Computational and Structural Biotechnology Journal 20 (2022) 721–732

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Predicting the capsid architecture of phages from metagenomic data

Diana Y. Lee [a,b], Caitlin Bartels [a,c], Katelyn McNair [a,b], Robert A. Edwards [a,b,c,d], Manal A. Swairjo [a,e], Antoni Luque [a,b,f,*]

[a] Viral Information Institute, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA
[b] Computational Science Research Center, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA
[c] Department of Biology, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA
[d] Flinders Accelerator for Microbiome Exploration, Flinders University, Bedford Park, GPO Box 2100, Adelaide 5001, South Australia, Australia
[e] Department of Chemistry and Biochemistry, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA
[f] Department of Mathematics & Statistics, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA

ARTICLE INFO

Article history:
Received 1 October 2021
Received in revised form 22 December 2021
Accepted 22 December 2021
Available online 5 January 2022

Keywords:
Tailed bacteriophages
Icosahedral capsids
Physical modeling
Machine learning
Metagenomes
Gut microbiome
Viral ecology
Physical virology

ABSTRACT

Tailed phages are viruses that infect bacteria and are the most abundant biological entities on Earth. Their ecological, evolutionary, and biogeochemical roles in the planet stem from their genomic diversity. Known tailed phage genomes range from 10 to 735 kilobase pairs thanks to the size variability of the protective protein capsids that store them. However, the role of tailed phage capsids' diversity in ecosystems is unclear. A fundamental gap is the difficulty of associating genomic information with viral capsids in the environment. To address this problem, here, we introduce a computational approach to predict the capsid architecture (T-number) of tailed phages using the sequence of a single gene—the major capsid protein. This approach relies on an allometric model that relates the genome length and capsid architecture of tailed phages. This allometric model was applied to isolated phage genomes to generate a library that associated major capsid proteins and putative capsid architectures. This library was used to train machine learning methods, and the most computationally scalable model investigated (random forest) was applied to human gut metagenomes. Compared to isolated phages, the analysis of gut data reveals a large abundance of mid-sized (T = 7) capsids, as expected, followed by a relatively large frequency of jumbo-like tailed phage capsids ($T \geq 25$) and small capsids (T = 4) that have been under-sampled. We discussed how to increase the method's accuracy and how to extend the approach to other viruses. The computational pipeline introduced here opens the doors to monitor the ongoing evolution and selection of viral capsids across ecosystems.

1. Introduction

Tailed phages are viruses that infect bacteria and have evolved an extremely diverse set of protein capsid architectures to protect their infective genome [82,12]. Tailed phage capsids sizes range from 40 nm to 180 nm in diameter [95,118,51]. The internal volumes of these capsids accommodate genomes spanning three orders of magnitude in length, from 5 kilobase pairs (kbp) to 735 kbp [82,62]. The diversity in genome length and genomic content of tailed phages may explain their key role in regulating ecosystems [86,113], in promoting the evolution of microbes [129,67,121], in participating strongly in the planetary carbon cycle [74], and in becoming the most abundant biological entity on the planet [27]. However, the role of the diversity in tailed phage capsid architectures and genome lengths across ecosystems remains unclear.

A key challenge investigating the selection and evolution of tailed phage capsids is linking viral capsids with their viral genome in the environment [19]. The number of phages isolated and studied both genetically and structurally [34,70] represent a very small sample compared to the number of viruses evolving in the environment [27,6,29,38,109,50,103,11,107]. Electron microscopy can measure the morphology and size of tailed phages, but these observations do not include genomic information, limiting how to interpret the change in capsid size distributions observed across ecosystems[119,18]. There are trade-offs in selecting capsid sizes that are difficult to disentangle [37]. An increase in temperature may promote smaller genomes among viruses and other organisms

* Corresponding author at: Viral Information Institute, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA.
E-mail address: aluque@sdsu.edu (A. Luque).

https://doi.org/10.1016/j.csbj.2021.12.032

[90], but larger genomes encode more genes, which can enhance the survival of both phages and their hosts[120,110,112]. On the other hand, larger genomes and their associated larger capsids are more costly energetically, which can compromise their replication in limiting growth conditions [20,84]. Additionally, an increase in size reduces virus diffusivity [26], which can negatively impact their infectivity [66]. To link the capsid and genomic information of viruses in the environment, we introduced a new computational approach that builds on the established geometrical principles governing the capsid structure and genome packing of tailed phages [102,81,44,118,42,82].

The majority—80% to 90%—of tailed phage capsids are quasi-spherical [2,3]. The remaining tailed phages adopt elongated capsids with icosahedral caps [3,80]. Among tailed phages, the capsids are built from multiple copies of the major capsid protein, which systematically adopt the HK97-fold [124,98,34]. Capsid proteins in tailed phages are organized following hexagonal and trihexagonal icosahedral lattices, Fig. 1 [122,99,82], and the double-stranded DNA genome is packed in the capsid at quasicrystalline densities [36,78,82]. The number of capsid proteins is determined by the triangulation number or T-number, which is a discrete index determining the possible capsid surfaces compatible with icosahedral symmetry [24,122]. The number of major capsid proteins is 60 T_0 (Fig. 1), where T_0 represents the classic T-number:

$$T_0(h,k) = h^2 + hk + k^2 \quad . \tag{1}$$

In the generalized theory for icosahedral capsids, the T-number for the hexagonal lattice is $T_{hex} = T_0$, and the T-number for the trihexagonal lattice is $T_{tri} = 4/3\ T_0$ [122]. The factor $4/3 \approx 1.33$ accounts for the additional surface associated with 60 T_0 minor capsid proteins inserted as trimers in the trihexagonal lattice (Fig. 1). Experimental and bioinformatic studies indicate that tailed phages can adopt capsid architectures from putative T = 1.33 capsids to T = 52 capsids [118,82]. The T-number follows an allometric relationship with the genome length with an approximate exponent of $2/3 \approx 0.67$ because the T-number is proportional to the capsid surface and the genome length is proportional to the capsid volume [82]. Thus, the increase in genomic content is associated with larger tailed phage capsids built with more capsid proteins. Since the major capsid proteins conserve the HK97-fold while adopting a large diversity of sequences, here we propose that part of this sequence diversity is associated with the formation of different T-number capsids.

Confirming a direct relationship between major capsid protein sequences and T-number capsids would open the doors to predicting the capsid architecture of tailed phages (and genome lengths) from a single gene. This would facilitate inferring tailed phage capsids from sequenced environmental data that is now obtained routinely [17,112,103,77,105]. To test the capsid protein-to-T-number association, we developed a computational approach that can predict accurately the capsid architectures of tailed phages from the major capsid protein gene (Fig. 2). First, the genome-to-T-number model (G2T) was extended by training a power function physical model using a larger database of high-resolution tailed phage capsids than prior studies (Fig. 2a). Major capsid proteins (MCPs) adopting HK97-fold were obtained from tailed phage genome isolates, and the G2T model was applied to the genomes to obtain the putative capsid architectures among these phage isolates, generating the MCP/T library (Fig. 2b). The MCP/T library was used to train the major capsid protein-to-T-number (MCP2T) models using a proximity matrix approach (MCP2T-PM) and a random forest approach (MCP2T-RF) (Fig. 2c). Finally, these statistical

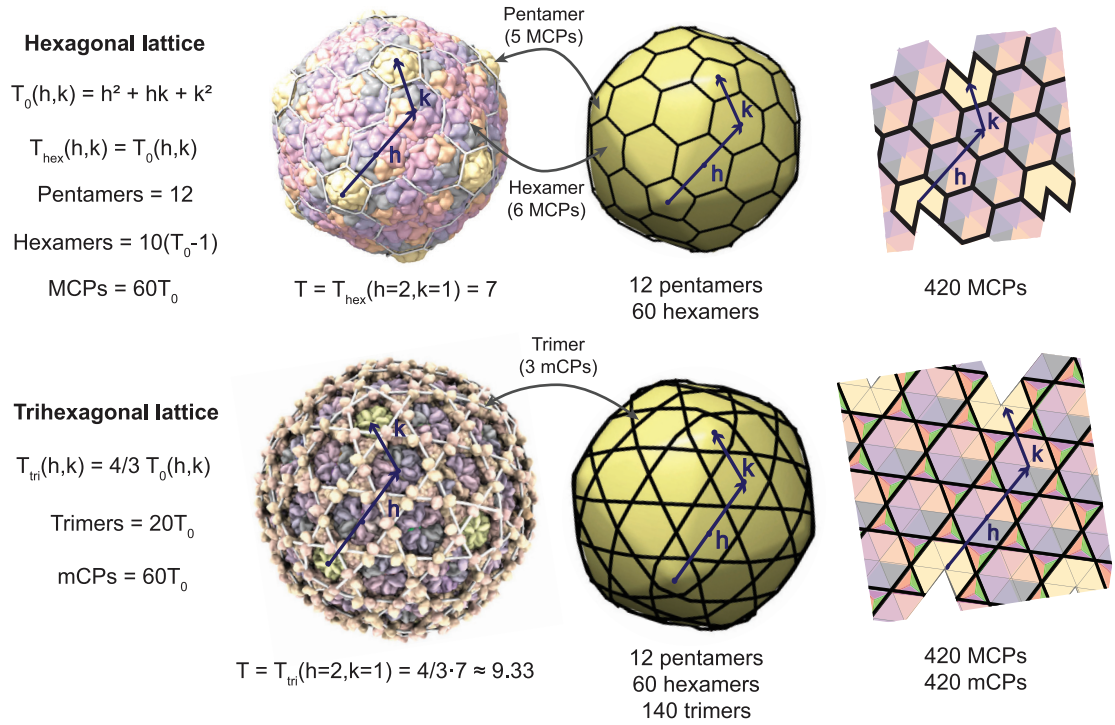

Fig. 1. Icosahedral capsids among tailed phages. The hexagonal (top) and trihexagonal (bottom) icosahedral lattices observed among icosahedral tailed phage capsids. In both lattices, major capsid proteins (MCPs) form clusters (capsomers) of five (pentamers) and six (hexamers) proteins. Two nearby pentamers are connected by *h* and *k* steps crossing over hexamers. The trihexagonal lattice also contains minor capsid proteins (mCPs) clustered in groups of three (trimers). The T-number is proportional to the number of major and minor capsid proteins. T_0 is the T-number defined by the classic icosahedral capsid theory [24]. T_{hex} and T_{tri} are the T-numbers associated, respectively, with the hexagonal and trihexagonal lattices defined by the generalized icosahedral capsid theory[122]. The top and bottom capsid examples correspond, respectively, to phage HK97 (PDB 2 fs3; [47]and phage patience (EMDB-21123; [99]. The capsids were rendered with ChimeraX[96]. The 3D icosahedral lattice models were produced with the generalized *hkcage* tool in ChimeraX[82].

