12-2016

Strategies for Identifying the Optimal Length of K-mer in a Viral Phylogenomic Analysis using Genomic Alignment-free Method

Qian Zhang
University of Tennessee, Knoxville, qzhang24@vols.utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by Qian Zhang entitled "Strategies for Identifying the Optimal Length of K-mer in a Viral Phylogenomic Analysis using Genomic Alignment-free Method." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Life Sciences.

Dave Ussery, Major Professor

We have read this thesis and recommend its acceptance:

Mike Leuze, Colleen Jonsson

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Strategies for Identifying the Optimal Length of K-mer in a Viral Phylogenomic Analysis using Genomic Alignment-free Method

A Thesis Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Qian Zhang

December 2016

I dedicate this work to my family, who supported my dream and pushed me toward higher education. For my parents who give me life and teach me for living well. To my husband Dao, who is always by my side and supports me through rough time. And also to my daughter Cassie, I wish you grow up healthy and happy (Kaixin), chase your dream, love and serve.

ACKNOWLEDGEMENTS

I would like to begin by thanking Dr. Dave Ussery, my advisor for all of his support and guidance through this process. I would like to thank Drs. Intawat Nookaew and Se-Ran Jun, who spent many hours helping me with every detail of this research. I would also like to acknowledge Dr. Mike Leuze, Miriam Land and Visanu Wanchai for their assistance.

I would like to thank Dr. Dave Ussery, Dr. Mike Leuze and Dr. Colleen Jonsson for serving on my review committee.

ABSTRACT

Whole genome sequencing has been rapidly developed and widely used, made possible by exponentially decreasing cost and computational advances in biological sequence analysis. Massive amount of viral sequences has been produced. By Oct 2016, over 102,000 of records has been archived in NCBI Viral Genome Project and 7730 genomes are RefSeq genomes. To better understand viral classification, phylogenomic analysis, which based on whole-genome information, provides the possibility of reconstructing a “tree of life”. However, there are difficulties to apply phylogenomic methods to large-scale viral genomes. In this study, we designed a 3-step strategy for identifying the optimal length of K-mer in a viral phylogenomic analysis using genomic alignment-free method. These three steps include: 1) Cumulative Relative Entropy, 2) Average Number of Common Features among genomes, and 3) Shannon Diversity Index. A dendrogram of 3905 RefSeq viral genomes has also been constructed by using the optimal K = 9. The resulting dendrogram shows consistency with the viral taxonomy and the Baltimore classification of viruses.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE
INTRODUCTION

NCBI Viral Genomes Project

Over the past decade, DNA sequencing technology has been rapidly developed and widely used, while the cost of DNA sequencing falls off exponentially [1]. Benefit by the reducing sequencing cost and the rising throughput, massive amount of microbial whole-genome sequences have been used in microbial identification and characterization [2,3]. For viral genomic research, the National Center for Biotechnology Information (NCBI) Viral Genomes Project has produced over 102,000 of records representing thousands of different species by October 6, 2016, and this number has increased explosively since the new millennium [4].

NCBI RefSeq Database

To better represent the complete sequence information for any given species, the viral NCBI Reference Sequence (RefSeq) database provides a curated, non-redundant sequence collection of viral genomes [5]. Among different complete genome sequences from various isolates and strains in the same species, only one sequence would be selected as a reference to work as a molecular standard [4]. As of October 2016, 7730 genomes have been archived in viral RefSeq database.

Phylogenetic Analysis vs. Phylogenomic Analysis

Phylogenetic analysis is the means of inferring or estimating evolutionary relationships among molecules, organisms or both [6]. It is widely used for microbial characterization [7,8], gene and protein function prediction [9,10] , drug development [11], and other biomedical areas. Generally, a basic phylogenetic analysis has four steps: alignment, model selection, tree building and tree evaluation [6]. The phylogenetic alignment is all about mapping the relationships between residues in

a set of DNA/RNA sequences or amino acid sequences, in order to produce plausible hypotheses of evolutionary homology among these residues [12]. The most popular methods of constructing phylogenetic trees fall into three categories: 1) distance-based methods: such as Neighbor Joining (NJ), Unweighted Pair Group Method with Arithmetic Mean (UPGMA); 2) Maximum parsimony; 3) maximum likelihood methods. The commonly used valuation methods are Bootstrap [13] and Jackknife[14].

However, for prokaryotes, phylogenetic trees based on small subunit ribosomal RNAs (SSU rRNAs) often do not agree with those based on different genes. More genes and genomes sequenced, more conflicts have been found among gene trees [15]. For viruses, proteins are very diverse and it is difficult to reconstruct phylogenetic tree based on conserved proteins among various viruses, especially when some viruses only have one or two genes. To get more robust information for phylogeny inference, phylogenomic trees have been constructed based on whole-genome/ whole-proteome information. Most phylogenomic methods are either sequence-based, such as multiple alignment and supertree/supermatrix construction, or based on whole-genome features like gene orders, gene content and DNA-string comparisons [15]. Nonetheless, these phylogenomic methods still have some problems with huge tree space, assessing the statistical confidence of trees and “divide-conquer” resolution, etc [15].

Alignment-free Methods

For phylogenomic analysis of large-scale genomes, especially highly diverse ones, alignment-free methods have been increasingly used in the past few years [16–19]. These alignment-free methods could be classified into two categories, according to different theoretical basis: one based on statistics of word frequency, the other on Kolmogorov complexity and chaos theory [20]. Comparing to alignment-based methods, these alignment-free methods are of a linear complexity and efficient [21].

Different from traditional model-based phylogenetic analysis, alignment-free phylogeny may not provide an evolutionary interpretation but perform as "dendrogram". However, alignment-free methods are essential to compare large-scale distant genomes, since they greatly accelerate the computation speed and solve the sequence comparison problem that cannot be otherwise done by alignment-based methods.

Feature Frequency Profile Method

Sims *et al.* [22] introduced an alignment-free method that uses a measure based on Jensen–Shannon Divergence between Feature Frequency Profiles (FFPs), where the features, called K-mers, are short nucleotide or amino acid sequences of length K. This FFP method has been applied in previous eukaryotic and prokaryotic studies [23,24], and shows great agreement with organism taxonomies. For viruses, this method was also applied to whole-proteome sequences of 142 large dsDNA viruses [25]. However, there is little work available on using FFP to determine the phylogeny of large-scale viral genomes [26,27].

A major challenge is identifying the optimal K-mer length when using the FFP method for comparing whole genomes. In previous studies [24,25,28], the optimal K has been identifies as the value when both Cumulative Relative Entropy (CRE) and Relative Sequence Divergence (RSD) decrease to less than 10% of their maximum values as K is increased. However, we found these two criteria cannot be achieved when we construct a phylogenomic tree of thousand viral genomes with various genome sizes. To solve this problem, we developed a comprehensive strategy for identifying the optimal length of k-mer in our large-scale viral phylogenomic analysis, which includes Cumulative Relative Entropy, Average Number of Common Features among genomes and Shannon Diversity Index to identify the optimal K-mer.

CHAPTER TWO

MANUSCRIPT: VIRAL PHYLOGENOMICS USING ALIGNMENT-FREE METHOD: HOW TO FIND AN OPTIMAL LENGTH OF LENGTH OF K-MER?

Qian Zhang[2], Se-Ran Jun[1], Michael Leuze[3,4], David Ussery[1], Intawat Nookaew[1, *]

[1] Department of Biomedical Informatics, College of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA

[2] UT-ORNL Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN 37996 USA

[3] Joint Institute for Computational Sciences, University of Tennessee, Knoxville, TN 37831, USA

[4] Computational Biomolecular Modeling and Bioinformatics Group, Computer Science and Mathematics Division, Oak Ridge National Laboratories, Oak Ridge, TN 37831, USA

* Corresponding Author

Phone: 501-603-1986

Fax: 501-526-5964

Email: INookaew@uams.edu

Abstract

Development of genome sequencing sheds a new light on the classification of viruses. The NCBI provides about two million nucleotide sequences of viruses, and thousands of viral reference sequences that cover a wide range of viral taxonomy in the RefSeq database. Whole genome information has been used to obtain a better classification, and it may open new possibilities for the viral “tree of life”. However, it is not feasible to build the tree of life using traditional phylogenetic methods based on conserved proteins due to the lack of evolutionary conservation among diverse viruses. In this study, we employed alignment-free method which uses K-mers as genomic features for large-scale comparison of complete viral genomes available in RefSeq. To determine optimal feature length K, which is essential step to obtain a good dendrogram, we designed a comprehensive strategy that uses a combination of key three strategies: 1) Cumulative Relative Entropy; 2) Average Number of Common Features among genomes 3) Shannon Diversity Index to identify the optimal K-mer. Ultimately, we derived a procedure to decide the optimal feature length for the comparison of all 3905 complete viral genomes. The optimal dendrogram showed great consistency with viral taxonomy of ICTV and Baltimore classification.

Introduction

Whole genome sequencing (WGS) is now commonly used [29–31], made possible by exponential reductions in the cost of sequencing [32] and computational advances in biological sequence analysis [33,34]. Viral taxonomy, in particular, has benefited from the availability of many new viral genome sequences, enabling improved classification of viruses. In support of viral genomics research, the NCBI Viral Genome Project [35] provides thousands of viral reference sequences that cover a wide range of viral taxonomic species in the NCBI Reference Sequence Database. The classification of viruses is maintained by the International Committee on

Taxonomy of Viruses (ICTV), which considers multiple viral properties and consensus data [4], including similarities in genome structures, host ranges, and the presence of homologous genes and various phylogenetic features [36]. Although viral taxa have been continuously updated by the virus research community [37,38], there are still many misclassifications in ICTV viral taxonomy [39]. Further, sequencing of viral metagenomics samples often results in many viral genomes that are of unknown origin [40,41].

Phylogenetic analysis is widely used for taxonomic identification, characterization, and revision [42,43]. However, for prokaryotic genomes, phylogenetic trees based on SSU rRNAs often do not agree with those based on different genes. Conflicts among gene trees have increased as more genes and genomes are sequenced [15]. This incongruence is caused by many reasons, including tree-building errors, incomplete lineage sorting, hidden paralogy, and horizontal gene transfer (HGT). For viruses, as early as 1996, inconsistent phylogenetic trees were obtained when using different numbers of isolates or different lengths of aligned sequences in a study of hepatitis C viruses [44]. Similar inconsistencies have been reported for human papillomaviruses [45], SARS coronavirus [46], and some plant viruses [47].

Phylogenomic trees constructed using whole-genome sequences are based on a more complete set of genomic information than phylogenies based on individual genes [48]. For large-scale comparisons of genome-scale sequences, especially highly diverse ones, alignment-free methods of phylogeny construction have been increasingly used in the past few years [16–19]. There are two categories of alignment-free methods for phylogenomic analysis: one based on statistics of word frequency, the other on Kolmogorov complexity and chaos theory[20]. The primary advantage of these methods is that they enable quick genome-scale comparisons with linear time complexity (O(n))[21], more efficiently than minimum likelihood or Bayesian alignment methods with sub-quadratic time complexity ($o(n^2)$). Another

advantage of alignment-free methods is that they can be used to compare sequences from unfinished genomes, with information loss proportional to the number of discontinuities in a genome. However, alignment-free methods do not capture the nuances of evolutionary models that incorporate site-dependent substitution patterns. Therefore, it is not possible to interprete branch lengths of alignment-free based trees in terms of mutation rates, even though alignment-free trees constructed from whole genome sequences capture taxonomic classification (which reflects the evolutionary history of organisms) better than 16S rRNA alignment based trees for prokarytoes[21].

Sims *et al.* [22] introduced an alignment-free method that uses a measure based on Jensen–Shannon Divergence between Feature Frequency Profiles (FFPs), where the features, called K-mers, are short nucleotide or amino acid sequences of length K. Applied in eukaryote and prokaryote sytems, this approach shows great agreement with taxonomic information accepted by scientific community [23,24]. For viruses, Wu *et al.* [25] applied the FFP method to whole-proteome sequences of 142 large dsDNA eukaryote viruses, and Huang *et al*. used this approach when evaluating different methods for phylogenetic analysis of multiple-segmented viruses [49,50]. To date, however, relatively little work has been done using FFP to determine the phylogeny of virus genomes [51], and there are only a few reports [26,27] on construction of phylogenetic trees from thousands of viral genomes.

In general, genome-scale phylogenetic trees can be built using either whole-genome sequences or whole-proteome sequences. However, some viruses have only one or two genes from which protein sequences can be predicted, and viral proteins tend to be very diverse. As a consequence, it is not feasible to build a viral "tree of life" based on conserved proteins. We have, therefore, used an FFP approach applied to complete viral genome sequences and have built a dendogram of viruses.

A major challenge in using the FFP method for comparing whole genomes is determining the optimal K-mer length. In previous studies of dsDNA eukaryote viruses [19, 22, 23], the optimal feature length was based on Cumulative Relative Entropy (CRE) and Relative Sequence Divergence (RSD). For each individual genome and a value of K, the CRE, determined by a comparison of the observed FFP and the expected FFP from a second-order Markov model, captures how much information of the whole genome sequence is encoded in the FFP. In other words, CRE indicates the power of the FFP to reconstruct the whole genome sequence. Smaller CRE values, which result from longer K-mers, are indicitave of the ability to better identify individual genomes. For a whole genome, the RSD for a value of K is a measure of the relatedness of the genome sequence (in terms of FFP) to a random sequence of the same length. According to Wu *et al.* [25], the optimal value of K is the value when both CRE and RSD decrease to less than 10% of their maximum values as K is increased.

Determining RSD values becomes increasingly computationally complex as the number of genomes grows. This increase in complexity is due, in part, to an increase in the density of the K-mer feature space. We found RSD cannot monotonically decrease when k increases, which is probably because this huge dimersional K-mer space can cover artificial K-mers (K-mers derived from random sequences), even though their probability are quite low. However, calculation of RSD values becomes increasingly complex as the number of genomes grows. This increase in complexity is due, in part, to an increase in the density of the K-mer feature space.

In this study, we consider 3905 complete viral genomes available in the NCBI Reference Sequence Database (RefSeq) [52]. We show that CRE is significantly influenced by genome size as well as K-mer composition. Genomes of different sizes show different trend CRE curves. For small viral genomes (~3kb), CRE

values drop to zero around K value of 6; for large viral genomes (1Mb or more), the drop increases to K value of 10. Consequently, CRE values for genomes of greatly various size cannot simultaneously be decreased to less than 10% of maximum values at the fixed feature length as suggested by Wu et al [23]. Accordingly, we first group viral genomes by genome size. For each group, we propose the optimal K-mer length considering several genomic features, including the CRE value, the number of K-mers shared by genomes, and the total number of K-mers observed, and construct a dendrogram at its optimal K-mer length. Finally, we derive a procedure to decide the optimal feature length for the comparison of all 3905 complete viral genomes. The tree of life of viral whole-genomes constructed by our precedure of alignment-free method is visualized using the optimal feature length for the global view.

