Phylogeny and disease links of a widespread and ancient gut phage lineage

Patrick A. de Jonge[1,2,3], Bert-Jan H. van den Born[1,2,3], Aeilko H. Zwinderman[4], Max Nieuwdorp[1], Bas E. Dutilh[5,6†], Hilde Herrema[1,2,3†]*

1: Department of Internal and Experimental Vascular Medicine; Amsterdam UMC; Location AMC; University of Amsterdam; Amsterdam; the Netherlands.

2: Amsterdam Gastroenterology, Endocrinology & Metabolism; Endocrinology, metabolism & nutrition; Amsterdam UMC; Amsterdam; the Netherlands.

3: Amsterdam Cardiovascular Sciences; Diabetes & Metabolism; Amsterdam UMC; Amsterdam; the Netherlands.

4: Department of Clinical Epidemiology; Biostatistics and Bioinformatics; Amsterdam UMC; Location AMC; University of Amsterdam; Amsterdam; the Netherlands.

5: Theoretical Biology and Bioinformatics; Science for Life; Utrecht University; Utrecht; the Netherlands.

6: Institute of Biodiversity; Faculty of Biological Sciences; Cluster of Excellence Balance of the Microverse; Friedrich-Schiller-University Jena; Jena; Germany.

†: These authors contributed equally.

*: Correspondence: h.j.herrema@amsterdamumc.nl

Abstract

Viruses are a core component of the human microbiome, impacting health and disease through interactions with gut bacteria[1] and the immune system[2]. Most viruses in the human microbiome are bacteriophages, which exclusively infect bacteria. Individual gut bacteriophages can affect bacterial bile acid deconjugation[3], and can alter their infection strategy based on dietary content[4]. Up to recently, most studies of the gut virome have focused on low taxonomic scales (e.g., viral operational taxonomic units), hampering population-level analyses. We previously identified the expansive and widespread Candidatus *Heliusviridae* bacteriophage family in a cohort with inhabitants of Amsterdam,

the Netherlands. Here, we study their biodiversity and evolution in a wide variety of human populations. With a more detailed phylogeny based on sequences from six viral genome databases, we now propose the Candidatus order *Heliusvirales* to accommodate these viruses. We identify *Ca. Heliusvirales* viruses in 80% of 5,441 individuals across 39 studies, and also in thirteen out of sixteen analyzed metagenomes from ancient humans that lived in Europe and North America between 1,000 and 5,000 years ago. We show that a large *Ca. Heliusvirales* lineage has diversified starting with human dispersion across the globe some 50,000 years ago. Ancient peoples and modern hunter-gatherers further have distinct *Ca. Heliusvirales* populations that are characterized by lower richness than modern urbanized people. Within urbanized people, those suffering from type 1 and type 2 diabetes, as well as inflammatory bowel disease, have higher *Ca. Heliusvirales* richness than healthy controls. We thus conclude that these ancient core members of the human gut virome have thrived with increasingly westernized lifestyles of the human population.

Introduction

Bacteriophages are the second most common biological entities in the human gut microbiome, after their bacterial targets[5]. As obligate parasites of bacteria, phages play a key role in regulating the structure and function of the gut microbiome in health and disease[6]. For example, phage community alterations have been linked to diseases such as inflammatory bowel disease[1,7], metabolic syndrome[8], and malnutrition[9]. Individual human gut phages have further been found to mediate diet-induced bacterial lysis[4], bacterial bile acid metabolism[3], and gut colonization[10]. But while phages are evidently crucial to the functioning of the gut microbiome, their study at higher taxonomic levels was long restricted to morphology- rather than phylogeny-based taxonomy.

This knowledge gap is partly due to the absence of a universal viral marker gene akin to ribosomal RNA in cellular organisms. Bacteriophage taxonomy was originally based on nucleic acid composition (with the majority being dsDNA viruses) and viral particle

morphology[11], with the bulk of known bacteriophage diversity being classified as *Siphoviridae, Myoviridae,* or *Podoviridae*. These families, as well as the order *Caudovirales*, proved to be so genomically diverse that they were recently abolished and reclassified in the class *Caudoviricetes*[12,13]. A striking example of the discovery of higher taxonomic lineage was the highly prevalent and abundant human gut phage crAssphage[14] and its relatives[15] in metagenomics data. These viruses are now accepted as the phylogenetics-based *Crassvirales* order by the International Committee for the Taxonomy of Viruses (ICTV). But while the description of the *Crassvirales* is a step forward in the taxonomy of the human gut virome, many other important lineages remain unclassified.

We recently identified the putative phage family *Ca. Heliusviridae* in a study of gut virome perturbations associated with metabolic syndrome, a collection of risk factors for cardiometabolic disease[8]. The expansive family was detected in over 90% of almost 200 study participants and may thus be part of a persistent core of human gut bacteriophage lineages[16,17]. But as our previous study focused on a single cohort, it remains unknown how widespread these phages are in the general population, as well as in other environments. Furthermore, their relation to human health beyond metabolic syndrome is yet to be elucidated. Their further adoption in virome studies necessitates a robust taxonomic classification based on comprehensive phylogenetic analysis of sequences from diverse sources[18–21].

Here, we present a comprehensive study of *Ca. Heliusviridae* proteomic and phylogenetic diversity. Comparisons of member phages derived from several large metagenomic databases show both the distinctness of this family from known ICTV taxonomic lineages, and their phylogenetic structure. We further provide evidence that increasing ecological richness of this phage family in the gut is linked to both urbanized lifestyles and several (cardiometabolic) diseases.

Results

Helius-phages are ubiquitous in bacteriophage databases

We previously reported an expansive and widespread clade of human gut phages, tentatively christened as the family *Ca. Heliusviridae*, after the HELIUS (Healthy Life In an Urban Setting) cohort that we studied[8]. These phages shared nine marker genes: four structural proteins (terminase large subunit or TerL, portal protein, major capsid protein, and a head maturation protease), three replication-related ones (DNA polymerase I, helicase, and nuclease), and two proteins of unknown function (**Figure 1a**)[8]. To study these phages more extensively, we analyzed 1,314,988 contigs of over 25 kbp from seven phage databases with profile hidden Markov models (HMMs) of all nine marker genes (**Figure 1b**). Of the 111,534 contigs with at least one homolog to a marker gene, 2,322 were deemed complete by checkv[22]. Notably, the 1,007 complete genomes with hits to all marker genes distinctly had bit-scores of over 600 in homology searches against the terminase large subunit (TerL) marker gene, while this threshold was seldomly passed by genomes carrying fewer marker genes (**Figure 1c**). The TerL, which encodes an ATP-driven molecular motor involved in genome translocation into the viral capsid, is often highly conserved among *Caudoviricetes* phages and consequently frequently used as marker gene[13,15]. We thus selected the 1,253 complete viral genomes with bit-scores of over 600 against the TerL HMM for further study. Phage-specific gene annotation of all 1,253 genomes with pharokka[23] refined the function of several marker genes: the DNA polymerase I was revealed to instead be an exonuclease, while the ones of unknown function encode a DNA polymerase and ssDNA-binding protein (**Figure 1a**). We next determined which genes were most common among these phages by forming clusters of homologous proteins (i.e., protein clusters). The only universal protein cluster was the TerL, and the next eight most common protein clusters represented the other marker genes (**Figure 1d**, **Supplementary Table 1**). Genes with hits against the HMMs of the major capsid protein, head maturation protease, helicase, and DNA polymerase genes were spread out over multiple protein clusters, indicating some heterogeneity in their sequences. Overall, all nine marker genes were found in 1000-1253 of

the genomes (**Supplementary Table 2**). Thus, the nine originally identified marker genes are extremely common, though not universal, among *Ca. Heliusviridae*.

Helius-phages are taxonomically and ecologically distinct

We next sought to establish more detailed phylogenies of these phages. We previously divided the *Ca. Heliusviridae* into three groups: *alpha, beta,* and *gamma*. In an approximate maximum-likelihood tree of TerL genes, we again discerned three large well-supported clades (Shimodaira–Hasegawa-like approximate likelihood ratio test[24] ≥80 and ultrafast bootstrap approximation[25] ≥95, **Figure 2a**), but these did not completely align with the previous putative groups. The largest clade contained all but one of the partial genomes from the previously proposed *gamma*-group, but groups *alpha* and *beta* were spread out across the other two clades. This reflects the two main clades in group *beta* that were evident in the earlier analysis[8], while their different architecture in the TerL-based tree versus the earlier concatenated tree likely reflects horizontal gene transfer of at least some of the core genes among these groups. Furthermore, large subclades in all three clades contained no representative of the earlier groups, and neither did a number of smaller clades across the tree which were not part of the three main clades. In the face of this expanded biodiversity, we set out to draw more detailed taxonomic boundaries.

The TerL tree also made it evident that these phages are better classified at the order-level. The reasoning behind this is as follows: the three largest clades were divisible into smaller clades that at minimum shared 5-57% of their protein content, too low for them to be considered different viral genera (**Figure 2b, Supplementary Figure 1**)[26]. In line with the viral taxonomic hierarchy[27], we classify these clades as subfamilies. The higher-level clades are then necessarily families, and the entirety of the tree an order. Consequently, we renamed the Ca. *Heliusviridae* as the Candidatus viral order *Heliusvirales.* Because Helios was the ancient Greek solar deity, we named families after various solar deities. The three largest families are the *Ca. Utuviridae* (Sumerian), *Ca. Aineviridae* (Irish), and *Ca. Hathorviridae* (Egyptian). Several smaller well-supported and high-level clades were also

placed at the family rank due to their very low (<15%) shared protein content (**Supplementary Figure 1)**. The subfamilies within the three largest families, and three small high-level clades classified as subfamilies, all shared at least 30% of their protein content with each other (**Figure 2b, Supplementary Figure 1)**. All subfamilies were named after words for sun in languages from around the world (**Figure 2a, Supplementary Table 3**).

In order for a group of phages to be deemed a defined lineage, they need to be a monophyletic group that is distinct from established taxonomy[26]. To establish this, we performed proteomic-based analyses on the 1,237 complete genomes versus the national center for biotechnological information (NCBI) reference sequence database[28]. In a genomic protein-sharing network, our genomes formed a cluster separate from RefSeq phages (**Supplementary Figure 2a**). This was confirmed in a phage proteomic tree, were they also formed a separate monophyletic clade distinct from ICTV-recognized and RefSeq sequences (**Supplementary Figure 2b**). Together, these results show that the *Ca. Heliusviriales* are a cohesive lineage that is unrelated to any of the viral families that are currently accepted by the ICTV.

A genomic network of the *Ca. Heliusvirales* showed that several subfamilies formed distinct clusters, notably the second- and third-largest subfamilies: the *Ca. Mataharivirinae* (*Ca. Hathorviridae*) and *Ca. Mehrvirinae* (*Ca. Aineviridae)* (**Figure 2c**). Their distinctness was reflected by the environments in which they were found and the bacteria that they infect (**Figure 3a and b**), as well as their divergent GC content when compared to other members of their family (**Supplementary Figure 3**). Although animal-, terrestrial-, and aquatic-derived phages dominate some *Ca. Heliusvirales* lineages, the *Lachnospiraceae* phages in the *Ca. Mataharivirinae* were enriched in human gut-derived phages (hypergeometric test, $p = 2.8\text{x}10^{-21}$). Within this family, these phages seem to have a broad host range, as on average they had CRISPR-spacer hits (0-1 mismatch) against 3.18±2.75 genera (median 2), with the most common host predictions being *Acetatifactor sp900066365, Agathobacter faecis,* and *Blautia_A wexlerae_A* (**Supplementary Figure 4**). The *Ca. Mehrvirinae* were similarly enriched in phages from the human oral cavity (hypergeometric test, $p = 1.5 \text{ x } 10^{-17}$). This

environmental preference was also reflected in their most common hosts: the facultatively anaerobic *Streptococcaceae*. These two subfamilies are thus specialized for distinct environments. Similarly, the *Ca. Nitavirinae* formed a separate cluster in the genomic network, reflecting their hosts in the *Actinobacteriota* rather than the *Firmicutes* infected by other *Ca. Heliusvirales*. So too did the distinctness of the *Ca. Zonviridae* reflect their hosts in the *Negativicutes*. Unlike the recently defined *Crassvirales*, which all infect within the *Bacteroidetes* phylum[29], the *Ca. Heliusvirales* as a collective thus have a very broad host range across multiple phyla.

The majority of complete *Ca. Heliusvirales* genomes (56.9%) carried identifiable genes involved with integration and excision, meaning they were temperate phages. This also differed among the subfamilies, with notably higher amounts among the *Ca. Ra-* and *Matahariviriinae* in the *Ca. Hathorviridae* (96.7% and 82% of genomes), as well as the *Ca. Mehrvirinae* (85% of genomes, **Supplementary Figure 5**). Considerably lower fractions of temperate phages were present among the subfamilies of the *Ca. Utuviridae* (overall 47.6%). It is thus possible that besides contrasting environments and host species, the different *Ca. Heliusvirales* lineages also differ in lifestyles.

Putative auxiliary metabolic genes were very common among *Ca. Heliusvirales*, being present in 78.1% of genomes (**Figure 3c**), though we found no clear family- or subfamily specific auxiliary metabolic genes. Many were part of cysteine and methionine metabolism, with the most common genes being DNA (cytosine-5)-methyltransferase 1 (K00558, present in 67% of genomes) and S-adenosylmethionine synthetase (K00789, present in 58.9% of genomes). Both of these genes likely aid the phage in defending against bacterial restriction-modification defenses[30,31]. After cysteine and methionine metabolism, sulfur metabolism-associated genes were most commonly found. This is in line with previous findings that bacteriophages can use sulfur metabolism to reprogram bacterial metabolism for additional energy production during phage particle formation[32].

Helius-phages have been associated with humans since ancient times

Across our datasets, we found *Ca. Heliusvirales* TerL genes from every continent (**Figure 4a**), indicating that they are widespread. Only two countries had human gut-associated samples without any detected *Ca. Heliusvirales*: El Salvador and Kazakhstan. Indeed, analysis of 7,166 human gut metagenomes of 5,441 individuals from 38 studies[33] found *Ca. Heliusvirales* TerL genes in 4,491 individuals (80%, **Figure 4b**). Outside the gut, prevalence of *Ca. Heliusvirales* was markedly lower: we detected their presence in 18/54 skin samples, 17/61 vaginal samples, and 115/312 oral samples. While the *Ca. Heliusvirales* thus occur in various body sites, they are especially prevalent in the gut. This is in line with the finding that their most common hosts are in the *Firmicutes* (e.g., the *Lachnospiraceae*), which is one of the main bacterial phyla in the human gut.

We hypothesized that high *Ca. Heliusvirales* prevalence reflected its ancient association to the human lineage. To explore this, we re-analyzed whole metagenome shotgun sequencing datasets from paleofeces in Austria[34] and North America[35], and intestinal biopsies of Ötzi the iceman[36]. Among the North-American samples, we found 36 *Ca. Heliusvirales* TerL genes across all eight of the studied samples (**Figure 4a**). These samples were from three different locations across the USA and Mexico, and were dated between 10-920 CE. We furthermore identified four *Ca. Heliusvirales* TerL genes in small- and large-intestinal samples of Ötzi the iceman, who lived around 3,200 BCE[37]. In the Austrian samples, we however found only four TerL sequences in one of the four Austrian samples, which was dated at between 652-544 BCE. Since all other samples from both North America and Ötzi the iceman did contain *Ca. Heliusvirales*, their absence from 3/4 ancient Austrian samples may reflect higher levels of degradation or fragmentation of these samples rather than their true absence among these individuals.