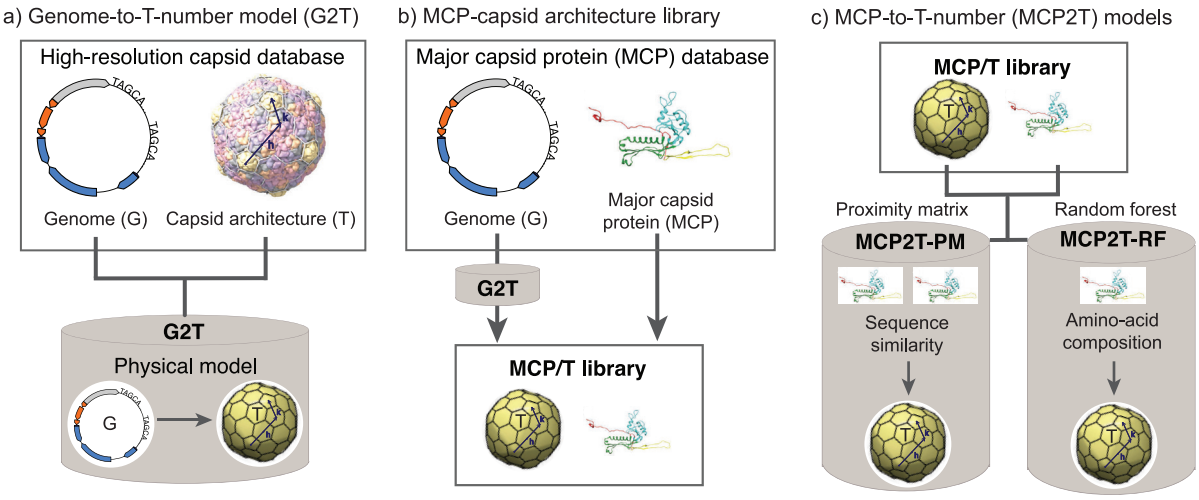

Fig. 2. Computational approach to predict capsid architecture from genomic information. a) A database containing tailed phage genomes and their associated high-resolution capsid reconstructions was used to validate the physical genome-to-T-number (G2T) model. b) A database containing isolated tailed phage genomes and encoded HK97-fold major capsid proteins (MCPs) was curated. The G2T model was applied to identify the putative T-number capsid architectures associated with each HK97-fold MCP, obtaining the MCP/T library. c) The MCP/T library was used to train statistical learning methods to predict the capsid architecture of tailed phages from information in the MCP sequence, leading to the major capsid protein-to-T-number (MCP2T) models. The MCP2T-PM model was built on a proximity matrix (PM) algorithm using protein sequence similarity. The MCP2T-RF model was built on a random forest algorithm using MCP amino-acid composition as features.

learning models were applied to metagenomic data to infer the capsid architecture of uncultured tailed phages in the human gut.

2. Methods

The GitHub repository http://github.com/Luquelab/Lee_etal_CSBJ_2022 contains the codes and instructions necessary to implement the methods and replicate the research. The supplementary section SI-1 contains the description of the supplementary Data Files referenced in the Methods and Results sections.

Genome-to-T-number (G2T) model. The genome-to-T-number (G2T) model is a physical model that predicts the capsid architecture (T-number) of a tailed phage from its genome length (Fig. 2a). The G2T model was introduced in [82]. The model relies on the empirically and theoretically justified physical allometric relation between the genome length and capsid architecture of tailed phages [82]. Here, the G2T model was revised, increasing the database of high-resolution structure to train and test the model (from 23 to 37 structures) as detailed in the *Data acquisition* section below. Another novelty was the error analysis of the model and error prediction when increasing the training data set, as detailed in the *Model accuracy* section below.

Data acquisition. Tailed phages containing high-resolution capsid data were initially identified from a review article in the field [118], the icosahedral capsid database VIPERdb [89], and four recently reconstructed tailed phages displaying new T-numbers: the jumbo tailed phage SCTP2 [59] and P74-26, P23-45, and Mic1 [116,8,65]. The capsid protein stoichiometry and high-resolution structures were revised to update the T-numbers according to the generalized quasi-equivalence icosahedral framework, including hexagonal and trihexagonal lattices observed among tailed phages [122]. The final high-resolution database included n_{HR} = 37 tailed phage capsid structures (Table 1 and Data File 1).

Statistical model. A power function model $T(G) = b\left(\frac{G}{G_0}\right)^a$ related the T-number as a function of the genome length, G. Here, b was the prefactor constant, a the allometric exponent, and G_0 the reference units of G, G_0 = 1 kbp. This allometric relationship was empirically and theoretically established previously for a smaller number of tailed phages [60,82]. The allometric relationship is a consequence of the constant density of the genome stored in tailed

Table 1
High-resolution capsid database. See additional information in Data File 1.

Phage	T	Genome (kbp)	Reference
C1	4	16.7	[5]
HSTV-1	7	32.2	[98]
P2	7	33.6	[31]
TP901-1	7	37.7	[9]
Sf6	7	39.0	[92]
ε15	7	39.7	[7,64]
HK97	7	39.7	[48,55,124]
T7	7	39.9	[4,53,61]
CUS-3	7	40.2	[93]
HK022	7	40.8	[100]
Pf-WMP4	7	40.9	[132]
BPP-1	7	42.9	[127]
P22	7	43.5	[25,91]
80α	7	43.9	[115]
K1E/K1-5	7	44.7	[75]
P-SSP7	7	45 . 0	[79]
Gifsy-2	7	45.8	[40]
Syn5	7	46.2	[49,131]
Λ	7	48.5	[73]
CW02	7	49. 4	[108]
SPP1	7	49.5	[123]
SIO-2	12	80.0	[72]
P74-26	9.33	83.0	[116]
P23-45	9.33	84.2	[8]
Basilisk	12	81. 8	[52,122]
Mic1	13	92.6	[65]
T5	13	121. 8	[39]
SPO1	16	132. 6	[35]
ΦM12	19	194.7	[117]
N3	19	207.0	[118,59]
PAU	25	219.0	[118,59]
ΦRSL1	27	240.0	[41]
PBS1	27	252.0	[118,59]
ΦKZ	27	280.0	[45]
121Q	28	348.5	[118]
SCTP2	39	440.0	[59]
G	52	498.0	[118,59]

phage capsids and constant surface of the major protein on the capsid exterior [82]. The theory predicts an allometric exponent a_{th} = 2/3 because the T-number scales like the capsid surface and the genome scales with the capsid volume. A derivation of the theoretical prediction is provided in the supplementary section SI-2. The model was linearized using a logarithmic transformation:

$$ln(T) = a\, ln(G/G_0) + ln(b)\ . \quad (2)$$

The slope, a, and intercept, ln b, of best fit were obtained using the least squares method in the *Linear Regression* function from the Scikit learn package for Python [94]. The residual bias and coefficient of determination of this model were compared with alternative models (exponential, quadratic, reciprocal, logarithmic) for quality control, confirming the adequacy of the power function model (see supplementary section SI-3 and Fig. S1).

Model accuracy. The accuracy of the G2T model was investigated statistically using different training sets. This estimated the expected model's error and facilitated making projections to judge if increasing the data set would improve the model. The approach was as follows. The best fit values for the G2T model, Eq. (2), were obtained using different training data sets of size n, ranging from n = 5 to n = 30. The n data points in a training data set were chosen randomly from the high-resolution tailed phage capsid database (Table 1). For each model, the T-number was predicted from the genome length of the remaining capsid structures (n_{HR} – n, that is, 37 – n). The relative error was defined as the model's residual (difference between the predicted T-number and the empirical T-number) divided by the empirical T-number. This process was repeated 10,000 times for each n to estimate the G2T's mean relative error (MRE) as a function of the training data set size, n. To predict the accuracy of the model for data sets larger than the current database, (n > n_{HR}), the mean relative error was fitted to the exponential model

$$MRE(n) = pe^{-qn} + w\,. \quad (3)$$

The values of best fit for the parameters p, q, and w were obtained applying the robust least squares method from the least squares function in the Python's SciPy optimize package [125]. The confidence interval of the parameters was estimated by bootstrapping 10,000 random subsets and fitting Eq. (3) in each case. A genome length was associated with a T-number in the hexagonal or trihexagonal lattice if the uncertainty of the predicted T value, that is, T ± ΔT, contained such T-number. The uncertainty ΔT was calculated based on the mean relative error projected from Eq. (3) for the size of the high-resolution database, n = n_{HR} = 37, that is, ΔT = T·MRE(n_{HR}).

MCP/T library. Major capsid protein amino acid sequences associated with tailed phages were obtained from isolated genomes accessed on the phantome.org website in January 2017 [88,97]. Genomes listed as *Caudovirales* (the taxonomic order of tailed phages) in the GenBank *ORGANISM* field were filtered. Among the 2,996 *Caudovirales* genomes identified, protein-coding genes (CDS) containing the term "major capsid" as a product keyword were selected, leading to 669 putative tailed phage major capsid proteins. The folded structures for the selected major capsid proteins were obtained investigating structural relatives in HHpred using the PDB database and submitting the top candidates (above 95% probability) to Modeller [128,46,114,56,88]. The folded models were inspected visually. Only those major capsid proteins displaying the canonical features of the HK97-fold were selected [118]. Major capsid proteins identified in phage genomes from the high-resolution capsid database were also included. The protein sequences associated with open reading frames (ORFs) in these genomes were retrieved from NCBI. Structural functions were identified from the protein sequences using the Phage Artificial Neural Networks (PhANNs) web server [21]. Sequences predicted to be major capsid protein as the most likely function and displaying a score ≥ 2 (98% true positive confidence) were selected. The HK97-fold in these proteins was validated combining HHpred and Modeller as described above. HK97-fold MCP proteins were obtained for 31 out of 37 phages in the high-resolution database. The exceptions were Gifsy-2, SIO-2, Basilisk, ΦRSL1, ΦKZ, and SCTP2. This led to a final library of n_{lib} = 635 HK97-fold MCPs associated with tailed phage genome lengths (Data File 2 and Fig. 2b).

The distribution of genome lengths was investigated using the non-parametric kernel density estimation method. To capture accurately the multimodal nature of the phage genome length distribution, the kernel bandwidth was investigated independently for four distinctive genome length groups identified visually: 17–130 kbp, 130–210 kbp, 210–270 kbp, and 270–498 kbp (Supplementary Fig. S2). The Scikit grid search 5-fold cross-validation method [94] was applied to obtain the most likely Gaussian kernel's bandwidth for each group, leading to 1.78 kbp, 3.33 kbp, 1.39 kbp, and 20 kbp, respectively. The four distributions were combined and normalized to obtain a single probability density function of tailed phage genome lengths. The peaks of the distribution were obtained using the find peaks function from the SciPy signal package for Python [125].

The library containing MCPs and the associated T-numbers (MCP/T library) was built as follows (Fig. 2b). For the 31 MCPs found in the high-resolution capsid database, the T-number used in the library was the one associated with the 3D capsid architecture. For the rest of the isolated tailed phage genomes, the T-number was predicted applying the G2T model to the genome length. If the predicted T-number fell within the ranges of one or several overlapping T-numbers regions, the T-number selected was closest to the mean predicted T-number, and the alternative T-numbers were tallied. For T-numbers associated with multiple lattices (for example, T = 12 trihexagonal versus T = 12 hexagonal), each architecture was considered as a potential structure. If the predicted T-number was not within the error margin of a valid icosahedral T-number, the architecture was categorized as "elongated."

