RUITHER ARTHUR LOCH GOMES

AUTOMATION TOOL FOR TAXONOMIC CLASSIFICATION OF VIRUSES IN THE FAMILY Geminiviridae

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
G633a
2018

Gomes, Ruither Arthur Loch, 1990-
Automation tool for taxonomic classification of viruses in the family *Geminiviridae* / Ruither Arthur Loch Gomes. – Viçosa, MG, 2018.
viii, 39f. : il. (algumas color.) ; 29 cm.

Texto em inglês.
Inclui apêndices.
Orientador: Francisco Murilo Zerbini Júnior.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 34-35.

1. Begomovírus - Identificação. 2. Micro-organismos patogênicos. 3. Vírus - Classificação. 4. Bioinformática. I. Universidade Federal de Viçosa. Departamento de Fitopatologia. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 579.28

RUITHER ARTHUR LOCH GOMES

AUTOMATION TOOL FOR TAXONOMIC CLASSIFICATION OF VIRUSES IN THE FAMILY *Geminiviridae*

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Magister Scientiae*.

APROVADA: 31 de julho de 2018.

Tiago Antônio de Oliveira Mendes

Cosme Damião Cruz

Francisco Murilo Zerbini Júnior
(Orientador)

Dedico à minha companheira e futura esposa Rayane por todo seu amor, companheirismo e carinho. Dedico à toda minha família, que sempre me apoiou e auxiliou nos momentos bons e nos difíceis.

AGRADECIMENTOS

Ao amor da minha vida e futura esposa Rayane, sempre ao meu lado me incentivando a crescer, com todo carinho, amor e compreensão que pode dar;

A meus pais, Ivete e Luiz Pereira, que tanto já fizeram por mim, me educaram e apoiaram em todos os desafios que surgiram na vida, sempre com muito amor e dedicação;

Às minhas irmãs, Rayana e Rayara, por todo amor fraterno e companheirismo que só irmãos sabem ter;

A todos os amigos do laboratório de Virologia Vegetal Molecular, Cesar, Angélica, Tarsiane, Anelise, Ayane, João Paulo, João Wesley, Patrícia, pelo companheirismo;

Aos meus grandes amigos do Laboratório de Biologia Molecular de Plantas, Cleysinho e Otavio, por toda ajuda no trabalho, amizade e boas conversas de descontração;

Ao meu orientador, Prof. Murilo Zerbini, pela disponibilidade, paciência, incentivo, dedicação e profissionalismo;

À Universidade Federal de Viçosa, ao programa de Genética e Melhoramento e à FAPEMIG pela oportunidade de realização deste curso;

A todos que ajudaram e contribuíram direta ou indiretamente para a realização deste trabalho.

SUMÁRIO

ABSTRACT

GOMES, Ruither Arthur Loch, M.Sc., Universidade Federal de Viçosa, July, 2018. **Automation tool for taxonomic classification in the family Geminiviridae.** Advisor: Francisco Murilo Zerbini.

Pathogenic microorganisms have the potential to cause serious problems for humankind. Their precise taxonomic classification is an important step for understanding and combating the diseases caused by them. Several technologies were created to make it easier to classify microorganisms, and with the emergence of high-throughput sequencing technologies this process has been hugely accelerated. However, this led to another problem, because extremely large volumes of genetic sequence information are generated, making the bioinformatic analysis of sequences a time-consuming process. In the case of viruses classified in the family Geminiviridae, this problem is compounded by the large amount of new sequences that are deposited in public databases. Geminiviruses are responsible for large losses of production in economically important crops worldwide, which makes them the focus of much research leading to the constant discovery of new viruses. Although there are several ways of performing the taxonomic classification of microorganisms, the use of the percentage of identity obtained from the alignment between individuals has been increasingly applied. In the case of viruses with small genomes, the use of percent identities obtained from pairwise alignments has been applied for decades, so that several algorithms have already been created to accomplish this goal. However, none of the algorithms developed until today carries out the classification of the virus, leaving to the researcher the work of deciding the taxonomic classification, one virus at a time. Here we present a tool that will carry out the classification of viruses in the Geminiviridae. This tool is capable of acquiring the sequences as they are added to public databases or receiving the sequences given by the user. It then filters the added sequences to eliminate those already classified and parses the remaining sequences based on their percentage of pairwise identity with classified viruses . It also updates the values of taxonomic demarcation thresholds used to classify species and strains. Using this tool, it was possible to analyze all viruses added to public databases from January 2017 until July 2018. A total of 27 new species were identified. We also suggest revised demarcation

thresholds for the genera Becurtovirus, Capulavirus, Curtovirus, Grablovirus and Mastrevirus.

RESUMO

GOMES, Ruither Arthur Loch, M.Sc., Universidade Federal de Viçosa, julho de 2018. **Ferramenta de automatização para classificação taxonômica da família Geminiviridae**. Orientador: Francisco Murilo Zerbini.

Os microrganismos patogênicos são a causa de diversos problemas para a humanidade. Sua classificação é um importante passo para o entendimento e combate às doenças por eles causadas. Diversas tecnologias foram criadas para facilitar a classificação, e com o surgimento das tecnologias de sequenciamento de alto rendimento, esse trabalho foi imensamente acelerado. Entretanto isso gerou outro problema, pois volumes extremamente grandes de informação de sequências genéticas passaram a ser gerados, tornando a análise das sequências um processo que demanda muito tempo. No caso da família Geminiviridae, esse problema é somado à grande quantidade de novas sequências que são depositadas periodicamente nos bancos de dados públicos. Esses vírus são responsáveis por grandes perdas na produção de diversas de culturas de grande importância econômica em todo o mundo, o que leva à constante descoberta de novos geminivírus. Apesar de existirem diversas formas de classificar os microrganismos, a utilização da porcentagem de identidade obtida do alinhamento entre os indivíduos vem sendo cada dia mais aplicada. No caso de vírus com genoma pequeno, a utilização da identidade obtida de alinhamentos par-a-par já é aplicada há décadas, de modo que diversos algoritmos já foram criados para realizar essa tarefa. Entretanto nenhum dos algoritmos desenvolvidos até o presente realizam a classificação taxonômica dos vírus, deixando para o pesquisador o trabalho de realizar a classificação, vírus por vírus. Neste trabalho apresenta-se uma ferramenta que realiza a classificação de vírus da família Geminiviridae. A ferramenta é capaz de adquirir as sequências à medida que são adicionadas nos bancos de dados públicos, ou de recebê-las diretamente do usuário. Em seguida, filtra as sequencias adicionadas a fim de eliminar aquelas correspondentes a vírus já classificados, e analisa as restantes com base nas porcentagens de identidade obtidas do pareamento com os vírus já classificados. A ferramenta também atualiza os valores de limites de demarcação taxonômica utilizados para classificar esses vírus aos níveis de espécie e estirpe. Utilizando essa ferramenta foi possível analisar todos os vírus adicionados nos bancos de dados desde janeiro de 2017 até julho de 2018. Um total de 27 novas espécies foram identificadas. Sugere-se também

