Linear Algebraic and Boolean Analysis of Genomic Topology

Shengyun Peng[1], Cai Huang[1], Peter Audano[1], and Fredrik O. Vannberg[1]

[1]Parker H. Petit Institute for Bioengineering and Biosciences, School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

ABSTRACT

To date the comparison of genomic DNA sequences[12] have routinely utilized shorter conserved regions for comparative genomics. Current phylogenetic analysis therefore can create divergent results based on which genetic loci are utilized for this analysis[13]. Sequence similarity is also commonly determined by first carrying out gap penalty pairwise alignments[2][15][10] for a set of sequences, and the similarity is quantified based upon this alignment. This approach, however, has limitations and is primarily utilized to compare relatively conserved sequences and is also dependent upon the algorithm utilized to create the alignment. From first principles we here develop a framework for encoding nucleic acid sequences into fixed length sections (k-mers)[4][7] which we utilize to create an invariant pairwise distance of total information content of over $\sim 4,000$ bacteria[6] and viruses, revealing important cryptic relationships not previously reported. To date k-mer strategies have been utilized by nearly all short read DNA alignment strategies, including de Bruijn-type[5] combinatorial mathematics and Burrows-Wheeler transform[1] (BWT) algorithm to efficiently align sequences. Our framework adds another layer of abstraction to this field by

creating an invariant encoding heuristic for k-mers that allows for the efficient analysis and computation of genome similarities. This invariant encoding framework is computable by linear algebra and also diverse Boolean, logic and bit operations of discrete mathematics. In addition to orthogonal transformation of this dataset we also show pairwise differential occupancy (which we term $\Delta\Gamma_k$) of these data structures for hierarchical clustering of genomic sequence. To our knowledge we define the first global clustering of entire bacterial and viral genomes, which we define here as genomic topology, which led to the discovery of a number of novel cryptic genomic associations between specific viral phages[14] and bacteria. This formalization of genomic relationships yields an invariant pairwise differential occupancy ($\Delta\Gamma_k$) metric between all species based upon global information content of each genome, irrespective of Kingdom, and finally allows a universal complete encoding for species that was lacking in previous approaches.

The exponential growth in sequencing information of organisms of all Kingdoms has greatly increased our understanding of the diversity of genomes and genomic complexity. In order to help formalize and explore genomic encoding we have created a sparse k-mer encoding system wherein genomic data is stored in an identical format that allows for global computations across all genomic data, regardless of homology. To date there has been a distinct lack of fundamental organization of genomic data, with primarily ad hoc analysis of genome similarity using subsets of evolutionarily conserved genes.

We created a universal encoding system to store k-mer data in sparse linear vector form as well as an equivalent $n\times$n square sparse matrix form (Supporting Online Material Fig. 1). Briefly, our system utilizes a sparse vector that contains 4^k elements, arrayed in ascending order by the binary representation of the k-mer, where $\{A = 00, C = 01, G = 10, T = 11\}$. For these sparse vectors we define tau (Γ_k) by the number of non-zero elements divided by 4^k elements , where k represents the k-mer length. We routinely also transform this binary k-mer representation into a decimal number as seen in Fig. 1A.In order to optimize this k-mer analysis we authored KAnalyze [3], which is a fast and extensible k-mer suite with APIs specific for this process.

From these first principles, we sought ways to globally analyze these sparse linear vectors. Us-

ing standard linear algebra approaches it is possible to complete orthogonal transformations of this data (Fig. 1B). In one manifestation we carry out hierarchical principle component analysis (hPCA) [11] (see Methods) of the combined genomic sparse vectors of multiple species, allowing for distance calculations in n-dimensional space, with the closest shared information content giving the smallest distance using three (or more) components. These first three components can then be used to create a three dimensional plot of the distances between the sparse vectors for each genome sequence (Fig. 1B). In a similar manner, it is possible to carry out global Boolean analysis on these sparse vectors in a pairwise fashion. By arraying the sparse vectors in an invariant format, with full encoding, it is trivial to carry out pairwise discrete functions for any logic gate functions including AND, OR, XOR as seen in Figure 1B. One simple metric is the resultant value of the XOR Boolean operator between two sparse vectors, which we here define as delta tau ($\Delta\Gamma_k$), or differential occupancy, where k defines the k-mer length (see Methods).