MCP-to-capsid model based on similarity (proximity matrix): MCP2C-PM. Protein-protein sequence similarities were obtained for the MCPs in the library using NCBI blastp [13,85], applying the default algorithm parameters except for the e-value threshold, which was chosen to be 0.001 to increase the quality and decrease the effects of randomness for the matches. In any instance where blastp returned more than one score for any pair of phages, the higher similarity score was chosen for the pair. In the MCP/T library, 80% of the data was selected randomly as the training set and the remaining 20% was used as the test dataset (80/20 split). For statistical robustness, 1000 different 80/20 training and test splits were generated. For each major capsid protein sequence in the test set, the T-number predicted corresponded to the T-number associated with the most similar major capsid protein sequence in the training set (proximity matrix). A prediction was considered correct if the T-number predicted coincided with the T-number associated with the major capsid protein in the MCP/T library. The model accuracy was defined as the fraction of correct predictions in the full test dataset. The accuracy was investigated as a function of different minimum similarity thresholds, from 0% to 100% similarity in increments of 10%. In each case, the fraction of predicted architectures was tallied.

MCP phylogenetic tree. The protein sequences of the MCPs in the MCP/T library were aligned using the Clustal Omega webserver (default settings) at https://www.ebi.ac.uk/Tools/msa/clustalo/ [111]. The resulting phylogenetic tree (ClustalW format) was visualized and analyzed using the Interactive Tree of Life (iToL) webserver at https://itol.embl.de [76]. Each MCP in the tree included the associated phage genome length (using the log-linear transfor-

mation G_t = $\log_{10}$(genome length in kbp)-3) and the capsid architecture predicted by the G2T model. Clades displaying common properties were identified qualitatively from the tree internal structure and phage phenotypic data.

MCP-to-capsid model based on random forest: MCP2C-RF. The similarity model introduced above has two important limitations. First, the method cannot predict the capsid architecture for major capsid proteins that have no similarity in the MCP/T library. This is a bottleneck for environmental analysis because most uncultured tailed phage genes have little similarity to genes in public databases [131,71]. Second, the matrix similarity is a computational search method of quadratic order, $O(n_{lib}^2)$, which limits the scalability of the model when increasing the size of the training library, n_{lib}. To circumvent these foreseeable challenges when characterizing environmental data, an alternative machine learning method was investigated and compared. The approach chosen was random forest because it offers a rapid learning process when the training sets are small with respect to the dimensionality of data, and the cost of prediction is independent of the training data set's size [63].

Random forest regression is an ensemble statistical learning algorithm that generates multiple decision trees using a collection of input features as nodes and the value of the dependent variable (output) at the end of the node. To create each of these decision trees, m random observations and f random features are selected from the original data and the corresponding labels used as targets. A final sorting decision is made based on the trees formed by the training data and can then be used to generate a proposed label for each test data point [57,16]. A total of 22 MCP features were used to train the random forest model: protein sequence length, the protein's isoelectric point, and the frequency of each amino acid in the MCP sequence (20 features referred to as the amino acid composition). The isoelectric point was calculated using Biopython's sequence utility package [28]. These protein features have been previously used to identify functions of viral proteins efficiently in machine learning approaches [106,21]. The T-number associated with each major capsid protein in the MCP/T library was used as the label for the random forest classification. T-numbers that were overlapping based on the confidence interval of the G2T model were combined in single classes. Due to the small density of large genomes, architectures T $\geq$ 25 were groped as a single class. An 80/20 training/test split was applied to the library to test the random forest model. The random forest parameters were optimized for accuracy using Scikit's GridSearchCV function [94] using 80% of the library. The top 10 estimators were run 100 times each to verify the aggregate highest average accuracy. This led to a maximum number of 4 features per tree, 250 estimators, a max depth of 20, 1 minimum sample in a leaf, and a minimum sample split of 46, with data bootstrapping, and using a balanced weight distribution. To ensure statistical robustness, the random forest model was then tested selecting 1000 different randomly generated training datasets from the MCP/T library. Given a major capsid protein sequence, a predicted capsid architecture was considered correct if the predicted T-number was within the margin of error (9%) expected associated with the T-number in the MCP/T library. The number of correctly predicted phages was tallied and used to calculate a percentage accuracy for that test set. Both permutation and dropout analysis were performed on all features. The randomization or omission of no single feature caused deviation greater than 8 %. To gain insight in the interpretation of the random forest model, the 22 features of the model were analyzed for the main clades identified in the phylogenetic analysis and compared to the average features in the MCP/T library. Those features departing on average more than a standard deviation from the reference value were identified as significant.

To determine the impact of increasing the training library in the accuracy of the random forest model, the accuracy of the model was assessed for different library sizes and fitted to a mathematical model. The different sizes for the training set were defined as n_i =- n_{lib} i/20 for a total of twenty training sizes, i = 1 to 19. The size of the testing set was n_{lib} – n_i = n_{lib} (1– i/20). For statistical robustness, 1000 different training sets were generated for each size n_i and the mean accuracy was measured in each case, $MACC_i$. The mean accuracy values were fitted to the logarithmic model

$$MACC(n) = g\log_{10} n + h \ . \quad (4)$$

The values of best fit for the parameters g and h were obtained using the robust least squares method. The confidence intervals of the values of best fit were obtained by bootstrapping 10,000 subsets generated randomly from the estimated mean accuracies, $MACC_i$.

Computational performance of MCP2C models. The computational scalability of the proximity matrix similarity (MCP2C-PM) and random forest (MCP2C-RF) models was estimated generating larger artificial libraries. The original MCP/T library (n_{lib} = 617) was sequentially used 15 times, generating 15 artificial libraries with 617 to 10,035 entries. Both models were trained (80/20 split) for 100 different randomly selected training sets for each library size. For each training, the elapsed training time was recorded, and the statistics of the training time were obtained for each model and library size. Then, the T-number of 50 major capsid protein sequences were predicted to tally in each case the elapsed time for the prediction. These time-searches were averaged for each generated model and library size. Linear and quadratic models were fitted to the average times as a function of the library size using least-squares method via numpy polyfit [54]. These fitted models were used to extrapolate the scalability of the two methods for libraries as large as 1,000,000 entries. The elapsed times were obtained on Lenovo laptop with an intel i7 processor and 16 GB RAM.

Capsid architecture prediction from gut metagenomes. 3,173 metagenomically assembled circular genomes (direct terminal repeats $\geq$ 50 bp) and at least two canonical tailed phage markers published in [11] were accessed at ftp://ftp.ncbi.nih.gov/pub/yutinn/benler_2020/gut_phages/ in the NCBI server. The open reading frame sequences (putative proteins) were input to the PhANNs web server [21]. Proteins that displayed major capsid protein function as the highest score were selected. Those proteins with score $\geq$ 2 were further selected (expected accuracy—true positives—of using this score is 98%). When the same MCP was found in similar metagenomic assembled genomes, one representative was kept (dereplication). These selected putative major capsid proteins were run in the MCP2T-RF model to predict capsid architectures.

3. Results

Genome length predicts capsid architecture with 90% accuracy. The power function model, Eq. (2), relating the capsid architecture, T, as a function of the genome length, G, explained 98% of the variance (R^2 = 0.98, n = 37, Fig. 3a). This model is referred to as the genome-to-T-number (G2T) model. In the high-resolution capsid database, the genome lengths, G, ranged from G = 16.7 kilobase pairs (kbp) to 498.0 kbp. The capsid architectures ranged from T = 4 to 52 (see Data File 1). The fitted allometric exponent was 0.71 ± 0.03. This value was consistent with a prior analysis using a smaller dataset (0.68 ± 0.09, n = 23) [82]. The value was also close to the theoretical value, 2/3 $\approx$ 0.67 , expected for quasi-spherical shells packing a genome at a constant density (see supplementary

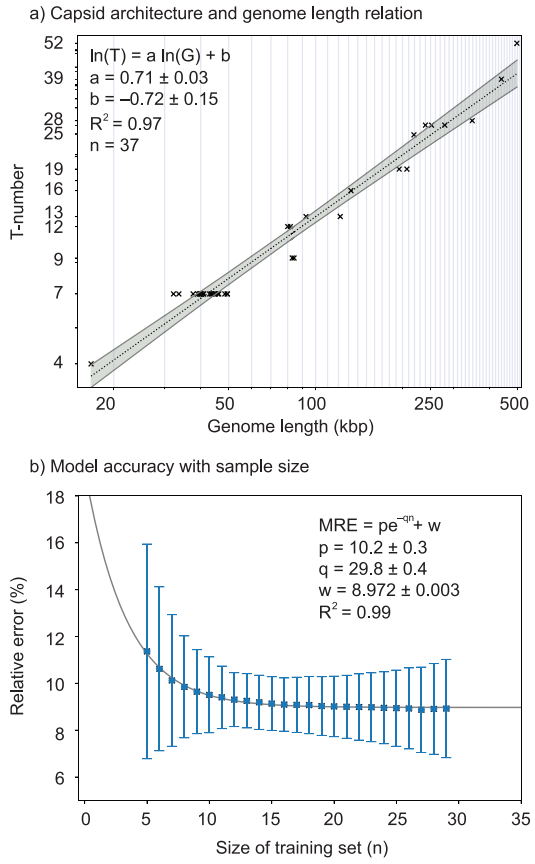

Fig. 3. Genome-to-T-number (G2T) model and accuracy. a) T-number as a function of genome length in log–log scale (natural log) obtained from n = 37 tailed phage capsid 3D reconstructions (black product signs). The data is available in Data File 1. Vertical lines are displayed every 10 kbp as guide to the eye. The dotted black line corresponds to the linear regression of the power function (G2T) model in log–log scale (Eq. (2)). The gray band indicates the 95% confidence interval of the regression. b) Mean relative error, MRE, of the G2T model as a function of the size of the training set, n (blue squares). The error bars represent the standard deviation of the mean relative error. The solid, gray line corresponds to the fitted exponential decay model. a-b) The equations fitted, values of best fit, and coefficient of determination (R^2) are displayed in each legend. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