Results

Dataset and information content evaluation

The non-redundant dataset includes 3905 complete genomes of RefSeq viruses as summarized in Table. S1. The smallest genome is the *Anguilla anguilla* circovirus (NC_023421), with a length of 1,378 nt and the largest genome is *Pandoravirus salinus* (NC_022098), which is 2,473,870 nt long. The distribution of genome sizes is depicted as the density plot in Figure 1. The long tail is on the right shows there are some large genome sizes as outliers such as *Pandoraviruses*, *Megaviruses*, *Mimiviruses* and other giant viruses. It is worth mentioning that, after determining the Cumulative Relative Entropy (CRE) values as shown in Figure 2, we noted that the recommended range for K-mer length varies greatly, depending on genome size, and divided the dataset into 4 arbitary subgroups (Q1 - Q4) using the 25%, 50% and 75% quartiles of 6,407, 12,141 and 45,242 bp, respectively.

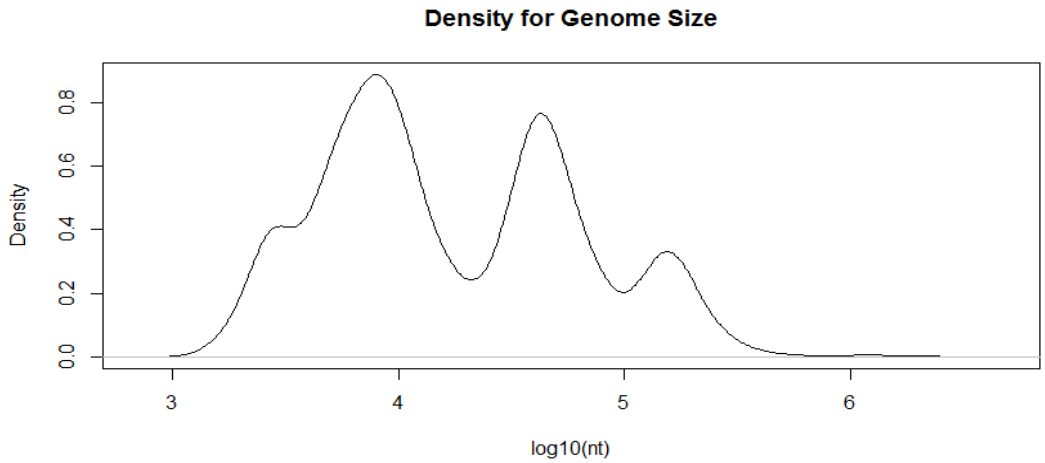

Figure 1 Distribution of genome size for 3905 viral genomes in semi- logX scale.

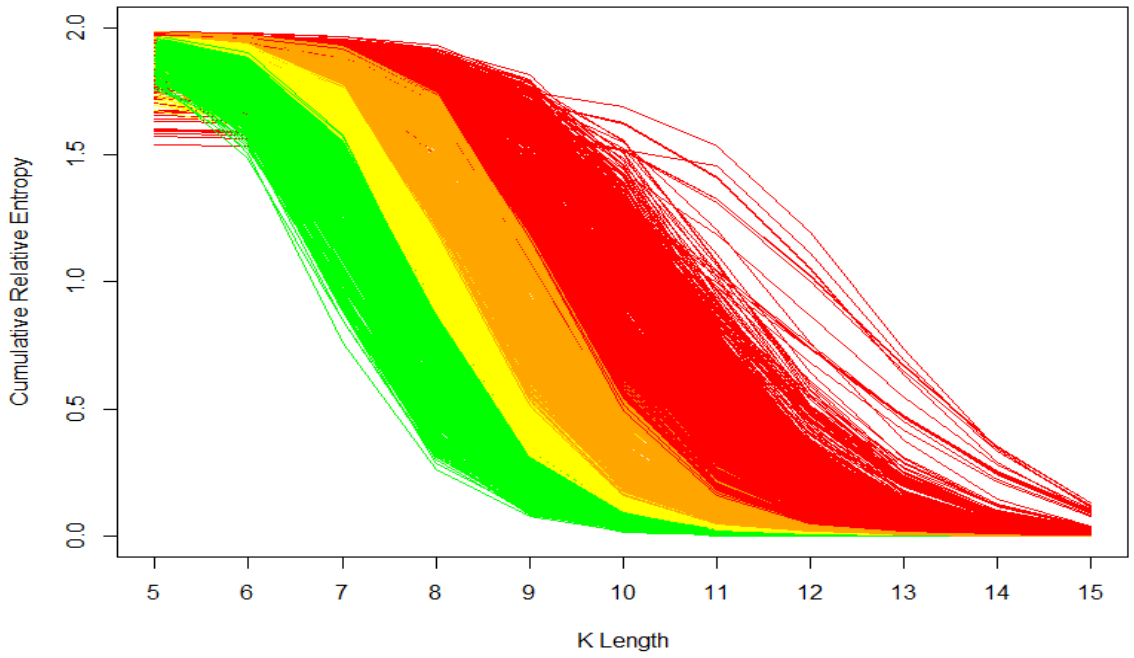

Figure 2 Cumulative Relative Entropy curves for 3905 viral RefSeq genomes. The curves start to fall below 10% of the maximum at k = 9 and most genomes satisify the criteria at k=13. Subgroups Q1, Q2, Q3 and Q4 are colored as green, yellow, orange and red.

Assessment of Optimal Feature Length (K)

Since the criteria used by Wu *et al.* [25] are not directly applicable to our large-scale virus dataset, due to the dependence of CRE on genome size, we determined optimal feature length based on three criteria: 1) from an individual genome perspective, using Cumulative Relative Entropy (CRE) to find the minimum feature length: where the genome curves reach zero CRE or fall to <10% of their CRE maximum values; this CRE value is the original criterion of optimal feature lengths in previous published papers[24,25,28]; 2) from a pairwise comparison perspective, Average Number of Common Features (ACF) among genomes is applied to determine the maximum feature length: the length prior to ACF dropping to a lower value; this ACF criterion is defined as the average number of common features when comparing pairwise to each of the other genomes at a specific feature length; 3) from an "all genomes comparison" perspective, we measure commonness of K-mers among all genomes in our dataset in terms of diversity index to narrow the range of optimal feature length down. Shannon Diversity Index is used to quantity the diversity of commonness of K-mers using fraction of K-mers shared by genomes. The preferred length is the one with higher Shannon Diversity Index value (which represents more diversity of commonness of K-mers) in the range suggested from criteria (1) and (2); 4) additionally, the tree stability, which is based on Robinson-Foulds distance, is also considered as supporting information, especially when multiple lengths in the range are suggested (see Materials and Methods Section for more details).

Cumulative Relative Entropy (CRE)

For each individual genome, CRE values were calculated by increasing K-mer length from 5 to 15. We plotted CRE values for 3905 reference viral genomes, illustrated in Figure 2, which is colored by genome size and is ordered from smallest to largest genome. Cumulative Relative Entropy (CRE) curves do not simultaneously drop to <10% of maximum CRE for all genomes, which is the

selection criterion that Wu *et al.* [25] recommend. When curves for smaller genomes achieve that goal, some curves for larger genomes are still at a plateau. At K = 9, the curves of small genomes start to fall below 10% of maximum CREs, and roughly 50% of all CREs drop below 20% of their maxima. At larger values of K (K = 10, 11 and 12), more genome CREs satisfy the less than 10% of maximum criterion. When K = 13, the CRE values of most genomes fall below 10% of maximum CREs. However, K = 13 cannot be simply chosen as the optimal feature length, because it might be too large (no information left) for small genomes. By quartile, the optimal K-mer lengths for subgroups Q1, Q2, Q3, and Q4 are determined to be 9 to 11, 10 to 12, 11 to 13 and 12 to 15, respectively. Therefore, we initially determined the optimal range of K-mer lengths for the entire set of 3905 genomes to be 9 to 13. This range will be refined in the following steps.

Average Number of Common Features (ACF)

Previously computed RSD values were found to not work as expected (that is, they did not converge to zero after reaching the optimal feature length). Because of this, we did not use the comparison with random feature space, and instead we only used the denominator of RSD to explore the common features between pairwise genomes, which we call the 'Average Number of Common Features' (ACF). For each genome, the Average Number of Common Features is defined as the average number of common features from a pairwise comparison of all the other genomes at a specific feature length (See Materials and Methods). Because FFP is a pairwise-comparing method, the ACF is not expected to be very low at the specific feature length. Otherwise, the obtained information will tend to be randomized, which means it could produce a random phylogeny.

First, in order to reveal the shared degree of features at different length, we calculated ACF among 3905 RefSeq viral genomes by comparing each genome with the other 3904 ones at different feature lengths, as plotted in Figure 3. The

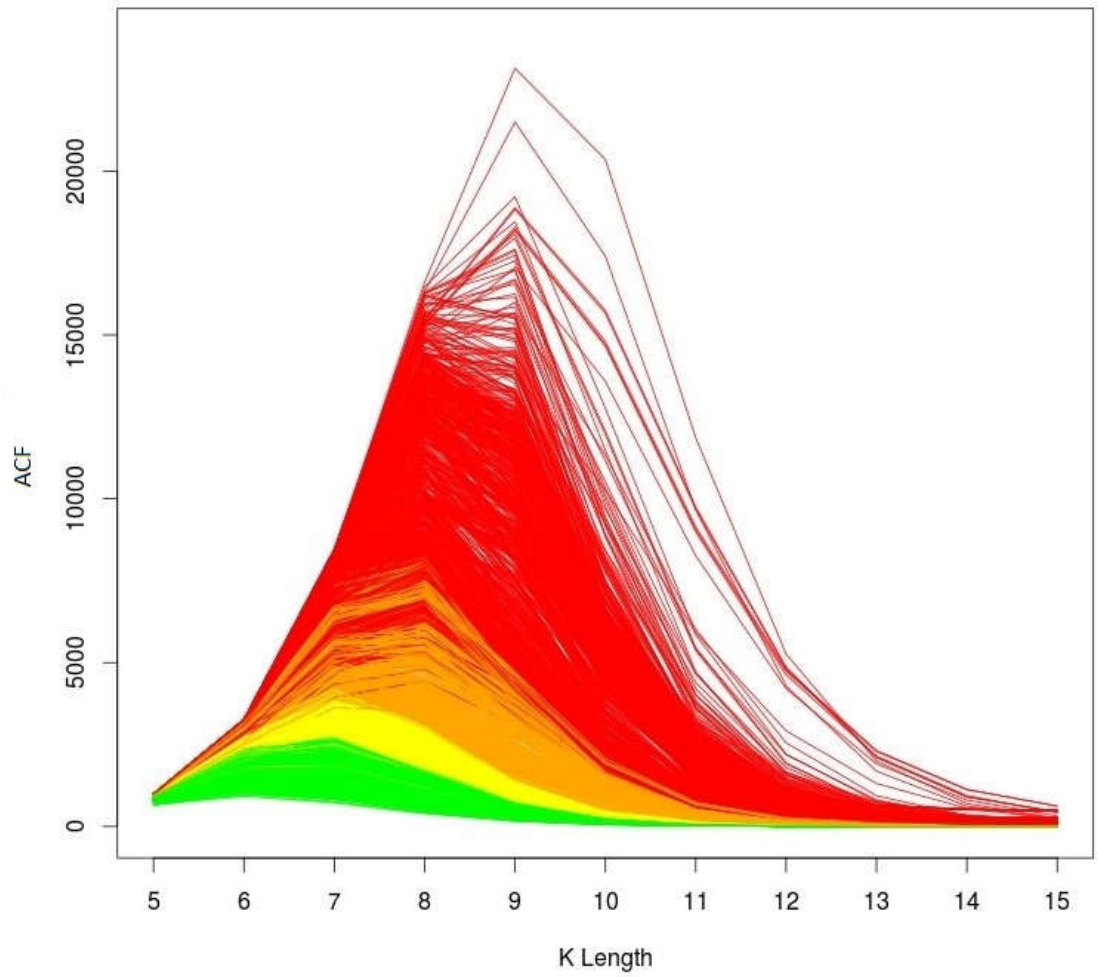

Figure 3 Average Number of Common Features (ACF) for 3905 viral RefSeq genomes. Each curve shows the ACF numbers between this individual genome and other 3904 genomes. Subgroups Q1, Q2, Q3 and Q4 are colored by green, yellow, orange and red.

ACF plot demonstrated that few features are shared when the feature length is larger than 11 (k >11). As a result, the maximal feature length for 3905 genomes should be 11 nucleotides. So, the range based on CRE values is reduced to the range between 9 to 11. These curves were also colored by different levels of genome sizes, as in subgroups Q1, Q2, Q3 and Q4. Apparently, the ACFs stack up with increase of genome size. As we estimated, when k = 13, many of the features of small genomes (in Q1 subgroup) are shared, which implies that we cannot only consider only CRE criterion to choose the optimal k.

Finally, we also calculated ACF values for subgroups (Figure 4), by comparing each genome with the other 995 or 996 ones in the same quartile. The maximal optimal feature lengths for Q1, Q2, Q3 and Q4 are found to be 10, 11, 12 and 13. As a result, the optimal feature ranges are reduced to 9-10, 10-11, 11-12 and 12-13.

All observed feature occurrences in genomes

The unions of all observed features at different lengths have been calculated and compared with theoretical occurrences, as shown in Table 1. Obviously, the numbers of observed non-redundant features increase exponentially as powers of alphabetical size (4 for nucleotide sequences); when k <13, the total redundant feature number (165,838,971) largely covers the expected feature space. However, when k > 13, the numbers of observed non-redundant features grow more slowly in subgroups, all of the numbers also present the similar trends.

