Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

Article

The Missing Tailed Phages: Prediction of Small Capsid Candidates

Antoni Luque[1,2,3,*], Sean Benler[4], Diana Lee [1,2], Colin Brown[1,5], and Simon White[6]

[1] Viral Information Institute, San Diego State University, San Diego, CA, USA.

[2] Computational Science Research Center, San Diego State University, San Diego, USA.

[3] Department of Mathematics and Statistics, San Diego State University, San Diego, USA.

[4] National Center for Biotechnology Information (NCBI), Bethesda, MD, USA.

[5] Department of Physics, San Diego State University, San Diego, USA.

[6] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA.

* Correspondence: aluque@sdsu.edu.

© 2020 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

Abstract: Tailed phages are the most abundant and diverse group of viruses on the planet. Yet, the smallest tailed phages display relatively complex capsids and large genomes compared to other viruses. The lack of tailed phages forming the common icosahedral capsid architectures T = 1 and T = 3 is puzzling. Here, we extracted geometrical features from high-resolution tailed phage capsid reconstructions and built a statistical model based on physical principles to predict the capsid diameter and genome length of the missing small tailed phage capsids. We applied the model to 3,348 isolated tailed phage genomes and 1,496 gut metagenome-assembled tailed phage genomes. Four isolated tailed phages were predicted to form T = 3 icosahedral capsids, and twenty-one metagenome-assembled tailed phages were predicted to form T < 3 capsids. The smallest capsid predicted was a T = 4/3 ≈ 1.33 architecture. No tailed phages were predicted to form the smallest icosahedral architecture, T = 1. We discuss the feasibility of the missing T = 1 tailed phage capsids and the implications of isolating and characterizing small tailed phages for viral evolution and phage therapy.

Keywords: Bacteriophage**,** tailed phages; icosahedral capsids; capsid modeling; statistical learning; isolated genomes; metagenome-assembled genomes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

1. Introduction

Tailed phages are viruses that infect bacteria and are the most abundant biological entity on Earth [1]. They are responsible for the regulation of biogeochemical processes at a planetary scale [2,3], the control of microbial populations [4,5], and the mobility of genes across hosts and ecosystems [6,7]. The broad functionality of tailed phages is facilitated by their vast reservoir of genes, which results in a large range of genome lengths, from 10 kilobase pairs (kbp) to 500 kbp [8,9]. Tailed phages can accommodate these genomes because the proteins building their protective capsids display a high degree of plasticity, allowing for a broad range of size capsids, from about 30 nm to 160 nm in diameter [10]. The majority of tailed phages, about 80–90%, form capsids with icosahedral symmetry [11], while the remaining 10–20% form elongated capsids with icosahedral caps [12,13]. The formation of icosahedral capsids is not unique to tailed phages; it is the most common viral capsid architecture in the virosphere [14]. However, despite the abundance and diversity of tailed phages, even the smallest tailed phages form far more complex capsids and store far larger genomes than most icosahedral viral families [15]. Furthermore, the major capsid proteins of tailed phages adopt the HK97-fold, which is also found in the building blocks of bacterial and archeal enzymatic cellular nanocompartments called encapsulins [16,17]. Intriguingly, encapsulins can form the smallest icosahedral protein shells that are absent among tailed phages (Figure 1).

Icosahedral capsids are characterized by the triangulation number T, which determines the number of quasi-equivalent proteins (complexity) and the total number of proteins forming the capsid [18,19]. As illustrated in Figure 2, icosahedral capsids can organize their proteins in four different lattices, accommodating different stoichiometries of major and minor capsid proteins [19]. The generalized T-number is $T_i(h,k) = \alpha_i\, T_0(h,k)$, where h and k are the steps in the hexagonal sublattice joining two consecutive vertices in the icosahedral capsid. T_0 is the classic T-number associated with the hexagonal lattice and is given by the equation $T_0(h,k) = h^2 + hk + k^2$. The subindex i takes the values h, t, s, and r, associated, respectively, with the hexagonal, trihexagonal, snub hexagonal, and rhombitrihexagonal icosahedral lattices. The number of major capsid proteins

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

in a capsid is $60T_0(h,k)$. The hexagonal lattice case only contains major capsid proteins, that is, $T_h(h,k) = T_0(h,k)$ with $\alpha_h = 1$. If the size of the major capsid protein is conserved across lattices, the other three lattices must include minor capsid proteins occupying the secondary polygons (triangles and squares). This increases the surface of the capsid by a factor $\alpha_t = 4/3 \approx 1.33$ (trihexagonal), $\alpha_s = 7/3 \approx 2.33$ (snub hexagonal), and $\alpha_r = 4/3 + 2/\sqrt{3} \approx 2.49$ (rhombitrihexagonal). When combining all lattices, the first four elements of the generalized T-number are T = 1, 1.33, 2.33, and 2.49, containing 60 major capsid proteins each. The fifth element of the series is T = 3 and contains 180 major capsid proteins. Tailed phages have been observed to form capsids adopting the hexagonal and trihexagonal lattices [19,20]. But no tailed phages have been observed to form $T \leq 3$ capsids (Figure 1). The smallest characterized tailed phage structure corresponds to the *Bacillus* phage phi29, which adopts an elongated structure with icosahedral T = 3 caps and a Q = 5 body [12,21,22], and *Streptococcus* phage C1, which adopts a T = 4 icosahedral capsid [10,23].

There are several interrelated scientific observations that make the lack of small icosahedral capsids striking among tailed phages. First, T = 1 and T = 3 icosahedral capsid architectures are optimal thermodynamic configurations for protein shells and are the most kinetically favorable icosahedral capsids [13,24–27]. T = 1 and T = 3 are the most common architectures imaged among viruses [15]. Second, tailed phages belong to the *Duplodnaviria* viral realm, which also includes archaeal and eukaryotic viruses [28,29]. These viruses are characterized by building their capsids with a major capsid protein adopting the highly conserved HK97-fold, and packing their genome in the form of double-stranded DNA at quasi-crystalline densities (Figure 1). The structural elements of this realm appear to have emerged prior to the Last Universal Common Ancestor [30]. Modern tailed phages forming T = 4 and larger capsid shells are expected to have evolved from precursor capsids that used T = 1 and T = 3 architectures. Third, encapsulins organize the HK97-fold in T = 1, T = 3, and T = 4 architectures, showing that the HK97-fold is capable of forming small icosahedral capsids [16,17] (Figure 1). Finally, non-tailed icosahedral phages in the *Microviridae* family adopt T = 1 capsids, storing a small genome capable of executing a lytic replication strategy analogous to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

tailed phages [31,32]. This suggests that tailed phages should be able to store the instructions for their lytic lifestyle in T = 1 capsids.

This body of evidence indicates that T = 1 and T = 3 should exist among tailed phages. Our working hypothesis is that these small tailed phages have not been isolated because they are relatively low in abundance in most environments sampled. Larger tailed phage capsids capable of storing longer genomes may have been favored by incorporating genes that can alter the host's physiology, protect the bacterial host against other phages, and overcome the host's resistance mechanisms [6,33–35]. Here, we applied modeling and bioinformatics to narrow the search for small tailed phages, overcoming the current limitations of sampling. As a proof of concept, we analyzed the capsid architecture and genome size of twenty-three tailed phage that have had high-resolution structures of their capsid determined. We used an allometric model based on conserved structural characteristics of tailed phages and trained the model to infer the genome size and capsid diameter size of tailed phages as a function of the T-number icosahedral architecture. The predictions of the model were compared with more than 3,348 isolated phage genomes and 1,496 gut metagenome-assembled tailed phage genomes. Four isolated tailed phages and twenty-one metagenome-assembled tailed phages were predicted to form, respectively, T = 3 and T< 3 icosahedral capsids architectures, currently missing among tailed phages.

2. Materials and Methods

Structural features of tailed phage capsids. High-resolution tailed phage structures were obtained from twenty-three cryo-EM maps [23,36–57]. UCSF Chimera software was used to visualize the structures and measure the geometrical properties of the capsids [58]. The icosahedral tool was fitted manually to the internal and external surface of the capsids by adjusting the radius and sphericity properties. Error measurements were assigned by comparing the outputs at one discrete step below and above the final measurement. The direct measurements were the internal radius, sphericity, surface, and volume, as well as external radius, sphericity, surface, and volume.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

The shell thickness was estimated by subtracting the internal radius from the external radius. The number of major capsid proteins was calculated using the icosahedral T-number associated with each capsid (Suhanovsky and Teschke, 2015; Caspar and Klug, 1962; Twarock and Luque, 2019). The internal and external major capsid protein surfaces were obtained by dividing the fraction of internal and external surfaces associated with major capsid proteins with respect to the total number of major capsid proteins. The genome density was estimated by dividing the genome size by the internal volume. The Electron Microscopy Data Base ID (EMDB) and measurements obtained for each tailed phage are provided in Data File 1. The correlation with the capsid diameter of all variables measured was evaluated as a function of the interior and external capsid diameters using the non-parametric Spearman coefficient correlation and a two-tailed test. The normality of the variables that were not correlated with capsid diameter was assessed using the Shapiro-Wilk test.

