Predicting the capsid architecture of phages from metagenomic data

Diana Y. Lee1,2, Caitlin Bartels1,3, Katelyn McNair1,2, Robert A. Edwards1,2,3,4, Manal A. Swairjo1,5, and Antoni Luque1,2,6,\*

1 Viral Information Institute, San Diego State University, 5500 Campanile Drive, San Diego, CA, 92182, USA.

2 Computational Science Research Center, San Diego State University, 5500 Campanile Drive, San Diego, CA, 92182, USA.

3 Department of Biology, San Diego State University, 5500 Campanile Drive, San Diego, CA, 92182, USA.

4 Flinders Accelerator for Microbiome Exploration, Flinders University, Bedford Park, GPO Box 2100, Adelaide 5001, South Australia, Australia.

5 Department of Chemistry, San Diego State University, 5500 Campanile Drive, San Diego, CA, 92182, USA.

6 Department of Mathematics & Statistics, San Diego State University, 5500 Campanile Drive, San Diego, CA, 92182, USA.

\* Corresponding author: [aluque@sdsu.edu](mailto:aluque@sdsu.edu)

**Abstract**

Tailed phages are viruses that infect bacteria and are the most abundant biological entity 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≥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.

**Keywords**: tailed bacteriophages, icosahedral capsids, physical modeling, machine learning, metagenomes, gut microbiome, viral ecology, physical virology.

**INTRODUCTION**

Tailed phages are viruses that infect bacteria and have evolved an extremely diverse set of protein capsid architectures to protect their infective genome (Luque et al. 2020; Berg and Roux 2021). Tailed phage capsids sizes range from 40 nm to 180 nm in diameter (Petrovsky et al. 2011; Suhanovsky and Teschke 2015; Gonzalez et al. 2020). The internal volumes of these capsids accommodate genomes spanning three orders of magnitude in length, from 5-10 kilobase pairs (kbp) to 735 kbp (Luque et al. 2020; Iyer et al. 2021). The diversity in genome length and genomic content of tailed phages may explain their key role in regulating ecosystems (Maurice 2019; Silveira et al. 2021), in promoting the evolution of microbes (Zinder and Lederberg 1952; Keen et al. 2016; Touchon et al. 2017), in participating strongly in the planetary carbon cycle (Lara et al. 2017), and in becoming the most abundant biological entity on the planet (Cobián-Güemes et al. 2016). 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 (Brum et al. 2016). The number of phages isolated and studied both genetically and structurally (Duda and Teschke 2019; Krupovic and Koonin 2017) represent a very small sample compared to the gargantuan number of viruses evolving in the environment (Cobián-Güemes et al. 2016; Aylward et al. 2017; Coutinho et al. 2017; Edwards et al. 2019; Shkoporov et al. 2019; Gregory et al. 2020; Roux et al. 2021; Benler et al. 2021; Shamash and Maurice 2021). 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 (Sulcius et al 2011; Brum et al. 2013). There are trade-offs in selecting capsid sizes that are difficult to disentangle (Edwards et al. 2021). An increase in temperature may promote smaller genomes among viruses and other organisms (Nifong and Gilooly, 2016), but larger genomes encode more genes, which can enhance the survival of both phages and their hosts (Sullivan et al. 2006; Sieradzki et al. 2019; Silveira et al. 2020). On the other hand, larger genomes and their associated larger capsids are more costly energetically, which can compromise their replication in limited growth conditions (Bryan et al. 2016; Mahmoudabadi et al. 2017). Additionally, an increase in size reduces virus diffusivity (Cobarrubia et al. 2021), which can negatively impact their infectivity (Joiner et al. 2019). 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 (Roos et al. 2010; Luque and Reguera 2013; Fokine and Rossmann 2014; Suhanovsky and Teschke 2015; Evilevitch 2018; Luque et al. 2020).

The majority—80% to 90%—of tailed phage capsids are quasi-spherical (Ackermann 2007). The remaining tailed phages adopt elongated capsids with icosahedral caps (Ackermann 2007; Luque and Reguera 2010). Among tailed phages, the capsids are built from multiple copies of the major capsid protein, which systematically adopt the HK97-fold (Wikoff et al. 2000; Pietilä et al. 2013; Duda and Teschke 2019). Capsid proteins in tailed phages are organized following hexagonal and trihexagonal icosahedral lattices, Figure 1 (Twarock and Luque 2019; Podgorski et al. 2020; Luque et al. 2020), and the double-stranded DNA genome is packed in the capsid at quasicrystalline densities (Earnshaw and Casjens 1980; Liu et al. 2014; Luque et al. 2020). 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 (Caspar and Klug 1962; Twarock and Luque 2019). The number of major capsid proteins is 60T0 (Figure 1), where T0 represents the classic T-number:

|  |  |
| --- | --- |
| . | (1) |

In the generalized theory for icosahedral capsids, the T-number for the hexagonal lattice is Thex = T0, and the T-number for the trihexagonal lattice is Ttri = 4/3 T0 (Twarock and Luque 2019). The factor 4/3 ≈ 1.33 accounts for the additional surface associated with 60T0 minor capsid proteins inserted as trimers in the trihexagonal lattice (Figure 1). Empirical and bioinformatic studies indicate that tailed phages can adopt capsid architectures from putative T = 1.33 capsids to T = 52 capsids (Suhanovsky and Teschke 2015; Luque et al. 2020). The T-number follows an allometric relationship with the genome length with an approximate exponent of 2/3 ≈ 0.67 because the T-number is proportional to the capsid surface and the genome length is proportional to the capsid volume (Luque et al. 2020). 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 of 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 (Breitbart et al. 2002; Silveira et al. 2020; Roux et al. 2021; Liang and Bushman 2021; Santos-Medellin et al. 2021). 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 (Figure 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 (Figure 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 (Figure 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) (Figure 2c). Finally, these statistical learning models were applied to metagenomic data to infer the capsid architecture of uncultured tailed phages in the human gut.

**METHODS**

The GitHub repository <https://github.com/luquelab/Lee_et_al_journal_yyyy> 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.** Thegenome-to-T-number (G2T) model is a physical model that predicts the capsid architecture (T-number) of a tailed phage from its genome length (Figure 2a). The G2T model was introduced in Luque et al. 2020. The model relies on the empirically and theoretically justified physical allometric relation between the genome length and capsid architecture of tailed phages (Luque et al. 2020). 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 (Suhanovsky and Teschke 2015), the icosahedral capsid database VIPERdb (Montiel-Garcia et al 2021), and four recently reconstructed tailed phages displaying new T-numbers: the jumbo tailed phage SCTP2 (Hua et al. 2017) and P74-26, P23-45, and Mic1 (Stone et al. 2019; Bayfield et al. 2019; Jin et al. 2019). 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 (Twarock and Luque 2019). The final high-resolution database included nHR = 37 tailed phage capsid structures (Table 1 and Data File 1).

