ADVANCED BIOSYSTEMS
www.adv-biosys.com

Mini-Metagenomics and Nucleotide Composition Aid the Identification and Host Association of Novel Bacteriophage Sequences

Jonathan Deaton, Feiqiao Brian Yu, and Stephen R. Quake*

A broad spectrum of metagenomic and single cell sequencing techniques have become popular for dissecting environmental microbial diversity, leading to the characterization of thousands of novel microbial lineages. In addition to recovering bacterial and archaeal genomes, metagenomic assembly can also produce genomes of viruses that infect microbial cells. Because of their diversity, lack of marker genes, and small genome size, identifying novel bacteriophage sequences from metagenomic data is often challenging, especially when the objective is to establish phage–host relationships. The present work describes a computational approach that uses supervised learning to classify metagenomic contigs as phage or non-phage as well as assigning phage taxonomy based on tetranucleotide frequencies. Furthermore, the method assigns phage–host relationships using co-occurrence statistics derived from a recently developed mini-metagenomic experimental technique. This work evaluates method performance at identifying viral contigs and predicting taxonomic classification using publicly available references. Then, using two mini-metagenomic datasets, over 100 novel phage contigs from hot spring samples of Yellowstone National Park are identified and assigned to putative microbial hosts. Results of this work demonstrate the value of combining viral sequence identification with mini-metagenomic experimental methods to understand the microbial ecosystem.

1. Introduction

Bacteriophages (phages) play important roles in microbial communities, often driving lateral gene transfer,[1] regulating the recycling of microbial biomass,[2] and are the most abundant biological entities on the planet at an estimated 10^{31} viral particles.[3] Despite this abundance, our understanding of phage diversity is limited to 10^5 metagenomic assembled viral sequences and a few thousand isolate sequences. Before the advent of high-throughput sequencing, our understanding of phage genomics was limited to lineages that could be cultured in plaques.[4] Although metagenomics alleviates this limitation,[5] the typically small size of phage genomes, the lack of a universal marker gene, and the high genetic diversity complicate the identification of phage genomic sequences.[6]

Common approaches of phage identification use Hidden Markov Models based gene annotation to search for coding regions that are homologous to known viral genes, including "hallmark genes" such as the terminase large and small subunits, major capsid, coat, tail, and portal proteins.[7–9] Approaches based on hallmark genes perform well for identifying genomes closely related to known phage sequences in terms of genetic content but may fail at identifying phage sequences with few known genetic homologues. A less common approach employs machine-learning techniques on frequencies of short oligonucleotides of length k (k-mers).[10] Frequencies of k-mers vary among species, and can be used to predict phylogeny and taxonomy. For instance, k-mer frequencies of a phage's genome are predictive of the phage's host.[5,11,12] A k-mer frequency based approach is advantageous when no high-similarity alignments exist. Such an approach is capable of handling metagenome assembled contigs significantly shorter than the complete viral genome and reduces computational expense compared to alignment-based methods. Other than identifying phage genomes, machine-learning techniques have been used for molecular applications such as identifying phage virion proteins[13,14] and phage proteins[15] that may be located inside host cells, possibly during the lysogenic phase. Finally, identification of potential host for phage contigs assembled from metagenomic data typically requires comparison to reference phage genomes with known host,[16] matching to CRISPR repeats found in bacterial genomes,[17] or comparing sequence content such as tetranucleotide frequencies.[18,19]

Here we analyze mini-metagenomic assembled contigs from three hot spring locations in Yellowstone National Park (YNP) for phage representation. We create a custom algorithm, called

J. Deaton, Dr. F. B. Yu, Prof. S. R. Quake
Department of Bioengineering
Stanford University
443 Via Ortega, Stanford, CA 94305, USA
E-mail: brian.yu@czbiohub.org
Dr. F. B. Yu, Prof. S. R. Quake
Chan Zuckerberg Biohub
499 Illinois St, San Francisco, CA 94158, USA
The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adbi.201900108.

DOI: 10.1002/adbi.201900108

 © 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

PhaMers (Phage *k*-Mers), that uses tetranucleotide frequency and a combination of two supervised learning algorithms to identify as well as to classify viral sequences from metagenomic sequencing data. We first characterize performance of PhaMers using references viral and microbial sequences from NCBI. We then apply PhaMers to hotspring metagenomic contigs and compare the results to those obtained from VirSorter, an automated phage identification pipeline based on alignment to hallmark genes,[7] and DeepVirFinder, another *k*-mer based viral identification tools based on convolutional neural network (CNN), a different machine-learning framework. In total, we identify 1165 putative phage sequences longer than 5 kbp and 313 putative sequences longer than 10 kbp supported by at least one viral identification tool. Less than half of the phage sequences PhaMers identifies is also identified as phage by DeepVirFinder.[20] We use Joint Genome Institute's Integrated Microbial Genomes & Microbiomes (IMG/ER) pipeline[21] to generate functional annotations for all phage sequences and demonstrate that more putative phage sequences identified by PhaMers contain viral genes compared to DeepVirFinder. Finally, using mini-metagenomic co-occurrence patterns, we identify and propose putative hosts for more than 100 novel phage sequences.

2. Results and Discussions

2.1. Tetranucleotide Frequency Differentiates Phage Taxonomy

Tetranucleotide frequency has been used to group prokaryotic genomes, to assign taxonomy,[19] and to distinguish finer scale compositional biases such as those around the replication origins of bacteria.[22] Due to viral host specificity, we argue that tetranucleotide frequencies can also serve to distinguish phage genomes.[12,23] To assess relationships in tetranucleotide frequencies among known phage sequences, we collected 2255 phage genomes from RefSeq in October 2015 belonging to various taxonomies and having different genome lengths clustered around 40–50 kbp and 170–180 kbp (Figure S1, Table S1, Supporting Information), calculated tetranucleotide frequencies, and visualized phage genome relations in two dimensions using t-Distributed Stochastic Neighbor Embedding (t-SNE) (**Figure 1**a).[24] We then clustered phage genomes using DBSCAN,[25] and observed enriched taxa within clusters (Figure 1b). 60% of the viral genomes are assigned to clusters containing 10 or more sequences (Table S2, Supporting Information). All clusters containing more than 10 sequences contain almost exclusively viral sequences from a single category under Baltimore classification, demonstrating the effectiveness of DBSCAN clustering of viral sequences to enrich for sequences of similar taxonomy (Figure 1b).

