJ, Draper LA, Forde A, Stockdale SR, Daly KM, McDonnell SA, Nolan JA, Sutton TDS, Dalmasso M *et al*: **Reproducible protocols for metagenomic analysis of human faecal phageomes**. *Microbiome* 2018, **6**(1):68.

15.   Antipov D, Raiko M, Lapidus A, Pevzner PA: **Plasmid detection and assembly in genomic and metagenomic data sets**. *Genome Research* 2019, **29**(6):961-968.

16.   Antipov D, Raiko M, Lapidus A, Pevzner PA: **metaviralSPAdes: assembly of viruses from metagenomic data**. *Bioinformatics* 2020.

17.   Gideon SG, Wright A: **DNA Segregation in Bacteria**. *Annual Review of Microbiology* 2000, **54**(1):681-708.

18.   De Sordi L, Lourenço M, Debarbieux L: **The Battle Within: Interactions of Bacteriophages and Bacteria in the Gastrointestinal Tract**. *Cell Host & Microbe* 2019, **25**(2):210-218.

19.   Beller L, Matthijnssens J: **What is (not) known about the dynamics of the human gut virome in health and disease**. *Current Opinion in Virology* 2019, **37**:52-57.

20.   Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B *et al*: **Virus taxonomy in the age of metagenomics**. *Nature Reviews Microbiology* 2017, **15**(3):161-168.

21.   Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS *et al*: **CDD/SPARCLE: the conserved domain database in 2020**. *Nucleic Acids Research* 2019, **48**(D1):D265-D268.

22.   Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification**. *BMC Bioinformatics* 2010, **11**(1):119.

23.   Lowe TM, Eddy SR: **tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence**. *Nucleic Acids Research* 1997, **25**(5):955-964.

24.   Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, Archie EA, Turnbaugh PJ, Seed KD, Blekhman R *et al*: **Megaphages infect Prevotella and variants are widespread in gut microbiomes**. *Nature Microbiology* 2019, **4**(4):693-700.

25.   Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, Visel A, Woyke T, Kyrpides NC, Rubin EM: **Stop codon reassignments in the wild**. *Science* 2014, **344**(6186):909-913.

26.   Auslander N, Gussow AB, Benler S, Wolf YI, Koonin EV: **Seeker: Alignment-free identification of bacteriophage genomes by deep learning**. *bioRxiv* 2020:2020.2004.2004.025783.

27.   Kans J: **Entrez Direct: E-utilities on the Unix Command Line**. Bethesda (MD): National Center for Biotechnology Information; 2010.

28.   Olm MR, Brown CT, Brooks B, Banfield JF: **dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication**. *The ISME Journal* 2017, **11**(12):2864-2868.

29.   Ondov BD, Starrett GJ, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM: **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biology* 2019, **20**(1):232.

30.   Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S: **High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries**. *Nature Communications* 2018, **9**(1):5114.

31.   Steinegger M, Söding J: **Clustering huge protein sequence sets in linear time**. *Nature Communications* 2018, **9**(1):2542.

32.   Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Research* 1997, **25**(17):3389-3402.

33.   Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, Dolja VV, Koonin EV: **Origins and Evolution of the Global RNA Virome**. *mBio* 2018, **9**(6):e02329-02318.

34.   Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Research* 2004, **32**(5):1792-1797.

35.   Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J: **HH-suite3 for fast remote homology detection and deep protein annotation**. *BMC Bioinformatics* 2019, **20**(1):473.

36.   Yutin N, Makarova KS, Mekhedov SL, Wolf YI, Koonin EV: **The Deep Archaeal Roots of Eukaryotes**. *Molecular Biology and Evolution* 2008, **25**(8):1619-1630.

37.   Price MN, Dehal PS, Arkin AP: **FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments**. *PLOS ONE* 2010, **5**(3):e9490.

38.   Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R *et al*: **Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks**. *Nature Biotechnology* 2019, **37**(6):632-639.

39.   Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks**. *Genome Research* 2003, **13**(11):2498-2504.

40.   Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M: **Uniclust databases of clustered and deeply annotated protein sequences and alignments**. *Nucleic Acids Research* 2016, **45**(D1):D170-D176.

41.   Ye Y: **Identification of Diversity-Generating Retroelements in Human Microbiomes**. *International Journal of Molecular Sciences* 2014, **15**(8).

42.   Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012, **9**(4):357-359.

43.   Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics* 2018, **34**(17):i884-i890.

