viruses

Article

Modeling of Mutational Events in the Evolution of Viruses

Akhtar Ali [1] **and Ulrich Melcher** [2,*]

[1] Department of Biological Sciences, University of Tulsa, Tulsa, OK 74104, USA; akhtar-ali@utulsa.edu
[2] Department of Biochemistry & Molecular Biology, Oklahoma State University, Stillwater, OK 74078-3035, USA
* Correspondence: ulrich.melcher@okstate.edu; Tel.: +1-405-744-6210

Received: 4 March 2019; Accepted: 2 May 2019; Published: 5 May 2019

Abstract: Diverse studies of viral evolution have led to the recognition that the evolutionary rates of viral taxa observed are dependent on the time scale being investigated—with short-term studies giving fast substitution rates, and orders of magnitude lower rates for deep calibrations. Although each of these factors may contribute to this time dependent rate phenomenon, a more fundamental cause should be considered. We sought to test computationally whether the basic phenomena of virus evolution (mutation, replication, and selection) can explain the relationships between the evolutionary and phylogenetic distances. We tested, by computational inference, the hypothesis that the phylogenetic distances between the pairs of sequences are functions of the evolutionary path lengths between them. A Basic simulation revealed that the relationship between simulated genetic and mutational distances is non-linear, and can be consistent with different rates of nucleotide substitution at different depths of branches in phylogenetic trees.

Keywords: virus–host co-divergence; endogenous viral elements; virus networks; speciation

1. Virus Evolution Introduction

Viruses are often thought of as agents that have evolved to threaten humans and other manifestations of life. This view is strengthened by the impression that many disease epidemics (of humans, other animals, and plants) are of viral instigation. Nevertheless, the relationships between viruses and their hosts cover the full range, from mutualists or symbionts to commensals to pathogens [1–4]. As life forms have evolved, so too have the viruses associated with them. These observations generate the hypothesis that many viruses have co-diverged with their hosts during evolution. Here, we discuss ways to reconcile conflicting claims of high mutation rates in short-term evolution studies with low nucleotide substitution rates at longer time scales.

Dating the emergence of viral lineages is important for understanding the epidemiology of viral outbreaks [5]. Dating has been complicated by the observation that the rates of substitution of nucleotide and/or amino acid residues in viral genes or gene products, respectively, are dependent on the time scales of the events being dated [6–8]. A variety of approaches to dating phylogenetic trees, covering different time periods, have been taken. In one approach, the rates of nucleotide substitution in viral genomes have been determined in prospective experiments. In these, hosts are inoculated with a genetically homogeneous viral genome, and hosts are then sampled at various times to determine how many nucleotides have changed [9–11]—a process whose usefulness is debated [12]. The typical substitution frequencies observed by such techniques are 10^{-3} to 10^{-4} substitutions per residue per year.

Phylogenetic inference methods are based on comparisons of sequence differences among contemporaneous viral isolates. In heterochronous sampling, the sequences obtained from viruses isolated at different times [13–16] provide calibration points for phylogenetic comparisons. Specimens from more distant times [13] occasionally yield additional sequences that can be used to calibrate phylogenies [17–19]. Similarly, paleo-sequences, or endogenous viral elements (EVEs) [20], can also be used in the calibration of trees. The fact that phylogenetic trees, for numerous viral taxa, resemble, in topology and branch lengths, the trees constructed for their predominant hosts, argues for the co-divergence of nucleotide sequences of the host and virus during the evolution of host lineages [21], and provides a further calibration approach.