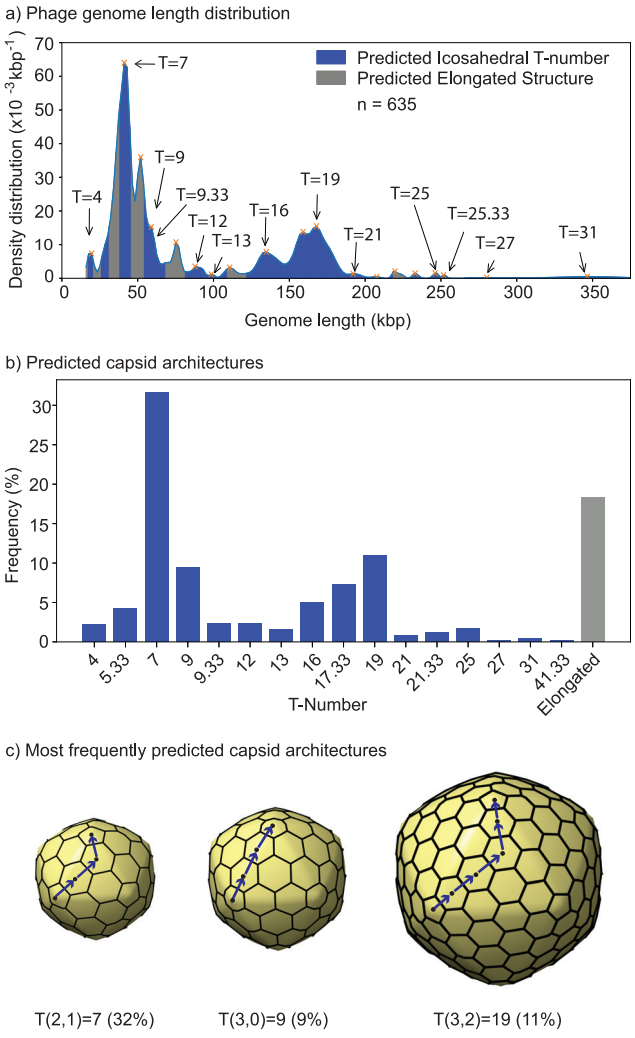

Fig. 4. Putative capsid architectures among phage isolates in the MCP library. a) Probability density distribution of genome lengths (black line). The density was built with Gaussian kernels using multiple bandwidths (see methods). The genome length peaks in the probability density function are indicated with red product signs. Genome length regions predicted to form icosahedral capsids (G2T model) are shaded in blue. Regions associated with putative elongated capsids are shaded in gray. The T-numbers associated with peaks are displayed. b) Frequency of predicted architectures. The bar colors are associated with the shaded regions in panel a). c) 3D models for the three most common predicted capsid architectures. The labels at the bottom display the T-number, *h* and *k* steps, and frequency in percentage. Blue arrows and black dots highlight the steps in the hexagonal lattice. The models were generated with the *hkcage* function in Chimera X [96,82]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

section SI-2 for derivation). The mean relative error of the G2T model was 9% when testing the model using 30 structures for training and 7 for testing, 80/20 split. The analysis of the relative error using different training sizes revealed an initial exponential decay with training size, n, saturating at ~ 9% for n ≥ 25 (R2 = 0.99, Fig. 3b). This implied that the genome length can predict the capsid architecture with 91% accuracy, each T-number is associated with a range of genome lengths that may overlap with nearby T-numbers (Data File 3), and this accuracy is not expected to improve when increasing the number of high-resolution capsid architectures.

Phage isolates display multimodal genome lengths dominated by T = 7, 9, and 19 architectures. The genome length distribution of tailed phage genomes (n_{lib} = 635) displayed a multimodal distribution with 19 peaks (Fig. 4a). The densest genome regions were around ~ 40 kbp and ~ 160 kbp. The G2T model revealed that 15 out of the 19 peaks (55%) were associated with T-number architectures. Several possible T-number ranges overlap, thus yielding more than one possible T-number assignment for 38 % of phages (Supplementary Fig. S3). The remaining four peaks (21 %) were associated with alternative capsid architectures, which were interpreted as elongated architectures. The peak densities of elongated architectures, however, were far less prominent than those associated with icosahedral architectures. The total fraction of elongated architectures among isolates was predicted to be 18 % (Fig. 4b). This number was consistent with the observation of 10% to 20% of elongated architectures among isolates imaged with transmission electron microscopy [3]. Among the remaining 82 % of capsid architectures, which were predicted to be icosahedral, the most frequent capsids were T = 7 (32%), T = 9 (9 %), and T = 19 (11 %) (Fig. 4c). These three architectures combined accounted for 51 % of the putative structures. In the high-resolution database (Data File 1) 20 capsids were T = 7 (54%), no capsids were T = 9 (0%) and two capsids were T = 19 (5%). Therefore, with respect to tailed phage isolates, T = 7 has been over sampled in high-resolution capsid studies, while T = 9 and T = 19 have been under sampled. No tailed phage capsids were predicted to adopt the following T-numbers: 1, 1.33, 3, 25.33, 28, 33.33, 36, 37, 37.33, and 39.

Protein sequence similarity can predict capsid architecture with 75% accuracy when requiring protein–protein similarities above 80%. The analysis of the MCP/T library curated from phage isolates (n_{lib} = 635) revealed that MCPs sharing more than 80% similarity were associated with similar T-number architectures, with a mean relative difference in T-number of 2% (Fig. 5a). The relative T-number difference ranged from 0% to 7% for these highly similar MCPs. As the MCP similarity dropped below 60% the range of associated architectures increased substantially (Fig. 5a and Supplementary Table S1). In the last group, MCP similarities below 20%, the mean relative difference in T-number was 63 % with a broad range ranging from 0% to 699 %. A subset of 14.6 % of the MCPs that shared less than 20% similarity were predicted to form the same capsid architecture. This implies that high protein sequence similarity is a good predictor of capsid architecture, but very distant protein sequences can form the same capsid architecture.

The phylogenetic analysis of the major capsid protein sequences confirmed the observation derived from the initial MCP-MCP similarity analysis. The tree was very divergent due to the overall dissimilarity between proteins (Fig. 5b). Protein clusters displayed similar predicted architectures, but such architecture was not unique and could be find in independent clusters. Three clades contained a larger number of similar proteins, displaying each similar phage genome lengths and capsid architectures (see highlighted groups in Fig. 5c and Data File 5). Clade one (n_{c1} = 95) adopted T = 19 capsids or slightly larger; clade two (n_{c2} = 47) adopted T = 16 and T = 17.33 capsids, and clade three (n_{c3} = 64) adopted mostly elongated architectures and some T = 9 and T = 9.33 architectures (all with similar genome lengths). These architectures were not exclusive of these clades; other small divergent clusters also adopt them. The frequent T = 7 capsids were distributed in multiple groups across the tree. The phylogenetic tree suggested that similar capsid architectures have emerged independently several times during tailed phage evolution. The alignment, Newick format tree, vectorial render of the tree, and nodes associated to the three highlighted clades are provided in Data Files 4–7.

The prediction of capsid architectures based on MCP-MCP similarity (MCP2T-PM model) assigned T-numbers to 98 % of the test set with 70 % accuracy when the proximity did not require a minimum similarity threshold to make a prediction (Fig. 5c). As the similarity percentage required to make a prediction increased, the accuracy increased slightly, reaching 75 % when requiring 90% similarity. However, above similarity thresholds of 20%, the number of possible predictions decreased substantially, reaching 61 % of the test dataset when requiring 90% similarity (Fig. 5b).

MCP amino-acid composition predicts capsid architecture with 74% accuracy. The random forest model (MCP2T-RF) trained using the MCP/T library (n = 508 out of 635 in a 80/20 split) successfully identified on average 95.2% ± 2.2% of icosahedral structures as icosahedral, and 53.4% ± 9.6% of the elongated structures as elongated. (Fig. 6a). The accuracy varied across T-numbers (Fig. 6b). For T= , 7, 16–17.33, and 19, the accuracy was above 80%, while for T = 4 , the accuracy was just below 50%. The average accuracy was 74%. (see Supplementary Fig. S4 for further details on the T-number confusion matrix). The most relevant amino acid sequence features classifying the T-number were the amino acid length (len) and frequencies of glycine (G), alanine (A), and phenylalanine (F) (Supplementary Fig. S5).

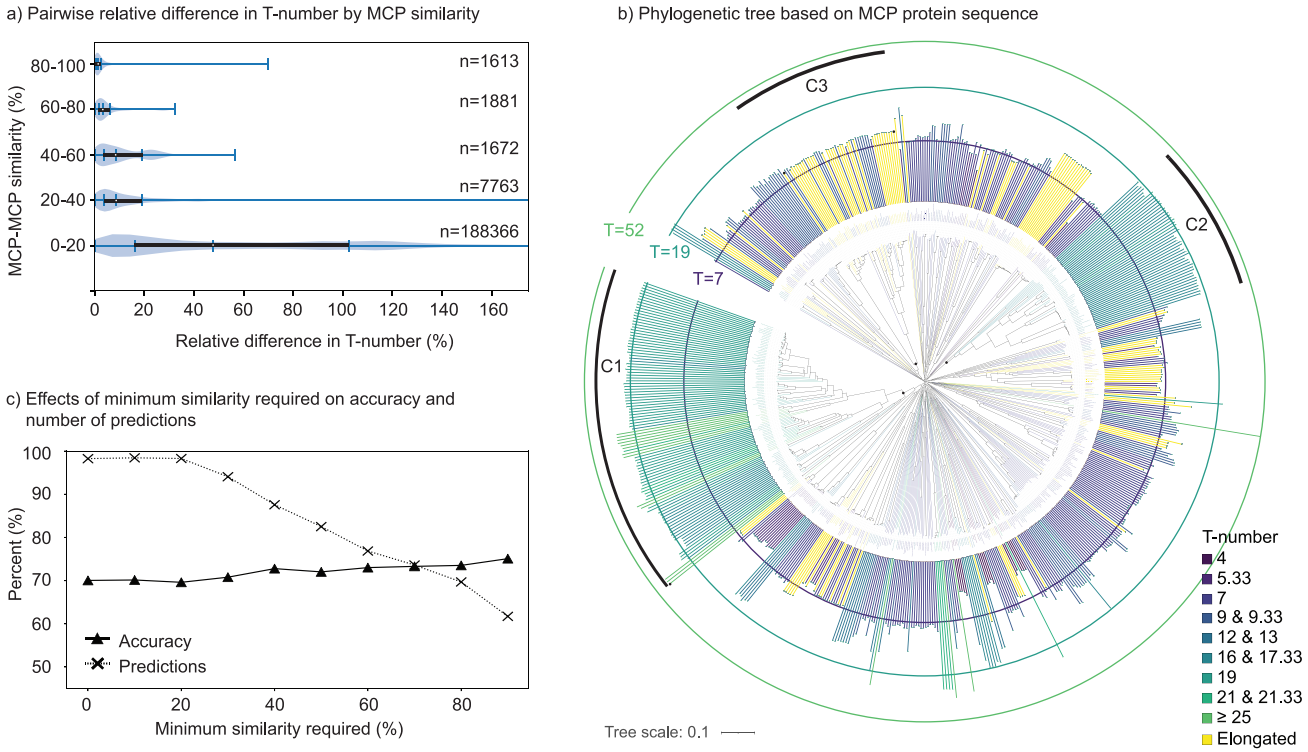

Fig. 5. Association between major capsid protein similarity and capsid architecture. a) Violin plots for the distribution of relative differences in T-number (blue shade) for major capsid protein groups based on protein–protein similarity. The horizontal black line include ticks associated with the 25th quantile, median, and 75th quantile The blue lines capture the full range for each group. b) Unrooted circular phylogenetic tree obtained for the MCPs in the MCP/T library. The inner circle contains the phage names. The bars correspond to the genome length of the associated phage. The colors correspond to the predicted T-numbers and the three circumferences represent the mean genome length associated to the three more frequent T-number architectures. Three large clades are highlighted with black arcs and solid dots on the clade node. Data File 6 contains the tree in vectorial format. c) The percentages of total capsid architectures predicted (product signs) and accurate predictions (black triangles) are plotted as a function of the minimum protein–protein similarity required. The lines connecting points provide a guide to the eye. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

