Chapter 2: Methods

2.1 Chapter 3: The Gut Phage Database

2.1.1 Metagenome assembly

Sequencing reads from 28,060 human gut metagenomes were obtained from the European Nucleotide Archive (Leinonen et al., 2011). Paired-end reads were assembled using SPAdes v3.10.0 (Bankevich et al., 2012) with option '--meta', while single-end reads were assembled with MEGAHIT v1.1.3 (Li et al., 2015) both with default parameters.

2.1.2 Viral sequence prediction

To identify viral sequences among human gut metagenomes, VirFinder v1.1 (Ren et al., 2017) which relies on k-mer signatures to discriminate viral from bacterial contigs, and VirSorter v1.0.5 (Roux et al., 2015) which exploits sequence similarity to known phage and other viral-like features such as GC skew were used. While VirFinder is only able to classify whole contigs, VirSorter can also detect prophages and thus classifies viral sequences as 'free' or integrated. Since obtaining high-quality genomes was paramount for downstream analyses, conservative settings for both tools were used. Only metagenome assembled contigs >10 kb in length were analysed for viral prediction. With VirSorter, only predictions classified as category 1, 2, 4 or 5 were considered. In the case of VirFinder, contigs with a score >0.9 and P <0.01 were selected.

Contigs were further quality-filtered to remove host sequences using a blast-based approach. Briefly, the 'blastn' function of BLAST v2.6.0 (Altschul et al., 1990) was used to query each contig against the human genome GRCh38 using the following parameters: '-word_size 28 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -dust yes -evalue 0.0001 -min_raw_gapped_score 100 -penalty -5 -perc_identity 90 -soft_masking true'. Contigs with positive hits across >60% total length were excluded.

2.1.3 Sequence clustering

Dereplication of the filtered contigs was performed with CD-HIT v4.7 (Li and Godzik, 2006) using a global identity threshold of 99% ('-c 0.99'). This was performed first on contigs obtained within the same ENA study, and afterwards among those obtained across studies. A final set of representative viral sequences was generated by clustering these resulting contigs at a 95% nucleotide identity over a local alignment of 75% of the shortest sequence (options '-c 0.95 -G 0 -aS 0.75').

2.1.4 Quality control of GPD predictions

In order to ensure a high-quality of GPD predictions I removed integrative and conjugative elements by using a machine learning approach.

The training set consisted of all experimental ICEs with intact sequence retrieved from ICEberg 2.0 (Bi et al., 2012) and the phage RefSeq genomes from NCBI (Brister et al., 2015). The test set was downloaded from the Intestinal microbiome mobile elements database (ImmeDB) (Jiang et al., 2019) corresponding to the "ICEs" and "Prophages" datasets. By parsing GFF files with custom Python scripts, for each sequence I calculated 3 high-level features, namely number of genes/kb, number of hypothetical proteins/total genes, and 5-kmer relative frequency (4^5 = 1024 kmers). I used Keras with the TensorFlow (Abadi et al., 2016) backend to train a feedforward neural network with an initial hidden layer of size 10 (ReLU activation), followed by another hidden layer of size 5 (ReLU activation) and a final neuron with a sigmoid activation function. Model selection was carried out with 5-fold cross-validation. I trained the network using the Adam optimizer and the binary cross entropy as the loss function.

I carried out the classification by allowing a false positive rate of 0.25% with a recall of 91%. Finally, I excluded genomes that were predicted to belong to non-phage taxa (82 predictions)

The code for the classifier can be found here:
https://github.com/cai91/GPD

2.1.5 Genome completeness and contamination

Genome completeness and contamination was evaluated by running CheckV v0.5.1 (Nayfach et al., 2020) with the "end_to_end" program.

2.1.6 Viral taxonomic assignment

Viral taxonomic assignment of contigs was performed using a custom database of phylogenetically informative profile HMMs (ViPhOG v1, available here: ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/viral-pipeline/hmmer_databases), where each model is specific to one viral taxon. First, protein-coding sequences of each viral contig were predicted with Prodigal v2.6.3 (Hyatt et al., 2010). Thereafter, 'hmmscan' from HMMER v3.1b2 (Eddy, 1998) was used to query each protein sequence against the ViPhOG database, setting a full-sequence E-value reporting threshold of 10^{-3} and a per-domain independent E-value threshold of 0.01. Resulting hits were analysed to predict the most likely and specific taxon for the whole contig based on the following criteria: (i) a minimum of 20% of genes with hits against the ViPhOG database, or at least two genes if the contig had less than 10 total genes; and (ii) among those with hits against the ViPhOG database, a minimum of 60% assigned to the same viral taxon.