In particular, single stranded RNA (ssRNA) viruses cluster into two enriched groups belonging to the genera *Allolevirus* and *Levivirus* respectively, generally classified as *Enterobacteriophge MS2* and *Enterobacteriophage Qß*,[26] Six enriched clusters of single stranded DNA (ssDNA) viruses are also identified. The largest of these clusters is enriched for the well-characterized 5.4 kbp *Enterobacteria phage phiX174 sensu lato*. The second

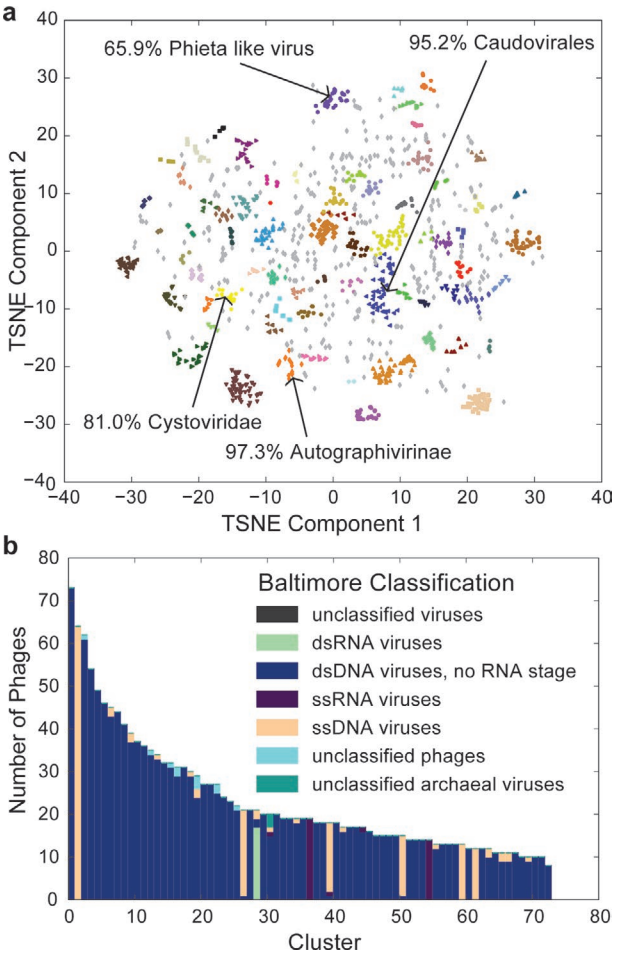

Figure 1. RefSeq phage tetranucleotide characteristics and taxonomy. a) t-SNE representation based on tetranucleotide frequency of 2255 phage sequences from RefSeq. Clusters assigned with reduced dimensionality (2D) embedded tetramer frequency vectors and clustered using DBSCAN (epsilon = 1.5, min points per cluster = 10). Some clusters enriched with phage of a single taxon are labeled with percentages denoting the proportion of phage sequences in that cluster belonging to that enriched taxon. A cluster is considered enriched with a taxon if the proportion of phage sequences belonging to that taxon in the cluster is greater than 50% and the enriched abundance is statistically significant compared to background abundance in the reference dataset, as tested using Pearson's chi-squared test. b) Compositions of taxa for phage sequences assigned to the clusters shown in Figure 1a at the Baltimore Classification depth.

largest of these clusters is enriched for the 5.5 kbp *Enterobacteria phage G4 sensu lato*. There are limited number of double stranded RNA (dsRNA) virus sequences in the database we used. These viruses are deposited into the same cluster based on our algorithm, demonstrating that dsRNA viruses are distinct from other types of viruses. Finally, most viral clusters are enriched for double stranded DNA (dsDNA) viruses, to which phage sequences belong. Furthermore, clusters enriched for dsDNA viruses represent enriched viral groups at lower taxonomies. Examples include *Caudovirales* (81.0% of cluster), *Autographivirinae* (97.3% of cluster), *Cystoviridae* (81.0% of cluster), and *Phieta like virus* (65.9% of cluster) (Figure 1a). Taken together, we show that tetranucleotide frequency based

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

phage taxonomy identification using t-SNE and DBSCAN is an effective phage differentiation scheme, generating groups of viral genomes enriched for a single taxon.

For comparison purposes, we also used *k*-means clustering on tetranucleotide frequencies to perform similar taxonomic classification on reference viral sequences (Figure S2, Supporting Information). Unlike DBSCAN, *k*-means clustering associates every sequence into a cluster based on the number of cluster centers provided. Using $k = 40$ results in less clusters but at the cost of some viral clusters becoming less pure. Similar to DBSCAN results, most clusters are enriched for dsDNA viruses. However, ssRNA and dsRNA virus clusters now contain significant proportion of dsDNA viruses as well (Figure S2b, Supporting Information). Although enriched clusters at the family level exist, such as *Autographivirinae*, *Caudovirales*, and *Podoviridae*, most clusters are no longer enriched for a single group, demonstrated by the example where dsDNA *Cellulophaga phage phiSM*, dsDNA *Lactococcus phage 936 sensu lato*, and *Skunalikevirus* are classified into the same cluster (Figure S2, cluster 10 in Table S3, Supporting Information). However, when *k* is increased to 86, results obtained become similar to those obtained using DBSCAN.

2.2. PhaMers Uses Machine-Learning Techniques to Identify Phage Sequences from Reference Genomes

In addition to differentiating phage sequences from each other, tetranucleotide frequencies can differentiate phage from prokaryotic sequences. This feature is useful since the abundance of prokaryotic contigs from metagenomic datasets often hampers discovery of novel viral genomes. Hence, we sought to understand how tetranucleotide frequencies could best distinguish phage sequences from prokaryotic sequences and developed a machine-learning algorithm that compares tetranucleotide frequencies of unknown sequences to those of known phage, bacterial and archaeal sequences (Figure S3, Supporting Information). Reference phage genomes from RefSeq (Table S1, Supporting Information) and bacterial and archaeal genomes from GenBank (Table S4, Supporting Information) were used. Testing a set of machine-learning algorithms and their combinations using 20-fold cross validation on reference phage genomes, we found that a linear combination of KNN and a cluster centroid proximity metric (SI Text, Figure S4, Supporting Information) performed best when considering both area under the curve (AUC = 0.992) and true positive rate (TP = 91%) (**Figure 2**a). We