Time-Dependent Rate Phenomena

The integration of results from these diverse studies led to the recognition that the rates observed are dependent on the time scale being investigated—with short-term studies giving fast substitution rates, with orders of magnitude having lower rates for deep calibrations [20]. The current issue is how the number of evolutionary steps needed to evolve from an ancestor is related to the phylogenetic distance between the tip and ancestor. Power law and exponential relationships have been proposed and tested for the relationship of rates to time spans [22]. Artefacts and biases have been proposed as explanations for the discrepancies [12]. Various ways of correcting for time-dependent rate phenomena (TDRP) [8] have been proposed, including different scales for different biological epochs [23,24], avoiding sequencing errors, correcting for tree imbalances [25], and a better treatment of the influence of rate heterogeneity among sites. Accurate correction requires an understanding of the root causes of TDRP.

A variety of factors that may produce TDRP have been suggested. These include, hypermutation, recombination, and variations in selective forces [26]. Although each of these factors may contribute to the TDRP, a more fundamental cause should be considered. We sought to test computationally whether the basic phenomena of virus evolution (mutation, replication, and selection) can explain the relationships between the evolutionary and phylogenetic distances. Specifically, we aimed to test the hypothesis that the phylogenetic distances between the pairs of sequences are functions of the evolutionary path lengths between them.

2. Materials and Methods

To test the hypothesis, we defined the phylogenetic distance as the number of loci (sequence positions) that differ between the two sequences and evolutionary distances, as the number of modeled mutations used to transition between the two sequences, given the mutation, replication, and selection processes. We devised a simple program (Mevolve.c4d; Supplementary Material) to generate evolutionary distances using a network similar to the neutral network of Manrubia and Cuesta [27], except that we focused on non-neutral mutations. Genomes were represented as strings of an arbitrary length of 12, representing variable positions in the sequence. Each position in the string is diallelic [28], having just two states, "0" and "1", that can mutate from one to the other. A "1" in a position contributes a positive effect on the fitness of the virus in a particular environment, the latter assumed to be held constant. The program incorporates mutation, replication, and selection in cycles. Inputs are a "seed" string and a target string. The initial homogeneous seed population of size ten is subjected to the mutation of a single digit per string—the position of the mutation being chosen randomly.

After this mutation step, the population is allowed to expand to 100 members by replication. The number of offspring produced by each string is determined by the fitness of the string (determined by the number of "1" in the chain, the string's fitness index). To implement the selection, the population size is then lowered to the original ten by a bottleneck, in which ten random selections of pairs are compared by fitness index values, keeping the more fit for the next round of 10 strings. The process of mutation, replication, and selection is repeated for as many rounds as desired (a variable set by the user). The calculation is continued for as many times as needed to produce the target string of one of the ten selected population members. When the target string is obtained, the routine reports the number of repetitions (cycles) used. The routine was executed in Chipmunk Basic [29].

3. Results

The preliminary examination of the simulation script tested the effect of the relative order of string elements. It established that varying the positions of the "0" and "1" elements of the seed and target strings was without effect on the numbers of simulation cycles. The mean numbers of evolutionary cycles needed to increase the fitness index from a given level to the various higher fitness levels are illustrated in Figure 1. The simulation revealed that the relationship between the simulated genetic and mutational distances is non-linear, and can be consistent with different rates of nucleotide substitution observed at different depths of branches in phylogenetic trees. As expected because of the random calculation steps integral to the simulation script, considerable variation in the extent of mutation distances (computational cycles) occurred, as reflected in the substantial standard errors. Despite these large standard errors for numbers, the mutation distance increased as the genetic distance increased from 1 to 5. Also, as expected, the mutation distance increased with the genetic distances. Not unexpectedly, using targets with more than six negative alleles (genetic distance = 6) resulted in the highest number of computational cycles to reach a string consisting of all positive alleles. Considering transitioning from a distance of two viral substitutions to eight, such a change was accompanied by close to 6000 cycles of mutation, replication, and selection. Approximately 3000 cycles underlie a genetic distance leap of three substitutions (from two to five).