*Statistical model.* A power function model related the T-number as a function of the genome length, *G*. Here, *b* was the prefactor constant, *a* the allometric exponent, and *G0* the reference units of G, *G0* = 1 kbp. This allometric relationship was empirically and theoretically established previously for a smaller number of tailed phages (Hua et al. 2015; Luque et al, 2020). The allometric relationship is a consequence of the constant density of the genome stored in tailed phage capsids and constant surface of the major protein on the capsid exterior (Luque et al, 2020). The theory predicts an allometric exponent *ath = 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:

|  |  |
| --- | --- |
|  | (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 (Pedregosa et al. 2011). 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 Figure 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 (nHR – 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 > nHR), the mean relative error was fitted to the exponential model

|  |  |
| --- | --- |
|  | (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 (Virtanen et al. 2020). 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 = nHR = 37, that is, ΔT = T·MRE(nHR).

**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 (PhAnToMe 2017). 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 (Zimmermann et al 2018, Gabler et al 2020, Söding J. 2005, Hildebrand et al 2009, Meier and Söding 2015). The folded models were inspected visually. Only those major capsid proteins displaying the canonical features of the HK97-fold were selected (Suhanovsky and Teschke, 2015). 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 (Cantú et al. 2020). 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 *nlib* = 635 HK97-fold MCPs associated with tailed phage genome lengths (Data File 2 and Figure 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 Figure S2). The Scikit grid search 5-fold cross-validation method (Pedregosa et al. 2011) 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 (Virtanen et al. 2020).

The library containing MCPs and the associated T-numbers (MCP/T library) was built as follows (Figure 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. , andt 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 (Madden 2013), 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/ (Sievers et al. 2011). The resulting phylogenetic tree (ClustalW format) was visualized and analyzed using the Interactive Tree of Life (iToL) webserver at https://itol.embl.de (Letunic and Bork, 2021). Each MCP in the tree included the associated phage genome length (using the log-linear transformation Gt = log10(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 (Krishnamurthy and Wang D, 2017). Second, the matrix similarity is a computational search method of quadratic order, *O(nlib2)*, which limits the scalability of the model when increasing the size of the training library, *nlib*. 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 (James et al. 2013).

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. (Ho 1995, Breiman 2001). 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 (Cock et al, 2009. These protein features have been previously used to identify functions of viral proteins efficiently in machine learning approaches (Seguritan et al. 2012, Cantú et al. 2020). 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 ≥ 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 (Pedregosa et al. 2011) 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 ni = nlib i/20 for a total of twenty training sizes, *i* = 1 to 19. The size of the testing set was nlib – ni = nlib (1– i/20). For statistical robustness, 1000 different training sets were generated for each sizeni and the mean accuracy was measured in each case, MACCi . The mean accuracy values were fitted to a logarithmic model

|  |  |
| --- | --- |
|  | (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, *MACCi*.

**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 (*nlib* = 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 (Harris et al. 2020). 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 16GB RAM.

**Capsid architecture prediction from gut metagenomes.** 3,173 metagenomically assembled circular genomes (direct terminal repeats ≥ 50 bp) and at least two canonical tailed phage markers published in Benler et al. 2021 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 (Cantú et al. 2020). Proteins that displayed major capsid protein function as the highest score were selected. Those proteins with score ≥ 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.

**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 (R2 = 0.98, n = 37, Figure 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) (Luque et al. 2020). The value was also close to the theoretical value, 2/30.67, expected for quasi-spherical shells packing a genome at a constant density (see supplementary 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, Figure 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 (nlib =635) displayed a multimodal distribution with 19 peaks (Figure 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 Figure 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% (Figure 4b). This number was consistent with the observation of 10% to 20% of elongated architectures among isolates imaged with transmission electron microscopy (Ackermann 2007). 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%) (Figure 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 (nlib = 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% (Figure 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 (Figure 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 (Figure 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 Figure 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 (Figure 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 (Figure 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. (Figure 6a). The accuracy varied across T-numbers (Figure 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 Figure 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 Figure S5).

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 (R2 = 0.996, Figure 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 milliseconds/datum, R2 = 1.00, Supplementary Figure S6a). Training the random forest model with a training set of size 2,330 (predicted to be 90% accurate) would take about 20 seconds. The increase in training time was about two times less costly than for the similarity model (slope = 4 milliseconds/datum, R2 = 1.00, Figure SI-7a). In the random forest model, a single prediction was independent of the training size, approximately 1 millisecond for a single search (Figure 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(n2)* (Supplementary Figure S6b). The crossover time-search was around training size sets of size 10,000, with a search time on the order of 1 millisecond. 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 (Figure 7a). The MCP2T-RF modelpredicted 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%) (Figure 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 (Figures 4b and 7b). This was interpreted due to the large presence of integrated prophages in bacterial genomes in the gut (Howard-Varona et al. 2017; Luque and Silveira 2020). The genome length of phages that can integrate as prophages is typically around 45 kbp (Bobay et al. 2014), 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 (Berg and Roux 2021).

**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 (Suhanovsky and Teschke 2015), but the number of hits observed in gut metagenomes exceeded the initial expectation. Our interpretation is that the high frequency of T=7 capsids is associated with the prevalence of lysogeny in gut bacteria (Shkoporov et al. 2019; Luque and Silveira, 2021). Temperate tailed phages can integrate in bacterial genomes as prophages, forming lysogenic bacteria that can alter the functionality of microbiomes (Knowles et al. 2016; Howard-Varona et al. 2017). 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 (Bobay et al 2014), which, based on our model, are expected to be associated with T=7 capsids, as observed in lambda and other temperate lambdoids (Casjens and Hendrix 2015). Prophages in bacteria can be domesticated and shortened in genome length (Bobay et al 2014), 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 (Figure 7b). This was an unexpected result because these group of capsids are relatively rare among tailed phage isolates (Figure 4). Nonetheless, these capsids are considered jumbo phages (above 200 kbp), and recent studies have discovered that they are far more common than initially expected across ecosystems (Fokine et al. 2005; Iyer et al 2021; Berg and Roux 2021). Our analysis indicates that jumbo tailed phages might be particularly prevalent in gut microbiomes and the detailed genomic and structural characterization of this group 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 (Figure 7b) than expected from phage isolates. This also aligns with recent bioinformatic studies indicating that these groups of capsids have been under sampled (Luque et al. 2020; Berg and Roux 2021). These small capsids could be the key to understand the evolution of tailed phages and cellular compartments like encapsulins (Luque et al. 2020). 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 < 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 (Luque et al. 2020).

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 (Figure 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 (Figure 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 (Figure 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 (Casjens and Gilcrease 2009; Hua et al. 2017). These variations may explain that the empirical exponent in the power-function model is slightly larger than the theoretical prediction (Figure 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 (Lander et al 2008; Fokine and Rossmann 2014; Dearborn et al. 2017; Podgorski et al 2020). 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 (Cantú *et al.* 2020).