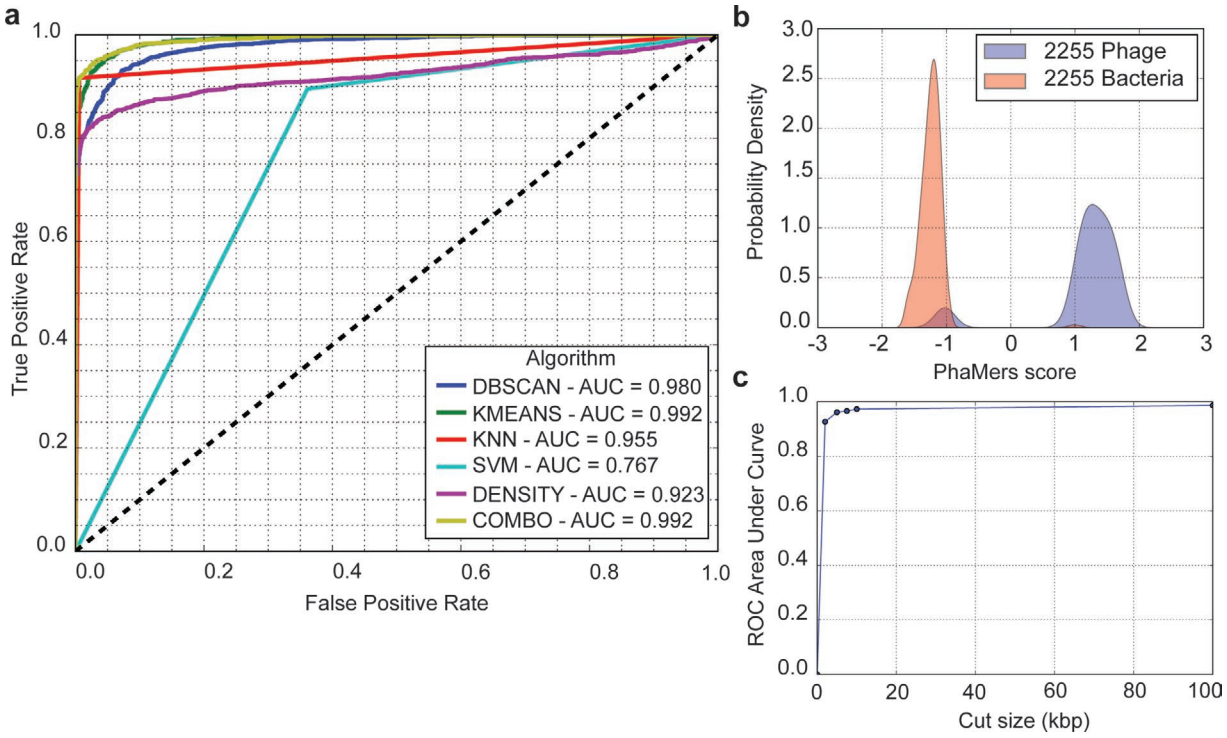

Figure 2. *N*-fold cross validation of PhaMers scoring algorithm. a) Receiver operator characteristic (ROC) curves for various supervised learning algorithms used on the reference datasets of phage and bacterial sequences in 20-fold cross validation. "DBSCAN" and "KMEANS" scoring algorithms apply the method described in S1 Text to calculate a metric of proximity to the nearest clusters of phages and bacteria assigned by DBSCAN and *k*-means clustering with $k = 86$, respectively. "DENSITY" stands for a custom algorithm that uses Kernel Density Estimation to approximate the probability density of phage and bacteria data points, and then gives a score as the log-ratio of the two probabilities. "KNN" represents a K-Nearest-Neighbors algorithm, and "SVM" represents a Support Vector Machine approach. The "COMBO" algorithm is a linear combination of K-Nearest Neighbors and "KMEANS" and is described in S1 Text. b) Distributions of PhaMers scores, given by "COMBO", for 2255 phage and 2255 bacterial genomes, as calculated with 20-fold cross validation. The small blue population centered at score = −1 are the phages in the datasets which were misclassified as bacteria (false negatives). There is also a smaller population of misclassified bacteria shown at score = 1. c) Predictive performance (AUC) of PhaMers as a function of reference genome length. The reference datasets of phage and bacterial genomes were cut randomly to sizes of 2.5, 5, 7.5, 10, and 100 kbp. Predictive performance, shown on the *y*-axis, is given by the area beneath the ROC, which drops as sequences were cut to lengths shorter than 5 kbp. Dataset of bacteria used for this analysis are included in Table S6, Supporting Information.

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

decided to use this combined approach for phage identification in PhaMers. The underlying method used by PhaMers to distinguish phage sequences from other microbial sequences differs from other *k*-mer based viral identification tools. For example, VirFinder uses a logistic regression model to build a binary classifier. The more recent DeepVirFinder developed by the same group uses a convolutional neural network (CNN). Because different machine-learning frameworks can lead to distinct outcomes, PhaMers offer a novel machine-learning model for phage contig identification. In practice, PhaMers scores predict phage sequences with positive values and non-phage sequences with negative values (Figure 2b). A reasonable cutoff to distinguish phage and non-phage sequences is 0, yielding 91.8% sensitivity, 99.3% specificity, and 99.2% positive predictive value (PPV).

When classifying reference sequences in the test dataset using a "leave one out" approach, PhaMers correctly classified phage sequences belonging to *Enterobacteria phage T4*, T4-like and T7-like viruses, *Propionibacterium phage*, and *Lactococcus phage ASCC*. Misclassified phage sequences (negative scores) were generally from *unclassified Siphoviridae*, *unclassified Podoviridae*, and *unclassified Myoviridae*. Misclassification occurred because there were few similar relatives in the dataset. To discount the effect of heavily represented phages on performance estimation, we performed cross validation with a reduced dataset of phages containing only one member from each taxonomy, generating 342 phage sequences. Tetranucleotide frequencies distinguished this set of phage sequences from bacteria with only marginally decreased accuracy (Figure S5, Supporting Information). We also assessed how sequence length affected PhaMers' performance by performing 20-fold cross validation with random subset from the test dataset. We observed that the area under the curve (AUC) of the receiver operator characteristic (ROC) dropped most significantly as sequence length decreased below 5 kbp (Figure 2c), demonstrating that 5 kbp is a useful contig length cutoff. The 5 kbp contig length also limits us to using tetranucleotide frequencies because longer *k*-mers tend to generate sparse vectors that reduce classification specificity.

2.3. PhaMers Identifies Putative Phage Sequences from Yellowstone Hot Springs

We analyzed two sets of assembled metagenomic contigs longer than 5 kbp with PhaMers, VirSorter, DeepVirFinder, and visualized the resulting viral contigs identified by each tool (**Figure 3**a, Table S5, Supporting Information). These datasets were prepared using a microfluidic-based mini-metagenomic method on two hot spring samples from Bijah and Mound Springs of the Yellowstone National Park (Experimental Section).[27] Briefly, microfluidic-based mini-metagenomics is a method developed to analyze complex microbial communities by dividing each sample into smaller subsamples containing 5–10 cells each. Whole genome amplification, metagenomic sequencing, and de-novo assembly of combined sequencing data produce a set of metagenomic contigs, most of which belong to microbial genomes.[27,28] In this work, we explore viral contigs and their relationships with microbial genomes from the same samples. Most contigs in both samples represent microbial genome fragments and are not identified as phage-like by any tool. PhaMers identified 836 of the 5594 contigs longer than 5 kbp as potential phage with positive scores. Similarly, out of all contigs over 5 kbp, VirSorter identified 23