The optimal K-mer length necessary for construction of a good dendrogram should give the balance of overlap and unique features among the genome dataset. To illustrate the relationship between “all features” and “all genomes”, the distribution of feature occurrences in genomes is calculated and plotted. As shown in Figure 5. When the feature length is small (k = 5, 6), most features can be found in most genomes; when feature length is large (k = 14, 15), most features (>50% or 80%)

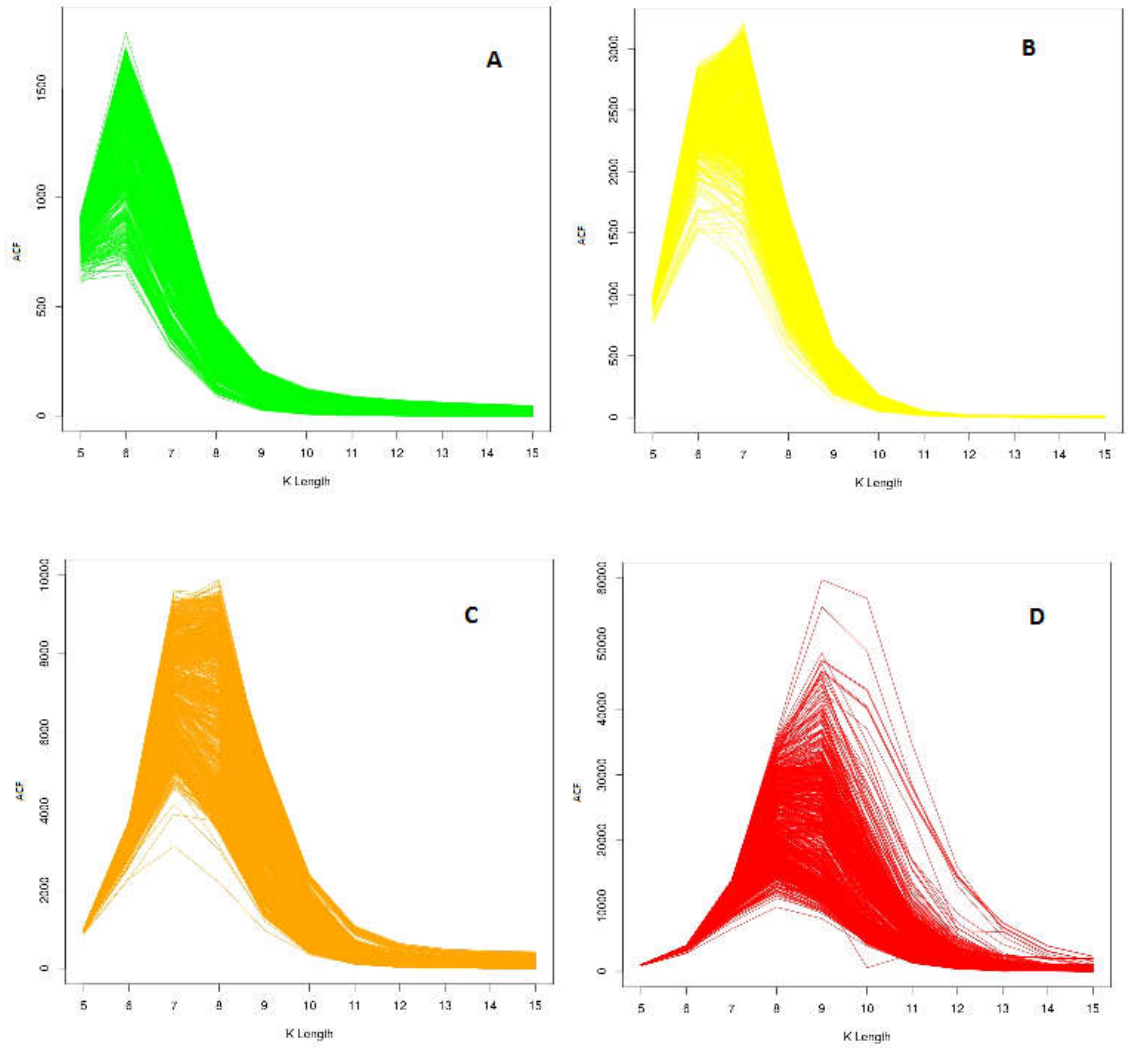

Figure 4 Average Number of Common Features (ACF) for viral RefSeq genomes in four subgroups. A) Q1 subgroup (genome size < 25% quartile): 976 genomes, colored by green; B) Q2 subgroup (genome size in 25% -50% quartiles): 977 genomes, colored by yellow; C) Q3 subgroup (genome size in 50%-75% quartiles): 977 genomes, colored by orange; D) Q4 subgroup (genome size > 75% quartile): 977 genomes, colored by red.

Table 1 Numbers of all observed non-redundant features in 3905 genomes and in subgroups.

K	Expected (4^k)	Observed	Observed in subgroups			
			Q1	Q2	Q3	Q4
5	1,024	1,024	1,024	1,024	1,024	1,024
	%obs/exp	100	100	100	100	100
6	4,096	4,096	4,096	4,096	4,096	4,096
	%obs/exp	100	100	100	100	100
7	16,384	16,384	16,384	16,384	16,384	16,384
	%obs/exp	100	100	100	100	100
8	65,536	65,536	65,536	65,536	65,536	65,536
	%obs/exp	100	100	100	100	100
9	262,144	262,144	261,744	262,135	262,144	262,144
	%obs/exp	100	99.84	99.99	100	100
10	1,048,576	1,048,576	927,225	1,028,114	1,048,272	1,048,576
	%obs/exp	100	88.42	98.04	99.97	100
11	4,193,940	4,193,940	1,983,092	3,133,972	4,011,469	4,191,555
	%obs/exp	99.99	47.28	74.72	95.64	99.94
12	16,777,216	16,405,985	2,691,077	5,776,434	10,767,534	15,878,890
	%obs/exp	97.79	16.04	34.43	64.17	94.64
13	67,108,864	48,841,160	2,999,146	7,352,145	17,313,110	41,880,927
	%obs/exp	72.78	4.46	10.95	25.79	62.40
14	268,435,456	87,268,900	3,134,521	7,979,080	20,718,374	67,931,028
	%obs/exp	32.51	1.16	2.97	7.71	25.30
15	1,073,741,824	111,123,028	3,211,835	8,210,153	22,064,213	83,014,712
	%obs/exp	10.35	0.29	0.76	2.05	7.73

*Total number of redundant features for 3905 genomes is 165,838,971; all percentages are calculated based on expected ones. %obs/exp = percent of obserbed/expected K-mer

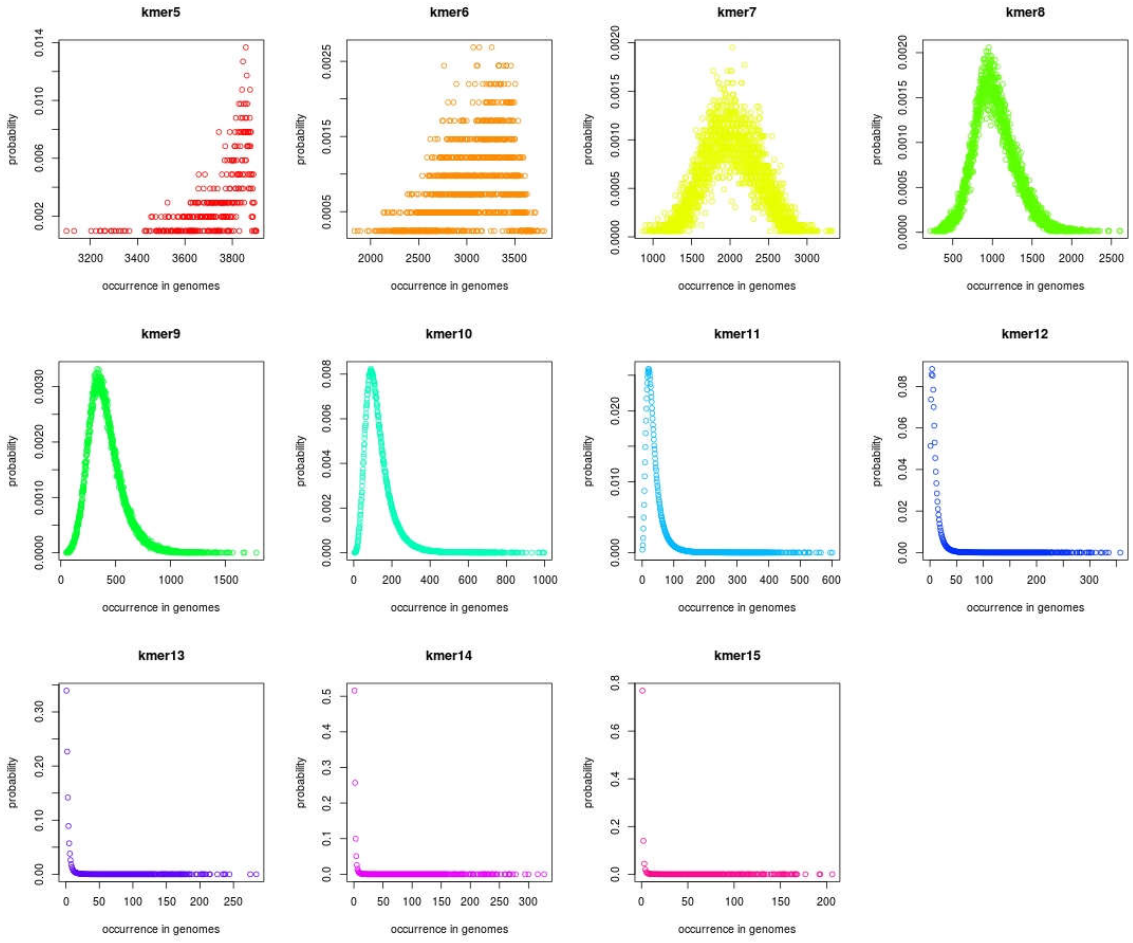

Figure 5 Distribution of feature occurrences in genomes. A dot represents a unique kmer. Y axis represents proability (kmer fraction) calculated from the observed frequency of individual kmer divided by total number of observed kmer, X axis represents number of genomes that share the same kmers.

are unique (occurrence = 1). In both these scenarios, FFP cannot work efficiently. After all, the feature occurrences should be diverse to balance the similarity and dissimilarity when comparing all genomes. For this purpose, Shannon Diversity Index was applied and plotted with different feature lengths (Figure 6). From the curve, the diversity of feature occurrence peaks at k = 7, and then steadily. In this regard, k = 9 is more appropriate than 10 and 11 within our previous optimal feature range.

For each of the four subgroups, we repeated the same process, and obtained Figure S1-S4 for distributions and Figure 7 for Shannon Diversity Index. Finally, the optimal feature length for Q1, Q2, Q3 and Q4 was determined as 9, 10, 11 and 12, respectively.

What is the optimal feature length?

All results for above criteria have been summarized in Table 2. For the dendrogram of 3905 viral genomes, either 9 or 11 can be chosen as the optimal feature length. k = 10 has lower ACF and Shannon diversity indicating non-linear relationship in the dataset. When k = 9, CRE values have not dropped to <10% of their maximum, the other two criteria perform well. And when k = 11, most of CRE values drop to <10% of their maximum, while the Average Number of Common Features (ACF) is not good for small viral genomes. In this case, it is hard to choose between 9 and 11, because neither of them can perfectly satisfy our three criteria. So it makes sense to check the tree stability and use it as a supporting information for this study. To evaluate the tree stability, we calculated Robinson-Foulds distances between k (5, 6, 7…) and k+1 at different feature lengths. When the Robinson-Foulds distances drop to a low value, it means the tree stability starts at this k point and tree topology does not change much as feature lengths increase. As shown in Figure 8, trees start to converge at k =9, so we will choose k = 9 as the optimal feature length of this dendrogram. Furthermore, since we want to obtain a global

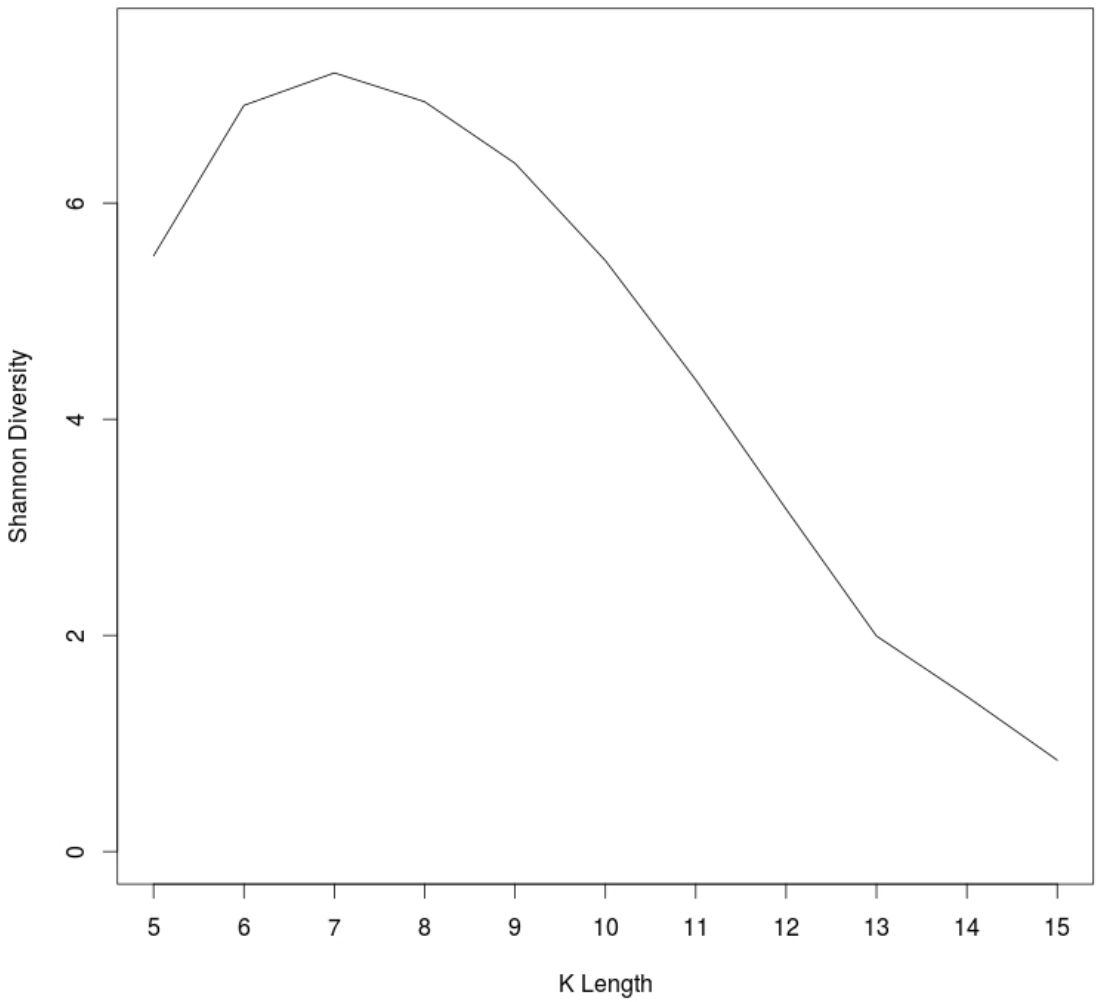

Figure 6 Shannon Diversity Index for feature occurrence in genomes as a function of kmer length.