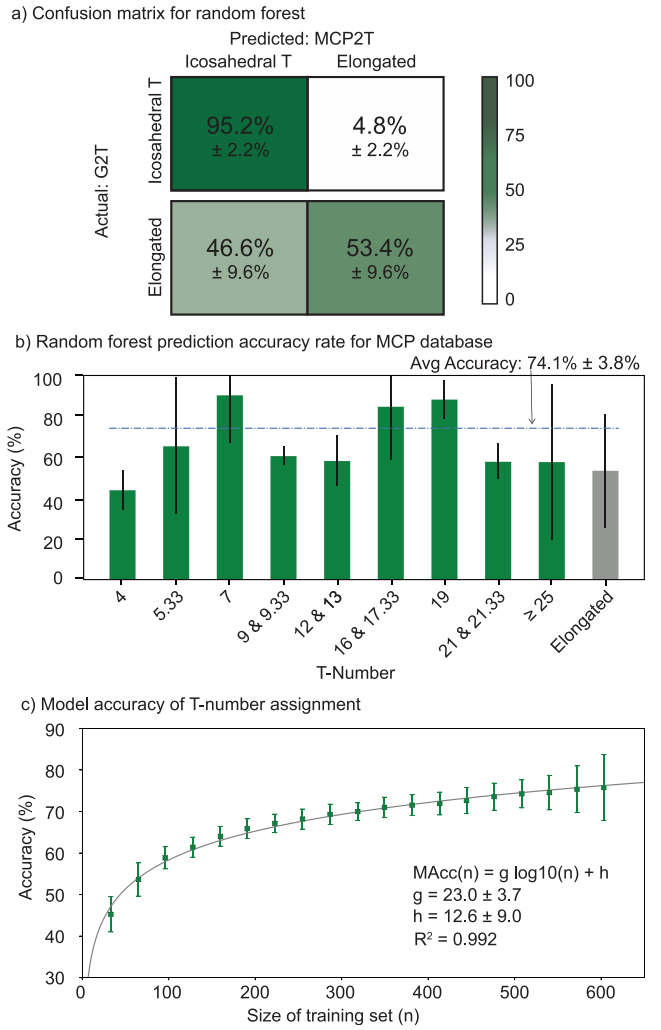

Fig. 6. Capsid architecture prediction from major capsid protein sequence composition. a) Confusion matrix (mean and standard deviation) comparing actual capsid morphologies and predicted capsid morphologies for the major capsid protein-to-T-number random forest (MCP2T-RF) model. The green gradient scale reflects the mean values. b) Accuracy of the MCP2T-RF model predicting different architectures. Bars represent the mean accuracy (green for T-number architectures and gray for elongated architectures). Error bars display the standard deviation. The dashed line indicates the average accuracy. c) Mean (green squares) and standard deviation (error bars) accuracy of the MCP2T-RF model as a function of the size of the training set, n. The solid, gray line is the fitted logarithmic model displayed in the legend (equation, parameters, and coefficient of determination, R^2). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

To obtain further insight about the random forest model, the amino acid sequence features were investigated for the three large clades identified in the phylogenetic analysis. Clade 1 and clade 3 displayed, respectively, three and two amino acid sequence features in their MCPs that departed significantly from the average features in the MCP/T library (Supplementary Table S2). Clade 1 (characterized by T = 19 capsid structures) displayed MCP sequences with a larger average number of amino acids (498 versus 391), average glycine enrichment (9.6% of the sequence versus 7.8%), and average impoverishment of leucine (6.4% of the sequence versus 7.9%). Clade 3′s MCPs (characterized by elongated structures near T = 9 architectures) were on average enriched in tryptophan (1.5% of the sequence versus 0.9%) and impoverished in tyrosine (2.0% of the sequence versus 3.0%).

The accuracy of the model was investigated as a function of the size of the training data set. This identified a logarithmic increase of accuracy with the training size (R^2 = 0.996, Fig. 6c). The accuracy model predicts that reaching a 90% accuracy would require a training set of 2,330 , that is, a library of 2,588 major capsids proteins and putative capsid architectures.

The training time of the random forest model increased linearly with the size of the training data set (slope = 2 ms/datum, R^2 = 1.00, Supplementary Fig. S6a). Training the random forest model with a training set of size 2,330 (predicted to be 90% accurate) would take about 20 s. The increase in training time was about two times less costly than for the similarity model (slope = 4 ms/datum, R^2 = 1.00, Fig. SI-7a). In the random forest model, a single prediction was independent of the training size, approximately 1 ms for a single search (Fig. SI-7b). For the similarity model, the search time was faster for small training sizes, but it increased quadratically with the training size, that is, $O(n^2)$ (Supplementary Fig. S6b). The cross-over time-search was around training size sets of size 10,000, with a search time on the order of 1 ms. Therefore, the random forest model provided a scalable approach.

T = 7 capsids dominate among uncultured gut phages. A total of distinct 1,479 HK97-fold major capsid proteins annotations were identified among 3,181 metagenomically assembled genomes from gut samples containing tailed phage markers and direct terminal repeats (Fig. 7a). The MCP2T-RF model predicted the presence of capsid architectures ranging from T = 4 to T = 31 (Data File 8). The most frequent predicted capsid architecture was T = 7 (48.9%), followed by T ≥ 25 (12.4%), T = 4 (9.4%), and T = 5.33 (9.2%) (Fig. 7b). The frequency of predicted elongated capsids was 1.9% (see Data File 9). The frequency of putative T = 7 capsid architectures in gut metagenomes was ten points larger than those predicted among tailed phage isolates (Fig. 4b and 7b). This was interpreted due to the large presence of integrated prophages in bacterial genomes in the gut[58,83]. The genome length of phages that can integrate as prophages is typically around 45 kbp [14], which is within genome length that we predict to be associated with T = 7 capsids. The large frequency of T ≥ 25 architectures in gut metagenomes was unexpected based on tailed phage isolates, probably because the observation of jumbo phages has been particularly elusive until the emergence of sequencing [12].

4. Discussion

The computational model introduced here confirmed a strong association between the information encoded in the major capsid protein and the capsid architecture of tailed phages. The application of this model to metagenomic data facilitated surveying the putative capsid architectures of tailed phages in the human gut microbiome. The most frequent capsid predicted was T = 7. High-resolution studies have revealed that this architecture is common among tailed phages [118]. Our interpretation is that the high frequency of T = 7 capsids is associated with the prevalence of lysogeny in gut bacteria [109,83]. Temperate tailed phages can integrate in bacterial genomes as prophages, forming lysogenic bacteria that can alter the functionality of microbiomes [68,58]. These prophages are expected to be present in gut metagenomes in addition to free tailed phages. Temperate phages are characterized by adopting genomes around 45 kbp [14], which, based on our model, are expected to be associated with T = 7 capsids, as observed in lambda and other temperate lambdoids [23]. Prophages in bacteria can be domesticated and shortened in genome length [14], but the remaining major capsid protein would indicate that the free version of the prophage was encoding a T = 7 capsid.

The gut metagenome analysis also identified a significant presence of T ≥ 25 capsids with predicted genome lengths above 206 kbp (Fig. 7b). This was an unexpected result because these group of capsids are relatively rare among tailed phage isolates (Fig. 4). Nonetheless, these capsids are considered jumbo phages (above

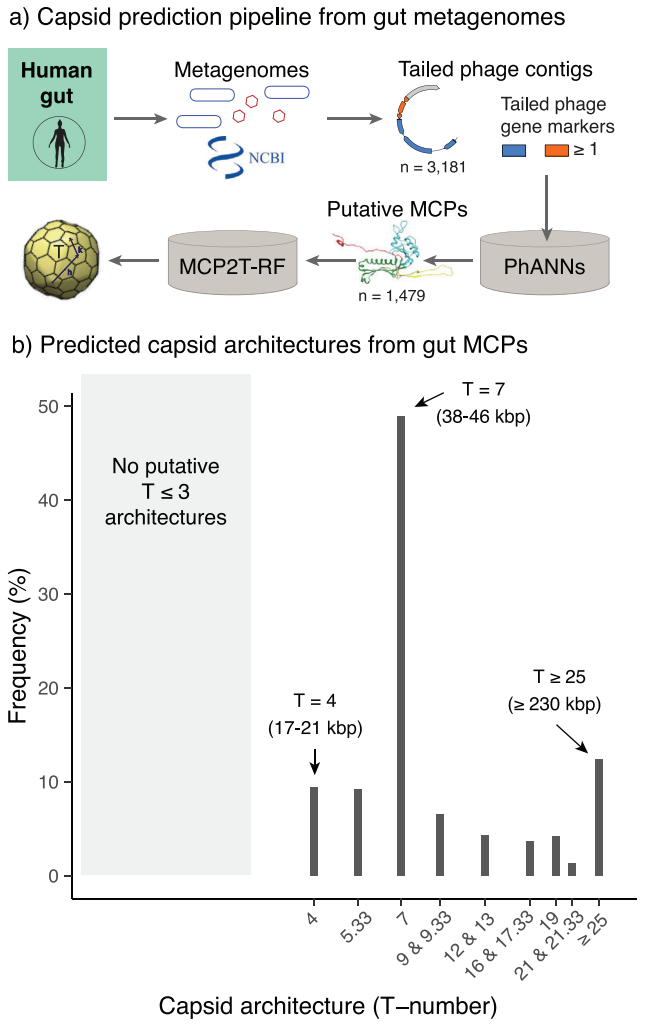

Fig. 7. Capsid architectures predicted in gut metagenomes. a) Bioinformatic pipeline displaying the key steps and tools used to predict tailed phage capsids from gut metagenomic data. b) Frequency of predicted icosahedral capsid architectures. The arrows highlight the three most frequent T-numbers, including the putative genome length range in parenthesis.

200 kbp) [126], and recent studies have discovered that they are far more common than initially expected across ecosystems [45,62,12]. Our analysis indicates that jumbo tailed phages might be particularly prevalent in gut microbiomes, in agreement with recent studies [33] The detailed genomic and structural characterization of jumbo phages might be key to understanding the ecology of phage and bacteria in the human gut. Additionally, the model also predicted more frequent small capsids, T = 4 and T = 5.33 (Fig. 7b) than expected from phage isolates. This also aligns with recent bioinformatic studies indicating that these groups of capsids have been under sampled [82,12]. These small capsids could be the key to understand the evolution of tailed phages and cellular compartments like encapsulins [82]. The application of the G2T model to the circular genome lengths indicated that 43 genomes predicted T = 4 and T = 5.33 capsids in agreement with the MCP2T-RF model; only 56 genomes led to this agreement for gut phage genomes. This result suggests that these 43 candidates are probably complete phage genomes that could provide a great source of information to investigate the structure and evolution of small, tailed phage capsids. The MCP2T-RF model did not predict smaller capsids (T less than 4) because such putative capsids were not present in the MCP/T library, but the observation of small circular genomes among tailed phages suggest that that they could exist [82].

The computational model introduced here is a first step to bridge viral genomic information with viral structural phenotype in microbiomes. However, there are important steps ahead to improve the accuracy of the models. The MCP2T-RF model is projected to reach an accuracy of 90% using a library of 2,600 MCPs and putative T-number architectures (Fig. 6c). However, to go beyond this accuracy, it would be necessary first to improve the underlying genome-to-T-number (G2T) model responsible for building the MCP/T library (Fig. 2). The G2T model currently has an accuracy of 91%, but this error is not projected to be reduced when increasing the number of structures in the high-resolution database (Fig. 3b). This implies that at least one more genome feature would be necessary in addition to the genome length. One compelling direction would be to add the tailed phage packing strategy. Head-full mechanisms tend to pack more DNA than encoded in the genome, while packing signal mechanisms pack exactly the genome length [22,59]. These variations may explain that the empirical exponent in the power-function model is slightly larger than the theoretical prediction (Fig. 3a).