atualizações nos limites de demarcação de espécies e estirpes para os gêneros Becurtovirus, Capulavirus, Curtovirus, Grablovirus e Mastrevirus.

1. INTRODUCTION

Pathogenic microorganisms such as viruses have the potential to cause great harm to humankind. Thus, since the birth of microbiology, better methods of study and classification of these agents have been sought. Various technologies have been developed to this end. Nowadays, techniques involving high-throughput sequencing (HTS) have the capacity of generating extremely large amounts of data with high sensitivity and reduced time. The huge volume of data generated by HTS (and other modern techniques) creates difficulties for accurate analysis, mostly due to the inexistence of appropriate computational tools. An example of such bottleneck is the large amount of unclassified viral sequences deposited in public databases. The procedures for taxonomic assignment, based on manual analysis of each sequence (which in theory corresponds to one viral isolate), simply cannot keep up with the number of sequences generated. The existence of so many unclassified viruses poses serious problems. For example, control measures based on quarantine cannot be applied if the correct identity of the agent is undetermined.

The family Geminiviridae, comprised of nine genera of plant-infecting viruses, contains some of the most economically important viruses in agriculture, attacking crops such as cassava, tomato and corn in tropical and subtropical regions. Viruses in this family have small circular, single-stranded DNA genomes, varying between 2.5 and 5.2 kb, and their particles have a unique morphology of two geminated icosahedrons, which gave the name to the family. Geminiviruses are transmitted by insect vectors belonging to the order Homoptera (aphids, leafhoppers, treehopers and whiteflies, depending on the genus) and possess a broad genetic diversity, being abundant in several ecosystems. Due to their economical importance, genetic diversity and ease of cloning and sequencing their small DNA genomes, this family is the largest of all virology, with more than 400 member

species. The large number of new geminiviruses described every year poses a challenge to their taxonomy.

Several classification schemes have been proposed to accurately study and catalog the various existing pathogens. Recently, classification based on nucleotide sequence comparisons has been given preference due to its practicality and accuracy, especially in the case of viruses with small genomes (less than 30 kbp). In the case of geminiviruses, taxonomic classification takes into account characteristics such as genome organization, type of vector, host range, phylogenetic relationships and, with increasing emphasis, the complete genomic sequence. For classification based on genomic sequence, pairwise alignments should be performed and then the percent identity between two sequences is calculated. The percent identities of different individuals within taxonomic units such as genus and species are then compared in order to establish taxonomic demarcation thresholds that can divide the viruses in a precise and stable way. This procedure is reviewed in a systematic way by the Geminiviridae Study Group of the International Committee on Taxonomy of Viruses (ICTV; https://talk.ictvonline.org).

The large number of sequences deposited periodically on the databases has been making the classification process difficult, as each sequence must be analyzed independently in order to determine its taxonomic position. Since taxonomic assignment of new geminiviruses to genera and species is based on sequence comparisons, computational tools could greatly facilitate the process. Thus, seeking to expedite the taxonomic classification process of new whole genome sequences belonging to the family Geminiviridae, this work was conducted to develop and validated the Automation tool for Taxonomic classification of viruses in the family Geminiviridae (AutoTaxa), at the genus and species levels.

2. LITERATURE REVIEW

2.1. Genetic sequencing and the "data deluge"

Viruses are the cause of several diseases with high impact on human society, especially in the health and agricultural sectors. Classic methods of detection, characterization and taxonomic classification developed in the last century, based on filtration and purification of viral particles from cultures of infected cells, aided in the discovery and study of several important viruses, thus allowing the treatment and prevention of many viral diseases. Hence, until the beginning of the second half of the 20th century, it was believed that most viral diseases had been classified. However, anthropological factors such as globalization, deforestation and rapid unplanned urbanization have led different human (and agricultural) populations to exposure to new pathogens, and thus to the emergence of new diseases (Datta et al., 2015).

With the progress of science, other techniques dependent on the nucleic acid sequence have been developed, rendering cell culture-dependent methods obsolete for certain purposes. Techniques such as polymerase chain reaction (PCR)-based sequencing or microarrays were much faster compared to classical techniques, and enabled the discovery of several new genotypes of known viruses. Nevertheless, these new techniques had some limitations, such as the dependence of previous information on the sequence for the design of the oligonucleotide primers and hybridization probes (Datta et al., 2015). Precisely because of these limitations, the number of new viruses discovered using these techniques has been small.

For the discovery of truly "new" viruses, ie, viruses unrelated to those already cataloged, the use of unbiased amplification techniques such as rolling-circle amplification (RCA; Inoue-Nagata et al., 2004) and metagenomics methods (which are based on the amplification, cloning and mass sequencing of nucleic acids in environmental samples,

such as water from rivers and lakes) was necessary, since they do not require any previous knowledge about the sequenced organism. By 2005, new high-throughput sequencing (HTS) technologies (Figure 1) became available, leading to a new era in genetic sequencing. HTS technologies have high sensitivity and are able to generate complete genomic sequences from less sample material compared to the conventional sequencing method (Sanger). Together with unbiased amplification techniques, HTS considerably decreased the required time and raised the sensitivity in the detection of new viruses (Datta et al., 2015).