2.1.7 Clustering of phages into VCs

I first created a BLAST database (makeblastdb with options -parse_seqids -dbtype nucl) of all the nucleotide sequences stored in GPD and then carried out all the pairwise comparisons by blasting GPD against itself (I only kept hits with evalue<=0.001). Then, for every pairwise comparison, I calculated the coverage by merging the aligned fraction length of the smaller sequence that shared at least 90% sequence similarity. I kept only the results with a coverage >75%. Finally, I carried out a graph-based clustering by running the Markov Clustering Algorithm (MCL) (Dongen (S.M.), 2000) with an inflation value of 6.0

2.1.8 Bioinformatics tools

The code for the tools developed in this work can be found here:

DotBlast: https://github.com/cai91/dotBlast
HyperVir: https://github.com/cai91/hyperVir
vMatch: https://github.com/cai91/vMatch

2.2. Chapter 4: Function, phylogeny and host assignment of gut phages

2.2.1 Detection of function in gut phages

KEGG pathways, modules, and orthologs were predicted with eggNOG-mapper V2.0.0 (Huerta-Cepas et al., 2017) . Annotation of predictions was carried out using Prokka v. 1.5-135 (Seemann, 2014).

2.2.2 Clustering of proteins into protein clusters (PCs)

I predicted the whole proteome of GPD with Prodigal v2.6.3 (metagenomic mode) (Hyatt et al., 2010) and masked the low-complexity regions with DustMasker. I then created a BLAST (Altschul et al., 1990) database of all the protein sequences and carried out all the pairwise comparisons by blasting the GPD proteome against itself (E-value<=0.001). Then, for every pairwise comparison, I calculated a similarity metric as defined by Chan et al (Chan et al., 2013). Finally, I ran the Markov Clustering Algorithm (MCL) (van Dongen, 2000) with an inflation value of 6.0 and removed clusters with only 1 member.

2.2.3 Phylogenetic analyses

The phylogenetic tree comparing Gubaphage against crAss-like phages was constructed by aligning the corresponding large terminase genes with MAFFT v7.453 (Katoh et al., 2002) –auto mode, followed by FastTree v2.1.10 (Price et al., 2010). The results tree was visualized on iTOL (Letunic and Bork, 2007). For studying the phylogenetic structure of Gubaphage and *Picovirinae*, I calculated the fraction of shared protein clusters among all the Gubaphage genomes and then carried out hierarchical clustering with average linkage and Euclidean metric.

2.2.4 Taxonomic assignment of bacterial genomes

Bacterial isolate genomes were taxonomically classified with the Genome Taxonomy Database Toolkit (GTDB-Tk) v0.3.1 (Chaumeil et al., 2019) (https://github.com/Ecogenomics/GTDBTk) (database release 04-RS89) using the

'classify_wf' function and default parameters. Taxa with an alphabetic suffix represent lineages that are polyphyletic or were subdivided due to taxonomic rank normalization according to the GTDB reference tree. The unsuffixed lineage contains the type strain whereas all other lineages are given alphabetic suffixes, suggesting that their labelling should be revised in due course.

2.2.5 Host assignment

I predicted CRISPR spacer sequences from the 2898 gut bacteria using CrisprCasFinder-2.0.2 (Couvin et al., 2018). I only used spacers found in CRISPR arrays having evidence levels 3 and 4. I assigned a host to a prediction only if the putative host CRISPR spacer matched perfectly to the phage prediction (100% sequence identity across whole length of CRISPR spacer). I carried out the screen by blasting all the predicted CRISPR spacers against the nucleotide GPD BLAST database using the following custom settings (task: blastn-short, -gapopen 10, -gapextend 2, penalty -1, -word_size 7m -perc_identity 100). I retained only hits that matched across the whole length of the spacer with a custom script. In addition, prophages were assigned to the bacterial assembly from which they were predicted. In order to assess the prevalence of false positives due to random chance, I generated 100 sets of CRISRPR random spacers and mapped them against the GPD.