The research introduced here does not clarify the structural reasons why features such as amino acid sequence length as well as glycine and threonine frequencies are so relevant in predicting capsid architecture. Nonetheless, two out of the three big MCP clades identified phylogenetically displayed a few characteristic features, including larger amino acid sequence lengths and a larger content of glycine and tryptophan, which could be important to MCPs associated with large capsids. However, follow-up structural analyses would be necessary to reveal the origin of the selection of MCPs to form specific T-numbers. Additionally, information from other proteins involved in the assembly of tailed phages (like scaffold, minor capsid proteins, and reinforcement proteins) will be necessary to predict more accurately the capsid architecture as well as alternative capsid architectures formed by the same major capsid protein [73,44,32,99]. It is now possible to predict these protein functions from genomic data, but the accuracy is typically lower than for major capsid proteins, and some categories are still hard to predict correctly, like minor capsid proteins [21].

The method described in Fig. 2 could be adapted to also predict the capsid architecture of other viruses. The first key step would be identifying strong allometric relationships between the genome length and capsid architecture of those viruses (Fig. 2a). The analysis of allometric relationship between virion volume and genome length combining all virus types has led to non-optimal statistical results due to the variance between virus groups [30,15,37]. Prior studies indicate that the allometric exponent would vary strongly depending on the virus group [10,37]. A strategy to improve the accuracy of this relationship is separating viruses that use the same capsid protein fold and genome storage strategy[1,70,122,69,101]. The second step would be generating the MCP/T library of capsid proteins and capsid architectures using isolated genomes (Fig. 2b), and the third would be using these libraries to train similar statistical learning methods as those presented here (Fig. 2c). Sequencing technologies are now capable of identifying both DNA and RNA viruses [104,43]. Nonetheless, the diversity of capsid architectures among viruses different than tailed phages is smaller [89]. Thus, other phenotypical features might be more interesting to include in the MCP/T library. The development of bioinformatic pipelines as the one used here would facilitate constant monitoring and analysis of viral capsids of different virus groups in the environment (Fig. 7a).

5. Conclusion

The protein-to-capsid model introduced here predicts the architecture of tailed phages from just one gene (the major capsid pro-

tein) with 74% accuracy. Increasing the library of proteins and putative architectures around 2,600 could increase this accuracy to 90%. The application of this approach in human gut metagenomes predicted the abundance of T = 7 capsids probably associated to temperate phages followed by an unexpected abundance of jumbo capsid architectures ($T \geq 25$) and small architectures (T = 4 and T = 5.33) that have been under sampled among phage isolates and high-resolution tailed phage capsid studies. The method introduced here will facilitate bridging the evolution and selection of tailed phage genomic data with capsid architecture. This would eventually help identify the functions associated with capsids beyond storage capacity.

Funding

The research of D.Y.L., C. B., and A.L. was supported by the National Science Foundation award #1951678 and the Gordon and Betty moore Foundation, GBMF9871, grant https://doi.org/10.37807/GBMF9871 and D.Y.L.'s research was also supported by the a STEM scholarship award funded by the National Science Foundation grant DUE-1259951. The research of M.A.S. was supported by the National Institutes of Health GM110588 and the California Metabolic Research Foundation.

CRediT authorship contribution statement

Diana Y. Lee: Conceptualization, Data curation, Methodology, Writing – original draft. **Caitlin Bartels:** Data curation. **Katelyn McNair:** Data curation. **Robert A. Edwards:** Data curation. **Manal A. Swairjo:** Data curation, Writing – review & editing. **Antoni Luque:** Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank Simon White, Marina Chugunova, Allon Percus, Peter Salamon, Anca Segall, Marcelo Sevilla, Chao Zhi, and Spencer Lank for their insight at different stages of the research project. This article is dedicated to the memory of Donald L. D. Caspar and Aaron Klug for their seminal contribution to the geometrical architecture of viral capsids.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.12.032.

References

[1] Abrescia NGA, Bamford DH, Grimes JM, Stuart DI. Structure Unifies the Viral Universe. Annu Rev Biochem 2012;81(1):795–822. https://doi.org/10.1146/annurev-biochem-060910-095130.
[2] Ackermann HW. Sad State of Phage Electron Microscopy. Please Shoot the Messenger. Microorganisms 2014;2(1):1–10. https://doi.org/10.3390/microorganisms2010001.
[3] Ackermann HW. 5500 Phages Examined in the Electron Microscope. Arch Virol 2007;152(2):227–43. https://doi.org/10.1007/s00705-006-0849-1.
[4] Agirrezabala X, Velázquez-Muriel JA, Gómez-Puertas P, Scheres SHW, Carazo JM, Carrascosa JL. Quasi-Atomic Model of Bacteriophage T7 Procapsid Shell: Insights into the Structure and Evolution of a Basic Fold. Structure 2007;15(4):461–72. https://doi.org/10.1016/j.str.2007.03.004.
[5] Aksyuk AA, Bowman VD, Kaufmann B, Fields C, Klose T, Holdaway HA, et al. Structural investigations of a Podoviridae streptococcus phage C1, implications for the mechanism of viral entry. PNAS 2012;109(35):14001–6. https://doi.org/10.1073/pnas.1207730109.
[6] Aylward FO, Boeuf D, Mende DR, Wood-Charlson EM, Vislova A, Eppley JM, et al. Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. Proc Natl Acad Sci 2017;114(43):11446–51. https://doi.org/10.1073/pnas.1714821114.
[7] Baker ML, Hryc CF, Zhang Q, Wu W, Jakana J, Haase-Pettingell C, et al. Validated near-atomic resolution structure of bacteriophage epsilon15 derived from cryo-EM and modeling. PNAS 2013;110(30):12301–6. https://doi.org/10.1073/pnas.1309947110.
[8] Bayfield OW, Klimuk E, Winkler DC, Hesketh EL, Chechik M, Cheng N, et al. Cryo-EM structure and in vitro DNA packaging of a thermophilic virus with supersized T=7 capsids. PNAS 2019;116(9):3556–61. https://doi.org/10.1073/pnas.1813204116.
[9] Bebeacua C, Lai L, Vegge CS, Brøndsted L, van Heel M, Veesler D, et al. Visualizing a complete Siphoviridae member by single-particle electron microscopy: the structure of lactococcal phage TP901-1. J Virol 2013;87(2):1061–8. https://doi.org/10.1128/JVI.02836-12.
[10] Belyi VA, Muthukumar M. Electrostatic origin of the genome packing in viruses. Proc Natl Acad Sci 2006;103(46):17174–8. https://doi.org/10.1073/pnas.0608311103.
[11] Benler S, Yutin N, Antipov D, Rayko M, Shmakov S, Gussow AB, et al. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. Microbiome 2021;9(1). https://doi.org/10.1186/s40168-021-01017-w.
[12] Berg M, Roux S. Extreme dimensions — how big (or small) can tailed phages be? 407 Nat Rev Microbiol 2021;19(7):407. https://doi.org/10.1038/s41579-021-00574-z.
[13] Blastp [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi.
[14] Bobay L-M, Touchon M, Rocha EPC. Pervasive domestication of defective prophages by bacteria. Proc Natl Acad Sci 2014;111(33):12127–32. https://doi.org/10.1073/pnas.1405336111.
[15] Brandes N, Linial M. Gene overlapping and size constraints in the viral world. Biology Direct 2016;11:26. https://doi.org/10.1186/s13062-016-0128-3.
[16] Breiman L. Random Forests. Machine Learning 2001;1:5–32. https://doi.org/10.1023/A:1010933404324.
[17] Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, et al. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci 2002;99(22):14250–5. https://doi.org/10.1073/pnas.202488399.
[18] Brum JR, Schenck RO, Sullivan MB. Global Morphological Analysis of Marine Viruses Shows Minimal Regional Variation and Dominance of Non-Tailed Viruses. ISME J 2013;7(9):1738–51. https://doi.org/10.1038/ismej.2013.67.
[19] Brum, J.R., Ignacio-Espinoza, J.C., Kim, E., Trubl, G., Jones, R.M., Roux, S., VerBerkmoes, N.C., Rich, V.I., Sullivan, M.B.. Structural proteins in marine viral communities. Proc Natl Acad Sci 113 (9) 2436-2441. **2106.** https://doi.org/10.1073/pnas.1525139113
[20] Bryan D, El-Shibiny A, Hobbs Z, Porter J, Kutter EM. Bacteriophage T4 Infection of Stationary Phase E. coli: Life after Log from a Phage Perspective. Front Microbiol 2016;7:1391. https://doi.org/10.3389/fmicb.2016.01391.
[21] Cantu VA, Salamon P, Seguritan V, Redfield J, Salamon D, Edwards RA, et al. PhANNs, a fast and accurate tool and web server to classify phage structural proteins. PLoS Comput Biol 2020;16(11):e1007845. https://doi.org/10.1371/journal.pcbi.1007845.
[22] Casjens SR, Gilcrease EB. Determining DNA Packaging Strategy by Analysis of the Termini of the Chromosomes in Tailed-Bacteriophage Virions. Bacteriophages. 2009. https://doi.org/10.1007/978-1-60327-565-1_7.
[23] Casjens SR, Hendrix RW. Bacteriophage lambda: Early pioneer and still relevant. Virology 2015;479–480:310–30. https://doi.org/10.1016/j.virol.2015.02.010.
[24] Caspar DLD, Klug A. Physical Principles in the Construction of Regular Viruses. Cold Spring Harb Symp Quant Biol 1962;27(0):1–24.
[25] Chen DH, Baker ML, Hyrc CF, DiMaio F, Jakana J, Weimin W, et al. Structural basis for scaffolding-mediated assembly and maturation of a dsDNA. PNAS 2011;108:1355–60. https://doi.org/10.1073/pnas.1015739108.
[26] Cobarrubia A., Tall J., Crispin-Smith A. & Luque A. Empirical and theoretical analysis of particle diffusion in mucus. *Front. Phys.*; 2021. 9:594306. https://doi.org/10.3389/fphy.2021.594306.
[27] Cobián Güemes AG, Youle M, Cantú VA, Felts B, Nulton J, Rohwer F. Viruses as Winners in the Game of Life. Annu Rev Virol 2016;3(1):197–214. https://doi.org/10.1146/annurev-virology-100114-054952.
[28] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25(11):1422–3. https://doi.org/10.1093/bioinformatics/btp163.
[29] Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. Nat Commun 2017;8(1). https://doi.org/10.1038/ncomms15955.
[30] Cui J, Schlub TE, Holmes EC. An Allometric Relationship between the Genome Length and Virion Volume of Viruses. J Virol 2014;88(11):6403–10. https://doi.org/10.1128/JVI.00362-14.

[31] Dearborn AD, Laurinmaki P, Chandramouli P, Rodenburg CM, Wang S, Butcher SJ, et al. Structure and size determination of bacteriophage P2 and P4 procapsids: function of size responsiveness mutations. J Struct Biol 2012;178(3):215–24.

[32] Dearborn AD, Wall EA, Kizziah JL, Klenow L, Parker LK, Manning KA, et al. Competing scaffolding proteins determine capsid size during mobilization of Staphylococcus aureus pathogenicity islands. eLife. 2017;6. https://doi.org/10.7554/eLife.30822.