The presence of *Ca. Heliusvirales* in multiple pre-colonial North Americans imply that they were part of the human gut microbiome before human migration to the Americas (about 15,000 years ago[38]). We confirmed this hypothesis with a time-measured phylogenetic tree of the *Ca. Hathorviridae*, which we chose for this analysis because it contains the large and highly human-gut-associated *Ca. Mataharivirinae*. A tree with ancient North-American

sequences and a selection of modern sequences that reflect *Ca. Hathorviridae* tree topology (**Supplementary Figure 6**) showed that these phages began diversifying about 65,000 years ago, with a 95% highest posterior density (HPD) interval of 44,000-87,000 years ago (**Figure 4c**). This is notably around the time that modern humans started to disperse worldwide from Africa (40,000-60,000 years ago)[39]. Further diversification of the *Ca. Hathorviridae* into its two largest subfamilies (*Ca. Mataharivirinae* and *Ca. Ravirinae*) occurred about 34,000 years ago (95% HDP 24,000-45,000 years ago), indicating that these phages adapted to specific human populations. Indeed, a concatenated tree of all universal genes in the complete *Ca. Mataharivirinae*, showed notable clustering of sequences from China (**Supplementary Figure 7**).

Ancient gut microbiomes have similar diversity to those of modern populations consuming non-urbanized diets[35], which are more diverse than those consuming urbanized diets[40,41]. By extension, this is likely true for the gut virome as well. To further explore the relation between ancient and modern *Ca. Heliusvirales*, we compared them to samples derived from urbanized and non-urbanized populations. To minimize interstudy batch-effects, we selected two studies that included both hunter-gatherers (from Tanzania or Peru) and urbanized people (from Italy or the USA)[40,41]. Within both studies, we observed a significantly higher *Ca. Heliusvirales* richness among urbanized people (Wilcoxon signed-rank test, Benjamini-Hochberg-adjusted p=0.026 and 0.01, **Supplementary Figure 8a**). Combining the values of modern urbanized and hunter-gatherer populations with ancient ones showed significantly higher *Ca. Heliusvirales* richness in urbanized people than both hunter-gatherers (linear mixed-effect model, Benjamini-Hochberg-adjusted $p=4.5\times10^{-5}$) and ancients (linear mixed-effect model, Benjamini-Hochberg-adjusted $p=1.2\times10^{-5}$), but not between hunter-gatherers and ancients (linear mixed-effect model, Benjamini-Hochberg-adjusted p=0.076, **Figure 4d**). This interestingly is the reverse of overall gut bacterial richness, which decreases with urbanization[42]. Abundance, measured as the fraction of reads that mapped to *Ca. Heliusvirales* TerL genes, was not significantly different between

the three populations (**Supplementary Figure 8b**). *Ca. Heliusvirales* thus seem to diversify in tandem with urbanization.

We next analyzed *Ca. Heliusvirales* β-diversity among urbanized, hunter-gatherer, and ancient samples with a phylogenetic tree of TerL genes from these samples and those from complete genomes as depicted in Figure 2a. Non-metric multidimensional scales (NMDS) on unweighted UniFrac distances intriguingly showed that all three populations (urban, hunter-getherers, ancients) had *Ca. Heliusvirales* populations that were distinct from each other permanova q=0.001, **Figure 4e**). Among ancient samples, the *Ca. Xiheviridae*, *Ca. Mithraviridae, Ca. Tilevirinae*, and *Ca. Ghrianvirinae* were all more prevalent than in both hunter-gatherers and urban populations (Fisher's exact test, Benjamini-Hochberg-adjusted $p < 0.05$, **Figure 4f**). Meanwhile, the *Ca. Sontsevirinae* and *Ca. Dzuwavirinae* were more prevalent, and the *Ca. Kuarahyvirinae* and *Ca. Nayviriane* were less prevalent among modern urban populations. With the exception of the *Ca. Ghrianvirinae*, which target *Bacteroidacaeae*, these significantly altered lineages notably did not seem to infect aberrant hosts than the other subfamilies. These results are, however, at least partially caused by the greater sequencing depth of the ancient samples. In fact, taxonomic richness expressed as the number of observed *Heliusvirales* subfamilies per billion sequenced bases, showed the same pattern as overall richness (**Supplementary figure 8c**). Yet despite a higher sequencing depth, ancient samples were notably mostly lacking the *Ca. Matahariviriinae*, which is the largest of the subfamilies. This intriguingly suggests that this subfamily in particular is associated with modern and urbanized human gut microbiomes. Our results could reflect that *Ca. Heliusvirales* hosts in the *Lachnospiraceae* thrive in the urbanized gut microbiome, or, since the samples we analyzed are not virome-specific, that these phages have increased rates of lysogenation.

Helius-phage richness is associated with disease

Our previous research focused on gut virome alterations in metabolic syndrome (MetS). To study a wider range of illnesses, we now identified *Ca. Heliusvirales* sequences in

assemblies of fourteen studies of diverse human-derived samples[33] by searching for the TerL marker gene. Within-study analyses of richness alterations identified significantly altered *Ca. Heliusvirales* richness in four out of twelve illnesses (Wilcoxon signed-rank test Benjamini-Hochberg adjusted $p < 0.05$): type 1 diabetes (T1D), type 2 diabetes (T2D), inflammatory bowel disease (IBD), and liver cirrhosis (**Figure 5a).** In the former three, richness was elevated when compared to healthy controls, while in liver cirrhosis it was decreased. These disease linkages are interesting in light of the *Lachnospiraceae* hosts of many *Ca. Heliusvirales*, because this bacterial family includes species involved in short-chained fatty acid production (e.g., *Roseburia intestinalis, Anaerobutyricum hallii, Coprococcus eutactus,* and *Blautia wexlerae* **Supplementary Figure 4**)[43], which are often associated with the healthy gut. Indeed, increased *Lachnospiraceae* abundance in T1D-[44] and T2D-patients[45] has been reported, though a study of in vitro incubations in a relatively small cohort size (n=26) linked IBD to decreased *Lachnospiraceae* abundance[46]. For the former two diseases, this could explain the greater richness and prevalence in *Ca. Heliusvirales* phages among such patients. The alterations in the *Ca. Heliusvirales* order are notably counter to ASV-level analysis of the entire gut virome, which was found to be unchanged in T1D[47], T2D[48], and IBD[1]. This underscores the value of analyzing higher taxonomic ranks in the context of the gut virome.

Besides differences in *Ca. Heliusvirales* richness, the populations of these phages also differed in each of the four studies (unweighted UniFrac, PERMANOVA Benjamini-Hochberg adjusted $p < 0.05$, **Figure 5b**). These population-wide differences between healthy controls and people with T1D, T2D, IBD, and cirrhosis expressed themselves in distinct patterns of prevalence at the subfamily level (**Figure 5c**). The *Ca. Uternevirinae* in particular were significantly more prevalent among T1D patients than healthy controls, but more prevalent in controls than IBD and cirrhosis patients (Fisher's exact test, Benjamini-Hochberg adjusted $p < 0.05$). Additionally, the *Ca. Nayvirinae* were more prevalent in controls in the IBD cohort, but less so in the T2D cohort, while the *Ca. Oorunvirinae* were more prevalent in controls in cirrhosis and less so in T2D. These subfamilies are all part of

the *Ca. Utuviridae*, consistent with our previous linkages of viruses in this family to metabolic syndrome[8]. Thus, while some of these results could be explained by technical differences (e.g., different sample sizes between controls and T1D patients), the findings across five different cohorts linking members of the *Ca. Utuviridae* to different cardiometabolic diseases robustly show they are associated with a wide range of disease states. The *Ca. Nayvirinae* were notably less likely temperate than other *Ca. Heliusvirales* (31% temperate phage versus 56.9% overall, **Supplementary Figure 5**), as where the *Ca. Atenviridae*. Their association with healthy over IBD patients is thus in line with previous findings that this disease is correlated with a loss of lytic phages[1]. The *Ca. Nayvirinae* were also depleted in urbanized populations compared to hunter-gatherer and ancient ones (**Figure 4f**), perhaps indicating that they are the viral component of VANISH (volatile and/or associated negatively with industrialized societies of humans) gut microbiome taxa[49]. Conversely, its positive associations with T1D (non-significant) and T2D (significant) points toward a role in disease, underscoring the complex interrelations between the human gut virome and health.

Discussion

Elucidating higher-order structures among gut phages is essential for gaining a deeper understanding of the viromes in the human gut and beyond. This process is underway with the re-evaluation of viral taxonomy along genomic lines[26]. Recently, the prevalent *Crassvirales* phage order has undergone the arduous process of classification[29,50], and here we add the *Ca. Heliusvirales* order as a second genome-based classification of phages widespread in the human gut. For this, we used a combination of marker genes and proteomics-based approaches, though our phylogeny of the TerL gene was the main driver for taxonomic classification. As this gene is well-conserved among the *Duplodnaviria*, to which dsDNA bacteriophages belong[51], we considered the phylogeny built with this gene is the most reliable method for defining *Ca. Heliusvirales* taxonomy. Our usage of the TerL gene to define taxonomy is akin to the *Crassvirales*[12], but unlike the concatenated phylogeny

of marker genes used to define the *Herelleviridae*[13]. Naturally, the absence of any other universal marker gene beside the TerL gene among the *Ca. Heliusvirales* invalidates a concatenated phylogeny approach. This absence of universal marker genes also indicates that relatively large amounts of horizontal gene transfer events occur within the *Ca. Heliusvirales*, meaning that they may be interesting targets to study evolutionary dynamics of human gut viruses.

The *Ca. Heliusvirales* order is highly diverse in environmental origin, geographic distribution, and bacterial hosts. While we found *Ca. Heliusvirales* phages in terrestrial, marine, and gut microbiomes of both cattle and wild animals, the majority of genomes derived from human gut samples. While at face-value this may be assigned to anthropocentric sampling bias, the IMGVR database included in our analysis contains about six times more aquatic and terrestrial phages than human-derived ones[52]. Furthermore, non-animal-associated *Ca. Heliusvirales* phages are largely consigned to specific families, indicating environmental differentiation.

Ca. Heliusvirales phages associated with the human gut were identified in countries from every permanently inhabited continent, and in about 80% of the population. This percentage could well be an underestimation, as shallower sequencing depths may lead to false negatives. In either case, this order is clearly a core component of the human gut virome. This has been the case since the deep past of modern humans, as evidenced by the presence of these phages in palaeofaeces and our accompanying evolutionary studies. We dated the diversification of the *Ca. Hathorviridae* family roughly to the dispersal of human populations out of Africa 60,000 years ago[39]. This dating is notably more recent than the previously studied diversification of the human-associated bacterial species *Methanobrevibacter smithii*[35] and *M.oralis*[53] at respectively 85,000 and 126,000 years ago. During their rapid evolution, these phages perhaps diversified in response to microbiome alterations influenced by changing human geography and diet, both main drivers of human microbiome composition[54,55]. Earlier studies showed co-linearity of *Crassvirales* between humans and great apes[56], and estimated a diversification of these phages in the past several

centuries[57]. This estimate of much more recent diversification was, however, noted for its uncertainty and focused only on the *crAssphage* species rather than the family-level analysis presented here.

We also found that higher *Ca. Heliusvirales* richness is associated with increased urbanization. This is remarkable, as microbiome diversity as a whole is lower with increasing urbanization[40,41]. The opposite observations in *Ca. Heliusvirales* could reflect the success of their bacterial hosts to replicate in urbanized microbiomes: indeed, Hadza hunter-gatherers from Tanzania and Matses hunter-gatherers from Peru have lower levels of *Lachnospiraceae* bacteria such as *Blautia* and *Ruminococcus* than people in urbanized environments[41,58].

While the *Ca. Heliusvirales* as a whole are near universally present in the human gut, no single viral genome in this putative order is. The linkages to lifestyle and disease demonstrated here are thus only evident when studying higher viral taxonomic levels. In line with the recent reassessment of viral taxonomy along genomic lines[12], this study thus shows how future definition of additional viral orders and classes can provide key insights into the ecology and disease-associations of the gut virome.

References

1. Clooney, A. G. *et al.* Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host Microbe* **26**, 764-778.e5 (2019).
2. Fluckiger, A. *et al.* Cross-reactivity between tumor MHC class I–restricted antigens and an enterococcal bacteriophage. *Science* **369**, 936–942 (2020).
3. Campbell, D. E. *et al.* Infection with Bacteroides Phage BV01 Alters the Host Transcriptome and Bile Acid Metabolism in a Common Human Gut Microbe. *Cell Rep.* **32**, 108142 (2020).
4. Oh, J.-H. *et al.* Dietary Fructose and Microbiota-Derived Short-Chain Fatty Acids Promote Bacteriophage Production in the Gut Symbiont Lactobacillus reuteri. *Cell Host Microbe* **25**, 273-284.e6 (2019).

389 5. Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The “Known Unknown” of the
390 Microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
391 6. Cao, Z. *et al.* The gut virome: A new microbiome component in health and disease. *eBioMedicine*
392 **81**, 104113 (2022).
393 7. Norman, J. M. *et al.* Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel
394 Disease. *Cell* **160**, 447–460 (2015).
395 8. de Jonge, P. A. *et al.* Gut virome profiling identifies a widespread bacteriophage family associated
396 with metabolic syndrome. *Nat. Commun.* **13**, 3594 (2022).
397 9. Reyes, A. *et al.* Gut DNA viromes of Malawian twins discordant for severe acute malnutrition.
398 *Proc. Natl. Acad. Sci.* **112**, 11941–11946 (2015).
399 10. Brown, E. M. *et al.* Gut microbiome ADP-ribosyltransferases are widespread phage-encoded
400 fitness factors. *Cell Host Microbe* **29**, 1351-1365.e11 (2021).
401 11. Simmonds, P. *et al.* Four principles to establish a universal virus taxonomy. *PLOS Biol.* **21**,
402 e3001922 (2023).
403 12. Turner, D. *et al.* Abolishment of morphology-based taxa and change to binomial species
404 names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch. Virol.* **168**, 74
405 (2023).
406 13. Barylski, J. *et al.* Analysis of Spounaviruses as a Case Study for the Overdue Reclassification
407 of Tailed Phages. *Syst. Biol.* **69**, 110–123 (2020).
408 14. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of
409 human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).
410 15. Koonin, E. V. & Yutin, N. The crAss-like Phage Group: How Metagenomics Reshaped the
411 Human Virome. *Trends Microbiol.* **28**, 349–359 (2020).
412 16. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual
413 Specific. *Cell Host Microbe* **26**, 527-541.e5 (2019).

414 17. Manrique, P. *et al.* Healthy human gut phageome. *Proc. Natl. Acad. Sci.* **113**, 10400–10405
415 (2016).
416 18. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive
417 expansion of human gut bacteriophage diversity. *Cell* **184**, 1098-1109.e9 (2021).
418 19. Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome
419 Diversity in the Human Gut. *Cell Host Microbe* **28**, 724-740.e8 (2020).
420 20. Tisza, M. J. & Buck, C. B. A catalog of tens of thousands of viruses from human metagenomes
421 reveals hidden associations with chronic diseases. *Proc. Natl. Acad. Sci.* **118**, e2023202118 (2021).
422 21. Roux, S. *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for
423 interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* **49**, D764–D775 (2021).
424 22. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled
425 viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
426 23. Bouras, G. *et al.* Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* **39**,
427 btac776 (2023).
428 24. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood
429 Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
430 25. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving
431 the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
432 26. Turner, D., Kropinski, A. M. & Adriaenssens, E. M. A Roadmap for Genome-Based Phage
433 Taxonomy. *Viruses* **13**, 506 (2021).
434 27. Gorbalenya, A. E. *et al.* The new scope of virus taxonomy: partitioning the virosphere into 15
435 hierarchical ranks. *Nat. Microbiol.* **5**, 668–674 (2020).
436 28. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status,
437 taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

438 29. Smith, L., Goldobina, E., Govi, B. & Shkoporov, A. N. Bacteriophages of the Order
439 Crassvirales: What Do We Currently Know about This Keystone Component of the Human Gut
440 Virome? *Biomolecules* **13**, 584 (2023).
441 30. Samson, J. E., Magadán, A. H., Sabri, M. & Moineau, S. Revenge of the phages: defeating
442 bacterial defences. *Nat. Rev. Microbiol.* **11**, 675–687 (2013).
443 31. Guo, X. *et al.* Structure and mechanism of a phage-encoded SAM lyase revises catalytic
444 function of enzyme family. *eLife* **10**, e61818.
445 32. Howard-Varona, C. *et al.* Phage-specific metabolic reprogramming of virocells. *ISME J.* **14**,
446 881–895 (2020).
447 33. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over
448 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-
449 662.e20 (2019).
450 34. Maixner, F. *et al.* Hallstatt miners consumed blue cheese and beer during the Iron Age and
451 retained a non-Westernized gut microbiome until the Baroque period. *Curr. Biol.* **31**, 5149-
452 5162.e6 (2021).
453 35. Wibowo, M. C. *et al.* Reconstruction of ancient microbial genomes from the human gut.
454 *Nature* **594**, 234–239 (2021).
455 36. Maixner, F. *et al.* The 5300-year-old Helicobacter pylori genome of the Iceman. *Science* **351**,
456 162–165 (2016).
457 37. Bonani, G., Ivy, S. D., Hajdas, I., Niklaus, T. R. & Suter, M. Ams 14C Age Determinations of
458 Tissue, Bone and Grass Samples from the Ötztal Ice Man. *Radiocarbon* **36**, 247–250 (1994).
459 38. Willerslev, E. & Meltzer, D. J. Peopling of the Americas as inferred from ancient genomics.
460 *Nature* **594**, 356–364 (2021).
461 39. Bergström, A., Stringer, C., Hajdinjak, M., Scerri, E. M. L. & Skoglund, P. Origins of modern
462 human ancestry. *Nature* **590**, 229–237 (2021).

