Shallow metagenomics with Nanopore sequencing in canine fecal microbiota improved bacterial taxonomy and identified an *uncultured CrAssphage*

Anna Cuscó[1*], Anna Salas[2], Celina Torre[2], Olga Francino[3]

[1]Vetgenomics, Ed Eureka, Parc de Recerca UAB, Barcelona, Spain

[2]Affinity Petcare, Research and Development, Pl d'Europa, 54, 08902 L'Hospitalet de Llobregat, Barcelona, Spain

[3]Molecular Genetics Veterinary Service (SVGM), Veterinary School, Universitat Autònoma de Barcelona, Barcelona, Spain

*Corresponding author: Anna Cuscó, anna.cusco@vetgenomics.com

Keywords: MinION, nanopore, microbiota, metagenomics, dog, canine, fecal, crAssphage, bacteriophage

Abstract

Long-read metagenomics –using single-molecule sequencers– has the potential to assembly entire genomes, even from complex metagenomics datasets. Using long-read metagenomics with Nanopore sequencing in pooled samples, we aim to improve the individual taxonomic profiles obtained with V4 16S rRNA massive sequencing and to assemble the fecal metagenome of healthy dogs.

Fecal samples from healthy dogs were sequenced individually using V4 16S rRNA gene and in pools using a shallow metagenomics approach with Nanopore sequencing. Long-read metagenomics allowed us refining the V4 16S rRNA results up to the species level and determining the main bacterial species inhabiting on fecal microbiota of our cohort of healthy dogs. Among them, the most abundant were *Fusobacterium varium*; *Megamonas hypermegale*; *Bacteroides vulgatus*; *Blautia hansenii*; *Clostridium perfringens*; and *Clostridoides difficile*. Once performed the metagenome assembly, one contig was suggested to be circular and hit to an *uncultured crAssphage*.

To conclude, shallow long-read metagenomics with pooled samples using MinION allowed characterizing the dog fecal microbiota at lower taxonomic level, such as bacterial species. The assembly of the reads retrieved a contig that represents a circular draft genome of an uncultured *CrAssphage* from dog fecal samples that is one of the most abundant bacteriophages in the human gut.

Introduction

Metagenomics allows observing microorganisms from the whole tree of life, going beyond bacteria. The most common approaches rely on shotgun metagenomics and massive sequencing technologies, which sample the DNA from a specific environment, fragment it and sequence the small fragments with the aim to solve the puzzle (1). Single molecule sequencing platforms such as those developed by PacBio and Oxford Nanopore Technologies allow sequencing with longer reads, despite presenting lower accuracy (2).

Long-read metagenomics allows easier assembly of entire genomes, even from complex metagenomics datasets. Nanopore sequencing combined with short-read polishing has allowed the assembly of complete or nearly complete circular genomes retrieved from several metagenomics samples. On one hand, this approach has been tested in bacterial (3,4) and viral mock communities (5). On the other hand, it has also been applied on real samples: i) to characterize human fecal metagenome (4,6,7); ii) to detect viruses on clinical samples (8); and iii) to characterize other environments, such as an enrichment reactor (9) or the cow rumen (10).

Until the date, two studies have characterized dog fecal microbiota using shotgun metagenomics, and both of them showed similarities to human gut microbiota (11,12). We aim to profile the canine fecal microbiota at lower taxonomical level -such as

species- performing low deep metagenomics with Nanopore long-reads to complement the individual taxonomic profiles obtained with V4 16S rRNA sequencing. We performed the long-read metagenomics with pooled samples rather than individual ones to improve the cost/benefit ratio of our approach.

After assembling the fecal metagenome of healthy dogs, we retrieved several contigs, one of them suggested being circular and blasting to an *uncultured crAssphage*. *Uncultured crAssphage* and crAss–like bacteriophages are the most abundant bacteriophages in the human gut (13–16). This bacteriophage was discovered *in silico* in 2014 when screening multiple metagenomics sets (13). A crass-like bacteriophage was isolated for the first time in 2018 (17).