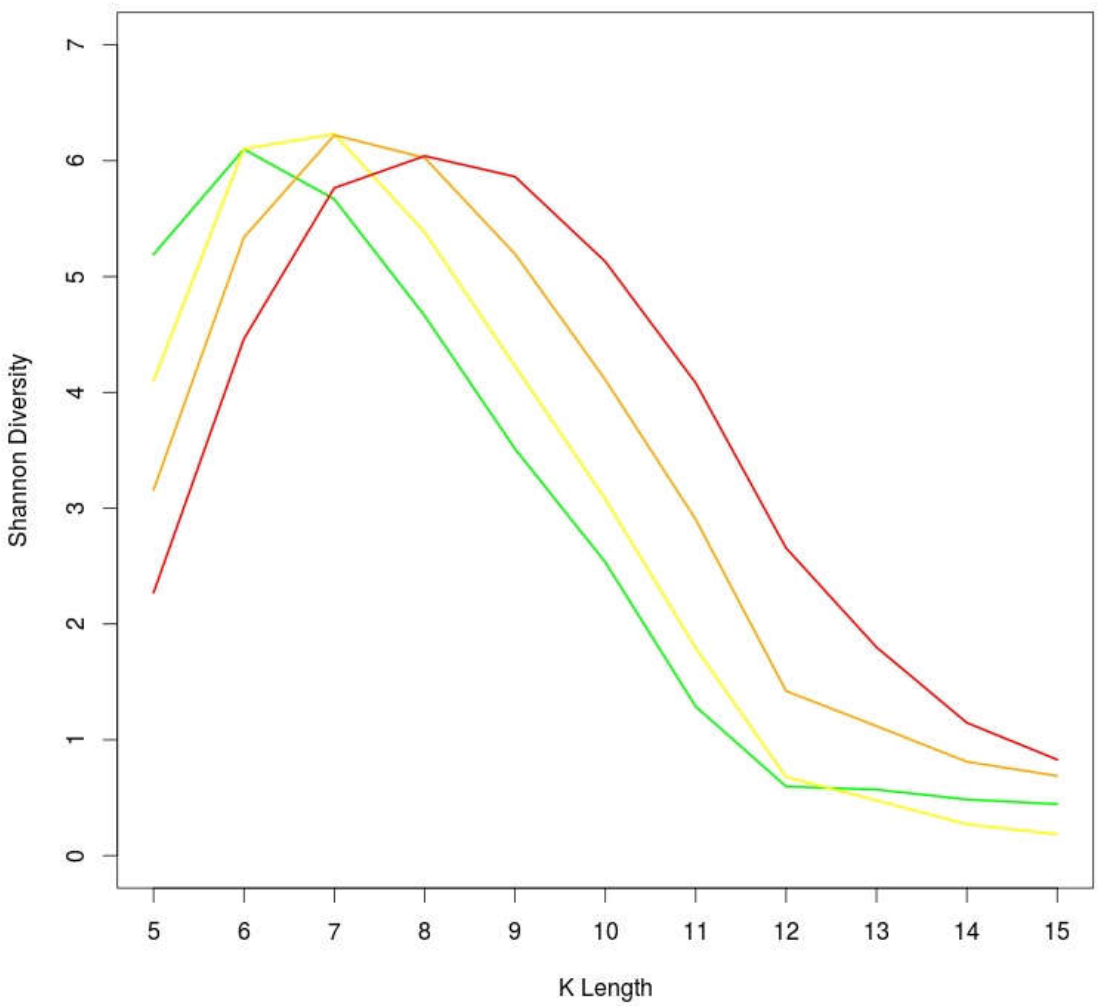

Figure 7 Shannon Diversity Index for feature occurrence in four subgroups a function of kmer length. Q1 subgroup (genome size < 25% quartile): 976 genomes, colored by green; Q2 subgroup (genome size in 25% -50% quartiles): 977 genomes, colored by yellow; Q3 subgroup (genome size in 50%-75% quartiles): 977 genomes, colored by orange; Q4 subgroup (genome size > 75% quartile): 977 genomes, colored by red.

view of the relationship among RefSeq viral genomes, the 'pairwise comparison perspective' and 'all genome comparison perspective' are considered more important in this research, than exactly estimation of individual genomes, especially when all sequences are RefSeq whole genomes (not so similar and sensitive). For dendrograms of 4 subgroups Q1, Q2, Q3 and Q4, the optimal feature lengths have been identified as k = 9, 10, 11 and 12, respectively.

Table 2 Summary for optimal feature length.

	Whole database	**Q1**	**Q2**	**Q3**	**Q4**
Step 1: CRE	9, 10, 11, 12,13	9, 10, 11	10, 11, 12	11, 12, 13	12, 13, 14
Step 2: ACF	9, 10, 11	9,10	10, 11	11, 12	12, 13
Step 3: feature Occurrence in genomes	9 or 11*	9	10	11	12
Optimal feature length	9 or 11*	9	10	11	12

*k = 9 performs best in step 3 and k = 11 performs best in step 1

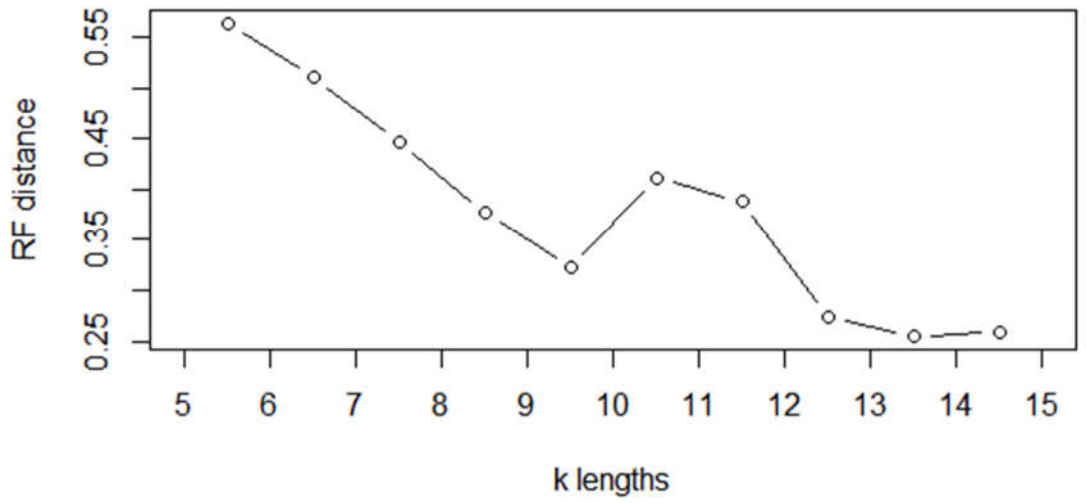

Figure 8 Robinson-Foulds distance between trees at feature length k (5, 6, 7, ...) and k +1.

Phylogenomic Analysis of 3905 viral RefSeq genomes

Based on the 3 steps assessment, the dendrogram of all 3905 RefSeq viruses (k =9) is shown in Figure 9. This dendrogram is built by Neighbor Joining method using all FFP values as pairwise distances. As a whole, the taxonomic groupings of 3905 viral whole genomes agree well with the reference taxonomy. The dendrogram is colored by Baltimore Classification, viral orders, kingdom of hosts and different levels of genome sizes. From this dendrogram, a global view of all relationships among 3905 viral RefSeq genomes is demonstrated. With hundreds whole-genomes of Ebola viruses sequenced in 2015 West Africa Outbreak. This dendrogram was used as the preliminary step to show the global view of clustering when compare the diverse set of viral taxa, and then rigorous analysis based on traditional methods were employed to analyze the genomic variation of among Ebola virus[53].

As shown in Figure 9, all branches of the dendrogram are colored by Baltimore Classification, including dsDNA viruses, dsRNA viruses, Retro-transcribing viruses, ssDNA viruses, ssRNA positive-strand viruses, ssRNA negative-strand viruses. In our dendrogram, dsDNA viruses, the largest taxon, are classified into five major groups, which are one large group, one middle size, and three small groups. The second major group, ssRNA(+) virus, forms multiple small clades and interlaces among other groups. ssDNA viruses also form five groups, which are one large group and four small groups. ssRNA(-) viruses and Retro-transcribing viruses organize two relatively independent clades, respectively.

The innermost circle of the dendrogram is colored by reference taxonomy at different orders, including Caudovirales, Herpesvirales, Ligamenvirales, Mononegavirales, Nidovirales, Picornavirales, Tymovirales and unclassified ones. From Table S1, around the reference order of 60% viruses is Caudovirales in our

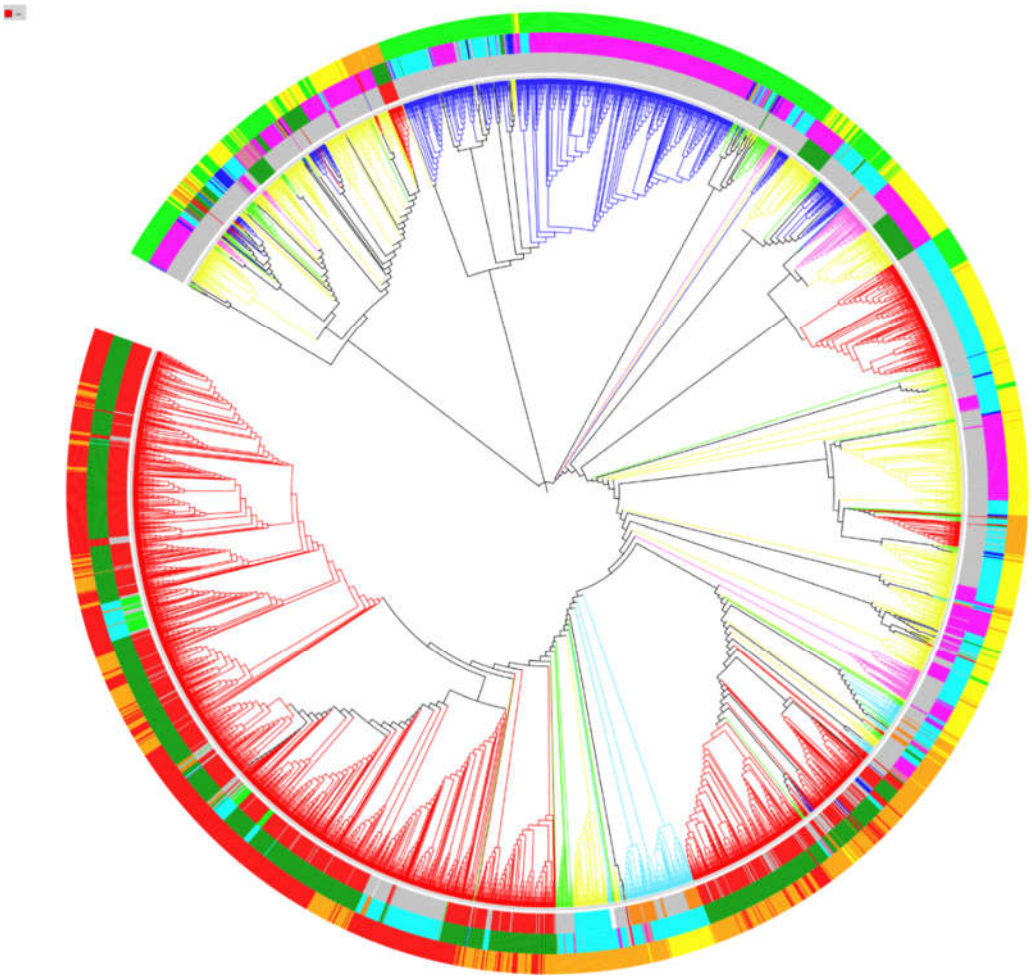

Figure 9 Optimal dendrogram of 3905 RefSeq viral genomes (k = 9). The braches are colored by Baltimore Classifications. The circles, from inside to outside, are colored by different orders, hosts and genome sizes. [Color information: (A) Branch: Baltimore Classification; dsDNA, no RNA stage: red; dsRNA: green; Retro-transcribing viruses: pink; ssDNA: blue; ssRNA negative-strand: bright blue; ssRNA positive-strand: yellow. (B) From inside to outside, first circle: Order; Caudovirales: red; Herpesvirales: green; Ligamenvirales: blue; Mononegavirales: orange; Nidovirales: cyan: Picornavirales: pink; Tymovirales: dark green; unclassified: silver; (C) From inside to outside, second circle: Host; protest: orange; archaea: red; bacteria: dark green; fungi: blue; animal: cyan; animal and plants: pale violet red; plant: pink; environment or NA: silver. (D) From inside to outside, third circle: genome size: Q1: Green, Q2: Yellow, Q3: Orange, Q4: Red.]

database, excluding 2171 viruses whose reference orders are unclassified or unassigned.Ignoring the unclassified part, those Caudovirales viruses group well, with a few membership discrepancies. It is interesting to note, Herpesvirales viruses form a small clade to split the largest clade of Caudovirales. Other Herpesvirales viruses also groups inside Caudovirales clades as discrepancies. Ligamenvirales, Mononegavirales, Nidovirales, Picornavirales and Tymovirales separate from each other to form small sporadic groups.

The second circle shows the kingdoms of hosts, including archaea, bacteria, fungi, animal, plants, protist and environment. As can be seen, the host kingdom of most dsDNA viruses is bacteria. The plant viruses mainly remain in ssDNA viruses and ssRNA(+) viruses. The animal viruses distribute around the whole dendrogram, and response to various sequence structures and reference orders, which suggests their possible origins from transmission. The outside circle is colored by different levels of genome sizes. The overall trend is that genomes with similar sizes are easier to get together, although colors mix as local changes.

We observed form the figure 9 that, there are a correlation between length of genome and dendrogram grouping as seen in the outer circle. So the dendrogram of subgroup base one the optimal K-mer as reported in the Table 2 will give a better taxonomic resolution.

Statistical Analysis for Grouping Uncertainty

The RefSeq dataset of 3905 genomes contains 97 known families (by the ICTV annotation), and 59 genomes do not have information about their families (missing or “unassigned” in GenBank). The ten largest families, as listed in material and methods, were evaluated for grouping uncertainty (Huang *et al* [50]). Considering the dendrogram derived from the optimal K = 9 the descriptive statistics of within-group and between-group distances of different viral families were calculated by the Kruskal-Wallis one-way analysis of variance and the Wilcoxon rank sum test..

For the Kruskal-Wallis one-way analysis of variance, the null hypothesis, which is that the within-group and between-group distances of the largest ten families have equal means, is rejected ($p\text{-}value < 2.2 \times 10^{-16}$). The pairwise Wilcoxon rank sum test shows the within-group distances are smaller than the between-group distance for each viral family (almost $p\text{-}values < 2.2 \times 10^{-16}$). Both statistical results strongly indicate the good grouping of the constructed dendrogram and its consistency with ICTV annotation. Detailed results of the statistical analysis are provided in Supplementary table S2.

Subgroup Dendrograms

The dendrogram (k = 9) of 976 RefSeq viral genomes in subgroup Q1 (genome size < 25%) is shown in Figure S5. In this dendrogram, ssDNA viruses make up a large majority, and most of them are clustered together to form a large clade (which branches colored by blue). This clade has been separated by two main kinds of viral hosts, plants and animals. The other large clade of animal viruses is formed by two independen clusters of ssDNA and dsDNA. ssRNA(+) , dsRNA and RT viruses also can be observed. These three classes form independent small clades respectively, and then cluster with each other. Also, likewise with the host information. The orders of most viruses in subgroup Q1 are unclassified, except some from Tymovirales.

In Figure S6, the dendrogram (k = 10) of 977 RefSeq viral genomes in subgroup Q2 (genome size: 25%-50%), ssRNA(+) viruses roughly forms three clusters at different scales. The largest cluster of ssRNA(+) has been interrupted by a few RT viruses and ssDNA viruses, and then forms two clades. These two clades can be distinguished by host features, which means animal and plant ssRNA(+) viruses are separated in this cluster. Also, Tymovirales viruses in this cluster are grouped well. The medium cluster of ssRNA(+) viruses is made up of plant viruses, and Tymovirales viruses are distingushed with Picornavirales viruses.

As shown in the dendrogram (k = 11) of 977 RefSeq viral genomes in subgroup Q3 (genome size: 50%-75%) (Figure S7), more than 60% viruses are dsDNA viruses. They are clustered together in this dendrogram, and most of them are in Caudovirales Family and bacterial viruses, while some special cases are either archaeaviruses in Ligamenvirales Family or unclassified animal viruses. The other 40% viruses in this dendrogram are mainly ssRNA(+) viruses, ssRNA(-) viruses and dsRNA viruses. Each of them forms a few small clusters and then grouped with others. It is worth noting that animal ssRNA(+) viruses are closer to animal dsRNA viruses than to plant ssRNA(+) viruses, although the latter ones are in the same classification. Also, in this dendrogram, Mononegavirales viruses have a independent clade with different hosts.

For the largest viruses, all most all of them are dsDNA viruses (Figure S8). The Caudovirales viruses, most of which are bacterial viruses, form three large clades. Among these three clades are animal viruses with a few protist viruses, which orders are *Herpesvirales* or unknown.

Discussion

Identifying optimal feature length in a alignment-free phylogenomic method is the most important but challenging process, especially when we construct phylogenomic trees for large-scale datasets of divergent genomes of various size. In this study, we have developed a comprehensive strategy to find the optimal length of K-mer in alignment-free phylogenomic analysis, and built phylogenomic dendrogram for all complete viral genomes in NCBI RefSeq as of December, 2014 [54].