463 40. Rampelli, S. *et al.* Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota.
464 *Curr. Biol. CB* **25**, 1682–1693 (2015).
465 41. Obregon-Tito, A. J. *et al.* Subsistence strategies in traditional societies distinguish gut
466 microbiomes. *Nat. Commun.* **6**, 6505 (2015).
467 42. Sonnenburg, E. D. & Sonnenburg, J. L. The ancestral and industrialized gut microbiota and
468 implications for human health. *Nat. Rev. Microbiol.* **17**, 383–390 (2019).
469 43. Vacca, M. *et al.* The Controversial Role of Human Gut Lachnospiraceae. *Microorganisms* **8**,
470 573 (2020).
471 44. Kostic, A. D. *et al.* The Dynamics of the Human Infant Gut Microbiome in Development and
472 in Progression toward Type 1 Diabetes. *Cell Host Microbe* **17**, 260–273 (2015).
473 45. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes.
474 *Nature* **490**, 55–60 (2012).
475 46. Sasaki, K. *et al.* Construction of a Model Culture System of Human Colonic Microbiota to
476 Detect Decreased Lachnospiraceae Abundance and Butyrogenesis in the Feces of Ulcerative
477 Colitis Patients. *Biotechnol. J.* **14**, 1800555 (2019).
478 47. Zhao, G. *et al.* Intestinal virome changes precede autoimmunity in type I diabetes-
479 susceptible children. *Proc. Natl. Acad. Sci.* **114**, E6166–E6175 (2017).
480 48. Ma, Y., You, X., Mai, G., Tokuyasu, T. & Liu, C. A human gut phage catalog correlates the gut
481 phageome with type 2 diabetes. *Microbiome* **6**, 24 (2018).
482 49. Fragiadakis, G. K. *et al.* Links between environment, diet, and the hunter-gatherer
483 microbiome. *Gut Microbes* **10**, 216–227 (2019).
484 50. Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most
485 abundant viruses from the human gut. *Nat. Microbiol.* **3**, 38–46 (2018).
486 51. Krupovic, M., Dolja, V. V. & Koonin, E. V. The LUCA and its complex virome. *Nat. Rev.*
487 *Microbiol.* **18**, 661–670 (2020).

488 52. Camargo, A. P. *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes
489 within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids*
490 *Res.* **51**, D733–D743 (2023).
491 53. Weyrich, L. S. *et al.* Neanderthal behaviour, diet, and disease inferred from ancient DNA in
492 dental calculus. *Nature* **544**, 357–361 (2017).
493 54. He, Y. *et al.* Regional variation limits applications of healthy gut microbiome reference
494 ranges and disease models. *Nat. Med.* **24**, 1532–1535 (2018).
495 55. Zmora, N., Suez, J. & Elinav, E. You are what you eat: diet, health and the gut microbiota.
496 *Nat. Rev. Gastroenterol. Hepatol.* **16**, 35–56 (2019).
497 56. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human
498 gut virus crAssphage. *Nat. Microbiol.* **4**, 1727–1736 (2019).
499 57. Honap, T. P. *et al.* Biogeographic study of human gut-associated crAssphage suggests
500 impacts from industrialization and recent expansion. *PLOS ONE* **15**, e0226930 (2020).
501 58. Schnorr, S. L. *et al.* Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* **5**, 3654
502 (2014).
503 59. Snijder, M. B. *et al.* Cohort profile: the Healthy Life in an Urban Setting (HELIUS) study in
504 Amsterdam, The Netherlands. *BMJ Open* **7**, e017873 (2017).
505 60. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut
506 microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
507 61. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
508 identification. *BMC Bioinformatics* **11**, 119 (2010).
509 62. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
510 63. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the
511 analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

512 64. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and
513 curation of microbial viruses, and evaluation of viral community function from genomic
514 sequences. *Microbiome* **8**, 90 (2020).
515 65. Nishimura, Y. *et al.* ViPTree: the viral proteomic tree server. *Bioinformatics* **33**, 2379–2380
516 (2017).
517 66. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is
518 enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
519 67. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular
520 Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
521 68. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree
522 display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
523 69. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated
524 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
525 70. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference
526 in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
527 71. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
528 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–
529 589 (2017).
530 72. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut
531 microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
532 73. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
533 *Bioinformatics* **34**, i884–i890 (2018).
534 74. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile
535 metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
536 75. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient
537 alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

76. Borry, M., Hübner, A., Rohrlach, A. B. & Warinner, C. PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA de novo assembly. *PeerJ* **9**, e11845 (2021).

77. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

78. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).

79. Bouckaert, R. R. OBAMA: OBAMA for Bayesian amino-acid model averaging. *PeerJ* **8**, e9460 (2020).

Methods

Data

All bioinformatic tools were used with default settings, unless explicitly stated otherwise. Statistical analyses were done in R v4.2.1. All p-values were adjusted for multiple testing with the Benjamini-Hochberg method.

The primary dataset of *Heliusviridae* marker genes were derived from 298 contigs that were used in our earlier study[8], which studied the gut virome in the general population-based multi-ethnic Healthy Life in an Urban Setting (HELIUS) cohort[59]. Contig datasets that were used to identify *Heliusviridae* consisted of 63,481 genomes from this previous study. This included all viral contigs before removing redundant ones. Further data sources consisted of four recently constructed databases of gut bacteriophages: the gut phage database (GPD, 142,809 genomes)[18], cenote-taker 2 human virome database (CHVD, 93,860 genomes)[20], gut virome database (GVD, 33,242 genomes)[19], and metagenomic gut virus catalogue (MGV, 189,680 genomes)[60]. Further contigs were derived from fourth version of the integrated microbial genomes viral dataset (IMG/VR, 15,722,824 genomes)[21] and bacterial viruses from the NCBI viral reference sequence dataset (RefSeq, 4,220

genomes) release 216. For studies of urbanization- and disease-links, metagenome assemblies were obtained from Pasolli *et al*[33]. of 7,166 human gut metagenomes of 5,441 individuals from 38 studies.

Identification of putative Heliusviridae sequences

Since previous findings showed *Heliusviridae* genomes to be around 50,000-100,000 bp long, we first selected contigs of at least 25,000 bp without any ambiguous bases from all datasets. This resulted in an effective search space of 1,314,988 contigs. Open reading frames (ORFs) were then predicted using prodigal v2.6.3[61] (option --meta).

For each of the nine *Heliusviridae* marker genes, a profile hidden Markov model (HMM) was constructed from protein sequences derived from our previous study using hmmbuild v3.3[62]. The resulting profile HMMs were used to search against predicted ORFs using hmmsearch v3.3[62]. For phylogenetic studies, we selected contigs with an hmmsearch hit (bit-score ≥600) against the *Heliusviridae* terminase large subunit (TerL). Duplicates were removed with dedupe from BBTools v38.84 (option minidentity=100), resulting in a non-redundant dataset of 6,359 contigs. These were then analyzed for completeness with checkv v1.0.1[22] (using database version 1.4), and we selected contigs that had 100% completeness, no warnings, and were not recognized as prophages. These genomes were fully annotated with pharokka v1.2.1[23] (options -m, -g prodigal), and their ORFs were grouped in protein clusters using mmseqs2 v14.7e284[63]. GC contents were derived from pharokka output, and temperate phages were defined as those with at least one ORF with a predicted function in the “integration and excision” category. Potential auxiliary metabolic genes encoded by the genomes were identified with VIBRANT v1.2.1[64].

Viral sequence clustering and proteomic analysis

Separation between our phages and those in the NCBI refseq database was analyzed with both ViPTreeGen v1.1.3[65] and vContact2 v0.11.3[66]. The vContact2 network was visualized with Cytoscape v3.9.1[67], while the ViPTreeGen tree was visualized with the interactive tree

of life (iTOL) webtool[68]. To analyze the grouping of genomes among the high quality contigs, we used them to constructed a second vContact2 network.

Phylogenetic tree construction

Before phylogenetic tree construction, incomplete protein sequences were removed by selecting only sequences that were longer than the 25th percentile minus 1.5 times the inter-quantile ranges of all lengths. As outgroup, the homologous TerL sequence from *Staphylococcus* phage LH1 (NC055041) was added. Marker genes were aligned using MAFFT v4.753 (options --maxiterate, --localpair), after which positions with more than 90% gaps were trimmed with trimal v1.4.rev15[69] (option -gt 0.1). Finally, a tree was built using IQTree v2.2.0.3[70], using model finder[71] and 1000 iterations of the ultrafast bootstrap approximation[25] and SH-like approximate likelihood ratio test. IQTree performed ten separate tree iterations and selected the one with the best log likelihood score (options -B 1000 -alrt 1000 --runs 10). The tree was visualized with the interactive tree of life (iTOL) webtool[68].

Heliusvirales hosts

We determined the hosts infected by the *Heliusvirales* with two complementary methods, adapted from previous work[60]. Firstly, we identified phage-host links by matching CRISPR spacers. For this, we downloaded 1,846,441 CRISPR spacers identified[60] among the Bacteria and Archaea in the Unified Human Gastrointestinal Genomes (UHGG) dataset[72]. We then created a subset of 1,742,522 spacers with lengths between 30 and 40 bp and without undefined bases. These we used for a blastn search against *Heliusvirales* genomes (options -task blastn-short). Hits were finally filtered for those with 0-2 mismatches across the entire length. Our second host-linking approach focused on identifying phages as prophages. We downloaded all 286,997 bacterial and archaeal genomes from the Unified Human Gastrointestinal Genomes (UHGG) dataset[72]. Viral genomes were then used for a

blastn search against these genomes, after which hits with over 96% identity and a length of over 1000 bp were selected. Prophage and CRISPR-spacer hits were combined and the most common phage-host link was determined for each viral genome. If one bacterial lineage was responsible for 50+% of all phage-host linkage hits, it was considered as the host of that phage.

Analysis of ancient samples

For analysis of ancient metagenomes, reads were downloaded from two separate studies[34,35]. These were trimmed with fastp v.0.23.2[73] (options --detect_adapter_for_pe) and error-corrected with tadpole from the bbmap package v38.90 (https://jgi.doe.gov/data-and-tools/bbtools, options mode=correct, ecc=t, prefilter=2). Error corrected reads were assembled with metaSPAdes v3.15.5[74]. The ancient origins of assembled contigs were determined by mapping reads to contigs of ≥1,000 bp with bowtie2 v2.4.2[75], and determining C-to-T degradation at the 5' end of sequencing reads with PyDamage v0.70[76]. Contigs with a predicted accuracy of over 0.5 were considered of ancient origins.

Heliusvirales prevalence and relation to urbanization

To determine *Heliusvirales* prevalence, we firstly gathered geographic and environmental metadata from all genomes used to build the main phylogenetic tree. Second, contigs obtained from Pasolli et al[33] were analyzed for the presence of *Heliusvirales* TerL genes. To do this, ORFs were predicted with Prodigal v2.6.3[61] (option –meta), and used for an hmmsearch with the profile hmm of the TerL constructed from sequences identified in our previous study. Hits with bit-scores of over 600 were selected. The same approach was used to identify *Heliusvirales* TerL genes among ancient populations. Enrichment of phages from certain environments among selected subfamilies was calculated with the phyper function from base R.

To analyze the relation of *Heliusvirales* richness to urbanization, richness was calculated by dividing the number of identified TerL genes by the number of reads in billions. Significance testing of richness differences after combining studies was done with linear mixed effects models where the study was entered as a random effect, using the lme function from the nlme R package v3.1-162 and Anova function from the rstatix R package v0.7.2. The same approach was used to calculate significant richness differences among pooled disease-related cohorts.

Assignment of TerL sequences from ancient- and disease-related samples was done by constructing a phylogenetic tree with TerL genes from all high-quality *Heliusvirales* genomes that had been used in the main phylogeny. For the ancient, hunter-gatherers, and urban populations, a single such tree was built, while for the disease-related datasets one tree was constructed per project. Trees were constructed with IQTree v2.2.0.3 as described above, and pairwise cophenetic distances were calculated from them with the ape R package v5.7-1. Each sequence was then assigned to the same taxonomy as its closest neighbor of which the taxonomy was known.

For the ancient, hunter-gatherers, and urban populations, taxonomic assignment at the family level was used to calculate UniFrac distances, from which a non-metric multidimensional scaling (NMDS) was constructed. Permutated ANOVA significance levels were calculated with the adonis2 function in the vegan R package v2.6-4. To determine significant differences in subfamily prevalence between populations with different lifestyles, Fisher exact tests were calculated using the fisher.test function from base R.

Timed tree construction

To reconstruct an evolutionary timed tree of the *Ca. Hathorviridae*, we first constructed a regular tree on members of this family. For this, we selected the TerL genes of all *Ca. Hathorviridae* from both the complete genomes and the ancient samples. This was used to build a phylogenetic tree with IQTree as described above. From this tree, we selected

thirteen complete genomes and all three ancient genome fragments that combined reflected the architecture of the *Ca. Hathorviridae*. All ORFs from these genomes were then clustered with mmseqs2 v14.7e284[63], after which the six protein clusters that were universally present were selected. This number was relatively low because the ancient sequences were genome fragments of 15,076, 19,664, and 67,881 bp long. The six universal PCs were a TerL, head maturation protease, major head protein, portal protein, a hypothetical protein, and terminase small subunit. For each protein, an alignment was made with MAFFT v4.753 (options --maxiterate, --localpair)[77], from which a timed tree was inferred using BEAST v2.7.3[78]. BEAST was run with the following models: Birth death model, Coalescent constant population, Coalescent exponential population, Coalescent bayesian skyline, Coalescent extended bayesian skyline, each until the estimated sample size was above 200. The coalescent exponential population was the best-fitting model. BEAST was run with OBAMA v1.1.1[79] for amino-acid model averaging.

Acknowledgements

P.A.d.J. was supported by a Talent Development grant of the Amsterdam Gastroenterology, Endocrinology, and Metabolism research school at the Amsterdam UMC. MN was supported by a personal ZONMW-VICI grant 2020 (09150182010020). HH was supported by an Aspasia premium (015.017.050) and Dutch Diabetes Research Foundation (2019.82.004). The HELIUS study is conducted by the Amsterdam University Medical Centers, location AMC and the Public Health Service of Amsterdam. Both organizations provided core support for HELIUS. The HELIUS study is also funded by the Dutch Heart Foundation, the Netherlands Organization for Health Research and Development (ZonMw), the European Union (FP-7), and the European Fund for the Integration of non-EU immigrants (EIF). We are most grateful to the participants of the HELIUS study and the management team, research nurses, interviewers, research assistants and other staff who have taken part in gathering the data of this study.

Author contributions

P.A.d.J. conducted data analysis and wrote the manuscript; B.J.v.d.B., A.H.Z., and M.N. assisted with experimental design and data interpretation; P.A.d.J., B.E.D., and H.H. designed the study. All authors read and provided input on the manuscript.