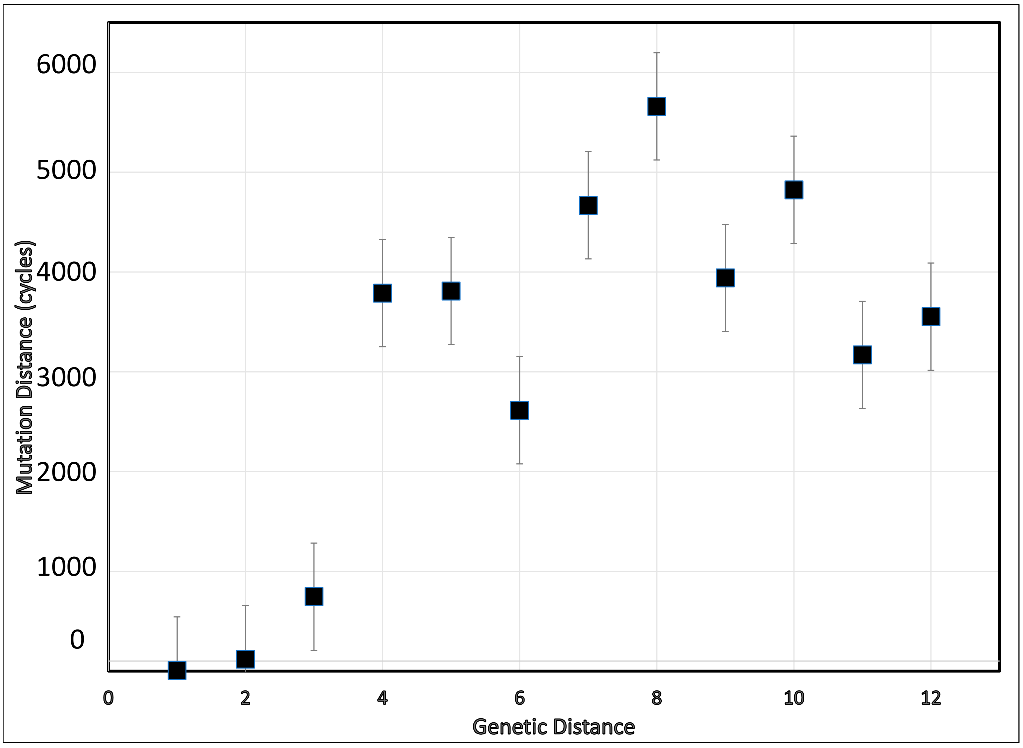

Figure 1. For hypothetical genomes containing 12 differentiating loci, the relation between the genetic distance (number of loci at which the members of a pair being analyzed differ) and mutational distance (number of mutational events needed to convert one member of the pair into the other) in a simulation of cycles of repeated random mutation, replication, and competition for 12 loci, allowing for only two alternative states, modeled as ″0″ and "1". Lines extending from the markers represent the standard error of five repetitions of each possible seed string in conversion to "111111111111".

4. Discussion

Viral populations are often thought of as quasispecies. They are networks consisting of viral genome segments as edges, connected by nodes at the edge ends. These are capable of mutational changes. In the present context, the mutation of a node residue to one that no longer interacts with a partner closes the edge to substitution. Conversely, a change in another network node can open an interaction with another network module (Figure 2).

The concept of a network of functional loci connected by mutational paths, as playing a significant role in the evolution of viruses in host environments, may help to imagine viral sequence space as a random network in which each node represents a viable viral genome sequence. Each node is directly connected to all of the other nodes by a bidirectional edge, so that it can reach by a single nucleotide mutation (Figure 2a). Conversely, a change in another network node can open an interaction with another species network. In the computational simulation, the recorded substitutions that do not contribute to the genetic distance may correspond to the networks of changes that open or close the gates connecting the submodules of the population.