We start by carrying out the hPCA analysis of 578 bacterial genomes in the NCBI bacterial database. And we plot log transformed pairwise distance of these bacterial genomes base on hPCA. By z-transforming the pairwise $-\log(\Delta\Gamma_k)$ data we see a desirable and expected association between species-based similarities between k-mer encoding, and define $z >= 2$ as closely related species. We formalize this association and speculate that this relationship follows a power law association (Fig. 2A), which appear to be inherent in natural genomic topologies of biological systems. We here show punctuate peaks (Fig. 2B) of closely related species within a sea of random or near random associations. Similarly, in calculating genomic topologies of all pairwise $-\log(\Delta\Gamma_k)$ species (Fig. 2C, D) will be defined as those that reside within the k-mer topology within the exponential function of arrayed pairwise relationships, indicating a "null" and "positive"space. For these closely related species we see R^2 value of 0.87 between the hPCA and Boolean analysis (Fig. 2E), for the species with a pairwise relationship of $z >= 2$. We show the density plot of all pairwise comparisons (extended Fig. 1, Supporting Online Material Table 1). Using this $\Delta\Gamma_k$ we apply hierarchical clustering and obtain expected clustering of these bacterial genomes (Fig. 3).

This technique enables the comparison of highly divergent genomic sequences and we utilize

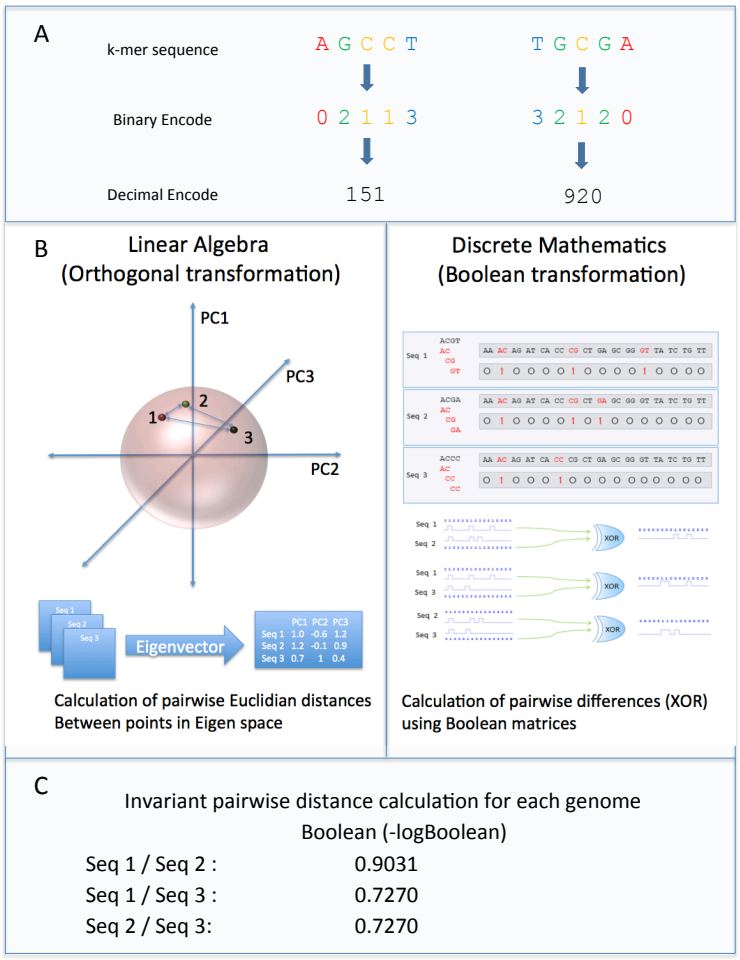

Figure 1: Universal K-mer encoding and computation. A) Universal encoding defined using the Quaternary-Decimal encoding system. B) Linear algebra forms analyzes a variety of orthogonal transformation / eigenvector calculations. Discrete mathematics enables Boolean analysis using the full complement discrete functions including AND, OR, and XOR.

