Chapter 3: The Gut Phage Database

3.1 Introduction and aims

The first metagenomic studies revealed that the majority of the viral gut diversity is novel (81%-93%) (Manrique et al., 2016; Reyes et al., 2010), and since only recently their bacterial hosts started to be cultured (Browne et al., 2016), gut phage host assignment and host range have remained largely uncharacterized. An exception has been crAssphage, a phage discovered in 2014 by computational analysis of metagenomic reads and found in >50% of Western human gut microbiomes (Dutilh et al., 2014). A surprising finding was that the majority of phage sequences uncovered by metagenomics could not be classified into any known viral taxonomy laid out by the International Committee on Taxonomy of Viruses (ICTV) (e.g. species, genus, family), prompting many researchers to organize phage predictions from metagenomic datasets into custom grouping schemes based solely on genomic features (Bin Jang et al., 2019).

More recently, gut metagenomes have been mined in order to compile a more comprehensive list of gut phage genomes (Gregory et al., 2019; Paez-Espino et al., 2019). Nevertheless, the limited number (<700) of metagenomes used to construct these databases, and the median fragment size of their predictions (<15 kb as opposed to ~50 kb for an average *Caudovirales* phage genome), suggests that we have yet to capture a globally representative gut phage diversity and the current phage genomes are likely far from complete. Indeed, a recent report estimated that the IMG/VR database, which contains viral sequences from a wide range of environments including the human gut, showed that only 1.9% of the predictions were complete, and 2.5% high-quality (Nayfach et al., 2020). These issues highlight the need for a comprehensive resource of longer and complete reference phage genomes to enable genome-resolved metagenomics for virome studies.

In this chapter, I describe the construction of the largest database to date that harbours the human gut phage sequences, which were product of mining 28,060 metagenomes and 2898 isolate genomes derived from the human gut microbiota. I investigate ways to organise the huge viral diversity uncovered in this work in order to improve the characterisation of gut

phages in the following chapters. I also developed tools that can aid in the exploratory analysis of viral genomes that will be presented in this chapter.

The aims of the research presented in this chapter are:

- generate the Gut Phage Database (GPD), a high-quality and comprehensive database of the human gut bacteriophage sequences;
- group viral diversity into meaningful clusters to enable more powerful downstream analyses;
- Develop tools for the high-throughput analysis of genome synteny, hypervariation, and phylogeny of viral genomes.

3.2 Results and discussion

3.2.1 Construction of the gut phageome database (GPD)

In order to uncover the diversity of human gut bacteriophages, the biggest datasets of human gut metagenomes (n=28,060) and reference genomes of cultured gut bacteria (n=2,898) were mined. In addition, the metagenomes had a worldwide distribution, as they originated from 28 different countries spanning six major continents (Africa, Asia, Europe, North America, South America and Oceania). To identify viral sequences among human gut metagenomes, over 45 million contigs were assembled and screened with VirFinder (Ren et al., 2017), which relies on *k*-mer signatures to discriminate viral from bacterial contigs, and VirSorter (Roux et al., 2015), which exploits sequence similarity to known phage and other viral-like features such as GC skew. Since obtaining high-quality genomes was paramount for downstream analyses, conservative settings were used for both tools and only predictions that were at least 10 kb long were kept. After removing contamination with a machine learning approach (see below) and dereplicating the final set of filtered sequences at a 95% nucleotide identity threshold (over a 75% aligned fraction), a database of 142,809 gut phage sequences was generated (the gut phage database, hereafter referred to as GPD) (Figure 3.1).

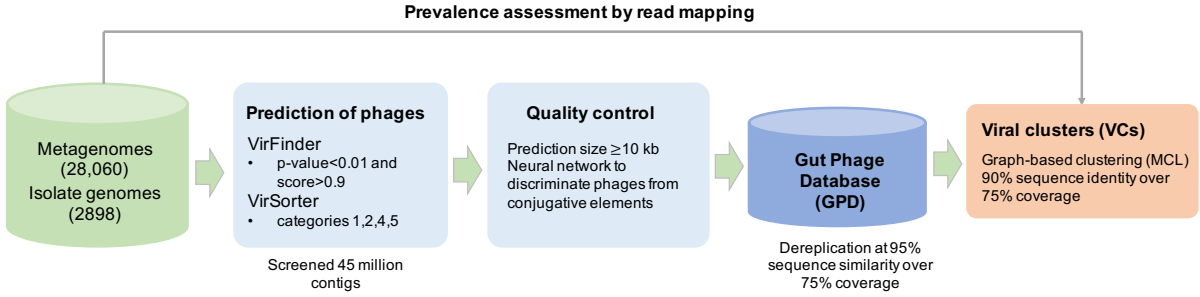


Figure 3.1. Generation of the Gut Phage Database (GPD). An initial dataset composed of 28,060 public human gut metagenomes and 2898 gut bacteria isolate genomes were mined to identify phage genomes. After assembling 45 million contigs, predictions were carried out with VirFinder and VirSorter. Whereas the former is only able to process whole contigs, the latter can also detect integrated viral sequences or prophages. In order to minimize false positives, conservative setting were used for both tools and only fragments > 10 kb were kept. A neural network was trained to remove further contamination caused by ICEs. Predictions were dereplicated at 95% nucleotide identity and they were stored in the gut phage database. In order

to further organize viral diversity, predictions were grouped into viral clusters (VCs). Finally, read mapping was used to quantify prevalence of VCs in the original metagenomes (epidemiology results in Chapter 5).

3.2.2 Decontamination using a machine learning approach

Many false positives (FPs) gene predictions coded for type IV secretion systems and relaxases, suggesting contamination by conjugative mobile elements (Guglielmini et al., 2013). Although plasmids can encode Type IV machinery, I decided to focus on integrative and conjugative elements (ICEs) as conjugation is an inherent feature of their lifestyle (Delavat et al., 2017). In a sense, ICEs behave like temperate "intracellular phages": they integrate into a bacterial genome, can excise from the chromosome and encode a tail-like structural machinery necessary for injecting their DNA into another host. Thus, it's understandable that some of them can be predicted as phages. However, given the widespread use of VirFinder and VirSorter, it came as a surprise that previous reports that used these tools never discussed or raised a warning about potential contamination by conjugative elements. This contamination issue was further exacerbated because many predictions contained truncated ICEs and uncharted diversity, making difficult to discriminate by a marker gene approach.

In order to automate the detection of FPs, I devised a machine learning approach to carry out a further round of decontamination. A feedforward neural network was used to discriminate phages from ICEs. Gene density (genes/kb), kmer signature (pentanucleotide composition), and fraction of hypothetical proteins (hypothetical genes/total genes) were selected as machine learning features, since these metrics can be computed for incomplete sequences and do not rely on direct specific homology (Figure 3.2A and 3.2B). In general, phages had higher densities of genes and hypothetical proteins. The former could be attributed to a selective pressure of phages of fitting their genome into the capsid, while the latter could be explained by poor annotation of phage structural proteins due to their lack of conservation (Seguritan et al., 2012). The extent of discrimination of phages from ICEs by computing these two metrics can be appreciated in Figure 3.2C where they clearly segregate (phages in blue and ICEs in red). The classifier was trained with validated experimental sequences of phages (RefSeq, n=2,387) and ICEs (ICEberg 2.0, n=113). Model selection was carried out with 5-fold cross-validation and the classifier showed an excellent performance in an independent test set

(AUC>0.97) harbouring human gut mobile genetic elements (MGEs) (Figure 3.2D). I carried out the classification by allowing a false positive rate of 0.25% with a recall of 91%.

