We are becoming accustomed to the idea that healthy humans are colonized by a rich diversity of microorganisms — the microbiome. However, less well known is the fact that the human body also hosts vast numbers of different viruses, collectively termed the ‘virome’. Viruses are believed to be the most abundant and diverse biological entities on our planet, with an estimated $10^{31}$ particles on Earth. The human virome is similarly vast and complex, consisting of approximately $10^{13}$ particles per human individual, with great heterogeneity. In recent years, studies of the human virome using metagenomic sequencing and other methods have clarified aspects of human virome diversity at different body sites, the relationships to disease states and mechanisms of establishment of the human virome during early life. Despite increasing focus, it remains the case that the majority of sequence data in a typical virome study remain unidentified, highlighting the extent of unexplored viral ‘dark matter’. Nevertheless, it is now clear that viral community states can be associated with adverse outcomes for the human host, whereas other states are characteristic of health. In this Review, we provide an overview of research on the human virome and highlight outstanding recent studies that explore the assembly, composition and dynamics of the human virome as well as host–virome interactions in health and disease.
and these have been the most studied. This site is rich in cells of both the human gut and prokaryotic microbiota, providing a rich variety of hosts for viral growth. Most other sites have sparser microbiota, at least in healthy individuals, and so sparser viral populations as well; however, recent work has defined rich viral populations at many locations throughout the human body. A high degree of inter-individual variation is seen in the human virome, paralleling findings from bacterial and fungal communities, raising questions of to what extent inter-individual differences in phenotypes are attributable to differences in the virome (TABLE 1).

Recent studies have uncovered numerous factors that show associations with virome structure in addition to anatomical location, such as diet, age and geographic location of the individual sampled. Disease is another prominent influence, with emerging studies suggesting associations between virome structure and inflammatory bowel disease, diabetes, hypertension and cancer (box 1). Most studies have only reported associations of the virome and diseases — thus, more work is needed to understand the directions of causal relationships.

In this Review, we provide an overview of research on the human virome, focusing both on secure conclusions and on areas where additional work would be helpful. We first discuss the virome at different body sites in healthy humans. Next, we summarize nascent work on the origin of these populations, that is, the initial assembly of the human virome in neonates and infants after delivery. We then summarize data on interactions between the host and the virome, including links to disease states.

**Human virome diversity**

The number of bacterial cells associated with the human body is estimated to be roughly the same as the number of human cells, \( \sim 10^{13} \) in total\(^{20} \). One estimate suggests that each gram of metazonal tissue is associated, on average, with \( 3 \times 10^9 \) bacteria\(^2 \), and the complexity of these communities increases with body size\(^{22} \). VLP counts in humans reveal that the ratios between viruses and bacteria range from roughly 0.1 to 10, suggesting that the total number of viruses in the human body is of a similar order to the number of bacterial and human cells (reviewed in ref \(^{19} \)).

Viruses that are found in humans can be categorized by various features. Viral genomes can be either RNA or DNA, and either double-stranded or single-stranded. Genome sizes can be as low as a few kilobases or as large as hundreds of kilobases. All viruses have a protein shell (or capsid) surrounding the genome; viruses can also be enclosed in one or more lipid membranes. Viral particles can be classified according to their morphology — spherical (usually icosahedral), filamentous, bullet-shaped, pleomorphic or, for phages, tailed.

Most constituents of the human virome are inferred to be phages. This is an inference because, in most cases, the majority of sequences uncovered in a virome metagenomic sequencing experiment do not align with any information present in existing databases (BOX 1), so it is unknown whether they are phages or some other virus types. The major taxa of phages are typically Caudovirales (tailed phages) and Microviridae (icosahedral non-tailed phages) of the group including phage ΦX174. Phages can have multiple relationships with their bacterial hosts, which constrain ideas about virome dynamics (fig. 1). Lytic phages inject their nucleic acid into cells, leading to the synthesis of new viral components, the assembly of particles and the lysis of host cells, thus releasing progeny phages. Another mode of replication, carried out by temperate phages, can spare the host initially. Phages inject DNA into cells, after which the DNA can then become integrated into the host cell chromosome, forming a prophage. Prophages are maintained in a quiescent state by repressor proteins\(^{31} \). However, should conditions become unfavourable in the host cell — for example, by damage to DNA — phage DNA can become excised from the cellular chromosome leading to lytic growth, resulting in cell lysis and the production of viral progeny. Phages can also have other relationships with their hosts — one is pseudolysogeny, in which phage genomes persist as episomes without

**Table 1** | **Examples of viral population alterations in human disorders**

| Human disease | Sample | Major virome alteration | Ref. |
|--------------|--------|-------------------------|------|
| Severe acute malnutrition | Faeces | Reduced viral diversity | \(^{10} \) |
| Crohn’s disease and ulcerative colitis | Faeces | Increased Caudovirales richness | \(^{14} \) |
| Crohn’s disease | Faeces and biopsies | Moderate alterations | \(^{152} \) |
| AIDS | Faeces | Increased enteric adenoviruses | \(^{91} \) |
| Type 1 diabetes | Faeces | Reduced viral diversity | \(^{16} \) |
| Hypertension | Faeces | Erwinia phage ΦEaH2 and Lactococcus phage 1706 may be associated with hypertension | \(^{18} \) |
| Type 2 diabetes | Faeces | Increased putative phage scaffolds | \(^{153} \) |
| DOCK8 deficiency | Skin swabs | Increased skin virome, especially human papillomavirus | \(^{71} \) |
| Colorectal cancer | Faeces | Increased viral diversity | \(^{19} \) |
| Crohn’s disease and ulcerative colitis | Faeces | Increased Caudovirales abundance | \(^{11} \) |
| Type 1 diabetes during pregnancy | Faeces | Increased picobirnaviruses and tobamoviruses | \(^{17} \) |
| Bacterial vaginosis | Vaginal swabs | Viral population structures correlated with bacterial vaginosis | \(^{74} \) |
| Early-diagnosed Crohn’s disease and ulcerative colitis | Gut biopsies | Increased Hepadnaviridae and Hepeviridae; reduced Polydnaviridae, Tymoviridae and Virgaviridae | \(^{154} \) |
| Coeliac disease autoimmune | Faeces | Increased enteroviruses | \(^{169} \) |
| Crohn’s disease | Faeces | The virulent phages are replaced with temperate phages | \(^{13} \) |
| Ulcerative colitis | Gut biopsies | Increased Caudovirales, phage and bacteria virulence functions, and loss of viral-bacterial correlations | \(^{15} \) |
| HIV viraemia | Seminal fluid | Increased human cytomegalovirus | \(^{75} \) |
| Very early-onset inflammatory bowel disease | Faeces | Increased ratio of Caudovirales to Microviridae | \(^{12} \) |
| Haematopoietic stem cell transplantation | Faeces | Increased picobirnaviruses | \(^{155} \) |
integration. In another mode, phages can bud out of infected cells, sparing the host cell from lysis and death.