Material and methods

Sample collection and DNA extraction

Fecal samples were collected and frozen at -80ºC until the DNA extraction. Bacterial DNA was extracted from 200 mg of fecal sample using the ZymoBIOMICS™ DNA Microprep Kit under manufacturer's conditions. To assess for contamination from the laboratory or reagents, one blank sample was processed at the same time.

V4 16S rRNA PCR, massive sequencing and bioinformatics analyses

V4 hypervariable region of 16S rRNA gene was amplified for each dog fecal sample using the Phusion Hot Start II High-Fidelity DNA polymerase (Thermo Scientific). The samples were amplified using the barcoded forward primer 515F and reverse primer R806, with sequencing adaptors at the 5′ end. One no template control (NTC) sample was included in each PCR reaction to assess if there was contamination. Each PCR reaction contained RNAse and DNAse free water, 5x Phusion Buffer HF (5 μl), dNTPs 2mM (2.5 μl), Primer Fw 10μM (1.25 μl), Primer Rv 10μM (1.75 μl), Phusion High Fidelity Taq Polymerase 2 U/μl (0.25 μl) and 5 ng of DNA. The thermal profile consisted of an initial denaturation of 30 sec at 98˚C, followed by 30 cycles of 15 sec at ˚98 C, 15 sec at 50˚C, 20 sec at 72˚C, and a final extension of 7 min at 72˚C.

Quality and quantity of PCR products were determined using Agilent Bioanalyser 2100 and Qubit™ fluorometer, respectively. Samples were sequenced on an Ion Torrent Personal Genome Machine (PGM) with the Ion 318 Chip Kit v2 (Life Technologies) under manufacturer' conditions. Supplementary Table 1 contains information about the sequencing performance.

Demultiplexed fastq reads were analyzed using Quantitative Insight Into Microbial Ecology 2 (QIIME 2) software (18) (https://qiime2.org). DADA2 was used as quality filtering method to denoise, dereplicate single-end sequences and remove chimeras (19). Amplicon Sequence Varians (ASVs) were used to classify the sequences and assign taxonomy, using Greengenes 13.8 (20) at 99% of similarity to reduce redundancy of the

3

database. Finally, chloroplasts were removed from the sequences. Taxonomic composition was plotted using ampvis2 package (21) from R software.

Long read metagenomics, single molecule sequencing and bioinformatics analyses

Long read metagenomics was performed in two pools of DNA. Pool 1 compresses six fecal microbiota samples, while pool 2, three samples. Two samples from both pools come from the same dogs, but at two different time points. DNA samples were pooled considering the initial DNA quality and quantity to obtain an even representation of the samples included. Libraries were prepared with the Rapid Barcoding kit (SQK-RBK004) and sequenced using MinION, following the instructions of the Rapid Barcoding Sequencing protocol of Oxford Nanopore Technologies.

The samples were run using MinKNOWN software. Fast5 files generated by the software were basecalled and demultiplexed (sorted by barcode) using Albacore v2.3.1, obtaining fastq files. By default, sequences with a q-score lower than 7 are sent to fail folder. A second round of demultiplexing was done with Deepbinner (Wick et al., 2018) in which barcodes that agreed with Albacore were kept and the remaining ones were removed. Supplementary Table 1 contains information about the sequencing performance.

Reads shorter than 1,000 bp were removed for further analyses. To assign taxonomy we used *What's in my Pot* (WIMP) (22), a workflow from EPI2ME in the Oxford Nanopore Technologies cloud (based on Centrifuge (23) software) that uses the NCBI RefSeq database. A final filtering step was applied and we kept only those hits with a classification score greater than 300 (23). Taxonomic composition was plotted using ampvis2 package (21) in R software.

Using the reads from both pools to obtain a sufficient number of sequences we performed the assembly using canu (24). We polished the contigs obtained using Nanopolish (25). We focused this first analysis on the predicted circular contig corresponding to an uncultured crAssphage. For this contig, Phaster (26) was used to predict the CDS. ORFs were blasted using pBLAST. Descriptions of the protein names were retrieved using UniProt (https://www.uniprot.org/uploadlists/). This contig was mapped using Minimap2 (27) against the crass-like genomes reported by Guerin and collaborators (14).