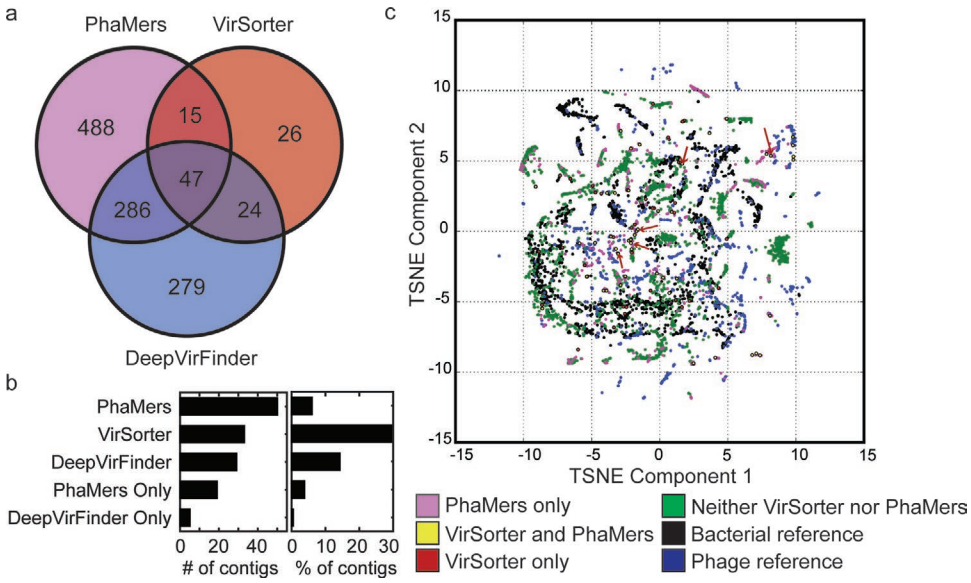

Figure 3. Phage contig identification and taxonomy prediction. a) Venn diagram showing the number of contigs from Mound Spring and Bijah Spring classified as phage by PhaMers, VirSorter, and DeepVirFinder. b) Bar graph illustrating either the number (left) or percentage (right) of contigs extracted by the respective phage identification method that contain at least one viral gene. The difference between PhaMers and PhaMers Only (or between DeepVirFinder and DeepVirFinder Only) is that contigs identified by PhaMers may also be identified by other tools, whereas contigs identified by PhaMers Only are not identified by any other tool. c) t-Distributed Stochastic Neighbor Embedding (t-SNE) representation of tetramer frequency vectors of reference genomes in conjunction with contigs longer than 5 kbp from the Mound Spring dataset. Contigs that are identified as phage by neither PhaMers nor VirSorter represent microbial contigs. Red arrows indicate contigs identified as phages which were assigned a taxonomic classification based on their proximity to clusters of phages enriched for a single taxon.

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

putative phage contigs from the Bijah Spring sample and 89 from the Mound Spring sample. Six phage contigs from the Mound Spring dataset were identified as potentially prophage. DeepVirFinder identified 636 contigs as potentially phage, out of which 71 were also identified by VirSorter, and 47 identified by all 3 tools (Figure 3a). Of the 112 contigs identified as potentially phage-like by VirSorter, 62 were given positive PhaMers scores. Even though PhaMers and DeepVirFinder each identified a similar number of contigs as phage-like, less than half (333) were identified by both tools. This result was surprising considering both tools use machine-learning methods and tetranucleotide frequency of DNA sequences to predict putative phages. We postulate that the difference results from the distinct algorithms used by machine-learning models underlying both phage prediction tools (Experimental Section).

Following phage contig identification we assess whether putative phage contigs identified by PhaMers are phage-like. Since many putative phage contigs identified by PhaMers and DeepVirFinder contain only unannotated open reading frames, we tabulated the number of contigs containing at least one annotated gene with predicted viral function (Figure 3b). Putative phages identified by PhaMers contained the highest number of contigs with viral genes (Figure 3b left panel), which may not be surprising since PhaMers also identified the highest number of sequences as phage-like. After normalizing by the total number of contigs identified by the respective methods, VirSorter results had the highest percentage of putative phage contigs with viral genes (Figure 3b right panel). However, out of putative phage contigs identified by PhaMers or DeepVirFinder only, PhaMers's predictions included 5x the number of contigs with viral genes, demonstrating that these were likely phage contigs that would not be identified via other tools. Taken together, PhaMers identifies phage contigs from metagenomic datasets that VirSorter and DeepVirFinder miss.

To deepen our analysis of novel phage sequences, we used t-SNE to visualize tetranucleotide frequencies of all contigs combined with phage and bacterial reference sequences (Figure 3c, Figure S6, Supporting Information). Microbial contigs from Yellowstone metagenomic samples formed tight clusters, indicating that they originate from closely related bacterial genomes. We focused on a set of putative phage contigs predicted by VirSorter and PhaMers that lie in proximity to clusters of known phage genomes (19 contigs from Bijah Spring and 83 contigs from Mound Spring).

To predict phage taxonomy from tetranucleotide frequencies, we clustered known phage tetramer frequency vectors with those of novel phages. We looked for novel phage contigs assigned to clusters enriched with reference phage sequences of a single taxon. Clusters were labeled as enriched for a taxon if members of that taxon constituted more than half of the cluster and their prevalence was significantly greater than the proportion that the taxon was represented in the reference dataset. A contig meeting these criteria and lying within a standard deviation of the mean cluster silhouette value was assigned the taxon of that cluster. Of the 24 contigs (five from Bijah Spring, 18 from Mound Spring, and one from Mammoth Geyser Basin) that met these criteria, most were assigned to clusters enriched for *Siphoviridae, Podoviridae,* and *Myoviridae* (Figure 3c red arrows). These three families of dsDNA phages belong to the order *Caudovirales* and are differentiated by their tail morphology. *Podoviridae* have short non-contractile tails while *Myoviridae* and *Siphoviridae* have long tails, contractile for the former and non-contractile for the ladder.[29] A 13504 base pair phage contig (Contig 1753) from Mound Spring (category 2 "quite sure" prediction by VirSorter and 1.14 by PhaMers) is similar in tetranucleotide frequency to *Siphoviridae* and contains many genes with viral functions, including phage tail and portal proteins typically associated with *Siphoviridae* (**Figure 4**a–c). Another phage contig (Contig 677) 20664 base pairs in length (category