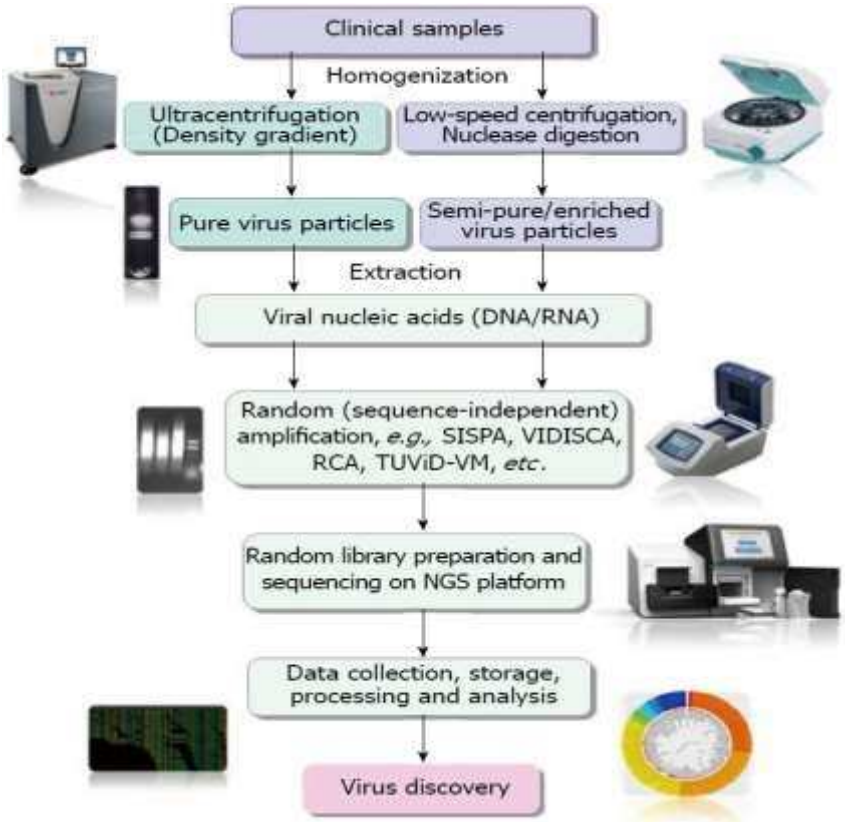


Figure 1. Diagram indicating the main steps in the detection of new viruses using HTS technology (Datta et al., 2015).

With the enhancement of HTS technology, there has been a huge increase in genetic sequencing efficiency, resulting in the production of a volume of genetic data on a scale that exceeds the evolution in the computing power of modern computers (Figure 2), a phenomenon known as "data deluge". From comparisons of the available statistics in GenBank, it is noted that the number of nucleotides present in this database is growing by approximately 43% annually, while the number of viral sequences grows by approximately 21% annually. The amount of data generated makes it difficult and tiresome for researchers to perform manual analyzes, generating a demand for computational tools to assist in this task (Datta et al., 2015; Zhao et al., 2017).

Figure 2. Comparison chart of Moore's Law (in green) with the cost of sequencing the human genome (in blue). With the development of high-throughput sequencing technologies, the cost of obtaining genomic sequences fell at a rate much higher than the cost of computer processors. Reproduced from http://www.a2apple.com/cracking-the-code/

Within this data deluge, we can highlight the progress in recent years in sequencing viruses from the family Geminiviridae. The viruses of this family are among the most economically important in the agricultural sector, which generates much interest in their characterization. Furthermore, geminiviruses have certain attributes that facilitate their

characterization, such as small circular DNA genome (which allows the use of the RCA technique), genome replication through double-stranded DNA intermediates (easy to manipulate by classical molecular methods), and the existence of vector-independent methods for plant inoculation (Varsani et al., 2014b; Zerbini et al., 2017; Zhao et al., 2017). As a result, there are already more than 400 known geminivirus species.

2.2. Family Geminiviridae

The family Geminiviridae is one of the largest and most successful family of plant viruses. It is composed of small non-enveloped viruses that have circular, single-stranded (ss) DNA genomes with one or two segments, varying between 2.5 and 5.2 kb. They have a unique particle morphology of twinned (geminate) icosahedra (Figure 3). Nowadays the family has nine genera. Viruses classified in the genus Becurtovirus, Capulavirus, Curtovirus, Eragrovirus, Grablovirus, Mastrevirus, Topocuvirus and Turncurtovirus have monosegmented genomes, whereas those classified in the genus Begomovirus can have mono or bisegmented genomes (Fauquet et al., 2008; Hanley-Bowdoin et al., 2013; Silva et al., 2017; Varsani et al., 2017; Zerbini et al., 2017).

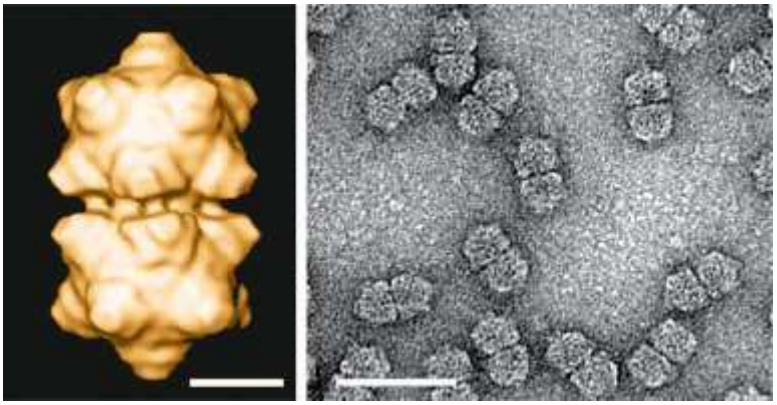


Figure 3. (Left) Cryo-electron microscopic reconstruction of the maize streak virus (MSV) particle. Bar, 10 nm. (Right) Purified particles of MSV stained with uranyl acetate. Bar, 50nm (Zerbini et al., 2017).

Within this family, the genus Begomovirus stands out. Members of this genus infect a wide range of dicotyledonous plants, mainly in the tropics and subtropics. Whereas in the old world (OW, Africa, Asia, Australasia and Europe) the begomoviruses are mostly monosegmented, in the New World (NW, the Americas) these viruses are mostly bissegmented. The two segments are known as DNA-A and DNA-B, each having a length between 2.5 and 2.6 kb (Brown et al., 2015; Zerbini et al., 2017).