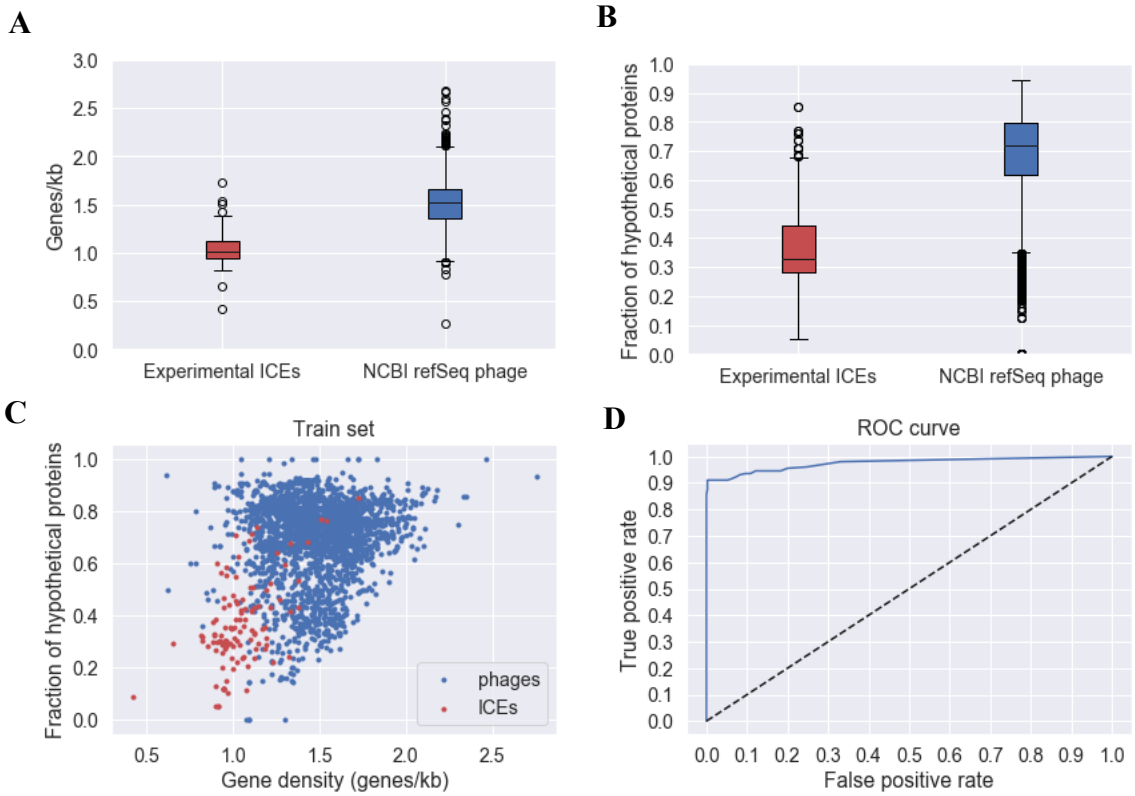


Figure 3.2 – A machine learning approach to distinguish phages from ICEs. In order to discriminate ICEs from phages I relied on three features: kmer signature, gene density, and fraction of hypothetical proteins. Kmer signature has already been exploited as a way to discriminate phages from host DNA. Generally, gene density **A)** and fraction of hypothetical proteins **B)** were lower for ICEs than for phages. **C)** When experimental sequences of ICEs (in red, n =113) and genomes of NCBI phages (in blue, n=2,387) are described by these two features, they clearly segregate. I trained a feed forward neural network that harnessed the 3 features described using experimental sequences from ICEs and phages and benchmarked it with a dataset of gut phages (n=201) and ICEs (n=405). **D)** The classifier had an excellent performance in an independent dataset with an AUC>0.97.

3.2.3 GPD significantly expands gut bacteriophage diversity

In order to assess the viral diversity of the GPD at high taxonomic levels, I used a graph-based clustering approach to group genetically related phages. Merging GPD with RefSeq and two other human gut phage databases (GVD and IMG/VR) (Gregory et al., 2019; Paez-Espino et al., 2019), resulted in the generation of 21,012 non-singleton viral clusters (VCs) with at least 1 GPD prediction (GPD VCs). A VC corresponds to a viral population sharing approximately 90% sequence identity over ~75% aligned fraction.

Comparison of GPD against RefSeq phage genomes, revealed only 171 out 21,012 VCs overlaps. Phages from these 171 VCs mainly infect *Escherichia*, *Enterobacter*, *Staphylococcus*, and *Klebsiella* genera, reflecting the bias of the RefSeq database to harbour phages from well-known clinically important and traditionally culturable bacteria. Consistent with previous reports of phage predictions from metagenomic datasets (Hoyles et al., 2014), I was not able to confidently assign a family to the majority (~80%) of GPD VCs, while the rest corresponded mainly to the *Podoviridae*, *Siphoviridae* and *Myoviridae* families (Figure 3.3A). These 3 viral families belong to the *Caudovirales* order (phages characterized by having tails and icosahedral capsids) which from microscopic studies have been found to be enriched in human faeces (Hoyles et al., 2014; Roux et al., 2012).

For comparison purposes, in addition to GPD VCs, I also considered VCs without GPD predictions (Figure 3.3B). Analysis of VCs composed from only GPD and IMG/VR genomes showed 3,699 overlaps, while I found 3,206 VCs composed of only GPD and GVD genomes. Moreover, GPD harboured the highest number of unique VCs with 12,731 novel clusters. On the other hand, 1099 VCs, and 113 VCs were unique to IMG/VR and GVD, respectively. In addition, 1205 VCs were shared by the three databases. Interestingly, the number of VCs with an assigned phage taxon was lower in the VCs that were unique to GPD as opposed to those shared with GVD and IMG/VR (18.74% vs 27.8%) (P = 1.96e-9, χ^2). Thus, GPD considerably increases the known gut phage diversity in the human gut. This phage diversity expansion is likely driven by the high number of gut metagenomes mined and their global distribution which allows the retrieval of rarer gut phage clades.

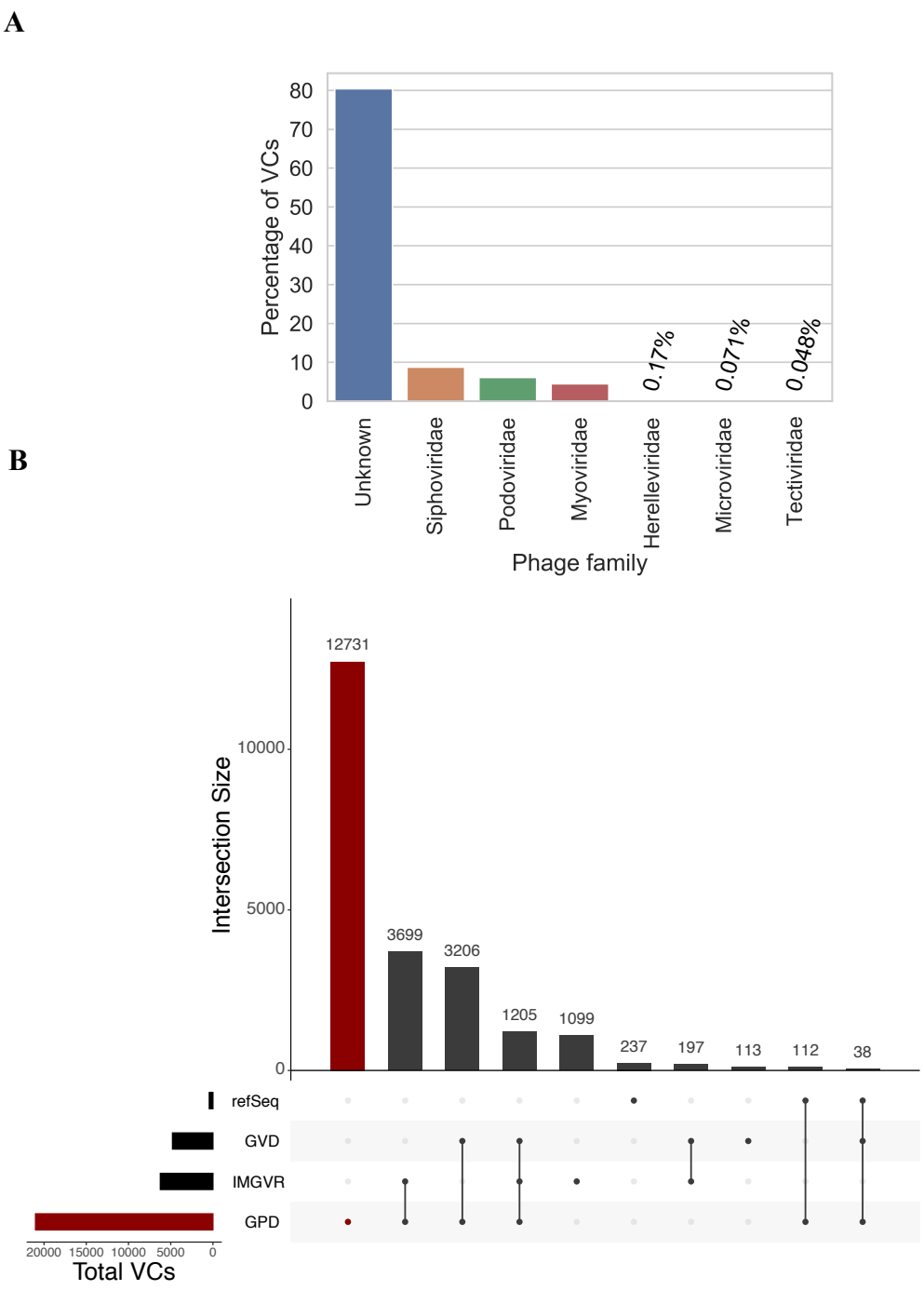


Figure 3.3. GPD taxonomy assignment and comparison against other gut phage databases. **A)** Most of GPD VCs (~80%) could not be assigned to a phage family. The assigned fraction corresponded to mainly families of the *Caudovirales*. **B)** UpSet plot comparing GPD against other public gut phage databases. GPD captures the greatest unique diversity of phage genomes that inhabit the human gut.

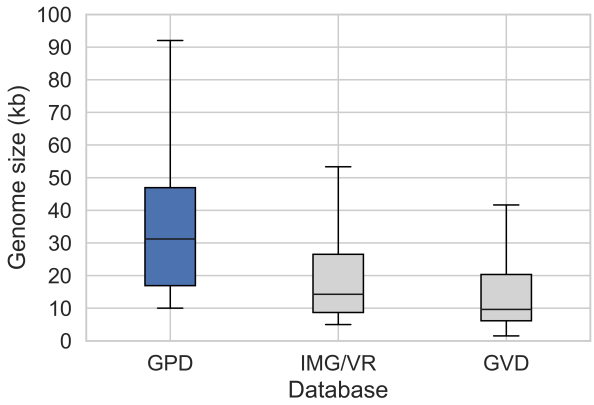
3.2.4 Genome completeness of GPD

Genome completeness is another important feature of a high-quality reference genome database. Unlike prokaryotic genomes, there is no current consensus tool to assess phage completeness and contamination, thus multiple complementary approaches were explored to assess the GPD genome completeness. First, I assessed genome size. The *Caudovirales* order, which is considered a dominant group of the human gut phageome, possesses an average genome size of ~50 kb (Ackermann, 1998). Based on this criteria, GPD harbours the most complete gut phage genomes as it has the largest median genome size with ~31 kb, followed by IMG/VR and GVD with 15 and 11 respectively (Figure 3.4A).

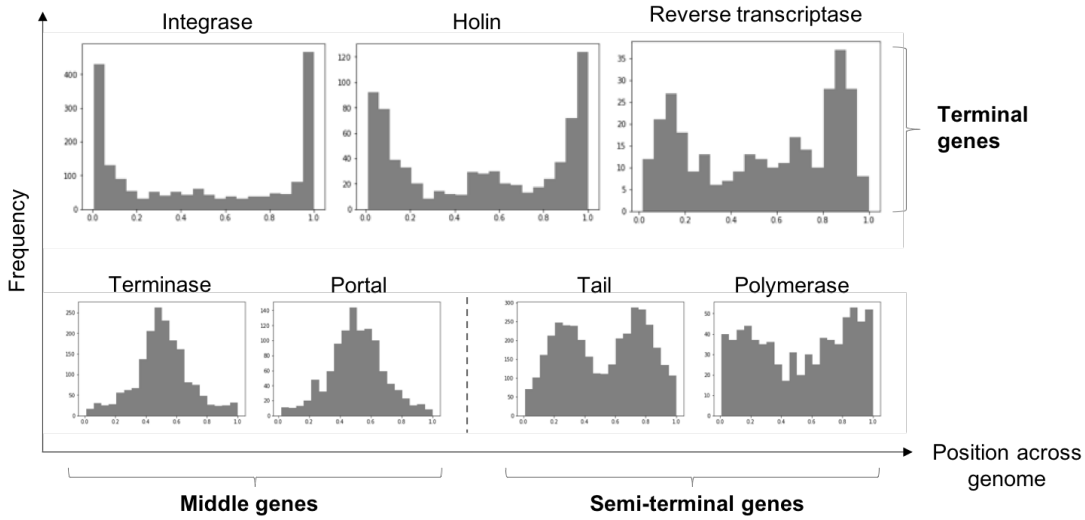
I further assessed completeness by studying the genome organisation of the GPD phage. Figure 3.4B shows the consensus position of marker genes along GPD genomes. I found that key marker genes localized at their expected positions within the predictions. For instance, integrases were more often found at the edges (terminal genes), terminases in the middle, and polymerases in between (semi-terminal genes). This observation reflects the highly complete nature of the GPD genomes. Moreover, this result highlighted the large number of linear genomes which can be a result of prophages or an inherent feature of a phage clade (e.g. *Caudovirales*)

Finally, I estimated the level of completeness of each viral genome using CheckV (Nayfach et al., 2020) (Figure 3.4C). This tool estimates the expected genome length of a viral prediction based on the average amino acid identity to a database of complete viral genomes from NCBI and environmental samples. In total, 41,248 (29%) of the viral genomes were classified as high quality (of which 13,249 were predicted to represent complete genomes), 38,574 (27.01%) as medium quality, 53,116 (37.19%) as low quality, and 9,691 (6.78%) as non-determined. The median genome completeness of all genomes stored in the GPD was estimated to be 63.5% (interquartile range, IQR= 34.68%–95.31%) (Figure 3.4D). Estimation of non-viral DNA by checkV showed that 73.5% of GPD predictions had no contamination whereas 84.13% had a predicted contamination <10%.