Statistical model. Allometric models were used to relate the genome length (G) and the capsid diameter (D) as a function of the capsid architecture, T-number. To reduce the bias associated with the oversampling of some T architectures, such as T = 7 capsids, the model was built using the average genome lengths and capsid diameters for each T-number. The models were, respectively, $G(T) = a_G T^{b_G}$ and $D(T) = a_D T^{b_D}$. Each model had two parameters: the pre-factor a_i and the allometric exponent b_i . The subindex i took the values G for the genome length model and D for the capsid diameter model. The variables were logged on base 10 to obtain the best fit using a linear regression least-squares approach for the parameters $\log(a_i)$ and b_i. The residuals were analyzed to evaluate the accuracy of the standard errors and confidence intervals extracted from the linear regression analysis, which relied on the standard assumptions of linearity, normality, homoscedasticity, and low leverage (Kutner et al. 2004; Witten et al. 2013). The linear statistical analysis and predictions were performed using the *lm* and *predict* functions in the statistical computing language R [59].

3D icosahedral capsid models. Small capsid 3D model architectures with T-number T ≤ 4 were generated for hexagonal (h), trihexagonal (t), snub hexagonal (s), and rhombitrihexagonal (r) icosahedral lattices. These Archimedean lattices minimize the structural information necessary for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

building icosahedral capsids and are the basis for the generalized theory of icosahedral capsids [19]. The 3D geometrical models were generated computationally using the new prototype of the hkcage bundle in ChimeraX, which can produce any element of the infinite series of icosahedral capsids for the four fundamental lattices and their associated dual Laves lattices [60].

Predicted capsid architectures among tailed phage isolates. The genomes of isolated tailed phages were obtained from the NCBI Reference Sequence Database [61]. The genome information was filtered by the viral realm (*Duplodnaviria*) and host (bacteria). The database was accessed in October 2020 and downloaded in table format. For 95% of the tailed phages, the NCBI id and genome length were extracted using a Unix shell script. The remaining 5% of tailed phages had information that was not standardized with the NCBI format and was extracted manually. The final file analyzed included the NCBI identifiers and genome lengths (Data File 2). The probability density distribution of the genome lengths was obtained using Gaussian kernels with the default bandwidth for the *density* function in the statistical computing language R [59]. The predicted icosahedral architecture for each tailed phage was obtained by identifying the nearest average genome length obtained in the statistical T-number model described above. A frequency analysis of the predicted T-numbers was obtained. The tailed phages associated with architectures $T \leq 4$ were filtered and analyzed in depth.

Predicted structures of metagenome-assembled tailed phages. The genomes of uncultured human gut tailed phages were obtained from ftp://ftp.ncbi.nih.gov/pub/yutinn/benler_2020/. Genomes smaller than 10 kb were checked for potential errors in assembly as follows. The largest parent metagenome for each phage genome was downloaded from the NCBI SRA database using fasterq-dump (v. 2.10.8). The raw sequencing reads were mapped to the genome using Bowtie2 set with default parameters [62]. Phage genomes with fewer than five paired-end reads aligned across the 5′ and 3′ contig termini were considered misassembled, and all others considered complete. The contig identifier and genome length of the selected gut metagenome-assembled tailed phages are provided in Data File 3. The analysis of the distribution of complete circular genomes and predicted

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

T-number architectures was conducted following the same protocol described above for isolated tailed phage genomes.

3. Results

3.1. Structural properties of characterized tailed phage capsids

The range of structural properties obtained from the twenty-three capsids studied is summarized in Table 1. The T-number ranged from 4 to 27, containing 240 to 1620 major capsid proteins, although five of those major capsid proteins are replaced by portal proteins in the virion. The external capsid surface was relatively angular, with sphericity ranging from 0.14 to 0.55. The interior capsid surfaces were more spherical, with sphericity ranging from 0.25 to 0.75. The capsid sphericity was negatively correlated with capsid diameter (rho = –0.68, p-value < 0.001). The largest capsid had a diameter three times bigger than the smallest capsid (143 nm versus 49 nm). The capsid thickness ranged from 3 to 8 nm and was positively correlated with capsid diameter (rho = 0.66, p-value = 0.0010). The largest interior capsid volume was about 25 times larger than the smallest one ($3.36 \cdot 10^4$ to $8.22 \cdot 10^5$ nm^3 range). The largest genome (280 kbp) was about 16 times larger than the smallest one (17 kbp). The average genome packing density within the capsid was around ~0.5 bp/nm^3, ranging from 0.34 to 0.62 bp/nm^3. The values for the interior and exterior capsid surface areas spanned 8.5-fold ($5.08 \cdot 10^3$–$4.32 \cdot 10^4$ nm^2) and 8-fold ($6.55 \cdot 10^3$–$5.19 \cdot 10^4$ nm^2), respectively, with the exterior capsid surface area being 15% to 43% larger than the interior area. The average surface area for each major capsid protein in the interior and exterior parts of the capsid was ~23 nm^2 (19 to 26 nm^2) and ~30 nm^2 (24 to 35 nm^2) per major capsid protein, respectively. The genome density (rho = –0.16, p-value = 0.47) and major capsid exterior surface area (rho = 0.38, p-value = 0.082) were the only variables that did not display a significant correlation with capsid size (see Supplementary Table 1).

3.2. Tailed phage capsids: statistical models and predictions

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

The genome lengths and capsid diameters obtained from the high-resolution phage structures were used to build a statistical model relating the icosahedral capsid architecture with these two accessible quantities. The allometric model for the average genome length, G, as a function of the architecture index, T, explained 98.5% (R^2=0.985, n = 7) of the variance (Figure 3**a**). The pre-factor was $\log a_G = 0.37 \pm 0.10$ (S.E.) with a_G in kbp units. The allometric exponent was $b_G = 1.47 \pm 0.09$. The coefficient of variation (CV = S.E./mean) for the intercept was 27%, significantly larger than the CV associated with the power exponent, 6.1%. Similarly, the allometric model for the average capsid diameter, D, as a function of the architecture index, T, explained 98.6% (R^2=0.985, n = 7) of the variance (Figure 3**b**). The pre-factor was $\log a_D = 1.38 \pm 0.34$ (S.E.) with a_D in nm units and coefficient of variation (CV) of 25%. The allometric exponent was $b_D = 0.52 \pm 0.03$ with a CV of 5.8%. The qualitative diagnostic of the residuals was similar for both models, consistent with the standard assumptions associated with the statistics of the linear regression analysis (Figures S1 and S2). The residuals were scattered around zero with a near-normal distribution and a standardized range on the order of ±1 with relatively homoscedastic variance and relatively low leverage.

The allometric exponents obtained in the statistical models were compared with theoretical scaling relationships. The capsid structural analysis revealed that the genome density was constant. This implied that the genome length (G) was proportional to the capsid volume (V). The volume of a quasi-spherical capsid depends on the third power of the diameter (D), leading to the scaling $G \sim D^3$ between the genome length and capsid diameter. The T-number, by definition, is proportional to the capsid surface, and the units are implicitly related to the major capsid protein surface [19,63]. The exposed surface of the major capsid protein obtained in the structural analysis was constant. Since the capsid surface of a quasi-spherical shell depends on the square of the diameter, this leads to the scaling $T \sim D^2$ relating to the T-number and the capsid diameter. The scaling relationships derived led to the allometric relationships $G \sim T^{3/2}$ and $D \sim T^{1/2}$. The theoretical exponent for the genome length versus capsid architecture relationship was $b_G^{th} = 3/2 = 1.5$, which was within the empirical range obtained in the statistical model, $b_G = 1.47 \pm 0.09$ (Figure 3**a**). The theoretical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

exponent for the capsid diameter versus capsid architecture yielded $b_D^{th} = 1/2 = 0.5$, which was also within the empirical range obtained in the statistical model $b_D = 0.52 \pm 0.03$ (Figure 3**b**). The agreement between the statistical model and the theoretical allometric exponents provides confidence in using the statistical model to predict the properties of capsids for T-number architectures outside the range used to train the statistical model.

The average genome length predicted for small tailed phage architectures was 2.34 kbp (T = 1), 3.57 kbp (T ≈ 1.33), 8.14 kbp (T ≈ 2.33), 8.94 kbp (T ≈ 2.49), 11.78 kbp (T = 3), and 17.98 kbp (T = 4). The capsid architectures and 95% confidence intervals for the genomes are shown in Figure 4. The average capsid diameter predicted for these architectures was 24.20 nm (T = 1), 28.09 nm (T ≈ 1.33), 37.54 nm (T ≈ 2.33), 38.81 nm (T ≈ 2.49), 42.76 nm (T = 3), and 49.63 nm (T = 4). The 95% confidence intervals are also displayed in Figure 4. The confidence intervals for the predicted genome lengths and capsid diameters of these small capsids were relatively large and overlapped. This was a consequence of the relatively large coefficient of variation of the intercepts in statistical models due to the limited number of different T-number capsid architectures used to generate the models (n = 7).