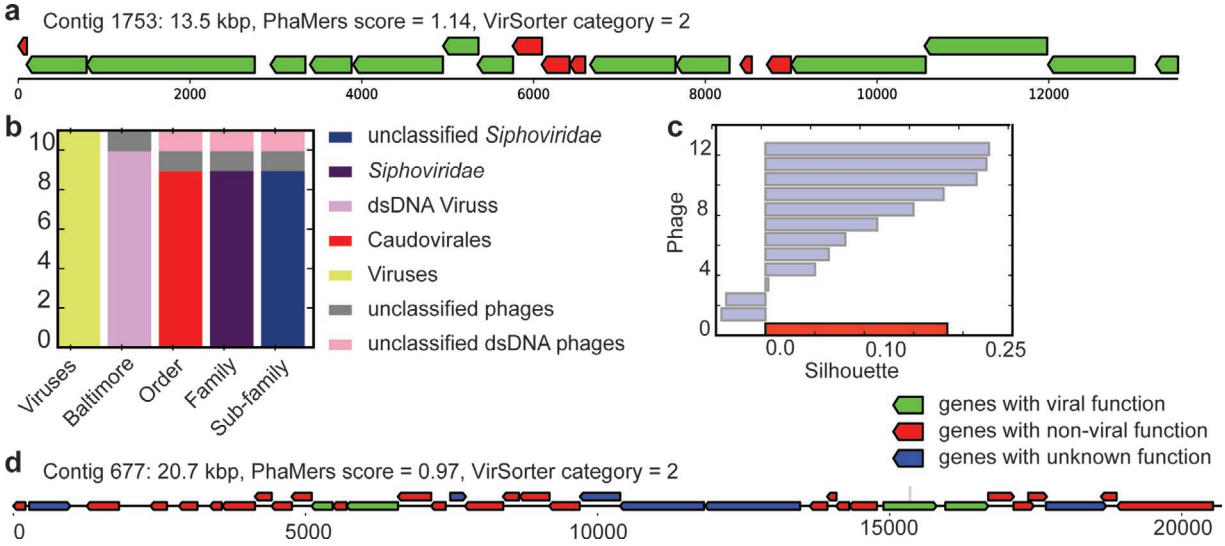

Figure 4. Selected novel phage contigs identified by both PhaMers and VirSorter. a) Diagram of 13.5 kbp contig (Contig 1753) from Mound Spring showing putative coding regions identified by VirSorter. b) Bar graph representing the composition of taxa for phages in the cluster to which contig 1753 was assigned by *k*-means ($k = 86$) on the basis of tetranucleotide frequencies. c) Cluster silhouette values for reference phage sequence (blue) and contig 1752 (red) assigned by *k*-means clustering. d) Diagram of contig 677 showing VirSorter annotations.

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

2 by VirSorter and 0.97 by PhaMers) contains four viral genes but is also enriched (33/36) for genes homologous to the thermophilic bacterium *Hydrogenobacter*, in the phylum *Aquificae* (Figure 4d). The enrichment for predicted genes homologous to *Hydrogenobacter* is an indication that *Hydrogenobacter* is a natural host. Finally, we characterized a single novel viral contig of length 35211 base pairs (Figure S6, Supporting Information). This contig was not identified as a phage by VirSorter but was given a score of 0.90 by PhaMers. Top hits using NCBI BLAST revealed similarities at both nucleotide and protein level to the thermophilic archaeal phage genera *Sulfobus filamentous* and *Acidianus filamentous*, belonging to the *Betalipothrixvirus* genus of the family *Lipothrixviridae*.[29] This contig was assigned on the basis of tetranucleotide frequencies to a cluster enriched with *Lactococcus phage 936 sensu lato*, an *unclassified Siphoviridae*. Protein blast revealed a 596 amino acid putative protein with 73% identity to a Holiday junction branch migration helicase from *Acidianus filamentous virus* 9, as well as a 563 amino acid putative protein with 72% identity to a helicase from *Acidianus filamentous virus* 9. Another 1038 amino acid putative protein had 48% identity to a conserved hypothetical protein of *Acidianus filamentous virus 3*,[8] not appearing in the *Sulfobus filamentous* genome.[30] These features indicate that, as a new phage in the *Betalipothrixvirus* genus, this contig has shorter phylogenetic distance to *Acidianus filamentous* than to *Sulfobus filamentous*.

2.4. Co-Occurrence from Mini-Metagenomics Enable Microbial Host Predictions

Mini-metagenomics distribute single bacterial cells randomly into subsamples containing 5–10 cells each. The occurrence of cells across subsamples that belong to a particular bacterial species or phage genome is valuable for binning microbial contigs[31] and assigning phage to potential host. To explore the ability of assigning putative phage contigs predicted by PhaMers, VirSorter, and DeepVirFinder to potential microbial host, we focused on 313 phage-like sequences over 10 kbp identified by one of the three tools from the Mound Spring sample. We clustered all contigs over 10 kbp from Mound Spring using t-SNE based on the contig's occurrence patterns across 93 mini-metagenomic subsamples (**Figure 5**a). Occurrence patters