Data availability

The datasets analyzed during the current study are available in the NCBI Sequence Read Archive, under the Bioproject accession number PRJNA527983.

Results and discussion
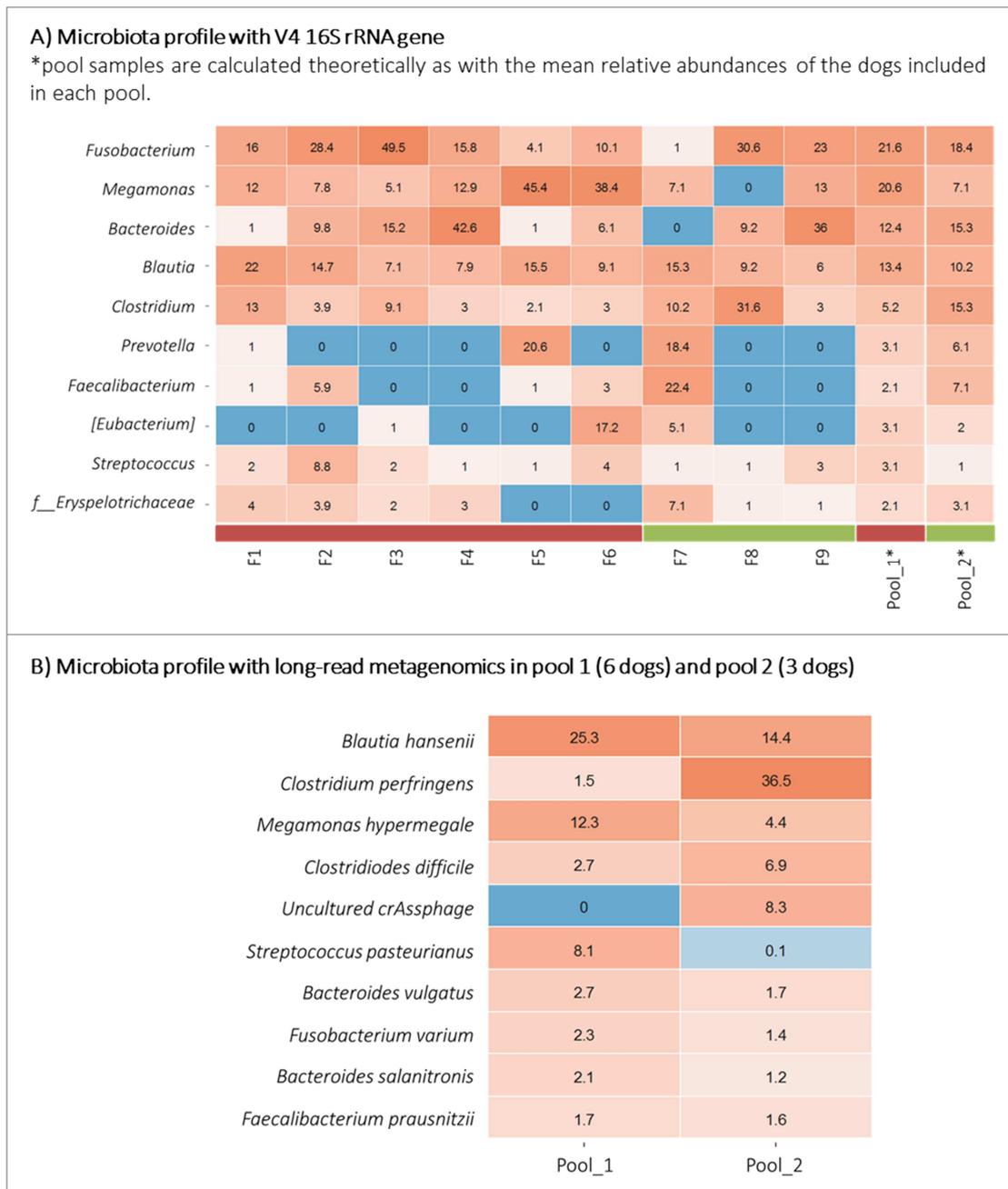
Fecal microbiota of healthy dogs

Fecal samples from healthy dogs were sequenced individually using V4 region amplicons of the 16S rRNA gene (massive sequencing) and in pools using a metagenomics approach with long-reads (single molecule sequencing by ONT). Results for both approaches are shown in Figure 1.