The method described in Figures 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 (Figure 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 (Cui 2014, Brandes and Linial 2016, Edwards et al. 2021). Prior studies indicate that the allometric exponent would vary strongly depending on the virus group (Belyi and Muthukumar 2006; Edwards *et al.* 2020). A strategy to improve the accuracy of this relationship is separating viruses that use the same capsid protein fold and genome storage strategy (Abrescia et al. 2012; Krupovic and Koonin 2017; Twarock and Luque, 2019; Koonin et al. 2020). The second step would be generating the MCP/T library of capsid proteins and capsid architectures using isolated genomes (Figure 2b), and the third would be using these libraries to train similar statistical learning methods as those presented here (Figure 2c). Sequencing technologies are now capable of identifying both DNA and RNA viruses (Roux et al. 2016; Fitzpatrick et al. 2021). Nonetheless, the diversity of capsid architectures among viruses different than tailed phages is smaller (Montiel-Garcia *et al.* 2021). 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 (Figure 7a).

**CONCLUSION**

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

**Author Contributions:** Conceptualization, D.Y.L. and A.L.; data curation, D.Y.L., C.B., K.M., R.E., M.S.; methodology, D.Y.L. and A.L.; analysis, D.Y.L. and A.L.; writing—original draft preparation, D.Y.L.; writing—review and editing, A.L. and M.S.

**Funding.** The research of D.Y.L., C. B., and A.L.’s was supported by National Science Foundation Award 1951678 from the Mathematical Biology Program in the Division of Mathematical Sciences. The research of M.A.S. was supported by the National Institutes of Health GM110588 and the California Metabolic Research Foundation.

**Acknowledgements**: We would like to thank Marina Chugunova, Allon Percus, Peter Salamon, Anca Segall, Marcelo Sevilla, Chao Zhi, and Spencer Lank for their insight at different stages of the project.

**REFERENCES**

Abrescia, N.G.A.; Bamford, D.H.; Grimes, J.M.; Stuart, D.I.. Structure Unifies the Viral Universe. Annual Review of Biochemistry, 81(1), 795–822. 2012. https://doi.org/10.1146/annurev-biochem-060910-095130

Ackermann H.W. Sad State of Phage Electron Microscopy. Please Shoot the Messenger. *Microorganisms.* 2(1):1-10. **2014**. https://doi.org/10.3390/microorganisms2010001

Ackermann, H.W. 5500 Phages Examined in the Electron Microscope. *Arch. Virol.*, *152* (2), 227–243. **2007**. https://doi.org/10.1007/s00705-006-0849-1.

Agirrezabala, X., Velázquez-Muriel, J.A., Gómez-Puertas, P., Scheres, S. H. W., Carazo, J. M., Carrascosa, J. L., Quasi-Atomic Model of Bacteriophage T7 Procapsid Shell: Insights into the Structure and Evolution of a Basic Fold. *Structure*, 15(4) 461-472. **2007.** https://doi.org/10.1016/j.str.2007.03.004

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. USA*, 109,14001–14006. **2012.** https://doi.org/10.1073/pnas.1207730109

Aylward, F.O., Boeuf, D., Mende, D.R., Wood-Charlson, E.M., Vislova, A., Eppley, J.M., Romano, A.E. DeLong, E.F.. Diel cycling and long-term persistence of viruses in the ocean’s euphotic zone. *Proc. Natl. Acad. Sci.,* 114 (43) 11446-11451. **2017**. https://doi.org/10.1073/pnas.1714821114

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., Chiu, W., Validated near-atomic resolution structure of bacteriophage epsilon15 derived from cryo-EM and modeling. *Proc.Natl.Acad.Sci. USA,* 110,12301–12306. **2013** https://doi.org/10.1073/pnas.1309947110

Bayfield, O.W., Klimuk, E., Winkler, D.C., Hesketh, E.L., Chechik, M., Cheng, N., Dykeman, E.C., Minakhin, L., Ranson, N.A., Severinov, K., Steven, A.C., Antson, A.A., Cryo-EM structure and in vitro DNA packaging of a thermophilic virus with supersized T=7 capsids, *Proc.Natl.Acad.Sci. USA,* 116 (9) 3556-3561.**2019**. https://doi.org/10.1073/pnas.1813204116

Bebeacua, C., Lai, L., Vegge, C.S., Brondsted, L., vanHeel, M., Veesler, D., Cambillau, C., Visualizing a complete Siphoviridae member by single-particle electron microscopy: the structure of lactococcal phage TP901-1. *Journal of Virology,* 87,1061–1068. **2012**. <https://doi.org/10.1128/JVI.02836-12>

Belyi, V.A. and Muthukumar, M. Electrostatic origin of the genome packing in viruses. *Proceedings of the National Academy of Sciences*, 103(46), 17174-17178. **2006**. [https://doi.org/10.1073/pnas.0608311103](about:blank)

Benler, S., Yutin, N., Antipov, D., Rayko, M., Shmakov, S., Gussow, A.B., Pevzner, P., Koonin, E.V., Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome*, 9, 78. **2021**. https://doi.org/10.1186/s40168-021-01017-w

Berg, M., Roux, S. Extreme dimensions — how big (or small) can tailed phages be?. *Nat Rev* *Microbiol.* **2021.** 19, 407. https://doi.org/10.1038/s41579-021-00574-z

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

Bobay, L-M., Touchon, M., Rocha, E.P.C. Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci.*, 111 (33) 12127-12132. **2014**. https://doi.org/10.1073/pnas.1405336111

Brandes N. & Linial M., Gene overlapping and size constraints in the viral world. *Biology Direct*, 11:26. **2016**. https://doi.org/10.1186/s13062-016-0128-3

Breiman L., Random Forests. *Machine Learning*, (1): 5–32. **2001**. https://doi.org/10.1023/A:1010933404324

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., Rohwer, F.. Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci.*, 99 (22) 14250-14255. **2002**. https://doi.org/10.1073/pnas.202488399

Brum J. R., Schenck R. O. & Sullivan M. B. Global Morphological Analysis of Marine Viruses Shows Minimal Regional Variation and Dominance of Non-Tailed Viruses. *ISME J,* 7 (9), 1738–1751. **2013**. https://doi.org/10.1038/ismej.2013.67.

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

Bryan, D., El-Shibiny, A., Hobbs, Z., Porter, J., Kutter, E.M.. Bacteriophage T4 Infection of Stationary Phase E. coli: Life after Log from a Phage Perspective. *Frontiers in Microbiology* 7, 1391. **2016**. https://doi.org./10.3389/fmicb.2016.01391

Cantú, V. A., Salamon, P., Seguritan, V., Redfield, J., Salamon, D., Edwards, R. A., Segall, A., PhANNs, a fast and accurate tool and web server to classify phage structural proteins. *PLoS Comput Biol,* 16(11): e1007845. **2020**. https://doi.org/10.1371/journal.pcbi.1007845

Casjens, S.R., Gilcrease, E.B., 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

Casjens, S.R.; Hendrix, R.W.. Bacteriophage lambda: Early pioneer and still relevant. *Virology*, 479-480, 310–330. **2015**. doi:10.1016/j.virol.2015.02.010

Caspar D. L. & Klug A. Physical Principles in the Construction of Regular Viruses. *Cold Spring Harb. Symp. Quant. Biol.*, *27*, 1–24 **1962**. https://doi.org/10.1101/sqb.1962.027.001.005

Chen, D. H., Baker, M. L., Hyrc, C. F., DiMaio, F., Jakana, J., Weimin, W., Dougherty, M., Haase-Pettingell, C., Schmid, M. F., Jiang, W., Baker, D., King, J., Chiu, W., Structural basis for scaffolding-mediated assembly and maturation of a dsDNA, *Proc. Natl. Acad. Sci. USA*, 108, 1355–1360. **2011.** https://doi.org/10.1073/pnas.1015739108

Cobarrubia A., Tall J., Crispin-Smith A. & Luque A. Empirical and theoretical analysis of particle diffusion in mucus. *Front. Phys.*, 9:594306*.* <https://doi.org/10.3389/fphy.2021.594306>.