44.   Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV: **The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes**. *mBio* 2017, **8**(5):e01397-01317.

45.   Shmakov SA, Wolf YI, Savitskaya E, Severinov KV, Koonin EV: **Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes**. *Communications Biology* 2020, **3**(1):321.

46.   Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA: **Database indexing for production MegaBLAST searches**. *Bioinformatics* 2008, **24**(16):1757-1764.

47.   Gussow AB, Park AE, Borges AL, Shmakov SA, Makarova KS, Wolf YI, Bondy-Denomy J, Koonin EV: **Machine-learning approach expands the repertoire of anti-CRISPR protein families**. *Nature Communications* 2020, **11**(1):3784.

48.   Hynes AP, Rousseau GM, Agudelo D, Goulet A, Amigues B, Loehr J, Romero DA, Fremaux C, Horvath P, Doyon Y *et al*: **Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins**. *Nature Communications* 2018, **9**(1):2919.

49.   Marino ND, Zhang JY, Borges AL, Sousa AA, Leon LM, Rauch BJ, Walton RT, Berry JD, Joung JK, Kleinstiver BP *et al*: **Discovery of widespread type I and type V CRISPR-Cas inhibitors**. *Science* 2018, **362**(6411):240-242.

50.   Pawluk A, Davidson AR, Maxwell KL: **Anti-CRISPR: discovery, mechanism and function**. *Nature Reviews Microbiology* 2018, **16**(1):12-17.

51.   Hwang S, Maxwell KL: **Meet the Anti-CRISPRs: Widespread Protein Inhibitors of CRISPR-Cas Systems**. *The CRISPR Journal* 2019, **2**(1):23-30.

52.   Osuna BA, Karambelkar S, Mahendra C, Christie KA, Garcia B, Davidson AR, Kleinstiver BP, Kilcher S, Bondy-Denomy J: **Listeria Phages Induce Cas9 Degradation to Protect Lysogenic Genomes**. *Cell Host & Microbe* 2020.

53.   Osuna BA, Karambelkar S, Mahendra C, Sarbach A, Johnson MC, Kilcher S, Bondy-Denomy J: **Critical Anti-CRISPR Locus Repression by a Bi-functional Cas9 Inhibitor**. *Cell Host & Microbe* 2020.

54.   Stanley SY, Borges AL, Chen K-H, Swaney DL, Krogan NJ, Bondy-Denomy J, Davidson AR: **Anti-CRISPR-associated proteins are crucial repressors of anti-CRISPR transcription**. *Cell* 2019, **178**(6):1452-1464. e1413.

55.   Koonin EV, Yutin N: **The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome**. *Trends in Microbiology* 2020, **28**(5):349-359.

56. Iranzo J, Krupovic M, Koonin EV: **The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing**. *mBio* 2016, **7**(4):e00978-00916.

57. Aiewsakun P, Adriaenssens EM, Lavigne R, Kropinski AM, Simmonds P: **Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy**. *Journal of General Virology* 2018, **99**(9):1331-1343.

58. Barylski J, Enault F, Dutilh BE, Schuller MB, Edwards RA, Gillis A, Klumpp J, Knezevic P, Krupovic M, Kuhn JH *et al*: **Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Phages**. *Systematic Biology* 2019, **69**(1):110-123.

59. Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, Koonin EV: **Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut**. *Nature Microbiology* 2018, **3**(1):38-46.

60. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK *et al*: **A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes**. *Nature Communications* 2014, **5**(1):4498.

61. Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, Dinsdale EA, Cinek O, Aziz RK, McNair K, Barr JJ *et al*: **Global phylogeography and ancient evolution of the widespread human gut virus crAssphage**. *Nature Microbiology* 2019, **4**(10):1727-1736.

62. Schumacher MA, Tonthat NK, Kwong SM, Chinnam Nb, Liu MA, Skurray RA, Firth N: **Mechanism of staphylococcal multiresistance plasmid replication origin assembly by the RepA protein**. *Proceedings of the National Academy of Sciences* 2014, **111**(25):9121-9126.

63. Botstein D, Lew KK, Jarvik V, Swanson CA: **Role of antirepressor in the bipartite control of repression and immunity by bacteriophage P22**. *Journal of Molecular Biology* 1975, **91**(4):439-462.