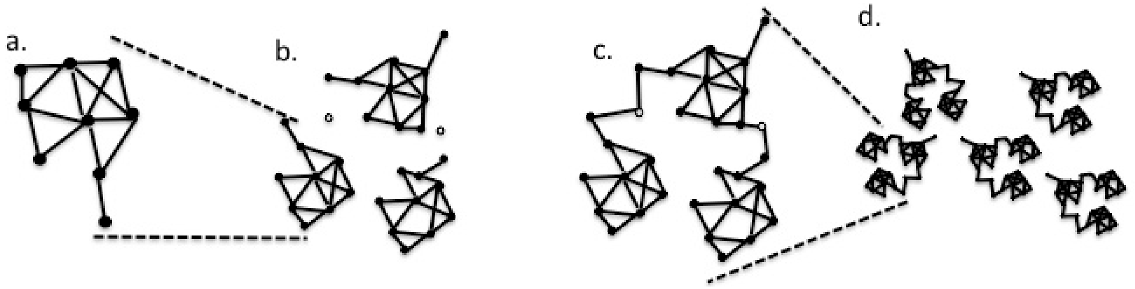

Figure 2. Diagram of the network concept of the viral evolution, where solid nodes represent viable viral genome sequences, and edges represent single nucleotide changes that connect the nodes. Open nodes represent sequences that are conditionally viable, namely: (**a**) basic single module, perhaps a viral isolate; (**b**) three such modules near one another; and (**c**) modules of (**b**) now connected through the formation of a supermodule by joining at conditional nodes. The supermodule may be equivalent to a viral species; (**d**) a potentially higher-level module composed of supermodules, potentially forming a viral genus. The network should form a module with a high connectivity. A plot of the frequencies of the node connectivity values (measured as k, degree, which are determined by the number of edges attached to a node) should follow a Poisson-like distribution [30]. There will be multiple paths connecting many pairs of nodes to each other, yet all of the nodes must be connected to the network by at least one edge.

The appropriateness of the network model is supported by the success of pairwise sequence comparison (PASC) [1,31–36] and the DEmARC analysis [37]. PASC is a web-based tool (http://www.ncbi.nlm.nih.gov/sutils/pasc) originally devised to both support the virus taxonomy and to facilitate it. A PASC plot (Figure 3) displays the frequencies of the sequence pair similarities in the bins of the nucleotide similarities for all pairs of taxa being considered, and thus represents the average connectivity within a network cluster.

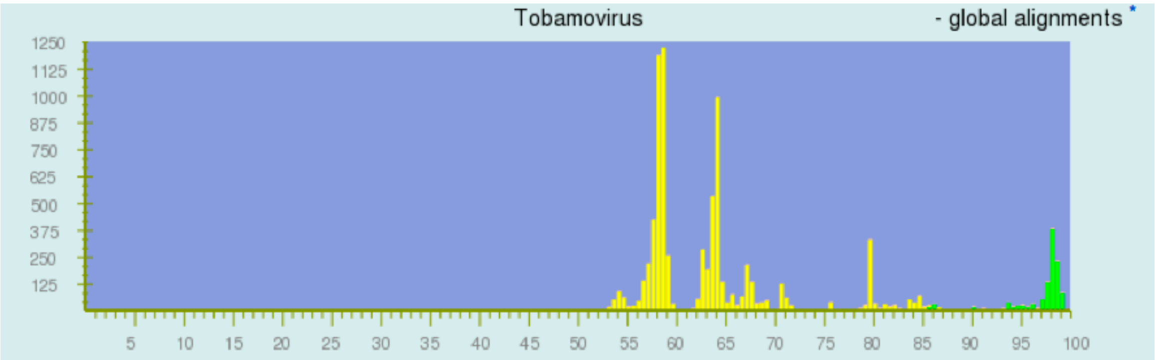

Figure 3. Pairwise sequence comparison (PASC) display of the distance relationships among the *Tobamovirus* sequences (Screenshot, March 2015). A histogram of the numbers of sequence pairs yielding binned similarity values. Green bars are as a result of within species comparisons, and the yellow bars are between members of different species.

