Chapter 6: Summary and future work

6.1 Summary

6.1.1 Development of the GPD

In this thesis, I carried out the largest genomic analysis of the human gut phageome by examining more than 142,000 phage genomes derived from 28,060 worldwide distributed human gut metagenomes and 2898 gut bacteria isolates.

In Chapter 3, I introduced the Gut Phage Database (GPD). Although several databases harbouring phage sequences from gut viromes have been published (Gregory et al., 2019; Paez-Espino et al., 2019), to my knowledge, this set represents the largest collection of human gut phage genomes analysed to date. Given the scale of the analyses, not only I was able to identify completely novel viral lineages, but also longer, more complete representatives of known phage genomes. Importantly, this work shows that it is possible to recover high-quality phage genomes from shotgun metagenomes without the need to previously enrich for viral-like particles (VLPs). With this approach, I not only recovered non-integrative phages like *Picovirinae* phages, but also prophage sequences which may rarely enter the lytic cycle and form VLPs. As shotgun metagenomes are far more readily available than VLP metagenomes, I had access to an unparalleled amount of DNA sequences which enabled me to obtain more complete and diverse genomes.

In Chapter 3 I also carried out quality control (QC) and developed methods to handle the massive nature of the dataset. An important finding was the presence of false positives that corresponded to conjugative elements, which highlighted the need for stringent QC when generating thousands of predictions from metagenomic datasets. Even the use of conservative settings of available bioinformatics tools should not preclude the use of extensive QC on phage predictions. As the field moves towards the analysis of larger datasets, manual curation becomes impractical, and I believe that machine learning (ML) approaches (such as the classifier developed here) can be harnessed to help mitigate contamination and significantly boost the quality of the final set of predictions. ML is an extremely fast-paced field and

biologists should take advantage of recent breakthroughs (e.g. deep learning) to make sure that the increasing large volume of biological data submitted to repositories is of high quality (Webb, 2018).

A challenge of this project was the organisation of the large number of predictions into meaningful groups. On one hand, a set of dereplicated predictions at 95% nucleotide identity can be analysed without any further clustering, however patterns can be missed due to underpowering. On the other hand, organising predictions into viral clusters (VCs) allowed me to better generalize my findings. Predictions can be clustered at any defined threshold (e.g. sequence identity), however in order to use a more objective criterion, I benchmarked cluster growth at different thresholds and found that at 90% nucleotide identity most clusters stopped growing (reflecting a more natural threshold). Ideally, clustering by taxonomy proposed by the International Committee on Taxonomy of Viruses (ICTV) should be used (e.g. genus, subfamily), however the majority of my predictions could not be assigned a low level rank or no rank at all. Using a very high-level taxonomy such as order (e.g. *Caudovirales*) also causes to miss patterns because of loss of signal resolution. I expect that as genomic and phenotypical features of the VCs generated are further studied, it's going to be possible to classify them into at least one of the 15 hierarchical ranks recommended by the ICTV (Gorbalenya et al., 2020).

6.1.2 Characterising phage functions and host range

In Chapter 4, I capitalized on the vast number of predictions in GPD to gain knowledge about functions carried out by gut phages. I detected other auxiliary metabolic genes (AMGs) including those involved in nucleotide and sulphur metabolism. Targeted searches also revealed phage reverse transcriptases (RTs) and nutrient transporters.

Mining of function in phages requires a stringent quality control to avoid overestimating their functional potential due to contamination by bacterial genes. Special attention should be paid to genes found at the ends of prophages and contamination assessment should be always carried out. Fortunately, decontamination of phage contigs is becoming automatized with recently published tools such as CheckV (Nayfach et al., 2020) and DRAM-v (Shaffer et al., 2020), facilitating the large-scale annotation of phages from metagenomes. Once a set of clean contigs

are generated, other annotation tools can be used to further characterize the functional potential of phages.

Decontaminated phage contigs still do not guarantee a comprehensive functional annotation as a large fraction of phage proteins are labelled as hypothetical. This limitation highlights our lack of our understanding of protein function which is not exclusive of phages, as recently it was reported that ~27% of proteins derived from gut bacteria do not match any database (Almeida et al., 2020). The number of hypothetical proteins in phages can also be exacerbated by their structural proteins which due to poor conservation are challenging to annotate by conventional methods. However, novel approaches which rely on compositional and physicochemical features such as VIRALpro (Galiez et al., 2016), PVP-SVM (Manavalan et al., 2018), and DeepCapTail (Abid and Zhang, 2018) have showed promise in recognizing them.

The second objective of Chapter 4 was to study relevant gut phage clades. The data-driven discovery of the Gubaphage clade suggests a strategy to identify important clades of phages in metagenomic datasets, as the same approach re-discovered the p-crAssphage as one of the most prevalent clades of human gut phages. Analysis of the *Picovirinae* subfamily illustrated how metagenomics datasets can also help fill-in gaps in viral diversity.

An important element of this work was bacterial host assignment of the majority of gut phages. Both methods used here, exact matches and CRISPR, rely on cultured gut bacteria isolates and highlight the importance of culturing bacteria when studying the viral diversity of ecosystems. The existence of broad host range phages in the human gut suggests that phages have the potential to act as vehicles for horizontal gene transfer (HGT) across distant bacterial clades. The conservative settings used here (100% match and coverage) while highly specific, may have been very stringent and future work could be benefited by allowing a small number of mismatches while maintaining a high specificity.