64. Argov T, Sapir SR, Pasechnek A, Azulay G, Stadnyuk O, Rabinovich L, Sigal N, Borovok I, Herskovits AA: **Coordination of cohabiting phage elements supports bacteria–phage cooperation**. *Nature Communications* 2019, **10**(1).

65. Davis-Richardson AG, Ardissone AN, Dias R, Simell V, Leonard MT, Kemppainen KM, Drew JC, Schatz D, Atkinson MA, Kolaczkowski B *et al*: **Bacteroides dorei dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes**. *Frontiers in Microbiology* 2014, **5**(678).

66. Campbell A: **Comparative Molecular Biology of Lambdoid Phages**. 1994, **48**(1):193-222.

67. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, Khan H, Liu M, Pohlschroder M, Shaw KL *et al*: **Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey**. *Nucleic Acids Research* 2017, **46**(1):11-24.

68. Handa S, Jiang Y, Tao S, Foreman R, Schinazi RF, Miller JF, Ghosh P: **Template-assisted synthesis of adenine-mutagenized cDNA by a retroelement protein complex**. *Nucleic Acids Research* 2018, **46**(18):9711-9725.

69. Miller JL, Coq JL, Hodes A, Barbalat R, Miller JF, Ghosh P: **Selective Ligand Recognition by a Diversity-Generating Retroelement Variable Protein**. *PLOS Biology* 2008, **6**(6):e131.

70. Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, Preston A, Maskell DJ, Simons RW, Cotter PA, Parkhill J *et al*: **Reverse Transcriptase-Mediated Tropism Switching in Bordetella Bacteriophage**. *Science* 2002, **295**(5562):2091-2094.

71. Xu J, Hendrix RW, Duda RL: **Conserved Translational Frameshift in dsDNA Bacteriophage Tail Assembly Genes**. *Molecular Cell* 2004, **16**(1):11-21.

72. Levin ME, Hendrix RW, Casjens SR: **A Programmed Translational Frameshift is Required for the Synthesis of a Bacteriophage λ Tail Assembly Protein**. *Journal of Molecular Biology* 1993, **234**(1):124-139.

73.   Baranov PV, Fayet O, Hendrix RW, Atkins JF: **Recoding in bacteriophages and bacterial IS elements**. *Trends in Genetics* 2006, **22**(3):174-181.

74.   McNair K, Zhou C, Dinsdale EA, Souza B, Edwards RA: **PHANOTATE: a novel approach to gene identification in phage genomes**. *Bioinformatics* 2019, **35**(22):4537-4542.

75.   Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, Simons RW, Zimmerly S, Miller JF: **Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements**. *Nature* 2004, **431**(7007):476-481.

76.   Meineke B, Shuman S: **Determinants of the cytotoxicity of PrrC anticodon nuclease and its amelioration by tRNA repair**. *RNA* 2012, **18**(1):145-154.

77.   de Jonge PA, von Meijenfeldt FAB, van Rooijen LE, Brouns SJJ, Dutilh BE: **Evolution of BACON Domain Tandem Repeats in crAssphage and Novel Gut Bacteriophage Lineages**. *Viruses* 2019, **11**(12):1085.

78.   Pell LG, Gasmi-Seabrook GMC, Morais M, Neudecker P, Kanelis V, Bona D, Donaldson LW, Edwards AM, Howell PL, Davidson AR *et al*: **The Solution Structure of the C-Terminal Ig-like Domain of the Bacteriophage λ Tail Tube Protein**. *Journal of Molecular Biology* 2010, **403**(3):468-479.

79.   Fraser JS, Yu Z, Maxwell KL, Davidson AR: **Ig-Like Domains on Bacteriophages: A Tale of Promiscuity and Deceit**. *Journal of Molecular Biology* 2006, **359**(2):496-507.

80.   Barr JJ, Auro R, Sam-Soon N, Kassegne S, Peters G, Bonilla N, Hatay M, Mourtada S, Bailey B, Youle M *et al*: **Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters**. *Proceedings of the National Academy of Sciences* 2015, **112**(44):13675-13680.

81.   Nasko DJ, Chopyk J, Sakowski EG, Ferrell BD, Polson SW, Wommack KE: **Family A DNA Polymerase Phylogeny Uncovers Diversity and Replication Gene Organization in the Virioplankton**. *Frontiers in Microbiology* 2018, **9**(3053).

82.   Campbell TL, Brown ED: **Characterization of the Depletion of 2-C-Methyl-d-Erythritol-2,4-Cyclodiphosphate Synthase in Escherichia coli and Bacillus subtilis**. *Journal of Bacteriology* 2002, **184**(20):5609-5618.