These frequencies, for many collections of taxa, exhibit a series of peaks and valleys. Each peak has a Poisson-like distribution, as expected from the network theory. The peaks are interpreted as substitution distances for a particular taxonomic level (within strain, within species, within genus, within family, etc.). In this way, pairs can be assigned as members of the same strain of a species, or as different strains of the same species, or as different species of the same genus, and so on, depending on the peak order. Such a plot is predicted by the network theory, where each peak corresponds to a module of related sequences with other modules of the same divergence.

Recall that nodes are defined as viable genome sequences, genomes whose fitness is greater than some arbitrary value. "Viable" or "fit" are used here to characterize a genome molecule that, when placed in a given environment, will generate multiple copies of itself ($R_o > 1$). Some nucleotide residues in viral genomes can mutate without major effects on the genome fitness. Groupings of nodes interconnected by a high density of edges, called modules, can be imagined as describing a strain of a virus. A module for strain A of a virus will be distinct from a module formed by strain B sequences, with few, if any, edges joining the two modules.

There will be fewer nodes at the periphery of a module. Connections from strain A outlier nodes to similar nodes of strain B may exist. These connections provide an evolutionary path between strains A and B. Such connections likely will require further mutational changes (node connections). Over time, the taxa will wander their landscape networks. Occasionally, one will wander to the vicinity of a node shared with another module, representing another taxon or potential taxon—a boundary.

The addition of further strain modules will result in a supermodule, depicting the possible evolutionary paths within a viral species. At a still larger scale, the species supermodules may connect to create genus super-supermodules. The process is extensible to larger and larger scales. Thus, the structure of the connectivity diagrams is independent of scale.

The simple computational strategy developed here to address the explanation of the TDR can also be modified to model events, other than replication, mutation, and selection, that affect evolution. Such events can include new hosts, environmental changes, host defenses, and agricultural practice. Introducing into models modified 12mer strings that are defined to signal path closure, or new options for a particular host environment, could reveal new pathways of evolution.

Supplementary Materials: The following are available online: http://www.mdpi.com/1999-4915/11/5/418/s1 (Mevolve.c4d).

Author Contributions: Conceptualization, software, and drafting the text, U.M.; writing (review and editing), A.A.