With the development of sequencing technologies, whole-genome information presents new possibilities for microbial classification [55]. Comparing to traditional

gene trees, whole-genome phylogenies use completed genomic information and solve the incongruence generated by gene trees from various studies. The alignment-free method with K-mers is useful for comparing genomes with low homology and has been applied to various microbial studies. However, it is still not clear how to find the optimal feature length of K-mer in alignment-free phylogenomic analysis especially for large-scale comparison of viral genomes. CRE and RSD values have been used as criteria in previous studies[22,24,25,28], but these studies used at most hundreds of genomes and their lengths do not change greatly. However, thousands of viral genomes in NCBI RefSeq showed a great difference in size which ranged from the smallest one (*Anguilla anguilla* circovirus) 1,378 to the largest one (*Pandoravirus salinus*) 2,473,870. As a result, their CRE curves cannot simultaneously drop to <10% of maximum as required in previous study. Furthermore, CRE reflects the ability to identify individual whole genomes at various lengths of K. More details should be taken into consideration when dealing with such highly-diverse data, such as pairwise comparison information and shared K-mers among all genomes. Hence, we divided our dataset into four subgroups by 25%, 50% and 75% quantiles of genome size.

In this study, we designed a comprehensive strategy to find the optimal length of K-mer for alignment-free FFP phylogenomic analysis. This comprehensive strategy combines three steps: 1) an individual genome perspective: Cumulative Relative Entropy (CRE) to find the minimum feature length; 2) pairwise comparison perspective where Average Number of Common Features (ACF) among genomes is applied to determine the maximum feature length; 3) an all-genome comparison perspective where Shannon Diversity Index of all observed feature occurrences in genomes to find the optimal feature length between the minimum and the maximum. And then, tree stability information, which obtained from Robinson-Foulds distance, can be used to determine the optimal length K if results are not unique. Based on these criteria above, the optimal feature lengths for each

subgroup has been identified shown in Table2. To get a hint of the global relationship of all 3905 viral whole genomes, we chose the smallest K (K=9) among the optimal feature lengths for subgroups as an acceptable feature length and constructed a dendrogram of all viral whole genomes.

In conclusion, our 3-step comprehensive strategy was successfully applied to identify the optimal feature length K in an alignment-free phylogenomic analysis for thousands of whole-genomes with highly-diverse sizes. Moreover, our dendrogram with the optimal feature length derived from all complete viral genomes gives a global view of classification in good agreement with the current viral taxonomy reported by ICTV and Baltimore classification. Moreover, this overall dendrogram can also be used as a preliminary step to show the global view of clustering of the diverse viral taxa and further analyze the genomic variation by traditional methods of specific viruses, especially Ebola viruses responsible for the recent outbreak in 2015 West Africa [53].

Materials and Methods

Dataset

5326 RefSeq viral genomes were downloaded from the RefSeq: NCBI Reference Sequence Database[54] (http://www.ncbi.nlm.nih.gov/refseq/) by the end of 2014. After merged all multiple-segmented genomes from the same virus, 4300 genomes were obtained. Viroid and satellite data has been excluded from the dataset, and then 3905 genomes were determined for this research. All genome data was converted to k-mer feature counts by using Jellyfish[56]. The database was also divided into four subsets by 25%, 50% and 75% quantiles of genome size, in order to fit different optimal feature lengths.

Feature Frequency Profile (FFP) and Phylogenomic Trees

All phylogenomic trees are calculated based on Feature Frequency Profile (FFP)-based distance matrices[22]. All criteria, which are related to optimal feature lengths, have been computed in parallel by Python 2.7. Phylogenomic trees are calculated from distance matrices based on Neighbor Joining method, by using R package phytools [57]. All dendrograms were plotted by the ITOL online tool (http://itol.embl.de/itol.cgi), and the other figures were generated by R software.

Optimal feature lengths

As shown in Figure 10, the optimal feature lengths have been determined by three criteria: 1) from individual genome perspective using Cumulative Relative Entropy (CRE); 2) from pairwise comparison perspective: Average Number of Common Features (ACF) among genomes; 3) from all genome comparison perspective: all observed feature occurrences in genomes. If multiple values of feature lengths are determined after this process, tree stability will be used to find the optimal length.

Cumulative Relative Entropy (CRE): A general description of CRE can be found in previously published paper [28], and the optimal feature length K was considered as where genome curves start having zero CRE or falling to <10% of their CRE maximum values. The CRE has been calculated as[25]:

$$CRE(l) = \sum_{k=l}^{\infty} RE(F_k, \hat{F}_k) \quad (1)$$

and

$$RE(F_l, \hat{F}_k) = \sum_i f_i log_2 \frac{f_i}{\hat{f}_i} \quad (2)$$

Where l is the feature length, f_i is the observed feature frequency, and $\hat{f}_i$ is the expected frequency formulated from K-2 Markov chain as in the previous publication [58] . Since the Relative Entropy (Kullback–Leibler divergence) [59] is always non-negative value, the function of CRE is monotonically decreasing.

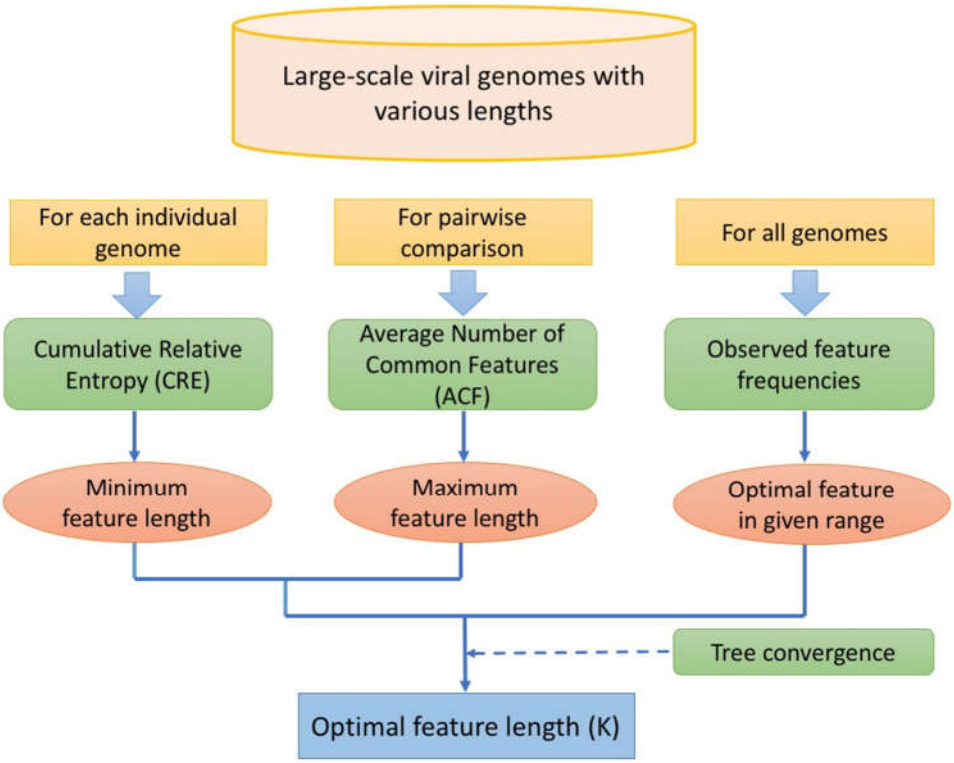

Figure 10 The 3-step assessment to obtain optimal feature lengths (k).

In previous published papers [22,25], Relative Sequence Divergence (RSD) has also been used to determine the optimal feature length. However, RSD cannot be applied for this research. Because our 3905 genomes provide a huge feature space, the overlap in feature space between the viral genomes and random sequence does not reduce. As a result, not all RSD values decrease to zero as expected. From another aspect, the random sequences are only generated once, without any iteration, and the iteration can be time-costing. So, RSD was failed to be used in this research. But enlightened by this value, we developed Average

Number of Common Features (ACF) to check the overlap in feature space among genomes.

Average Number of Common Features (ACF): For pairwise genomes, the similarity in FFP method is actually held by the common features between them. When the K is small, most features in one viral genome can be shared by the other one. However, the all possible feature number is small (4^K), so the average number should be low. On the other side, when the K is very large, because the features are long, only a few features can be shared between pairwise genomes. In this case, FFP may not provide enough signals for phylogeny and may show a random phylogeny. Therefore, the optimal K should be chosen before the ACF dropping to low values. The ACF can be defined as:

$$ACF(l) = \sum_{j \neq i} c(s_i, s_j, l)/(N-1) \quad (3)$$

where $c(s_i, s_j, l)$ is the number of common feature of length l between sequences s_i and s_j, and N is the genome number is the database. We used 10% of the maximum ACF of the considered population as suggestive cut-off similar to the suggestion on RSD [22,25].

All observed feature occurrences in genomes: From the perspective of all genomes, to balance the similarity and dissimilarity, neither of these situations is acceptable in FFP: 1) most features can be found in most genomes (when feature length is too small); 2) most features are unique (when feature length is too large). In this purpose, the unions of all observed features at different k were calculated in our dataset, and also their occurrence in genomes. Theoretically, the number of all possible features is 4^K. However, the biological sequence is not a random combination of alphabets. As a result, the percentage of observed ones decreases with feature length increasing, in our 3905 genomes. To balance the measure of similarity and dissimilarity, the occurrence for all observed features can be measured by Shannon Diversity Index[60]:

$$H' = -\sum_{i=1}^{N} p_i \ln p_i \quad (4)$$

Where p_i is the probability of features can be found in i genomes and N is total genome number in the database. With the specific length *k*, the number of observed kmers is O_k ($O_k \leq 4^k$). C_i kmers, can be found in *i* genomes ($1 \leq i \leq N$). The p_i can be calculated as $p_i = C_k / O_k$. For example, to calculate the Shannon Diversity Index of the K=9 dendrogram, the O_k = 262,144. We assume there are C_5 kmers that can be found in 5 genomes, which means any of these C_5 kmers exists in 5 genomes among the 3905 genomes. Here $p_5 = C_5 / 262144$ (i = 5). The Shannon Diversity Index can be calculated by adding values from p_1 to p_{3905}.

Tree Stability: Although 3-step process is applied to check the optimal feature length, it is still possible that inconsistent results can be obtains from three criteria. To strengthen the feasibility of our method, we use tree stability as an additional information to determine the optimal feature length. Tree stability is estimated by calculating the topology difference between trees at feature length k (k = 5, 6, 7, ……) and k + 1 using Robinson-Foulds distance [61], which is a metric to compare differences between two phylogenies. Therefore, when the Robinson-Foulds distances between tree at feature length k and k+1 decrease to a low value, it means the tree stability starts at this k point and tree topology does not change much as k increases. In our case, trees start to converge at k =9, so k = 9 has been chosen as the optimal feature length of the global dendrogram.

Evaluation of grouping uncertainty

The dendrogram (k=9) was evaluated for grouping uncertainty by viral family annotation, based on ICTV classification, using the statistical methods described by Huang [50]. Kruskal-Wallis one-way analysis of variance test was employed to evaluate the difference of the distance mean between within-groups and between-groups. Wilcoxon rank sum test was employed to evaluate the difference of distance mean between within-group and between-group for each group. The top

10 highest members of viral families which are [1] Siphoviridae (657 viruses), Geminiviridae (364 viruses), Myoviridae (307 viruses), Podoviridae (218 viruses), Papillomaviridae (125 viruses), Potyviridae (119 viruses), Parvoviridae (81 viruses), Picornaviridae (73 viruses), Flaviviridae (70 viruses) and Betaflexiviridae (66 viruses) were selected to perform the statistical analyses.

Acknowledgements

We gratefully thank Visanu Wanchai and Miraim Land for their technical assistance.

Author contribution Statement

QZ, SJ, ML collected and cleaned viral Refseq data set. QZ, SJ performed data analysis and draft the manuscript. IN, DU, SJ supervised QZ. IN designed, conceived and conduct the project. All authors discussed the results and implications and commented on the manuscript at all stages.

Competing financial interests

The authors declare no competing financial interests.

CHAPTER THREE
CONCLUSIONS

This thesis designed a 3-step comprehensive strategy for identifying the optimal length of K-mer in a viral phylogenomic analysis using genomic alignment-free method. This comprehensive strategy consists of three steps: 1) an individual genome perspective: CRE value to find the minimum feature length; 2) pairwise comparison perspective where ACF value among genomes is applied to determine the maximum feature length; 3) an all-genome comparison perspective where Shannon Diversity Index of all observed feature occurrences in genomes to find the optimal feature length between the minimum and the maximum. Also, tree stability information, which obtained from Robinson-Foulds distances, has been used as an assistant criterion to determine the optimal length K if results are not unique. By applying this strategy, we determined the optimal K-mer length (K=9) and reconstructed the dendrogram of 3905 completed viral RefSeq genomes in NCBI. This dendrogram gives a global view of classification in good agreement with the current viral taxonomy reported by ICTV and Baltimore classification. Additionally, statistical analysis was also done to test the grouping uncertainty.

LIST OF REFERENCES

1. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–1145 (2008).
2. Bentley, D. R. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16,** 545–552 (2006).
3. Hasman, H. *et al.* Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J. Clin. Microbiol.* **52,** 139–46 (2014).
4. Bao, Y. *et al.* National center for biotechnology information viral genomes project. *J. Virol.* **78,** 7291–8 (2004).
5. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33,** D501-4 (2005).
6. Brinkman, F. S. L. & Leipe, D. D. in *Bioinformatics: a practical guide to the analysis of genes and proteins* **2,** 349 (Wiley-Interscience. Nueva York, 2001).
7. Gao, Y. *et al.* Phylogenetic analysis of porcine epidemic diarrhea virus field strains prevailing recently in China. *Arch. Virol.* **158,** 711–715 (2013).
8. Jacques, M.-A. *et al.* Phylogenetic analysis and polyphasic characterization of Clavibacter michiganensis strains isolated from tomato seeds reveal that nonpathogenic strains are distinct from C. michiganensis subsp. michiganensis. *Appl. Environ. Microbiol.* **78,** 8388–402 (2012).
9. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41,** D377-86 (2013).
10. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236–40 (2014).
11. Casali, N. *et al.* Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat. Genet.* **46,** 279–86 (2014).
12. Morrison, D. A. *et al.* L. A. S. JOHNSON REVIEW No. 8. Multiple sequence

alignment for phylogenetic purposes. *Aust. Syst. Bot.* **19,** 479 (2006).