Results obtained for the 16S-V4 amplicon approach show that the fecal microbiota of healthy dogs from our cohort is composed by bacteria from the genera *Fusobacterium*, *Megamonas*, *Bacteroides*, *Blautia* and *Clostridium,* with relative abundance greater than 1% in each individual (except F7 and F8, for *Bacteroides* and *Megamonas* respectively). Other genera with lower abundance are *Prevotella*, *Faecalibacterium, Eubacterium, Streptococcus* and the family Eryspelotrichaceae.

We can refine microbial profiles retrieved by sequencing short fragments of the 16S rRNA approach down to species level either using long-amplicons of 16S rRNA gene or the 16S-ITS-23S amplicons (28) or using shallow long-read metagenomics. Here we applied the second approach in pools of samples. We pooled the DNA samples that had been previously used for the 16S rRNA gene amplification and short-read sequencing (V4) and we sequenced them with MinION device from ONT in a metagenomics approach. After quality filtering and basecalling, we assigned taxonomy with WIMP (22).

The results obtained in this long-read metagenomics approach revealed that the main bacterial species on the fecal microbiota of this cohort of dogs are: *Fusobacterium varium*; *Megamonas hypermegale*; *Bacteroides vulgatus*; *Blautia hansenii*; and *Clostridium perfringens*, together with *Clostridoides difficile,* as representatives of *Clostridium* genus. Other species are *Streptococcus pasteurianus, Bacteroides salanitronis and Faecalibacterium prausnitzii* (see Figure 1B).

For bacterial genomes that present similar genome sizes, WIMP could be considered a semi-quantitative approach. Even though, we observed some divergences on the relative abundances when comparing V4 16S rRNA gene approach and long-read metagenomics. Some of the divergences could also be due to the variable number of 16S rRNA gene copies depending on the bacterial species (29) or to the bias associated with the universal primers chosen for the amplification (30).

5

A) Microbiota profile with V4 16S rRNA gene
*pool samples are calculated theoretically as with the mean relative abundances of the dogs included in each pool.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	Pool_1*	Pool_2*
Fusobacterium	16	28.4	49.5	15.8	4.1	10.1	1	30.6	23	21.6	18.4
Megamonas	12	7.8	5.1	12.9	45.4	38.4	7.1	0	13	20.6	7.1
Bacteroides	1	9.8	15.2	42.6	1	6.1	0	9.2	36	12.4	15.3
Blautia	22	14.7	7.1	7.9	15.5	9.1	15.3	9.2	6	13.4	10.2
Clostridium	13	3.9	9.1	3	2.1	3	10.2	31.6	3	5.2	15.3
Prevotella	1	0	0	0	20.6	0	18.4	0	0	3.1	6.1
Faecalibacterium	1	5.9	0	0	1	3	22.4	0	0	2.1	7.1
[Eubacterium]	0	0	1	0	0	17.2	5.1	0	0	3.1	2
Streptococcus	2	8.8	2	1	1	4	1	1	3	3.1	1
f__Eryspelotrichaceae	4	3.9	2	3	0	0	7.1	1	1	2.1	3.1

B) Microbiota profile with long-read metagenomics in pool 1 (6 dogs) and pool 2 (3 dogs)

	Pool_1	Pool_2
Blautia hansenii	25.3	14.4
Clostridium perfringens	1.5	36.5
Megamonas hypermegale	12.3	4.4
Clostridiodes difficile	2.7	6.9
Uncultured crAssphage	0	8.3
Streptococcus pasteurianus	8.1	0.1
Bacteroides vulgatus	2.7	1.7
Fusobacterium varium	2.3	1.4
Bacteroides salanitronis	2.1	1.2
Faecalibacterium prausnitzii	1.7	1.6

Figure 1. Heatmap representing fecal microbiota composition of healthy dogs. Each value represents the relative abundance of a specific taxa. A) Microbiota composition at the genus level as detected by V4 16S rRNA amplification and sequencing. The last two columns represent the theoretical mean value of each pool. B) Microbiota composition at the species level of pool 1 (6 dogs) and pool 2 (3 dogs) as detected by long-read metagenomics.