Viruses that infect human cells are also an important part of the human virome. Some may cause acute infections, and others may establish long-term latency. Some viruses engage in benign colonization and cannot be associated with any particular disease, appearing to be long-term ‘passengers’ or ‘commensals’. Virome sequencing studies have unveiled some new lineages of human viruses that appear to be such commensals. For example, the Anelloviridae are a family of non-enveloped, single-stranded DNA viruses with quite small circular genomes (2–4 kb), including torque teno virus, torque teno mini virus and torque teno mini virus⁶³. The genomes of viruses in this family encode what appears to be a single large protein, although gene expression has not been

---

**Box 1 | Methods for studying the virome and the ‘dark matter’**

Isolating virus-like particles (VLPs) from samples of interest is the first step of a virome study. Focusing on VLPs avoids wasting resources on sequencing of non-viral nucleic acids. Filtration is a typical first step to purify VLPs, taking advantage of the fact that viruses are smaller than cells. Commonly used filter pore sizes include 0.2 μm and 0.45 μm. Filtering VLPs using a 0.2 μm filter may lose large viruses⁶⁰, whereas 0.45 μm filters may fail to retain some bacteria, as rare bacteria are <0.45 μm⁶⁰. However, these bacteria appear to be uncommon in humans¹⁶⁰, so use of 0.45 μm filters seems optimal. Unpublished comparisons of the two filter types in our laboratory showed no differences in contamination with bacterial 16S rRNA gene sequences, supporting use of 0.45 μm filters.

Other fractionation steps can provide purer VLPs, but often at the expense of losing viral material. Chloroform extraction can be used to disrupt cell membranes, but results in the loss of membrane-enclosed viruses. Comparison of VLP samples purified from human stool with or without chloroform treatment showed that chloroform modestly reduced the recovery of contaminating bacterial DNA. For stool, membrane-enclosed viruses are rare¹⁶⁰, so chloroform treatment may allow superior VLP purification.

For virome nucleic acid preparations, absolute yields are typically low, so multiple displacement amplification using phage Φ29 DNA polymerase has been used to increase DNA amounts. However, this can result in unequal amplification of different DNA forms, with small circular single-stranded DNA genomes amplifying preferentially⁶²,²⁶, distorting the ratios of different viral lineages. Nevertheless, this is often the method of choice.

Numerous algorithms, databases and pipelines have been developed to process virome sequencing data, supporting quality control, read dereplication, contig assembly, viral identification, dark matter exploration and other steps (see the table). Despite steadily improving methods, it remains challenging to identify unstudied viruses in complex mixtures. Researchers can align individual sequence reads to databases, allowing identification of many candidate viruses, but attributions are often tentative and a large proportion of reads are typically not attributed to anything (that is, the ‘dark matter’). Assembling reads into contigs can provide more opportunity to explore previously unstudied sequences, but still a substantial fraction of contigs may find no attribution. Combining contig analysis with cataloguing gene types, and quantifying matches to gene families known to be viral, can help with attribution. Another helpful step can be assessing circularity in contigs, which commonly indicates completion of a metagenomic viral sequence (although this too can arise as an artefact of sequence repetition in a contig assembly).

No standards of dark matter annotation have been established, so the proportion of viral dark matter varies between studies. Values range from 13 to 95% (reviewed in RefS¹⁶⁸). In some cases, a community will be dominated by a well-known virus, driving down the unknown proportion, but commonly the unknown sequences represent a majority. A recent study constructed 33,242 viral species from 32 available public human virome studies, and >32,000 species were predicted to be phages, which is 12-fold higher than the phage number in the RefSeq database¹⁵⁵. Thus, metagenomic sequencing boosts the discovery of novel viral genomes enormously. The vast majority of these probable viruses have yet to be grown in cell culture, classically a key step in establishing viral taxa, but today viral lineages are commonly specified as sequences before successful culture. As a consequence, for many members of the virome, it is not known which organisms serve as hosts for replication. The field is developing stand-alone virome-derived contig databases that can be updated regularly, such as the Gut Virome Database (GVDB)¹⁶⁸, which are useful for updating new discoveries and establishing standards for annotating the dark matter.

RNA viruses of the human virome have generally been less studied. RNA appears to be less stable in samples typically used for metagenomics, so the RNA viruses may be undersampled for technical reasons. Many studies focus on phages, and only a few RNA phages have been reported. Recent studies have highlighted a much greater diversity of RNA phages than previously thought in certain ecological niches³¹⁶,³¹⁷, focusing interest in further studies of RNA viruses.

Another challenge is DNA contamination in low-biomass samples. Studies have found that low levels of environmental DNA are usually present in laboratory reagents and that this DNA can predominate in samples that start with low amounts of nucleic acid³⁰,³¹⁶,³¹⁷. The yields of nucleic acids from virome preparations are typically modest, so it is important to consider contamination when interpreting virome data. It is essential to analyse multiple blank samples in parallel with experimental samples so as to document the contamination background — commonly, at least a little of this background will annotate as viral, providing considerable challenges in data analysis. Sequence data from negative controls should be uploaded to databases along with virome data as part of the publication process. One useful tool for evaluating virome enrichment in virome sequence data is ViromeQC¹⁶⁸. Overall, it is important to be mindful of the potential contribution of contamination and to carefully consider the potential impact on interpretation of experimental results.

---

**Software for analysis of the human virome**

| Name | Function | Ref. |
|------|----------|------|
| ViromeQC | Detects contamination | 108 |
| CD-HIT | Reads clustering | 109 |
| grabseqs | Public data downloading | 170 |
| VirSorter | Viral contig annotation | 171 |
| VirFinder | Viral contig annotation | 172 |
| FastViromeExplorer | Viral contig annotation | 173 |
| Metavir | Viral contig annotation | 174 |
| MetaPhinder | Viral contig annotation | 175 |
| ViromeScan | Viral contig annotation | 176 |
| vConTACT | Taxonomic classification | 177 |
| NCBI Virus | Database | 178 |
| UniProt | Database | 179 |
| pVOG | Database | 180 |
| Pfam | Database | 181 |
| Vfam | Database | 182 |
| Sunbeam | Analysis pipeline | 183 |
| Cenote-Taker | Analysis pipeline | 184 |
| VirusSeeker | Analysis pipeline | 185 |
| PHACT | Phage life cycle classification | 186 |
| PHASTER | Prophage prediction | 187 |
well studied because representative *Anelloviridae* have not yet been grown in pure culture. *Anelloviridae* are extremely diverse and can be found in many human body sites in a large fraction of all humans examined. No specific pathogenic effects have been linked to *Anelloviridae* so far (reviewed in refs 27,28). Greater abundance of *Anelloviridae* has been found in individuals who are immunocompromised, including the recipients of lung transplants, individuals who are HIV positive and individuals on immunosuppressive medications owing to inflammatory bowel disease, indicating that *Anelloviridae* are normally under host immune control12,14,29–31. The newly discovered *Redondoviridae* are another family of small circular DNA viruses that appear to be widespread commensals commonly found in the respiratory tract32–34.