13. Soltis, D. E. & Soltis, P. S. Applying the Bootstrap in Phylogeny Reconstruction. *Stat. Sci.* **18,** 256–267 (2003).
14. Lapointe, F.-J., Kirsch, J. A. W. & Bleiweiss, R. Jackknifing of Weighted Trees: Validation of Phylogenies Reconstructed from Distance Matrices. *Mol. Phylogenet. Evol.* **3,** 256–267 (1994).
15. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **6,** 361–75 (2005).
16. Comin, M. & Verzotto, D. Alignment-free phylogeny of whole genomes using underlying subwords. *Algorithms Mol. Biol.* **7,** 34 (2012).
17. Horwege, S. *et al.* Spaced words and kmacs: fast alignment-free sequence comparison based on inexact word matches. *Nucleic Acids Res.* **42,** W7-11 (2014).
18. Leimeister, C.-A. & Morgenstern, B. Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics* **30,** 2000–8 (2014).
19. Huang, H. H. & Yu, C. Clustering DNA sequences using the out-of-place measure with reduced n-grams. *J. Theor. Biol.* **406,** 61–72 (2016).
20. Vinga, S. & Almeida, J. Alignment-free sequence comparison-a review. *Bioinformatics* **19,** 513–23 (2003).
21. Bonham-Carter, O., Steele, J. & Bastola, D. Alignment-free genetic sequence comparisons: A review of recent approaches by word analysis. *Brief. Bioinform.* **15,** 890–905 (2013).
22. Sims, G. E., Jun, S.-R., Wu, G. A. & Kim, S.-H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 2677–82 (2009).
23. Sims, G. E. & Kim, S.-H. Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). *Proc. Natl. Acad. Sci. U. S. A.* **108,** 8329–34 (2011).

24. Sims, G. E., Jun, S.-R., Wu, G. A. & Kim, S.-H. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 17077–82 (2009).
25. Wu, G. A., Jun, S.-R., Sims, G. E. & Kim, S.-H. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 12826–31 (2009).
26. Furuse, Y., Suzuki, A., Kamigaki, T. & Oshitani, H. Evolution of the M gene of the influenza A virus in different host species: large-scale sequence analysis. *Virol. J.* **6,** 67 (2009).
27. Shi, W. *et al.* Identification of novel inter-genotypic recombinants of human hepatitis B viruses by large-scale phylogenetic analysis. *Virology* **427,** 51–9 (2012).
28. Jun, S.-R., Sims, G. E., Wu, G. A. & Kim, S.-H. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 133–8 (2010).
29. Royer-Bertrand, B. & Rivolta, C. Whole genome sequencing as a means to assess pathogenic mutations in medical genetics and cancer. *Cell. Mol. Life Sci.* **72,** 1463–71 (2015).
30. Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15,** 749–63 (2014).
31. Wyres, K. L. *et al.* WGS Analysis and Interpretation in Clinical and Public Health Microbiology Laboratories: What Are the Requirements and How Do Existing Tools Compare? *Pathog. (Basel, Switzerland)* **3,** 437–58 (2014).
32. Chrystoja, C. C. & Diamandis, E. P. Whole genome sequencing as a diagnostic test: challenges and opportunities. *Clin. Chem.* **60,** 724–33 (2014).
33. Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L. & Nolan, G. P. Computational solutions to large-scale data management and analysis. *Nat.*

Rev. Genet. **11,** 647–57 (2010).

34. Braun, R. Systems analysis of high-throughput data. *Adv. Exp. Med. Biol.* **844,** 153–87 (2014).
35. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43,** D571-7 (2015).
36. Simmonds, P. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J. Gen. Virol.* **96,** 1193–206 (2015).
37. Adams, M. J., Hendrickson, R. C., Dempsey, D. M. & Lefkowitz, E. J. Tracking the changes in virus taxonomy. *Arch. Virol.* **160,** 1375–83 (2015).
38. Radoshitzky, S. R. *et al.* Past, present, and future of arenavirus taxonomy. *Arch. Virol.* **160,** 1851–74 (2015).
39. CALISHER, C. H. & MAHY, B. W. J. TAXONOMY: GET IT RIGHT OR LEAVE IT ALONE. *Am J Trop Med Hyg* **68,** 505–506 (2003).
40. Hannigan, G. D. *et al.* The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *MBio* **6,** e01578-15 (2015).
41. Skvortsov, T. *et al.* Metagenomic Characterisation of the Viral Community of Lough Neagh, the Largest Freshwater Lake in Ireland. *PLoS One* **11,** e0150361 (2016).
42. Seto, D., Chodosh, J., Brister, J. R. & Jones, M. S. Using the whole-genome sequence to characterize and name human adenoviruses. *J. Virol.* **85,** 5701–2 (2011).
43. Brown, J. K. *et al.* Revision of Begomovirus taxonomy based on pairwise sequence comparisons. *Arch. Virol.* **160,** 1593–619 (2015).
44. Ohno, T. *et al.* Usefulness and limitation of phylogenetic analysis for hepatitis C virus core region: application to isolates from Egyptian and Yemeni patients. *Arch. Virol.* **141,** 1101–1113 (1996).
45. Narechania, A., Chen, Z., DeSalle, R. & Burk, R. D. Phylogenetic

incongruence among oncogenic genital alpha human papillomaviruses. *J. Virol.* **79,** 15503–10 (2005).

46. Holmes, E. C. & Rambaut, A. Viral evolution and the emergence of SARS coronavirus. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **359,** 1059–65 (2004).
47. Wu, B. *et al.* Assessment of codivergence of mastreviruses with their plant hosts. *BMC Evol. Biol.* **8,** 335 (2008).
48. Rokas, A., Williams, B. L., King, N. & Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425,** 798–804 (2003).
49. Huang, H. H. *et al.* Global comparison of multiple-segmented viruses in 12-dimensional genome space. *Mol. Phylogenet. Evol.* **81,** 29–36 (2014).
50. Huang, H. H. An ensemble distance measure of k-mer and Natural Vector for the phylogenetic analysis of multiple-segmented viruses. *J. Theor. Biol.* **398,** 136–144 (2016).
51. Labonté, J. M. & Suttle, C. A. Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J.* **7,** 2169–77 (2013).
52. Tatusova, T. *et al.* Update on RefSeq microbial genomes resources. *Nucleic Acids Res.* **43,** D599-605 (2015).
53. Jun, S. R. *et al.* Ebolavirus comparative genomics. *FEMS Microbiol. Rev.* **39,** 764–778 (2015).
54. Pruitt, K., Brown, G., Tatusova, T. & Maglott, D. The Reference Sequence (RefSeq) Database. (2012).
55. Varghese, N. J. *et al.* Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* **43,** 6761–6771 (2015).
56. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27,** 764–70 (2011).
57. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3,** 217–223 (2012).
58. Sadovsky, M. G. Comparison of Real Frequencies of Strings vs. the

Expected Ones Reveals the Information Capacity of Macromoleculae. *J. Biol. Phys.* **29,** 23–38 (2003).

59. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **22,** 79–86 (1951).
60. Shannon, C. E. A mathematical theory of communication. *ACM SIGMOBILE Mob. Comput. Commun. Rev.* **5,** 3 (2001).
61. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53,** 131–147 (1981).

APPENDIX

Supplement Materials

Table S 1 Baltimore classification and ICTV Orders Information

Baltimore Classification	**counts**	**ICTV Order**	**counts**
dsDNA viruses, no RNA stage	1826	*Caudovirales*	1208
(+)ssRNA viruses	911	*Picornavirales*	157
ssDNA viruses	649	*Tymovirales*	141
dsRNA viruses	192	*Mononegavirales*	91
(-)ssRNA viruses	180	*Herpesvirales*	67
Retro-transcribing viruses	131	*Nidovirales*	58
Unclassified viruses	8	*Ligamenvirales*	12
Unclassified virophages	5	Unassigned or Unclassified	2171
Unassigned ssRNA viruses	3		

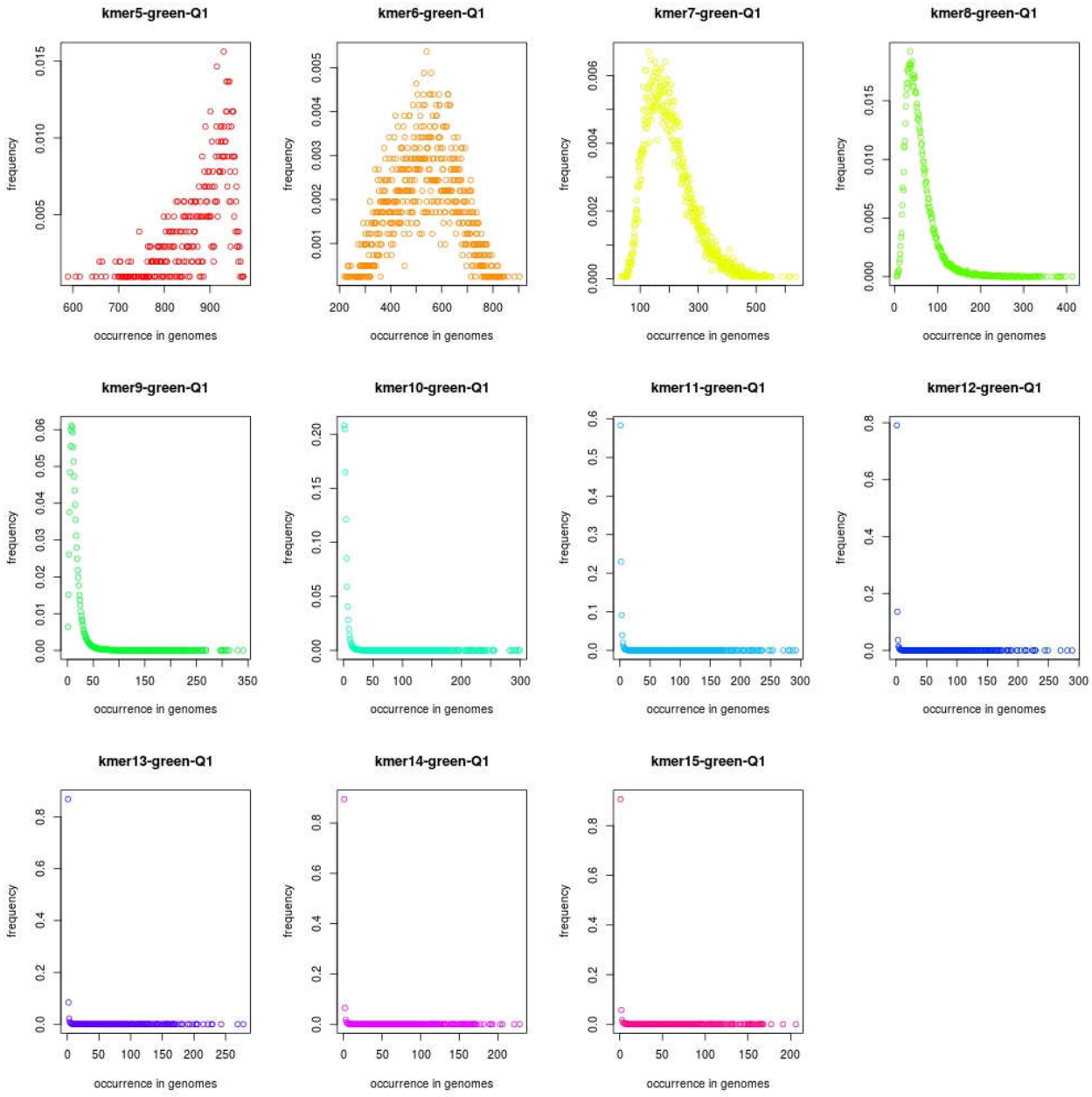

Figure 11 Distribution of feature occurrences in subgroup Q1 (size < 25%)

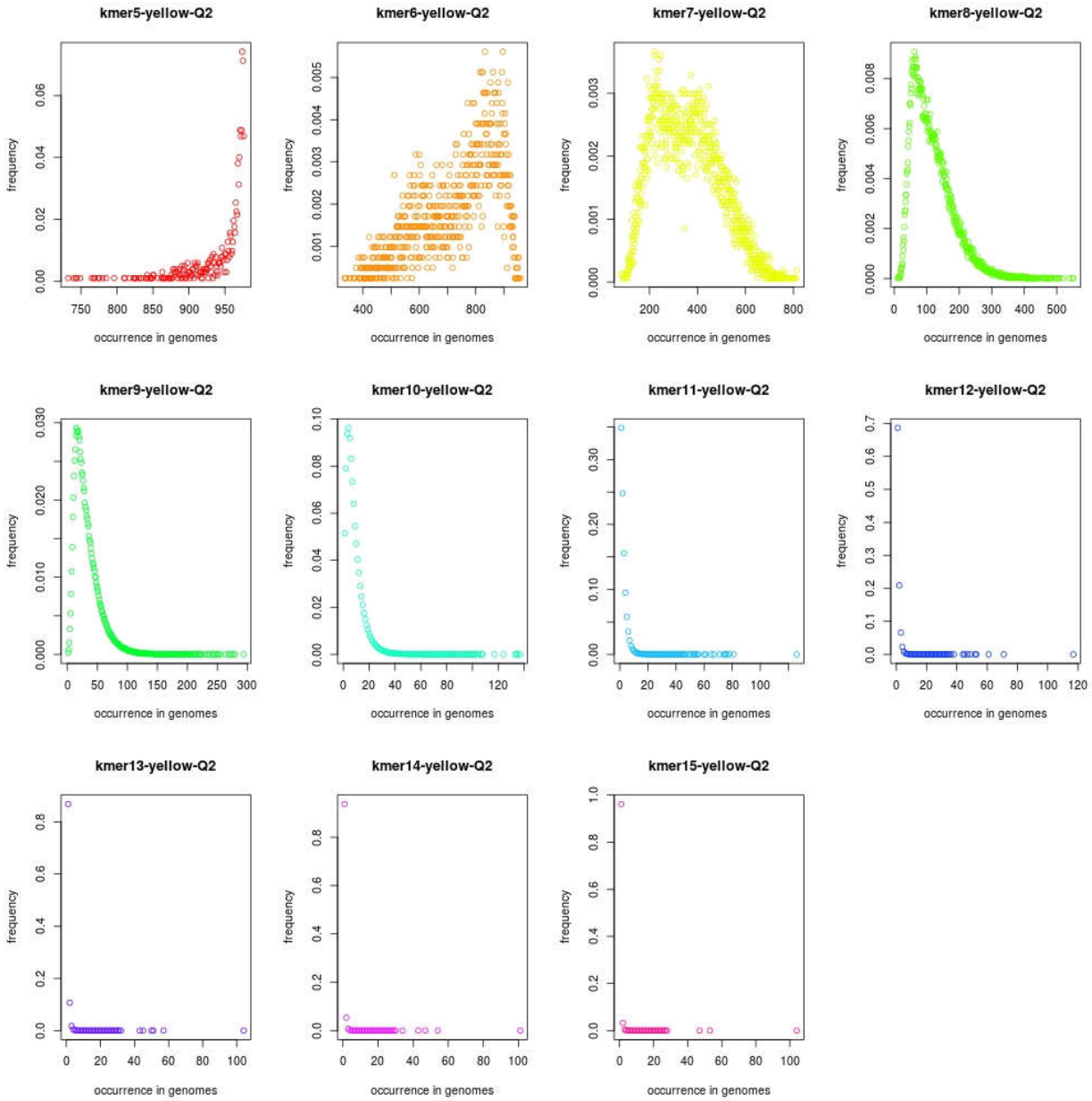

Figure 12 Distribution of feature occurrences in subgroup Q2 (25% < size < 50%)