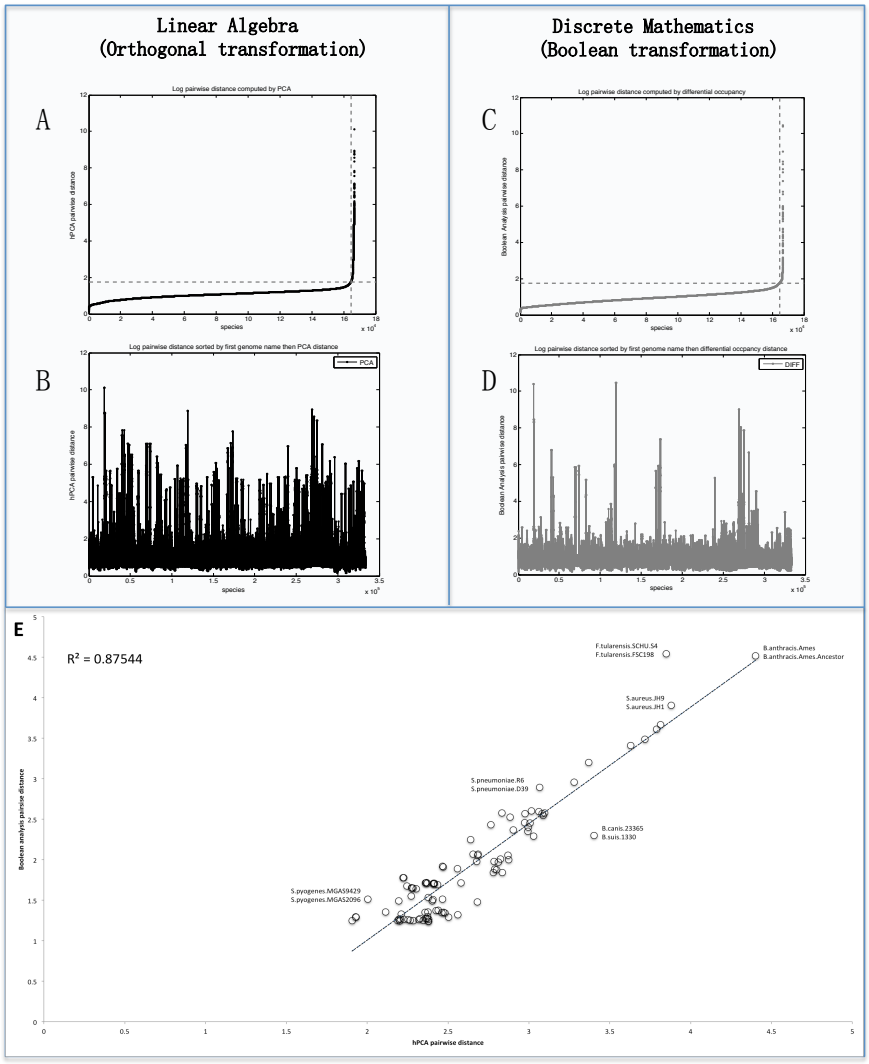

Figure 2: Pairwise distance. A) Sorted Log pairwise distance computed by hPCA; B) Unsorted Log pairwise distance computed by hPCA; C) Sorted Log pairwise distance computed by Boolean Analysis; D) Unsorted Log pairwise distance computed by Boolean Analysis; E) Pairwise distance from hPCA and Boolean analysis relationship.