3.3. Putative small tailed phage candidates

3.2.1. Predictions from isolated tailed phages

The genome lengths of 3,348 isolated tailed phages were investigated. The distribution of genome lengths was multimodal (Figure 5**a**). The most dominant peak was at 42.0 kbp. The second dominant peak was among the largest genomes with a length of 158.9 kbp. The shortest genomes displayed a minor peak at 18.3 kbp. The minimum genome length was 11.6 kbp, while the maximum was 497.5 kbp. The median genome length was 50.3 kbp, while the mean was 72.7 kbp. The associated icosahedral structures predicted ranged from T = 3 to T = 39 (Figure 5**b**). The median

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

architecture was at T ≈ 7.46, and the mean architecture value was T ≈ 9.95. The three genome length peaks were associated with predicted T ~ 4 like structures (18.3 kbp peak), T ~ 7 like structures (42.0 kbp peak), and T ~ 16–19 like structures (158.9 kbp peak). Among the smallest tailed phages, seventy-seven were predicted to form T = 4 architectures, and six were initially predicted to form T = 3 architectures. The identifiers for the isolated tailed phages predicted to form T ≤ 4 capsid architectures are provided in Data File 4.

3.2.2. Predictions from metagenome-assembled tailed phages

The genome lengths of 1,496 metagenome-assembled tailed phages obtained from gut metagenomes were investigated. The distribution of genome lengths was also multimodal (Figure 5**c**). The most dominant peak was at 42.9 kbp, similar to the dominant peak among isolated genomes. The second dominant peak was at 98.2 kbp, which was absent in the isolated genomes. The third dominant peak was among the largest genomes with a length of 160.8 kbp, similar to the isolated genomes. The shortest genomes displayed a minor peak at 12.6 kbp, which was slightly shorter in length compared to the analogous peak among isolated genomes. The minimum genome length was 4.5 kbp, while the maximum was 294.5 kbp. The median genome length was 44.8 kbp, while the mean was 55.8 kbp. The associated icosahedral structures predicted ranged from T ≈ 1.33 to T = 27 (Figure 5**d**). The median architecture was at T ≈ 7.46, and the mean architecture value was T ≈ 8.34. The four genome length peaks were associated with predicted T ~ 3 like structures (12.6 kbp peak), T ~ 7 like structures (42.9 kbp peak), T ~ 12-13 like structures (98.2 kbp peak), and T ~ 16–19 like structures (160.8 kbp peak). Among the smallest tailed phages, two were predicted to form T ≈ 1.33 architectures, nine were predicted to form T ≈ 1.33 capsids, nine were predicted to form T ≈ 2.33 architectures, ten were predicted to form T ≈ 2.49 architectures, forty-three were predicted to form T = 3 architectures, and forty-one were predicted to form T = 4 architectures. The identifiers for the metagenome-assembled tailed phages predicted to form T ≤ 4 capsid architectures are provided in Data File 5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

3.2.3. Small tailed phage capsid candidates

Six entries in the isolated tailed phage database were predicted to form T = 3 capsid architectures. This was the smallest architecture predicted for this group. However, only four out of those six were kept in the pool of predicted T = 3 phage capsids after further scrutiny. *Enterobacteria phage P4* (NC_001609 and genome length 11.6 kbp) was discarded because it is a satellite phage that relies on the infection and genes of phage P2 to produce particles, forming a larger capsid than predicted, T = 4 [64]. *Enterobacteria phage BF23* was discarded because the entry identified in NCBI (NC_042564) was associated with its tRNA gene region (gene length 14.5 kbp). The complete genome was not deposited, but the phage is T5-like, encoding a 100-120 kbp genome. The remaining four were complete phage genomes and infected hosts from four different phyla (Figure 6**a**). *Salmonella* phage astrithr (11.6 kbp, NC_48862) was a *Podoviridae* infecting *Salmonella enterica* in the phylum *Proteobacteria. Rhodococcus* phage RRH1 (14.3 kbp, NC_016651) was a *Siphoviridae* infecting *Rhodococcus rhodochrous* in the phylum *Firmicutes*. The two more distant phages were *Lactococcus* phage bIL311 (14.5 kbp, NC_002670), a *Siphoviridae* infecting *Lactococcus lactic* in the phylum *Actinobacteria*, and *Mycoplasma* virus P1 (11.7 kbp, NC_002515), a *Podoviridae* infecting *Mycoplasma pulmonis* in the phylum *Tenericutes*.

Twenty-one metagenome-assembled tailed phages analyzed were predicted to form T ≤ 3 capsid architectures. No candidates, however, were predicted to form the smallest icosahedral architecture, T = 1 (Figure 6**b**). The 3D capsid models and labels of the contigs associated with T ≈ 1.33 (two contigs), T ≈ 2.33 (nine contigs), and T ≈ 2.49 (ten contigs) are displayed in Figure 6**b**. The genome architectures were diverse (Supplementary Figure S3). The genomes associated with T ≈ 1.33 did not encode major capsid proteins. The smallest genome containing a major capsid protein was OLNE01004159.1, which had a genome length of 7.4 kbp and a predicted architecture of T ≈ 2.33 (Figure 6**c**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

4. Discussion

Our hypothesis was that tailed phages adopting small icosahedral structures do exist, but their low abundance in the environment has precluded isolating them to be characterized in high-resolution capsid reconstruction studies. The modeling and bioinformatic approach introduced here supports this hypothesis. The allometric models trained using tailed phage with high-resolution structures predicted capsids among isolated and metagenome-assembled tailed phages that were smaller than the icosahedral tailed phages characterized to date. Among isolated tailed phages, we predicted four T = 3 icosahedral capsids. Among metagenome-assembled tailed phages, the study predicted twenty-one potential T < 3 architectures, including two T = 4/3 ≈ 1.33 icosahedral capsids.

The four isolated phages predicted to form T = 3 capsid architectures infected hosts from four distantly related phyla, suggesting that small tailed phage capsids might be prevalent across bacterial hosts. The two isolated phages with the smallest genomes adopting T = 3 capsid architectures were *Podoviridae*: *Salmonella* phage astrithr (11.6 kbp, NC_48862) and *Mycoplasma* virus P1 (11.7 kbp, NC_002515). Their genome lengths were very similar to the predicted average genome length for T = 3 tailed phage capsids, 11.8 kbp. The other two phages were *Siphoviridae* and had genomes ~ 3 kbp longer: *Rhodococcus* phage RRH1 (14.3 kbp, NC_016651) and *Lactococcus* phage bIL311 (14.5 kbp, NC_002670). Their genome lengths were well within the confidence limit predicted for T = 3 tailed phage capsids, 8.1 – 17.1 kbp. But their genome lengths also overlapped with the lower limit predicted for T = 4 tailed phage capsids, 13.2 – 24.5 kbp range. It is not unlikely, thus, that these two phages could adopt instead T = 4 instead of T = 3 capsids.

Among the gut metagenome-assembled tailed phages, the putative phages OGQL01007720.1 (4.5 kbp) and OMEC01003054.1 (5.4 kbp) were predicted to adopt T = 4/3 ≈ 1.33 icosahedral capsids. The trihexagonal lattice associated with T = 4/3 ≈ 1.33 has been observed among higher T-numbers for tailed phages [19,20]. However, no major capsid protein or portal protein was annotated in those

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

two genomes (Supplementary Figure S3). These two phages, thus, could be satellites of other phages, adopting larger capsids, as occurs with *Enterobacteria phage P4* (NC_001609 and genome length 11.6 kbp), which is a satellite phage of P2. Phage P4 was predicted to adopt a T = 3 capsid, but it forms a T = 4 capsid [64].

Nine metagenome-assembled tailed phages that displayed genomes from 7.3 kbp to 8.5 kbp and were predicted to adopt a T = 7/3 ≈ 2.33 icosahedral capsid (Figure 6**b**). This group of diverse phages encoded major capsid proteins (Supplementary Figure S3), providing confidence to this prediction. Nonetheless, the capsid predicted adopted a snub hexagonal lattice, which has not been observed among larger tailed phages [19]. The same applies to the next group of ten tailed phages, which were predicted to adopt the $T = 4/3 + 2/\sqrt{3} \approx 2.49$ capsid architecture. This capsid is associated with the rhombitrihexagonal lattice, which has not been observed among tailed phages either.

If these predictions are confirmed, these would be the first viruses known to adopt regular snub hexagonal and rhomitrihexagonal lattices [19]. Alternatively, these two groups of phages could form elongated T = 1 capsids or icosahedral T = 3 capsids because the genome length of these phages is close to the lower 95% confidence predicted for T = 3 capsids, 8.1 kbp (Figure 4).