83.   Heuston S, Begley M, Gahan CGM, Hill C: **Isoprenoid biosynthesis in bacterial pathogens**. *Microbiology* 2012, **158**(6):1389-1401.

84.   Zhang J-R, Idanpaan-Heikkila I, Fischer W, Tuomanen EI: **Pneumococcal licD2 gene is involved in phosphorylcholine metabolism**. *Molecular Microbiology* 1999, **31**(5):1477-1488.

85.   Lopez R, Garcia E, Garcia P, Ronda C, Tomasz A: **Choline-containing bacteriophage receptors in Streptococcus pneumoniae**. *Journal of Bacteriology* 1982, **151**(3):1581-1590.

86.   Odom AR: **Five Questions about Non-Mevalonate Isoprenoid Biosynthesis**. *PLOS Pathogens* 2011, **7**(12):e1002323.

87.   Breitbart M, Bonnain C, Malki K, Sawaya NA: **Phage puppet masters of the marine microbial realm**. *Nature Microbiology* 2018, **3**(7):754-766.

88.   Łobocka MB, Rose DJ, Plunkett G, Rusin M, Samojedny A, Lehnherr H, Yarmolinsky MB, Blattner FR: **Genome of Bacteriophage P1**. *Journal of Bacteriology* 2004, **186**(21):7032-7068.

89.   McNair K, Bailey BA, Edwards RA: **PHACTS, a computational approach to classifying the lifestyle of phages**. *Bioinformatics* 2012, **28**(5):614-618.

90.   Holmfeldt K, Solonenko N, Shah M, Corrier K, Riemann L, VerBerkmoes NC, Sullivan MB: **Twelve previously unknown phage genera are ubiquitous in global oceans**. *Proceedings of the National Academy of Sciences* 2013, **110**(31):12798-12803.

91.   Kyrkou I, Byth Carstens A, Ellegaard-Jensen L, Kot W, Zervas A, Djurhuus AM, Neve H, Hansen M, Hestbjerg Hansen L: **Expanding the Diversity of Myoviridae Phages Infecting Lactobacillus**

plantarum—A Novel Lineage of Lactobacillus Phages Comprising Five New Members. *Viruses* 2019, **11**(7):611.

92.	Rocha ER, Selby T, Coleman JP, Smith CJ: **Oxidative stress response in an anaerobe, Bacteroides fragilis: a role for catalase in protection against hydrogen peroxide**. *Journal of Bacteriology* 1996, **178**(23):6895-6903.

93.	Betteken MI, Rocha ER, Smith CJ: **Dps and DpsL Mediate Survival In Vitro and In Vivo during the Prolonged Oxidative Stress Response in Bacteroides fragilis**. *Journal of Bacteriology* 2015, **197**(20):3329-3338.

94.	Gauss GH, Reott MA, Rocha ER, Young MJ, Douglas T, Smith CJ, Lawrence CM: **Characterization of the Bacteroides fragilis Gene Product Identifies a Bacterial DPS-Like Protein and Suggests Evolutionary Links in the Ferritin Superfamily**. *Journal of Bacteriology* 2012, **194**(1):15-27.

95.	Germain E, Castro-Roa D, Zenkin N, Gerdes K: **Molecular Mechanism of Bacterial Persistence by HipA**. *Molecular Cell* 2013, **52**(2):248-254.

96.	Huang CY, Gonzalez-Lopez C, Henry C, Mijakovic I, Ryan KR: **hipBA toxin-antitoxin systems mediate persistence in Caulobacter crescentus**. *Scientific Reports* 2020, **10**(1):2865.

97.	Rahmsdorf HJ, Pai SH, Ponta P, Herrlich P, Roskoski R, Schweiger M, Studier FW: **Protein Kinase Induction in Escherichia coli by Bacteriophage T7**. *Proceedings of the National Academy of Sciences* 1974, **71**(2):586-589.

98.	Gone S, Alfonso-Prieto M, Paudyal S, Nicholson AW: **Mechanism of Ribonuclease III Catalytic Regulation by Serine Phosphorylation**. *Scientific Reports* 2016, **6**(1):25448.

99.	Severinova E, Severinov K: **Localization of the Escherichia coli RNA Polymerase β′ Subunit Residue Phosphorylated by Bacteriophage T7 Kinase Gp0.7**. *Journal of Bacteriology* 2006, **188**(10):3470-3476.