The two segments do not have sequence identity with each other, except in a region of approximately 200 bases, known as the common region (CR) (or long intergenic region, LIR, in the case of monosegmented viruses). Within the CR there is an extremely conserved nonanucleotide among all geminiviruses (5'-TAATATTAC-3'), which is the origin of viral replication (Zerbini et al., 2017).

Although geminiviruses have a very conserved genomic organization (Figure 4), there is considerable variation among the different genera in the number of genes, ranging from two to seven. In the case of begomoviruses, the DNA-A of the NW viruses normally encodes five proteins: the capsid protein (CP); the transcriptional activator protein (TrAP), which also acts in the suppression of gene silencing; the replication enhancer protein (REn); the replication-associated protein (Rep); and the AC4 protein, involved in suppression of gene silencing and symptom induction. The DNA-B encodes two proteins involved in extracellular and intracellular transport of viral DNA (known respectively as movement protein, MP, and nuclear shuttling protein, NSP) (Briddon et al., 2010; Silva et al., 2017; Zerbini et al., 2017).

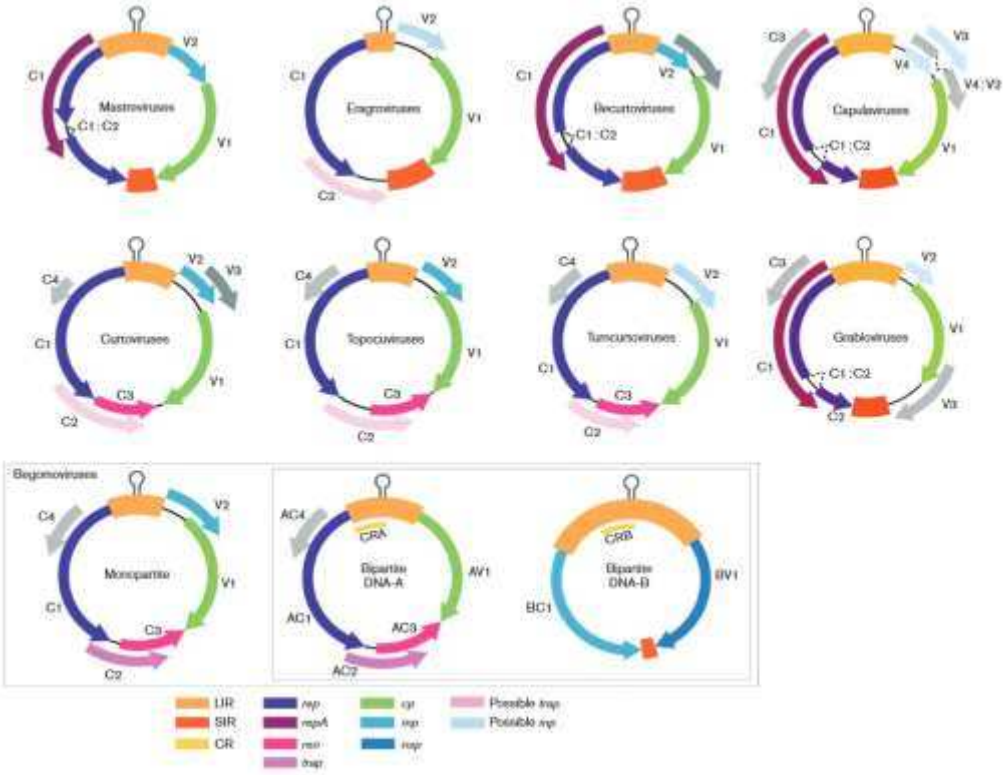


Figure 4. Genome organization of different genera in the family Geminiviridae (Zerbini et al., 2017).

Geminiviruses are transmitted by insects of the order Homoptera (aphids, leafhoppers, treehoppers and whiteflies). Members of the genera Mastrevirus, Curtovirus, Becurtovirus and Turncurtovirus are transmitted by specific leafhoppers, one member of the genus Capulavirus is transmitted by an aphid, members of the genera Grablovirus and Topocuvirus are transmitted by treehoppers, and members of the genus Begomovirus are transmitted by whiteflies. The vector of eragroviruses remains unknown (Zerbini et al., 2017).

Geminiviruses have high mutation, recombination and pseudo-recombination rates, which makes them exhibit a high degree of genetic diversity. Thus, there is a rapid adaptation to new hosts and environments, leading to a large distribution of these viruses

across diverse habitats. As they have the ability to infect cultivated and ornamental plants, as well as innumerable species of non-cultivated plants, geminiviruses cause high losses in crops worldwide. As an example, Tomato yellow leaf curl virus causes one of the viral diseases with greatest impact in the cultivation of tomatoes on a global scale. In Africa and Asia the geminiviruses Maize streak virus, East African cassava mosaic virus and Cotton leaf curl virus cause great losses in corn, cassava and cotton crops, respectively (Hanley-Bowdoin et al., 2013).

2.3. Taxonomic classification based on complete sequences

When a new organism is discovered, one of the first difficulties that a researcher encounters is to determine its taxonomic classification. In the dawn of microbiology, more than a century ago, classification was based mainly on morphology and biochemical characteristics. With the development of the first techniques of nucleic acid sequencing, a number of classification methods were stablished based on specific regions such as the 16S ribosomal RNA gene, which has been widely used in the classification of bacteria and archaea. However, as it is an extremely conserved region of the genome, sometimes it does not offers enough resolution to separate different species and even, in some cases, genera. In addition, this gene represents only a tiny fraction of the genomes of these organisms, greatly limiting the classification (Larsen et al., 2014).

With the constant evolution of HTS techniques, the volume of biological data available, such as whole-genome sequences (WGS), increases on a steady basis. These advances have revolutionized the study of the ecology and diversity of prokaryote communities in the last decades, and raised the interest of researchers in using these sequences to evaluate the taxonomic relationships between species. Already in the final years of the 20th century, the first attempts were made to use WGS for taxonomic purposes. A considerable number of different methods were proposed, which can be

divided between those that uses gene information and those that use nucleotide sequences (Coenye et al., 2005; Larsen et al., 2014; Luo et al., 2014; Muhire et al., 2014).