Competing interests

M.N. owns stock in, consults for, and has intellectual property rights in Caelus Health. He consults for Kaleido. None of these are directly relevant to the current paper. The remaining authors declare no competing interests.

Figure legends

Figure 1: Presence of near-universal marker genes is linked to TerL homology. **a:** Annotation of NatCom_15117, a complete genome with inverse terminal repeats. The nine marker genes are denoted above the genome line, with re-annotated functions in parentheses. Other annotated genes are below the genome line. **b:** Overview of the procedure by which genomes were selected from seven phage databases. **c:** density plots of bit-scores from a search with the TerL HMM, separated by the total amount of marker genes found in the genome. Dots are individual complete genomes. **d:** prevalence of the most common protein clusters. Red denotes marker genes, grey others.

Figure 2: the *Ca. Heliusvirales* are a large order of gut phages with distinct subgroups. **a:** a rooted approximate maximum-likelihood tree of *Ca. Heliusvirales* TerL genes. The outgroup

at the extreme left-hand side of the tree is *Triavirus LH1* (NC_055041). **b:** chart depicting the pairwise percent of shared genes among the three main families (top) and their subfamilies (bottom). The n reflects the number of genomes in the lineage. For an overview of all families and subfamilies, see **Supplementary Figure 1**. Box plots show the median (middle line), 25th, and 75th percentile (box), with the 25th percentile minus and the 75th percentile plus 1.5 times the interquartile range (whiskers), and outliers (single points). **c:** protein-sharing network of all sequences in **a**. Node edge color denotes family, while fill denote subfamilies, both according to the same colors as use in **a**.

Figure 3: high *Ca. Heliusvirales* diversity in environmental origin and host species. **a:** environmental origins of the *Ca. Heliusvirales* subfamilies. **b:** bacterial host assignment of *Ca. Heliusvirales* subfamilies. **c:** pathway assignment of putative auxiliary metabolic genes (AMGs) and morons encoded by the *Ca. Heliusvirales*. For brevity, the 25 pathways that were most commonly found were selected.

Figure 4: the *Ca. Heliusvirales* are widespread and ancient. **a:** map showing in which countries *Ca. Heliusvirales* TerL-containing sequences were found. Data pertains to the seven bacteriophage databases described in the methods. Red dots show the locations of ancient gut microbiomes that were analyzed for *Ca. Heliusvirales* TerL presence, with their dates, gene-, and sample-counts. **b:** prevalence of *Ca. Heliusvirales* TerL genes among individuals of 39 gut metagenome studies. The red bar is the prevalence among the totality of all studies. **c:** tip-dated phylogenetic tree of *Ca. Hathorviridae* sequences from modern and ancient samples. Bold sequences are from ancient samples. Bar plots denote 95% confidence, values denote mean and 95% confidence intervals **d:** boxplots of the number of *Ca. Heliusvirales* TerL genes found per person, per billion sequenced bases, with jittered points separated by study. Significance is according to nested linear mixed-effect models

where the study was used as a random effect. Box plots show the median (middle line), 25th, and 75th percentile (box), with the 25th percentile minus and the 75th percentile plus 1.5 times the interquartile range (whiskers). Comparisons of richness in urbanized vs hunter-gatherer populations within each study can be found in Supplementary Figure 2. **e:** non-metric multidimensional scaling (NMDS) on unweighted UniFrac distances, colored by lifestyle. **f:** Prevalence of *Ca. Heliusvirales* subfamilies in ancient peoples, hunter-gatherers, and urbanized peoples, with significance according to Fisher exact tests for subfamilies with significant values. All p-values were adjusted for multiple testing with the Benjamini-Hochberg procedure, and are denoted as follows * ≤ 0.05, ** ≤0.01, *** ≤ 0.001, **** ≤ 0.0001. Lineages without any representatives among the samples were not displayed.

Figure 5: disease associations of *Ca. Heliusvirales* phages in gut microbiomes. **a:** four studies in which a significant difference was found in the number of *Ca. Heliusvirales* TerL genes per person per billion sequenced bases. Significance is according to Wilcoxon signed-rank tests. Box plots show the median (middle line), 25th, and 75th percentile (box), with the 25th percentile minus and the 75th percentile plus 1.5 times the interquartile range (whiskers). **b:** non-metric multidimensional scaling (NMDS) on unweighted UniFrac distances, colored by disease status. **c:** the same tree as in **Figure 1a**, but with all subfamilies collapsed. Boxes show the difference in prevalence between controls and people with disease. All p-values were adjusted for multiple testing with the Benjamini-Hochberg procedure.

Supplementary Figure 1: shared gene content for all families (top) and subfamilies (bottom).

Supplementary Figure 2: **a**: vContact2 genome network of the *Ca. Heliusvirales* and all viral sequences in the NCBI RefSeq database. **b:** VipTree proteome tree of the same sequences in **a**.

Supplementary Figure 3: GC-content of the *Ca. Heliusvirales* by subfamily.

Supplementary Figure 4: the most common host predictions in the six largest subfamilies.

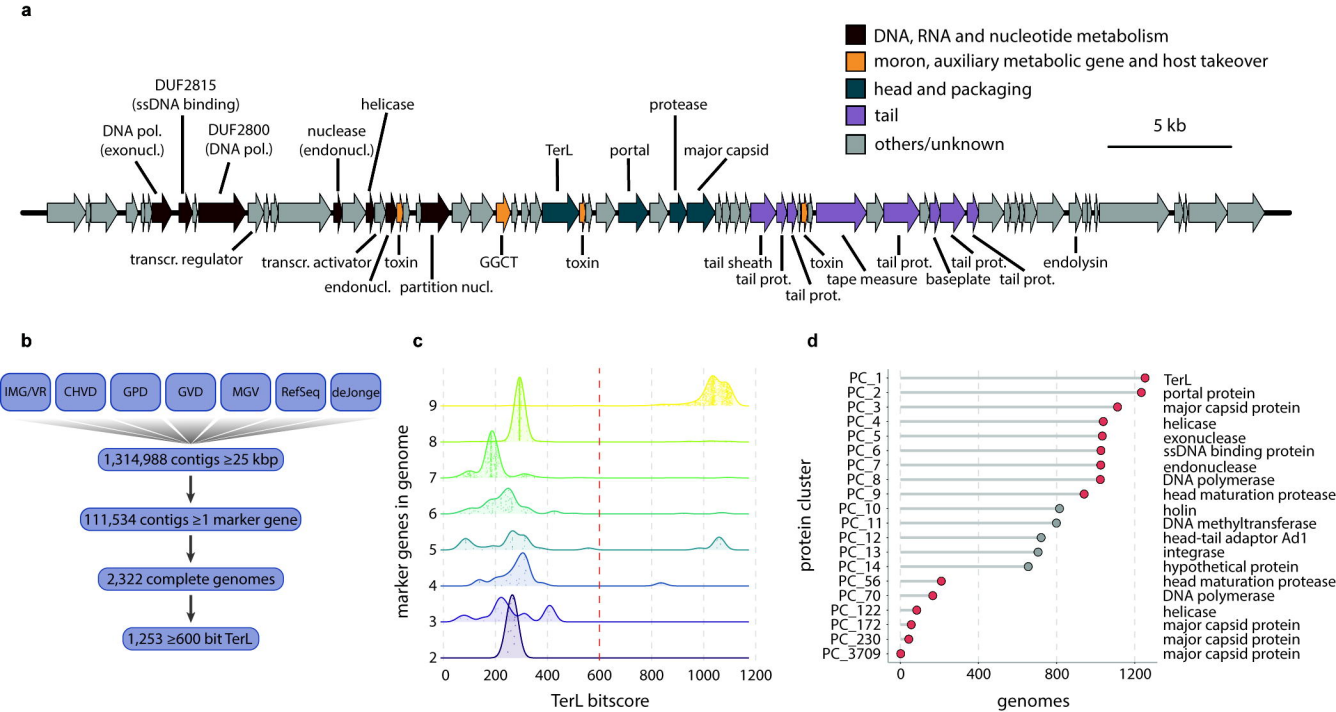
Supplementary Figure 5: the percentage of temperate phages across the subfamilies, as determined by the presence of integration and excision genes in pharokka annotations. The dashed red line is the average for all *Ca. Heliusvirales*

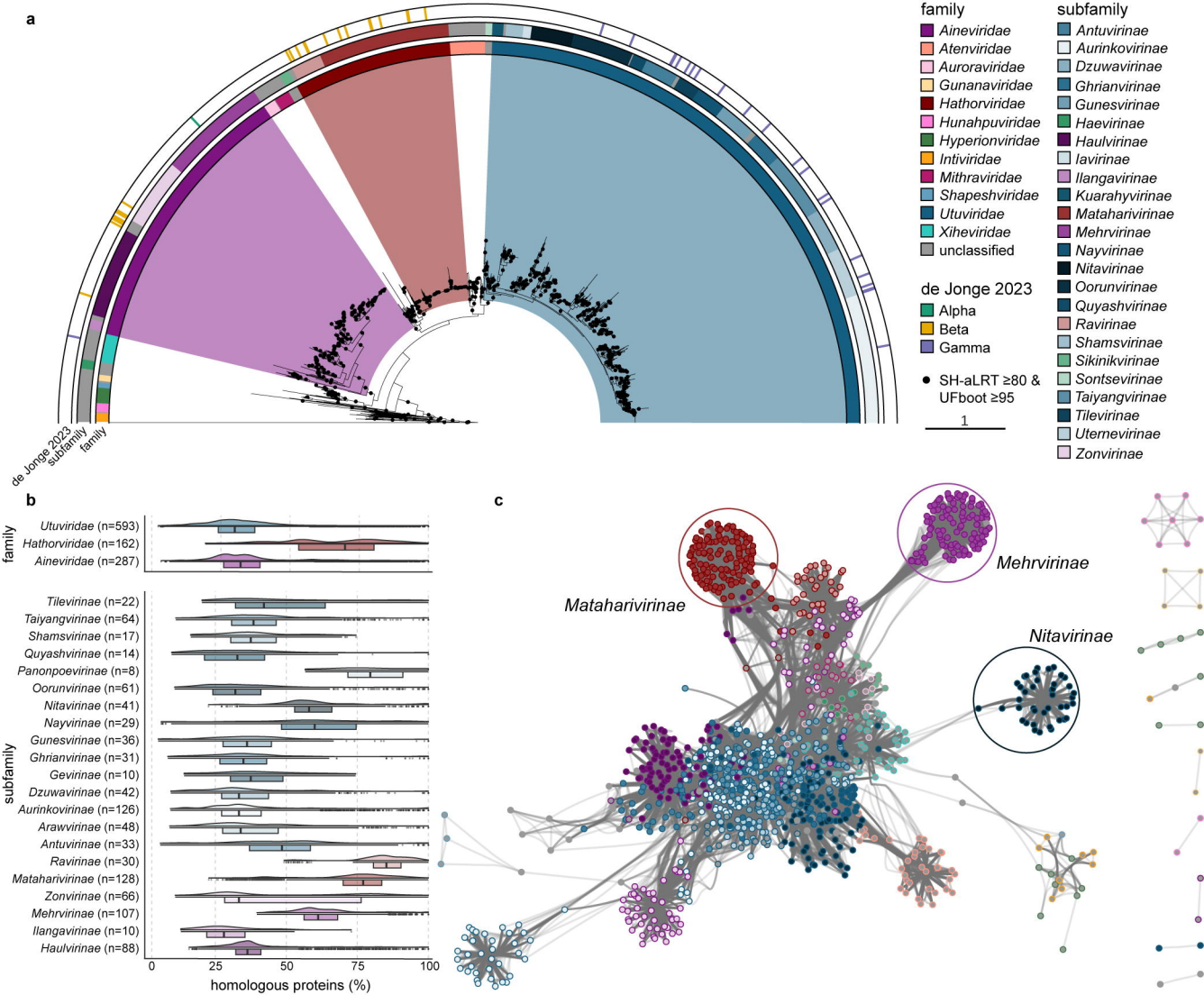
Supplementary Figure 6: An approximate maximum-likelihood tree of *Ca. Hathorviridae* TerL genes, with color coding conform genome origin. Dots show branches with SH-like approximate likelihood ratio test ≥80 and ultrafast bootstrap approximation ≥95.

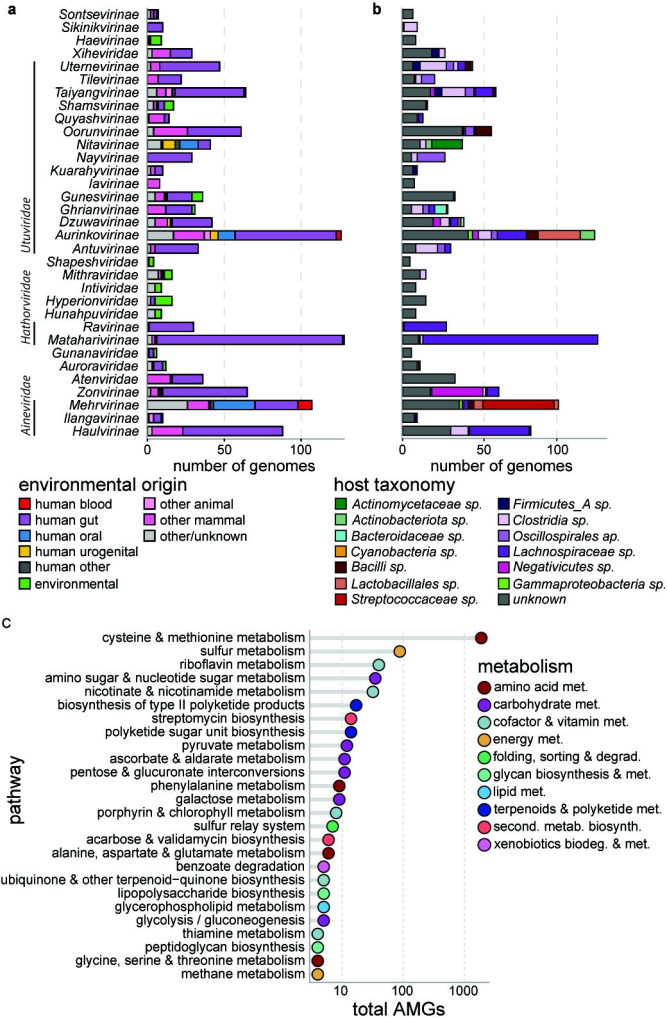
Supplementary Figure 7: An approximate maximum-likelihood tree of *Ca. Matahariviridae* TerL genes, with color coding conform genome origin. Dots show branches with SH-like approximate likelihood ratio test ≥80 and ultrafast bootstrap approximation ≥95.