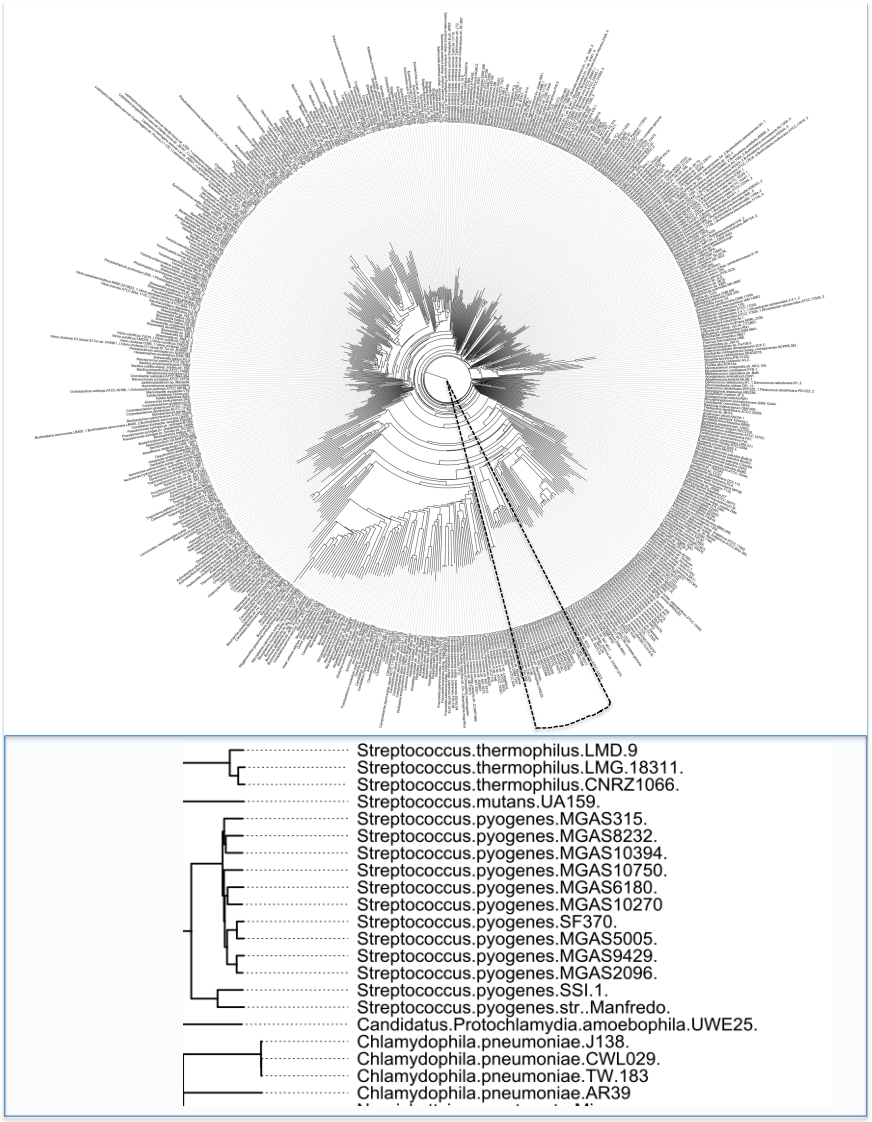

Figure 3: Hierarchical clustering. Zoom in part of the hierarchical clustering generated by differential occupancy $completetreeinSupportingOnlineMaterial, 578_10mer_diff_tree$.