No tailed phages were predicted to adopt the smallest icosahedral capsid, T = 1. Additional studies will be necessary to determine if such small tailed phage capsids exist. The average genome predicted to adopt a T = 1 tailed phage capsid was 2.3 kbp, with a lower and upper 95% confidence of 1.2 kbp and 4.4 kbp, respectively. This range was consistent with small icosahedral DNA viruses in the *Monodnaviria* realm, which store the genome as single-stranded DNA. Phages in the *Circoviridae* and *Microviridae* families, for example, form T = 1 capsids and store genomes ranging, respectively, from 1.8 to 3.8 kb and from 4.4 to 6.1 kb [65]. The lower range of the *Microviridae* viruses is particularly appealing because these viruses follow an analogous life cycle as tailed phages: translocating the genome through the host cell upon infection, assembling an empty procapsid, packing the genome to form the mature capsid, and lysing the host to release the new

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

mature virions [31,32]. The absence of a tail in *Microviridae* viruses suggest that tailed phages in the *Podoviridae* family might be more likely to form T = 1 capsid architectures or small phage capsids. This is consistent with the fact that the smallest isolated tailed phages predicted to form T = 3 capsids were *Podoviridae*.

It has been observed that the genome length of viruses, bacteria, and archaea are smaller at higher-temperatures [66–68]. This suggests that T ≤ 3 architectures would be more likely to be present in hot environments, such as hydrothermal vents, which can reach temperatures of ~100°C or higher. Warm-blooded animals may also be a good source of small tailed phage capsids. The temperature is lower than in hydrothermal vents but higher than in most other environments. Due to medical and economic reasons, metagenomes from warm-animal sources are more accessible and abundant, increasing the odds of finding small tailed phage capsids. This is consistent with our finding of small capsids predicted among human gut tailed phages.

The discovery and study of small tailed phage capsids would help interrogate the origin of the ancient *Duplodnaviria* realm and the origin of the last universal common ancestor for phages [30]. The assumption is that the smallest icosahedral capsids were precursors of tailed phage capsids. Therefore, the molecular information and divergence of small tailed phages could provide insightful clues on how viruses emerged on Earth. These small tailed phages could also help uncover a potential evolutionary relationship between tailed phages and cellular compartments like encapsulins [14,16,17].

The discovery and characterization of small tailed phages would also have important biomedical implications. Their shorter genomes would contain fewer genes. It is expected that most of the genes would be involved in structural functions. The conserved folds of these proteins, such as the major capsid protein and the portal, would help narrow the functions of genes in these small genomes. This would facilitate the annotation and functional characterization of these phages, circumventing one of the main issues of phage therapy nowadays. Even in scenarios that phages

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

have been used successfully as therapeutics, the number of unknown gene functions encoded in those phage genomes was significant and may be involved in increasing bacterial virulence [69,70]. Small tailed phages would make the application of phage therapy more predictable and easier to regulate.

5. Conclusions

Small tailed phages with icosahedral architectures $T \leq 3$ have not yet been observed by high-resolution imaging techniques. Here we proposed that these structures exist in the environment but are challenging to sample due to low abundances. The modeling and bioinformatic approach applied here predicted small icosahedral capsids among isolated and metagenome-assembled tailed phages. The smallest capsid predicted was a trihexagonal $T = 4/3 \approx 1.33$. No candidate was found to form the smallest icosahedral capsid $T = 1$, but we proposed that *Podoviridae* in high-temperature environments, like warm-blooded animals and hydrothermal vents, would be potential candidates for the missing $T = 1$ tailed phage capsids. The discovery of these small tailed phages would be transformative for the study of viral evolution as well as biomedical and biotechnological applications.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Residual diagnostics for the genome length model, Figure S2: Residual diagnostics for the capsid diameter model, Figure S3: Annotation of small metagenome-assembled tailed phage genomes. Table S1: Correlation analysis for the capsid structural properties as a function of the capsid diameter. Data File 1: Geometrical measurements of high-resolution tailed phage capsids. Data File 2: NCBI identifiers and genome length for associated isolated genomes. Data File 3: Contig identifiers and genome lengths of gut metagenome-assembled tailed phage genomes. Data File 4: Associated isolated tailed phages predicted to form $T \leq 3$ capsid architectures. Data File 5: Associated gut metagenome-assembled tailed phages predicted to form $T \leq 3$ capsid architectures.

Author Contributions: Conceptualization, A.L. and S.W.; methodology, A.L, D.L, S.B.; software, A.L., C.B., S.B.; validation, A.L., S.B., S.W.; formal analysis, A.L.; investigation, A.L.; resources, A.L. and S.B.; data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

curation, A.L.; writing—original draft preparation, A.L.; writing—review and editing, S.W., S.B., D.L., C.B.; visualization, A.L. and C.B. All authors have read and agreed to the published version of the manuscript.

Funding: A.L. acknowledges that this research was funded by the National Science Foundation Award 1951678 in the Division of Mathematical Sciences. S.B. is supported by the Intramural Research Program of the National Institutes of Health (National Library of Medicine).

Acknowledgments: Molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

References

1. Cobián Güemes, A.G.; Youle, M.; Cantú, V.A.; Felts, B.; Nulton, J.; Rohwer, F. Viruses as winners in the game of life. *Annu. Rev. Virol.* **2016**, *3*, 197–214, doi:https://doi.org/10.1146/annurev-virology-100114-054952.

2. Wommack, K.E.; Colwell, R.R. Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **2000**, *64*, 69–114, doi:10.1128/MMBR.64.1.69-114.2000.

3. Danovaro, R.; Corinaldesi, C.; Dell'Anno, A.; Fuhrman, J.A.; Middelburg, J.J.; Noble, R.T.; Suttle, C.A. Marine viruses and global climate change. *FEMS Microbiol. Rev.* **2011**, *35*, 993–1034, doi:10.1111/j.1574-6976.2010.00258.x.

4. Knowles, B.; Silveira, C.B.; Bailey, B.A.; Barott, K.; Cantu, V.A.; Cobián-Güemes, A.G.; Coutinho, F.H.; Dinsdale, E.A.; Felts, B.; Furby, K.A.; et al. Lytic to temperate switching of viral communities. *Nature* **2016**, doi:10.1038/nature17193.

5. Luque, A.; Silveira, C. Quantification of lysogeny caused by phage coinfections in microbial communities from biophysical principles. *mSystems* **2020**, *5*, e00353-20, doi:10.1128/mSystems.00353-20.

6. Silveira, C.B.; Coutinho, F.H.; Cavalcanti, G.S.; Benler, S.; Doane, M.P.; Dinsdale, E.A.; Edwards, R.A.; Francini-Filho, R.B.; Thompson, C.C.; Luque, A.; et al. Genomic and ecological attributes of marine bacteriophages encoding bacterial virulence genes. *BMC Genomics* **2020**, *21*, 126, doi:10.1186/s12864-020-6523-2.

7. Paez-Espino, D.; Eloe-Fadrosh, E.A.; Pavlopoulos, G.A.; Thomas, A.D.; Huntemann, M.; Mikhailova, N.; Rubin, E.; Ivanova, N.N.; Kyrpides, N.C. Uncovering Earth's virome. *Nature* **2016**, 1–21, doi:10.1038/nature19094.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

8. Hua, J.; Huet, A.; Lopez, C.A.; Toropova, K.; Pope, W.H.; Duda, R.L.; Hendrix, R.W.; Conway, J.F. Capsids and genomes of jumbo-sized bacteriophages reveal the evolutionary reach of the HK97 fold. *MBio* **2017**, *8*, doi:10.1128/mBio.01579-17.

9. Briani, F.; Dehò, G.; Forti, F.; Ghisotti, D. The plasmid status of satellite bacteriophage P4. *Plasmid* **2001**, *45*, 1–17, doi:10.1006/plas.2000.1497.

10. Suhanovsky, M.M.; Teschke, C.M. Nature's favorite building block: Deciphering folding and capsid assembly of proteins with the HK97-fold. *Virology* **2015**, *479–480*, 487–97, doi:10.1016/j.virol.2015.02.055.

11. Ackermann, H.W. 5500 Phages examined in the electron microscope. *Arch. Virol.* **2007**, *152*, 227–243, doi:10.1007/s00705-006-0849-1.

12. Luque, A.; Reguera, D. The structure of elongated viral capsids. *Biophys. J.* **2010**, *98*, 2993–3003, doi:10.1016/j.bpj.2010.02.051.

13. Luque, A.; Zandi, R.; Reguera, D. Optimal architectures of elongated viruses. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 5323–8, doi:10.1073/pnas.0915122107.

14. Krupovic, M.; Koonin, E. V Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl. Acad. Sci.* **2017**, *114*, E2401–E2410, doi:10.1073/pnas.1621061114.

15. Ho, P.T.; Montiel-Garcia, D.J.; Wong, J.J.; Carrillo-Tripp, M.; Brooks III, C.L.; Johnson, J.E.; Reddy, V.S. VIPERdb: A Tool for Virus Research. *Annu. Rev. Virol.* **2018**, *5*, 477–488, doi:10.1146/annurev-virology-092917-043405.