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* *3* (1), 197–214. **2016**. https://doi.org/10.1146/annurev-virology-100114-054952.

Cock P.A., Antao T., Chang J.T., Chapman B.A., Cox C.J., Dalke A., Friedberg I., Hamelryck T., Kauff F, Wilczynski B., de Hoon M.J.L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25, 1422-1423. 2009. https://doi.org/10.1093/bioinformatics/btp163

Coutinho, F.H., Silveira, C.B., Gregoracci, G.B., Thompson, C.C., Edwards, R.A., Brussaard, C.P.D., Dutilh, B.E., Thompson, F.L.. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun*. 8, 15955. **2017**. https://doi.org/10.1038/ncomms15955

Cui J., Schlub T. E., & Holmes E. C. An Allometric Relationship between the Genome Length and Virion Volume of Viruses. *J Virol*, *88* (11), 6403–6410. **2014.** https://doi.org/10.1128/JVI.00362-14.

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. *Journal of Structural Biology,* 178,215–224. **2012**. https://doi.org/10.1016/j.jsb.2012.04.002

Dearborn, A.D., Wall, E.A., Kizziah, J.L., Klenow, L., Parker, L.K., Manning, K.A., Spilman, M.S., Spear, J.M., Christie, G.E., Dokland, T.. Competing scaffolding proteins determine capsid size during mobilization of Staphylococcus aureus pathogenicity islands. eLife. **2017**. https://doi.org/10.7554/eLife.30822

Devoto A.E., Santini J.M., Olm M.R., Anantharaman K., Munk P., Tung J., Archie E.A., Turnbaugh P.J., Seed K.D., Blekhman R., Aarestrup F.M., Thomas B.C., & Banfield J.F.. Megaphages infect prevotella and variants are widespread in gut microbiomes. *Nature Microbiology*, 4(4), **2019**. https://doi.org/10.1038/s41564-018-0338-9

Duda R. L. & Teschke C. M. The Amazing HK97 Fold: Versatile Results of Modest Differences. *Current Opinion in Virology*, *36*, 9–16. **2019.** https://doi.org/10.1016/j.coviro.2019.02.001.

Duda, R. L., Hendrix, R.W., Huang, W. M., Conway, J. F., Shared architecture of bacteriophage SPO1 and herpesvirus capsids. *Curr. Biol.*, 16, R11–R13. **2006.** https://doi.org/10.1016/j.cub.2005.12.023

Earnshaw, W.C. and Casjens S.R.. DNA packaging by the double-stranded DNA bacteriophages. *Cell*, 21(2) 319-331. **1980**. https://doi.org/10.1016/0092-8674(80)90468-7

Edwards, K. F., Steward, G. F., & Schvarcz, C. R.. Making sense of virus size and the tradeoffs shaping viral fitness. *Ecology Letters.* **2020**. https://doi.org/10.1111/ele.13630

Edwards, R. A., Vega, A. A., Norman, H. M., Ohaeri, M., Levi, K., Dinsdale, E. A., … Barr, J. J.. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nature Microbiology* 4, 1727–1736. **2019**. https://doi.org/10.1038/s41564-019-0494-6

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

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 lambda. *Virology*, 402, 355–365. **2010.** https://doi.org/10.1016/j.virol.2010.03.041

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 PhiRSL1 infecting the phytopathogen Ralstonia solanacearum. *Structure*, 21, 298–305. **2013.** https://doi.org/10.1016/j.str.2012.12.017

Evilevitch, A., The mobility of packaged phage genome controls ejection dynamics. *eLife*. **2018**. https://doi.org/10.7554/eLife.37345

Fitzpatrick, A.H., Rupnik A., O'Shea H., Crispie, F., Keaveney, S., Cotter, P.. High Throughput Sequencing for the Detection and Characterization of RNA Viruses. *Frontiers in Microbiology,* 12, 190. **2021**. https://doi.org/10.3389/fmicb.2021.621719

Fokine A. & Rossman M.G.. Molecular architecture of tailed double-stranded DNA phages. *Bacteriophage,* 4(1), **2014**. https://doi.org/10.4161/bact.28281

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., Rossmann, M. G., Mesyanzhinov, V. V., A three-dimensional cryo-electron microscopy structure of the bacteriophage phiKZ head. *J. Mol. Biol.*, 352, 117–124. **2005**. https://doi.org/10.1016/j.jmb.2005.07.018

Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, Lupas AN, Alva V., Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr Protoc Bioinformatics.,*Dec;72(1):e108. **2020** https://doi.org/10.1002/cpbi.108

Gan L, Speir JA, Conway JF, Lander G, Cheng N, Firek BA, Hendrix RW, Duda RL, Liljas L, Johnson JE. Capsid conformational sampling in HK97 maturation visualized by X-ray crystallography and cryo-EM. *Structure*, 14:1655-65. **2006.** https://doi.org/10.1016/j.str.2006.09.006

Gertsman, I., Gan, L., Guttman, M., Lee, K., Speir, J. A., Duda, R. L., Hendrix, R. W., Komives, E. A., Johnson, J. E., An unexpected twist in viral capsid maturation. *Nature*,458, 646–650. **2009.** https://doi.org/10.1038/nature07686

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*, 5, 4278. **2014.** https://doi.org/10.1038/ncomms5278

Gregory, A.C., Zablocki, O., Ahmed A. Zayed, A.A., Howell, A., Bolduc, B., Sullivan, M.B.. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host & Microbe*, 28(5) P724-740.E8. **2020**. https://doi.org/10.1016/j.chom.2020.08.003

Gonzalez, B., Monroe, L., Li, K., Yan, R., Wright, E., Walter, T., Kihara, D., Weintraub, S.T., Thomas, J.A., Serwer, P., Jiang, W.. Phage G structure at 6.1 Å resolution, condensed DNA, and host identity revision to a *Lysinibacillus. Journal of Molecular Biology*, 432(14): 4139–4153. **2020**. https://doi.org/10.1016/j.jmb.2020.05.016

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*, 88, 11846–11860*.* **2014.** https://doi.org/10.1128/JVI.01364-14