these heuristics to perform hPCA clustering of bacterial and viral sequences (extended Fig. 2). In addition to finding a number of known relationships between bacteria and viruses, we also found evidence of previously cryptic relationships between these bacterial and viral genomes. From linear algebra genomic topology analysis we find examples that make sense in terms of likely horizontal gene transfer such as a close link between *Sulfolobus virus STSV1* and *Sulfolobus islandicus M*; *Sulfolobus virus STSV1* and *Sulfolobus solfataricus*; *Sulfolobus virus STSV1* and *Sulfolobus acidocaldarius*; *Sulfolobus virus STSV1* and *Sulfolobus tokodaii* [16]; *Salmonella phage ST64B* and *Escherichia coli ED1a*; *Klebsiella phage phiKO2* and *Escherichia coli O26*. However, this analysis also revealed several unexpected relationships between *Klebsiella phage phiKO2* and *Enterobacter cloacae EcWSU1*; *Enterobacteria phage N15* and *Pectobacterium atrosepticum SCRI1043*, among others. These relationships highly suggestive of shared genomic content through horizontal gene transfer, and provide a new perspective to examine these putative new relationships between of bacteria and viruses. We prepared a K-ermized data set of 2569 bacteria and 1754 virus sequences and hPCA (Supporting Online Material) analysis result for further study.

We demonstrate that our approach can analyze an ancient genome such as *Acidianus bottle shaped virus*, which infect archaea [9], in the context of other viruses. By using grep function, we output pairwise relationship between Acidianus bottle shaped virus and all other species in our database, then we sorted by the hPCA distance. We find out close virus spices have stronger relationship than other random picked and bacteria (Fig. 4). Also, we can to study the relationship between this virus and other bacteria. So we show a circus plot of several bacteria, which have low hPCA distance with the virus, and five random picked *Methylobacterium* family bacteria (extended Fig. 3). We can see the *Methylobacterium* family are far away from our virus, because they are less related than the *Vulcanisaeta moutnovskia* [8] to the virus, and it is an Archaea, which might be infected by *Acidianus bottle shaped virus*. Besides, *Caldivirga maquilingensis, Ignisphaera aggregans, Vulcanisaeta distribute, Desulfurococcus fermentans,* are all Archaea. This suggests that there is relationship between virus and infected bacteria on genome perspective. For another example, we can study on an unclassified sequence, and suggest classification. We pick an unclassified virus,

Pyrococcus abyssi virus 1. By using *grep* function, we output relationship between this virus and all other species in our database. After sorting the output by hPCA distance, we study the top 12 species (log-transformed hPCA distance higher than 2.5) and their lineage (extended table 1). From the table, we can see that most of these known species are belong to *Viruses; dsDNA viruses, no RNA stage; Caudovirales* family. So we can positively suggest that Pyrococcus abyssi virus 1 can be classified to *Viruses; dsDNA viruses, no RNA stage; Caudovirales*.

In the past it was not possible to adequately compare complete genomic content between two species due to the inability to align dissimilar sequence. Therefore although claims of genomic topology creating certain maxima and trough minima, this was not possible to properly test. Using our analysis, it does indeed appear that there is a specific power law effect that describes the overall genomic topology between species, with a sharp exponential decay in genomic topology between two different species. This, again, has been suggested but we feel that our result is one of the clearest attempts to show such a genomic topology.

We demonstrate the ability to cluster within species from first principles using genomic topology calculations without additional levels of abstraction. We show that optimized discrete calculations on global k-mer space qualitatively work as well as traditional linear algebra-based orthogonal transformations eigen value-based principle component analysis. Although laboratories can and will use analysis of higher levels of abstraction such as the case of PCA and multiple dimensional scaling, it is our opinion that it is easier to intuit intellectually discrete calculations of these higher order matrices verses taking into account the eigen vector space. An important distinction, which is always troubling with PCA and MDS type of analysis is that the origin of the variation is lost in the calculation, with the inability to determine the origin of the variation. Discrete systems do not face such restrictions, and our system creates the ability to real time cluster and reveals the origin of the variation at the same time, creating a use case that is relevant to pharmacological development. Optimizing our k-mer size to allow for the global creation of siRNA sequences against a known and novel virus we demonstrate that our k-mer system does well to not only process this but also be able to compare sequences in real time with approach we present. The simplicity of