[33] Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, et al. Megaphages infect Prevotella and variants are widespread in gut microbiomes. Nat Microbiol 2019;4(4):693–700.

[34] Duda RL, Teschke CM. The Amazing HK97 Fold: Versatile Results of Modest Differences. Current Opinion in Virology 2019;36:9–16. https://doi.org/10.1016/j.coviro.2019.02.001.

[35] Duda RL, Hendrix RW, Huang WM, Conway JF. Shared architecture of bacteriophage SPO1 and herpesvirus capsids. Curr Biol 2006;16(1):R11–3. https://doi.org/10.1016/j.cub.2005.12.023.

[36] Earnshaw WC, Casjens SR. DNA packaging by the double-stranded DNA bacteriophages. Cell 1980;21(2):319–31. https://doi.org/10.1016/0092-8674(80)90468-7.

[37] Edwards KF, Steward GF, Schvarcz CR, Ostling A. Making sense of virus size and the tradeoffs shaping viral fitness. Ecol Lett 2021;24(2):363–73. https://doi.org/10.1111/ele.13630.

[38] Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, Dinsdale EA, et al. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nat Microbiol 2019;4(10):1727–36. https://doi.org/10.1038/s41564-019-0494-6.

[39] Effantin G, Boulanger P, Neumann E, Letellier L, Conway JF. Bacteriophage T5 structure reveals similarities with HK97 and T4 suggesting evolutionary relationships. J Mol Biol 2006;361(5):993–1002. https://doi.org/10.1016/j.jmb.2006.06.081.

[40] Effantin G, Figueroa-Bossi N, Schoehn G, Bossi L, Conway JF. The tripartite capsid gene of Salmonella phage Gifsy-2 yields a capsid assembly pathway engaging features from HK97 and lambda. Virology 2010;402:355–65. https://doi.org/10.1016/j.virol.2010.03.041.

[41] Effantin G, Hamasaki R, Kawasaki T, Bacia M, Moriscot C, Weissenhorn W, et al. Cryo-electron microscopy three-dimensional structure of the jumbo Phage PhiRSL1 infecting the phytopathogen Ralstonia solanacearum. Structure 2013;21:298–305. https://doi.org/10.1016/j.str.2012.12.017.

[42] Evilevitch A. The mobility of packaged phage genome controls ejection dynamics. eLife. 2018;7. https://doi.org/10.7554/eLife.37345.

[43] Fitzpatrick AH, Rupnik A, O'Shea H, Crispie F, Keaveney S, Cotter P. High Throughput Sequencing for the Detection and Characterization of RNA Viruses. Front Microbiol 2021;12:190. https://doi.org/10.3389/fmicb.2021.621719.

[44] Fokine A, Rossmann MG. Molecular architecture of tailed double-stranded DNA phages. Bacteriophage 2014;4(2):e28281. https://doi.org/10.4161/bact.28281.

[45] Fokine A, Kostyuchenko VA, Efimov AV, Kurochkina LP, Sykilinda NN, Robben J, et al. A three-dimensional cryo-electron microscopy structure of the bacteriophage phiKZ head. J Mol Biol 2005;352:117–24. https://doi.org/10.1016/j.jmb.2005.07.018.

[46] Gabler F, Nam S, Till S, Mirdita M, Steinegger M, Söding J, et al. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr Protoc Bioinformat 2020;72(1):e108. https://doi.org/10.1002/cpbi.108.

[47] Gan L, Speir JA, Conway JF, Lander G, Cheng N, Firek BA, et al. Capsid conformational sampling in HK97 maturation visualized by X-ray crystallography and cryo-EM. Structure 2006;14(11):1655–65. https://doi.org/10.1016/j.str.2006.09.006.

[48] Gertsman I, Gan L, Guttman M, Lee K, Speir JA, Duda RL, et al. An unexpected twist in viral capsid maturation. Nature 2009;458(7238):646–50. https://doi.org/10.1038/nature07686.

[49] Gipson P, Baker ML, Raytcheva D, Haase-Pettingell C, Piret J, King JA, et al. Protruding knob-like proteins violate local symmetries in an icosahedral marine virus. Nat Commun 2014;5(1). https://doi.org/10.1038/ncomms5278.

[50] Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. Cell Host Microbe 2020;28(5):724–740.e8. https://doi.org/10.1016/j.chom.2020.08.003.

[51] Gonzalez B, Monroe L, Li K, Yan R, Wright E, Walter T, et al. Phage G structure at 6.1 Å resolution, condensed DNA, and host identity revision to a *Lysinibacillus*. J Mol Biol 2020;432(14):4139–53. https://doi.org/10.1016/j.jmb.2020.05.016.

[52] Grose JH, Belnap DM, Jensen JD, Mathis AD, Prince JT, Merrill BD, et al. The genomes, proteomes, and structures of three novel phages that infect the Bacillus cereus group and carry putative virulence factors. J Virol 2014;88(20):11846–60. https://doi.org/10.1128/JVI.01364-14.

[53] Guo F, Liu Z, Fang P-A, Zhang Q, Wright ET, Wu W, et al. Capsid expansion mechanism of bacteriophage T7 revealed by multistate atomic models derived from cryo-EM reconstructions. Proc Natl Acad Sci 2014;111(43):E4606–14. https://doi.org/10.1073/pnas.1407020111.

[54] Harris CR, Millman KJ, van der Walt StéfanJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585(7825):357–62. https://doi.org/10.1038/s41586-020-2649-2.

[55] Helgstrand C, Wikoff WR, Duda RL, Hendrix RW, Johnson JE, Liljas L. The Refined Structure of a Protein Catenane: The HK97 Bacteriophage Capsid at 3.44Å Resolution. J Mol Biol 2003;334(5):885–99. https://doi.org/10.1016/j.jmb.2003.09.035.

[56] Hildebrand A, Remmert M, Biegert A, Söding J. Fast and accurate automatic structure prediction with HHpred. Proteins 2009;77(Suppl 9):128–32. https://doi.org/10.1002/prot.22499.

[57] Ho, Tin Kam. Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC*, pp. 278–282. August **1995**.

[58] Howard-Varona C, Hargreaves KR, Abedon ST, Sullivan MB. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. The ISME Journal 2017;11(7):1511–20. https://doi.org/10.1038/ismej.2017.16.

[59] Hua J, Huet A, Lopez CA, Toropova K, Pope WH, Duda RL, et al. Capsids and Genomes of Jumbo-Sized Bacteriophages Reveal the Evolutionary Reach of the HK97 Fold. mBio 2017;8(5). https://doi.org/10.1128/mBio.01579-17.

[60] Hua, J., *Capsid Structure and DNA Packing in Jumbo Bacteriophages*, University of Pittsburgh, **2016**, http://d-scholarship.pitt.edu/27666/.

[61] Ionel A, Velázquez-Muriel JA, Luque D, Cuervo A, Castón JoséR, Valpuesta JoséM, et al. Molecular Rearrangements Involved in the Capsid Shell Maturation of Bacteriophage T7. J Biol Chem 2011;286(1):234–42. https://doi.org/10.1074/jbc.M110.187211.

[62] Iyer LM, Anantharaman V, Krishnan A, Burroughs AM, Aravind L. Jumbo Phages: A Comparative Genomic Overview of Core Functions and Adaptions for Biological Conflicts. Viruses 2021;13(1):63. https://doi.org/10.3390/v13010063.

[63] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: springer; **2013** Feb 11.

[64] Jiang W, Baker ML, Jakana J, Weigele PR, King J, Chiu W. Backbone structure of the infectious epsilon 15 virus capsid revealed by electron cryomicroscopy. Nature 2008;451:1130–4. https://doi.org/10.1038/nature06665.

[65] Jin H, Jiang Y-L, Yang F, Zhang J-T, Li W-F, Zhou K, et al. Capsid Structure of a Freshwater Cyanophage Siphoviridae Mic1. Structure 2019;27(10):1508–1516.e3. https://doi.org/10.1016/j.str.2019.07.003.

[66] Joiner KL, Baljon A, Barr J, Rohwer F, Luque A. Impact of bacteria motility in the encounter rates with bacteriophage in mucus. Sci Rep 2019;9:16427. https://doi.org/10.1038/s41598-019-52794-2.

[67] Keen EC, Bliskovsky VV, Malagon F, Baker JD, Prince JS, Klaus JS, et al. Novel "superspreader" bacteriophages promote horizontal gene transfer by transformation. mBio 2017;8(1). https://doi.org/10.1128/mBio.02115-16.

[68] Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, Cobián-Güemes AG, et al. Lytic to temperate switching of viral communities. Nature 2016;531(7595):466–70. https://doi.org/10.1038/nature17193.

[69] Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, et al. Global Organization and Proposed Megataxonomy of the Virus World. Microbiol Mol Biol Rev 2020;84(2). https://doi.org/10.1128/MMBR.00061-19.

[70] Krupovic M, Koonin EV. Multiple origins of viral capsid proteins from cellular ancestors. PNAS 2017;114(12):E2401–10. https://doi.org/10.1073/pnas.1621061114.

[71] Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. Virus Res 2017;239:136–42. https://doi.org/10.1016/j.virusres.2017.02.002.

[72] Lander GC, Baudoux AC, Azam F, Potter CS, Carragher B, Johnson JE. Capsomer dynamics and stabilization in the T¼12 marine bacteriophage SIO-2 and its procapsid studied by CryoEM. Structure 2012;20:498–503. https://doi.org/10.1016/j.str.2012.01.007.

[73] Lander GC, Evilevitch A, Jeembaeva M, Potter CS, Carragher B, Johnson JE. Bacteriophage lambda stabilization by auxiliary protein gpD: timing, location, and mechanism of attachment determined by cryo-EM. Structure 2008;16(9):1399–406. https://doi.org/10.1016/j.str.2008.05.016.

[74] Lara E, Vaqué D, Sà EL, Boras JA, Gomes A, Borrull E, et al. Unveiling the role and life strategies of viruses from the surface to the dark ocean. Sci Adv 2017;3(9). https://doi.org/10.1126/sciadv.1602565.

[75] Leiman PG, Battisti AJ, Bowman VD, Stummeyer K, Mühlenhoff M, Gerardy-Schahn R, et al. The structures of bacteriophages K1E and K1–5 explain processive degradation of polysaccharide capsules and evolution of new host specificities. J Mol Biol 2007;371(3):836–49. https://doi.org/10.1016/j.jmb.2007.05.083.

[76] Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 2021;49(W1):W293–6. https://doi.org/10.1093/nar/gkab301.

[77] Liang G, Bushman FD. The human virome: assembly, composition and host interactions. Nat Rev Microbiol 2021;19(8):514–27. https://doi.org/10.1038/s41579-021-00536-5.

[78] Liu T, Sae-Ueng U, Li D, Lander GC, Zuo X, Jonsson B, et al. Solid-to-fluid–like DNA transition in viruses facilitates infection. Proc Natl Acad Sci 2014;111(41):14675–80. https://doi.org/10.1073/pnas.1321637111.

[79] Liu X, Zhang Q, Murata K, Baker ML, Sullivan MB, Fu C, et al. Structural changes in a marine podovirus associated with release of its genome into Prochlorococcus. Nat Struct Mol Biol 2010;17(7):830–6. https://doi.org/10.1038/nsmb.1823.