The metagenomics approach allowed us detecting also non-bacterial members of the fecal microbiota. The most abundant one was *uncultured crAssphage*, representing around 8% of total relative abundance on Pool 2. *CrAssphage* infect bacteria from the *Bacteroides* genus and other Bacteroidetes members (13). Moreover, the first study culturing a *crAssphage* (ΦCrAss001) from human fecal samples showed that it infected *Bacteroides intestinalis* (17).

6

Previous studies on dog fecal shotgun metagenomics suggested that the canine gut microbiota resembles that from humans (11); more recently, Coelho and collaborators showed that canine gut microbiota resembles more to human gut microbiota than the most commonly used animal models, such as pig or mouse (12). Our results suggest another similarity: *uncultured crAssphage* was the most abundant virus in the fecal microbiota of one cohort of healthy dogs, similar to what has been previously reported in humans (13–15).

Assembly of an uncultured *crAssphage* from canine fecal metagenome

Long-read metagenomics allows easier assembly of entire genomes, even from complex metagenomics datasets, as it has been seen in previous studies (3–10). We used the long reads obtained by nanopore sequencing in the metagenomics approach to perform assembly and polishing. Using this strategy we were able to retrieve several contigs. One of the contigs with 107,597 bp and coverage of 6.8x was suggested to be circular, and it blasted to uncultured *crAssphage* (NC_024711.1) with a query coverage of 96% and identity of 96.72% (Figure 2 and Supplementary Table 2). We also aligned our contig with the crass-like genomes from Guerin and collaborators (14) and it mapped with the highest score to ERR844050_ms_1 and to ERR844032_ms_1. Both genomes belong to the candidate AlphacrAssvirinae subfamily that contain the prototypical *crAssphage* (Supplementary Table 2) and present an identity greater than 97% to the reference crAssphage genome (NC_024711.1).

The NCBI reference genome (NC_024711.1) for *uncultured CrAssphage* contained 88 CDS, while preliminary results with Phaster (26) predicted a total of 247 on our contig (Supplementary Table 3). Long-read sequencing technologies allow assembling draft genomes from metagenomes, but they present insertion/deletion errors that affect protein prediction, often truncating them (31), as seen here. This draft assembly will be further improved when combining with short reads in a hybrid assembly approach.