Guo, F., Liu, Z., Fang, P., Zhang, Q., Wright, E. T., Wu, W., Zhang, C., Vago, F., Ren, Y., Jakana, J., Chiu, W., Serwer, P., Jiang, W., Capsid expansion mechanism of bacteriophage T7 revealed by multistate atomic models derived from cryo-EM reconstructions, *Proc. Natl. Acad. Sci.*, 111 (43), E4606-E4614. **2014.** https://doi.org/10.1073/pnas.1407020111

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. *Nature***,** 585, 357–362.**2020.** https://doi.org/10.1038/s41586-020-2649-2

Helgstrand, C., Wikoff, W. R., Duda, R. L., Hendrix, R.W., Johnson, J. E., Liljas, L., The Refined Structure of a Protein Catenane: The HK97 Bacteriophage Capsid at 3.44Å Resolution. *Journal of Molecular Biology*,, Volume 334, Issue 5, 885-899. **2003**. https://doi.org/10.1016/j.jmb.2003.09.035

Hildebrand A., Remmert M., Biegert A., Söding J., Fast and accurate automatic structure prediction with HHpred. *Proteins*, 77 Suppl 9:128-32. **2009**. https://doi.org/10.1002/prot.22499

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

Howard-Varona, C.; Hargreaves, K.R; Abedon, S.T; Sullivan, M.B.. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *The ISME Journal*, 11, 1511–1520. **2017**. doi:10.1038/ismej.2017.16

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*, *8* (5). **2017.** https://doi.org/10.1128/mBio.01579-17.

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

Ionel, A., Velázquez-Muriel, J. A., Luque, D., Cuervo, A., Castón, J. R., Valpuesta, J. M., Martín-Benito, J., Carrascosa, J. L., Molecular Rearrangements Involved in the Capsid Shell Maturation of Bacteriophage T7, *Journal of Biological Chemistry*, Vol 286, Issue 1, 234-242, **2011**. <https://doi.org/10.1074/jbc.M110.187211>.

Iyer L.M., Anantharaman V., Krishnan A., Burroughs A.M., Aravind L.. Jumbo Phages: A Comparative Genomic Overview of Core Functions and Adaptions for Biological Conflicts. *Viruses,* 13(1):63. **2021**. <https://doi.org/10.3390/v13010063>

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

Jiang, W., Baker, M.L., Jakana, J., Weigele, P.R., King, J., Chiu, W. Backbone structure of the infectious epsilon 15 virus capsid revealed by electron cryomicroscopy. *Nature*, 451,1130–1134. **2008.** https://doi.org/10.1038/nature06665

Jin, H., Jiang, Y., Yang, F., Zhang, J., Li, W., Zhou, K., Ju, J., Chen, Y., Zhou, C., Capsid Structure of a Freshwater Cyanophage Siphoviridae Mic1, *Structure*, 27, 1508–1516. **2019.** https://doi.org/10.1016/j.str.2019.07.003

Joiner, K.L., Baljon, A., Barr, J., Rohwer, F., & Luque, A. Impact of bacteria motility in the encounter rates with bacteriophage in mucus. *Scientific Reports,* 9, 16427. **2019**. <https://doi.org/10.1038/s41598-019-52794-2>

Keen E.C., Bliskovsky V.V., Malagon F., Baker J.D., Prince J.S., Klaus J.S. & Adhya S.L. Novel “superspreader” bacteriophages promote horizontal gene transfer by transformation. *mBio* 8,e02115-16. **2017.** https://doi.org/10.1128/mBio.02115-16

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.; George, E. E.; Green, K. T.; Gregoracci, G. B.; Haas, A. F.; Haggerty, J. M.; Hester, E. R.; Hisakawa, N.; Kelly, L. W.; Lim, Y. W.; Little, M.; Luque, A.; McDole-Somera, T.; McNair, K.; de Oliveira, L. S.; Quistad, S. D.; Robinett, N. L.; Sala, E.; Salamon, P.; Sanchez, S. E.; Sandin, S.; Silva, G. G. Z.; Smith, J.; Sullivan, C.; Thompson, C.; Vermeij, M. J. A.; Youle, M.; Young, C.; Zgliczynski, B.; Brainard, R.; Edwards, R. A.; Nulton, J.; Thompson, F.; Rohwer, F.. Lytic to temperate switching of viral communities. *Nature*, 531, 466–470. **2016**. https://doi.org/10.1038/nature17193

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. *Microbiology and Molecular Biology Reviews,* 84(2). **2020**. https://doi.org/10.1128/MMBR.00061-19

Kozlowski, L., IPC - Isoelectric Point Calculator. Biology Direct, 11:55. 2016. https://doi.org.10/1186/s13062-016-0159-9

Krupovic, M. and Koonin, E.V.. Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl. Acad. Sci. USA,* 114 (12) E2401-E2410. **2017**. <https://doi.org/10.1073/pnas.1621061114>

Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. *Virus research*. 239:136-42. **2017**. <https://doi.org/10.1016/j.virusres.2017.02.002>

Lander, G. C., Baudoux, A. C., Azam, F., Potter, C. S., Carragher, B., Johnson, J. E., Capsomer dynamics and stabilization in the T¼12 marine bacteriophage SIO-2 and its procapsid studied by CryoEM. *Structure***,** 20, 498–503. **2012.** https://doi.org/10.1016/j.str.2012.01.007

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*, 16, 1399–1406. **2008**. https://doi.org/10.1016/j.str.2008.05.016

Lara E., Vaqué D., Sà E.L., Boras J.A., Gomes A., Borrull E., Díez-Vives C., Teira E., Pernice M.C., Garcia F.C. and Forn I.. Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Science advances*, *3*(9), p.e1602565. **2017.** https://doi.org/10.1126/sciadv.1602565

Leiman, P. G., Battisti, A. J., Bowman, V. D., Stummeyer, K., Muhlenhoff, M., GerardySchahn, R., Scholl, D., Molineux, I. J., The structures of bacteriophages K1E and K1-5 explain processive degradation of polysaccharide capsules and evolution of new host specificities. *J. Mol. Biol.*, 371, 836–849. **2007**. <https://doi.org/10.1016/j.jmb.2007.05.083>

Letunic I., and Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. **2021**. <https://doi.org/10.1093/nar/gkab301>

Liang, G. and Bushman, F.D.. The human virome: assembly, composition and host interactions. *Nature Reviews Microbiology*, 19, 514–527. **2021**. https://doi.org/10.1038/s41579-021-00536-5

Liu, T., Sae-Ueng, U., Li, D., Lander, G.C., Zuo, X., Jönsson, B., Rau, D., Shefer, I., Evilevitch, A.. Solid-to-fluid–like DNA transition in viruses facilitates infection. *Proc. Natl. Acad. Sci.*, 111 (41) 14675-14680. **2014**. https://doi.org/10.1073/pnas.1321637111

Liu, X., Zhang, Q., Murata, K., Baker, M., Sullivan, M. B., Fu, C., Dougherty, M. T., Schmid, M. F., Osburne, M. S., Chisholm, S. W., Chiu, W., Structural changes in a marine podovirus associated with release of its genome into Prochlorococcus. *Nat Struct Mol Biol*, 17, 830–836. **2010**. https://doi.org/10.1038/nsmb.1823

Luque, A., & Reguera, D. The structure of elongated viral capsids. *Biophysical journal*, *98*(12), 2993-3003. **2010.** <https://doi.org/10.1016/j.bpj.2010.02.051>

Luque, A., & Reguera, D. Theoretical Studies on Assembly, Physical Stability and Dynamics of Viruses. *Structure and Physics of Viruses*, 553–595. **2013**. doi:10.1007/978-94-007-6552-8\_19

Luque A., Benler S., Lee D.Y., Brown C., White S. The missing tailed phages: prediction of small capsid candidates. *Microorganisms*, 8, 1944. **2020.** https://doi.org/10.3390/microorganisms8121944

Luque, A. and Silveira, C.B. Quantification of Lysogeny Caused by Phage Coinfections in Microbial Communities from Biophysical Principles. *mSystems* 5(5) e00353-20. **2020**. https://doi.org/10.1128/mSystems.00353-20

Mahmoudabadi G., Milo R. & Phillips R. Energetic cost of building a virus *Proceedings of the National Academy of Sciences*, 114 (22) E4324-E4333. **2017**. https://doi.org/10.1073/pnas.1701670114

Madden T. The BLAST sequence analysis tool. InThe NCBI Handbook [Internet]. 2nd edition **2013** Mar 15. National Center for Biotechnology Information (US).

Maurice, C.F.. Considering the other half of the gut microbiome: bacteriophages. *mSystems*, 4(3) e00102-19. **2019.** https://doi.org/10.1128/mSystems.00102-19

McNair, Katelyn (2021): PHANOTOME June 2017 Backup. figshare. Dataset. https://doi.org/10.6084/m9.figshare.13557770.v1

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

Montiel-Garcia, D., Santoyo-Rivera, N., Ho, P., Carrillo-Tripp, M., Brooks III, C.L., Johnson, J.E., Reddy, V.S. VIPERdb v3.0: a structure-based data analytics platform for viral capsids, *Nucleic Acids Research*, 49(D1) D809–D816. **2021**. <https://doi.org/10.1093/nar/gkaa1096>

Nifong R. L. & Gillooly J. F. Temperature Effects on Virion Volume and Genome Length in DsDNA Viruses. *Biol. Lett.*, 12 (3), 20160023. **2016**. https://doi.org/10.1098/rsbl.2016.0023

Parent, K. N., Khayat, R., Tu, L. H., Suhanovsky, M. M., Cortines, J. R., Teschke, C. M., Johnson, J. E., Baker, T. S., P22 coat protein structures reveal a novel mechanism for capsid maturation: stability without auxiliary proteins or chemical crosslinks. *Structure*,18, 390–401. **2010.** https://doi.org/10.1016/j.str.2009.12.014

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,* 427,177–188. **2012.** https://doi.org/10.1016/j.virol.2012.01.040

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 tail spike receptor-binding domain. *Virology,* 464–465C, 55–66. **2014.** https://doi.org/10.1016/j.virol.2014.06.017

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12: 2825—2830. **2011**. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Petrovsky, S., Dyson, Z.A., Seviour, R.J. & Tillett, D. Small but Sufficient: the *Rhodococcus* Phage RRH1 Has the Smallest Known Siphoviridae Genome at 14.2 Kilobases. *Journal of Virology*, 86 (1) 358-363. **2011**. https://doi.org/10.1128/JVI.05460-11

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.,* Jan;30(1):70-82. **2021**. https://doi.org/10.1002/pro.3943

PhAnToMe: Phage Annotation Tools and Methods[http://www.phantome.org/] (accessed June 1th 2017).

Pietilä M.K., Laurinmaki P., Russell D.A., Ko C.C., Jacobs-Sera D., Hendrix R.W., Bamford .H., Butcher S.J. Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. *Proc. Nat. Acad. Sci. USA***,** 110,10604–10609. **2013.** https://doi.org/10.1073/pnas.1303047110

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*, 12:294. **2020.** https://doi.org/10.3390/v12030294

Pride, D.T., Wassenaar, T.M., Ghose, C., Blaser, M., Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics*, 7, 8. **2006.** https://doi.org/10.1186/1471-2164-7-8

Rohwer F., & Edwards R. The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of bacteriology*, 184(16), 4529–4535. **2002.** https://doi.org/10.1128/jb.184.16.4529-4535.2002

Roos, W. H., Bruinsma, R., & Wuite, G. J. L.. Physical virology. *Nature Physics*, 6(10), 733–743. **2010**. https://doi.org/10.1038/nphys1797

Roux, S., Páez-Espino, D., Chen, I.A., Palaniappan, K., Ratner, A., Chu, K., Reddy, T.B.K., Nayfach, S., Schulz, F., Call, L., Neches, R.Y., Woyke, T., Ivanova, N.N., Eloe-Fadrosh, E.A., Kyrpides, N.C.. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Research.* 49(D1) D764-D775. **2020**. https://doi.org/10.1093/nar/gkaa946

Roux, S., Solonenko, N.E., Dang, V.T., Poulos, B.T., Schwenck, S.M., Goldsmith, D.B., Coleman, M.L., Breitbart, M., Sullivan, M.B.. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ*, 4:e2777. **2016**. https://doi.org/10.7717/peerj.2777

Santos-Medellin, C., Zinke, L.A., ter Horst, A.M., Gelardi, D.L., Parikh, S.J., Emerson, J.B.. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *The ISME Journal*, 15, 1956–1970. **2021**. https://doi.org/10.1038/s41396-021-00897-y

Seguritan V., Alves Jr. N., Arnoult M., Raymond A., Lorimer D. , Burgin Jr. A.B., Salamon P., & Segall A.M.. Artificial neural networks trained to detect viral and phage structural proteins. *PLOS Computational Biology*, 8 (8), e1002657. **2012.** https://doi.org/10.1371/journal.pcbi.1002657

Shamash, M. and Maurice, C.F. 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

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

Shkoporov, A.N., Clooney, A.G., Sutton, T.D.S., Ryan, F.J., Daly, K.M., Nolan, J.A., McDonnell, S.A., Khokhlova, E.V., Draper, L.A., Forde, A., Guerin, E., Velayudhan, V., Ross, R.P., Hill, C.. The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host & Microbe*, 26(4) P527-541.E5. **2019**. https://doi.org/10.1016/j.chom.2019.09.009

Sieradzki, E., Ignacio-Espinoza, J.C., Needham, D., Fichot, E.B., Fuhrman, J.A.. Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nat Commun* 10, 1169. **2019**. <https://doi.org/10.1038/s41467-019-09106-z>

Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D. and Higgins D.G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539. **2011.** <https://doi.org/10.1038/msb.2011.75>

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., Rohwer F.L. & Thompson F., Genomic and ecological attributes of marine bacteriophages encoding bacterial virulence genes. *BMC Genomics* 21, 126. **2020**. https://doi.org/10.1186/s12864-020-6523-2