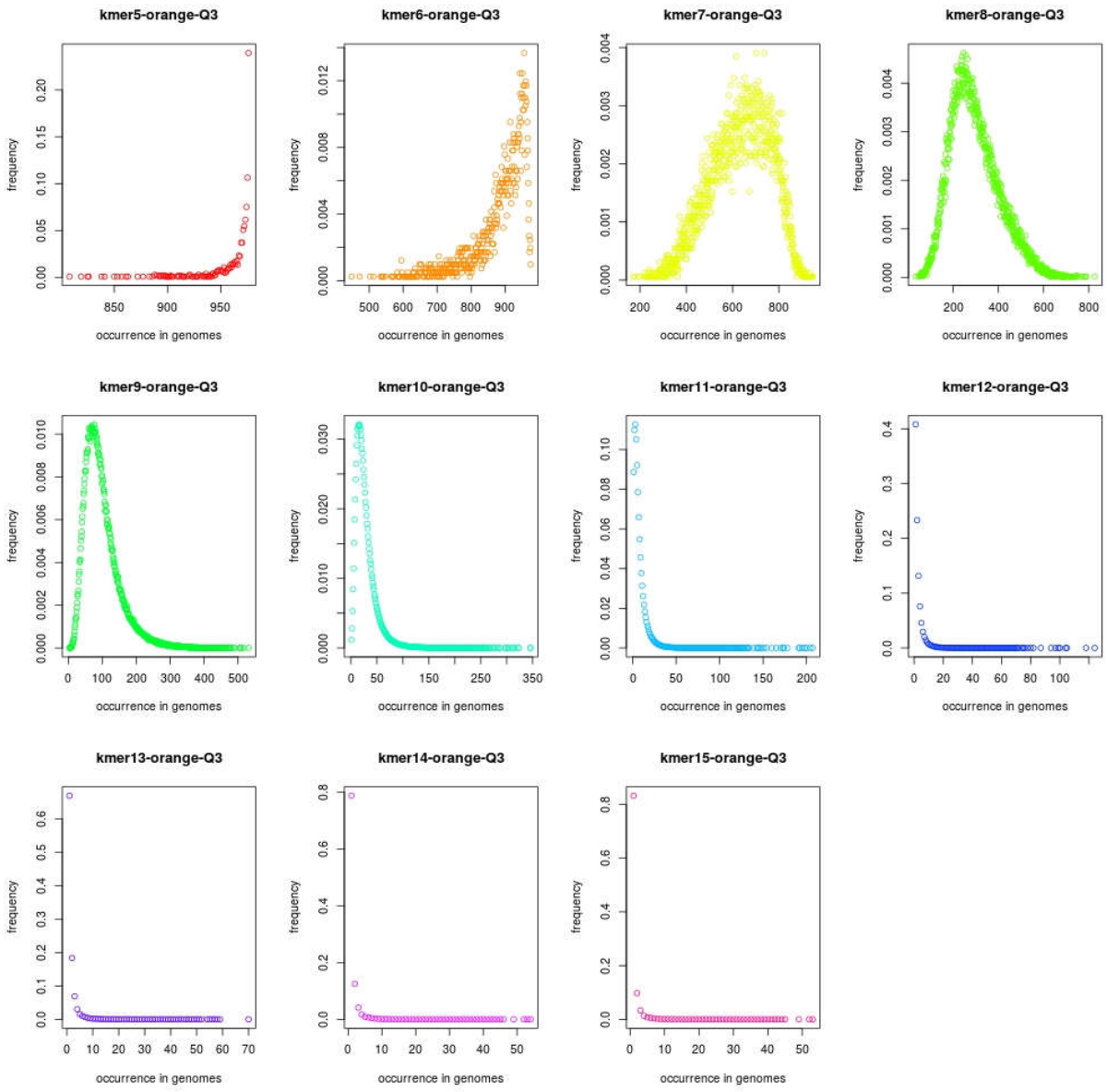

Figure 13 Distribution of feature occurrences in subgroup Q3 (50% < size < 75%)

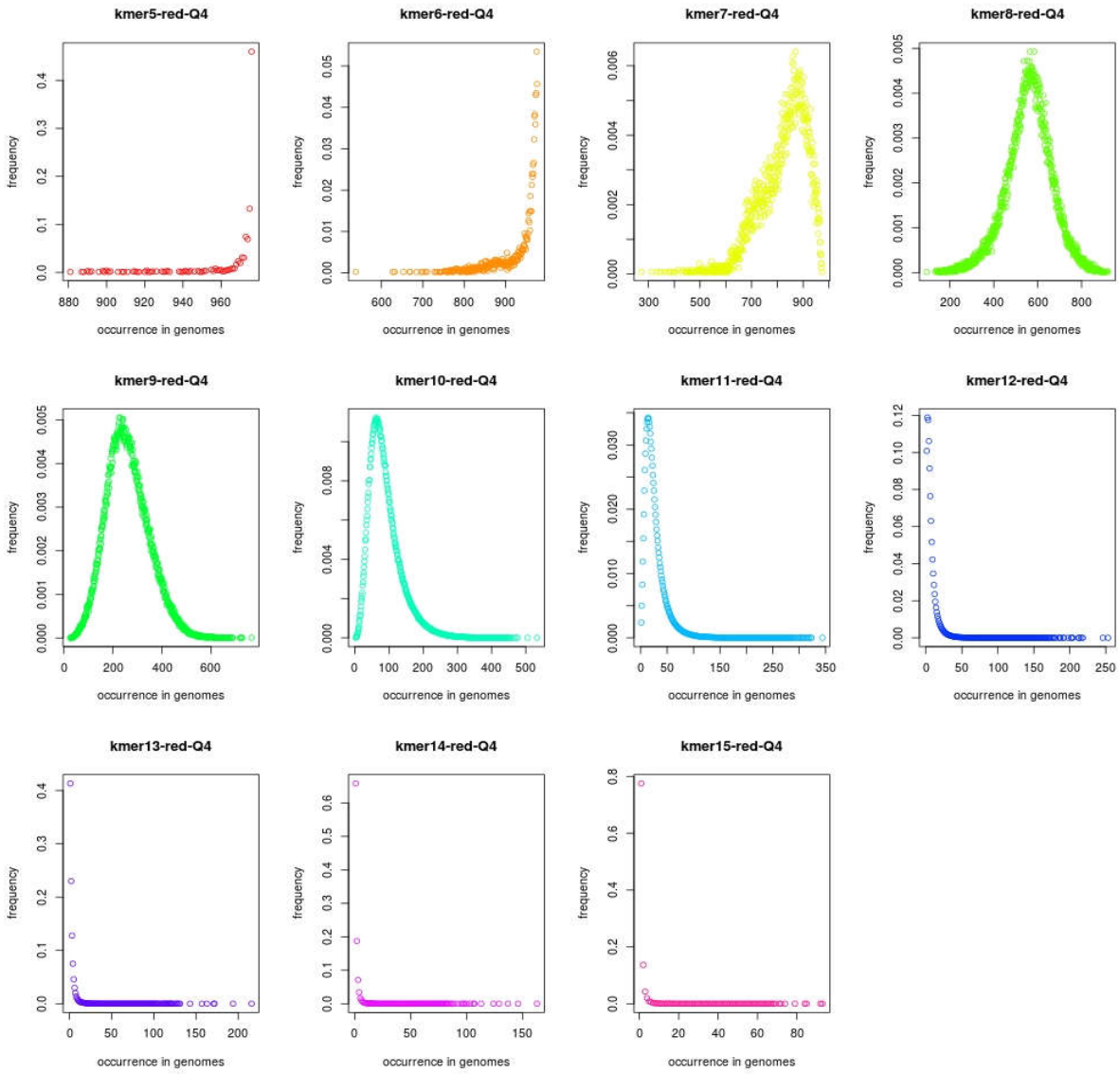

Figure 14 Distribution of feature occurrences in subgroup Q4 (size > 75%)

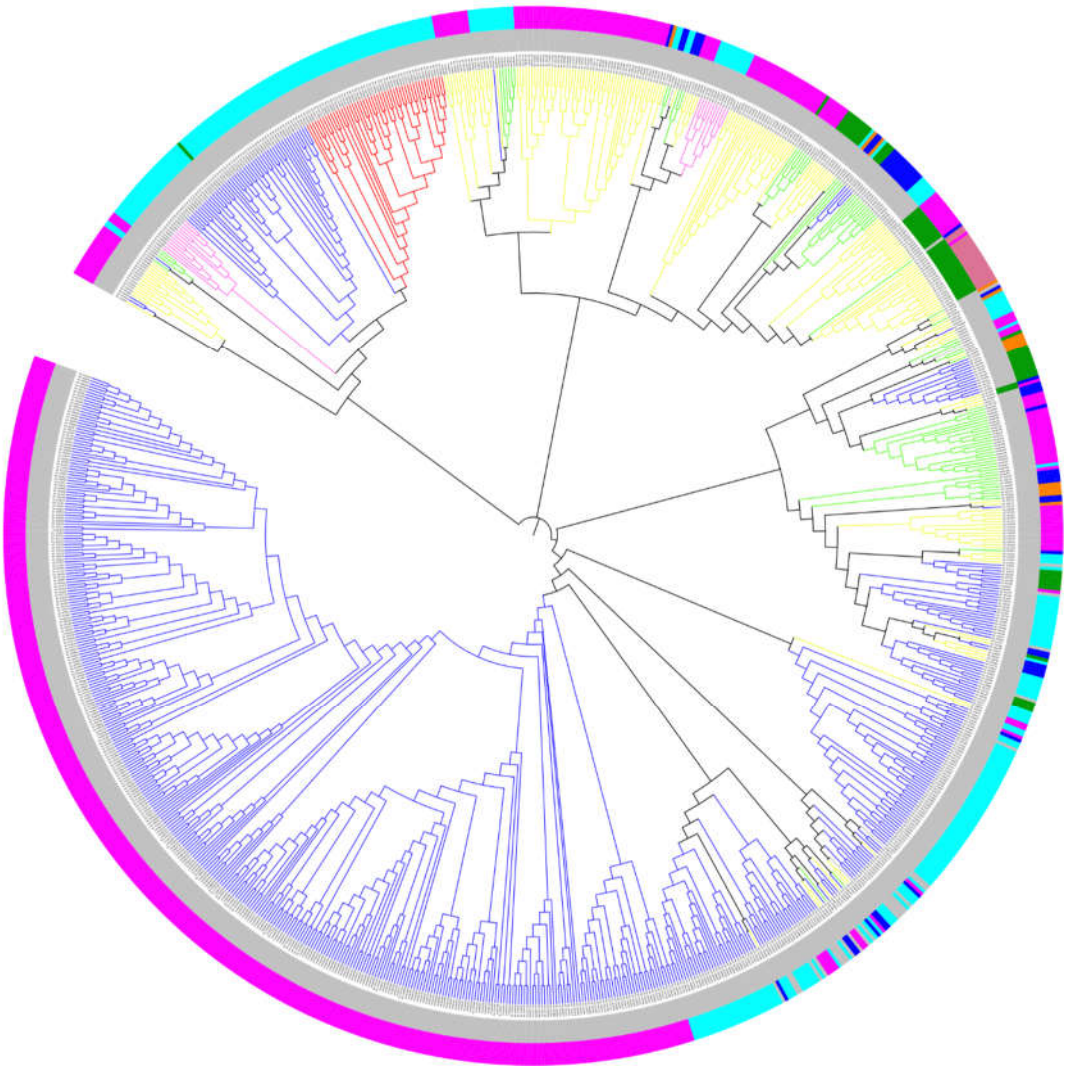

Figure 15 Dendrogram of 976 RefSeq viral genomes in subgroup Q1 (genome size < 25%), when k=9. The braches are colored by Baltimore Classifications. The circles, from inside to outside, are colored by different orders and hosts. [Color information: (A) Branch: Baltimore Classification; dsDNA, no RNA stage: red; dsRNA: green; Retro-transcribing viruses: pink; ssDNA: blue; ssRNA negative-strand: bright blue; ssRNA positive-strand: yellow. (B) From inside to outside, first circle: Order; Caudovirales: red; Herpesvirales: green; Ligamenvirales: blue; Mononegavirales: orange; Nidovirales: cyan: Picornavirales: pink; Tymovirales: dark green; unclassified: silver; (C) From inside to outside, second circle: Host; protest: orange; archaea: red; bacteria: dark green; fungi: blue; animal: cyan; animal and plants: pale violet red; plant: pink; environment or NA: silver.]

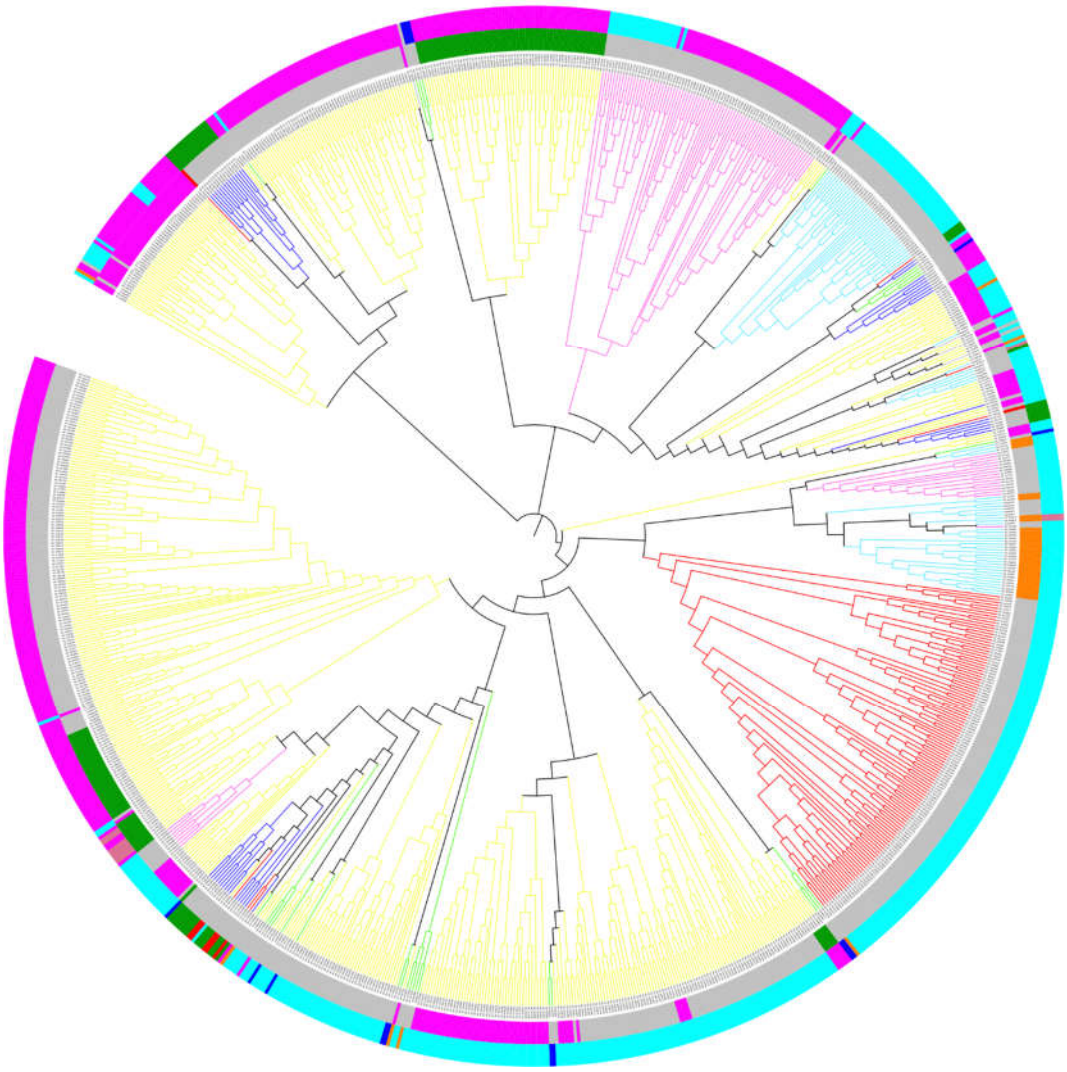

Figure 16 Dendrogram of 977 RefSeq viral genomes in subgroup Q2 (genome size: 25%-50%), when k=10. The braches are colored by Baltimore Classifications. The circles, from inside to outside, are colored by different orders and hosts. [Color information: (A) Branch: Baltimore Classification; dsDNA, no RNA stage: red; dsRNA: green; Retro-transcribing viruses: pink; ssDNA: blue; ssRNA negative-strand: bright blue; ssRNA positive-strand: yellow. (B) From inside to outside, first circle: Order; Caudovirales: red; Herpesvirales: green; Ligamenvirales: blue; Mononegavirales: orange; Nidovirales: cyan: Picornavirales: pink; Tymovirales: dark green; unclassified: silver; (C) From inside to outside, second circle: Host; protest: orange; archaea: red; bacteria: dark green; fungi: blue; animal: cyan; animal and plants: pale violet red; plant: pink; environment or NA: silver.]