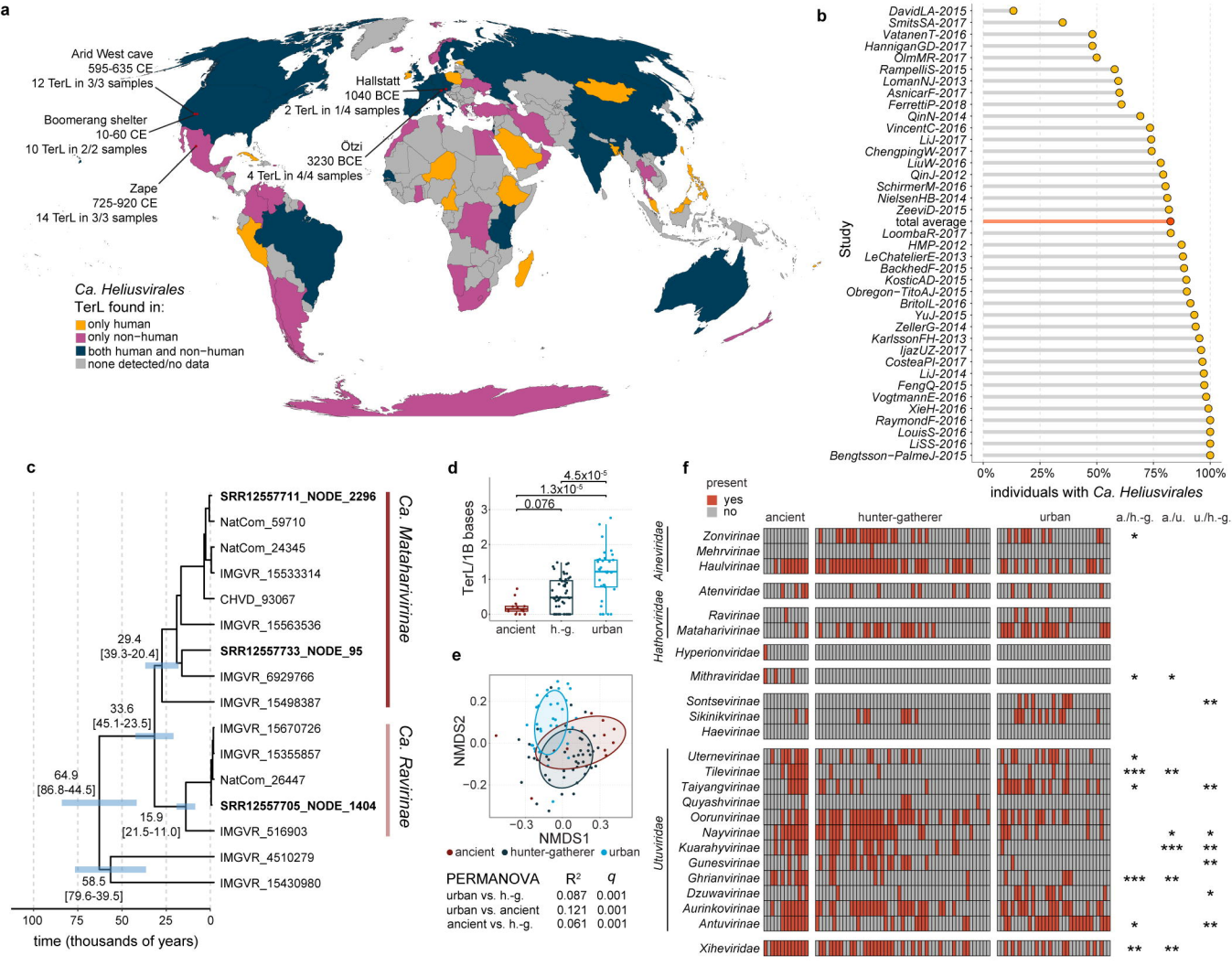
Supplementary Figure 8: **a**: comparisons of *Ca. Heliusvirales* richness between urban and hunter-gatherer populations in two studies. Q-values are according to Wilcoxon signed-rank tests. **b:** comparison of *Ca. Heliusvirales* TerL abundance between lifestyles. **c:** taxonomic *Ca. Heliusvirales* richness comparison between ancient, hunter-gatherer, and urban

populations. Q-values according to linear mixed-effect models with the study as a random effect. All p-values are adjusted for multiple testing with the Benjamini-Hochberg method.

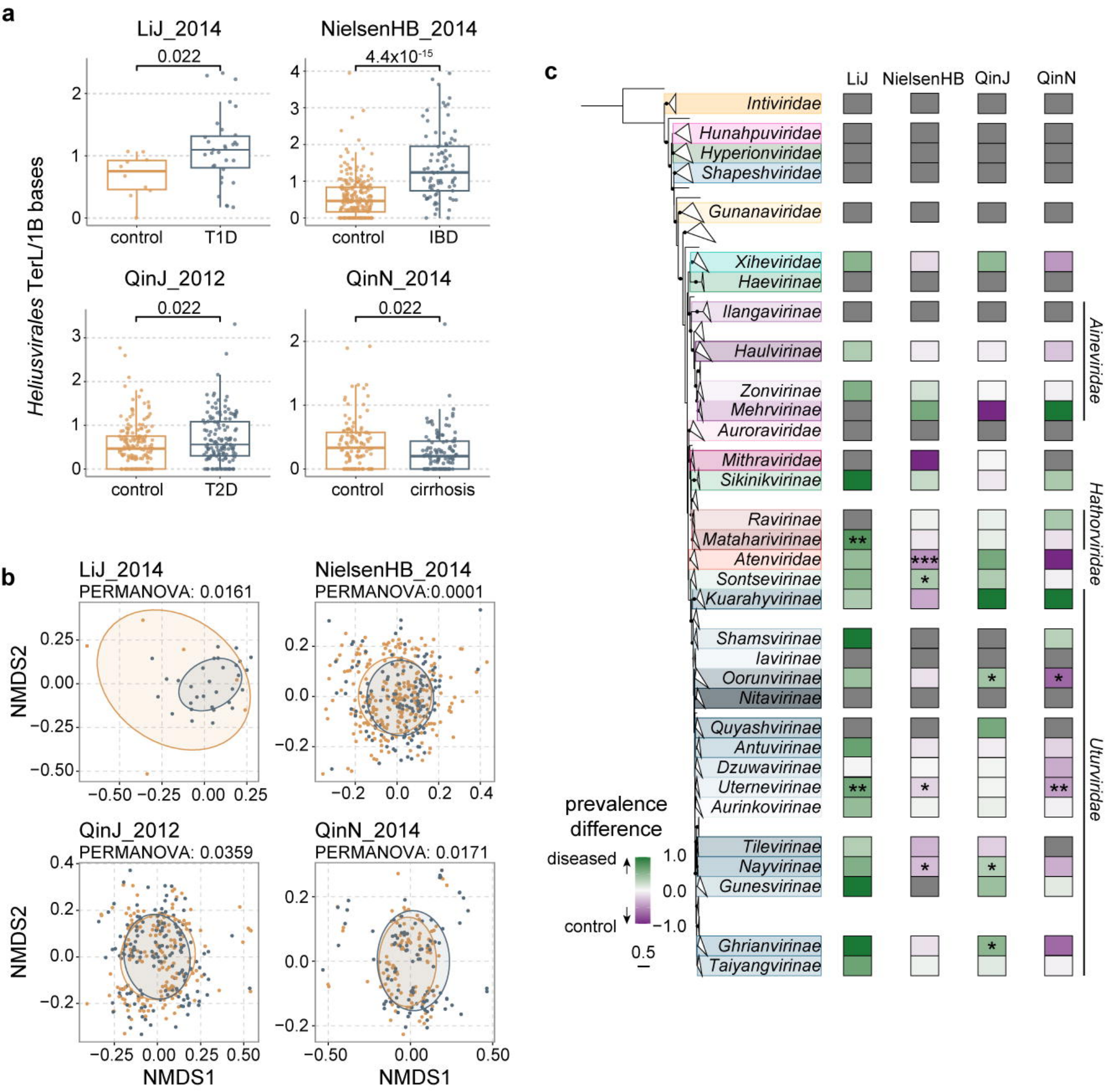

a
DNA, RNA and nucleotide metabolism
moron, auxiliary metabolic gene and host takeover
head and packaging
tail
others/unknown
5 kb
DNA pol. (exonucl.)
DUF2815 (ssDNA binding)
DUF2800 (DNA pol.)
nuclease (endonucl.)
helicase
TerL
portal
protease
major capsid
transcr. regulator
transcr. activator
toxin
endonucl.
partition nucl.
GGCT
toxin
tail sheath
tail prot.
tail prot.
toxin
tape measure
tail prot.
baseplate
tail prot.
tail prot.
endolysin
b
IMG/VR
CHVD
GPD
GVD
MGV
RefSeq
deJonge
1,314,988 contigs ≥25 kbp
111,534 contigs ≥1 marker gene
2,322 complete genomes
1,253 ≥600 bit TerL
c
marker genes in genome
TerL bitscore
d
protein cluster
genomes
PC_1 TerL
PC_2 portal protein
PC_3 major capsid protein
PC_4 helicase
PC_5 exonuclease
PC_6 ssDNA binding protein
PC_7 endonuclease
PC_8 DNA polymerase
PC_9 head maturation protease
PC_10 holin
PC_11 DNA methyltransferase
PC_12 head-tail adaptor Ad1
PC_13 integrase
PC_14 hypothetical protein
PC_56 head maturation protease
PC_70 DNA polymerase
PC_122 helicase
PC_172 major capsid protein
PC_230 major capsid protein
PC_3709 major capsid protein

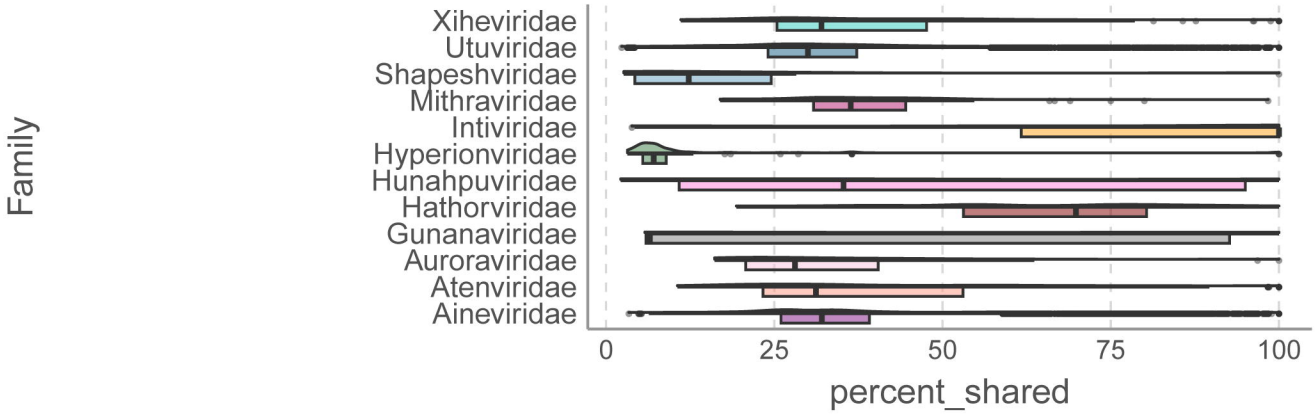

a
de Jonge 2023
subfamily
family
family
Aineviridae
Atenviridae
Auroraviridae
Gunanaviridae
Hathorviridae
Hunahpuviridae
Hyperionviridae
Intiviridae
Mithraviridae
Shapeshviridae
Utuviridae
Xiheviridae
unclassified
de Jonge 2023
Alpha
Beta
Gamma
SH-aLRT ≥80 & UFboot ≥95
1
subfamily
Antuvirinae
Aurinkovirinae
Dzuwavirinae
Ghrianvirinae
Gunesvirinae
Haevirinae
Haulvirinae
Iavirinae
Ilangavirinae
Kuarahyvirinae
Mataharivirinae
Mehrvirinae
Nayvirinae
Nitavirinae
Oorunvirinae
Quyashvirinae
Ravirinae
Shamsvirinae
Sikinikvirinae
Sontsevirinae
Taiyangvirinae
Tilevirinae
Uternevirinae
Zonvirinae
b
family
Utuviridae (n=593)
Hathorviridae (n=162)
Aineviridae (n=287)
subfamily
Tilevirinae (n=22)
Taiyangvirinae (n=64)
Shamsvirinae (n=17)
Quyashvirinae (n=14)
Panonpoevirinae (n=8)
Oorunvirinae (n=61)
Nitavirinae (n=41)
Nayvirinae (n=29)
Gunesvirinae (n=36)
Ghrianvirinae (n=31)
Gevirinae (n=10)
Dzuwavirinae (n=42)
Aurinkovirinae (n=126)
Arawvirinae (n=48)
Antuvirinae (n=33)
Ravirinae (n=30)
Mataharivirinae (n=128)
Zonvirinae (n=66)
Mehrvirinae (n=107)
Ilangavirinae (n=10)
Haulvirinae (n=88)
0
25
50
75
100
homologous proteins (%)
c
Mataharivirinae
Mehrvirinae
Nitavirinae

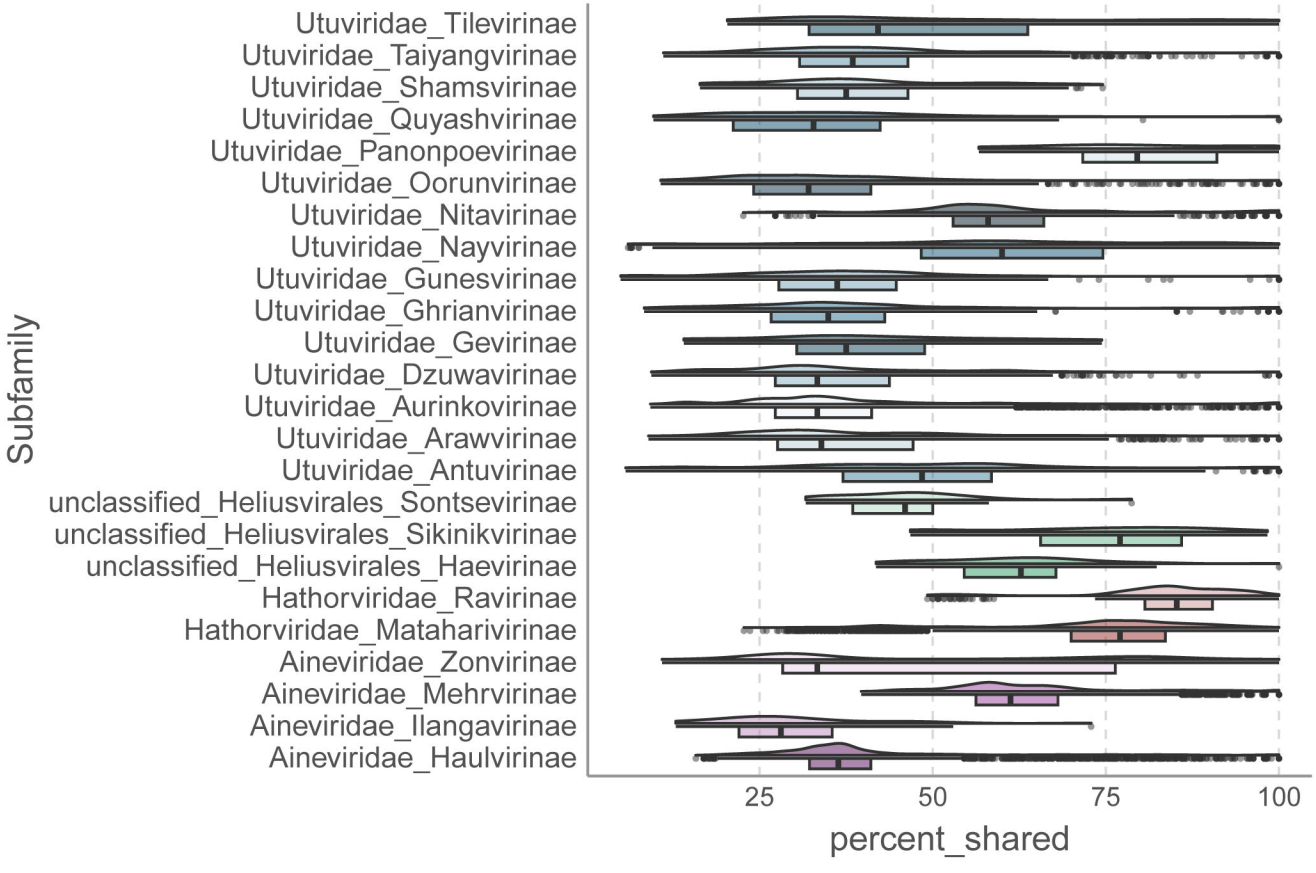

a
b
Sontsevirinae
Sikinikvirinae
Haevirinae
Xiheviridae
Uternevirinae
Tilevirinae
Taiyangvirinae
Shamsvirinae
Quyashvirinae
Oorunvirinae
Nitavirinae
Nayvirinae
Kuarahyvirinae
Iavirinae
Gunesvirinae
Ghrianvirinae
Dzuwavirinae
Aurinkovirinae
Antuvirinae
Shapeshviridae
Mithraviridae
Intiviridae
Hyperionviridae
Hunahpuviridae
Ravirinae
Matahariviridae
Gunanaviridae
Auroraviridae
Atenviridae
Zonvirinae
Mehrvirinae
Ilangavirinae
Haulvirinae
Utuviridae
Hathorviridae
Aineviridae
0
50
100
number of genomes
environmental origin
human blood
human gut
human oral
human urogenital
human other
environmental
other animal
other mammal
other/unknown
host taxonomy
Actinomycetaceae sp.
Actinobacteriota sp.
Bacteroidaceae sp.
Cyanobacteria sp.
Bacilli sp.
Lactobacillales sp.
Streptococcaceae sp.
Firmicutes_A sp.
Clostridia sp.
Oscillospirales ap.
Lachnospiraceae sp.
Negativicutes sp.
Gammaproteobacteria sp.
unknown
c
cysteine & methionine metabolism
sulfur metabolism
riboflavin metabolism
amino sugar & nucleotide sugar metabolism
nicotinate & nicotinamide metabolism
biosynthesis of type II polyketide products
streptomycin biosynthesis
polyketide sugar unit biosynthesis
pyruvate metabolism
ascorbate & aldarate metabolism
pentose & glucuronate interconversions
phenylalanine metabolism
galactose metabolism
porphyrin & chlorophyll metabolism
sulfur relay system
acarbose & validamycin biosynthesis
alanine, aspartate & glutamate metabolism
benzoate degradation
ubiquinone & other terpenoid−quinone biosynthesis
lipopolysaccharide biosynthesis
glycerophospholipid metabolism
glycolysis / gluconeogenesis
thiamine metabolism
peptidoglycan biosynthesis
glycine, serine & threonine metabolism
methane metabolism
pathway
10
100
1000
total AMGs
metabolism
amino acid met.
carbohydrate met.
cofactor & vitamin met.
energy met.
folding, sorting & degrad.
glycan biosynthesis & met.
lipid met.
terpenoids & polyketide met.
second. metab. biosynth.
xenobiotics biodeg. & met.

