In this review we provide an overview of current challenges and advances in bacteriophage research within the growing field of viromics. In particular, we discuss, from a human virome study perspective, the current and emerging technologies available, their limitations in terms of de novo discoveries, and possible solutions to overcome present experimental and computational biases associated with low abundance of viral DNA or RNA. We summarize recent breakthroughs in metagenomics assembling tools and single-cell analysis, which have the potential to increase our understanding of phage biology, diversity, and interactions with both the microbial community and the human body. We expect that these recent and future advances in the field of viromics will have a strong impact on how we develop phage-based therapeutic approaches.

Introduction
The human body is an ecosystem, a home to a complex network of microbial organisms comprised of bacteria, archaea, eukarya, and viruses. The majority (sometimes >90%) of viruses present in the human gut are those that infect bacterial hosts; these viruses are known as phages [1–3]. Phages can replicate through two major replication cycles: lysogenic or lytic. Other replication cycles, including pseudolysogeny and chronic infection, also exist. In the lytic cycle, phages infect the host and kill it shortly afterwards. The lysogenic cycle involves phages that stay dormant as part of the host genome; when inserted into the host genome the phage is known as a prophage [1,2]. The dynamics of bacteria-phages interactions varies between ecosystems [4,5], with phages of the human gut persisting for prolonged periods of time and seemingly promoting a stable, healthy gut microbiome [6–8]. Due to the rise in multidrug-resistant bacterial infections, there has been a renewed interest in phage-based therapies as an alternative antibacterial approach. However, despite the therapeutic use of phages being over a century old [9,10], and their high abundance in the body, they are among the least described components of the human microbiota, especially when compared with bacteria [11–13].

Of the estimated 10^{31} viruses on earth, the majority are phages, but only 2640 of their genomes are closed or fully sequenced (https://www.ncbi.nlm.nih.gov/refseq/), an example of how little we know about phage diversity (Box 1). The double-stranded DNA (dsDNA) tailed phages account for the majority of those characterized by electron microscopy and culture-based methods [14,15]. However, recent studies contradict the earlier ones and suggest that this dominance may be biased by the applied methods, rather than being a true representation of the human phageome, and it shows our limited understanding of phage diversity [14–17]. For example, a recent study identified more than 15 000 ssRNA phage sequences from public datasets, including over 1000 near-complete genomes, by optimizing a Hidden Markov Model (HMM)-based pipeline for the ssRNA phages’ discovery. This suggests that ssRNA phages have been overlooked within microbiome studies, and the current studies may have underestimated their contribution to phage diversity [17].
The lack of knowledge of phage diversity, lifestyle, and dynamics in the human body stems in part from a limited toolkit which, until recently, was restricted to classical microbiology methods, including microscopy and culture-dependent approaches [18], as well as a tendency to extrapolate data from one ecosystem to another [4,7]. Unlike classical microbiology, which isolates components of an ecosystem to explain them individually, the multi-omics approach allows the study of the organisms within a complex network of interactions [19]. However, even the advances of new high-throughput multi-omics technologies come with their own challenges and limitations related to sample and downstream processing, sequencing annotation, and in silico predictions [7,11,20]. In this review we describe current advances in the growing field of viromics.

Sample Processing and Downstream Analysis

Some clinical samples, such as skin swabs, are typically limited in volume and have a low abundance of viruses as well as a high background from the host microbiome [21,22]. The entire process, from sample collection to sequencing, will impact the detection of viral sequences and...
needs to be carefully tailored to sample type, origin, and volume [23,24]. Existing sampling tech-
niques have their advantages but can be biased toward recovering the most abundant commu-
nity members. For example, the use of 0.2 μm filters, a common approach for removing large
particles, such as host and bacterial cells from a sample, has been shown to also deplete large
viruses [25] and reduce the amount of recovered viral DNA by half [26]. Similarly, CsCl gradient
ultracentrifugation purification, depending on how the method is performed, can be biased
toward isolating specific phage types and those with atypical buoyancy, but it results in very
pure samples [27,28]. Viral quantification methods, such as epifluorescence microscopy, can
underestimate the actual number of virus-like particles (VLPs) in human samples (Feichtmeyer
et al., unpublished observations). Automated extraction platforms are now frequently used for
virus detection in combination with qPCR or droplet digital PCR due to their higher sensitivity
and high-throughput work capacity [29,30], while commercial kits work better with higher viral
loads (>10^6 copies/ml) and longer DNA fragments (>200 bp) [31]. If an amplification step of
viral nucleic acids is required, the most commonly used methods are: (i) random amplified
shotgun library (RASL) in which the template is restricted to dsDNA; (ii) linker-amplified shotgun
library (LASL), which requires a high template concentration [11,29]; and (iii) multiple displace-
ment amplification (MDA), which tends to overamplify circular single-stranded DNA (ssDNA)
and unevenly amplifies linear genomes [32,33]. As illustrated by these examples, the most
commonly used virus isolation methods have their specific drawbacks and/or biases.