16. Sutter, M.; Boehringer, D.; Gutmann, S.; Günther, S.; Prangishvili, D.; Loessner, M.J.; Stetter, K.O.; Weber-Ban, E.; Ban, N. Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nat. Struct. Mol. Biol.* **2008**, *15*, 939–947, doi:10.1038/nsmb.1473.

17. Giessen, T.W.; Orlando, B.J.; Verdegaal, A.A.; Chambers, M.G.; Gardener, J.; Bell, D.C.; Birrane, G.; Liao,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

M.; Silver, P.A. Large protein organelles form a new iron sequestration system with high storage capacity. *Elife* **2019**, *8*, e46070, doi:10.7554/eLife.46070.

18. Caspar, D.L.; Klug, A. Physical principles in the construction of regular viruses. *Cold Spring Harb. Symp. Quant. Biol.* **1962**, *27*, 1–24, doi:10.1101/SQB.1962.027.001.005.

19. Twarock, R.; Luque, A. Structural puzzles in virology solved with an overarching icosahedral design principle. *Nat. Commun.* **2019**, *10*, 1–9, doi:10.1038/s41467-019-12367-3.

20. Podgorski, J.; Calabrese, J.; Alexandrescu, L.; Jacobs-Sera, D.; Pope, W.; Hatfull, G.; White, S. Structures of three actinobacteriophage capsids: Roles of symmetry and accessory proteins. *Viruses* **2020**, *12*, 294, doi:10.3390/v12030294.

21. Tao, Y.; Olson, N.H.; Xu, W.; Anderson, D.L.; Rossmann, M.G.; Baker, T.S. Assembly of a tailed bacterial virus and its genome release studied in three dimensions. *Cell* **1998**, *95*, 431–7, doi:10.1016/S0092-8674(00)81773-0.

22. Choi, K.H.; Morais, M.C.; Anderson, D.L.; Rossmann, M.G. Determinants of bacteriophage phi29 head morphology. *Structure* **2006**, *14*, 1723–7, doi:10.1016/j.str.2006.09.007.

23. Aksyuk, A. a; Bowman, V.D.; Kaufmann, B.; Fields, C.; Klose, T.; Holdaway, H. a; Fischetti, V. a; Rossmann, M.G. Structural investigations of a Podoviridae streptococcus phage C1, implications for the mechanism of viral entry. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, doi:10.1073/pnas.1207730109.

24. Zandi, R.; Reguera, D.; Bruinsma, R.F.; Gelbart, W.M.; Rudnick, J. Origin of icosahedral symmetry in viruses. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 15556–60, doi:10.1073/pnas.0405844101.

25. Luque, A.; Reguera, D.; Morozov, A.; Rudnick, J.; Bruinsma, R. Physics of shell assembly: Line tension, hole implosion, and closure catastrophe. *J. Chem. Phys.* **2012**, *136*, 184507, doi:10.1063/1.4712304.

26. Hagan, M.F. Modeling viral capsid assembly. *Adv. Chem. Phys.* **2014**, *155*, 1,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

doi:10.1002/9781118755815.ch01.

27. Aznar, M.; Reguera, D. Physical ingredients controlling stability and structural selection of empty viral capsids. *J. Phys. Chem. B* **2016**, *120*, 6147–6159, doi:10.1021/acs.jpcb.6b02150.

28. Rice, G.; Tang, L.; Stedman, K.; Roberto, F.; Spuhler, J.; Gillitzer, E.; Johnson, J.E.; Douglas, T.; Young, M. The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 7716–20, doi:10.1073/pnas.0401773101.

29. Koonin, E. V; Dolja, V. V; Krupovic, M.; Varsani, A.; Wolf, Y.I.; Yutin, N.; Zerbini, F.M.; Kuhn, J.H. Global organization and proposed megataxonomy of the virus world. *Microbiol. Mol. Biol. Rev.* **2020**, *84*, doi:10.1128/MMBR.00061-19.

30. Krupovic, M.; Dolja, V. V; Koonin, E. V The LUCA and its complex virome. *Nat. Rev. Microbiol.* **2020**, 1–10, doi:10.1038/s41579-020-0408-x.

31. Doore, S.M.; Fane, B.A. The microviridae: diversity, assembly, and experimental evolution. *Virology* **2016**, *491*, 45–55, doi:10.1016/j.virol.2016.01.020.

32. Creasy, A.; Rosario, K.; Leigh, B.A.; Dishaw, L.J.; Breitbart, M. Unprecedented diversity of ssDNA phages from the family Microviridae detected within the gut of a protochordate model organism (Ciona robusta). *Viruses* **2018**, *10*, 404, doi:10.3390/v10080404.

33. Mavrich, T.N.; Hatfull, G.F. Evolution of Superinfection Immunity in Cluster A Mycobacteriophages. *Am Soc Microbiol* **2019**, doi:10.1128/mBio.00971-19.

34. Wiegand, T.; Karambelkar, S.; Bondy-Denomy, J.; Wiedenheft, B. Structures and Strategies of Anti-CRISPR-Mediated Immune Suppression. *Annu. Rev. Microbiol.* **2020**, *74*, doi:10.1146/annurev-micro-020518-120107.

35. Breitbart, M.; Thompson, L.R.; Suttle, C.A.; Sullivan, M.B. Exploring the vast diversity of marine viruses.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

Oceanography **2007**, *20*, 135–139, doi:10.5670/oceanog.2007.58.

36. Liu, X.; Zhang, Q.; Murata, K.; Baker, M.L.; Sullivan, M.B.; Fu, C.; Dougherty, M.T.; Schmid, M.F.; Osburne, M.S.; Chisholm, S.W.; et al. Structural changes in a marine podovirus associated with release of its genome into Prochlorococcus. **2010**, *17*, doi:10.1038/nsmb.1823.

37. Bebeacua, C.; Lai, L.; Vegge, C.S.; Brøndsted, L.; van Heel, M.; Veesler, D.; Cambillau, C. Visualizing a complete Siphoviridae member by single-particle electron microscopy: the structure of lactococcal phage TP901-1. *J. Virol.* **2013**, *87*, 1061–1068, doi:10.1128/JVI.02836-12.

38. Parent, K.N.; Tang, J.; Cardone, G.; Gilcrease, E.B.; Janssen, M.E.; Olson, N.H.; Casjens, S.R.; Baker, T.S. Three-dimensional reconstructions of the bacteriophage CUS-3 virion reveal a conserved coat protein I-domain but a distinct tailspike receptor-binding domain. *Virology* **2014**, *464*, 55–66, doi:10.1016/j.virol.2014.06.017.

39. Spilman, M.S.; Dearborn, A.D.; Chang, J.R.; Damle, P.K.; Christie, G.E.; Dokland, T. A conformational switch involved in maturation of Staphylococcus aureus bacteriophage 80α capsids. *J. Mol. Biol.* **2011**, *405*, 863–876, doi:10.1016/j.jmb.2010.11.047.

40. Effantin, G.; Figueroa-Bossi, N.; Schoehn, G.; Bossi, L.; Conway, J.F. The tripartite capsid gene of Salmonella phage Gifsy-2 yields a capsid assembly pathway engaging features from HK97 and λ. *Virology* **2010**, *402*, 355–365, doi:10.1016/j.virol.2010.03.041.

41. Shen, P.S.; Domek, M.J.; Sanz-García, E.; Makaju, A.; Taylor, R.M.; Hoggan, R.; Culumber, M.D.; Oberg, C.J.; Breakwell, D.P.; Prince, J.T. Sequence and structural characterization of great salt lake bacteriophage CW02, a member of the T7-like supergroup. *J. Virol.* **2012**, *86*, 7907–7917, doi:10.1128/JVI.00407-12.

42. Dai, W.; Hodes, A.; Hui, W.H.; Gingery, M.; Miller, J.F.; Zhou, Z.H. Three-dimensional structure of tropism-switching Bordetella bacteriophage. *Proc. Natl. Acad. Sci.* **2010**, *107*, 4347–4352,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

doi:10.1073/pnas.0915008107.

43. White, H.E.; Sherman, M.B.; Brasilès, S.; Jacquet, E.; Seavers, P.; Tavares, P.; Orlova, E. V Capsid Structure and Its Stability at the Late Stages of Bacteriophage SPP1 Assembly. *J. Virol.* **2012**, *86*, 6768–77, doi:10.1128/JVI.00412-12.

44. Grose, J.H.; Belnap, D.M.; Jensen, J.D.; Mathis, A.D.; Prince, J.T.; Merrill, B.D.; Burnett, S.H.; Breakwell, D.P. The Genomes, Proteomes, and Structures of Three Novel Phages That Infect the Bacillus cereus Group and Carry Putative Virulence Factors. *J. Virol.* **2014**, *88*, 11846–11860, doi:10.1128/JVI.01364-14.