Of the methods using nucleotide sequences, the application of pairwise alignments to obtain percent identities is considered one of the most promising approaches, and has been increasingly used to classify viruses and bacteria within practical and functional operational taxonomic units (OTUs), such as order, family, subfamily, genus and species. In the case of viruses, the ICTV suggests, among other criteria, the utilization of WGS for the taxonomic classification of new viruses in approximately half of the existing viral families (Muhire et al., 2014; Varsani et al., 2017).

Thus, each year more studies are presented proposing the use of the percent identity of genomic sequences as a method of taxonomic classification for different organisms. Thompson et al. (2013) evaluated the use of Streptococcus genomic sequences in its taxonomic classification, using 67 sequences and obtaining positive results. Walter et al. (2017) proposed a new taxonomic classification framework for cyanobacteria, finding 32 new species and 19 new genera. Varsani and Krupovic (2017) also proposed a taxonomic classification framework for the family Genomoviridae based on the percent identity obtained from pairwise alignment of WGSs. Brown et al. (2015) carried out a taxonomic review of the genus Begomovirus, proposing new guidelines for the nomenclature and classification of these viruses, as well as new identity thresholds for the classification of new species and strains of this genus.

For many viruses with small genomes (with less than 30 kb), the methods using nucleotide sequence comparisons stand out for their efficiency. For example, in the family Geminiviridae the identity obtained from WGS pairwise alignments has been used since the mid-1990s with great precision. In order to analyze these viruses, as well as the viruses of the other genuses of the Geminiviridae (Mastrevirus, Curtovirus, Eragrovirus, Becurtovirus e Turncurtovirus), guidelines and protocols for demarcation and species

classification have already been proposed based on the identity obtained from pairwise alignment of WGSs, defining fixed taxonomic demarcation thresholds according to the already classified viruses (Muhire et al., 2013; Muhire et al., 2014; Varsani et al., 2014a; Varsani et al., 2014b; Brown et al., 2015).

A significant amount of tools have already been developed by both researchers and organizations (such as the National Center for Biotechnology Information, NCBI) to aid in different steps of genetic sequence analysis. For the analysis of metagenomics data, the homology-based tool MyTaxa (Luo et al., 2014) classifies and quantifies all genes present in a given sequence, to perform a maximum likelihood analysis which allows the choice of the most probable taxonomic classification of that sequence. The DEmARC tool (DivErsity pArtitioning by hieRarchical Clustering; Lauber and Gorbalenya, 2012) uses evolutionary distance calculations by pairing and multiple alignments of conserved sequences to determine taxonomic demarcation criteria, using phylogeny and the known taxonomy of the virus under study to validate the results.

More commonly used than DEmARC, PASC (PAirwise Sequence Comparison; Bao et al., 2012; Bao et al., 2014) was developed as an online tool to assist in the determination of taxonomic demarcation thresholds of viral families and genera. This method uses identity results from all published WGS from a viral family/genus, both from BLAST (Altschul et al. 1990) and from Needleman-Wunsch alignments (Needleman and Wunsch, 1970), to create distribution graphs of these identities. These graphs are usually composed of identity peaks, and when well separated, the points of least identity occurrence between different OTU groups can be used as taxonomic demarcation thresholds for the family/genus.

Despite their qualities, none of these tools is ideally suited to classify new viruses in a precise and consistent manner, since they act in specific and punctual stages of the taxonomic classification problem. Moreover, while DEmARC requires the careful

elaboration of multiple alignments, PASC uses the sequence in the exact way that it was deposited in the databases, which may be a problem in the case of circular genomes such as those of geminiviruses (Muhire et al., 2014).

Seeking to deal with this problem, Muhire et al. (2014) developed the Sequence Demarcation Tool (SDT), which uses Needleman-Wunsch pairwise alignments (disregarding sequence gaps) to determine identities between the viruses analyzed. As an output, SDT releases colored matrices according to the identities obtained, with the sequences ordered according to the degree of phylogenetic relationship between them, facilitating the analysis for the user. This tool is already used by several ICTV Study Groups. However, it is limited to punctually assessing the data inserted, thus, it is necessary to carefully construct the input files for each viral sequence that will be analyzed and to manually evaluate the results, classifying the viruses based on the predefined demarcation thresholds of each OTU.

Considering that the volume of WGSs deposited in the databases grows rapidly with the advancements in unbiased amplification and HTS technologies, there is a need for automation in the taxonomic classification process. This would reduce the time necessary for the analysis of new viruses and for determining taxonomic demarcation thresholds. Such demarcation thresholds should be flexible, adaptable according to the addition of new viruses in the different OTUs, especially for viral families of great economic importance such as the Geminiviridae. To this end, we present here the Automation Tool for Taxonomic Classification (AutoTaxa) of the family Geminiviridae.

3. METHODS

AutoTaxa was written in the Python 3.7 language, using Pycharm (https://www.jetbrains.com/pycharm/) as IDE. The libraries used included Biopython, NumPy, SciPy, MatPlotLib, Pycountry, XLWT and XLRD. Initially, all data was obtained and used to create data sets, and then the algorithm started to be written divided in parts as depicted in the workflow (Figure 5). In the first step the user chooses to insert the data directly (in the GenBank format) or if the algorithm will search the NCBI database. Also, the user chooses whether AutoTaxa will consider more than three sequential ambiguous nucleotides for parsing. After that AutoTaxa works by itself, and may take some time to parse the new sequences depending on the quantity. First the information obtained is processed to the correct standard required. With the remaining sequences that represent new geminiviruses (DNA-A only for bipartite begomoviruses), the algorithm generates pairwise alignments with previously classified viruses and uses the maximum identity found to classify the new virus. An Excel file is then released with the taxonomic classifications. The program also updates the taxonomic classification thresholds based on both the new viruses and the previously classified ones.