The majority of the unpuriﬁed viral metagenome sequences are assigned to bacteria and eukaryotic
data [12,34]. Thus, removal of background contamination using VLP puriﬁcation methods is
essential to get a clear image of phage abundance in the human body [34]. Recently developed
flow-cytometry-based methods allow for the separation of VLPs from the background microbiota
by labelling phages with a fluorescent dye [35]. VLPs are then selected based on their size and
fluorescence level, and are removed from the sample using fluorescence-activated cell sorting
[35]. While this method still leads to the loss of VLPs, and decreases sensitivity of viral detection
[36], it signiﬁcantly reduces background contamination and can eliminate the need for whole-
genome ampliﬁcation before sequencing [35]. The classical VLP concentration and puriﬁcation
methods are described in depth elsewhere [34,37,38]. However, given that background contami-
nation is currently unavoidable [31,34], viral sequences should be checked against, and puriﬁed
from, contaminating host sequences before any further analysis. Because every available
sample-processing method has its limitations and biases, the study of less described phages is
dependent on a bioinformatics approach, which comes with its own set of advantages and
challenges (Figure 1).

Current Tools and Viral Databases

As phages lack a universal marker, such as the 16S ribosomal RNA genes in bacteria, they can be
hard to identify in a mixed sample [1,7]. Shotgun sequencing of VLP-derived DNA or RNA is one
solution to the issues of metabarcoding that relies on species- or group-specific markers.
Metagenomics allows for culture-independent sequencing of a complex microbial sample
without needing group-specific primers and can distinguish between the different species contained
within the sample. Metagenomic data are, however, prone to high background noise that
confounds current methods used for viral taxonomic characterization [39].

To address the poor, incorrect or insufficient, annotations present in public databases and limited
homology between viral sequences to reference databases, viromics studies rely on de novo as-
sembly to recover viral genomes from metagenomes [40]. However, this assembly can be chal-
lenging due to the specific characteristics of viral metagenomes: they are highly mosaic, include
many repeat regions within the genomes, and show high metagenomic complexity and strain-
Figure 1. Experimental and Computational Approaches for the Characterization of the Free Phage Fraction of the Human Microbiota. Lung and gut illustrations represent respiratory and gastrointestinal tracts that contain the highest number of virus-like particles (VLPs) in the body. The arc indicates two sequencing platforms that do not need an assembly step. Abbreviation: ORFs, open reading frames.
level diversity [19,36,40]. The microdiversity (high level of strain evenness and nucleotide diversity) of abundant phages can also complicate de novo assembly [36,40,41]. Protein-level assemblers, such as Plass [42], can be better tools to use on viral metagenomic data as they predict novel proteins from nucleotide sequences, increase sequence recovery, and improve protein function prediction. They also help avoiding the mismatches from synonymous single-nucleotide polymorphisms [42]. Yet, these assemblers cannot place the assembled protein sequences into a genomic context, and they are unable to resolve homologous proteins from closely related taxa with <95% sequence identity [42]. Long-read sequencers, such as Nanopore or PacBio, could potentially be used to recover a complete phage genome within a single read without the need for assembly. However, long-read sequencers require micrograms of DNA, that is orders of magnitude more than the nanograms usually isolated from a virome sample without amplification, and they still have a relatively high error rate and operating cost [36].