this system, along with the considerable advancements that can be made to further innovate from this adds a great deal we believe to the toolkit of mathematics for biologists and we hope that brings into sharper focus quantitative biology principles moving ahead. It is not lost on us that a similar encoding system using k-mer can also be carried out for amino acid sequences encoded within organisms, and the parity between such genomic and proteomic data will also be useful in comparative analysis of species. Indeed philosophically this brings to the fore arguments for such numerical taxonomy approaches whereby additional data streams can be added into such analysis, allowing for quantitative assessments of species, including morphological, developmental and even neuronal encoding. These to date can be seen as ad hoc, but formalization approaches abound, such as the National Institutes of Health Brain Mapping Initiative, which can allow for formalization of such systems. We present some preliminary analysis of this, but further more sophisticated analysis of genome, proteomic and other feature space is possible using a similar model of creating invariant data structures that allow for like for like comparisons of data sets. It should be noted that this is not closed to inter-species analysis, indeed use cases abound that make intra-species analysis informative. Examples include analysis of virulent E. coli strains, or viral outbreaks scenarios. Our laboratory is working closely with multiple Centers for Disease Control branches to implement such analysis to allow for near complete automation of such pairwise $\Delta\Gamma_k$ analytics during outbreak scenarios.

Methods

We provide a small data set to demonstrate the method we used to calculate the pairwise distance. The test data set is random picked 4 kmerized virus files with k length equal to 8, and stored as *csv* file. Each column is a sparse vector and represents one virus. sequence.

hPCA The hPCA analysis is base on PCA. So we perform a PCA calculation among the normalized dataset and use only the main components to calculate the distance between viruses. Here

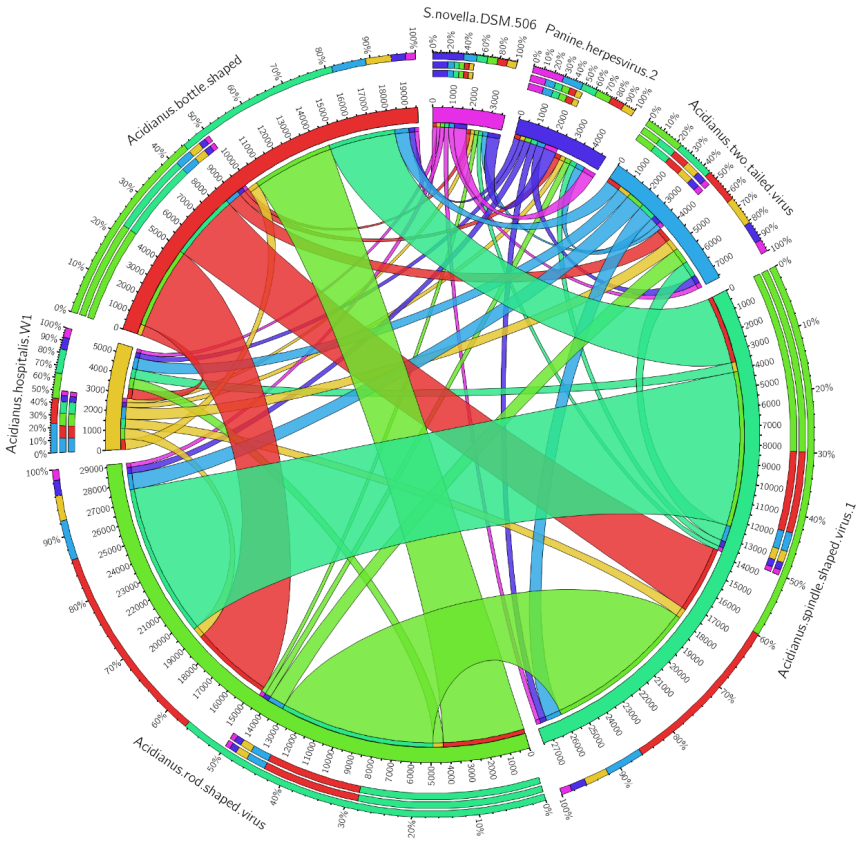

Figure 4: Circus plot of five Acidianus family virus, one random herpesvirus and one random Starkeya.novella bacteria. The pairwise relationship (extended table 3) are present by 10/(hPCA distance), so the thicker the connect line is, the closer they are.

we provide a R function to calculate the hPCA distance on kmerized data set. For normalization we subtract each column of data by its mean and divided by its stand deviation, and the distance funtion we uesd is *euclidean* distance.

Boolean Analysis We define the term differential occupancy as the XOR function between two sparse vectors that derives resultant sparse vector followed by summing the number of non-zero elements then dividing by 4^k. The more similar two genomes the smaller the number of resultant non-zero elements after the XOR calculation, resulting in a number that approaches zero for the same species. Herer is the matlab code for iterate the test data set and calculate the differential occupancy.

The pairwise distance reault shows in Table **??**.Each number represents the pairwise diffrential occupancy of two virus, and the smaller the number is, the closer they are.

References

[1] Adjeroh, D., Zhang, Y., Mukherjee, A., Powell, M., and Bell, T. (2002). Dna sequence compression using the burrows-wheeler transform. *Proc IEEE Comput Soc Bioinform Conf*, **1**, 303–13.

[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**(3), 403–10.

[3] Audano, P. and Vannberg, F. (2014). Kanalyze: A fast versatile pipelined k-mer toolkit. *Bioinformatics*.

[4] Chor, B., Horn, D., Goldman, N., Levy, Y., and Massingham, T. (2009). Genomic dna k-mer spectra: models and modalities. *Genome Biol*, **10**(10), R108.

[5] Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). How to apply de bruijn graphs to genome assembly. *Nat Biotechnol*, **29**(11), 987–91.

[6] Doolittle, R. F. (2002). Biodiversity: microbial genomes multiply. *Nature*, **416**(6882), 697–700.

[7] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nat Biotechnol*, **29**(7), 644–52.

[8] Gumerov, V. M., Mardanov, A. V., Beletsky, A. V., Prokofeva, M. I., Bonch-Osmolovskaya, E. A., Ravin, N. V., and Skryabin, K. G. (2011). Complete genome sequence of "vulcanisaeta moutnovskia" strain 768-28, a novel member of the hyperthermophilic crenarchaeal genus vulcanisaeta. *J Bacteriol*, **193**(9), 2355–6.

[9] Häring, M., Rachel, R., Peng, X., Garrett, R. A., and Prangishvili, D. (2005). Viral diversity in hot springs of pozzuoli, italy, and characterization of a unique archaeal virus, acidianus bottle-shaped virus, from a new family, the ampullaviridae. *J Virol*, **79**(15), 9904–11.

[10] Morgenstern, B., Dress, A., and Werner, T. (1996). Multiple dna and protein sequence alignment based on segment-to-segment comparison. *Proc Natl Acad Sci U S A*, **93**(22), 12098–103.

[11] Peterson, L. E. (2002). Clusfavor 5.0: hierarchical cluster and principal-component analysis of microarray-based transcriptional profiles. *Genome Biol*, **3**(7), SOFTWARE0002.

[12] Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **33**(Database issue), D501–4.

[13] Salichos, L. and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, **497**(7449), 327–31.

[14] Sampson, T., Broussard, G. W., Marinelli, L. J., Jacobs-Sera, D., Ray, M., Ko, C.-C., Russell, D., Hendrix, R. W., and Hatfull, G. F. (2009). Mycobacteriophages bps, angel and halo: comparative genomics reveals a novel class of ultra-small mobile genetic elements. *Microbiology*, **155**(Pt 9), 2962–77.

[15] Vingron, M. and Waterman, M. S. (1994). Sequence alignment and penalty choice. review of concepts, case studies and implications. *J Mol Biol*, **235**(1), 1–12.

[16] Xiang, X., Chen, L., Huang, X., Luo, Y., She, Q., and Huang, L. (2005). Sulfolobus tengchongensis spindle-shaped virus stsv1: virus-host interactions and genomic features. *J Virol*, **79**(14), 8677–86.