Acknowledgments: U.M.'s participation in this work was made possible by the Oklahoma Agricultural Experiment Station and the Robert J. Sirny Professorship at Oklahoma State University. Peggy Hill, Kay Scheets, Richard Eberle, and Francisco Ochoa-Corona are thanked for critiquing a draft of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Bao, X.D.; Roossinck, M.J. A life history view of mutualistic viral symbioses: Quantity or quality for cooperation? *Curr. Opin. Microbiol.* **2013**, *16*, 514–518. [CrossRef] [PubMed]
2. Márquez, L.M.; Redman, R.S.; Rodriguez, R.J.; Roossinck, M.J. A virus in a fungus in a plant: Three-way symbiosis required for thermal tolerance. *Science* **2007**, *315*, 513–515. [CrossRef] [PubMed]
3. Roossinck, M.J. The good viruses: Viral mutualistic symbioses. *Nat. Rev. Microbiol.* **2011**, *9*, 99–108. [CrossRef] [PubMed]
4. Roossinck, M.J. Plants, viruses and the environment: Ecology and mutualism. *Virology* **2015**, *479*, 271–277. [CrossRef] [PubMed]
5. Elena, S.F.; Agudelo-Romero, P.; Carrasco, P.; Codoner, F.M.; Martin, S.; Torres-Barcelo, C.; Sanjuan, R. Experimental evolution of plant RNA viruses. *Heredity* **2008**, *100*, 478–483. [CrossRef] [PubMed]
6. Duchêne, S.; Holmes, E.C.; Ho, S.Y.W. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. R. Soc. B Biol. Sci.* **2014**, *281*, 20140732.
7. Gibbs, A.J.; Fargette, D.; Garcia-Arenal, F.; Gibbs, M.J. Time–the emerging dimension of plant virus studies. *J. Gen. Virol.* **2010**, *91*, 13–22.
8. Ho, S.Y.W.; Lanfear, R.; Bromham, L.; Phillips, M.J.; Soubrier, J.; Rodrigo, A.G.; Cooper, A. Time-dependent rates of molecular evolution. *Mol. Ecol.* **2011**, *20*, 3087–3101. [CrossRef]
9. Ge, L.; Zhang, J.; Zhou, X.; Li, H. Genetic structure and population variability of tomato yellow leaf curl China virus. *J. Virol.* **2007**, *81*, 5902–5907. [CrossRef] [PubMed]
10. Schneider, W.L.; Roossinck, M.J. Evolutionarily related Sindbis-like plant viruses maintain different levels of population diversity in a common host. *J. Virol.* **2000**, *74*, 3130–3134. [CrossRef]
11. Schneider, W.L.; Roossinck, M.J. Genetic diversity in RNA virus quasispecies is controlled by host-virus interactions. *J. Virol.* **2001**, *75*, 6566–6571. [CrossRef] [PubMed]
12. Emerson, B.C.; Hickerson, M.J. Lack of support for the time-dependent molecular evolution hypothesis. *Mol. Ecol.* **2015**, *24*, 702–709. [CrossRef] [PubMed]
13. Harkins, K.M.; Stone, A.C. Ancient pathogen genomics: Insights into timing and adaptation. *J. Hum. Evol.* **2015**, *79*, 137–149. [CrossRef]
14. Pagan, I.; Firth, C.; Holmes, E.C. Phylogenetic analysis reveals rapid evolutionary dynamics in the plant RNA virus genus tobamovirus. *J. Mol. Evol.* **2010**, *71*, 298–307. [CrossRef]
15. Pagan, I.; Holmes, E.C. Long-term evolution of the Luteoviridae: Time scale and mode of virus speciation. *J. Virol.* **2010**, *84*, 6177–6187. [CrossRef] [PubMed]
16. Harkins, G.; Delport, W.; Duffy, S.; Wood, N.; Monjane, A.; Owor, B.; Donaldson, L.; Saumtally, S.; Triton, G.; Briddon, R.; et al. Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virol. J.* **2009**, *6*, 104. [CrossRef] [PubMed]
17. Fraile, A.; Escriu, F.; Aranda, M.A.; Malpica, J.M.; Gibbs, A.J.; García-Arenal, F. A century of tobamovirus evolution in an Australian population of *Nicotiana glauca*. *J. Virol.* **1997**, *71*, 8316–8320.
18. Malmstrom, C.M.; Shu, R.; Linton, E.W.; Newton, L.A.; Cook, M.A. Barley yellow dwarf viruses (BYDVs) preserved in herbarium specimens illuminate historical disease ecology of invasive and native grasses. *J. Ecol.* **2007**, *95*, 1153–1166. [CrossRef]
19. Smith, O.; Clapham, A.; Rose, P.; Liu, Y.; Wang, J.; Allaby, R.G. A complete ancient RNA genome: Identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. *Sci. Rep.* **2014**, *4*, 4003. [CrossRef] [PubMed]
20. Aiewsakun, P.; Katzourakis, A. Time dependency of foamy virus evolutionary rate estimates. *BMC Evol. Biol.* **2015**, *15*, 119. [CrossRef]
21. Stobbe, A.H.; Melcher, U.; Palmer, M.W.; Roossinck, M.J.; Shen, G. Co-divergence and host-switching in the evolution of tobamoviruses. *J. Gen. Virol.* **2012**, *93*, 408–418. [CrossRef] [PubMed]
22. Biek, R.; Pybus, O.G.; Lloyd-Smith, J.O.; Didelot, X. Measurably evolving pathogens in the genomic era. *Trends Ecol. Evol.* **2015**, *30*, 306–313. [CrossRef]