Once a complete or partial phage genome sequence has been assembled, functional annotation is conducted to understand the biological meaning of the predicted genes. For protein-coding genes, open reading frames (ORFs) are predicted using Prodigal [43] and Glimmer [44] or other tools, and they are aligned to protein databases for functional annotation [45]. Virus-specific databases, such as VOGDB [46], HVPC [47], pVOGs [48], GLUVAB [49], IMG/VR [50], Virus-Host DB [51], MVP [52], or general functional annotation databases such as PFAM [53], GO [54], EggNOG [55,56], COGNIZER [57], or KEGG [58], are commonly used for functional annotation (Table 1). As most sequences in general functional annotation databases are derived from the genomes of cellular organisms, this leads to a poor coverage for viral proteins. Meanwhile, virus-specific databases require significant improvement by adding more viral sequences.

Unknown Viruses and Discovery
With the advances in sequencing technology, the total number of uncultivated virus sequences that are identified each year is by far (e.g., five times between 2017 and 2019) more than sequences of virus isolates [59]. As a result, uncultivated viruses already represent the majority (≥95%) of the viral diversity in public databases [59]. Minimum Information about an Uncultivated Virus Genome (MIUViG) standards are being developed within the Genomic Standards Consortium framework to improve the reporting of uncultivated virus genomes. MIUViG asks for information about virus origin, genome quality, genome annotation, taxonomic classification, biogeographic distribution, and in silico host prediction for novel uncultivated viruses [59].

Most viral sequences show no significant homology to known reference sequences [7,60]. An alternative to database-dependent methods is clustering viral sequences by composition. For example, VirMap data processing can detect low coverage and highly divergent viruses and allows for recovery and reconstruction of viral information when closely related database entries are non-existent [39]. Similarly, virMine is not restricted by insufficient viral diversity represented in public databases, and instead scores contigs based on their comparison with both viral and nonviral sequences [61], while PHAST and PHASTER are two web server tools that use public databases for identification and annotation of prophages within bacterial genomes [62]. However, some of the detected prophages may be nonfunctional, secondary to deletions or mutations of essential genes [63].

Machine-learning methods can also be used to detect viral sequences. For example VirFinder [64] uses k-mer profiles to predict viral contigs, VIBRANT utilizes hybrid machine learning and a protein similarity approach that is independent of sequence features for viral sequence recovery [65], and Virsorter detects viral signals using a combination of reference-dependent and reference-independent approaches [66]. Likewise, MARVEL uses a random forest machine
Table 1. Software for Predicting Phage Hosts

<table>
<thead>
<tr>
<th>Tool name</th>
<th>Description</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tools for sample processing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amicon Ultra Centrifugal Filters</td>
<td>Using low-binding Ultracel regenerated cellulose membranes allows for high-throughput VLP concentration and recovery</td>
<td>[28]</td>
</tr>
<tr>
<td>PureLink viral RNA/DNA kit</td>
<td>Allows simultaneous extraction of high-quality DNA and RNA from biological material</td>
<td>[24]</td>
</tr>
<tr>
<td>eMAG</td>
<td>A fully automated nucleic acid extraction platform that enables simultaneous extraction of viral genomic material from 48 specimens</td>
<td>[29]</td>
</tr>
<tr>
<td>GenomePlex, WGA</td>
<td>A whole-genome amplification (WGA) kit for the rapid and highly representative amplification of genomic DNA from minimal amounts of starting material</td>
<td>[12]</td>
</tr>
<tr>
<td>Iron chloride</td>
<td>Useful for concentrating virus particles from large-volume samples</td>
<td>[38]</td>
</tr>
<tr>
<td>MAF</td>
<td>Uses a hydrolysed macroporous epoxy-based polymer system to concentrate and purify waterborne viruses</td>
<td>[30]</td>
</tr>
<tr>
<td>MagNA PURE96</td>
<td>Another fully automated extraction system that allows simultaneous extraction of 96 specimens using magnetic bead technology</td>
<td>[29]</td>
</tr>
<tr>
<td><strong>Tools for viral recovery</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeepVirFinder</td>
<td>An alignment-free tool that identifies viral sequences in metagenomes using machine learning</td>
<td>[68]</td>
</tr>
<tr>
<td>MARVEL</td>
<td>A tool that uses machine learning for prediction of dsDNA phages in metagenomes</td>
<td>[67]</td>
</tr>
<tr>
<td>PPR-Meta</td>
<td>A 3-class classifier that allows identification of phages from metagenomic assemblies with enhanced performance for short fragments</td>
<td>[66]</td>
</tr>
<tr>
<td>PHASTER</td>
<td>A web-based tool for identifying and annotating prophage sequences within bacterial genomes</td>
<td>[62]</td>
</tr>
<tr>
<td>VIBRANT</td>
<td>An automated tool that uses a hybrid machine-learning and protein-similarity approach to recover and annotate viruses of microbes</td>
<td>[65]</td>
</tr>
<tr>
<td>VirFinder</td>
<td>A novel k-mer-based tool that identifies viral sequences from assembled metagenomic data</td>
<td>[64]</td>
</tr>
<tr>
<td>VirMAP</td>
<td>Uses a combination of nucleotide and protein signals to taxonomically classify viral sequences independently of genome coverage or read overlap</td>
<td>[39]</td>
</tr>
<tr>
<td>virMine</td>
<td>Can identify viral genomes from collective raw reads within metagenomes of different environments</td>
<td>[61]</td>
</tr>
<tr>
<td>Virsorter</td>
<td>Can recover novel viruses in metagenomic data using both reference-dependent and reference-independent approaches</td>
<td>[66]</td>
</tr>
<tr>
<td><strong>Functional annotation databases</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COGNIIZER</td>
<td>A comprehensive stand-alone annotation framework that allows functional annotation of sequences from metagenomic data</td>
<td>[57]</td>
</tr>
<tr>
<td>EggNOG</td>
<td>A public source for orthologous groups (OGs) of proteins at different taxonomic levels, with integrated functional annotations</td>
<td>[56]</td>
</tr>
<tr>
<td>GO</td>
<td>A comprehensive knowledge-based resource of gene functions</td>
<td>[64]</td>
</tr>
<tr>
<td>KEGG</td>
<td>The gold standard database for understanding functions of the different biological systems from large-scale molecular datasets</td>
<td>[58]</td>
</tr>
<tr>
<td>PFAM</td>
<td>A database of manually curated protein families, containing 14 831 Pfam-A families</td>
<td>[53]</td>
</tr>
<tr>
<td><strong>Software for predicting phage hosts</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HostPhinder</td>
<td>Uses genomic similarity to a reference database of phages with known hosts to predict hosts for uncharacterized phages</td>
<td>[88]</td>
</tr>
<tr>
<td>IMFH-VH</td>
<td>Kernelized Logistic Matrix Factorization based on Similarity Network Fusion for predicting virus-host association</td>
<td>[89]</td>
</tr>
<tr>
<td>PHISDDetector</td>
<td>Uses several interaction signals, including CRISPR and protein–protein interaction, to predict novel phage–host pairings</td>
<td>[86]</td>
</tr>
</tbody>
</table>
learning approach to predict dsDNA phage sequences in metagenomic bins [67]. The recently developed DeepVirFinder improved the accuracy of viral identification for both long and short sequences using deep learning methods [68], while PPR-Meta uses deep-learning, Bi-path Convolutional Neural Network, to detect phage and plasmid sequences in metagenome assemblies simultaneously. PPR-Meta was specifically developed to improve the identification performance for short fragments [69]. However, a major drawback to viral discovery tools is that they are as efficient as the dataset they were trained on, which can lead to false positives in high confidence scored viral contigs.

For taxonomic classification, assembled contiguous sequences (contigs) are compared with annotated virus databases, either using a best-hit approach, BLAST [70], or a voting system that considers all ORFs, CAT and BAT [71]. The latter approach works best with contigs longer than 1 kb as they contain multiple ORFs [71]. For uncultivated virus sequences with no hits in reference databases, a gene-sharing network, such as vConTACT2, in which viruses are clustered together based on shared genes, can be used to automatically assign tentative taxonomy [72]. ViPTree is a web server that uses protein alignment for phylogenetic analysis and classification of viruses [73]. Concatenated protein phylogeny can also be used for classification of tailed dsDNA viruses [74]. One challenge to the taxonomic assignment of viruses is the dominance of predicted ORFs in which combined taxonomic signals may enhance the classification of unknown sequences [75].

### Identifying Unculturable Phages’ Host Range

How to determine a phages’ host range, that is, the different bacteria it can infect, is a contentious topic of discussion, starting with the definition of infection. The phage infection cycle consists of six main stages. The first step is absorption of the phage into the bacterial cell. Second, the phage ejects its DNA into the host cell. Third, defence mechanisms are evaded. Fourth, the bacterial machinery is hijacked, turning the host into a virocell [76]. Fifth, the phage replicates and builds a new generation of phages. The sixth and final step is lysis of the bacterial cell [77]. Up to seven different types of host range determination methods have been described in depth elsewhere [78,79]. Standard methods for host determination, such as efficiency of plating (EOP),

<table>
<thead>
<tr>
<th>Tool name</th>
<th>Description</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>viruses_classifier</td>
<td>An alignment-free approach for distinction between phages and eukaryotic viruses in metagenome data</td>
<td>[90]</td>
</tr>
<tr>
<td>VirHostMatcher</td>
<td>Uses similar oligonucleotide frequency patterns between phages and bacteria to predict host range for phages on a genus level</td>
<td>[88]</td>
</tr>
<tr>
<td>WsH</td>
<td>Fast and accurate, making it suitable for predicting phages’ host range from metagenomic data</td>
<td>[87]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Virus-specific databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>HVPC</td>
</tr>
<tr>
<td>An integrated reference database of both cultured and uncultured DNA/RNA viruses</td>
</tr>
<tr>
<td>IMG/VR</td>
</tr>
<tr>
<td>A microbe–phage interaction database with over 30 k viral clusters, gathered from public databases and microbiome sequences</td>
</tr>
<tr>
<td>MVP</td>
</tr>
<tr>
<td>A public webserver that provides information about the virus orthologous groups</td>
</tr>
<tr>
<td>pVOGS</td>
</tr>
<tr>
<td>A public webserver that provides information about the virus orthologous groups</td>
</tr>
<tr>
<td>VGDB</td>
</tr>
<tr>
<td>A public webserver that provides information about the virus orthologous groups</td>
</tr>
<tr>
<td>Virus-Host DB</td>
</tr>
<tr>
<td>Includes complete genomes of viruses and their hosts gathered from RefSeq, GenBank, UniProt, and ViralZone</td>
</tr>
</tbody>
</table>
are culture-dependent and the results vary between different methods [80], which makes the host-range determination for unculturable phages difficult (Figure 2). Alternatives to culture-dependent methods are viral-tagging or in silico abundance profiles, determination of tRNAs or prophages, or CRISPR recorded short phage segments [23].

There are a number of different culture-independent methods that can be used to measure phage host range. Viral-tagging uses fluorescence-activated cell sorting to separate out fluorescently labelled phages that are attached to a bacterial cell for further downstream applications and sequencing [18]. While attachment does not equal absorption or replication, it links to the first step of the phage infection cycle and has been demonstrated to successfully predict unique host–phage pairings in both the marine [18,81,82] and human [23] environments. For example, a recent study revealed a total of 363 unique bacteria–phages interactions within the faecal samples from 11 healthy volunteers [23].

Abundance profiles are another culture-independent approach to link phages to hosts by using (lagged) correlations in phage and bacterial abundance patterns. While promising in theory, the complex dynamics [7] underlying the interactions between phages and their hosts tend to defy straightforward correlation analysis, yielding low accuracy of this approach [1,83]. Genetic signatures can sometimes be used to link phages to their bacterial hosts; they are largely associated with the fifth step of the phage infection cycle. The most commonly used genetic signatures are: (i) horizontal gene transfer leading to genetic homology between phage and bacteria, though this is dependent on a comprehensive database [83]; (ii) prophage integration into host genomes, though this is limited to temperate phages [24]; (iii) the recording of a short segment of an infecting phage using CRISPRs to prevent reinfection, which can be used to identify the phage in question.
However, only ~10% of bacteria encode a CRISPR system in the first place which can be identified using state-of-the-art algorithms [83, 84]; and (iv) tracking of viral tRNA thought to originate with the host [68], though this is not specific at a species level and only 7% of known phages have tRNA sequences [23]. Due to these limitations, most in silico tools, such as VirHostMatcher [85], PHISDetector [86], and WISh [87], combine multiple genetic signatures to predict phages’ host range (Table 1). For more virus–host prediction tools see [88–90].

**Statistical Analysis of Multidimensional Data**

Analyzing multidimensional (-omics) data to elucidate species–environment relationships is a challenge currently faced by many disciplines [91]. Recent advances in computational and statistical approaches, such as machine learning, have helped to address this issue (including new tools that combine neural networks, random forests, and indicator species analysis to identify key players in driver-response relationships) [92]. These approaches, however, are data hungry, requiring a large number of observations. In cases where small sample size precludes the use of machine-learning approaches, canonical methods likewise offer a promising path forward in analyzing species–environment relationships [93, 94]. For instance, Multiple Factor Analysis (MFA) [95] was applied in a recent study [2] to examine the multivariate correlation between dominant bacterial and phage species, and environmental metadata. To reduce dimensionality prior to running the MFA, dominant bacterial and phage species were first determined by conducting Principal Component Analyses (PCAs) on each community matrix and identifying the species that contributed most to significant PC axes. This dimension reduction improved the interpretation of the MFA, which in turn helped to relate changes in the gut microbiome to several environmental factors (i.e., health status, diet, age, and sex). Ultimately, the appropriateness of an approach, with respect to the data and question, is likely to vary among studies, and, echoing the conclusions made [92], the integration of multiple, complementary statistical methods will likely offer the most robust conclusions and help to untangle complex, multidimensional data.

**Concluding Remarks**

The human body is one of the densest and most diverse microbial habitats known. The viral fraction alone accounts for ~10^{12} VLPs (it varies from high: ~10^9–10^{10}/g faecal content and ~10^5/ml of respiratory fluids, to low: ~10^8/cm^2 skin and ~10^5/ml blood), plus prophages in bacterial genomes [7, 21, 22]. Phages play critical roles in maintaining gut homeostasis through interacting with the bacterial community [96]. Understanding how phages regulate this complex microbial network will pave the way for the development of novel phage-based therapeutics to re-establish gut health in disease associated with dysbiosis such as inflammatory bowel disease. The novel advances in sequencing technology and bioinformatics have enabled rapid expansion in viral discovery. Yet, we have a way to go until the complete phage diversity (Box 1) in the environment and the human body has been revealed, and the functions of these phages have been elucidated. This is mainly due to isolation protocols and computational shortcomings, despite recent advances to better study phages. A current challenge is to develop suitable isolation methods and in silico analytics to better identify RNA phages. Future studies, in particular, should be adjusted to ensure that RNA phages are adequately represented.

The factors that are responsible for the bias found in phage metagenomics call for the scientific community to work together to improve the toolkits currently used in the field, in the laboratory, and in silico. For example, the application of single-cell technologies can significantly advance our understanding of phages in the human body by identifying their specific function and their interactions with host bacteria, and by revealing their impact on our health. We expect that the phage research field will benefit from near-future technological advancements as the world...
gets closer to completing the picture of global phage abundance, diversity, and distribution, as well as the interactions of phages with their bacterial hosts (see Outstanding Questions).

**Author Contributions**

All authors listed have made a significant, direct, and intellectual contribution to the work and approved it for publication.

**Acknowledgements**

The authors thank Bas E. Dutlih, E. Haggard-ljungquist, and H. Foroughi-Asl for their constructive comments. Kawtar Tiamari created the box image; the image is adapted with permission from Molina B. Dion et al. (10.1038/s41579-019-0311-8). Spot test image in Figure 2 is adapted from Wikimedia Commons; all other images were taken from Smart Medical Art (SMART): https://smart.servier.com. This work was funded by German Research Foundation (DFG Emmy Noether program, Project No. 273124240; and SFB 1371, Project No. 395357507) and European Research Council starting grant (ERC SIG 803077) awarded to L.D.

**References**