[80] Luque A, Reguera D. The structure of elongated viral capsids. Biophys J 2010;98(12):2993–3003. https://doi.org/10.1016/j.bpj.2010.02.051.

[81] Luque, A., Reguera, D. Theoretical Studies on Assembly, Physical Stability and Dynamics of Viruses. Struct Phys Virus 553–595. **2013**. doi:10.1007/978-94-007-6552-8_19

[82] Luque A, Benler S, Lee DY, Brown C, White S. The missing tailed phages: prediction of small capsid candidates. Microorganisms 2020;8:1944. https://doi.org/10.3390/microorganisms8121944.

[83] Luque A, Silveira CB. Quantification of Lysogeny Caused by Phage Coinfections in Microbial Communities from Biophysical Principles. mSystems 2020;5(5): e00353–20. https://doi.org/10.1128/mSystems.00353-20.
[84] Mahmoudabadi G, Milo R, Phillips R. Energetic cost of building a virus. Proc Natl Acad Sci 2017;114(22):E4324–33. https://doi.org/10.1073/pnas.1701670114.
[85] Madden T. The BLAST sequence analysis tool. InThe NCBI Handbook [Internet]. 2nd edition **2013** Mar 15. National Center for Biotechnology Information (US).
[86] Maurice CF. Considering the Other Half of the Gut Microbiome: Bacteriophages. mSystems 2019;4(3). https://doi.org/10.1128/mSystems.00102-19.
[87] McNair, Katelyn (2021): PHANOTOME June 2017 Backup. figshare. Dataset. Doi: 10.6084/m9.figshare.13557770.v1
[88] Meier A, Söding J. Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling., *PLoS Comput Biol.*, Oct 23;11(10). **2015**. https://doi.org/10.1371/journal.pcbi.1004343
[89] Montiel-Garcia D, Santoyo-Rivera N, Ho P, Carrillo-Tripp M, Brooks III CL, Johnson JE, et al. VIPERdb v3.0: a structure-based data analytics platform for viral capsids. Nucleic Acids Res 2021;49(D1):D809–16. https://doi.org/10.1093/nar/gkaa1096.
[90] Nifong RL, Gillooly JF. Temperature Effects on Virion Volume and Genome Length in DsDNA Viruses. Biol Lett 2016;12(3):20160023. https://doi.org/10.1098/rsbl.2016.0023.
[91] Parent KN, Khayat R, Tu LH, Suhanovsky MM, Cortines JR, Teschke CM, et al. P22 coat protein structures reveal a novel mechanism for capsid maturation: stability without auxiliary proteins or chemical crosslinks. Structure 2010;18 (3):390–401. https://doi.org/10.1016/j.str.2009.12.014.
[92] Parent KN, Gilcrease EB, Casjens SR, Baker TS. Structural evolution of the P22-like phages: comparison of Sf6 and P22 procapsid and virion architectures. Virology 2012;427(2):177–88. https://doi.org/10.1016/j.virol.2012.01.040.
[93] Parent KN, Tang J, Cardone G, Gilcrease EB, Janssen ME, Olson NH, et al. 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;464-465:55–66.
[94] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–30. , https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.
[95] Petrovsky S, Dyson ZA, Seviour RJ, Tillett D. Small but Sufficient: the *Rhodococcus* Phage RRH1 Has the Smallest Known Siphoviridae Genome at 14.2 Kilobases. J Virol 2011;86(1):358–63. https://doi.org/10.1128/JVI.05460-11.
[96] Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., Ferrin, T.E.. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. 30(1):70-82. **2021**. https://doi.org/10.1002/pro.3943
[97] PhAnToMe: Phage Annotation Tools and Methods[http://www.phantome.org/] (accessed June 1th 2017).
[98] Pietila MK, Laurinmaki P, Russell DA, Ko C-C, Jacobs-Sera D, Hendrix RW, et al. Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. Proc Nat Acad Sci 2013;110(26):10604–9. https://doi.org/10.1073/pnas.1303047110.
[99] Podgorski J, Calabrese J, Alexandrescu L, Jacobs-Sera D, Pope W, Hatfull G, et al. Structures of three actinobacteriophage capsids: Roles of symmetry and accessory proteins. Viruses 2020;12(3):294. https://doi.org/10.3390/v12030294.
[100] Pride DT, Wassenaar TM, Ghose C, Blaser M. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 2006;7:8. https://doi.org/10.1186/1471-2164-7-8.
[101] Rohwer F, Edwards R. The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol 2002;184(16):4529–35. https://doi.org/10.1128/jb.184.16.4529-4535.2002.
[102] Roos WH, Bruinsma R, Wuite GJL. Physical virology. Nat Phys 2010;6 (10):733–43. https://doi.org/10.1038/nphys1797.
[103] Roux S, Páez-Espino D, Chen I-M, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res 2021;49(D1):D764–75. https://doi.org/10.1093/nar/gkaa946.
[104] Roux S, Solonenko NE, Dang VT, Poulos BT, Schwenck SM, Goldsmith DB, et al. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. PeerJ 2016;4:e2777. https://doi.org/10.7717/peerj.2777.
[105] Santos-Medellin C, Zinke LA, ter Horst AM, Gelardi DL, Parikh SJ, Emerson JB. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. ISME J 2021;15(7):1956–70. https://doi.org/10.1038/s41396-021-00897-y.
[106] Seguritan V, Alves N, Arnoult M, Raymond A, Lorimer D, Burgin AB, et al. Artificial neural networks trained to detect viral and phage structural proteins. PLoS Comput Biol 2012;8(8):e1002657. https://doi.org/10.1371/journal.pcbi.1002657.
[107] Shamash M, Maurice CF. Phages in the infant gut: a framework for virome development during early life. ISME J 2021. https://doi.org/10.1038/s41396-021-01090-x.
[108] Shen PS, Domek MJ, Sanz-Garcia E, Makaju A, Taylor RM, Hoggan R, et al. Sequence and structural characterization of great salt lake bacteriophage CW02, a member of the T7-like supergroup. J Virol 2012;86(15):7907–17. https://doi.org/10.1128/JVI.00407-12.
[109] Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, et al. The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. Cell Host Microbe 2019;26(4):527–541.e5. https://doi.org/10.1016/j.chom.2019.09.009.
[110] Sieradzki E, Ignacio-Espinoza JC, Needham D, Fichot EB, Fuhrman JA. Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. Nat Commun 2019;10:1169. https://doi.org/10.1038/s41467-019-09106-z.
[111] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7(1):539.
[112] Silveira CB, Coutinho FH, Cavalcanti GS, Benler S, Doane MP, Dinsdale EA, et al. Genomic and ecological attributes of marine bacteriophages encoding bacterial virulence genes. BMC Genomics 2020;21(1). https://doi.org/10.1186/s12864-020-6523-2.
[113] Silveira CB, Luque A, Rohwer F. The landscape of lysogeny across microbial community density, diversity, and energetics. Environ Microbiol 2021;23 (8):4098–111. https://doi.org/10.1111/1462-2920.15640.
[114] Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21(7):951–60. https://doi.org/10.1093/bioinformatics/bti125.
[115] Spilman MS, Dearborn AD, Chang JR, Damle PK, Christie GE, Dokland T. A conformational switch involved in maturation of Staphylococcus aureus bacteriophage 80alpha capsids. J Mol Biol 2011;405:863–76. https://doi.org/10.1016/j.jmb.2010.11.047.
[116] Stone NP, Demo G, Agnello E, Kelch BA. Principles for enhancing virus capsid capacity and stability from a thermophilic virus capsid structure. Nat Commun 2019;10(1). https://doi.org/10.1038/s41467-019-12341-z.
[117] Stroupe ME, Brewer TE, Sousa DR, Jones KM. The structure of Sinorhizobium meliloti phage PhiM12, which has a novel T¼19 l triangulation number and is the founder of a new group of T4-superfamily phages. Virology 2014;450–451:205–12. https://doi.org/10.1016/j.virol.2013.11.019.
[118] Suhanovsky MM, Teschke CM. Nature's Favorite Building Block: Deciphering Folding and Capsid Assembly of Proteins with the HK97-Fold. Virology 2015;479–480:487–97. https://doi.org/10.1016/j.virol.2015.02.055.
[119] Sulcius S, Staniulis J, Paskauskas R. Morphology and distribution of phage-like particles in a eutrophic boreal lagoon. Oceanologia 2011;53(2):587–603. https://doi.org/10.5697/oc.53-2.587.
[120] Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW, et al. Prevalence and Evolution of Core Photosystem II Genes in Marine Cyanobacterial Viruses and Their Hosts. PLoS Biol 2006;4(8):e234. https://doi.org/10.1371/journal.pbio.0040234.
[121] Touchon M, Moura de Sousa JA, Rocha EPC. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. Curr Opin Microbiol 2017;38:66–73.
[122] Twarock R, Luque A. Structural Puzzles in Virology Solved with an Overarching Icosahedral Design Principle. Nat Commun 2019;10(1):4414. https://doi.org/10.1038/s41467-019-12367-3.
[123] White HE, Sherman MB, Brasilès S, Jacquet E, Seavers P, Tavares P, et al. Capsid structure and its stability at the late stages of bacteriophage SPP1 assembly. J Virol 2012;86(12):6768–77. https://doi.org/10.1128/JVI.00412-12.
[124] Wikoff WR, Liljas L, Duda RL, Tsuruta H, Hendrix RW, Johnson JE. Topologically Linked Protein Rings in the Bacteriophage HK97 Capsid. Science 2000;289(5487):2129–33. https://doi.org/10.1126/science.289.5487.2129.
[125] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods 2020;17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2.
[126] Yuan Y. & Gao M. Jumbo bacteriophages: An overview. Front Microbiol 8 (403), **2017**. https://doi.org/10.3389/fmicb.2017.00403
[127] Zhang X, Guo H, Jin L, Czornyj E, Hodes A, Hui WH, et al. A new topology of the HK97-like fold revealed in Bordetella bacteriophage by cryoEM at 3.5 Å resolution. eLife 2013;2:. https://doi.org/10.7554/eLife.01299.001e01299.
[128] Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Mol Biol 2018;430(15):2237–43. https://doi.org/10.1016/j.jmb.2017.12.007.
[129] Zinder ND, Lederberg J. Genetic exchange in Salmonella. J Bacteriol 1952;64 (5):679–99. https://doi.org/10.1128/JB.64.5.679-699.1952.
[130] Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, et al. Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy of Sciences 2002;99(22):14250–5. https://doi.org/10.1073/pnas.202488399.
[131] Pope WH, Weigele PR, Chang J, Pedulla ML, Ford ME, Houtz JM, et al. Genome sequence, structural proteins, and capsid organization of the cyanophage Syn5: a "horned" bacteriophage of marine synechococcus. Journal of molecular biology 2007;368(4):966–81. https://doi.org/10.1016/j.jmb.2007.02.046.
[132] Liu X, Shi M, Kong S, Gao Y, An C. Cyanophage Pf-WMP4, a T7-like phage infecting the freshwater cyanobacterium Phormidium foveolarum: complete genome sequence and DNA translocation. Virology 2007;366(1):28–39. https://doi.org/10.1016/j.virol.2007.04.019.