Even having an unpolished genome, we were able to identify the two main genome regions with opposite gene orientation of crAss-like phages (Figure 2B)(14,15) and their associated proteins despite being truncated –splitted in several CDS (Supplementary Table 3). The small region contains proteins involved in replication, while the bigger region encodes for proteins related to transcription and virion assembly. The predicted proteins were found in the same order as previously reported for *CrAssphage* (Supplementary Table 3) (14,15). Further polishing using short reads would allow a better characterization of the *crAssphage* genome found in canine fecal metagenome.
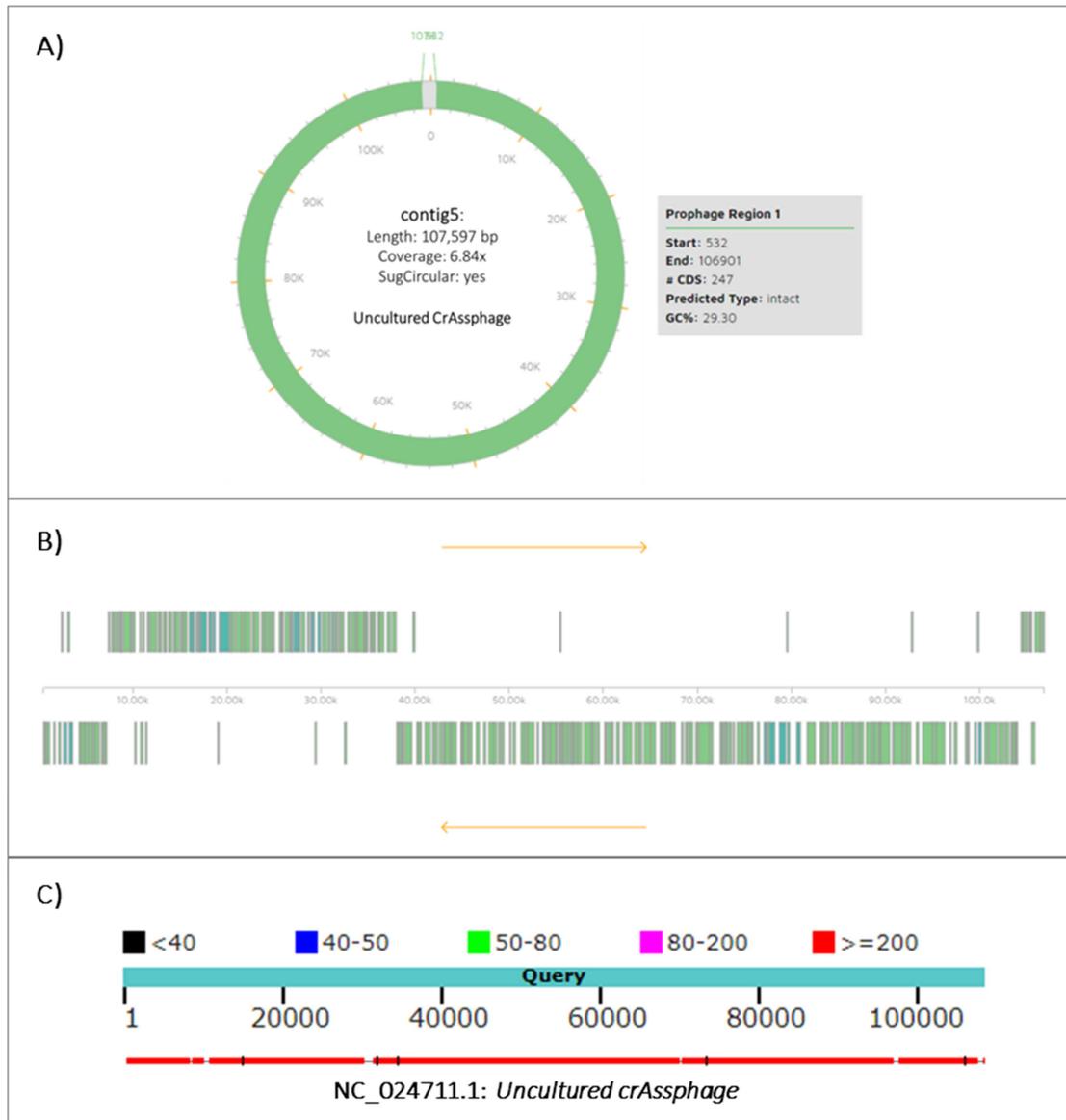
Figure 2. *Uncultured crAssphage* contig. A) Graphical representation of the uncultured *crAssphage* contig retrieved by Phaster, with some summary statistics associated from the assembly with canu. B) Genome representation of the *uncultured crAssphage* contig: green represents hypothetical proteins, and blue represents phage-like proteins as predicted by Phaster (see the detailed ORF predictions in Supplementary file 3). C) BLAST alignment with the Refseq database to *Uncultured crAssphage* (NC_024711.1), Query coverage of 96% and identity of 96.42% (see the complete alignment BLAST results in Supplementary File 2).

Uncultured crAssphage was the most abundant virus previously reported in human fecal metagenomes (13–15) and it is globally spread throughout the industrialized societies (32,33). Since the initial screening of *CrAssphage* via PCR in several animal cohorts was negative, it has been regarded as a potential marker for human fecal pollution, to distinguish the origin of the pollution from other animal sources (34,35). Using this methods, dog fecal samples have already been screened for *CrAssphage*, with negative results (35,36). However, our results suggest that uncultured *crAssphage* can be an abundant member of dog fecal metagenome.