2.2.6 Assessing viral diversity patterns

To compare viral diversity patterns across different gut bacteria, the number of VCs that targeted each bacterial genus was normalized by the total number of isolates from that genus. A VC was considered to target a gut isolate if at least 1 of the genomes from the cluster was predicted to infect it by either CRISPR matching or prophage assignment.

2.2.7 Host range analysis

The number of VCs restricted to target a bacterial taxonomic rank (e.g. species, genus, family) was calculated by predicting all the bacterial hosts associated to each VC and then computing the set for each rank. If the set was a singleton, then the VC was considered to be restricted to that bacterial taxonomic rank.

The gut bacteria isolate tree showing broad host range VCs was constructed by considering all the VCs not restricted to a single genus (cross-family). For each VC, a pair of bacteria assemblies that matched the different genera were picked. The tree was visualized on iTOL.

2.3 Chapter 5: Global distribution and epidemiology of gut phages

2.3.1. Metagenomic read mapping

To estimate the prevalence of each viral species, metagenomic reads were mapped using BWA-MEM v0.7.16a-r1181 ('bwa mem -M') (Li and Durbin, 2009) against the GPD database (clustered at 95% nucleotide identity). Mapped reads were filtered with samtools v1.5 (Li et al., 2009) to remove secondary alignments ('samtools view -F 256') and each viral species was considered present in a sample if the mapped reads covered >75% of the genome length.

2.3.2 Correlation of phages detected and sample sequencing depth

The number of phages detected was calculated by counting the number of GPD genomes that mapped to each of the 28,060 metagenomic samples and then associating it with the corresponding sample sequencing depth. Pearson's r was calculated with the function *stats.personr* from the Python package SciPy v1.3.1

2.3.3. Geographical distribution of metagenomic samples

Similarity between 2 samples was calculated by computing the number of shared VCs divided by the total number of VCs in both samples (Jaccard index). Only deeply sequenced samples (>50 million reads) and healthy samples were considered for the analysis. Distribution of samples was visualized with principal component analysis (PCA) using the *decomposition.PCA* function from scikit-learn v0.22.2. Confidence ellipses encompass 2 standard deviations for each lifestyle samples. PERMANOVA test was carried out with *stats.distance.permanova* function from the Python library scikit-bio v0.5.6

2.3.4 Calculation of phage carriage

Phage carriage was calculated by counting the number of different VCs found in each of the deeply sequenced samples (>50 million reads) for each continent. The Mann Whitney U-test was used to test significance with the *stats.mannwhitneyu* function from the Python package SciPy v. v1.3.1

2.3.5 Detection of enterotypes targeted by VCs

For each analysed region (North America, South America, Europe, Africa, Asia, Fiji and Australia), I predicted all the aggregate bacterial genera targeted by the corresponding genomes that mapped to each region. I then counted the number of genomes that targeted *Bacteroides* genera (*Bacteroides*, *Bacteroides* A, *Bacteroides* B) or the Prevotellaceae family (*Prevotella*, *Paraprevotella*) and normalized by total targeted genera found in each region. Statistical testing was carried out with the *stats.chisquare* function from SciPy v1.3.1.

2.3.6 Network of globally distributed phages

Globally distributed phages were detected by screening VCs for which at least 1 genome of the cluster was found in at least 5 continents. The host-phage network was generated by drawing an edge between each global VC and the predicted bacterial genera they infected. The network was visualized with Cytoscape v3.6.1.

2.3.7 Core virome analyses

In order to evaluate how many VCs covered a specific proportion of samples, I calculated how many samples contained at least 1 VC from a set of VCs. A VC was considered to be found in a sample if at least 1 of the genomes of a VC mapped to the sample. I repeated this procedure with sets sizes ranging from 1 to 500 VCs. Sets grew following the rank of the VCs from biggest to lowest (by number of genomes). When considering the crAss-like family, Gubaphage, and *Picovirinae* clades, I considered them present in a sample if any of the genomes associated to these clades mapped to the sample.

2.4 GPD resource and metadata

GPD genomes and associated metadata can be found here:
http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/gut_phage_database/