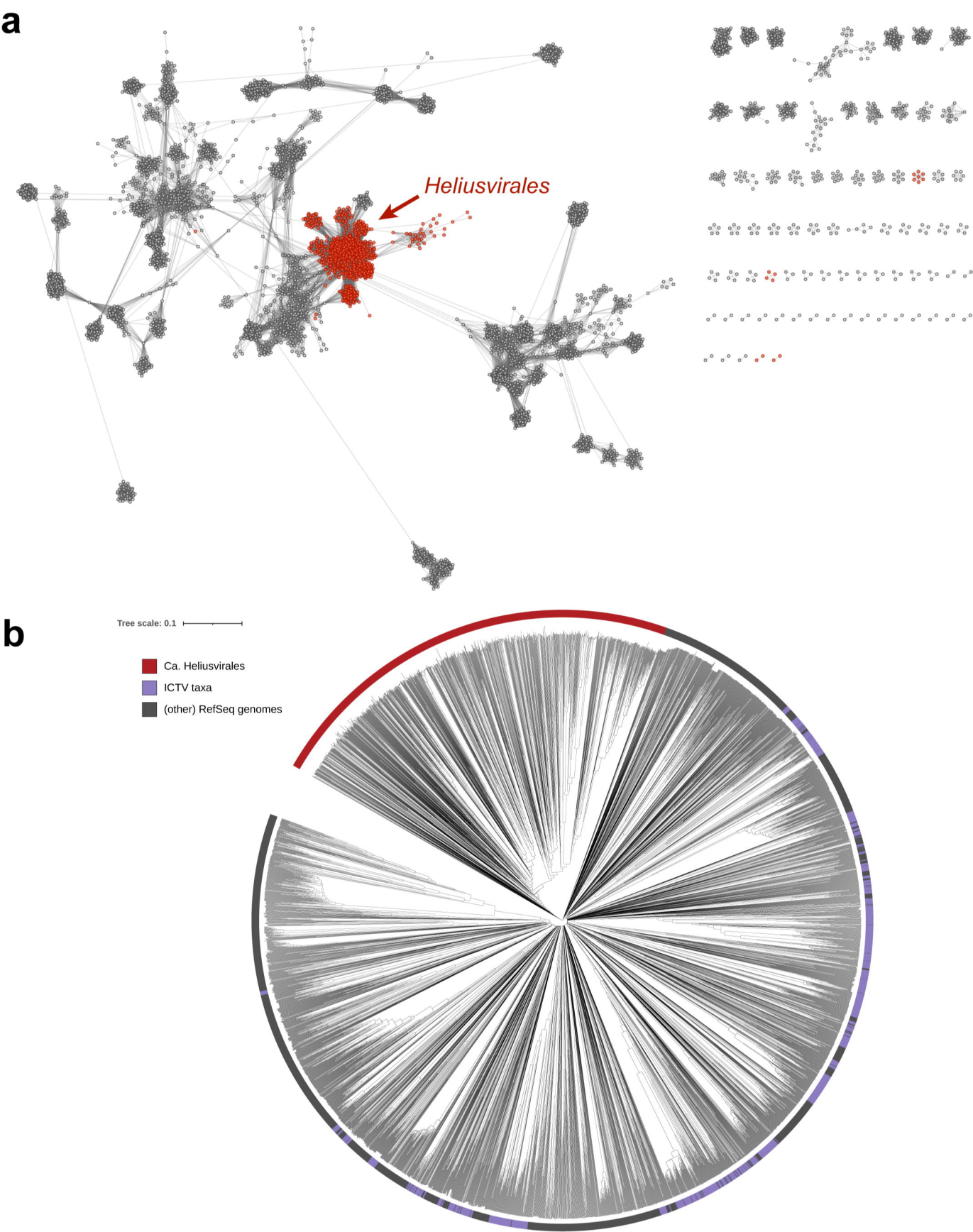


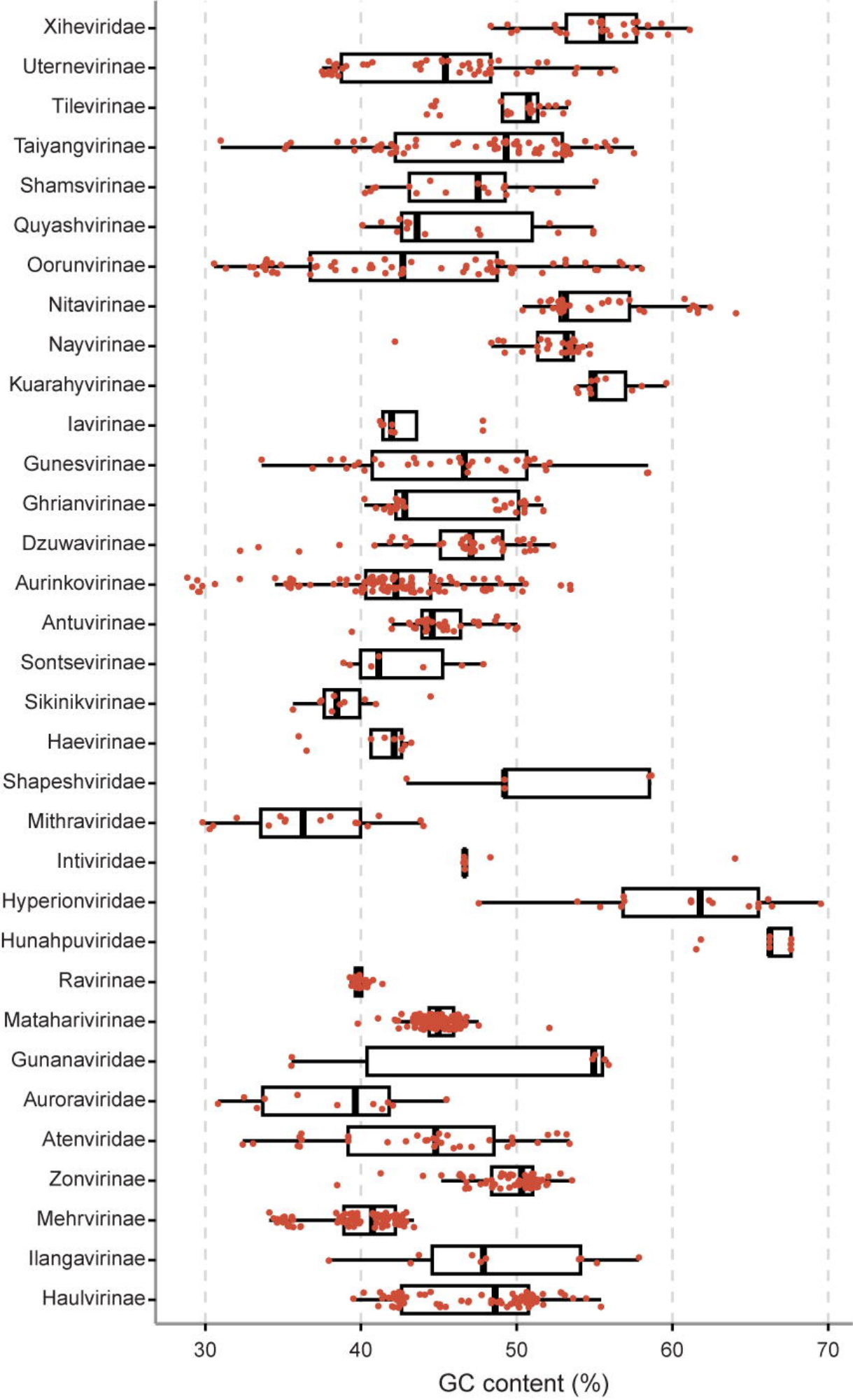
a
Arid West cave
595-635 CE
12 TerL in 3/3 samples
Boomerang shelter
10-60 CE
10 TerL in 2/2 samples
Zape
725-920 CE
14 TerL in 3/3 samples
Hallstatt
1040 BCE
2 TerL in 1/4 samples
Ötzi
3230 BCE
4 TerL in 4/4 samples
Ca. Heliusvirales
TerL found in:
only human
only non−human
both human and non−human
none detected/no data
b
Study
DavidLA-2015
SmitsSA-2017
VatanenT-2016
HanniganGD-2017
ÖlmMR-2017
RampelliS-2015
LomanNJ-2013
AsnicarF-2017
FerrettiP-2018
QinN-2014
VincentC-2016
LiJ-2017
ChengpingW-2017
LiuW-2016
QinJ-2012
SchirmerM-2016
NielsenHB-2014
ZeeviD-2015
total average
LoombaR-2017
HMP-2012
LeChatelierE-2013
BackhedF-2015
KosticAD-2015
Obregon−TitoAJ-2015
BritoIL-2016
YuJ-2015
ZellerG-2014
KarlssonFH-2013
IjazUZ-2017
CosteaPI-2017
LiJ-2014
FengQ-2015
VogtmannE-2016
XieH-2016
RaymondF-2016
LouisS-2016
LiSS-2016
Bengtsson−PalmeJ-2015
0%
25%
50%
75%
100%
individuals with Ca. Heliusvirales
c
SRR12557711_NODE_2296
NatCom_59710
NatCom_24345
IMGVR_15533314
CHVD_93067
IMGVR_15563536
SRR12557733_NODE_95
IMGVR_6929766
IMGVR_15498387
IMGVR_15670726
IMGVR_15355857
NatCom_26447
SRR12557705_NODE_1404
IMGVR_516903
IMGVR_4510279
IMGVR_15430980
29.4
[39.3-20.4]
33.6
[45.1-23.5]
64.9
[86.8-44.5]
15.9
[21.5-11.0]
58.5
[79.6-39.5]
Ca. Mataharivirinae
Ca. Ravirinae
100
75
50
25
0
time (thousands of years)
d
TerL/1B bases
4.5x10-5
1.3x10-5
0.076
ancient
h.-g.
urban
e
NMDS2
NMDS1
ancient
hunter-gatherer
urban
PERMANOVA R² q
urban vs. h.-g. 0.087 0.001
urban vs. ancient 0.121 0.001
ancient vs. h.-g. 0.061 0.001
f
present
yes
no
ancient
hunter-gatherer
urban
a./h.-g.
a./u.
u./h.-g.
Aineviridae
Zonvirinae
Mehrvirinae
Haulvirinae
Atenviridae
Hathorviridae
Ravirinae
Mataharivirinae
Hyperionviridae
Mithraviridae
Sontsevirinae
Sikinikvirinae
Haevirinae
Utuviridae
Uternevirinae
Tilevirinae
Taiyangvirinae
Quyashvirinae
Oorunvirinae
Nayvirinae
Kuarahyvirinae
Gunesvirinae
Ghrianvirinae
Dzuwavirinae
Aurinkovirinae
Antuvirinae
Xiheviridae

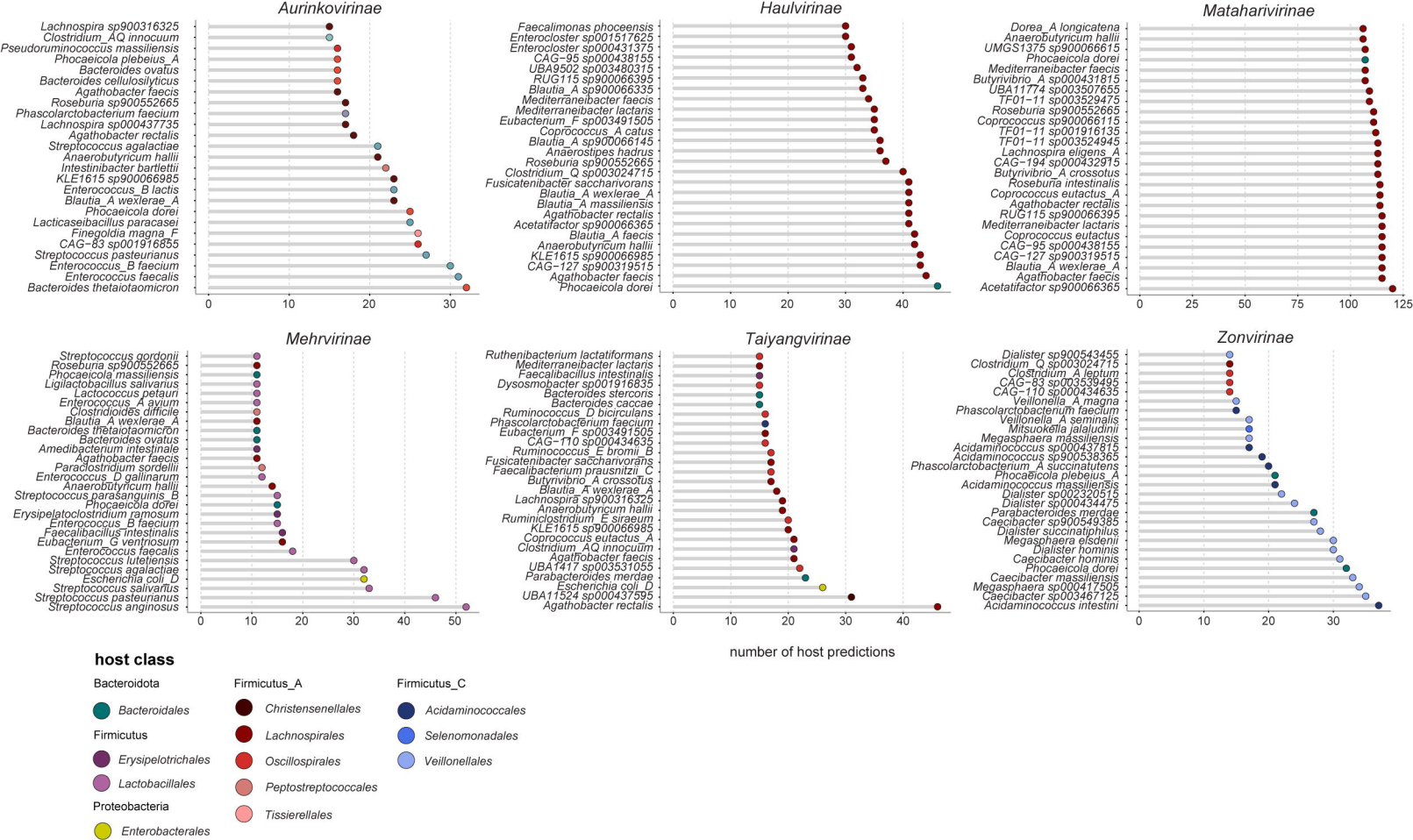

a
LiJ_2014
0.022
control
T1D
NielsenHB_2014
4.4x10-15
control
IBD
QinJ_2012
0.022
control
T2D
QinN_2014
0.022
control
cirrhosis
Heliusvirales TerL/1B bases
b
LiJ_2014
PERMANOVA: 0.0161
NielsenHB_2014
PERMANOVA:0.0001
QinJ_2012
PERMANOVA: 0.0359
QinN_2014
PERMANOVA: 0.0171
NMDS1
NMDS2
c
LiJ
NielsenHB
QinJ
QinN
Intiviridae
Hunahpuviridae
Hyperionviridae
Shapeshviridae
Gunanaviridae
Xiheviridae
Haevirinae
Ilangavirinae
Haulvirinae
Zonvirinae
Mehrvirinae
Auroraviridae
Mithraviridae
Sikinikvirinae
Ravirinae
Mataharivirinae
Atenviridae
Sontsevirinae
Kuarahyvirinae
Shamsvirinae
Iavirinae
Oorunvirinae
Nitavirinae
Quyashvirinae
Antuvirinae
Dzuwavirinae
Uternevirinae
Aurinkovirinae
Tilevirinae
Nayvirinae
Gunesvirinae
Ghrianvirinae
Taiyangvirinae
Aineviridae
Hathorviridae
Uturviridae
prevalence difference
diseased
control
1.0
0.0
−1.0
0.5