6.1.3 Epidemiology of gut phages

In Chapter 5, I investigated the epidemiology of gut phages. To my knowledge this is the most comprehensive analysis regarding the global distribution of gut phages given the diversity of the metagenomes (6 continents and 23 countries) and number of phages clades taken into account (21,012 VCs). At a global scale, I provided evidence that the composition of the gut phageome depends on the associated lifestyle of a sample, but also on the gut bacterial composition carried by an individual.

The general dependency of the gut phageome on bacterial composition does not preclude the idea of a global highly prevalent clade of phages (e.g. a VC with a very broad host range). Since its discovery in 2014 (Dutilh et al., 2014), the p-crAssphage has attracted the attention of the microbiome field and even taken as a biomarker of human faecal contamination. After analysing the most prevalent VCs per continent, I discovered that the p-crAssphage was not a highly prevalent clade in Africa and South America. This result provided evidence that p-crAssphage is not a highly prevalent phage in the gut of individuals with a non-Western lifestyle. However, when I analysed the whole crAss-like family, I found some of its members (particularly genera VI,VIII, and IX) in Africa and South America. Host prediction of these phage genera revealed that they prey on *Prevotella copri*. Therefore, it seems that the crAss-like family is a highly prevalent clade of gut phages around the world, raising questions of the biological adaptations that contribute to its success.

This result also highlighted the need to cluster phages into higher taxonomic groups (e.g. genus, subfamily, family) when studying general patterns in the gut phageome. The reason why many studies have not found a core phageome may be because they dereplicate contigs at the species level (e.g. 95% nucleotide identity). This threshold is too stringent; seemingly unrelated phages at the nucleotide sequence level (such as the members of the crAss-like family) may constitute a well-defined clade of phages that share a significant fraction of protein clusters.

When I analysed the concept of a core phageome using VCs, I couldn't find a single VC that was found in more than 50% of samples. However, when I analysed at the phage subfamily level, I found that the *Picovirinae* clade qualified to be a member of the core phageome. Importantly, this clade was found in over 80% of samples from Africa and South America which gut microbiomes are largely unexplored.

6.2 Main findings of this work

1. With proper QC measures, mining of shotgun metagenomes can generate highly complete representative phage genomes complementing VLP enriched metagenomes.
2. A large fraction of gut phages often encode reverse transcriptases (RTs) and auxiliary metabolic genes (AMGs) involved in nucleotide and sulphur metabolism.
3. The Gubaphage clade is a novel gut phage with reminiscent features to crAssphage and is globally distributed.
4. Metagenomics can be harnessed to expand and increase the resolution of previously defined phage subfamilies (*Picovirinae* subfamily).
5. A significant fraction of gut phages (~36%) are not restricted to infect a single species, potentially facilitating gene flow networks between phylogenetically distinct gut bacteria.
6. At a global scale, the gut phageome is associated to lifestyle and influenced by the gut bacterial composition.
7. P-crAssphage is not a highly prevalent phage in Africa and South America, but other members of the crAss-like family that infect *Prevotella copri*.
8. A group of core phages may exist at a global scale (such as the *Picovirinae* subfamily), and may become apparent when dereplicating at higher phage taxonomic ranks.

6.3 Future work

1. *Organizing phage diversity to improve knowledge transfer across metagenomic studies*

   With the current wealth of phage genomes stored in metagenomes, it's now possible to start organizing the large number of phage sequences into meaningful clusters which represent high level candidate viral clades (e.g. subfamilies). This organisation would facilitate the detection of common phage clades across conditions and environments (e.g. is there a phage shared by all body sites?)

2. *Elucidating the extent of active prophages in the human gut*

   An outstanding question is whether prophage sequences integrated in gut bacteria are active or not. Prophages can become "grounded" by mutations in integrases or can accumulate deleterious mutations in essential genes. Conversely, some prophage genes can be useful to bacteria and thus their function is conserved (domestication). Analysis of positive and negative selection on prophage genes from gut bacteria could shed light on this matter.

3. *Mining of phage-encoded antimicrobials*

   Phages represent a rich source of antimicrobials. Given that over 40,000 GPD phage genomes were assigned a host, custom phage encoded antimicrobials such as endolysins can be predicted for hundreds of gut bacteria species. This large-scale resource of anti-bacterial proteins could lead to the development of therapies that specifically modulate the composition of the human gut microbiota.

4. *Investigating diversity of Microviridae/RNA gut phages*

   Due to the minimum genome size imposed in GPD (10 kb), *Microviridae* phages were not investigated in this work. Smaller contigs could be re-analysed and further supported by other tools such as CheckV or an ensemble of predictions tools such as

the What the Phage workflow (Marquet et al., 2020). In the case of RNA gut phages, metatranscriptomics datasets could be harnessed for their discovery.

5. *Wet-lab validation of findings*

This thesis generated a vast amount of predictions that can guide experiments in the laboratory. Since many GPD phages are found in publicly available gut bacteria, further investigation in the wet lab can be carried out on the predicted host range of gut phages and functions conferred by phage-encoded auxiliary metabolic genes.