45. Lander, G.C.; Baudoux, A.; Azam, F.; Potter, C.S.; Carragher, B.; Johnson, J.E. Article Capsomer Dynamics and Stabilization in the T = 12 Marine Bacteriophage SIO-2 and Its Procapsid Studied by CryoEM. *Struct. Des.* **2012**, *20*, 498–503, doi:10.1016/j.str.2012.01.007.

46. Vernhes, E.; Renouard, M.; Gilquin, B.; Cuniasse, P.; Durand, D.; England, P.; Hoos, S.; Huet, A.; Conway, J.F.; Glukhov, A. High affinity anchoring of the decoration protein pb10 onto the bacteriophage T5 capsid. *Sci. Rep.* **2017**, *7*, 41662, doi:10.1038/srep41662.

47. Lander, G.C.; Tang, L.; Casjens, S.R.; Gilcrease, E.B.; Prevelige, P.; Poliakov, A.; Potter, C.S.; Carragher, B.; Johnson, J.E. The structure of an infectious P22 virion shows the signal for headful DNA packaging. *Science (80-.).* **2006**, *312*, 1791–1795, doi:10.1126/science.1127981.

48. Stroupe, M.E.; Brewer, T.E.; Sousa, D.R.; Jones, K.M. The structure of Sinorhizobium meliloti phage ΦM12, which has a novel T= 19l triangulation number and is the founder of a new group of T4-superfamily phages. *Virology* **2014**, *450*, 205–212, doi:10.1016/j.virol.2013.11.019.

49. Effantin, G.; Hamasaki, R.; Kawasaki, T.; Bacia, M.; Moriscot, C.; Weissenhorn, W.; Yamada, T.; Schoehn, G. Cryo-electron microscopy three-dimensional structure of the jumbo phage ΦRSL1 infecting the phytopathogen Ralstonia solanacearum. *Structure* **2013**, *21*, 298–305, doi:10.1016/j.str.2012.12.017.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

50. Fokine, A.; Kostyuchenko, V. a.; Efimov, A. V.; Kurochkina, L.P.; Sykilinda, N.N.; Robben, J.; Volckaert, G.; Hoenger, A.; Chipman, P.R.; Battisti, A.J.; et al. A three-dimensional cryo-electron microscopy structure of the bacteriophage φKZ head. *J. Mol. Biol.* **2005**, *352*, 117–124, doi:10.1016/j.jmb.2005.07.018.

51. Pietilä, M.; Laurinmäki, P.; Russell, D.A.; Ching-Chung, K.; Jacobs-Sera, D.; Hendrix, R.W.; Bamford, D.H.; Butcher, S.J. Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 10604–10609, doi:10.1073/pnas.1303047110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1303047110.

52. Guo, F.; Liu, Z.; Fang, P.-A.; Zhang, Q.; Wright, E.T.; Wu, W.; Zhang, C.; Vago, F.; Ren, Y.; Jakana, J. Capsid expansion mechanism of bacteriophage T7 revealed by multistate atomic models derived from cryo-EM reconstructions. *Proc. Natl. Acad. Sci.* **2014**, *111*, E4606–E4614, doi:10.1073/pnas.1407020111.

53. Parent, K.N.; Gilcrease, E.B.; Casjens, S.R.; Baker, T.S. Structural evolution of the P22-like phages: Comparison of Sf6 and P22 procapsid and virion architectures. *Virology* **2012**, *427*, 177–88, doi:10.1016/j.virol.2012.01.040.

54. Baker, M.L.; Hryc, C.F.; Zhang, Q.; Wu, W.; Jakana, J.; Haase-Pettingell, C.; Afonine, P. V; Adams, P.D.; King, J. a; Jiang, W.; et al. Validated near-atomic resolution structure of bacteriophage epsilon15 derived from cryo-EM and modeling. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 12301–6, doi:10.1073/pnas.1309947110.

55. Gipson, P.; Baker, M.L.; Raytcheva, D.; Haase-Pettingell, C.; Piret, J.; King, J.A.; Chiu, W. Protruding knob-like proteins violate local symmetries in an icosahedral marine virus. *Nat. Commun.* **2014**, *5*, 1–11, doi:10.1038/ncomms5278.

56. Leiman, P.G.; Battisti, A.J.; Bowman, V.D.; Stummeyer, K.; Mühlenhoff, M.; Gerardy-Schahn, R.; Scholl, D.; Molineux, I.J. The structures of bacteriophages K1E and K1-5 explain processive degradation of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

polysaccharide capsules and evolution of new host specificities. *J. Mol. Biol.* **2007**, *371*, 836–849, doi:10.1016/j.jmb.2007.05.083.

57. Lander, G.C.; Evilevitch, A.; Jeembaeva, M.; Potter, C.S.; Carragher, B.; Johnson, J.E. Bacteriophage lambda stabilization by auxiliary protein gpD: timing, location, and mechanism of attachment determined by cryo-EM. *Structure* **2008**, *16*, 1399–1406, doi:10.1016/j.str.2008.05.016.

58. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612, doi:10.1002/jcc.20084.

59. Team, R.C. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria: Vienna 2018.

60. Colin, B.; Luque, A. Extension of hkcage to generate icosahedral capsids with multiple lattices 2020.

61. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*, D733–D745, doi:10.1093/nar/gkv1189.

62. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357.

63. Aznar, M.; Luque, A.; Reguera, D. Relevance of capsid structure in the buckling and maturation of spherical viruses. *Phys. Biol.* **2012**, *9*, 036003, doi:10.1088/1478-3975/9/3/036003.

64. Dearborn, A.D.; Laurinmaki, P.; Chandramouli, P.; Rodenburg, C.M.; Wang, S.; Butcher, S.J.; Dokland, T. Structure and size determination of bacteriophage P2 and P4 procapsids: function of size responsiveness mutations. *J. Struct. Biol.* **2012**, *178*, 215–224, doi:10.1016/j.jsb.2012.04.002.

65. Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res.* **2011**, *39*, D576–D582,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

doi:10.1093/nar/gkq901.

66. Daufresne, M.; Lengfellner, K. Som-mer U. et al. Global warming benefits the small in aquatic ecosystems, 12788-12793. *PNAS* **2009**, *106*, 21, doi:10.1073/pnas.0902080106.

67. Morán, X.A.G.; López-Urrutia, Á.; Calvo-Díaz, A.; Li, W.K.W. Increasing importance of small phytoplankton in a warmer ocean. *Glob. Chang. Biol.* **2010**, *16*, 1137–1144, doi:10.1111/j.1365-2486.2009.01960.x.

68. Nifong, R.L.; Gillooly, J.F. Temperature effects on virion volume and genome length in dsDNA viruses. *Biol. Lett.* **2016**, *12*, 20160023, doi:10.1098/rsbl.2016.0023.

69. Schooley, R.T.; Biswas, B.; Gill, J.J.; Hernandez-Morales, A.; Lancaster, J.; Lessor, L.; Barr, J.J.; Reed, S.L.; Rohwer, F.; Benler, S. Development and use of personalized bacteriophage-based therapeutic cocktails to treat a patient with a disseminated resistant Acinetobacter baumannii infection. *Antimicrob. Agents Chemother.* **2017**, *61*, doi:10.1128/AAC.00954-17.

70. Hatfull, G.F. Actinobacteriophages: Genomics, Dynamics, and Applications. *Annu. Rev. Virol.* **2020**, *7*, 37–61, doi:10.1146/annurev-virology-122019-070009.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