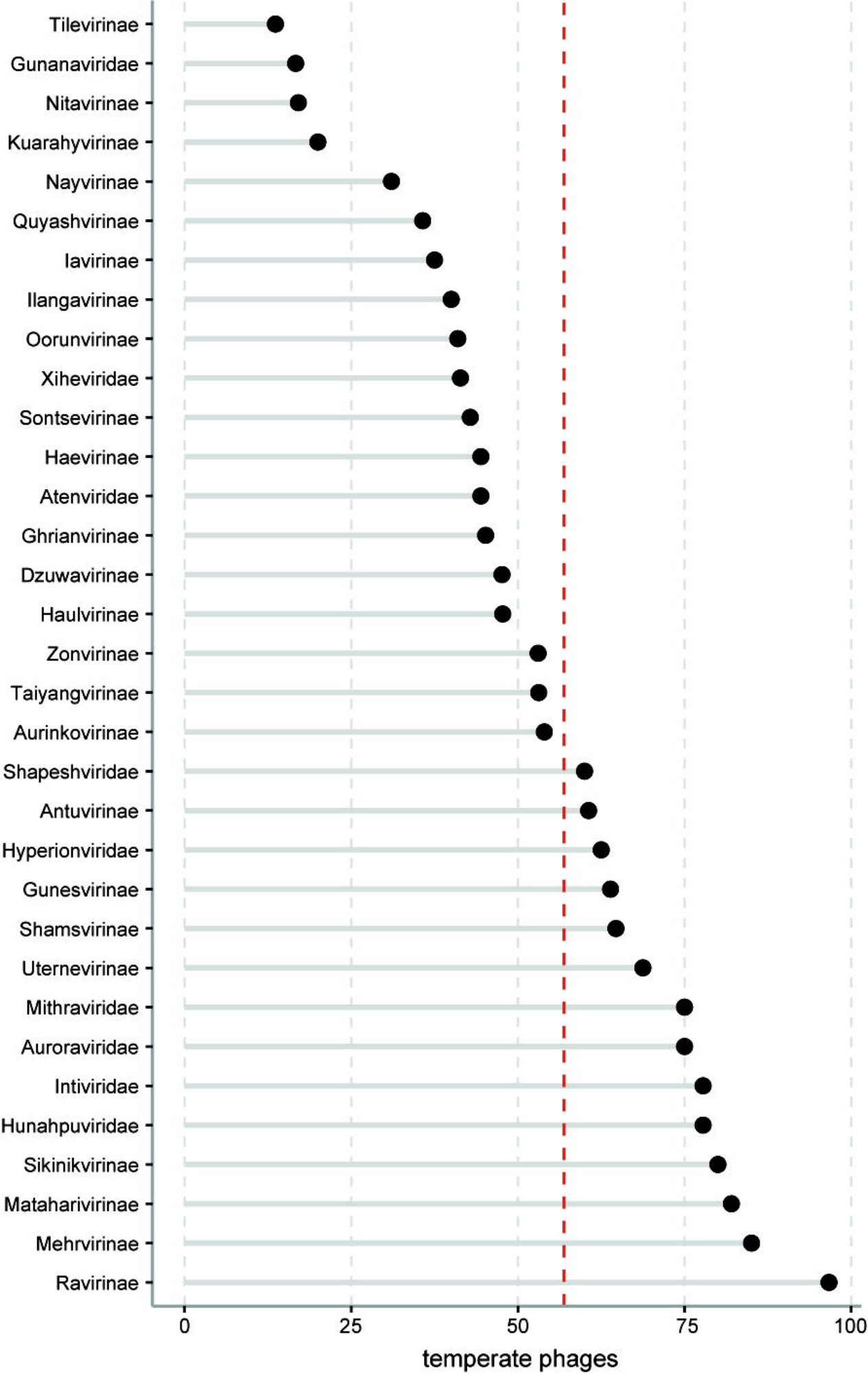

Family
Xiheviridae
Utuviridae
Shapeshviridae
Mithraviridae
Intiviridae
Hyperionviridae
Hunahpuviridae
Hathorviridae
Gunanaviridae
Auroraviridae
Atenviridae
Aineviridae
0
25
50
75
100
percent_shared

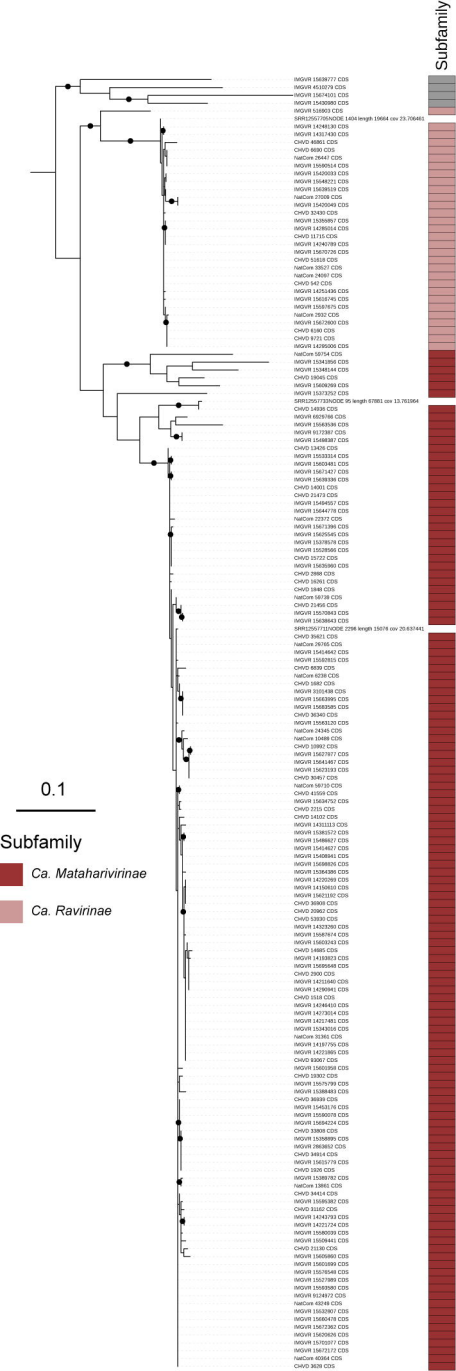

Subfamily
Utuviridae_Tilevirinae
Utuviridae_Taiyangvirinae
Utuviridae_Shamsvirinae
Utuviridae_Quyashvirinae
Utuviridae_Panonpoevirinae
Utuviridae_Oorunvirinae
Utuviridae_Nitavirinae
Utuviridae_Nayvirinae
Utuviridae_Gunesvirinae
Utuviridae_Ghrianvirinae
Utuviridae_Gevirinae
Utuviridae_Dzuwavirinae
Utuviridae_Aurinkovirinae
Utuviridae_Arawvirinae
Utuviridae_Antuvirinae
unclassified_Heliusvirales_Sontsevirinae
unclassified_Heliusvirales_Sikinikvirinae
unclassified_Heliusvirales_Haevirinae
Hathorviridae_Ravirinae
Hathorviridae_Mataharivirinae
Aineviridae_Zonvirinae
Aineviridae_Mehrvirinae
Aineviridae_Ilangavirinae
Aineviridae_Haulvirinae
25
50
75
100
percent_shared

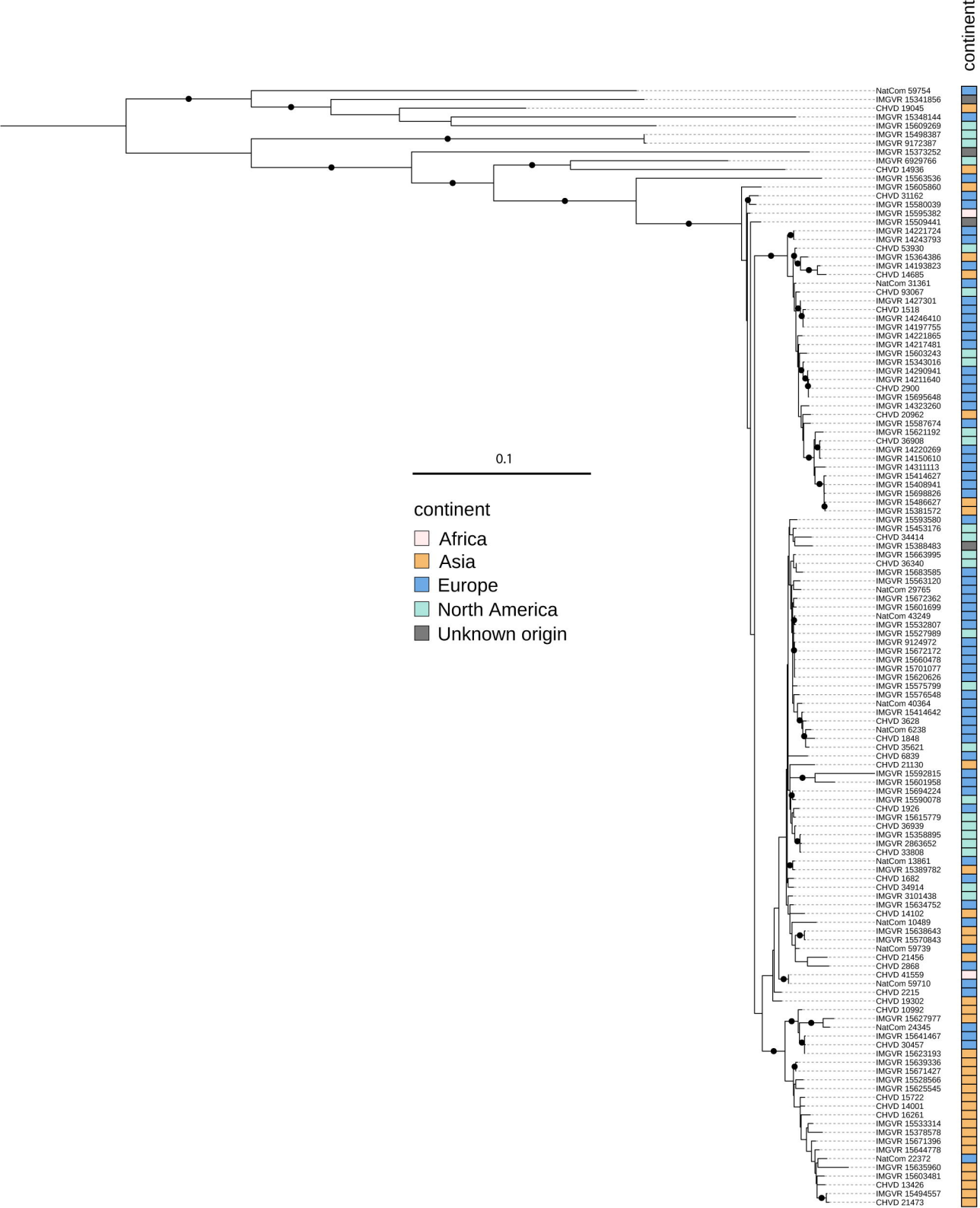

a
Heliusvirales
b
Tree scale: 0.1
Ca. Heliusvirales
ICTV taxa
(other) RefSeq genomes

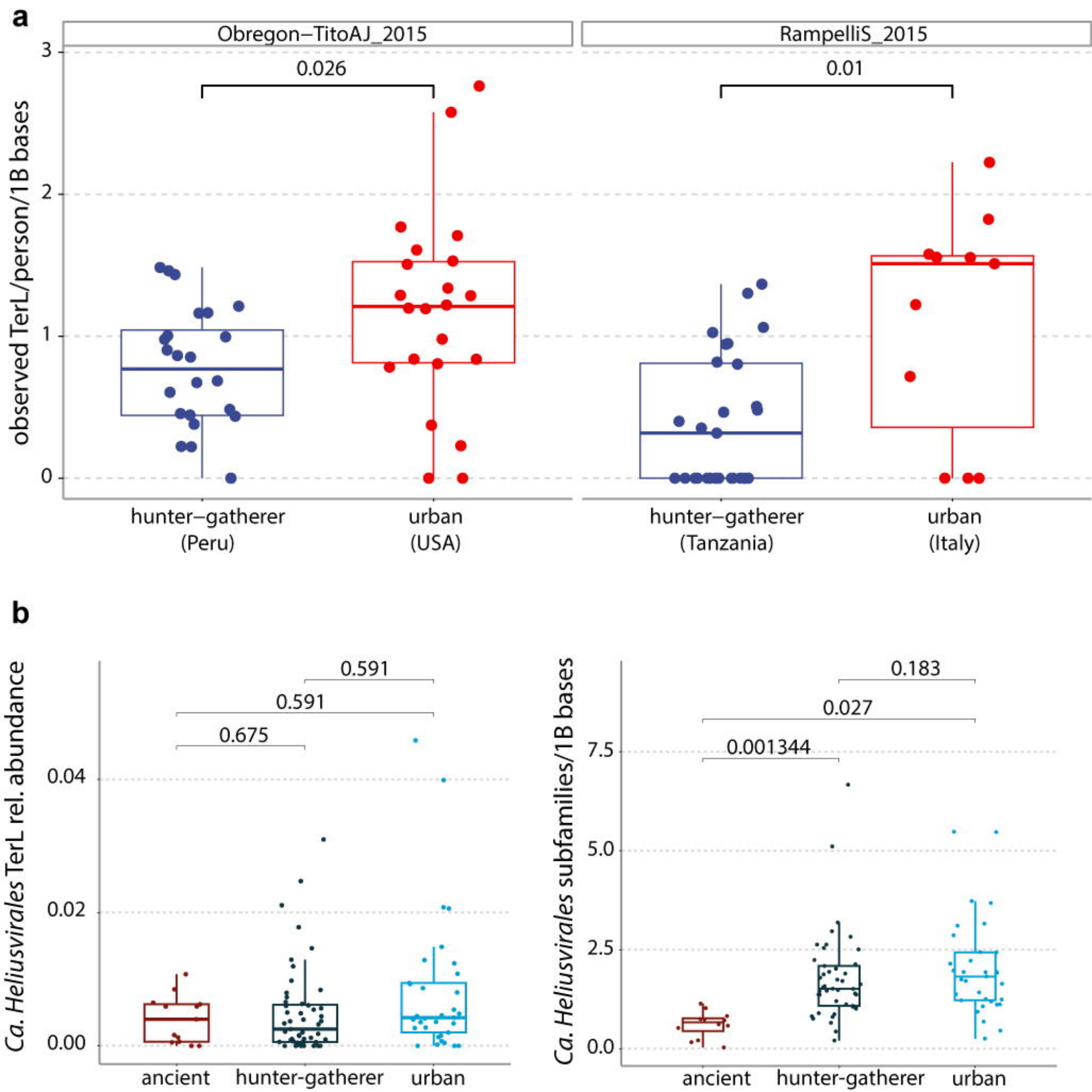

Xiheviridae
Uternevirinae
Tilevirinae
Taiyangvirinae
Shamsvirinae
Quyashvirinae
Oorunvirinae
Nitavirinae
Nayvirinae
Kuarahyvirinae
Iavirinae
Gunesvirinae
Ghrianvirinae
Dzuwavirinae
Aurinkovirinae
Antuvirinae
Sontsevirinae
Sikinikvirinae
Haevirinae
Shapeshviridae
Mithraviridae
Intiviridae
Hyperionviridae
Hunahpuviridae
Ravirinae
Matahariviriae
Gunanaviridae
Auroraviridae
Atenviridae
Zonvirinae
Mehrvirinae
Ilangavirinae
Haulvirinae
30
40
50
60
70
GC content (%)