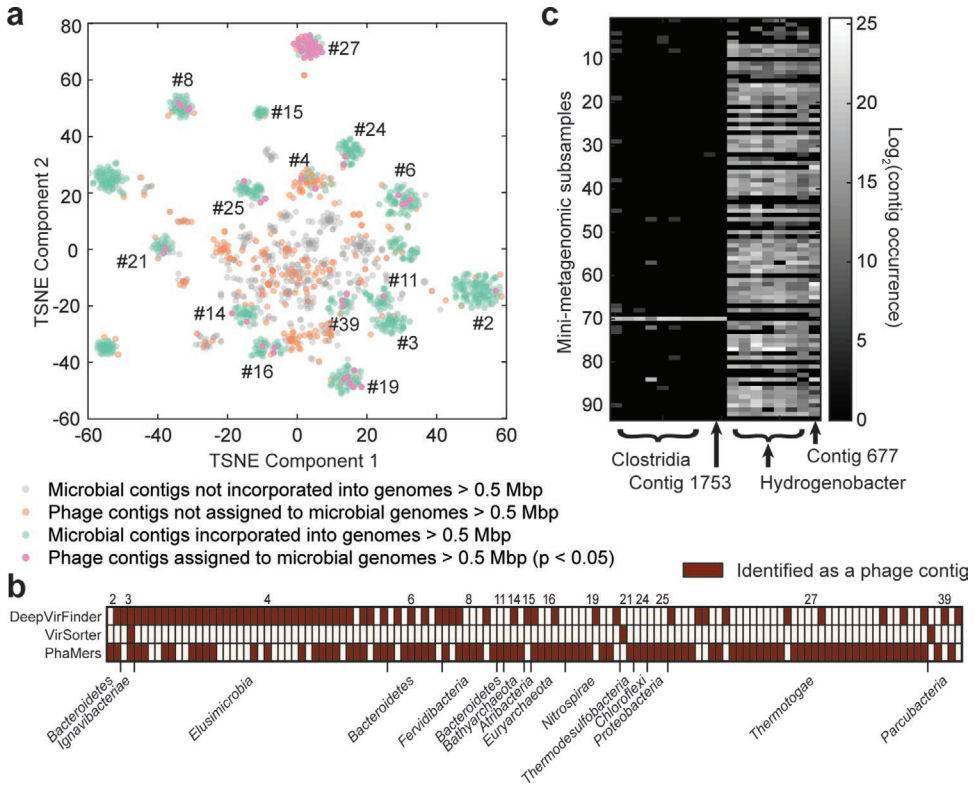

Figure 5. Assigning phage contigs to microbial host using mini-metagenomic co-occurrence. a) t-SNE plot of all contigs greater than 10 kbps (N = 1474) from Mound Spring clustered based on presence and absence patterns across 93 mini-metagenomic subsamples. Each green cluster represents binned microbial genomes larger than 0.5 Mbp. Pink dots represent phage contigs identified by PhaMers, VirSorter, DeepVirFinder, or a combination that are associated to a microbial host based on co-occurrence patterns ($p < 0.05$ using Fisher's Exact Test). Orange dots represent phage contigs identified by the same combination of tools that are either associated with microbial host genomes < 0.5 Mbp or are not associated with a microbial host ($p > 0.05$). b) Contigs (N = 125) in (a) identified as phage sequences that are associated with microbial host genomes > 0.5 Mbp. A brown rectangle shows that a contig is identified as phage by the corresponding method labeled to the left of the row. Numbers denote microbial genome clusters labeled in (a). Phylogeny of the microbial host is labeled below the graph. c) Contig occurrence patterns across 93 mini-metagenomic subsamples. Contig 1753 is plotted with related Clostridia contigs with similar occurrence patterns. Contig 677 is plotted along with a cluster of *Hydrogenobacter* contigs. Brightness represents base 2 logarithm of the contig abundance within each subsample, defined as the number of basepairs covered by at least 1 read in that subsample.

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

of each putative phage was compared with the occurrence pattern of the microbial genome using Fisher's exact test if size of the microbial genome was larger than 0.5 Mbp and the putative phage was clustered with said microbial genome (Experimental Section). Those with significant associations ($p < 0.05$) were plotted in pink while others are plotted in orange (Figure 5a). In addition, we identified microbial genomes based on gene annotation. Although some phage contigs were identified by both tools, majority of the phage contigs were either identified by PhaMers only or by DeepVirFinder only. Furthermore, PhaMers and DeepVirFinder identified putative phage contigs associated with different hosts (Figure 5b). Whereas DeepVirFinder identified more phage associated with *Elusimicrobia*, *Bacteroidetes*, and *Fervidibacteria*, PhaMers identified more archaeal viruses and phage associated with *Thermotogae*, *Nitrospirae*, *Atribacteria*, and *Fervidibacteria*. The identification of putative phages associated with different phyla of microbial hosts illustrates another advantage of developing and using PhaMers, a phage identification tool that uses complementary machine-learning methods as existing tools.

Based on co-occurrence patterns, some phage associate with high confidence ($p < 0.05$, Fisher's exact test) to microbial hosts whose genome is less complete (<5 Mbp). Two examples are contigs represented in Figure 4. Contig 1753 is only present in one mini-metagenomic subsample. We observe similar occurrence patterns across another set of nine contigs assigned to *Clostridiales* (Figure 5c). None of the other genomes appear only in the same mini-metagenomic subsample, suggesting that contig 1753 is more likely a *Siphoviridae* that infects *Clostridiales*, although we cannot tell if the *Siphoviridae* fragment is inserted into a *Clostridiales* genome because of its rare occurrence. Contig 677 was hypothesized to infect *Hydrogenobacter* based on predicted genes. This hypothesis is validated by the similar and abundant occurrence patterns between the phage contig and a partial *Hydrogenobacter* genome from the same environment ($p < 10^{-12}$, Fisher's exact test) (Figure 5c). While it may be possible that an infection is so prevalent that the phage is observed to co-occur with its host in a large fraction of hosts, it seems more likely that high co-occurrence statistics is a sign that the phage is either incorporated into the host's genome as a prophage or exists in the host as a plasmid.

3. Conclusion

The large numbers of recent metagenomic studies have generated exponentially increasing environmental sequencing data representing large prokaryotic and viral diversity. Although significant progress has been made in mining prokaryotic genomes from metagenomic datasets, finding phage genomes is still difficult, partly due to the lack of universal marker genes among phage genomes. The tetranucleotide frequency and machine-learning aspects of the PhaMers algorithm for both phage identification and classification may complement protein homology-based detection methods and other machine-learning based phage identification tools, enabling the discovery of broader classes of phage genomes. PhaMers permits phage classification and the identification of potential host by assessing proximity of nearby phage or prokaryotic genome clusters in the tetranucleotide frequency space. Combined with mini-metagenomic co-occurrence patterns, our method has the potential to assign putative phage to its microbial host and differentiate between a phage infection and bacterial cells that carry phage genomes. Although PhaMers's performance is affected by the comprehensiveness of reference datasets, most of which are dominated by few, well-characterized phage taxonomies, we have taken steps to mitigate such effects by using a reduced reference subset during cross validation. As the number of metagenomic datasets continues to increase, more phage genomes will be discovered, leading to more comprehensive reference phage databases. As phage databases grow to include more references, the combination of using multiple phage detection and classification methods with mini-metagenomics will be beneficial in elucidating the global viral diversity.

4. Experimental Section

Sample Collection: Environmental samples used in this study were collected from two separate hot springs from Yellowstone National Park under permit number YELL-2009-SCI-5788: sediments of the Bijah Spring in the Mammoth Norris Corridor and Mound Spring in the Lower Geyser Basin region. Samples were placed in 2-mL tubes and soaked in 50% ethanol onsite. Samples were spaced in 2-mL tubes without any filtering and soaked in 50% ethanol onsite. Upon returning, samples were transferred to −80 °C for long-term storage. Biosample information can be found from Joint Genomes Institute's GOLD system under Gb0114344 and Gb0114821, respectively.

Sample Preparation and Sequencing: Environmental samples were processed using the microfluidic-based mini-metagenomic protocol performed on a commercially available Fluidigm C1 Auto Prep IFC (Integrated Fluidic Circuit).[27] Steps performed on the automated microfluidic platform included cell partition, cell lysis, and genomic DNA amplification using MDA (Multiple Displacement Amplification). Amplified genomic DNA was harvested into a 96-well plate. The concentration was quantified independently using the high sensitivity large fragment analysis kit (AATI) and adjusted to 0.1–0.3 ng μL^{-1}, the input range of the Nextera XT library prep pipeline. Libraries were sequenced on the Illumina NextSeq (Illumina) platform using a 2 × 150 bp runs. Sequencing reads were filtered and assembled according to the methods described by Yu et al.[27] Assembly was performed via SPAdes V3.5.0 with *k*-mer values of 33, 55, 77, and 99. Contigs longer than 5 kbp were retained for phage analyses.