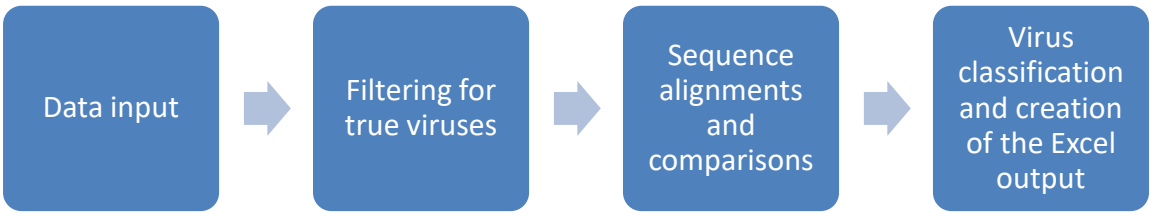


Figure 5. AutoTaxa workflow. In the first step the data is inserted either directly by the user or downloaded from the NCBI database. Then, all sequences obtained are filtered to remove those that do not possess the features found in viruses from the family *Geminiviridae*. The remaining sequences are aligned in a pairwise fashion with the previously classified viruses. Percent identities are calculated and used to classify the viruses based on the maximum identity found. After all viruses are parsed, an Excel file is generated with the new viruses' classification.

3.1. Data retrieval and new sequences input

All genomic sequence data annotated as belonging to the family Geminiviridae available at the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore) as of January 2017 was used for the data set assembly (these sequences will be hereafter referred to as "the reference data set"). The data retrieved in this step were manually analyzed to improve the program, with cycles of running the program, checking for errors in the classification in comparison with known information and correcting them for a new cycle. The genomic sequence data from that date until July 2018 (hereafter referred to as "the query data set") was used to validate the program and to demonstrate AutoTaxa's operation and output. Both the reference and the query data sets were downloaded as files in the GenBank standard format. Also, a taxonomic classification table containing the genera, species and strains recognized by the ICTV as of January 2017, prepared by the Geminiviridae Study Group of the ICTV, was used as a basis for the reference data set assembly, and also used as the standard for the AutoTaxa output.

3.2. Extraction, organization and validation of the data sets

As the parameter used for the NCBI search was "Geminiviridae" (for a broad approach, not excluding any sequences annotated as belonging to the Geminiviridae), many of the sequences initially assembled in the reference and query data sets consisted of DNA-B components, genes or gene fragments, geminivirus-associated DNA satellites, etc. Since the taxonomy of the family is based on the DNA-A, both data sets had to be filtered. This was done while the data was extracted from the GenBank file, by selecting only the sequences which had a length between 2500 and 3300 nucleotides, thus excluding genes/gene fragments and DNA satellites. Also, to avoid including DNA-B segments in the data sets, the algorithm searched for patterns in the sequence annotation that demarcates these segments, such as "DNA-B", "B component", "B segment", "nuclear shuttle", "BR-1", "BV-1" or "NSP" (for the DNA-B-encoded nuclear shuttle protein) and "movement protein", "BL-1", "BC-1" or "MP" (for the DNA-B-encoded movement protein).

If the sequence corresponded to a full-length DNA-A segments, AutoTaxa saved the information related to that sequence, including its accession numbers, genus, species name, the isolate name given by the researcher, the collection date and location, and the nucleotide sequence itself.

Since the sequences deposited in GenBank may not always follow the established standards used by the ICTV for the taxonomy of the Geminiviridae family, specially the site that marks the "beginning" of the circular genome, it was necessary to develop steps in the algorithm for the homogenization of the available information. For that, AutoTaxa searched the sequence for the conserved nonanucleotide located at the origin if replication (5'-TAATATTAC-3' in the case of begomoviruses, curtoviruses, mastreviruses, turncurtoviruses, topocuviruses, grabloviruses and capulaviruses, and 5'-TAAGATTCC-3' in the case of eragroviruses and becurtoviruses). When it was found, AutoTaxa broke the nonanucleotide between the 7^{th} and 8^{th} nucleotide position and then reorganized the

sequence to begin with "AC" (or "CC") and finish with "TAATATT" (or "TAAGATT"). Also, in rare occurrences the nonanucleotide may have an altered base or the sequence may be the complementary (or anti-sense) strand. Thus, if the nonanucleotide was not found in the entire sequence, it was necessary to search for a nonanucleotide with one changed base, preventing that the lack of a standard sequence would negatively affect the alignment. When a different nonanucleotide was found, this information, along with all other sequence information listed in the previous paragraph, was saved in a file in the .txt format named with the sequence accession number followed by log (e.g., AB007990 log; Supplementary figure S1).

3.3. Pairwise alignments and sequence comparisons

For the determination of the percent identities between each sequence in the query data set and the sequences in the reference data set, AutoTaxa employed the same methodology used by the SDT program, which is recommended for such analysis by the Geminiviridae Study Group of the ICTV. The program used the MUSCLE algorithm (Edgar, 2004) to perform true pairwise alignments (not multiple sequence alignments), and the percent identity was calculated by comparing nucleotides at the same position. Gaps introduced by the alignment algorithm were not considered, a logic which is also applied in the calculation of evolutionary distances for the construction of phylogenetic trees. After that, both the percent identity value and the corresponding accession number were saved. At the end of all pairwise comparisons for each sequence in the query data set, AutoTaxa searches for the sequence in the reference data set with the maximum identity, saving the alignment with the positions of nucleotide similarity between the sequences marked with asterisks and the positions of dissimilarity marked with empty spaces. If the query sequence had 100% identity with a sequence in the reference data set, it was removed from the query data set and the accession number was saved, so the algorithm doesn't parse it

again in the future. All identities obtained and the alignment were saved in the log file mentioned above.

3.4. Optimization of the time required for sequence comparisons

Considering the large number of geminiviruses already sequenced, the number of pairwise alignments to be performed reaches millions, require non-feasible analysis times. Thus, it was necessary to elaborate methodologies within AutoTaxa to filter the sequences that will be aligned and compared, avoiding unnecessary comparisons against all other sequences. For this, in a first step, only alignments against one sequence for each species within each genus were performed. In a second step, comparisons against all sequences within the species which had the highest percent identity were performed, hugely reducing the number of comparisons performed. Furthermore, when a percent identity value was above the strain demarcation threshold, the algorithm performed the comparisons only against the viruses within that particular species to check which member of the species had the highest overall identity, as suggested by Brown *et al.* (2015) as a conflict-resolution criteria.

3.5. Classification and nomenclature applied to new viruses

After identifying the sequence from the reference data set with the highest percent identity with a given sequence in the query data set, AutoTaxa compared the identity value with the demarcation threshold for the genus, determining whether the query sequence was classified as a new species, as a new strain of an existing species, or as a new isolate of an existing strain of an existing species.