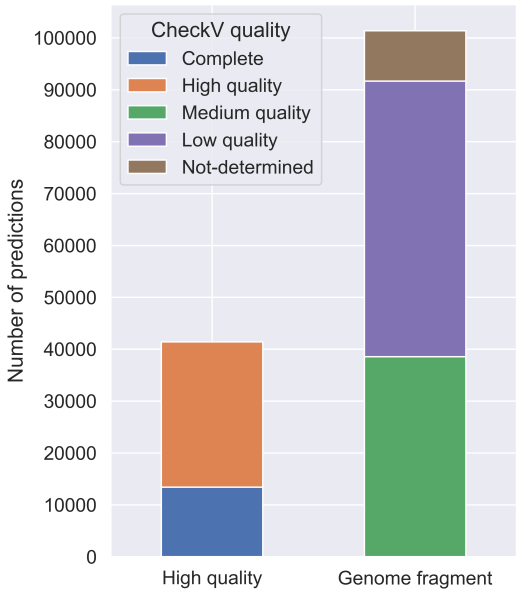
A


Genome size (kb)
100
90
80
70
60
50
40
30
20
10
0
GPD
IMG/VR
GVD
Database

B


Integrase
Holin
Reverse transcriptase
Terminal genes
Frequency
Terminase
Portal
Tail
Polymerase
Position across genome
Middle genes
Semi-terminal genes

C


CheckV quality
Complete
High quality
Medium quality
Low quality
Not-determined
Number of predictions
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
High quality
Genome fragment

D

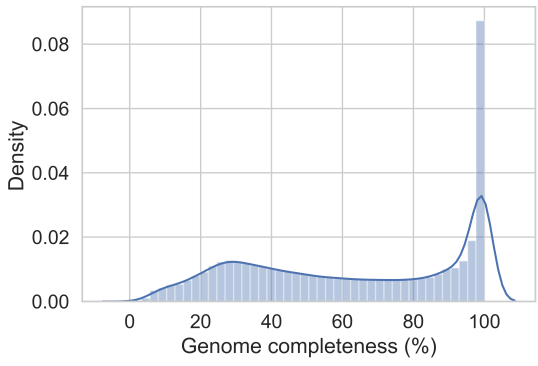

Density
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
Genome completeness (%)

Figure 3.4. Genome completeness of GPD. **A)** Compared to other public databases, GPD harbours the longest genomes with a median of 31 kb as opposed to 14 kb from IMG/VR and 11 kb from GVD. **B)** Distribution of phage marker genes across GPD predictions. Three main types of consensus distributions were observed, namely terminal, semi-terminal, and middle genes. **C)** Genome completeness as judged by CheckV. Over 40,000 genomes were categorized as high-quality (28%) (genome completeness > 90%), while the rest were predicted to be genome fragments. **D)** The median genome completeness of the whole database was was estimated to be 63.5%.

3.2.5 Clustering of phages into VCs

As explained above, I further organized the viral diversity contained in GPD into VCs. Even though a 95% nucleotide identity threshold has been proposed to delineate species in bacterial viruses (Adriaenssens and Brister, 2017), when I examined the final set of predictions (142,809), I realised that many phage genomes were still very similar between each other. Different predictions had extensive synteny with nucleotide identity < 95% and thus shared the majority of genes.

I then decided to explore further clustering by computing how many genomes were related to a "bait" genome at different thresholds of Mash distance (Figure 3.5A). Most of the genomes related to the bait were already saturating at a Mash distance of 10 (~90% nucleotide identity), which I considered as a more appropriate clustering threshold than a Mash distance of 5 (~95% nucleotide identity) (Figure 3.5B).

Since Mash doesn't take into consideration alignment fraction, I switched to BLAST to enforce a minimum alignment fraction of 75% of the shortest sequence and allowed a minimum of 90% nucleotide identity between genomes. In order to automatize the generation of clusters, I relied on an unsupervised approach, namely the Markov Clustering Algorithm or MCL (Dongen, 2000) (see Methods). In short, MCL uses random walks to automatically identify highly connected nodes (phage genomes in this case). After MCL clustering, GPD diversity ended up encapsulated in 21,012 non-singleton VCs. Benchmarking against the RefSeq phages revealed that GPD VCs were equivalent to a subgenus level, as >99% of all VCs were contained within a genus and in some cases, multiple VCs were associated to a single genus (Figure 3.5C).