PhaMers Classification of Metagenomic Contigs: PhaMers' scoring algorithm was used to score assembled metagenomic contigs greater than 5 kbp. PhaMers uses BioPython to parse fasta formatted files of assembled contigs, and tabulates tetranucleotide frequencies before scoring. Results are written to file and are used for subsequent analysis. To assign putative taxonomic classifications to phage, *k*-means ($k = 86$) was used to cluster the tetranucleotide frequencies of each putative phage with those of the phages from the reference data set and examined the taxonomic composition of the phage sequences in the cluster that the contig was assigned to. A contig assigned to a taxonomically enriched cluster was considered if phage from a single taxon composed its cluster at a proportion greater than 50%. It was considered as evidence of a putative phage's taxonomic classification if a phage was assigned to an enriched cluster and if its cluster silhouette score was within one standard deviation of the mean cluster silhouette scores of the reference phage in the cluster.

Reference Database Generation: The reference dataset of genomic phage sequences was assembled using the Phage available on RefSeq in October of 2015. A complete list of all viral accession numbers made available on NCBI was downloaded and used to find accession numbers for all viruses that infect bacteria or archaea (Table S1, Supporting

© 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Information). This set of accession numbers was used to access and compile a set of all phage sequences in fasta format, subsequently used for tetranucleotide frequency analysis. The reference dataset of bacterial genomic sequences was generated from genomic assemblies available on GenBank (Table S4, Supporting Information). Bacterial species were selected at random from the subdirectories at ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/, and the latest genomic assembly fasta files were used for analysis in PhaMers. Tetranucleotide frequencies were calculated for all genomic fragments within each contig file, and a list of bacterial accession number a tetranucleotide counts.

PhaMers Classification of Known Phage and Bacterial Genomes: To verify PhaMers' predictive discrimination between phage and non-phage genomic sequences, a *k*-mer length of 4 (tetranucleotide) was selected and occurrences of each of the 256 tetramers in the 2255 phage and 2255 bacterial genomes of the reference dataset were counted. We then normalized each tetranucleotide count vector by the total number of tetramers counted to produce tetranucleotide frequency vectors, thereby discrediting differences in vectors due to sequence length. Then, 20-fold cross validation was performed on different scoring algorithms, wherein both the phage and bacterial datasets were divided into subdivisions and each subdivision was scored by using the remaining nineteen subdivisions as training data.

The following supervised learning algorithms were tested: Support Vector Machine, Kernel Density Estimation, K-Nearest Neighbors (KNN), and Nearest Centroid, and linear combinations of results from each (Figure 2a). KNN was chosen as the primary classifier because it performed with lowest false positive rate while maintaining >90% sensitivity. The parameter specifying number of neighbors used in KNN classification (K) was varied from 3 to 20 during cross validation on the reference datasets. Increasing K yielded marginally decreased performance, hence informing our choice of $K = 3$. To add additional information into the final PhaMers score, the initial classification by KNN was taken to be a −1 (non-phage) or 1 (phage) and added to it a parameter between −1 and 1 that quantified the proximity of a point to phage clusters, and distance away from bacterial clusters. (Figure S4, Supporting Information) This was chosen because this algorithm performed well on its own, and quantifies relative distances to large groups of reference data, whereas KNN classified based on more local data points.

To study the relationship between tetranucleotide frequencies and phage taxonomy, the dimensionality of this reference phage tetramer frequency vectors was reduced from 256 to two using t-SNE (Figure 1a). The reduced dimensionality tetramer frequency vectors were then clustered using density-based spatial clustering DBSCAN,[25] and the prevalence of each taxa in each cluster was quantified.

VirSorter and DeepVirFinder Analysis of Metagenomic Contigs: Samples were analyzed using the VirSorter 1.0.3 phage identification pipeline available through the iPlant Discovery Environment on the iPlant collaborative website, made available by CyVerse. (https://de.iplantcollaborative.org/de/) VirSorter used all bacterial and archaeal virus genomes in Refseq, as of January 2014 for the analysis of both metagenomic samples. DeepVirFinder version 1.0 (https://github.com/jessieren/DeepVirFinder) was used to analyze all mini-metagenomic contigs over 5 kbp using default options plus "-l 5000." Contigs receiving a score greater than 0.7 and a *p*-value less than 0.05 are taken as valid viral predictions.

Annotation of Putative Phage Contigs: Contigs were uploaded to JGI's Integrated Microbial Genomes Expert Review online database (IMG/ER). Annotation was performed via IMG/ER.[32] Briefly, structural annotations were performed to identify CRISPRs (pilercr), tRNA (tRNAscan), and rRNA (hmmsearch). Protein coding genes were identified with a four ab initio gene prediction tools: GeneMark, Prodigal, MetaGeneAnnotator, and FragGeneScan. Functional annotation was achieved by associating protein coding genes with COGs, Pfams, KO terms, and EC numbers. Phylogenetic lineage was assigned to each contig based on gene assignment. Annotations can be found under genome IDs 3300006068 and 3300006065.

Software Availability: Code and algorithms used by PhaMers were tested in MATLAB and implemented in the Python 2.7 programming language. Python 2.7 was used to write scripts for parsing of VirSorter, DeepVirFinder, and IMG output files and for integration with PhaMers data. All PhaMers scripts are available at https://github.com/jondeaton/PhaMers. The Python library Matplotlib was used for plot generation.

Statistical Analysis: p values were computed using Fisher's exact test to assign a putative phage to its potential microbial host. Based on occurrence patterns of both the phage sequence and the associated microbial genome across 93 mini-metagenomic subsamples, a 2×2 contingency table is created including the number of subsamples where both phage and microbe are present (a), phage is present (b), microbe is present (c), and both are absent (d). Fisher's exact test then uses the following equation:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$$

Associations are deemed valid if $p < 0.05$.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors of this work would like to acknowledge members of the Quake Lab Sequencing Facility: Norma Neff, Jennifer Okamoto, Gary Mantalas, and Ben Passarelli. The authors would also like to acknowledge generous support from the DOE JGI's Emerging Technologies Opportunities Program (ETOP), John Templeton Foundation, Stanford Graduate Fellowship, NSF GRFP, and the Stanford Bioengineering REU program. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. Conceptualization was done by J.D., F.B.Y., and S.Q.; Methodology was done by J.D., F.B.Y., and S.Q.; Software was handled by J.D. and F.B.Y.; Investigation was done by J.D., F.B.Y., and S.Q.; Writing was done by J.D., F.B.Y., and S.Q.; Supervision was done by F.B.Y. and S.Q.