In the case of new species, the algorithm used the species name as suggested by the researcher who submitted the sequence, and the sequence was inserted in the "new species" list in alphabetical order of the species name. In the case of a new strain, the sequence

received the corresponding species name, but the strain name depended on the standard strain name for the genus. If the standard includes a geographical location and year of collection followed by a bar and a letter (e.g., "Mexico-Yucatan-2004/A"), AutoTaxa checked for the currently used letters and then used the strain name followed by the next alphabetical letter (adding a second letter, as in "AA", when reaching "Z"). If the strain name doesn't follow any particular standard, AutoTaxa added the strain name as found in the sequence annotation. The same procedure was followed for new isolates.

3.6. Revision of species and strain demarcation thresholds for each genus

As new sequences are deposited in public databases, there is a steady increase in the number of sequences classified in each genera. Therefore, the species and strain demarcation thresholds may become outdated, requiring periodical revisions. To address this issue, a second part of AutoTaxa was developed to automate threshold acquisition. For that, the algorithm used the standard way applied in previous works, where it generated frequency graphs of the percent identity values obtained for each pairwise alignment of all sequences within each genus, and then determined the point of minimum occurrence of percent identities (the "valleys"). These graphs can be analyzed and the points can be compared to the previous demarcation threshold for that genus, easily allowing the selection of new demarcation thresholds.

4. RESULTS AND DISCUSSION

4.1. Validation

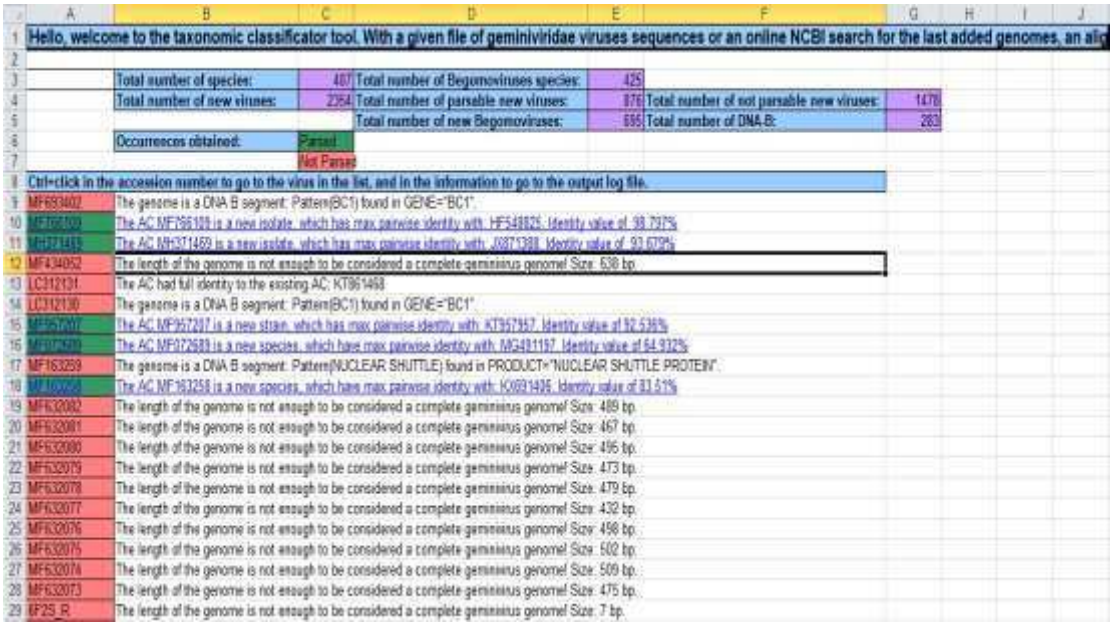
At first, the algorithm was validated by comparing the AutoTaxa classification of the reference data set with the existing classification for the same viruses present in the ICTV report. This was done by randomly dividing the 464 already classified viruses into five groups (four groups with 93 viruses and one group with 92 viruses) and then parsing each group separately, using the viruses in the other groups as the basis for AutoTaxa to compare. As a result, all viruses were classified by the algorithm in the same species in which they were classified by the ICTV, with the only difference being in the species name nomenclature adopted, where AutoTaxa used as species name the suggestion given by the researcher that first reported the virus.

4.2. Algorithm output

For the taxonomic classification, after all sequences in the query data set were classified, AutoTaxa released an Excel file containing eight sheets. The first sheet (Figure 6) contains all sequences in the initial query data set. The first column lists their accession numbers, with the ones that were not included in the final query data set (DNA-B, genes and gene fragments, etc.) marked in red, and the sequences included in the final query data set marked in green. The second column lists the highest percent identity value and the classification, or the reason that caused the exclusion of a given sequence from the final query data set. The second and the third sheets contain the full taxonomic list with the newly added sequences. In the third sheet (Supplementary figure S2) the newly added sequences are marked in yellow to facilitate their identification in the taxonomic list. In the fourth sheet (Supplementary figure S3) each line is a species representative with its information, followed by the number of sequences classified within the species and the

isolates that some strains may have. The fifth sheet (Figure 7) contains all sequences from the query data set that were added to the taxonomic list, with different shades of green marking whether they correspond to new species, new strains or new isolates. The last three sheets have only the new sequences that fit in the sheet's specific type of classification: new species in the sixth sheet (Supplementary figure S4), new strains in the seventh sheet and new isolates in the eighth sheet.

For the section of the algorithm that performs the update of species and strain demarcation thresholds, AutoTaxa releases a graph for each genus, where in the vertical axis are the proportion of pairwise identities and in the horizontal axis are the pairwise identity percentages.