Silveira C.B., Luque A., Rohwer F. The landscape of lysogeny across microbial community density, diversity, and energetics. *Environmental Microbiology,* 23: 4098-4111. **2021**. https://doi.org/10.1111/1462-2920.15640

Söding J., Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951-60. **2005**. https://doi.org/10.1093/bioinformatics/bti125

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 80alpha capsids. *J. Mol. Biol.*, 405, 863–876. **2011**. https://doi.org/10.1016/j.jmb.2010.11.047

Stone, N.P., Demo, G., Agnello, E. et al. Principles for enhancing virus capsid capacity and stability from a thermophilic virus capsid structure. *Nat Commun***,** 10, 4471. **2019.** https://doi.org/10.1038/s41467-019-12341-z

Stroupe, M. E., Brewer, T. E., Sousa, D. R., Jones, K. M., 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*, 450–451, 205–212. **2014.** https://doi.org/10.1016/j.virol.2013.11.019

Suhanovsky M. M. and Teschke C. M. Nature’s Favorite Building Block: Deciphering Folding and Capsid Assembly of Proteins with the HK97-Fold. *Virology*, *479–480*, 487–497. **2015.** https://doi.org/10.1016/j.virol.2015.02.055.

Sulcius S., Staniulis J. & Paskauskas R. Morphology and distribution of phage-like particles in a eutrophic boreal lagoon. *Oceanologia*, 53(2), 587-603.**2011.** https://doi.org/10.5697/oc.53-2.587

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., Chisholm, S.W.. Prevalence and Evolution of Core Photosystem II Genes in Marine Cyanobacterial Viruses and Their Hosts. *PLOS Biology* 4(8): e234. **2006**. https://doi.org/10.1371/journal.pbio.0040234

Touchon M., Moura de Sousa J.A. & Rocha E.P. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Curr Opin Microbiol*., 38:66-73. **2017.** https://doi.org/10.1016/j.mib.2017.04.010

Twarock R. & Luque A. Structural Puzzles in Virology Solved with an Overarching Icosahedral Design Principle. *Nature Communications*, *10* (1), 4414. **2019**. https://doi.org/10.1038/s41467-019-12367-3.

White, H. E., Sherman, M.B., Brasiles, 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.*, 86, 6768–6777. **2012.** https://doi.org/10.1128/JVI.00412-12

Wikoff, W. R., Liljas, L. Duda, R.L., Tsuruta, H., Hendrix, R.W., Johnson, J.E., Topologically Linked Protein Rings in the Bacteriophage HK97 Capsid. *Science*, 289(5487), 2129–2133. **2000**. doi:10.1126/science.289.5487.2129

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, CJ, Polat, I., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3), 261-272. **2020.** https://doi.org/10.1038/s41592-019-0686-2

Yuan Y. & Gao M. Jumbo bacteriophages: An overview. *Frontiers in Microbiology*, 8(403), **2017**. https://doi.org/10.3389/fmicb.2017.00403

Zhang, X., Guo, H., Jin, L., Czornyj, E., Hodes, A., Hui, W. H., Nieh, A. W., Miller, J. F., Zhou, Z. H., A new topology of the HK97-like fold revealed in Bordetella bacteriophage by cryoEM at 3.5 Å resolution. *eLife*, 2, e01299. **2013.** https://doi.org/10.7554/eLife.01299.001

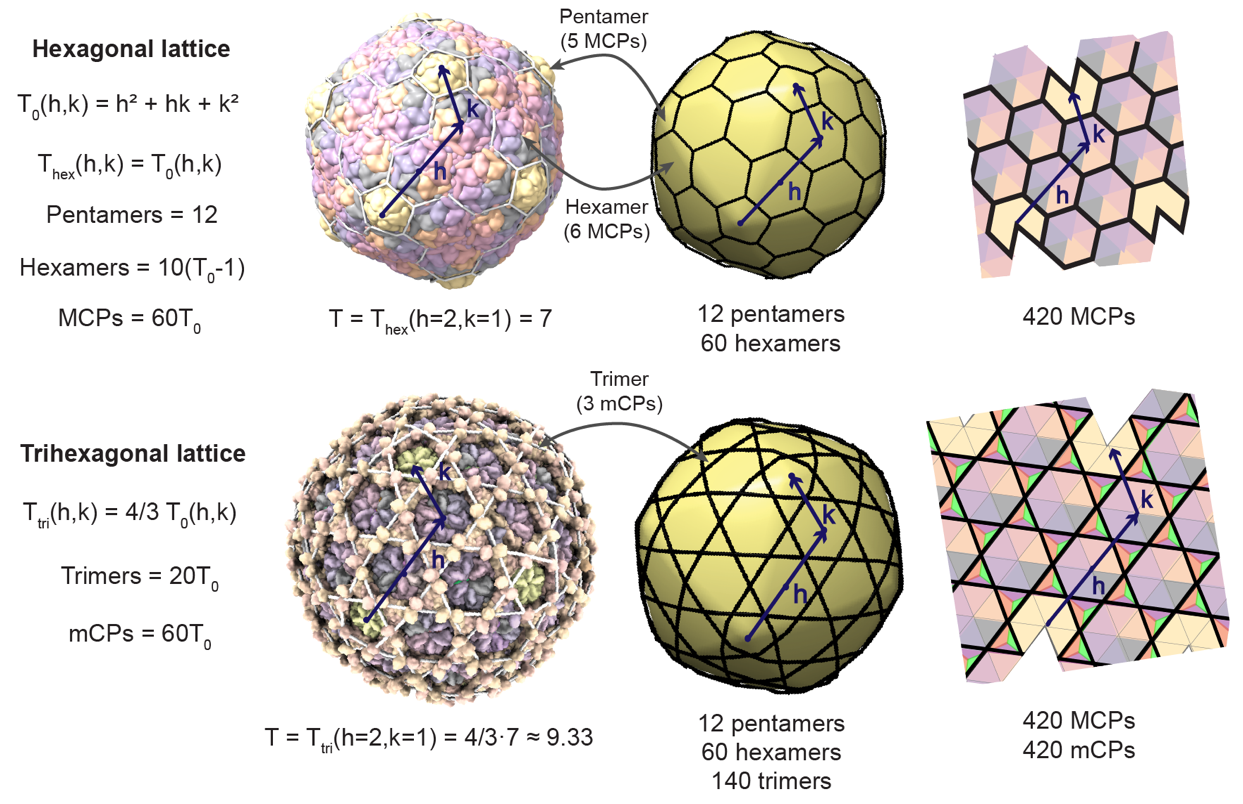
Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. J. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core., *Mol Biol.,* S0022-2836(17)30587-9. **2018**. https://doi.org/10.1016/j.jmb.2017.12.007

Zinder N. D., and Lederberg J. Genetic exchange in Salmonella. *Journal of bacteriology*, 64(5), 679–699. **1952.** https://doi.org/10.1128/JB.64.5.679-699.1952