8

Our results would be in line with previous articles that have detected *crAssphage* via qPCR in fecal samples from cats, and wastewater derived from cattle, pigs and poultry (36,37). Since these studies were PCR-based, they could not discard the possibility of a cross-reaction of the PCR to some similar genomic regions of other related phages. CrAssphage has also been detected in one Gorilla that had a close contact with humans (32). Here, we obtained a circular contig identified as an *uncultured crAssphage* and it presented a high average nucleotide identity with *crAssphage* genomes isolated in human feces (Supplementary Table 2). Larger studies screening more dogs and other mammals that have close contact with humans should be performed to confirm our preliminary results and to assess the potential of *crAssphage* to be used as a biomarker of fecal pollution from either human or other species origin.

In conclusion, shallow long-read metagenomics with pooled samples using MinION from Oxford Nanopore Technologies allowed characterizing the dog fecal microbiota at lower taxonomic level, such as bacterial species. The assembly of the reads retrieved a contig that represents a circular draft genome of an uncultured *CrAssphage* from dog fecal samples. Hybrid assembly using shorter reads from massive sequencing technologies will allow a better characterization of the *CrAssphage* found in dogs and approaches including Hi-C could allow determining the host of this bacteriophage in dogs.

Bibliography

1.   Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* (2017) **35**:833–844. doi:10.1038/nbt.3935

2.   Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* (2016) **17**:333–351. doi:10.1038/nrg.2016.49

3.   Nicholls SM, Quick JC, Tang S, Loman NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *bioRxiv* (2018)487033. doi:10.1101/487033

4.   Bertrand D, Shaw J, Narayan M, Ng HQA, Kumar S, Li C, Dvornicic M, Soldo JP, Kho JY, Ng OT, et al. Nanopore sequencing enables high-resolution analysis of resistance determinants and mobile elements in the human gut microbiome. *bioRxiv* (2018)456905. doi:10.1101/456905

5.   Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, Sullivan MB, Temperton B. Long-read viral metagenomics enables capture of abundant and microdiverse viral populations and their niche-defining genomic islands. *bioRxiv* (2018)345041. doi:10.1101/345041

6.   Moss EL, Bhatt AS. Generating closed bacterial genomes from long-read nanopore sequencing of microbiomes. *bioRxiv* (2018)489641. doi:10.1101/489641

7.   Leggett RM, Alcon-Giner C, Heavens D, Caim S, Brook TC, Kujawska M, Martin S, Hoyles L, Clarke P, Hall L, et al. Rapid profiling of the preterm infant gut microbiota using nanopore sequencing aids pathogen diagnostics. *bioRxiv* (2018)180406. doi:10.1101/180406

8.   Kafetzopoulou LE, Efthymiadis K, Lewandowski K, Crook A, Carter D, Osborne J, Aarons E, Hewson R, Hiscox JA, Carroll MW, et al. Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Eurosurveillance* (2018) **23**:1800228. doi:10.2807/1560-7917.ES.2018.23.50.1800228

9.   Arumugam K, Bessarab I, Liu X, Natarajan G, Drautz-Moses DI, Wuertz S, Lauro FM, Law YY, Huson DH, Williams RB. Improving recovery of member genomes from enrichment reactor microbial communities using MinION-based long read metagenomics. *bioRxiv* (2018)465328. doi:10.1101/465328

10.   Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. The genomic and proteomic landscape of the rumen microbiome revealed by comprehensive genome-resolved metagenomics. *bioRxiv* (2018)489443. doi:10.1101/489443

11.   Swanson KS, Dowd SE, Suchodolski JS, Middelbos IS, Vester BM, Barry KA, Nelson KE, Torralba M, Henrissat B, Coutinho PM, et al. Phylogenetic and gene-centric metagenomics of the canine intestinal microbiome reveals similarities with humans and mice. *ISME J* (2011) **5**:639–649. doi:10.1038/ismej.2010.162

12.   Coelho LP, Kultima JR, Costea PI, Fournier C, Pan Y, Czarnecki-Maulden G, Hayward MR, Forslund SK, Schmidt TSB, Descombes P, et al. Similarity of the dog and human gut microbiomes in gene content and response to diet. *Microbiome* (2018) **6**:72. doi:10.1186/s40168-018-0450-3