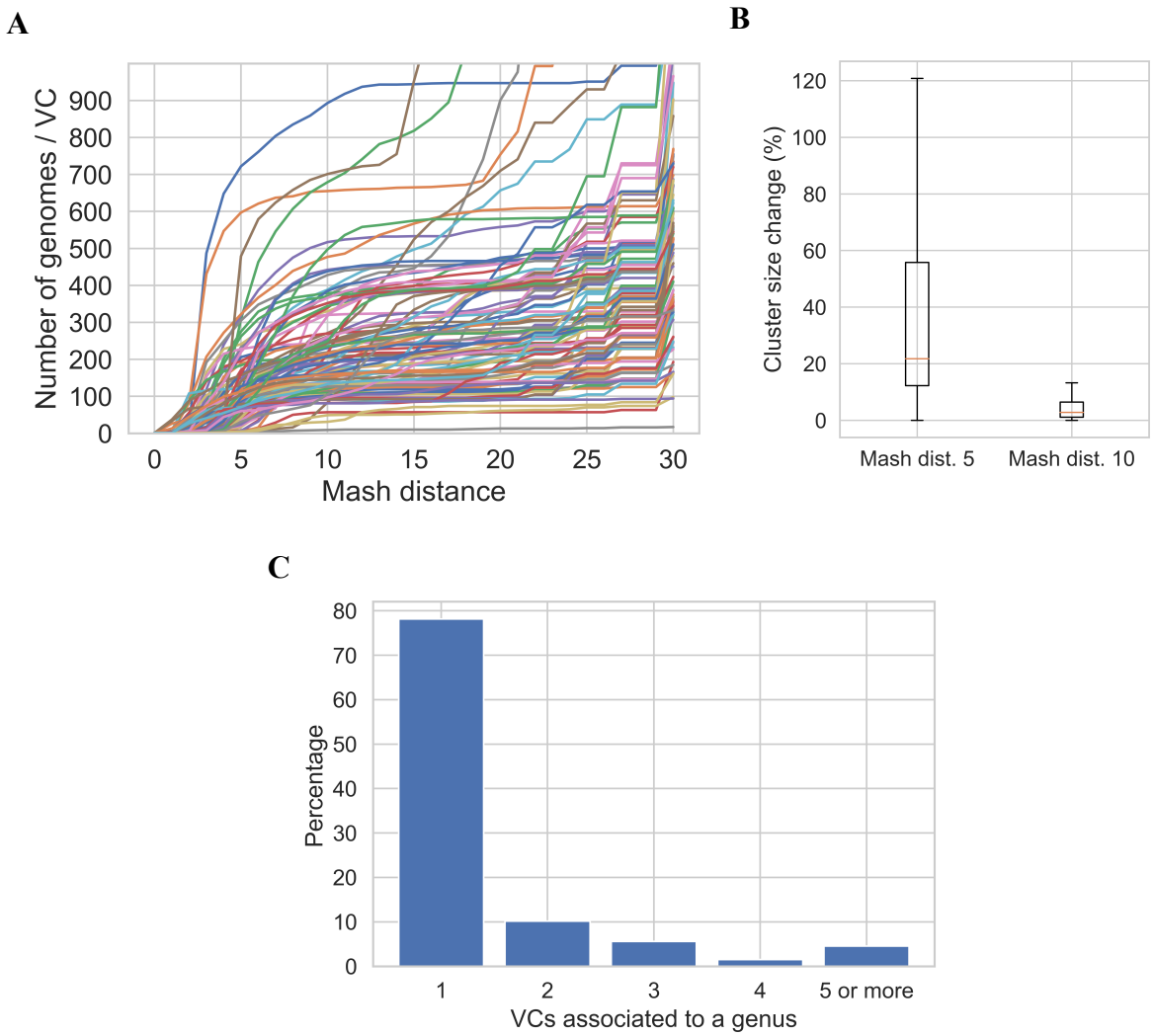


Figure 3.5. Clustering of phages into VCs. A) Even though 95% sequence similarity delineates species level in phages, I noticed extensive synteny between GPD predictions at that threshold. I explored other sequence identity thresholds by computing how many GPD genomes were related to a bait genome. **B)** Viral clusters started to saturate at a Mash distance of 10 (~90% sequence similarity), rather than 5 (~95% sequence similarity). **C)** Benchmarking against RefSeq phages showed that a single phage genus could be associated to several VCs, suggesting subgenus clustering.

3.2.6 Viral clusters reconstruct the phylogenetic structure of gut phages

The resultant VCs were not of uniform size but instead followed a negative exponential distribution with a few clusters (<50) composed of a large number of phage (>100 predictions) followed by a rapidly decreasing long tail of VCs with smaller membership size (Figure 3.6A).

This result suggested that genetic diversity is not evenly distributed in GPD. The number of genomes per VC could reflect inherent genetic diversity of a phage clade, however the most likely explanation here may be sampling bias (oversampled VCs will capture more genetic variation). The top VC was identified as the highly prevalent crAssphage (p-crAssphage), while the second contained a clade of phages characterized by a relatively long genome (~80kb), a BACON domain-containing protein, and *Bacteroidales* host range (hereafter referred to as the Gubaphage clade). The Gubaphage clade is a novel clade of gut phages proposed in this thesis and it is further characterized in Chapter 4. The phylogenetic structure of GPD could be visualized based on a network analysis of VCs (Figure 3.6B). Several VCs were highly inter-connected, forming super clusters and hinting to higher taxonomic clustering (e.g. viral subfamilies). On the other hand, isolated VCs may correspond to very genetically homogeneous viral clades.

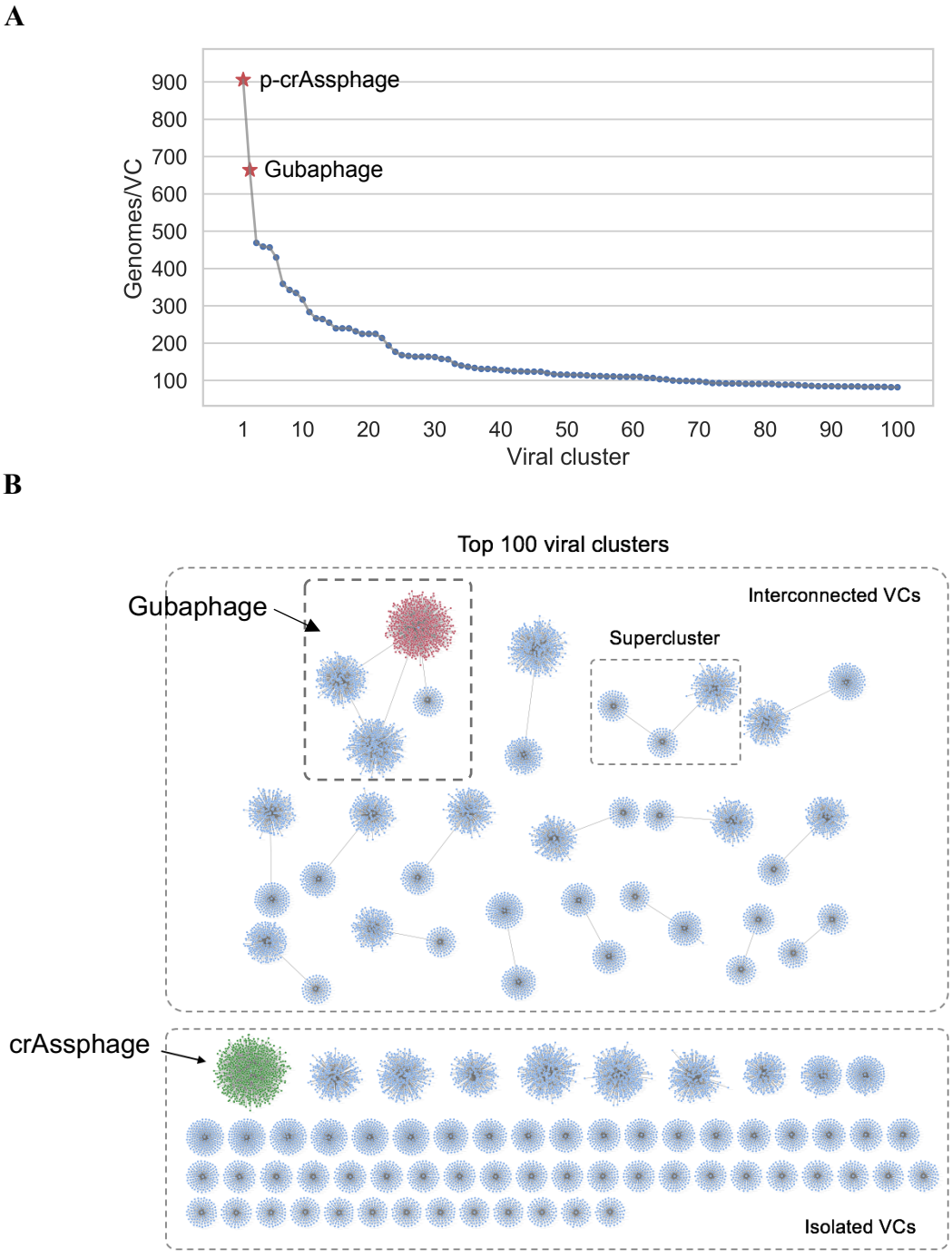


Figure 3.6. Distribution of genomes per VC and phylogenetic structure of GPD A) Distribution of genomes per VC. Only the 100 most prevalent VCs are shown. A member of the crAssphage family (p-crAssphage) was identified as the VC with the bigger cluster size, followed by a VC referred to as the Gubaphage. **B)** Visualization of the top 100 VCs reveal a subset of connected clusters and isolated ones. Inter-connection of VCs likely reflect higher phylogenetic structures such as subfamilies.

3.2.7 Bioinformatics tools

During the course of this work, I developed 3 bioinformatics tools that helped with the exploratory data analysis of GPD genomes, namely dotBlast (synteny analysis), hyperVir (visualization of hypervariable regions), and vMatch (classification of phage sequences). The development of these tools was motivated by the lack of ad-hoc bioinformatics tools to manage the sheer amount of genomes in GPD.