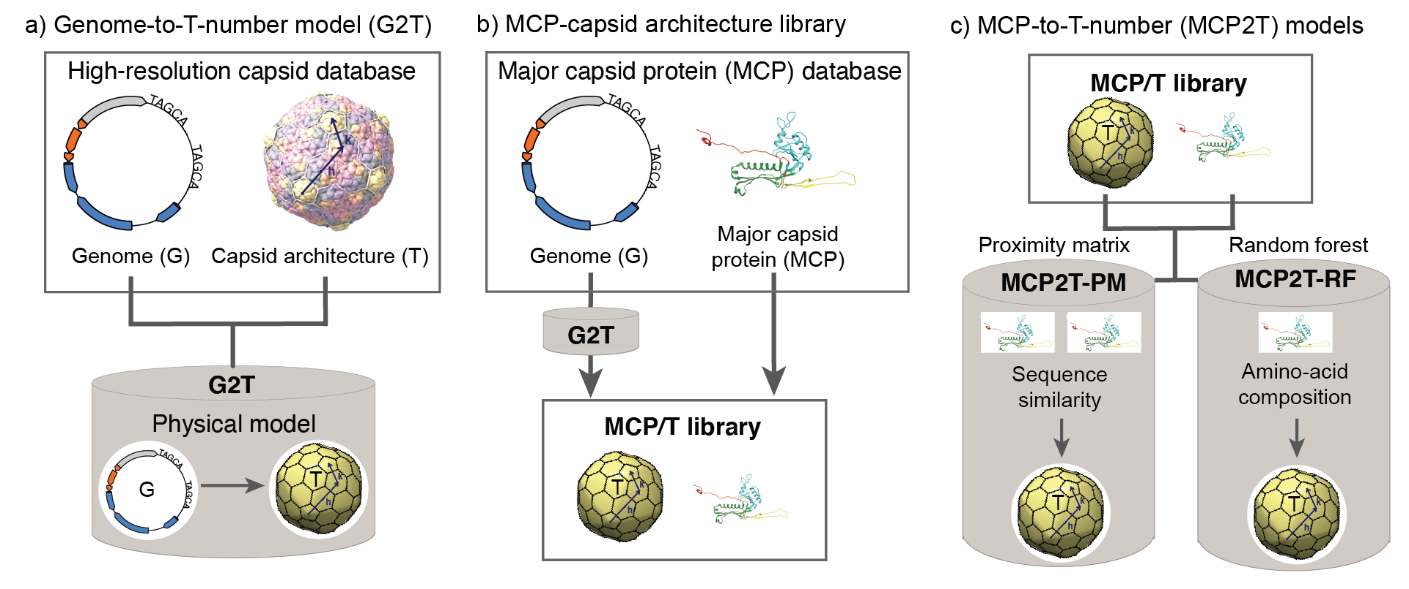
**TABLES AND FIGURES**

|  |  |  |  |
| --- | --- | --- | --- |
| **Phage** | **T** | **Genome (kbp)** | **Reference** |
| C1 | 4 | 16.7 | Aksyuk et al. 2012 |
| HSTV-1 | 7 | 32.2 | Pietila et al. 2013 |
| P2 | 7 | 33.6 | Dearborn et al. 2012 |
| TP901-1 | 7 | 37.7 | Bebeacua et al. 2013 |
| Sf6 | 7 | 39.0 | Parent et al. 2012 |
| ε15 | 7 | 39.7 | Baker et al. 2013; Jiang et al. 2008 |
| HK97 | 7 | 39.7 | Gertsman et al. 2009; Helgstrand et al. 2003; Wikoff et al. 2000 |
| T7 | 7 | 39.9 | Agirrezabala et al. 2007, Guo et al. 2014, Ionel et al. 2011 |
| CUS-3 | 7 | 40.2 | Parent et al. 2014 |
| HK022 | 7 | 40.8 | Pride et al. 2006 |
| Pf-WMP4 | 7 | 40.9 | Liu et al. 2007 |
| BPP-1 | 7 | 42.9 | Zhang et al. 2013 |
| P22 | 7 | 43.5 | Chen et al. 2011; Parent et al. 2010a |
| 80α | 7 | 43.9 | Spilman et al. 2011 |
| K1E/K1-5 | 7 | 44.7 | Leiman et al. 2007 |
| P-SSP7 | 7 | 45.0 | Liu et al. 2010 |
| Gifsy-2 | 7 | 45.8 | Effantin et al. 2010 |
| Syn5 | 7 | 46.2 | Gipson et al. 2014, and Pope et al. 2007 |
| Λ | 7 | 48.5 | Lander et al. 2008 |
| CW02 | 7 | 49.4 | Shen et al. 2012 |
| SPP1 | 7 | 49.5 | White et al. 2012 |
| SIO-2 | 12 | 80.0 | Lander et al. 2012 |
| P74-26 | 9.33 | 83.0 | Stone et al. 2019 |
| P23-45 | 9.33 | 84.2 | Bayfield et al. 2019 |
| Basilisk | 12 | 81.8 | Grose et al. 2014; Twarock and Luque 2019 |
| Mic1 | 13 | 92.6 | Jin et al. 2019 |
| T5 | 13 | 121.8 | Effantin et al. 2006 |
| SPO1 | 16 | 132.6 | Duda et al. 2006 |
| ΦM12 | 19 | 194.7 | Stroupe et al. 2014 |
| N3 | 19 | 207.0 | Suhanovsky and Teschke 2015; Hua et al. 2017 |
| PAU | 25 | 219.0 | Suhanovsky and Teschke 2015; Hua et al. 2017 |
| ΦRSL1 | 27 | 240.0 | Effantin et al. 2013 |
| PBS1 | 27 | 252.0 | Suhanovsky and Teschke 2015; Hua et al. 2017 |
| ΦKZ | 27 | 280.0 | Fokine et al. 2005 |
| 121Q | 28 | 348.5 | Suhanovsky and Teschke 2015 |
| SCTP2 | 39 | 440.0 | Hua et al. 2017 |
| G | 52 | 498.0 | Suhanovsky and Teschke 2015; Hua et al. 2017 |

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

****

**Figure 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. T0 is the T-number defined by the classic icosahedral capsid theory (Caspar and Klug, 1962). Thex and Ttri are the T-numbers associated, respectively, with the hexagonal and trihexagonal lattices defined by the generalized icosahedral capsid theory (Twarock and Luque, 2019). The top and bottom capsid examples correspond, respectively, to phage HK97 (PDB 2fs3; Gan et al. 2006) and phage patience (EMDB-21123; Podgorski et al. 2020). The capsids were rendered with ChimeraX (Pettersen et al. 2021). The 3D icosahedral lattice models were produced with the generalized *hkcage* tool in ChimeraX (Luque et al. 2020).

****

**Figure 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.

Chart, scatter chart

Description automatically generated

**Figure 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 (R2) are displayed in each legend.

**Chart

Description automatically generated**

**Figure 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 (Pettersen et al. 2021, Luque et al. 2020).

Chart, sunburst chart

Description automatically generated

**Figure 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.

Graphical user interface, diagram

Description automatically generated

**Figure 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, R2).

**Diagram

Description automatically generated**

**Figure 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.