13.   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. *Nat Commun* (2014) **5**:4498. doi:10.1038/ncomms5498

14.   Guerin E, Shkoporov A, Stockdale SR, Gonzalez-Tortuero E, Ross RP, Hill C. Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* (2018) **24**:653–664. doi:10.1016/j.chom.2018.10.002

15.   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. *Nat Microbiol* (2018) **3**:38–46. doi:10.1038/s41564-017-0053-y

16.   Tamburini FB, Sherlock GJ, Bhatt AS. Transmission and persistence of crAssphage, a ubiquitous human-associated bacteriophage. *bioRxiv* (2018)460113. doi:10.1101/460113

17.   Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, Hill C. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat Commun* (2018) **9**:4781. doi:10.1038/s41467-018-07225-7

18.   Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* (2010) **7**:335–336. doi:10.1038/nmeth.f.303

19.   Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* (2016) **13**:581–3. doi:10.1038/nmeth.3869

20.   DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* (2006) **72**:5069–72.

doi:10.1128/AEM.03006-05

21.   Andersen KSS, Kirkegaard RH, Karst SM, Albertsen M. ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. *bioRxiv* (2018)299537. doi:10.1101/299537

22.   Juul S, Izquierdo F, Hurst A, Dai X, Wright A, Kulesha E, Pettett R, Turner DJ. What's in my pot? Real-time species identification on the MinION. *bioRxiv* (2015)30742. doi:10.1101/030742

23.   Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* (2016) **26**:1721–1729. doi:10.1101/gr.210641.116

24.   Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. *Genome Res* (2017)722–736. doi:10.1101/gr.215087.116

25.   Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* (2015) **12**:733–735. doi:10.1038/nmeth.3444

26.   Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* (2016) **44**:W16–W21. doi:10.1093/nar/gkw387

27.   Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* (2018) **34**:3094–3100. doi:10.1093/bioinformatics/bty191

28.   Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and whole rrn operon. *F1000Research* (2018) **7**:1755. doi:10.12688/f1000research.16817.1

29.   Větrovský T, Baldrian P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS One* (2013) **8**:e57923. doi:10.1371/journal.pone.0057923

30.   Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, Knight R. PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* (2011) **27**:1159. doi:10.1093/BIOINFORMATICS/BTR087

31.   Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol* (2019) **37**:124–126. doi:10.1038/s41587-018-0004-z

32.   Edwards R, Vega A, Norman H, Ohaeri MC, Levi K, Dinsdale E, Cinek O, Aziz R, McNair K, Barr J, et al. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *bioRxiv* (2019)527796. doi:10.1101/527796

33.   Honap TP, Sankaranarayanan K, Schnorr SL, Ozga AT, Warinner C, Lewis CM. Biogeographic study of human gut associated crAssphage suggests impacts from industrialization and recent expansion. *bioRxiv* (2018)384677. doi:10.1101/384677

34.   Stachler E, Bibby K. Metagenomic Evaluation of the Highly Abundant Human Gut Bacteriophage CrAssphage for Source Tracking of Human Fecal Pollution. *Environ Sci Technol Lett* (2014) **1**:405–409. doi:10.1021/ez500266s

35.   Stachler E, Kelty C, Sivaganesan M, Li X, Bibby K, Shanks OC. Quantitative

CrAssphage PCR Assays for Human Fecal Pollution Measurement. *Environ Sci Technol* (2017) **51**:9146–9154. doi:10.1021/acs.est.7b02703

36.   Ahmed W, Payyappat S, Cassidy M, Besley C, Power K. Novel crAssphage marker genes ascertain sewage pollution in a recreational lake receiving urban stormwater runoff. *Water Res* (2018) **145**:769–778. doi:10.1016/J.WATRES.2018.08.049

37.   García-Aljaro C, Ballesté E, Muniesa M, Jofre J. Determination of crAssphage in water samples and applicability for tracking human faecal pollution. *Microb Biotechnol* (2017) **10**:1775–1780. doi:10.1111/1751-7915.12841