100.	Friedman DI, Mozola CC, Beeri K, Ko C-C, Reynolds JL: **Activation of a prophage-encoded tyrosine kinase by a heterologous infecting phage results in a self-inflicted abortive infection**. *Molecular Microbiology* 2011, **82**(3):567-577.

101.	Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Rüger W: **Bacteriophage T4 Genome**. *Microbiology and Molecular Biology Reviews* 2003, **67**(1):86-156.

102.	Casjens S: **Prophages and bacterial genomics: what have we learned so far?** *Molecular Microbiology* 2003, **49**(2):277-300.

103.	Whitfield C, Wear SS, Sande C: **Assembly of Bacterial Capsular Polysaccharides and Exopolysaccharides**. *Annual Review of Microbiology* 2020, **74**(1):521-543.

104.	Porter NT, Hryckowian AJ, Merrill BD, Fuentes JJ, Gardner JO, Glowacki RWP, Singh S, Crawford RD, Snitkin ES, Sonnenburg JL et al: **Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in Bacteroides thetaiotaomicron**. *Nature Microbiology* 2020, **5**(9):1170-1181.

105.	Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD: **Hypervariable loci in the human gut virome**. *Proceedings of the National Academy of Sciences* 2012, **109**(10):3962-3966.

106.	Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS et al: **Structure, function and diversity of the healthy human microbiome**. *Nature* 2012, **486**(7402):207-214.

107.	Hryckowian AJ, Merrill BD, Porter NT, Van Treuren W, Nelson EJ, Garlena RA, Russell DA, Martens EC, Sonnenburg JL: **Bacteroides thetaiotaomicron-Infecting Bacteriophage Isolates Inform Sequence-Based Host Range Predictions**. *Cell Host & Microbe* 2020, **28**(3):371-379.e375.

108.	Roux S, Paul BG, Bagby SC, Allen MA, Attwood G, Cavicchioli R, Chistoserdova L, Hallam SJ, Hernandez ME, Hess M et al: **Ecology and molecular targets of hypermutation in the global microbiome**. *bioRxiv* 2020:2020.2004.2001.020958.

109.   Cornuault JK, Petit M-A, Mariadassou M, Benevides L, Moncaut E, Langella P, Sokol H, De Paepe M: **Phages infecting Faecalibacterium prausnitzii belong to novel viral genera that help to decipher intestinal viromes**. *Microbiome* 2018, **6**(1):65.

110.   Benler S, Cobián-Güemes AG, McNair K, Hung S-H, Levi K, Edwards R, Rohwer F: **A diversity-generating retroelement encoded by a globally ubiquitous Bacteroides phage**. *Microbiome* 2018, **6**(1):191.

111.   Kaur G, Burroughs AM, Iyer LM, Aravind L: **Highly regulated, diversifying NTP-dependent biological conflict systems with implications for the emergence of multicellularity**. *Elife* 2020, **9**.

112.   Vallota-Eastman A, Arrington EC, Meeken S, Roux S, Dasari K, Rosen S, Miller JF, Valentine DL, Paul BG: **Role of Diversity-Generating Retroelements for Regulatory Pathway Tuning in Cyanobacteria**. *bioRxiv* 2020:2020.2005.2026.117283.

113.   Valguarnera E, Wardenburg JB: **Good Gone Bad: One Toxin Away From Disease for Bacteroides fragilis**. *Journal of Molecular Biology* 2020, **432**(4):765-785.

114.   Friedman ES, Bittinger K, Esipova TV, Hou L, Chau L, Jiang J, Mesaros C, Lund PJ, Liang X, FitzGerald GA *et al*: **Microbes vs. chemistry in the origin of the anaerobic gut lumen**. *Proceedings of the National Academy of Sciences* 2018, **115**(16):4170-4175.

115.   Albenberg L, Esipova TV, Judge CP, Bittinger K, Chen J, Laughlin A, Grunberg S, Baldassano RN, Lewis JD, Li H *et al*: **Correlation Between Intraluminal Oxygen Gradient and Radial Partitioning of Intestinal Microbiota**. *Gastroenterology* 2014, **147**(5):1055-1063.e1058.

116.   Donaldson GP, Lee SM, Mazmanian SK: **Gut biogeography of the bacterial microbiota**. *Nature Reviews Microbiology* 2016, **14**(1):20-32.