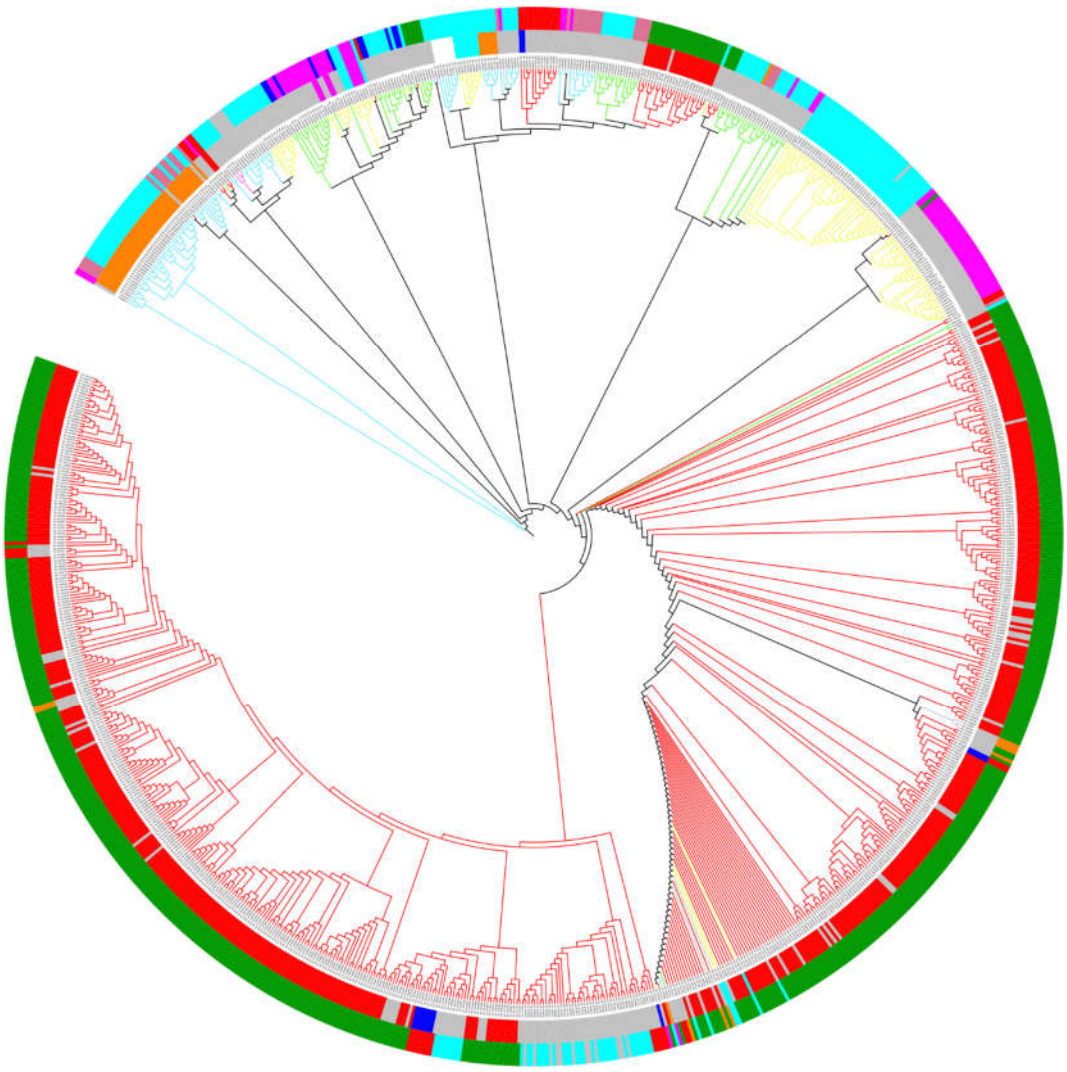

Figure 17 Dendrogram of 977 RefSeq viral genomes in subgroup Q3 (genome size: 50%-75%), when k=11. The braches are colored by Baltimore Classifications. The circles, from inside to outside, are colored by different orders and hosts. [Color information: (A) Branch: Baltimore Classification; dsDNA, no RNA stage: red; dsRNA: green; Retro-transcribing viruses: pink; ssDNA: blue; ssRNA negative-strand: bright blue; ssRNA positive-strand: yellow. (B) From inside to outside, first circle: Order; Caudovirales: red; Herpesvirales: green; Ligamenvirales: blue; Mononegavirales: orange; Nidovirales: cyan: Picornavirales: pink; Tymovirales: dark green; unclassified: silver; (C) From inside to outside, second circle: Host; protest: orange; archaea: red; bacteria: dark green; fungi: blue; animal: cyan; animal and plants: pale violet red; plant: pink; environment or NA: silver.]

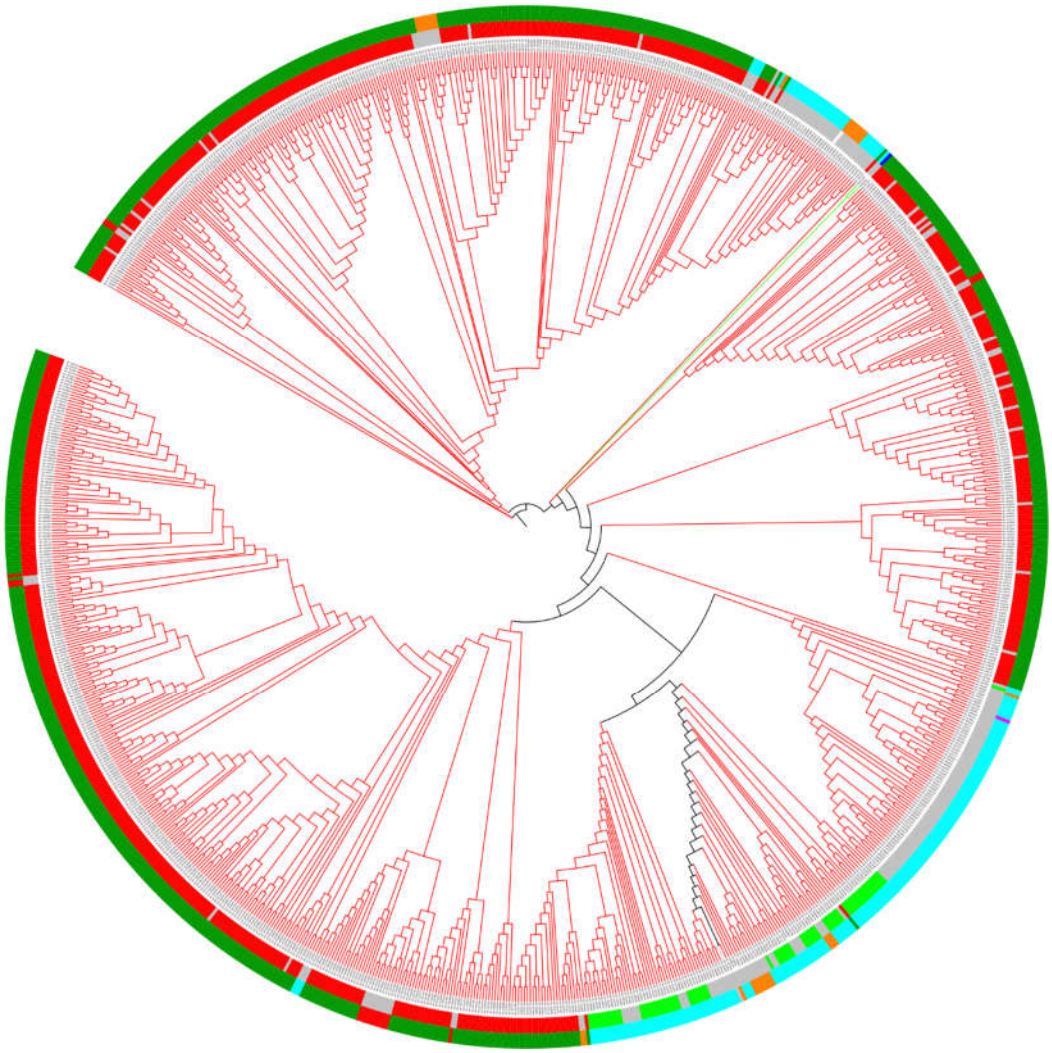

Figure 18 Dendrogram of 977 RefSeq viral genomes in subgroup Q4 (genome size: >75%), when k=12. The braches are colored by Baltimore Classifications. The circles, from inside to outside, are colored by different orders and hosts. [Color information: (A) Branch: Baltimore Classification; dsDNA, no RNA stage: red; dsRNA: green; Retro-transcribing viruses: pink; ssDNA: blue; ssRNA negative-strand: bright blue; ssRNA positive-strand: yellow. (B) From inside to outside, first circle: Order; Caudovirales: red; Herpesvirales: green; Ligamenvirales: blue; Mononegavirales: orange; Nidovirales: cyan: Picornavirales: pink; Tymovirales: dark green; unclassified: silver; (C) From inside to outside, second circle: Host; protest: orange; archaea: red; bacteria: dark green; fungi: blue; animal: cyan; animal and plants: pale violet red; plant: pink; environment or NA: silver.]

Table S 2 Wilcoxon rank sum test result of the top 10 highest members of viral family.

	Siphoviridae	Geminiviridae	Myoviridae	Podoviridae	Papillomaviridae	Potyviridae	Parvoviridae	Picornaviridae	Flaviviridae	Betaflexiviridae
Siphoviridae vs. Geminiviridae	< 2.2 E-16	< 2.2 E-16								
Siphoviridae vs. Myoviridae	< 2.2 E-16		< 2.2 E-16							
Siphoviridae vs. Podoviridae	< 2.2 E-16			< 2.2 E-16						
Siphoviridae vs. Papillomaviridae	< 2.2 E-16				< 2.2 E-16					
Siphoviridae vs. Potyviridae	< 2.2 E-16					< 2.2 E-16				
Siphoviridae vs. Parvoviridae	< 2.2 E-16						< 2.2 E-16			
Siphoviridae vs. Picornaviridae	< 2.2 E-16							< 2.2 E-16		
Siphoviridae vs. Flaviviridae	< 2.2 E-16								< 2.2 E-16	
Siphoviridae vs. Betaflexiviridae	< 2.2 E-16									< 2.2 E-16
Geminiviridae vs. Myoviridae		< 2.2 E-16	< 2.2 E-16							
Geminiviridae vs. Podoviridae		< 2.2 E-16		< 2.2 E-16						
Geminiviridae vs. Papillomaviridae		< 2.2 E-16			< 2.2 E-16					
Geminiviridae vs. Potyviridae		< 2.2 E-16				< 2.2 E-16				
Geminiviridae vs. Parvoviridae		< 2.2 E-16					< 2.2 E-16			
Geminiviridae vs. Picornaviridae		< 2.2 E-16						< 2.2 E-16		
Geminiviridae vs. Flaviviridae		< 2.2 E-16							< 2.2 E-16	
Geminiviridae vs. Betaflexiviridae		< 2.2 E-16								< 2.2 E-16
Myoviridae vs. Podoviridae			< 2.2 E-16	< 2.2 E-16						
Myoviridae vs. Papillomaviridae			< 2.2 E-16		< 2.2 E-16					
Myoviridae vs. Potyviridae			< 2.2 E-16			< 2.2 E-16				
Myoviridae vs. Parvoviridae			< 2.2 E-16				< 2.2 E-16			
Myoviridae vs. Picornaviridae			< 2.2 E-16					< 2.2 E-16		
Myoviridae vs. Flaviviridae			< 2.2 E-16						< 2.2 E-16	
Myoviridae vs. Betaflexiviridae			< 2.2 E-16							< 2.2 E-16
Podoviridae vs. Papillomaviridae				< 2.2 E-16	< 2.2 E-16					
Podoviridae vs. Potyviridae				< 2.2 E-16		< 2.2 E-16				
Podoviridae vs. Parvoviridae				< 2.2 E-16			< 2.2 E-16			
Podoviridae vs. Picornaviridae				< 2.2 E-16				< 2.2 E-16		
Podoviridae vs. Flaviviridae				< 2.2 E-16					< 2.2 E-16	
Podoviridae vs. Betaflexiviridae				< 2.2 E-16						< 2.2 E-16
Papillomaviridae vs. Potyviridae					< 2.2 E-16	< 2.2 E-16				
Papillomaviridae vs. Parvoviridae					< 2.2 E-16		< 2.2 E-16			
Papillomaviridae vs. Picornaviridae					< 2.2 E-16			< 2.2 E-16		
Papillomaviridae vs. Flaviviridae					< 2.2 E-16				< 2.2 E-16	
Papillomaviridae vs. Betaflexiviridae					< 2.2 E-16					< 2.2 E-16
Potyviridae vs. Parvoviridae						< 2.2 E-16	< 2.2 E-16			
Potyviridae vs. Picornaviridae						< 2.2 E-16		0.249381472		
Potyviridae vs. Flaviviridae						< 2.2 E-16			< 2.2 E-16	
Potyviridae vs. Betaflexiviridae						< 2.2 E-16				< 2.2 E-16
Parvoviridae vs. Picornaviridae							0.40400024	< 2.2 E-16		
Parvoviridae vs. Flaviviridae							< 2.2 E-16		< 2.2 E-16	
Parvoviridae vs. Betaflexiviridae							9.69E-14			< 2.2 E-16
Picornaviridae vs. Flaviviridae								0.017555005	< 2.2 E-16	
Picornaviridae vs. Betaflexiviridae								< 2.2 E-16		< 2.2 E-16
Flaviviridae vs. Betaflexiviridae									< 2.2 E-16	< 2.2 E-16

VITA

Qian Zhang was born in Jinan, China, where is a beautiful city well known for sweet springs and nice people. After graduating from Shandong Experimental High School in 2005, she attended Shandong University in the same city. Upon graduating with her double B.S. degrees in Preventive Medicine and Information & Computational Science in 2010, Qian obtained her M.S in Biostatistics from the same University. She enrolled at University of Texas, Houston in 2013 and transferred to University of Tennessee, Knoxville in 2014, because of family. She joined Comparative Genomics Group in ORNL in March 2014 under the guidance of Dr. Dave Ussery.