High inter-individual variation in the human virome has been reported in many studies (reviewed in refs 9,10,35–38). However, within a healthy adult, the virome is usually relatively stable over time, paralleling stability in the cellular microbiome. For example, one study found that ~80% of viral contigs present persisted over a span of 2.5 years in the gut of one individual6. Another recent study tracked the gut virome of 10 individuals and found that >90% of recognizable viral contigs persisted in each individual over 1 year9. Studies on the oral virome revealed similar stability40,41. As discussed below, destabilization of the virome is often associated with disease states.

### Virome of different body sites

Numerous recent studies have characterized the human virome at different body sites, revealing rich populations at numerous locations (FIG. 2). Phages are distributed widely across the human body, and different anatomical sites may have quite different phage composition due to the presence of different host bacteria. The distribution of eukaryotic viruses also differs at different body sites. Some notable site-specific features are summarized below.

#### Gastrointestinal tract.

The gastrointestinal tract is commonly the most abundant site of viral colonization, reaching ~10⁹ VLPs per gram of intestinal contents. Analysis of virome sequence data suggests that phages are the most abundant identifiable members of this population (reviewed in refs 9,10,35–38). Visualization of
stool VLPs using electron microscopy reveals that the majority of the phages belong to the order *Caudovirales* (tailed phages; reviewed in ref. 45). Metagenomic sequencing of the human gut also indicated that *Caudovirales* is commonly predominant, along with the spherical *Microviridae* (reviewed in refs 9,10,35–38).

It has been suggested that the most prevalent phage lineage in the human gut is often crAssphage (cross-assembly phage; members of the *Caudovirales*), specifically the short-tailed podoviruses, which infect bacteria of the phylum Bacteroidetes, common members of the gut microbiota45–46. crAssphages are commonly found in greater than 50% of human gut content samples and show a global distribution42–44. The abundance of crAssphages can be up to 90% of a human gut viral community42. Recently, the crAssphage dCrAss001 has been shown to infect *Bacteroides intestinalis*46, specifying one host species experimentally. Genomic analysis shows that phage genes important for lysogeny are rare or absent in crAssphage genomes45,46, and a lytic mode of replication has been suggested by infection studies in vitro46. However, curiously, proliferation of crAssphages does not appear to reduce the growth rate of their host cells in vitro46. One explanation is that crAssphages replicate via pseudolysogeny, in which the phage genome persists in a quiescent state as an episome (reviewed in ref. 46), only rarely lysing the host cell.

The healthy human gut usually contains relatively low proportions of eukaryotic viruses. DNA viral lineages occasionally detected include *Anelloviridae*, *Geminiviridae*, *Herpesviridae*, *Paroviridae*, *Polyomaviridae*, *Adenoviridae*, *Circoviridae* and *Papillomaviridae* (reviewed in ref. 48). The most commonly detected RNA viruses include *Caliciviridae*, *Picornaviridae*, *Reoviridae* and some plant viruses that appear to originate in food, such as *Virgaviridae*49,50.

A notable pattern in viruses of the gut is that most residents, including both phages and human viruses, are not enveloped. This makes sense — lipid envelopes are unlikely to survive the detergent effects of bile salts, dehydration in the large intestine and the conditions of the environment outside the gut required for transmission via the faecal–oral route. A recent statistical analysis showed a significant association between faecal–oral transmission and the absence of a lipid envelope42. It will be of interest to investigate virus structure–transmission relationships in more detail and at additional human body sites.

Surprisingly, coronaviruses are an exception; despite possessing a lipid envelope, many coronaviruses are known to be transmitted via the faecal–oral route. For severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), viral RNA has been widely reported in faeces, although the infectious potential of these viruses

---

**Table:**

| Body Site          | Eukaryotic Viruses | Phages |
|--------------------|--------------------|--------|
| Oral cavity        |                    |        |
| Eukaryotic viruses | *Herpesviridae*    | Phages |
| Anelloviridae      | Siphoviridae       | Podoviridae |
| Redoviridae        | Myoviridae         | Microviridae |
| Papillomaviridae   |                    |        |
| ~10^8 VLPs per millilitre of saliva |

| Eukaryotic viruses | Phages |
|--------------------|--------|
| *Herpesviridae*    | Siphoviridae |
| Podoviridae        | Myoviridae |
| Myoviridae         | Microviridae |

| Lung               | Eukaryotic viruses |
|--------------------|--------------------|
| Anelloviridae      | Siphoviridae       |
| Adenoviridae       | Podoviridae        |
| Picornaviridae     | Myoviridae         |
| Circoviridae       | Microviridae       |
| Virgaviridae       |                    |
| ~10^9 VLPs per gram of faeces |

| Gastrointestinal tract | Eukaryotic viruses |
|------------------------|--------------------|
| Anelloviridae          | Siphoviridae       |
| Adenoviridae           | Podoviridae        |
| Picornaviridae         | Myoviridae         |
| Circoviridae           | Microviridae       |
| Virgaviridae           |                    |
| ~10^9 VLPs per millilitre of cerebrospinal fluid |

| Nervous system | Eukaryotic viruses |
|---------------|--------------------|
| *Herpesviridae* | Siphoviridae |
| Phages         | Podoviridae |
| Myoviridae     | Microviridae |

| Blood          | Eukaryotic viruses |
|----------------|--------------------|
| *Herpesviridae* | Siphoviridae |
| Picornaviridae | Podoviridae |
| Polioviridae   | Myoviridae |
| ~10^9 VLPs per millilitre of cerebrospinal fluid |

| Skin           | Eukaryotic viruses |
|----------------|--------------------|
| *Adenoviridae* | Siphoviridae |
| *Circoviridae* | Podoviridae |
| *Herpesviridae*| Myoviridae |

| Urinary system | Eukaryotic viruses |
|----------------|--------------------|
| *Papillomaviridae* | Siphoviridae |
| *Herpesviridae*    | Podoviridae |
| ~10^9 VLPs per millilitre of urine |

| Semen          | Eukaryotic viruses |
|----------------|--------------------|
| *Anelloviridae* | Siphoviridae |
| *Papillomaviridae* | Podoviridae |
| *Herpesviridae* | Myoviridae |

| Phages | Unknown |
|--------|---------|
|        |         |

### Fig. 2 | The human virome at different body sites.

Summary of viruses found at each human body site. Viral types are summarized from published virome surveys2,5,10,12–13,16–19,40–45,47–53; those known at each body site are likely to increase as more human populations are surveyed and more viral types are discovered. In a few cases, viral lineages were excluded from the analysis because they likely represent contaminants or misattributions. These exclusions include mimivirus, phycodnavirus, marseillevirus, flaviviruses and poxviruses in blood, and baculovirus in the vagina. VLP, virus-like particle.
Multiple displacement amplification
A whole-genome amplification method, which starts by binding random primers to the template DNA and is then followed by strand displacement DNA synthesis performed by DNA polymerase, usually Φ29 DNA polymerase.

Primary immunodeficiencies
A group of rare immune disorders caused by genetic defects.

is uncertain[5]. One possible explanation is that coronaviruses such as SARS-CoV-2 can replicate in host cells of the lower gastrointestinal tract, so they do not need to traverse the entire gastrointestinal tract in order to appear in faeces; another possibility is that coronavirus particles are relatively stable for enveloped viruses. It will be useful to investigate these questions and clarify the infectivity of coronaviruses in faecal material systematically.

Oral cavity. The human oral cavity contains diverse viral communities as well as complex microbial populations. To date, saliva samples have been the primary source of material to characterize the oral virome[40,53], revealing abundant viral populations. Additional oral microenvironments, such as dental plaque[44], have also been studied, revealing high diversity in these environments as well. Staining of particles with a fluorescent dye that binds DNA, followed by visualization under a fluorescent microscope, shows approximately 108 VLPs per millilitre of saliva in healthy humans[45]. The most abundant taxon of phages in the oral virome is the Caudovirales[40,41,46].

Common eukaryotic viruses in the oral cavity of healthy adults include Herpesviridae, Papillomaviridae, Anelloviridae and Redondoviridae[40,57]. Anelloviridae are the most common, but, surprisingly, the newly discovered Redondoviridae[51] is the second most common virus family. However, it should be noted that prevalence measures are expected to be dependent on the sampling methods used, and the multiple displacement amplification steps typically used to amplify virome samples likely favour amplification of small DNA circles, potentially boosting multiple displacement amplification. Metagenomic analysis of skin swabs revealed the presence of multiple eukaryotic virus families including Polyomaviridae, Papillomaviridae and Circoviridae[68]. In a recent DNA virome study with VLP enrichment[18], ~95% of the viral sequences did not match a known viral genome and of the reads that could be assigned, many were associated with Caudovirales. This study also reported that the skin of healthy individuals harboured eukaryotic viruses including Adenoviridae, Anelloviridae, Circoviridae, Herpesviridae, Papillomaviridae and Polyomaviridae[59]. A notable expansion of these eukaryotic viruses was observed on the skin of individuals with primary immunodeficiencies, indicating the importance of immune surveillance in controlling viral colonization of the skin[5].

Respiratory tract. Virome analyses have been performed on respiratory tract samples including sputum, nasopharyngeal swabs and bronchoalveolar lavage, showing that the healthy human lung and respiratory tract can be populated by large viral communities[29,38-42]. Among human DNA viruses, Anelloviridae has been reported to be the most prevalent family of DNA viruses, followed by Redondoviridae[38,42]. Additional eukaryotic viruses frequently detected include Adenoviridae, Herpesviridae and Papillomaviridae[29,38-42]. Phages are commonly found, including Caudovirales, Microviridae and Inoviridae[20,38-42]. Phages, like most of the cellular microbiota, even when found in the lung appear to be derived mainly from the abundant bacterial populations in the mouth and upper respiratory tract.

Blood. Viruses of blood have been studied closely, to understand human health and also to assess the safety of donor blood supplies. A pioneering study found viral particles in blood using electron microscopy and identified genomic sequences related to several eukaryotic viruses, including Anelloviridae[29,51]. Other recent studies have reported Herpesviridae, Marseilleviridae, Mimiviridae, Phycodnaviridae and Picornaviridae families, with proportions varying with the geographical site sampled[20,63]. Some of these findings illustrate the challenges of virome studies, where samples with low levels of authentic viruses may be prone to contamination (BOX 1). It is important to note that Marseilleviridae, Mimiviridae and Phycodnaviridae are not known to replicate in human cells, and may represent environmental contamination. Phages belonging to the Myoviridae, Siphoviridae, Podoviridae, Microviridae and Inoviridae families have also been reported in blood, and, similarly, their origin is unclear (reviewed in REFS[18,60,63]). Some of these studies did not perform viral particle enrichment prior to sequencing, so that phage sequences integrated in bacterial genomes (rather than viral particles) may have been identified. A recent study reported that phage particles may be transported across gut epithelial cell layers by transcytosis, potentially reaching systemic circulation, with unknown consequences[46]. Thus, much remains to be learnt about the blood virome of healthy individuals; however, it does seem likely that at least Anelloviridae are common in blood and so must be expected to be present throughout the blood supply.

Skin. Compared with other body sites, the skin has a relatively low microbial biomass, which can, for some samples, make it difficult to distinguish the resident microbiome and virome from various forms of contamination. Metagenomic analysis of skin swabs revealed the presence of multiple eukaryotic virus families including Polyomaviridae, Papillomaviridae and Circoviridae[68]. In a recent DNA virome study with VLP enrichment[18], ~95% of the viral sequences did not match a known viral genome and of the reads that could be assigned, many were associated with Caudovirales. This study also reported that the skin of healthy individuals harboured eukaryotic viruses including Adenoviridae, Anelloviridae, Circoviridae, Herpesviridae, Papillomaviridae and Polyomaviridae[59]. A notable expansion of these eukaryotic viruses was observed on the skin of individuals with primary immunodeficiencies, indicating the importance of immune surveillance in controlling viral colonization of the skin[5].

Urogenital system. Urine samples from healthy humans have been reported to contain viruses in the region of 107 VLPs per millilitre[16]. Most of the identifiable viruses were phages; in addition, human papillomaviruses could be identified in >90% of subjects in some cohorts[2,7]. Virome analyses of healthy vaginal samples showed that the majority of identified viral sequences are derived from double-stranded DNA phages, with eukaryotic viruses contributing only 4% of total reads[41]. In seminal fluid, Anelloviridae, Herpesviridae and multiple genotypes of Papillomaviridae have been detected[5]. Thus, for these
body sites we again see mixtures of viruses that replicate in human cells and viruses from the resident microbiota.

**Nervous system.** Little information is available on virome populations in the nervous system in healthy humans. A recent study estimated the VLP number at ~10⁴ per millilitre of cerebrospinal fluid, with phages predominant, including *Myoviridae*, *Siphoviridae* and *Podoviridae*. *Herpesviridae* were also detectable. The clinical consequences of infection by herpesviruses in the nervous system have been well studied. Herpes simplex viruses, human cytomegalovirus and varicella zoster virus can establish latent infections in the central nervous system without symptoms; these viruses can later be reactivated and produce viral particles.

To summarize the virome over multiple human body sites, the human gut contains the most abundant viruses and has been the most frequently studied. Lower particle numbers are found at other body sites, but all seem to have detectable viral colonization. For sites with a resident microbiota, the viruses found are typically a mixture of viruses replicating in the local human cells and viruses infecting the local microbiota. The extent of circulation between sites is just starting to be assessed. Upper respiratory microbiota likely contribute much of the microbiome and virome to lower respiratory sites, at least for the phage population. Small circular DNA viruses (*Anelloviridae* and *Redoviridae*) are common in respiratory samples and can also be found in faecal samples, although it is not known whether they appear in the gut because of replication in the gastrointestinal tract or swallowing of saliva. *Anelloviridae* but not *Redoviridae* appear in blood, indicating systemic circulation. Even sites that are mostly cut off from normal microbial colonization, such as cerebrospinal fluid, show low levels of viruses including phages. How much of this colonization is due to true local viral replication, how much is due to systemic circulation of viruses and how much is due to technical mishaps associated with reagent contamination remains to be fully clarified. At all of these sites, unannotated ‘dark matter’ sequences are prominent, emphasizing how much remains to be learnt.

**Establishment of the human virome**

**Timing of the first microbial colonization.** The timing of virome establishment is linked to the question of establishment of the microbiome as a whole in human neonates and infants. Historically, the inability to culture microorganisms from samples from healthy deliveries has supported the idea that neonates are usually born sterile. Starting in 2014, several studies using metagenomic sequencing were carried out, leading to the proposal of a microbiome in the placenta, amniotic fluid and even the fetus, implying that microbial colonization may start in utero. However, multiple recent studies have indicated that these detections of microorganisms are likely to be false positives due to experimental contamination, and that no placental microbiome is present before rupture of membranes and delivery. This raises the question of when viral colonization takes place in human neonates. Vertical transmission of pathogenic viruses during pregnancy has been well documented for rubella virus, human cytomegalovirus, herpes simplex viruses, HIV, Zika virus and human papillomavirus (reviewed in refs [96][97]). However, these are characteristic of disease states and not health. Virome populations are robust in adults, so the question is when does the virome become established in healthy neonates.

**The virome at delivery.** An early study of virome colonization sampled meconium shortly after delivery, and failed to find VLPs using epifluorescent microscopy but did report ~10⁴ VLPs per gram at 1 week of life, suggesting that the neonate lacked a virome at birth but was quickly colonized. In a study of amniotic fluid, no evidence was found for the existence of a virome in healthy pregnancies before delivery. A later study of neonate stool samples, investigating both RNA and DNA viromes, taken a median of 2.6 days after birth found high diversity in the gut virome, consistent with rapid colonization after delivery. Another metaorganism study using stool samples collected a median of 37 h after birth again showed high viral diversity, supporting rapid virome acquisition. A more recent study of VLPs sampled a median of 17 h after birth reported that only 15% of samples were positive. Thus, the picture that emerges is that neonates usually lack a detectable virome at birth, but are rapidly colonized after rupture of membranes and delivery (FIG. 3).

**The first detectable viruses — a predominance of phages.** Several studies have investigated the nature of the human virome in samples taken very early in life. Recognizable early colonizers were mainly phages of the *Siphoviridae*, *Podoviridae* and *Myoviridae* families. Another phage family, the *Microviridae*, are less abundant during early life but rise in abundance with age. Early bacterial colonizers commonly include *Escherichia*, *Klebsiella*, *Enterococcus*, *Staphylococcus* and *Streptococcus* species and the phages of these bacteria are some of the most common early virome members.

A recent study investigated the production of the virome in gut samples of infants at 1 month of life and concluded that lytic phages were relatively rare. Ongoing replication of lytic phages could not be detected in direct infection assays; viral sequences were annotated primarily as lysogenic phages and not lytic phages; and *Microviridae* and *crAssphages*, which usually do not form lysogens, were rare or absent during the first month in infant guts. Instead, prophage induction was found to contribute most of the gut virome. Bacterial strains were isolated from infants’ stools, and many were found to produce viral particles at high levels. The viruses that were produced could be detected as prophage sequences in the bacterial genome sequences. Sequences of the induced phages were also commonly found in the infant stool virome, and the abundance of each type of VLP in stool was positively correlated with the abundance of the host bacteria in the same sample.

These experiments suggested a high rate of phage induction in the infant gut, raising the question of what constitutes the inducing signal. DNA damage is the most studied signal (reviewed in ref. [98]). In vitro studies suggested that spontaneous induction rates may be
relatively low \[^{99,97}\]. Bacterial metabolites, nutrients and bile salts have all been shown to induce prophages in some models \[^{10,101}\]. Thus, the signals (if there are any) are unclear and an interesting topic for future studies.

**Later evolution of the paediatric virome.** The paediatric virome continues to mature with age. Lytic phages appear to become more common later in life. For example, the crAssphages are mostly absent in the first month of life but become more prominent by month 4 (REF. \[^{49}\]). The gut virome also often undergoes a shift from **Caudovirales**-dominated to **Microviridae**-dominated \[^{12,102}\].

Another question is the possible influence of the mode of delivery. A metagenomic analysis of faecal virome samples from 20 infants at 1 year of age compared spontaneous vaginal delivery with caesarean section and found that the birth mode resulted in distinctly different viral communities, with infants born by spontaneous vaginal delivery having greater viral diversity \[^{103}\].

The effect of the delivery mode has been reproduced in some, but not all, cohorts in other studies \[^{49}\]. Studies linking the birth mode and long-term health outcomes reported greater risk for various diseases, including asthma, obesity and diabetes, in children delivered by caesarean section, although the results are controversial and research is ongoing \[^{104–108}\]. It may be useful to consider possible influences of the virome as well.

**Colonization of infants with viruses infecting human cells.** The viruses that replicate in human cells are also detected in metagenomic surveys of samples taken in early life. Gastroenteritis is one of the leading causes of childhood mortality, resulting in more than two million deaths every year \[^{109}\], and viral pathogens are primary causes (reviewed in REF. \[^{110}\]), highlighting the importance of paediatric virome studies. Viruses that have been reported to be associated with childhood diarrhoea include rotavirus, astrovirus, calicivirus, picornavirus,
polyomavirus and adenovirus (reviewed in REF.110). Less well known is the fact that these viruses are commonly found in healthy infant guts in metagenomic studies49,50. Other common viruses in healthy infant guts include parvovirus and anellovirus12,13,14.

Thus, recent data suggest that healthy infants are colonized in a stepwise fashion. In the first step, prophage induction from pioneering bacteria provides an initial population. Later, lytic phages become more common, and also viruses that replicate in human cells.

**Factors that shape the human virome**

Numerous factors have been reported to influence the human virome and, ultimately, affect health (FIG. 4), starting in infancy and extending throughout the life of the individual.

**Diet**. The infant virome, including both phages and eukaryotic viruses, can be affected by diet. Breastfeeding is well established to reduce viral gastroenteritis and infant mortality (reviewed in REFs 111,112). A recent metagenomic virome study showed lower accumulation of animal cell viruses in guts of infants fed with breast milk49 (FIG. 3). The protective effect was seen in cohorts from both the United States and Botswana50. Viruses affected included Adenoviridae, Picornaviridae and Caliciviridae. Breast milk contains multiple components that protect children from intestinal infections, such as maternal antibodies, oligosaccharides and lactoferrin (reviewed in REF.111). These antiviral components have been reported to inhibit viruses, such as rotavirus, norovirus, enterovirus, influenza virus and SARS-CoV (reviewed in REFs 113,113–115). The phage population structure in infant stool can also be influenced by breastfeeding — specific bacteria are known to increase in abundance with breastfeeding, and in a recent report their phages were increased in abundance as well16. Alterations of the viral population in infant stool samples owing to breastfeeding has been proposed in another study, although the small sample size was a limitation17. Furthermore, it has been suggested that the infant virome may, in fact, be at least partially transmitted from the mother via breast milk122,123.

Diet has also been reported to affect the virome in adults. For example, comparison of the virome structure in adult subjects on two different controlled diets showed that individuals on the same diet showed more similar viral compositions than those on different diets6.

**Host genetics and immunity**. Intense interest has focused on the potential influence of human genetics on the microbiome, raising the linked question of the role of human genetics in programming the virome. Several studies of twins, comparing monozygotic and dizygotic twin pairs, have suggested an influence of human genetics on microbiome composition because the monozygotic twins showed more similarity124–126. By contrast, a recent large-scale study of healthy adults (non-twins) reported little effect of genetics and emphasized environmental factors127. In studies of the infant virome, the gut virome compositions of co-twins were more similar than those between unrelated individuals128,129, but this similarity was not strongly affected by zygosity130,131, emphasizing the importance of the shared environment over human genetic composition. In an early study of adult twins, no greater similarity of virome samples was seen in twin pairs13. In a more recent study with a larger cohort of adult monozygotic twins, some were found to have greater similarity in virome composition compared with unrelated individuals132,133, but this similarity was no longer notably similar134. Collectively, studies of the gut virome in twin pairs so far show similarities early in life, but not stronger similarities in monozygotic twins versus dizygotic twins, thus emphasizing the importance of shared environmental factors in colonization versus contributions of genetic make-up.

However, in inherited diseases such as primary immunodeficiencies, the effects of genetics on the virome are well established. In some cases, phenotypes of mutations in human genes can only be understood by considering the interaction with the virome. One drastic example involves epidermodysplasia verruciformis (also known as treeman syndrome, which is a rare heritable skin disease). In the presence of a mutation in TMC6 (EVER1) or TMC8 (EVER2), skin papillomaviruses replicate aggressively and cause massive inappropriate outgrowth of skin cells, resulting in treeman syndrome (reviewed in REF.135). Other primary immunodeficiencies may similarly allow specific viruses or microorganisms to replicate unchecked, resulting in distinctive disorders. The effects of immunodeficiencies on the skin virome were mentioned above; also, virome studies of individuals undergoing treatment for X-linked severe combined
immunodeficiency (SCID-X1) revealed distinctive outgrowths of several viral lineages in the gut. An interesting question for future investigation is how much of the phenotypes of diverse human genetic diseases are a result of altered interactions with the normal virome.

**Geography and stochastics of colonization.** Large-scale virome studies have provided evidence that geographic location and stochastics of colonization have strong impacts on human virome variation. A study of human faecal samples from different regions within China reported variation of the phage population structure and found that geography had the strongest impact compared with other variables, including diet, ethnicity and medication. Geography has been associated with the eukaryotic virus populations as well. An early study found geographic variation in eukaryotic viromes in children with diarrhoea from two locations within Australia, with differential prevalence of Adenoviridae and Picornaviridae. A blood virome analysis of eukaryotic viruses in a Chinese population revealed a distinctive pattern for people living in the southern part of China. Moreover, one study of the infant virome found a higher prevalence of viruses infecting human cells in cohorts of people of African descent than in cohorts from the United States. A recent study collected public metagenomic data sets and built a virome database with >30,000 viral genomes, and found higher viral diversity in populations from non-western countries compared with populations from western countries. Similar results were reported in another study. Thus, effects of geography are quite prominent.

**Additional factors.** Additional factors have been tested and reported to influence the human virome. One study tested age using public data, and found that viral diversities in early life and in older individuals (>65 years of age) are lower than those in healthy adults (18–65 years of age), indicating another dimension of age-dependent patterns. Effects of ethnicity and medication were also found in a large Chinese cohort. Cohabitation is also a factor — in an early study, members of the same household shared more similarity in the oral virome compared with those from different households, suggesting transmission of the virome via close contact.

Thus, nascent studies of factors affecting the virome highlight diet, geography, age and health status as major correlates of virome structure. Genetics is less clear, at least in studies so far of healthy twins. More studies of large cohorts with linked comprehensive metadata will be helpful going forward.

**The virome in health and disease**

Viro populations can influence their human hosts in numerous ways. Eukaryotic viruses that infect human cells establish infections, trigger immune responses and, sometimes, cause diseases. Phages can affect the host indirectly via modulating of bacterial composition and bacterial fitness. Some phages and human cell viruses can be integrated into cells of their respective hosts, on occasion transferring new functionality to host cells. Some phages may also interact with human cells directly and trigger immune responses. Recent examples of host–virome interactions in health and disease are discussed below (for reviews, see refs 135,136).

**Interactions between bacteria, phage and their human hosts.** Relatively little is known about the impact of phage predation on human bacterial communities. A window on phage predation and host health is provided by phage therapy, in which phages are deliberately applied to human individuals to treat bacterial infections. As more drug-resistant bacteria are emerging, there is increased interest in this approach (reviewed in ref. 137). Recently, several studies have used phage cocktails to treat bacterial infection in a few individuals, with apparent success (reviewed in ref. 138) motivating larger clinical trials. Evidence that the phages are in fact creating evolutionary pressure on pathogenic bacteria in these studies comes from the observation that bacteria often mutated to become resistant to the therapeutic phage. Phages may even be useful to treat additional diseases. For example, a phage cocktail targeting adherent invasive Escherichia coli strains has been suggested recently as a treatment for Crohn’s disease.

Phages can move DNA between cells and, thereby, introduce new functionality to bacterial genomes, which may modify bacterial fitness and virulence (reviewed in ref. 139). In a recent study using a mouse model, the stool virome was analysed after antibiotic treatment, showing that there was an enrichment of phage-encoded genes for antibiotic resistance, which increased resistance in the bacterial community.

Phage induction can also lead to bacterial cell lysis, regulating bacterial abundance. A recent publication investigated diet-mediated phage induction of human-associated bacterial strains in vitro, revealing that several common food compounds can inhibit bacterial growth by inducing prophages, including artificial sweeteners.

---

**Fig. 5 | Host–virome interactions.** Eukaryotic viruses have both detrimental (red arrow) and beneficial (blue arrow) effects on host health. Phages interact with the host directly or indirectly via the host-associated bacterial community and pose undetermined effects (green arrow) on host health. Data based on refs 139,140,141,142. TLR, Toll-like receptor.
Recent in vitro and animal studies indicated that phages may interact with the host immune system directly. Phages may be taken up by immune cells and trigger immune responses, without the mediation of bacteria, via Toll-like receptor (TLR) signalling. A recent publication reported that phages produced by pathogenic bacteria can be taken up by immune cells (dendritic cells, B cells and monocytes) in mice to induce type I interferon responses via TLR3 signalling [44]. Another study showed that interferon-γ (IFNγ)-producing CD4+ T cells and CD8+ T cells were increased in mucosal sites in germ-free mice fed an E. coli phage isolated from the human gut [44]. Furthermore, this study also revealed that Lactobacillus, Escherichia and Bacteroides phages can stimulate the production of IL-12, IL-6, IL-10 and IFNγ via the nucleotide-sensing receptor TLR9 (REF [44]). Collectively, the interactions among phages, bacteria and the host immune system likely have important roles in host immune homeostasis.

**Human disease-associated virome signatures.** Studies of interactions between the virome with human diseases are just starting. Of course, numerous individual viruses are well known to cause morbidity and mortality. Studies of whole viral populations have now also begun to show patterns associated with disease states. Some recent examples are described below.

The virome has been considered to be a potential trigger of autoimmune diseases. In one study, changes in viral populations were both directly and inversely associated with the development of paediatric type 1 diabetes [66]. Virome signatures have been associated with paediatric and adult inflammatory bowel disease in several studies [15,14], including a reproducible expansion of Caulovirales and a reduction of Microviridae. Whether this is a consequence of altered dysbiotic bacterial populations or is more deeply involved in the disease process remains to be clarified. A metagenomic study showed that children with frequent exposure to enterovirus between 1 and 2 years of age have a higher risk of coeliac disease [66]. Recent studies revealed that the relationships between phages and bacterial populations may influence growth stunting in children [66,14]. Examples of proposed associations between diseases and viral population structure are listed in Table 1.

Studies using animal models have indicated that some eukaryotic viruses may even be beneficial to the host (reviewed in refs [15,14]). For example, persistent infection by a strain of murine norovirus can compensate for the absence of bacteria in gnotobiotic mice, allowing restoration of intestinal morphology and promoting lymphocyte differentiation [140]. Murine astrovirus protected immunodeficient mice from enteric norovirus infections through the induction of type III interferons in the intestinal epithelial barrier [146]. Depletion of murine gut viruses using an antiviral cocktail inhibited the development of the intestinal intraepithelial lymphocytes, at least in part [63]. So far, all of these studies showing positive effects were performed in model organisms — it will be of interest to determine how much beneficial immune instruction may be directed by the virome in humans.

Thus, both phages and eukaryotic viruses can promote host health through interactions with the host immune system. However, virome studies using human cohorts suggest that dysbiosis of the virome is associated with multiple diseases. These studies emphasize the balance between beneficial and harmful roles of viral populations in humans and other organisms.

**Conclusions and perspectives**

The virome field has come a long way since the first paper in 2002 that reported metagenomic sequencing of a viral specimen. The methods for study have advanced, although there are still many challenges associated with working with metagenomic dark matter. Human virome population diversity and composition is being documented for many body sites — one consistent conclusion is that each individual harbours diverse and distinctive viral communities. Future studies are needed to clarify the DNA and RNA viromes at different anatomical sites and to link the alterations of viral composition to specific diseases. We now have a detailed picture of the stepwise nature of the assembly of the human virome after birth. The impact of viral colonization during early life on long-term health outcomes is still unknown and warrants careful study. Breastfeeding appears to have an important role during viral colonization; how the antiviral components in breast milk interact with the different components of the virome warrants further investigation. Emerging data indicate that factors that influence the human microbiome also often influence the virome as well, so that sorting out the influence of each will be a challenge going forward. Resident viruses are not only actively interacting with other microorganisms but also with the mammalian immune system. Many intriguing conclusions have so far only been obtained in animal studies or in vitro experiments, focusing attention on translation to studies in humans. Associations between alterations of virome and disease states are being identified more commonly, but in many cases causality and molecular mechanisms remain to be worked out. The vast world of the human virome is beginning to be understood, laying the ground work for numerous future studies of its importance.

Published online 30 March 2021

---

1. Breitbart, M. et al. Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220–6225 (2003).
2. Zhang, T. et al. RNA viral community in human feces: prevalence of a pathogenic viruses. *PLoS Biol.* **4**, 0108–0118 (2006).
3. Reyes, A. et al. Viruses in the faecal microbiota of mononuclotic twins and their mothers. *Nature* **466**, 334–338 (2010).
4. Minot, S. et al. Rapid evolution of the human gut virome. *Proc. Natl Acad. Sci. USA* **110**, 12450–12455 (2013).
5. Minot, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Conservation of gene cassettes among diverse viruses of the human gut. *PLoS ONE* **7**, e42542 (2012).
6. Minot, S. et al. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625 (2011).
7. Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Hypervariable loci in the human gut virome. *Proc. Natl Acad. Sci. USA* **109**, 3962–3966 (2012).
8. Breitbart, M. et al. Genomic analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA* **99**, 14250–14255 (2002).
9. Aggarwala, V., Liang, G. & Bushman, F. D. Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mob. DNA* **8**, 12 (2017).
10. Shkoporov, A. N. & Hill, C. Bacteriophages of the human gut: the ‘known unknown’ of the microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
92. Arora, N., Sadovsky, Y., Dermody, T. S. & Coyne, C. B. Microbial vertical transmission during human pregnancy. Cell Host Microbe 21, 561–567 (2017).

93. Breitbart, M. et al. Viral diversity and dynamics in an infant gut. Res. Microbiol. 159, 567–3 (2008).

94. Maqbool, R. et al. Discordant transmission of bacterial and viral microbes to mothers at birth. Microbiome 7, 156 (2019).

95. Sausset, R., Petit, M. A., Gaboriau, V. & De Paepe, M. New insights into intestinal phages. Mucosal Immunol. 13, 205–210 (2020).

96. Nanda, A. M., Thornoman, K. & Frunke, J. Impact of spontaneous prophage induction on the fitness of bacterial populations and host–microbe interactions. J. Bacteriol. 197, 410–419 (2015).

97. Cortes, M. G., Krog, J. & Balasoiu, G. Optimality of the spontaneous induction process rate. J. Theor. Biol. 485, 110030 (2020).

100. Judd, R. et al. Modulation of enterohaemorrhagic Escherichia coli survival and virulence in the human gastrointestinal tract. Microorganisms 6, 115 (2018).

104. De Paepe, M. et al. Carriage of L. lactis is costly for its bacterial host due to frequent reactivation in mixed gastrointestinal infections. PLoS Genet. 12, e1005861 (2016).

105. Reyes, A. et al. Gut DNA viromes of Malasian twins discordant for severe acute malnutrition. Proc. Natl Acad. Sci. USA 112, 11941–11946 (2015).

109. McCann, A. et al. Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. PeerJ 12698 (2018).

110. Black, M., Bhattacharya, S., Philip, S., Norman, J. E. & McLernon, D. J. Planned cesarean delivery at birth affects the diversity and colonization pattern of the gut microbiome. Acta Paediatr. 116, 2346–2351 (2017).

114. Goodrich, J. K. et al. Human genetics shape the gut microbiome. Cell 159, 789–799 (2014).

119. Xu, H. et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. mSphere 5, e00284–e00219 (2020).

121. Rothchild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. Nature 551, 255–259 (2017).

130. Moreno-Gallego, J. L. et al. Virome diversity correlates with intestinal microbiome diversity in adult monozygotic twins. Cell Host Microbe 25, 261–272 e5 (2019).

144. Kortright, K. E., Chan, B. K., Koff, J. L. & Turner, P. E. Human–gut–DNA virome variations and pharmacology. Molecules 21, 561–567 (2016).

147. Salter, S. J. et al. Reagent and laboratory sequencing data. Bioinformatics 33, e379–e387 (2017).

163. Koff, J. L. et al. The eukaryotic gut virome in hematopoietic stem cell transplantation: new clues in enteric graft-versus-host disease. Nat. Med. 23, 1080–1085 (2017).

169. Liu, L. et al. Commensal viruses maintain intestinal intraepithelial lymphocytes via noncanonical RIG-I signaling. Nat. Immunol. 20, 1681–1691 (2019).

177. Pérez-Brocal, V. et al. Metagenomic analysis of Crohn’s disease patients identifies changes in the virome and microbiome related to disease status and therapy, and detects potential intestinal markers. Inflamm. Bowel Dis. 21, 2515–2532 (2015).

183. Ma, Y., Xou, Y., Mai, G., Tokuyasu, T. & Liu, C. A human gut phage catalog correlates with the microbiome with type 2 diabetes. Microbiome 24, 1 (2016).

185. Ungaro, F. et al. Metagenomic analysis of intestinal mucosa revealed a specific species–virome signature in early-diagnosed inflammatory bowel disease. Gut Microbes 10, 149–158 (2019).

191. Leppoff, J. et al. The eukaryotic gut virome in genetically at risk children: the TEDDY study. Nature 578, 452–456 (2020).

197. Wang, Y., Hammes, F. Duggelin, M. & Egli, T. Influence of size, shape, and oncolytic viral properties on bacterial passage through micropore membrane filters. Environ. Sci. Technol. 42, 6749–6754 (2008).

201. Weil, A. A., Becker, R. L. & Harris, J. B. Vibrio cholerae at the intersection of immunity and the microbiome mSphere 4, e00597-19 (2019).

202. Ryan, M. P. & Pembroke, J. T. Brevundimonas spp: emerging global opportunistic pathogens. Virulence 9, 480–495 (2018).

203. Roux, S. et al. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. PeerJ 2016, e2777 (2017).

207. Kim, K. H. & Bae, J. Amplification methods bias metagenomic libraries in single-stranded and double-stranded DNA viruses. Appl. Environ. Microbiol. 77, 7665–7668 (2011).

215. Krishnamurthy, S. R., Janowski, A. B., Zhao, G., Barouch, D. W. & Huang, J. Bacteriophage defence. PLoS Biol. 14, e1002409 (2016).

216. Callanan, J. et al. Expansion of known ssDNA phage genomes: from tens to over a thousand. Sci. Adv. 6, eaay5981 (2020).

217. Salter, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 12, 87 (2014).

218. Kim, D. et al. Optimizing methods and dodging pitfalls in microbiome research. Microbiome 5, 1–14 (2017).

220. Zaff, M. et al. Detecting contamination in viromes using Viromet. Nat. Biotechnol. 37, 1408–1412 (2019).

223. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, CD-HIT. accelerated clustering of large metagenomic and viromic sequencing data. Bioinformatics 28, 3150–3152 (2012).

224. Taylor, L. J., Abbas, A. & Bushman, F. D. Grabseq: simple downloading of reads and metadata from multiple next-generation sequencing data repositories Biostatistics 21, 3365–3369 (2020).

225. Roux, S., Enault, F., Hurwitz, L. & Sullivan, M. B. ViSorter: mining viral virome from microbial genomic data. PeerJ 3, e985 (2015).
172. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).

173. Tithi, S. S., Aylward, F. O., Jensen, R. V. & Zhang, L. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ* **2018**, e4227 (2018).

174. Roux, S., Tournaye, J., Mahul, A., Debrosa, D. & Enault, F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**, 76 (2014).

175. Jurtz, V. I., Villarroel, J., Lund, O., Voldby Larsen, M. & Nielsen, M. MetaPhinder — identifying bacteriophage sequences in metagenomic data sets. *PLoS ONE* **11**, e0163111 (2016).

176. Rampelli, S. et al. ViromeScan: a new tool for metagenomic viral community profiling. *BMC Genomics* **17**, 165 (2016).

177. Bolduc, B. et al. vConFACT: an iVirus tool to classify double-stranded DNA viruses that infect archaea and bacteria. *PeerJ* **2017**, e3243 (2017).

178. Hatcher, E. L. et al. Virus variation resource-improved response to emergent viral outbreaks. *Nucleic Acids Res.* **45**, D482–D490 (2017).

179. Bateman, A. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

180. Graziotin, A. L., Koxnin, E. V. & Kristensen, D. M. Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* **45**, D491–D498 (2017).

181. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

182. Skewes-Cox, P., Sharpton, T. J., Pollard, K. S. & DeRisi, J. L. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS ONE* **9**, e105067 (2014).

183. Clarke, E. L. et al. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. *Microbiome* **7**, 46 (2019).

184. Skewes-Cox, P. J., Sharpton, T. J., Pollard, K. S. & DeRisi, J. L. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS ONE* **9**, e105067 (2014).

185. Arndt, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).

Acknowledgements
The authors thank members of the Bushman laboratory for help and suggestions, and L. Zimmerman for artwork designs. This work was supported by the National Institutes of Health (NIH) (grants R61-HL137063, R01-HL115252), the Penn Center for AIDS Research (P50 AI 045008), the PennCHOP Microbiome Program and a Tobacco Formula grant under the Commonwealth Universal Research Enhancement (CURE) programme (grant number SAP # 4100068710), and the Crohn’s and Colitis Foundation.

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
The authors declare no competing interests.

Peer review information
*Nature Reviews Microbiology* thanks F. Maggi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher’s note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2021