Conflict of Interest

Dr. Stephen Quake is a shareholder of Fluidigm Corporation, whose instrument was used for mini-metagenomic sample preparation.

Keywords

hot spring, mini-metagenomics, phage, PhaMers, yellowstone

Received: May 12, 2019
Revised: July 10, 2019
Published online:

[1] C. Canchaya; G. Fournous; S. Chibani-Chennoufi; M.-L. Dillmann; H. Brüssow, *Curr. Opin. Microbiol.* **2003**, *6*, 417.
[2] S. W. Wilhelm, Corina P D Brussaard, F. Thingstad, M. G. Weinbauer, G. Bratbak, M. Heldal, S. A. Kimmance,

 © 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

M. Middelboe, K. Nagasaki, J. H. Paul, D. C. Schroeder, C. A. Suttle, D. Vaqué, K. E. Wommack, *ISME J.* **2008**, *2*, 575.
[3] G. F. Hatfull, *Curr. Opin. Microbiol.* **2008**, *11*, 447.
[4] R. A. Edwards, F. Rohwer, *Nat. Rev. Microbiol.* **2005**, *3*, 504.
[5] R. A. Edwards, K. McNair, K. Fraust, J. Raes, B. E. Dutilh, *FEMS Microbiol. Rev.* **2016**, *40*, 258.
[6] S. Roux, J. B. Emerson, E. A. Eloe-Fadrosh, M. B. Sullivan, *PeerJ* **2017**, *5*, e3817.
[7] S. Roux, F. Enault, B. L. Hurwitz, M. B. Sullivan, *PeerJ* **2015**, https://doi.org/10.7717/peerj.3817.
[8] D. Arndt, J. R. Grant, A. Marcu, T. Sajed, A. Pon, Y. Liang, D. S. Wishart, *Nucleic Acids Res.* **2016**, *44*, W16.
[9] D. Paez-Espino, G. A. Pavlopoulos, N. N. Ivanova, N. C. Kyrpides, *Nat. Protoc.* **2017**, *12*, 1673.
[10] J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, F. Z. Sun, *Microbiome* **2017**, https://doi.org/10.1186/s40168-017-0283-5.
[11] J. Villarroel, K. A. Kleinheinz, V. I. Jurtz, H. Zschach, O. Lund, M. Nielsen, M. V. Larsen, *Viruses* **2016**, *8*, 116.
[12] D. T. Pride, T. M. Wassenaar, C. Ghose, M. J. Blaser, *BMC Genomics* **2006**, https://doi.org/10.1186/1471-2164-7-8.
[13] J. X. Tan, F. Y. Dao, H. Lv, P. M. Feng, H. Ding, *Molecules* **2018**, *23*, 2000.
[14] B. Manavalan, T. H. Shin, G. Lee, *Front Microbiol* **2018**, https://doi.org/10.3389/fmicb.2018.00476.
[15] J. H. Cheng, H. Yang, M. L. Liu, W. Su, P. M. Feng, H. Ding, W. Chen, H. Lin, *Chemom. Intell. Lab. Syst.* **2018**, *180*, 64.
[16] H. Bin Jang, B. Bolduc, O. Zablocki, J. H. Kuhn, S. Roux, E. M. Adriaenssens, J. R. Brister, A. M. Kropinski, M. Krupovic, R. Lavigne, D. Turner, M. B. Sullivan, *Nat. Biotechnol.* **2019**, *37*, 632.
[17] D. Paez-Espino, E. A. Eloe-Fadrosh, G. A. Pavlopoulos, A. D. Thomas, M. Huntemann, N. Mikhailova, E. Rubin, N. N. Ivanova, N. C. Kyrpides, *Nature* **2016**, *536*, 425.
[18] C. Galiez, M. Siebert, F. Enault, J. Vincent, J. Soding, *Bioinformatics* **2017**, *33*, 3113.
[19] N. A. Ahlgren, J. Ren, Y. Y. Lu, J. A. Fuhrman, F. Z. Sun, *Nucleic Acids Res.* **2017**, *45*, 39.
[20] J. Ren, K. Song, C. Deng, N. A. Ahlgren, J. A. Fuhrman, Y. Li, X. Xie, F. Sun, *arXiv* **2018**, *DOI:arXiv:1806.07810*.
[21] M. Victor M., C. I-Min A., P. Krishna, C. Ken, S. Ernest, P. Manoj, R. Anna, H. Jinghua, W. Tanja, H. Marcel, A. Iain, B. Konstantinos, V. Neha, M. Konstantinos, P. Amrita, N. N. Ivanova, N. C. Kyrpides, *Nucleic Acids Res.* **2013**, *42*, D560.
[22] W. C. Li, E. Z. Deng, H. Ding, W. Chen, H. Lin, *Chemom. Intell. Lab. Syst.* **2015**, *141*, 100.
[23] P. Deschavanne, M. S. DuBow, C. Regeard, *Virol. J.* **2010**, *7*, 163.
[24] L. van der Maaten, G. Hinton, *J. Machine Learn. Res.* **2008**, *9*, 2579.
[25] M. Ester, H.-P. Kriegel, J. Sander, *Proc. 2nd Int. Conf. Knowledge Discovery Data Mining*, Portland, OR **1996**.
[26] J. Manlioff, H.-W. Ackermann, *Arch. Virol.* **1998**, *143*, 2051.
[27] F. Yu, P. C. Blainey, F. Schulz, T. Woyke, M. A. Horowitz, S. R. Quake, *eLife* **2017**, https://doi.org/10.7554/eLife.26580.001.
[28] B. A. Berghuis, F. B. Yu, F. Schulz, P. C. Blainey, T. Woyke, S. R. Quake, *PNAS* **2019**, 116, 5037.
[29] A. Fokine, M. G. Rossmann, *Bacteriophage* **2014**, *4*, https://doi.org/10.4161/bact.28281.
[30] G. Vestergaard, R. Aramayo, T. Basta, M. Häring, X. Peng, K. Brügger, L. Chen, R. Rachel, N. Boisset, R. A. Garrett, D. Prangishvili, *J. Virol.* **2008**, *82*, 371.
[31] C. Yu, T. Hernandez, H. Zheng, S.-C. Yau, H.-H. Huang, R. He, J. Yang, S. Yau, *PLoS One* **2013**, https://doi.org/10.1371/journal.pone.0064328.
[32] M. Huntemann, N. N. Ivanova, K. Mavromatis, J. H. Tripp, D. Paez-Espino, K. Tennessen, K. Palaniappan, E. Szeto, M. Pillay, I.-M. A. Chen, A. Pati, T. Nielsen, V. M. Markowitz, N. C. Kyrpides, *Stand Genomic Sci.* **2016**, *11*.

 © 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim