viruses

MDPI

Article

CoreGenes5.0: An Updated User-Friendly Webserver for the Determination of Core Genes from Sets of Viral and Bacterial Genomes

Patrick Davis [1], Donald Seto [2,*] and Padmanabhan Mahadevan [1,*]

[1] Department of Biology, The University of Tampa, Tampa, FL 33606, USA
[2] Department of Systems Biology, George Mason University, Manassas, VA 20110, USA
* Correspondence: dseto@gmu.edu (D.S.); pmahadevan@ut.edu (P.M.)

Abstract: The determination of core genes in viral and bacterial genomes is crucial for a better understanding of their relatedness and for their classification. CoreGenes5.0 is an updated user-friendly web-based software tool for the identification of core genes in and data mining of viral and bacterial genomes. This tool has been useful in the resolution of several issues arising in the taxonomic analysis of bacteriophages and has incorporated many suggestions from researchers in that community. The webserver displays result in a format that is easy to understand and allows for automated batch processing, without the need for any user-installed bioinformatics software. CoreGenes5.0 uses group protein clustering of genomes with one of three algorithm options to output a table of core genes from the input genomes. Previously annotated "unknown genes" may be identified with homologues in the output. The updated version of CoreGenes is able to handle more genomes, is faster, and is more robust, providing easier analysis of custom or proprietary datasets. CoreGenes5.0 is accessible at coregenes.org, migrating from a previous site.

Keywords: coregenes; webserver; bioinformatics; genomics; viruses; bacteria

Citation: Davis, P.; Seto, D.; Mahadevan, P. CoreGenes5.0: An Updated User-Friendly Webserver for the Determination of Core Genes from Sets of Viral and Bacterial Genomes. *Viruses* **2022**, *14*, 2534. https://doi.org/10.3390/v14112534

Academic Editor: Hernan Garcia-Ruiz

Received: 23 October 2022
Accepted: 11 November 2022
Published: 16 November 2022

1. Introduction

Core genes are the set of common genes in a set of genomes, in contrast to genes which are not common amongst these genomes (accessory genes) [1]. A better understanding of these core genes has led to the design and synthesis of a minimal bacterial genome [2] to address basic and applied (biotechnological) questions and needs. Core genes have also revealed insights into carbon cycling and carbohydrate metabolism in soil metagenomes [3]. The variation in core genes can also be used for epidemiological typing in bacteria [4]. Core genes have also been used in evolutionary studies of nucleo-cytoplasmic large DNA viruses of eukaryotes [5].

With the growing bacterial and viral genome databases, there has been an increase in the demand for easily accessible and user-friendly software for genomic analyses. Since its original development in 2002 [6], CoreGenes has been continuously updated in order to increase its ease of use, add functionality, and meet additional user suggestions [7]. In particular, it has been used extensively for characterizing and classifying bacteriophages, for example resolving taxonomic issues in the *Podoviridae*, *Myoviridae*, and *Siphoviridae* families [8–10] and characterizing newly sequenced bacteriophages [11]. The taxonomic approach pioneered by the CoreGenes application for the *Podoviridae*, *Myoviridae*, and *Siphoviridae* families has been extended by other researchers [12]. CoreGenes has demonstrated utility in the data mining of pathogenic viral and bacterial genomes [13]. However, a limiting factor with older versions of the software was its slower processing speed, which required a smaller allowance for accession numbers in each run and greatly limited the size of the input genomes. In this iteration of CoreGenes, these limitations have been greatly improved. Additionally, coding sequence retrieval can now be used on the web interface,

allowing the user to easily view the genome or retrieve new coding sequence information quickly, bypassing the NCBI webpage.

2. Methods

CoreGenes5.0 is written mainly using Python for processing. The webpage implementation is written using Python's Django module, HTML, CSS, and JavaScript. The reimplementation of the CoreGenes3.5 algorithm uses the same iterative algorithm to process a query genome against a reference genome, and to then create a new consensus genome as described previously [7]. The updated 3.5 version uses MMseqs2 [14] for rapid protein searches instead of Washington University BLAST (WU-BLAST). The genomes are retrieved from GenBank using the user-inputted accession numbers. The former option for the user-input blast score has been updated to accept e-values. While the former iteration of CoreGenes only allowed for up to five input accession numbers, both versions of the updated CoreGenes algorithm now allow for the quick input of twenty accession numbers or an uploaded .txt file of up to two hundred comma-separated accession numbers, using the file upload option. It is recommended that protein queries not exceed 50 input genomes for bacterial genomes, due to their larger size requiring longer processing.

CoreGenes5.0 uses the GET_HOMOLOGUES package [15,16] with BLAST+ [17,18] to perform group protein clustering using default options. The group protein clustering is supported by three common clustering algorithms: OrthoMCL [19], bidirectional best hit, and COGtriangles [20]. In this updated version, an optional input for the inclusion or exclusion of paralogs is available. This option excludes sequences that have significant matches in the same genome and is implemented by the GET_HOMOLOGUES package. The GET_HOMOLOGUES manual defines "inparalogs" as "sequences with best hits in its own genome and excludes clusters with these sequences". In addition, the user can also input a BLAST e-value in the web interface. The output table of either version is displayed on the webserver, with a link that is available for up to one month. A .csv file with a similar table formatting is available for download through the optional email input. For queries which may take longer to process, an optional email notification is now available with a link to the results page.

CoreGenes now includes a coding sequence (CDS) retrieval option. Genome accession numbers are input, then used to parse coding sequence data from the Genbank database. The parsed CDS files can be downloaded in a zipped folder containing the FASTA files (with a .fasta extension) of the input genomes via email. The Custom Dataset upload option has been updated to accept a standardized FASTA file, with a space-separated accession identifier and protein title.

The rewritten Iterative Comparison Algorithm with MMseqs2 allows for a higher number of accession numbers to be processed. While the iterative comparison algorithm is able to handle larger genomes than the previous version of CoreGenes, the algorithm still performs with greater speed when using viral and small bacterial genomes. The display table is formatted in an easy-to-read format, with hyperlinks for each accession number to the NCBI page. Hypothetical proteins are highlighted in red for an easy-to-locate comparison with annotated homologues. For larger bacterial genomes, the CoreGenes5.0 webserver is able to process multiple 5 Mb genomes in minutes and is able to handle genomes as large as 10 Mb. The main differences between CoreGenes5.0 and the previous version, CoreGenes4.0, are shown in Table 1 below.

Table 1. Additional functionality provided by CoreGenes5.0 compared to CoreGenes4.0.

Functionality	CoreGenes4.0	CoreGenes5.0
Additional clustering algorithms such as Bidirectional best hit, OrthoMCL and COGtriangles made available through the GET_HOMOLOGUES package	X	✓
Faster protein searches using MMseqs2 in the Iterative Comparison Algorithm	X	✓
Email results to user	X	✓
Easy CDS retrieval from GenBank	X	✓
More robust custom data input	X	✓

3. Results

The web interface of CoreGenes5.0 (Figure 1) enables the input of 20 genome accession numbers. We have designed the user interface to be intuitive and easy to use, without a lot of confusing options. Links on the left of the page lead to the file upload of accession numbers for batch processing of more than 20 genomes. Custom datasets can also be uploaded using the link on the left. The "old" CoreGenes3.5 algorithm can also be accessed by a link on the left (Iterative Comparison Algorithm). Figure 2 shows the partial CoreGenes5.0 output of five human adenovirus genomes. The output is clean and easy to read for the human eye. Links can be clicked on to access the complete genome or individual proteins in GenBank.

Figure 1. Web interface for CoreGenes5.0. Up to 20 accession numbers can be entered, as well as an optional email address for results. The bidirectional best hit, OrthoMCL, or COGTriangle algorithms can be chosen. The original CoreGenes3.5 iterative comparison algorithm can be chosen by clicking on the link on the left, as this has been used extensively in bacteriophage taxonomic analyses. Lists of accession numbers and custom datasets can also be uploaded by clicking on the links on the left.

The number of homologs in each column is :35

Human mastadenovirus F NC_001454	Human mastadenovirus E NC_003266	Human adenovirus 52 DQ923122	Human adenovirus B3 AY599834	Human adenovirus 1 MH183293
PI:NP_040845.1 Product:control protein E1A	PI:YP_068018.1 Product:E1A	PI:ABK35030.1 Product:E1A	PI:AAW33150.1 Product:29.1 kDa protein	PI:AYP21290.1 Product:E1A 32 kDa protein
PI:NP_040845.1 Product:control protein E1A	PI:YP_068018.1 Product:E1A	PI:ABK35030.1 Product:E1A	PI:AAW33151.1 Product:25 kDa protein	PI:AYP21290.1 Product:E1A 32 kDa protein
PI:NP_040845.1 Product:control protein E1A	PI:YP_068018.1 Product:E1A	PI:ABK35030.1 Product:E1A	PI:AAW33150.1 Product:29.1 kDa protein	PI:AYP21290.1 Product:E1A 32 kDa protein
PI:NP_040845.1 Product:control protein E1A	PI:YP_068018.1 Product:E1A	PI:ABK35030.1 Product:E1A	PI:AAW33150.1 Product:29.1 kDa protein	PI:AYP21291.1 Product:E1A 32 kDa protein
PI:NP_040848.1 Product:control protein E1B 19K	PI:YP_068019.1 Product:E1B 19K	PI:ABK35031.1 Product:E1B 19K	PI:AAW33152.1 Product:19 kDa small T antigen	PI:AYP21293.1 Product:E1B 19K
PI:NP_040850.1 Product:control protein E1B 55K	PI:YP_068020.1 Product:E1B 55K	PI:ABK35032.1 Product:E1B 55K	PI:AAW33153.1 Product:large T antigen	PI:AYP21294.1 Product:E1B 55K
PI:NP_040851.1 Product:capsid protein IX	PI:YP_068021.1 Product:IX	PI:ABK35033.1 Product:IX	PI:AAW33154.1 Product:hexon-associated protein IX	PI:AYP21295.1 Product:IX
PI:NP_040852.1 Product:encapsidation protein IVa2	PI:YP_068022.1 Product:IVa2	PI:ABK35034.1 Product:IVa2	PI:AAW33155.1 Product:maturation protein IVa2	PI:AYP21296.1 Product:IVa2
PI:NP_040853.1 Product:DNA polymerase	PI:YP_068023.1 Product:pol	PI:ABK35035.1 Product:pol	PI:AAW33156.2 Product:DNA polymerase	PI:AYP21297.1 Product:DNA polymerase
PI:NP_040854.1 Product:terminal protein precursor pTP	PI:YP_068024.1 Product:pTP	PI:ABK35036.1 Product:pTP	PI:AAW33160.1 Product:DNA terminal protein precursor	PI:AYP21299.1 Product:pTP
PI:NP_040855.1 Product:encapsidation protein 52K	PI:YP_068025.1 Product:52K	PI:ABK35037.1 Product:52k	PI:AAW33162.1 Product:55 kDa protein	PI:AYP21300.1 Product:L1 53 kDa protein
PI:NP_040856.1 Product:capsid protein precursor pIIIa	PI:YP_068026.1 Product:pIIIa	PI:ABK35038.1 Product:pIIIa	PI:AAW33163.1 Product:protein IIIa precursor	PI:AYP21302.1 Product:IIIa
PI:NP_040857.1 Product:penton base	PI:YP_068027.1 Product:III	PI:ABK35039.1 Product:III	PI:AAW33164.1 Product:penton protein	PI:AYP21301.1 Product:penton
PI:NP_040858.1 Product:core protein precursor pVII	PI:YP_068028.1 Product:pVII	PI:ABK35040.1 Product:pVII	PI:AAW33165.1 Product:protein VII precursor	PI:AYP21303.1 Product:pVII
PI:NP_040859.1 Product:core protein V	PI:YP_068029.1 Product:V	PI:ABK35041.1 Product:V	PI:AAW33166.1 Product:protein V precursor	PI:AYP21304.1 Product:pV

Figure 2. Partial CoreGenes output from five human adenovirus genomes. The total number of core genes found is 35. Links to the complete genomes and individual proteins are also provided for additional analysis.

CoreGenes5.0 can process small or large bacterial genomes in minutes, as shown in Table 2. Three 5 Mb bacterial genomes take less than 10 minutes to process, while three 1 Mb genomes take only a minute to process. It must be noted that these times will increase as the number of queried genomes increase. A partial core gene output of three 5 Mb bacterial genomes is shown in Figure 3.

Table 2. CoreGenes5.0 analysis of three bacterial genomes. Five runs were completed, with different but uniform genome sizes for each run. The data were processed using the Bidirectional Best Hit method, paralog inclusion, and an e-value of 1×10^{-5}.

Genome Size	1 Mb	2 Mb	3 Mb	4 Mb	5 Mb
Accession #s	NUHQ01000006.1 CAIT01000004.1 ASWA01000004.1	MTBP01000002.1 CP033822.1 UHGI01000001.1	UGNN01000001.1 CP035563.1 UKAD01000001.1	CP021892.1 CP012872.1 UGNN01000001.1	UGBR01000009.1 SILS01000001.1 RDRU01000001.1
Run Time	00 h:01 m:10 s	00 h:02 m:22 s	00 h:05 m:00 s	00 h:05 m:54 s	00 h:09 m:56 s
Number of Homologues	41	372	499	732	1213

The number of homologs in each column is :1213

Escherichia coli ***UGBR01000009***	**Rhizobium leguminosarum** ***SILS01000001***	**Paraburkholderia phenazinium** ***RDRU01000001***
PI:*STH53522.1* Product:chromosomal replication initiation protein	PI:*TBD57078.1* Product:chromosomal replication initiator protein DnaA	PI:*RMD36346.1* Product:chromosomal replication initiator protein DnaA
PI:*STH53525.1* Product:DNA polymerase III	PI:*TBD61537.1* Product:DNA polymerase III subunit beta	PI:*RMD36347.1* Product:DNA polymerase III beta subunit
PI:*STH53527.1* Product:DNA gyrase subunit B	PI:*TBD61191.1* Product:DNA topoisomerase (ATP-hydrolyzing) subunit B	PI:*RMD36348.1* Product:DNA gyrase subunit B
PI:*STH53529.1* Product:radical SAM domain-containing protein	PI:*TBD57883.1* Product:DUF937 domain-containing protein	PI:*RMD38249.1* Product:uncharacterized protein YidB (DUF937 family)
PI:*STH53533.1* Product:2-dehydro-3-deoxygalactonokinase	PI:*TBD57588.1* Product:2-dehydro-3-deoxygalactonokinase	PI:*RMD35810.1* Product:2-keto-3-deoxygalactonate kinase
PI:*STH53534.1* Product:2-dehydro-3-deoxy-6-phosphogalactonate aldolase	PI:*TBD57587.1* Product:2-dehydro-3-deoxy-6-phosphogalactonate aldolase	PI:*RMD35809.1* Product:2-keto-3-deoxy-phosphogalactonate aldolase
PI:*STH53535.1* Product:galactonate dehydratase	PI:*TBD60700.1* Product:galactonate dehydratase	PI:*RMD36871.1* Product:galactonate dehydratase

Figure 3. Partial CoreGenes output from three bacterial genomes approximately 5 Mb in size. The links to the complete genomes and individual proteins are shown. As noted, the number of core genes identified is 1213.

The annotation of hypothetical or previously annotated "unknown" proteins is made possible by highlighting all hypothetical proteins in red and by providing putative homologues across the output table. For example, in Figure 4, a hypothetical protein in the bacterium *Rhizobium leguminosarum* is annotated as a hypothetical protein and labeled in red in the right-hand column. The homologous protein in *E. coli*, which is annotated as a 2′-5′ RNA ligase (e-value threshold $\leq 1 \times 10^{-5}$), is located on the same row in the left-hand column. It is very likely that this hypothetical protein is also a 2′-5′ RNA ligase. Theoretically, a genome with hundreds of hypothetical proteins may be annotated using a closely related reference genome and CoreGenes5.0.

Escherichia coli ***UGBR01000009***	**Rhizobium leguminosarum** ***SILS01000001***
PI:*STH54755.1* Product:alanyl-tRNA synthetase	PI:*TBD59098.1* Product:alanine--tRNA ligase
PI:*STH54758.1* Product:inner membrane protein	PI:*TBD57832.1* Product:DedA family protein
PI:*STH54762.1* Product:2'-5' RNA ligase	PI:*TBD59575.1* Product:hypothetical protein
PI:*STH54763.1* Product:multidrug resistance protein B	PI:*TBD60211.1* Product:DHA2 family efflux MFS transporter permease

Figure 4. CoreGenes assists in the identification and annotation of hypothetical proteins. Here, the annotated hypothetical protein in *Rhizobium leguminosarum*, labeled in red, is putatively identified as 2′-5′ RNA ligase in a reference genome, *E. coli*. Flagging this to the user allows for a subsequent evaluation for identity.

4. Discussion

CoreGenes5.0 is a vastly improved version of its predecessor, CoreGenes3.5 [7], which is no longer supported at its previous home (http://binf.gmu.edu/genometools.html, accessed on 9 November 2022). It is strongly recommended that users migrate to this updated version. This version is more user-friendly, faster, more robust, and able to handle more genomes, all of which were suggested by users. The bacteriophage research community has used CoreGenes extensively to resolve taxonomic issues that were in question, based on traditional methods. Nucleotide sequence analysis for taxonomy has been improved with the application of this tool [21]. CoreGenes has been cited in 334 publications, as per Google Scholar. Frequent citations are in the International Committee on Taxonomy of Viruses (ICTV) Working Group publications, which are not recorded by Pubmed (https://pubmed.ncbi.nlm.nih.gov, accessed on 9 November 2022). For example, Kropinski A.M., Turner D., Tolstoy I., Moraru C., Adriaenssens E.M., and Mahony J. cited results from CoreGenes 3.5 in "Code assigned: 2022.001 B" as a report from the Bacterial Viruses Subcommittee, Caudoviricetes Study Group in 2022. This publication notes that "the genera Sfi21dtunalikevirus (2013.036 a-dB) and Sfi1unalikevirus (2013.034 a-dB) were renamed Moineauvirus and Brussowvirus, respectively through Taxonomy Proposal 2015.025 aB". These examples serve to underscore the usefulness and current application of CoreGenes.

CoreGenes5.0 provides large-sized bacterial genomes analyses in a shorter timeframe as well. In addition, results can be downloaded in .csv format for offline use. Batch processing is available by uploading a list of accession numbers, with the results emailed to the user. Hypothetical proteins can also be readily identified and annotated, using reference genomes. An added option of including/excluding paralogs may be of particular interest to CoreGenes5.0 users. This was conveniently a function of the GET_HOMOLOGUES package [15,16]. Its definition of "paralogs" is used, as noted in Methods. All of these features make CoreGenes5.0 an easy-to-use software tool for non-computationally savvy users.

Future work will involve replacing the BLAST+ portions of the pipeline with MMseqs2 [14] or DIAMOND [22] to perform fast protein searches. This will enable even more genomes to be processed at a faster rate. Eventually, we hope to transfer the application to a more powerful computer server or a cloud computing environment for higher throughput processing. It should be noted that CoreGenes has been in continuous use since the earliest version in 2002 [6,23].

Author Contributions: P.D. and P.M. implemented the webserver and wrote the manuscript. P.M. conceived the idea for the webserver. D.S. contributed to manuscript writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The genome data analyzed in this paper are available from GenBank at ncbi.nlm.nih.gov.

Acknowledgments: We thank Venkat Mahadevan for helpful comments on this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Availability and Requirements: Server's homepage: coregenes.org. Operating system(s): Platform-independent. Other requirements: A web browser such as Chrome, Firefox, Safari, or Microsoft Edge is needed to access the webserver.

References

1. Tettelin, H.; Masignani, V.; Cieslewicz, M.J.; Donati, C.; Medini, D.; Ward, N.L.; Angiuoli, S.V.; Crabtree, J.; Jones, A.L.; Durkin, A.S.; et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13950–13955. [CrossRef] [PubMed]
2. Hutchison, C.A., 3rd; Chuang, R.Y.; Noskov, V.N.; Assad-Garcia, N.; Deerinck, T.J.; Ellisman, M.H.; Gill, J.; Kannan, K.; Karas, B.J.; Ma, L.; et al. Design and synthesis of a minimal bacterial genome. *Science* **2016**, *351*, aad6253. [CrossRef] [PubMed]
3. Howe, A.; Yang, F.; Williams, R.J.; Meyer, F.; Hofmockel, K.S. Identification of the Core Set of Carbon-Associated Genes in a Bioenergy Grassland Soil. *PLoS ONE* **2016**, *11*, e0166578. [CrossRef] [PubMed]
4. Leekitcharoenphon, P.; Lukjancenko, O.; Friis, C.; Aarestrup, F.M.; Ussery, D.W. Genomic variation in Salmonella enterica core genes for epidemiological typing. *BMC Genom.* **2012**, *13*, 88. [CrossRef] [PubMed]
5. Yutin, N.; Koonin, E.V. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. *Virol. J.* **2012**, *9*, 161. [CrossRef] [PubMed]
6. Zafar, N.; Mazumder, R.; Seto, D. CoreGenes: A computational tool for identifying and cataloging "core" genes in a set of small genomes. *BMC Bioinform.* **2002**, *3*, 12. [CrossRef]
7. Turner, D.; Reynolds, D.; Seto, D.; Mahadevan, P. CoreGenes3. 5: A webserver for the determination of core genes from sets of viral and small bacterial genomes. *BMC Res. Notes* **2013**, *6*, 140. [CrossRef]
8. Lavigne, R.; Seto, D.; Mahadevan, P.; Ackermann, H.W.; Kropinski, A.M. Unifying classical and molecular taxonomic classification: Analysis of the Podoviridae using BLASTP-based tools. *Res. Microbiol.* **2008**, *159*, 406–414. [CrossRef]
9. Lavigne, R.; Darius, P.; Summer, E.J.; Seto, D.; Mahadevan, P.; Nilsson, A.S.; Ackermann, H.W.; Kropinski, A.M. Classification of Myoviridae bacteriophages using protein sequence similarity. *BMC Microbiol.* **2009**, *9*, 224. [CrossRef]
10. Adriaenssens, E.M.; Edwards, R.; Nash, J.H.; Mahadevan, P.; Seto, D.; Ackermann, H.-W.; Lavigne, R.; Kropinski, A.M. Integration of genomic and proteomic analyses in the classification of the Siphoviridae family. *Virology* **2015**, *477*, 144–154. [CrossRef]
11. Zhou, W.; Feng, Y.; Zong, Z. Two New Lytic Bacteriophages of the Myoviridae Family Against Carbapenem-Resistant Acinetobacter baumannii. *Front. Microbiol.* **2018**, *9*, 850. [CrossRef] [PubMed]
12. Bin Jang, H.; Bolduc, B.; Zablocki, O.; Kuhn, J.H.; Roux, S.; Adriaenssens, E.M.; Brister, J.R.; Kropinski, A.M.; Krupovic, M.; Lavigne, R.; et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **2019**, *37*, 632–639. [CrossRef] [PubMed]
13. Mahadevan, P.; King, J.F.; Seto, D. Data mining pathogen genomes using GeneOrder and CoreGenes and CGUG: Gene order, synteny and in silico proteomes. *Int. J. Comput. Biol. Drug Des.* **2009**, *2*, 100–114. [CrossRef] [PubMed]
14. Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **2017**, *35*, 1026–1028. [CrossRef]
15. Contreras-Moreira, B.; Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* **2013**, *79*, 7696–7701. [CrossRef]
16. Vinuesa, P.; Contreras-Moreira, B. Robust Identification of Orthologues and Paralogues for Microbial Pan-Genomics Using GET_HOMOLOGUES: A Case Study of pIncA/C Plasmids. In *Bacterial Pangenomics, Methods in Molecular Biology*; Mengoni, A., Galardini, M., Fondi, M., Eds.; Humana Press: New York, NY, USA, 2015; Volume 1231, pp. 203–232.
17. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef]
18. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl. Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
19. Li, L.; Stoeckert, C.J.; Roos, D.S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **2003**, *13*, 2178–2189. [CrossRef]
20. Kristensen, D.M.; Kannan, L.; Coleman, M.K.; Wolf, Y.I.; Sorokin, A.; Koonin, E.V.; Mushegian, A. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* **2010**, *26*, 1481–1487. [CrossRef]
21. Kropinski, A.M.; Lingohr, E.J.; Ackermann, H.W. The genome sequence of enterobacterial phage 7–11, which possesses an unusually elongated head. *Arch Virol.* **2011**, *156*, 149–151. [CrossRef]
22. Buchfink, B.; Xie, C.; Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **2015**, *12*, 59–60. [CrossRef] [PubMed]
23. Mazumder, R.; Kolaskar, A.; Seto, D. GeneOrder: Comparing the order of genes in small genomes. *Bioinformatics* **2001**, *17*, 162–166. [CrossRef] [PubMed]