Figures and Tables

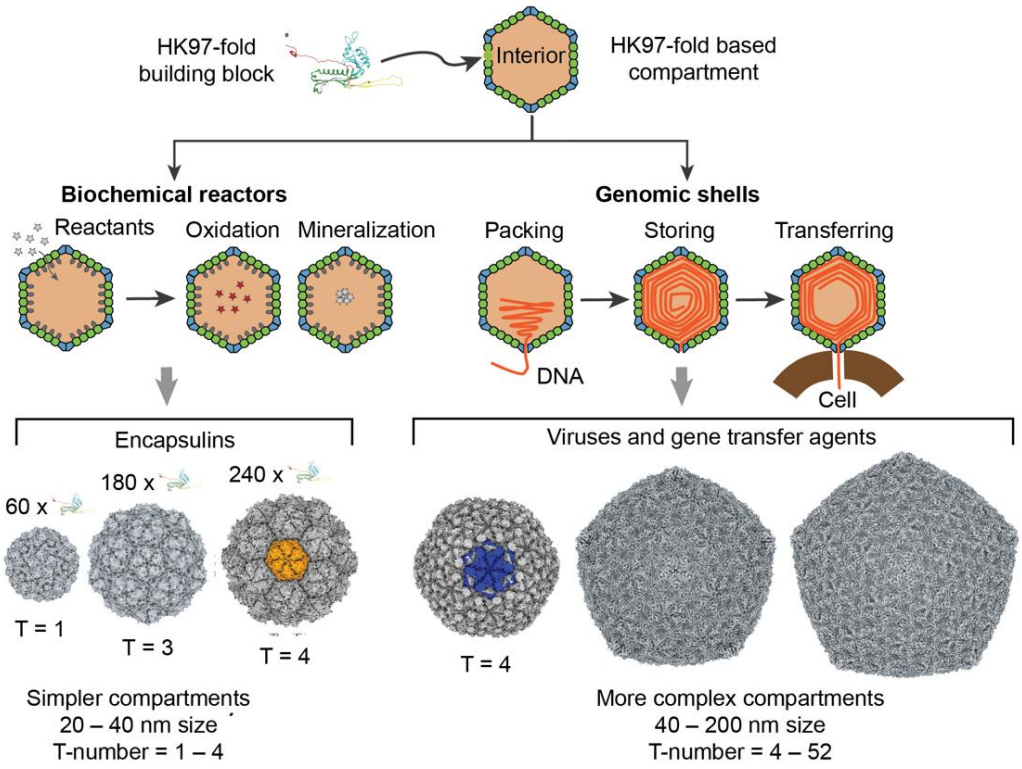

Figure 1. HK97-fold protein compartments. The panel on the left focuses on encapsulins, nanocompartments responsible for chemical storage and biochemical reactions in bacteria and archaea. The panel on the right side focuses on viruses and gene transfer agents. The viruses belong to the realm *of Duplodnaviria*. Tailed phages in the phylum *Uroviricota* and class *Cauvricetes* are the most diverse and abundant representatives of this group.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

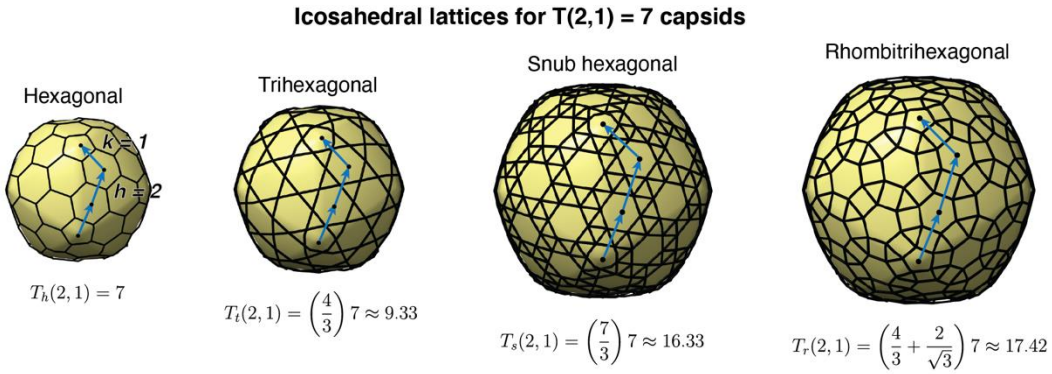

Figure 2. Icosahedral lattices for T(2,1) = 7 capsids. The label on the top displays the name of the lattice. The blue arrows and blue dots display the h and k steps in the hexagonal lattice, h = 2 and k = 1 in this case. The generalized T_h, T_t, T_s, and T_r numbers are obtained from the classic T-number multiplied by the lattice factor associated with the minor polygons (triangles and squares) of each lattice: h (hexagonal), t (trihexagonal), s (snub hexagonal), and r (rhombitrihexagonal).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

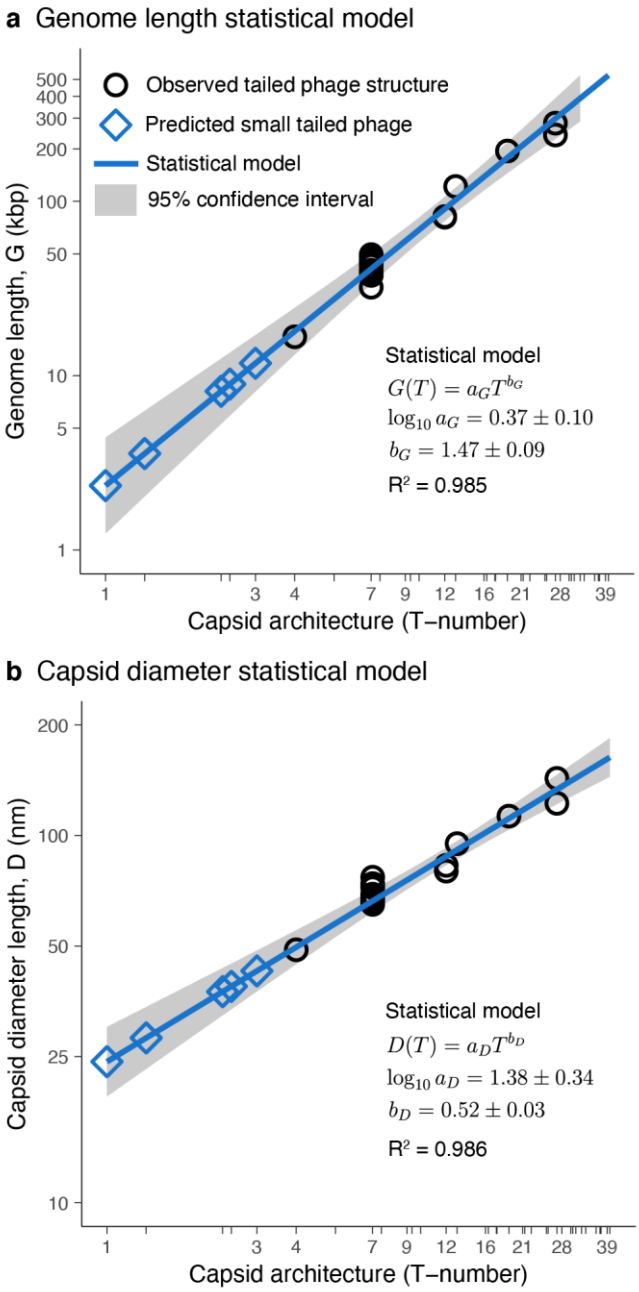

Figure 3. Tailed phage statistical models. Genome length (**a**) and capsid diameter (**b)** plotted as a function of capsid architecture for the studied tailed phage structures (black circles) and the predicted small tailed phage structures (blue diamonds). **a-b** The solid blue lines and grey areas correspond, respectively, to the mean values and 95% confidence interval predicted from the statistical model. The mean values and standard errors of the fitted parameters, as well as the coefficient of determination (R^2) are displayed.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

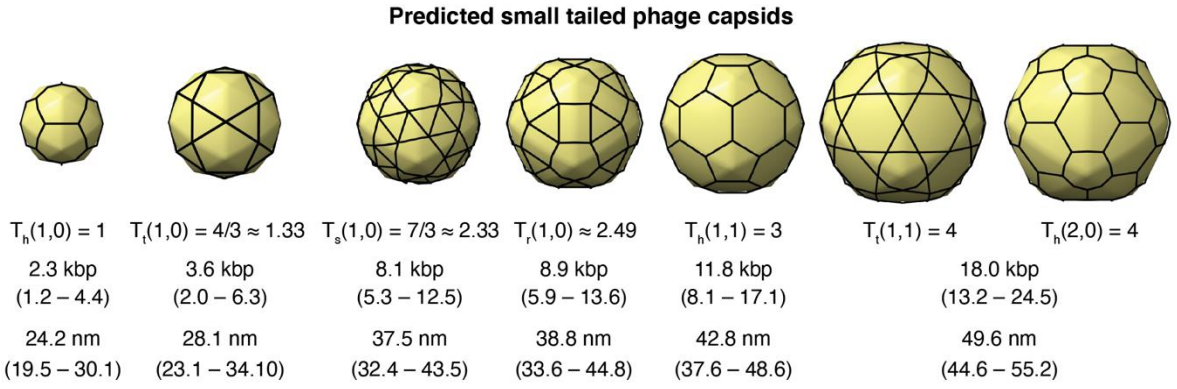

Figure 4. Predicted small tailed phage capsids. The sequence of 3D capsid icosahedral models generated with the new hkcage-prototype bundle in Chimera X [60]. The information below each structure corresponds to the T-number architecture, (h,k) steps, lattice (h: hexagonal, t: trihexagonal, s: snub hexagonal, and r: rhombitrihexagonal), the average predicted genome length (95% confidence interval), and average predicted capsid diameter (95% confidence interval).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