3.2.8 Synteny analysis for viral genomes (dotBlast)

During the exploratory analysis stage of this work I realised that I needed a high-throughput way to compare viral genomes. Sequence identity is a way forward, and adding coverage thresholds can lead to more robust strategies to assess similarity between two genomes. Nonetheless, the source of these two metrics (sequence identity and coverage) is the sequence alignment, and its inspection can help uncover more subtle differences such as insertions, deletions, and inversions.

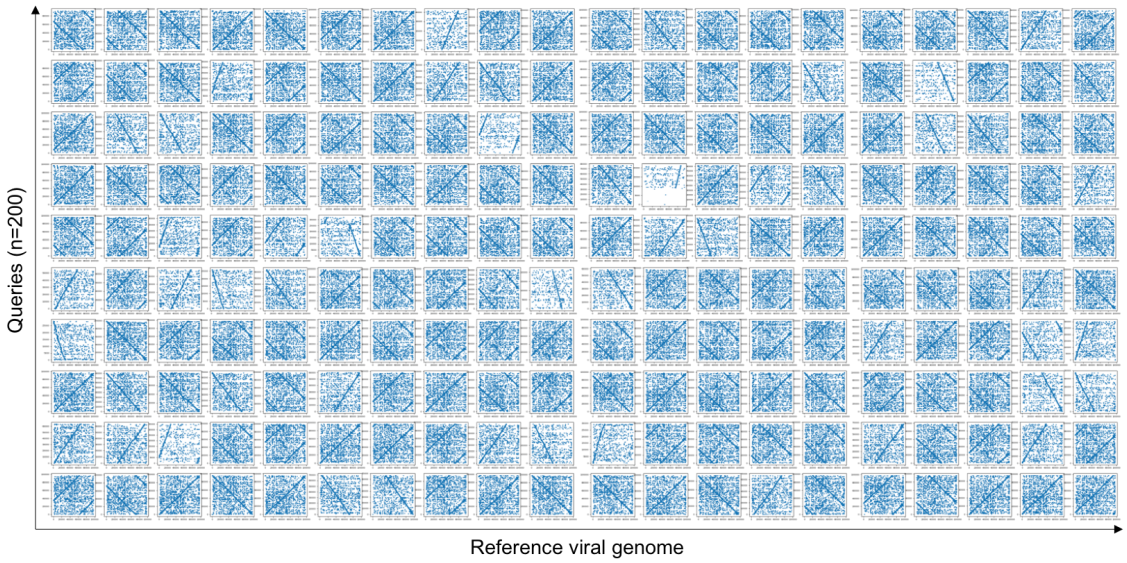
In bioinformatics, a dot plot (also known as a similarity matrix) is one way to efficiently visualize a pairwise sequence alignment. The dot plot was introduced in 1970 by Gibbs and McIntrye and it can be constructed by placing the bases of the first sequence as columns of a matrix, while the second sequence runs perpendicularly and thus fills up the rows of the matrix. Then we simply shade a cell in black if the residues in the corresponding column and row are identical. A consequence of this pattern is that matching subsequences appear as diagonal lines across the matrix.

If “n” and “m” are the lengths of the two sequences to analyse, then the number comparisons is n*m. However, generating the matrix this way is computationally inefficient (quadratic time complexity) and leads to a lot of noise. If a tool is meant to generate hundreds of dot plots in a reasonable amount of time, then this naïve strategy is not practical. A way around is simply to shade cells if they belong to a significant alignment. Fortunately, BLAST can readily process hundreds of queries in an efficient manner.

By incorporating the BLAST output of two aligned sequences, I developed dotBlast which given a blast reference viral genome and a set of queries, can quickly generate the coordinates for the generation of dot plots that compare each query to the reference (Figure 3.7A). In addition, in order to explore more conserved regions, the user can control the alignment significance threshold (Figure 3.7B). By generating dot plots, it's possible to have a quick glance of synteny across hundreds of queries against a reference (e.g. a member of a known viral subfamily). Analysis of dotplots can provide subtle details of genomic organisation e.g. a "broken" main diagonal may indicate circular genomes, a "jump" in the alignment can hint to an insertion or deletion.

With the increasingly large number of viral genomes mined from metagenomes, it is becoming more necessary to have high-throughput tools to easily visualize relationships between phage. DotBlast depends only on BLAST and Python, which are usually already available in a large number of bioinformatics systems or can be easily installed.

A



B

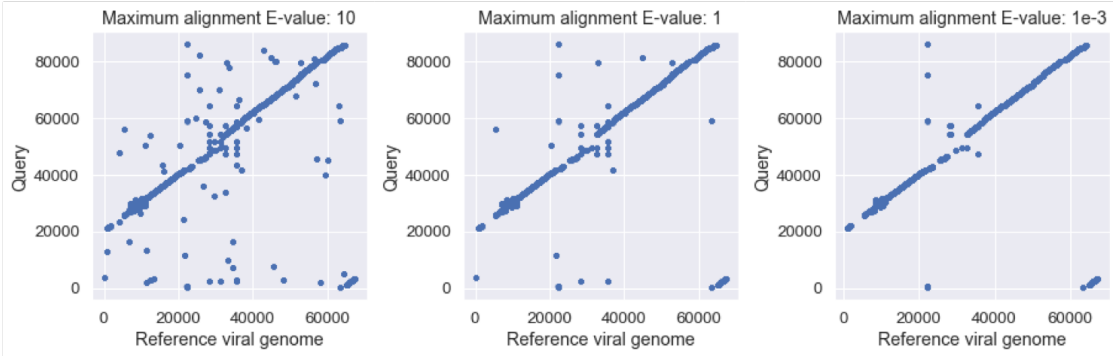


Figure 3.7. DotBlast tool. A) DotBlast can compare hundreds of viral genomes against a reference (e.g. a member of a viral subfamily) by generating dot plots. It uses BLAST to calculate significant alignments and plots them in a dot plot format in a fast manner. B) The significance of alignments can be controlled, allowing to identify highly conserved regions (or decrease noise).

3.2.9 Hypervariation analysis (hyperVir)

Having a large genetic diversity encapsulated in a clade of closely related viral genomes (e.g. species or genus) enables a large number of analyses. The discovery of hypervariation within proteins is particularly interesting because it can lead to the identification of genes with binding domains. These genes can be involved in recognition of bacterial receptors, binding of mucus, and even depolymerization of surface decorating polysaccharides by lytic phage enzymes. Analysis of gut viromes has suggested the existence of multiple hypervariable loci in gut phages (Minot et al., 2012), and thus the assessment of hypervariation in GPD phages can prove to be useful for their characterization. In order to facilitate hypervariation analysis in viral genomes I developed hyperVir which allows visualization of amino acid diversity and automatic detection of hypervariable regions in viral contigs.

The basic workflow (Figure 3.8A) involves an input FASTA file containing protein sequences, followed by a multiple sequence alignment with MAFFT, and finally the estimation of amino acid diversity at each position of the alignment by calculating Shannon's entropy. The signal is smoothed out by passing the Savitzky-Golay filter and hypervariable regions can be detected by a spike of amino acid diversity (Figure 3.8B).

HyperVir is thus a tool that conveniently can uncover viral genes with hypervariable domains which can help narrow down gene function. A more rigorous method involves the detection of positive selection with the Ka/Ks ratio. However, HyperVir is geared towards the detection of highly variable regions (hypervariation), speed, and high throughput visualization of results (Figure 3.8C).