23. Bielejec, F.; Lemey, P.; Baele, G.; Rambaut, A.; Suchard, M.A. Inferring Heterogeneous Evolutionary Processes Through Time: From Sequence Substitution to Phylogeography. *Syst. Biol.* **2014**, *63*, 493–504. [CrossRef]
24. Snir, S.; Wolf, Y.I.; Koonin, E.V. Universal Pacemaker of Genome Evolution. *PLoS Comput. Biol.* **2012**, *8*, e1002785. [CrossRef] [PubMed]
25. Duchene, D.; Duchene, S.; Ho, S.Y.W. Tree imbalance causes a bias in phylogenetic estimation of evolutionary timescales using heterochronous sequences. *Mol. Ecol. Resour.* **2015**, *15*, 785–794. [CrossRef] [PubMed]
26. Wertheim, J.O.; Smith, M.D.; Smith, D.M.; Scheffler, K.; Pond, S.L.K. Evolutionary Origins of Human Herpes Simplex Viruses 1 and 2. *Mol. Biol. Evol.* **2014**, *31*, 2356–2364. [CrossRef]
27. Manrubia, S.; Cuesta, J.A. Evolution on neutral networks accelerates the ticking rate of the molecular clock. *J. R. Soc. Interface* **2015**, *12*, 20141010. [CrossRef]
28. Park, S.C.; Szendro, I.G.; Neidhart, J.; Krug, J. Phase transition in random adaptive walks on correlated fitness landscapes. *Phys. Rev. E* **2015**, *91*, 042707. [CrossRef] [PubMed]
29. Nicholson, J. Ronald H Chipmunk Basic. Available online: http://www.nicholson.com/rhn/basic/ (accessed on 15 September 2018).
30. Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z.N.; Barabasi, A.L. The large-scale organization of metabolic networks. *Nature* **2000**, *407*, 651–654. [CrossRef] [PubMed]
31. Adams, M.J.; Antoniw, J.F.; Bar-Joseph, M.; Brunt, A.A.; Candresse, T.; Foster, G.D.; Martelli, G.P.; Milne, R.G.; Zavriev, S.K.; Fauquet, C.M. The new plant virus family Flexiviridae and assessment of molecular criteria for species demarcation. *Arch. Virol.* **2004**, *149*, 1045–1060. [CrossRef]
32. Adams, M.J.; Antoniw, J.F.; Fauquet, C.M. Molecular criteria for genus and species discrimination within the family Potyviridae. *Arch. Virol.* **2005**, *150*, 459–479. [CrossRef] [PubMed]
33. Bao, Y.; Kapustin, Y.; Tatusova, T. Virus classification by pairwise sequence comparison (PASC). In *Encyclopedia of Virology*; Mahy, B.W.J., Regenmortel, M.H.V.V., Eds.; Elsevier: Oxford, UK, 2008; Volume 5, pp. 342–348.
34. de Villiers, E.M.; Fauquet, C.; Broker, T.R.; Bernard, H.U.; zur Hausen, H. Classification of papillomaviruses. *Virology* **2004**, *324*, 17–27. [CrossRef] [PubMed]
35. Fauquet, C.M.; Stanley, J. Geminivirus classification and nomenclature: Progress and problems. *Ann. Appl. Biol.* **2003**, *142*, 165–189. [CrossRef]
36. Oberste, M.S.; Maher, K.; Kilpatrick, D.R.; Pallansch, M.A. Molecular evolution of the human enteroviruses: Correlation of serotype with VP1 sequence and application to picornavirus classification. *J. Virol.* **1999**, *73*, 1941–1948. [PubMed]
37. Lauber, C.; Gorbalenya, A.E. Partitioning the Genetic Diversity of a Virus Family: Approach and Evaluation through a Case Study of Picornaviruses. *J. Virol.* **2012**, *86*, 3890–3904. [CrossRef] [PubMed]