Aurinkovirinae
Lachnospira sp900316325
Clostridium_AQ innocuum
Pseudoruminococcus massiliensis
Phocaeicola plebeius_A
Bacteroides ovatus
Bacteroides cellulosilyticus
Agathobacter faecis
Roseburia sp900552665
Phascolarctobacterium faecium
Lachnospira sp000437735
Agathobacter rectalis
Streptococcus agalactiae
Anaerobutyricum hallii
Intestinibacter bartlettii
KLE1615 sp900066985
Enterococcus_B lactis
Blautia_A wexlerae_A
Phocaeicola dorei
Lacticaseibacillus paracasei
Finegoldia magna_F
CAG−83 sp001916855
Streptococcus pasteurianus
Enterococcus_B faecium
Enterococcus faecalis
Bacteroides thetaiotaomicron
0
10
20
30
Haulvirinae
Faecalimonas phoceensis
Enterocloster sp001517625
Enterocloster sp000431375
CAG−95 sp000438155
UBA9502 sp003480315
RUG115 sp900066395
Blautia_A sp900066335
Mediterraneibacter faecis
Mediterraneibacter lactaris
Eubacterium_F sp003491505
Coprococcus_A catus
Blautia_A sp900066145
Anaerostipes hadrus
Roseburia sp900552665
Clostridium_Q sp003024715
Fusicatenibacter saccharivorans
Blautia_A wexlerae_A
Blautia_A massiliensis
Agathobacter rectalis
Acetatifactor sp900066365
Blautia_A faecis
Anaerobutyricum hallii
KLE1615 sp900066985
CAG−127 sp900319515
Agathobacter faecis
Phocaeicola dorei
0
10
20
30
40
Mataharivirinae
Dorea_A longicatena
Anaerobutyricum hallii
UMGS1375 sp900066615
Phocaeicola dorei
Mediterraneibacter faecis
Butyrivibrio_A sp000431815
UBA11774 sp003507655
TF01−11 sp003529475
Roseburia sp900552665
Coprococcus sp900066115
TF01−11 sp001916135
TF01−11 sp003524945
Lachnospira eligens_A
CAG−194 sp000432915
Butyrivibrio_A crossotus
Roseburia intestinalis
Coprococcus eutactus_A
Agathobacter rectalis
RUG115 sp900066395
Mediterraneibacter lactaris
Coprococcus eutactus
CAG−95 sp000438155
CAG−127 sp900319515
Blautia_A wexlerae_A
Agathobacter faecis
Acetatifactor sp900066365
0
25
50
75
100
125
Mehrvirinae
Streptococcus gordonii
Roseburia sp900552665
Phocaeicola massiliensis
Ligilactobacillus salivarius
Lactococcus petauri
Enterococcus_A avium
Clostridioides difficile
Blautia_A wexlerae_A
Bacteroides thetaiotaomicron
Bacteroides ovatus
Amedibacterium intestinale
Agathobacter faecis
Paraclostridium sordellii
Enterococcus_D gallinarum
Anaerobutyricum hallii
Streptococcus parasanguinis_B
Phocaeicola dorei
Erysipelatoclostridium ramosum
Enterococcus_B faecium
Faecalibacillus intestinalis
Eubacterium_G ventriosum
Enterococcus faecalis
Streptococcus lutetiensis
Streptococcus agalactiae
Escherichia coli_D
Streptococcus salivarius
Streptococcus pasteurianus
Streptococcus anginosus
0
10
20
30
40
50
Taiyangvirinae
Ruthenibacterium lactatiformans
Mediterraneibacter lactaris
Faecalibacillus intestinalis
Dysosmobacter sp001916835
Bacteroides stercoris
Bacteroides caccae
Ruminococcus_D bicirculans
Phascolarctobacterium faecium
Eubacterium_F sp003491505
CAG−110 sp000434635
Ruminococcus_E bromii_B
Fusicatenibacter saccharivorans
Faecalibacterium prausnitzii_C
Butyrivibrio_A crossotus
Blautia_A wexlerae_A
Lachnospira sp900316325
Anaerobutyricum hallii
Ruminiclostridium_E siraeum
KLE1615 sp900066985
Coprococcus eutactus_A
Clostridium_AQ innocuum
Agathobacter faecis
UBA1417 sp003531055
Parabacteroides merdae
Escherichia coli_D
UBA11524 sp000437595
Agathobacter rectalis
0
10
20
30
40
Zonvirinae
Dialister sp900543455
Clostridium_Q sp003024715
Clostridium_A leptum
CAG−83 sp003539495
CAG−110 sp000434635
Veillonella_A magna
Phascolarctobacterium faecium
Veillonella_A seminalis
Mitsuokella jalaludinii
Megasphaera massiliensis
Acidaminococcus sp000437815
Acidaminococcus sp900538365
Phascolarctobacterium_A succinatutens
Phocaeicola plebeius_A
Acidaminococcus massiliensis
Dialister sp002320515
Dialister sp000434475
Parabacteroides merdae
Caecibacter sp900549385
Dialister succinatiphilus
Megasphaera elsdenii
Dialister hominis
Caecibacter hominis
Phocaeicola dorei
Caecibacter massiliensis
Megasphaera sp000417505
Caecibacter sp003467125
Acidaminococcus intestini
0
10
20
30
number of host predictions
host class
Bacteroidota
Bacteroidales
Firmicutus
Erysipelotrichales
Lactobacillales
Proteobacteria
Enterobacterales
Firmicutus_A
Christensenellales
Lachnospirales
Oscillospirales
Peptostreptococcales
Tissierellales
Firmicutus_C
Acidaminococcales
Selenomonadales
Veillonellales


Tilevirinae
Gunanaviridae
Nitavirinae
Kuarahyvirinae
Nayvirinae
Quyashvirinae
Iavirinae
Ilangavirinae
Oorunvirinae
Xiheviridae
Sontsevirinae
Haevirinae
Atenviridae
Ghrianvirinae
Dzuwavirinae
Haulvirinae
Zonvirinae
Taiyangvirinae
Aurinkovirinae
Shapeshviridae
Antuvirinae
Hyperionviridae
Gunesvirinae
Shamsvirinae
Uternevirinae
Mithraviridae
Auroraviridae
Intiviridae
Hunahpuviridae
Sikinikvirinae
Matahariviriinae
Mehrvirinae
Ravirinae
0
25
50
75
100
temperate phages


Subfamily
IMGVR 15638777 CDS
IMGVR 4510279 CDS
IMGVR 15674101 CDS
IMGVR 15430980 CDS
IMGVR 5189093 CDS
SRR12557705NODE 1404 length 19664 cov 23.706481
IMGVR 14248130 CDS
IMGVR 14317430 CDS
CHVD 48861 CDS
CHVD 8690 CDS
NatCore 26447 CDS
IMGVR 15680514 CDS
IMGVR 15420033 CDS
IMGVR 15644221 CDS
IMGVR 15638519 CDS
NatCore 27008 CDS
IMGVR 15420049 CDS
CHVD 32430 CDS
IMGVR 15359897 CDS
IMGVR 14285014 CDS
CHVD 11715 CDS
IMGVR 14240789 CDS
IMGVR 15670726 CDS
CHVD 51818 CDS
NatCore 33527 CDS
NatCore 24097 CDS
CHVD 542 CDS
IMGVR 14251436 CDS
IMGVR 15616745 CDS
IMGVR 15687675 CDS
NatCore 2932 CDS
IMGVR 15672800 CDS
CHVD 6190 CDS
CHVD 9721 CDS
IMGVR 14395006 CDS
NatCore 50754 CDS
IMGVR 15341856 CDS
IMGVR 15348144 CDS
CHVD 18045 CDS
IMGVR 15608269 CDS
IMGVR 15373252 CDS
SRR12557733NODE 95 length 67881 cov 13.761964
CHVD 14938 CDS
IMGVR 6029768 CDS
IMGVR 15683536 CDS
IMGVR 9172387 CDS
IMGVR 15468387 CDS
CHVD 13426 CDS
IMGVR 15533314 CDS
IMGVR 15603481 CDS
IMGVR 15671427 CDS
IMGVR 15639336 CDS
CHVD 14001 CDS
CHVD 21473 CDS
IMGVR 15494557 CDS
IMGVR 15644778 CDS
NatCore 22372 CDS
IMGVR 15671396 CDS
IMGVR 15625545 CDS
IMGVR 15378578 CDS
IMGVR 15528566 CDS
CHVD 15722 CDS
IMGVR 15636960 CDS
CHVD 2898 CDS
CHVD 18261 CDS
CHVD 1848 CDS
NatCore 58738 CDS
CHVD 21458 CDS
IMGVR 15520843 CDS
IMGVR 15638643 CDS
SRR12557711NODE 2258 length 15076 cov 20.637441
CHVD 35821 CDS
NatCore 26765 CDS
IMGVR 15414642 CDS
IMGVR 15682615 CDS
CHVD 6839 CDS
NatCore 6238 CDS
CHVD 1682 CDS
IMGVR 3101438 CDS
IMGVR 15683995 CDS
IMGVR 15683585 CDS
CHVD 36340 CDS
IMGVR 15563125 CDS
NatCore 24345 CDS
NatCore 10488 CDS
CHVD 10992 CDS
IMGVR 15627877 CDS
IMGVR 15641467 CDS
IMGVR 15623193 CDS
CHVD 30467 CDS
NatCore 58735 CDS
CHVD 41558 CDS
IMGVR 15634752 CDS
CHVD 2215 CDS
CHVD 14102 CDS
IMGVR 14311113 CDS
IMGVR 15281572 CDS
IMGVR 15486627 CDS
IMGVR 15414627 CDS
IMGVR 15408841 CDS
IMGVR 15688826 CDS
IMGVR 15364386 CDS
IMGVR 14320269 CDS
IMGVR 14150810 CDS
IMGVR 15621167 CDS
CHVD 36908 CDS
CHVD 20962 CDS
CHVD 53930 CDS
IMGVR 14323280 CDS
IMGVR 15687674 CDS
IMGVR 15603243 CDS
CHVD 14885 CDS
IMGVR 14183823 CDS
IMGVR 15685648 CDS
CHVD 2900 CDS
IMGVR 14211640 CDS
IMGVR 14290941 CDS
CHVD 1518 CDS
IMGVR 14244410 CDS
IMGVR 14273014 CDS
IMGVR 14217481 CDS
IMGVR 15343016 CDS
NatCore 31361 CDS
IMGVR 14197755 CDS
IMGVR 14221885 CDS
CHVD 93067 CDS
IMGVR 15603958 CDS
CHVD 33902 CDS
IMGVR 15575799 CDS
IMGVR 15388483 CDS
CHVD 36839 CDS
IMGVR 15453176 CDS
IMGVR 15580078 CDS
IMGVR 15684224 CDS
CHVD 33808 CDS
IMGVR 15358899 CDS
IMGVR 2863852 CDS
CHVD 34814 CDS
IMGVR 15615779 CDS
CHVD 1926 CDS
IMGVR 15389782 CDS
NatCore 13861 CDS
CHVD 34414 CDS
IMGVR 15595382 CDS
CHVD 31182 CDS
IMGVR 14343793 CDS
IMGVR 14221724 CDS
IMGVR 15580039 CDS
IMGVR 15609441 CDS
CHVD 21130 CDS
IMGVR 15605860 CDS
IMGVR 15603899 CDS
IMGVR 15578548 CDS
IMGVR 15627989 CDS
IMGVR 15593583 CDS
IMGVR 9124972 CDS
NatCore 43248 CDS
IMGVR 15632907 CDS
IMGVR 15680478 CDS
IMGVR 15627382 CDS
IMGVR 15620626 CDS
IMGVR 15705377 CDS
IMGVR 15672172 CDS
NatCore 40364 CDS
CHVD 3628 CDS
0.1
Subfamily
Ca. Matahariviriae
Ca. Ravirinae


continent
NatCom 59754
IMGVR 15341856
CHVD 19045
IMGVR 15348144
IMGVR 15609269
IMGVR 15498387
IMGVR 9172387
IMGVR 15373252
IMGVR 6929766
CHVD 14936
IMGVR 15563536
IMGVR 15605860
CHVD 31162
IMGVR 15580039
IMGVR 15595382
IMGVR 15509441
IMGVR 14221724
IMGVR 14243793
CHVD 53930
IMGVR 15364386
IMGVR 14193823
CHVD 14685
NatCom 31361
CHVD 93067
IMGVR 1427301
CHVD 1518
IMGVR 14246410
IMGVR 14197755
IMGVR 14221865
IMGVR 14217481
IMGVR 15603243
IMGVR 15343016
IMGVR 14290941
IMGVR 14211640
CHVD 2900
IMGVR 15695648
IMGVR 14323260
CHVD 20962
IMGVR 15587674
IMGVR 15621192
CHVD 36908
IMGVR 14220269
IMGVR 14150610
IMGVR 14311113
IMGVR 15414627
IMGVR 15408941
IMGVR 15698826
IMGVR 15486627
IMGVR 15381572
IMGVR 15593580
IMGVR 15453176
CHVD 34414
IMGVR 15388483
IMGVR 15663995
CHVD 36340
IMGVR 15683585
IMGVR 15563120
NatCom 29765
IMGVR 15672362
IMGVR 15601699
NatCom 43249
IMGVR 15532807
IMGVR 15527989
IMGVR 9124972
IMGVR 15672172
IMGVR 15660478
IMGVR 15701077
IMGVR 15620626
IMGVR 15575799
IMGVR 15576548
NatCom 40364
IMGVR 15414642
CHVD 3628
NatCom 6238
CHVD 1848
CHVD 35621
CHVD 6839
CHVD 21130
IMGVR 15592815
IMGVR 15601958
IMGVR 15694224
IMGVR 15590078
CHVD 1926
IMGVR 15615779
CHVD 36939
IMGVR 15358895
IMGVR 2863652
CHVD 33808
NatCom 13861
IMGVR 15389782
CHVD 1682
CHVD 34914
IMGVR 3101438
IMGVR 15634752
CHVD 14102
NatCom 10489
IMGVR 15638643
IMGVR 15570843
NatCom 59739
CHVD 21456
CHVD 2868
CHVD 41559
NatCom 59710
CHVD 2215
CHVD 19302
CHVD 10992
IMGVR 15627977
NatCom 24345
IMGVR 15641467
CHVD 30457
IMGVR 15623193
IMGVR 15639336
IMGVR 15671427
IMGVR 15528566
IMGVR 15625545
CHVD 15722
CHVD 14001
CHVD 16261
IMGVR 15533314
IMGVR 15378578
IMGVR 15671396
IMGVR 15644778
NatCom 22372
IMGVR 15635960
IMGVR 15603481
CHVD 13426
IMGVR 15494557
CHVD 21473
0.1
continent
Africa
Asia
Europe
North America
Unknown origin


a
Obregon−TitoAJ_2015
RampelliS_2015
observed TerL/person/1B bases
0.026
0.01
hunter−gatherer (Peru)
urban (USA)
hunter−gatherer (Tanzania)
urban (Italy)
b
Ca. Heliusvirales TerL rel. abundance
0.591
0.591
0.675
ancient
hunter-gatherer
urban
Ca. Heliusvirales subfamilies/1B bases
0.183
0.027
0.001344
ancient
hunter-gatherer
urban