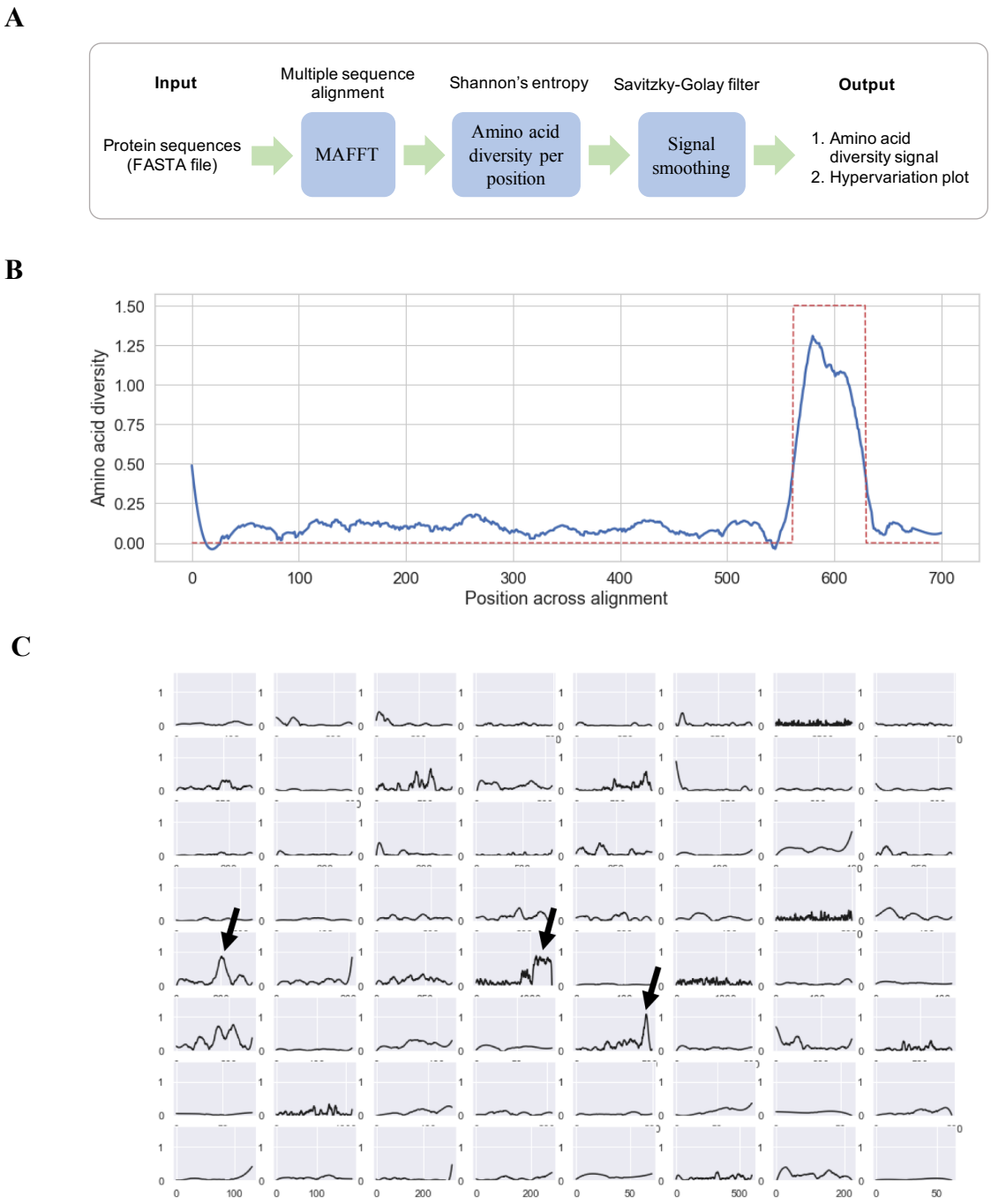


Figure 3.8. HyperVir tool. A) Pipeline to identify hypervariable genes. The input is a FASTA file containing a set proteins. After generating a multiple sequence alignment of the proteins, hyperVir calculates the amino acid diversity at each amino acid position by computing Shannon's entropy. Finally, the signal is smoothed with the Savitzky-Golay filter and the amino acid diversity plots visualized. **B)** Output of hyperVir. Amino acid variation is showed per position of the multiple sequence alignment. An hypervariable region is highlighted in red. **C)** hyperVir applied to 64 sets of proteins shows different hypervariation patterns. Pointed by arrows are examples of proteins with high hypervariation domains.

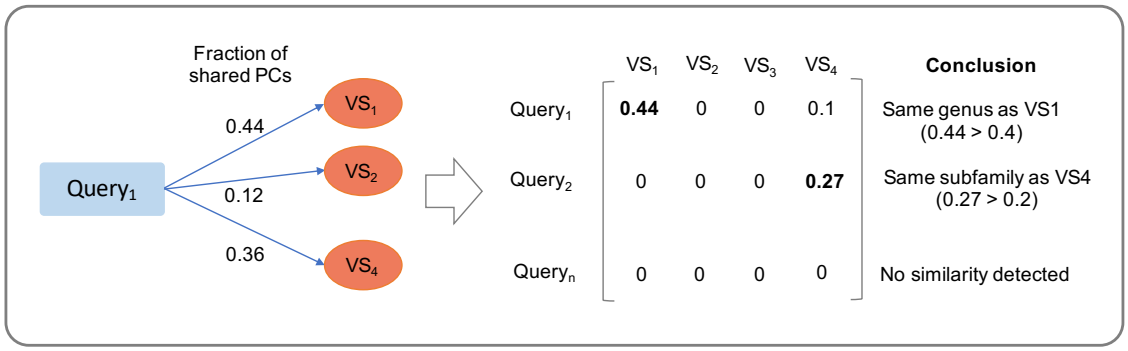
3.2.10 Exploring viral taxonomy through shared protein clusters (vMatch)

Large-scale classification of phage predictions is a recurrent challenge in metagenomic projects. Unlike bacteria, viruses lack a common marker gene and thus it's difficult to reliably estimate the phylogenetic distance between clades. This issue is compounded because phages often recombine and become mosaic, further blurring genetic distances between them. Finally, metagenomic projects often generate viral fragments which decrease the performance of methods that exploit specific-clade marker genes. The idea of using shared homologous proteins as a criterion to demarcate phage clades looked particularly promising e.g. the Phage Proteomic Tree (Rohwer and Edwards, 2002). In recent years, several tools were developed to harness the use of protein clusters to carry out phage taxonomy assignment. However, the majority of these methods were not implemented in packages, limiting their widespread use. A notable exception was the VICTOR tool, which was accessible online but had scalability issues (limit to 100 genomes) (Meier-Kolthoff and Göker, 2017). More recently, vContact2.0 combined a network approach with the idea of sharing protein clusters, and optimized it for the classification of viral predictions at the genus-level. Furthermore, vContact2.0 is also available as a standalone version, making it more accessible for custom datasets (Bin Jang et al., 2019).

Unfortunately, vContact2.0 is not scalable for huge datasets like GPD as the program could not finish processing the sheer volume of predictions (>140,000) submitted. Submission of shorter queries also failed to return taxonomy classification, but only the genus-like clusters. In addition, although useful, the genus scope of the program is a conservative taxonomy assignment. I believe that predictions can be more meaningfully placed into candidate viral subfamilies. This is particularly useful in metagenomes with huge novel viral diversity, as subfamilies can potentially bring together a multitude of novel genera that otherwise would be disconnected from known viral clades and deemed as "dark matter" of the dataset. Importantly, downstream analyses can be negatively affected, as hypothesis testing of associations of specific clades with another variable of interest (e.g. geographical distribution or disease) can end up underpowered. While the criteria for the inclusion of a phage into a specific viral subfamily varies, a sharing of at least 20% of homologous proteins between two genomes has been used to bioinformatically define viral subfamilies (Lavigne et al., 2008, 2009). This was the case of the crAss-like clade, in which the authors segregated all the crass-like sequences into viral subfamilies (20-40% sharing) and genera (>40% sharing) (Guerin et al., 2018).