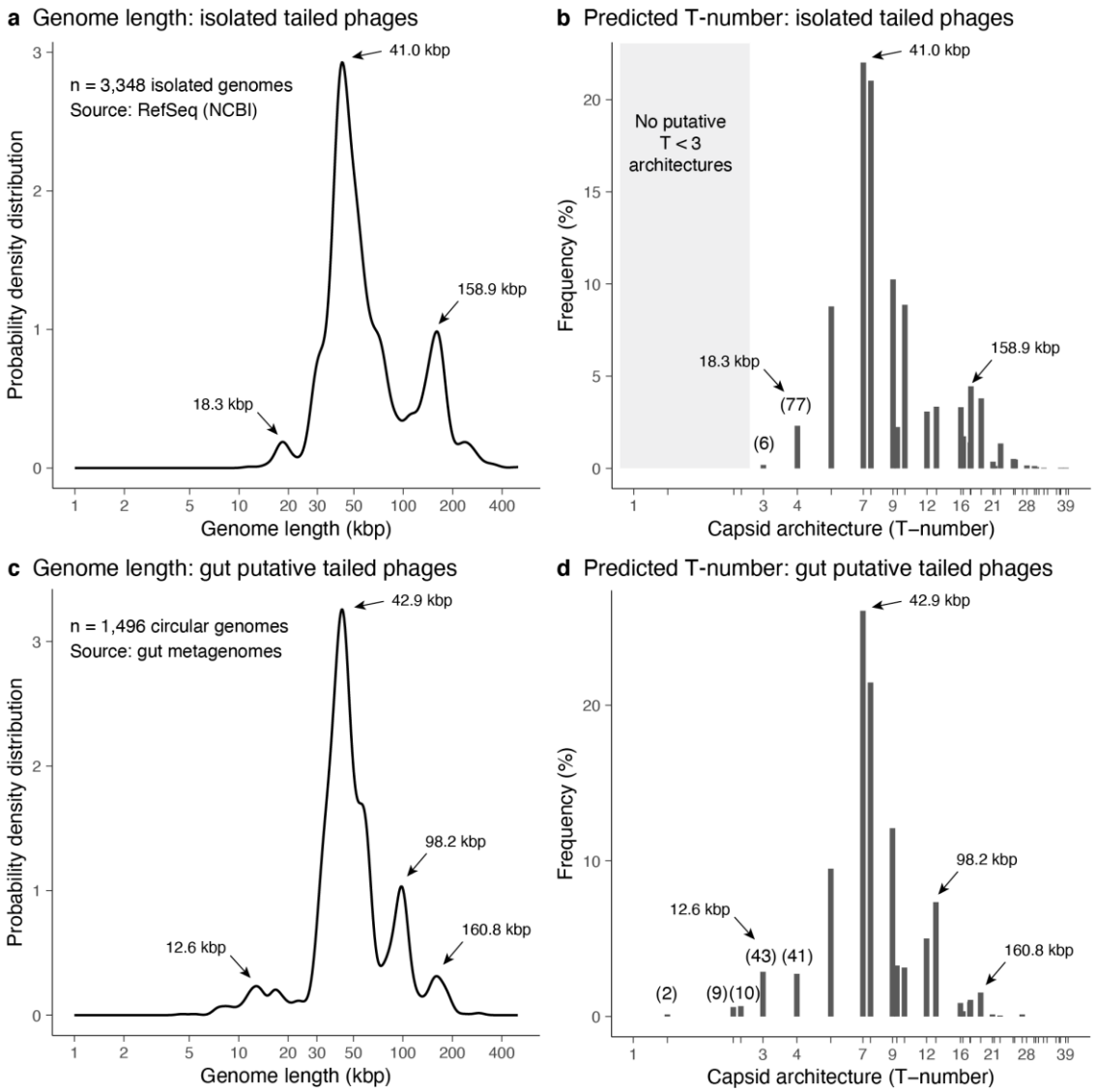

Figure 5. Predicted architectures among tailed phage isolates. (**a**) Genome length distribution for tailed phage isolates obtained from the NCBI Reference Sequence Database. (**b**) Frequency (percentage) of predicted T-number architectures from the genome lengths of isolated tailed phages. The grey area highlights the absence of T < 3 architectures. (**c**–**d**) Genome length distribution and predicted T-number architectures for putative gut tailed phages. (**a**–**d**) The arrows indicate the significant peaks of the genome length distribution and the associated T-number architectures. (**b**,**d**) The parenthesis indicates the number of predicted T ≤ 4 phage capsid architectures.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

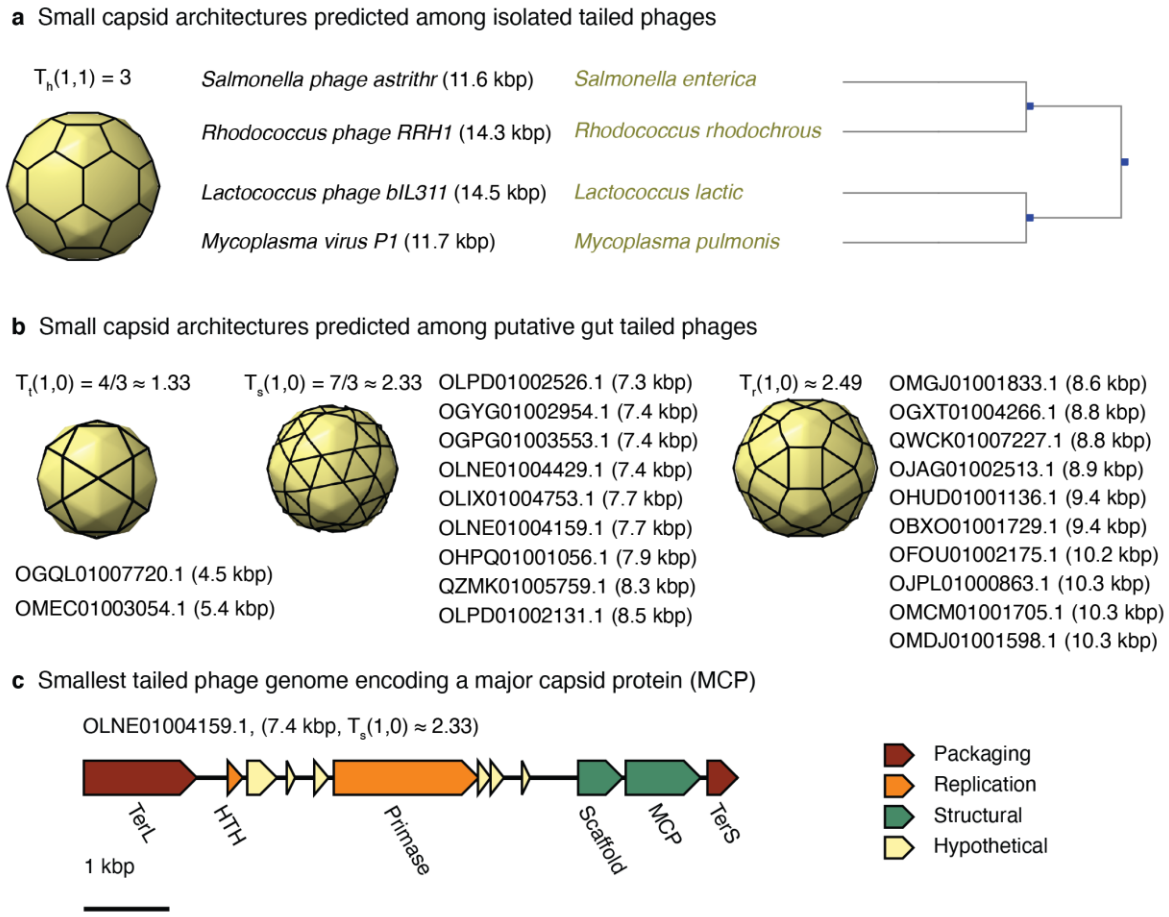

Figure 6. Predicted small tailed phages. (**a**) List of isolated tailed phages predicted to adopt a T = 3 capsid architecture. The panel displays the phage name, genome length, host, and hosts' phylogenetic tree. (**b**) List of putative gut tailed phage genomes predicted to adopt T ≤ 3 capsid architectures. The list displays the contig name and genome length in parenthesis. (**c**) Genome map of the smallest genome encoding a major capsid protein (MCP). The scale is 1 kbp. The genes are group by function: packaging (red), replication (orange), structural (green), and hypothetical (yellow).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0024.v1

Property	**Range**	**Property**	**Range**
Capsids analyzed	23	Genome size	17–280 kbp
T-number	4–27	Genome density	0.34–0.62 bp/nm^3
Interior sphericity*	0.25–0.75	Interior surface	$5.08 \cdot 10^3$–$4.32 \cdot 10^4$ nm^2
Exterior sphericity*	0.14–0.55	MCP interior area	19–26 nm^2
Capsid diameter†	49–143 nm	Exterior surface	$6.55 \cdot 10^3$–$5.19 \cdot 10^4$ nm^2
Capsid thickness	3–8 nm	MCP exterior area	24–35 nm^2
Interior volume	$3.36 \cdot 10^4$–$8.22 \cdot 10^5$ nm^3	MCP ratio (%)	15–43%

Table 1. Summary of measured structural properties. *The sphere factor ranged from 0 (polyhedral) to 1 (spherical). †Maximum (icosahedral) diameter determined from the vertex (5-fold) to vertex (5-fold).