With this in mind, my objective was to generate a tool for easy taxonomic exploratory data analysis of metagenomic datasets. I developed a standalone program (vMatch) for putative taxonomic assignment of metagenomic viral predictions against reference viral sequences (e.g. RefSeq) based on the principle of shared PCs to demarcate clades. vMatch takes in a file containing clusters of homologous proteins derived from pooling the proteome of the queries (e.g. metagenomic predictions) and reference viral sequences and then calculates the fraction of shared PCs between them. It then stores the results in a matrix in which the rows correspond to the queries and columns to the reference sequences (Figure 3.9A). Each entry corresponds to the pairwise mean of the shared PCs between the query and a reference. The matrix can then be visualized with a clustered heatmap. For instance, members of reference phage clades (*Skunavirus*, *Peduovirus*, *Pahexavirus*, *Teseptimavirus*) are columns of the heatmap, while rows are queries (Figure 3.9B). Clustering of the rows reveals a putative membership of the queries (e.g. metagenomic predictions). If the queries are also used as reference viral sequences, then visualization of the matrix enables the identification of novel clades (red boxes, Figure 3.9B).

A



B

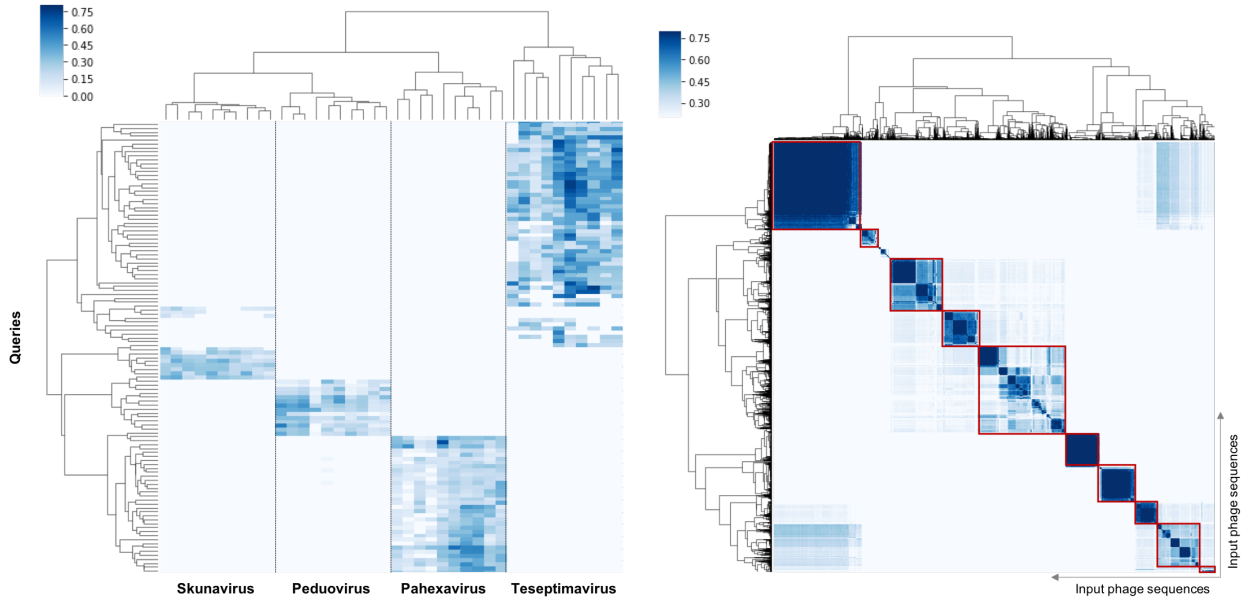


Figure 3.9. vMatch tool. A) Given a query and a set of viral sequences, vMatch calculates the fraction of shared protein clusters (PCs) between them as a proxy of their relationship. For instance, if two viral sequences share >20% of PCs, then they may belong to same candidate subfamily. **B)**. Visualization of vMatch results with clustered heatmap. On the left, a set of queries is compared against reference sequences, rows cluster according to their membership. On the right, the queries are also provided as reference sequences. The heatmap allows the easy identification of clades within the input sequences.

3.3 Conclusions

In this chapter, I presented the framework and rationale for the downstream analyses of human gut phages. By processing viral predictions from 28,060 gut metagenomes and 2898 bacterial isolate genomes, I generated a comprehensive and high-quality database of bacteriophage genomes, namely the gut phageome database (GPD). I showed that two popular tools for viral predictions (VirFinder and VirSorter) even with conservative settings, often predict integrative and conjugative elements (ICEs) as phages. I discovered that phages and ICEs significantly differ in gene density, fraction of hypothetical proteins, and kmer profile and thus these features can be exploited to segregate them. I trained a neural network to learn these differences and deployed it across thousands of predictions to minimize the number of false positives in GPD.

As reported in recent studies that analysed viromes from other environments, I uncovered an enormous amount of novel viral diversity in the human gut, which was particularly prominent when GPD is compared to the gold standard set of known viral genomes (RefSeq phages). This comparison highlighted three main things, namely the outstanding diversity of phages, the limited number of currently available high-quality phage genomes, and how mining of metagenomes can be harnessed to counter the lack of genomic data for phages. Comparing to other public phage databases, GPD outperformed in diversity and genome completeness by a wide margin. These improvements were due to the large number of metagenomes mined, and the diversity of samples which spanned all the 6 continents.

Even though viral predictions were non-redundant at 95% nucleotide identity (which roughly correspond to species level) (Adriaenssens and Brister, 2017), I noticed that at this threshold many predictions still had extensive synteny and nucleotide identity (>90%) to other predictions. For this reason, I decided to further group them into viral clusters (VCs) which consisted of more discrete viral populations. A recent study proposed to formalize the use of species-rank virus groups (Roux et al., 2019). This study found a cluster of genome pairs (suggestive of a species rank) that encompassed a large fraction of phage genomes with a nucleotide identity >90%, providing further support to a departure of the minimum 95% threshold. The generation of VCs is a powerful concept, because it enables to encapsulate highly related viruses into homogenous phage clades and allows to obtain better consensus of their inherent features such as their core and accessory genomes or average genome length. This becomes more evident in the next couple of chapters when I profile the biological

functions and epidemiology of gut phages. In addition, the quality of VCs defined in this work are benefited by the significantly longer genomes hosted by GPD (median>31kb), and provide more sensitivity to find distinctive features of a phage clade.

A critical step in this work was the exploratory data analysis. Unfortunately, none of the existing bioinformatic tools were suitable to handle the large number of GPD genomes. Thus, I decided to create standalone versions of programs that were useful during the development of this work. In addition, due to the large-scale nature of my dataset, processing speed was a priority and therefore all the tools are suitable for high-throughput analyses. The 3 programs developed here are suitable for the assessment of relatedness of viral genomes (dotBlast), study of hypervariation (hyperVir), and exploration of phage phylogeny by overlap of PCs (vMatch).