	A	B	C	D	E	F	G
1	Hello, welcome to the taxonomic classificator tool. With a given file of geminiviridae viruses sequences or an online NCBI search for the last added genomes, an alig						
2							
3		Total number of species:	407	Total number of Begomoviruses species:	425		
4		Total number of new viruses:	2264	Total number of parsable new viruses:	876	Total number of not parsable new viruses:	1478
5				Total number of new Begomoviruses:	695	Total number of DNA-B:	283
6		Occurrences obtained:	Parsed				
7			Not Parsed				
8	Ctrl+click in the accesion number to go to the virus in the list, and in the information to go to the output log file.						
9	MF693402	The genome is a DNA B segment: Pattern(BC1) found in GENE="BC1".					
10	MF766109	The AC MF766109 is a new isolate, which has max pairwise identity with: HF548826. Identity value of: 99.797%					
11	MH371469	The AC MH371469 is a new isolate, which has max pairwise identity with: J0871388. Identity value of: 93.079%					
12	MF434052	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 638 bp.					
13	LC312131	The AC had full identity to the existing AC: KT861468					
14	LC312130	The genome is a DNA B segment: Pattern(BC1) found in GENE="BC1".					
15	MF957207	The AC MF957207 is a new strain, which has max pairwise identity with: KT957357. Identity value of: 92.536%					
16	MF072683	The AC MF072683 is a new species, which have max pairwise identity with: MG491197. Identity value of: 64.932%					
17	MF163259	The genome is a DNA B segment: Pattern(NUCLEAR SHUTTLE) found in PRODUCT="NUCLEAR SHUTTLE PROTEIN".					
18	MF163258	The AC MF163258 is a new species, which have max pairwise identity with: KC693436. Identity value of: 83.51%					
19	MF632082	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 489 bp.					
20	MF632081	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 467 bp.					
21	MF632080	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 495 bp.					
22	MF632079	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 473 bp.					
23	MF632078	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 479 bp.					
24	MF632077	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 432 bp.					
25	MF632076	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 498 bp.					
26	MF632075	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 502 bp.					
27	MF632074	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 509 bp.					
28	MF632073	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 475 bp.					
29	KF25_R	The length of the genome is not enough to be considered a complete geminivirus genome! Size: 7 bp.					

Figure 6. First sheet of the Excel output file: parsing occurrences log indicating sequences which were not included in the final query data set as they were shorther than a complete geminivirus genomic component (lines 19-29) or were annotated as DNA-B (lines 9 and 14).

Figure 7. Fifth sheet of the Excel output file: all viruses that were added to the query data set and their taxonomic classification.

4.3. Taxonomic classification of new geminiviruses

Analyses of all geminiviruses added to the GenBank database after January 2017 was performed as a demonstration of AutoTaxa's operation and output. There were 2354 new geminivirus sequences available, from which 876 were within the expected size for a full-length genomic component and were annotated as DNA-A, comprising the query data set (Figure 6). As the detection of sequence size and component annotation occurs in the extraction step of the algorithm, it was possible to avoid the inclusion of short sequences, and of DNA-B components, in the data set. This saved time in subsequent steps, specially sequence alignments. Although several kinds of patterns were added to AutoTaxa seeking to avoid the inclusion of DNA-B components, is still possible that some of them went unnoticed. Since the annotation of sequences in GenBank is primarily done by the submitter, some may not use the standard nomenclature to define the two components, or may add a complex definition that involves words resembling the patterns that are searched, thus generating false positive DNA-A detection. This problem is difficult to

solve, since GenBank will only accept corrections in submitted sequences when done by the submitter.

After AutoTaxa parsed all sequences, 495 of them were identical (100% sequence identity) to viruses classified as members of previously existing species, and therefore were removed from the taxonomic list. From the remaining 381 sequences, 27 corresponded to new species: 18 new begomoviruses, three new mastreviruses, three new capulaviruses, two new grabloviruses, and one new becurtovirus (Figure 8). An additional 20 sequences corresponded to new strains of existing species, and 334 sequences corresponded to new isolates of previously existing species, where almost all of them were new begomoviruses (Figure 7), as expected, since it is the most studied genus, with greater economic importance.

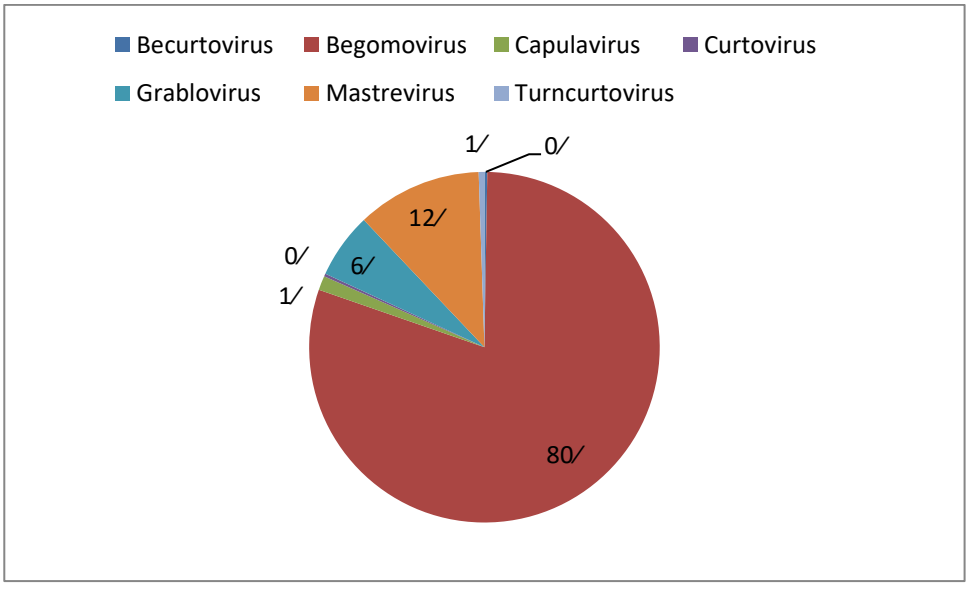


Figure 8. Genus assignment of the 381 sequences from the query data set that were different from sequences in the reference data set.

4.4. Determination of taxonomic demarcation thresholds

As mentioned before, as new viruses are classified into the existing genera, the increase in the number of species in a genus may necessite a revision of the species and strain demarcation thresholds. Thus, after classification, the second part of AutoTaxa was designed to create graphical depictions of the proportion of pairwise identities within each genus (with percent identities rounded up as whole numbers).

4.4.1. Genus *Becurtovirus*

There are currently two species in the genus *Becurtovirus*, *Spinach curly top Arizona virus* and *Beet curly top Iran virus* (BCTIV), with the latter subdivided into several strains. The algorithm added a third species, named *Exomis microphylla associated virus*, which had the highest identity (73.642%) with a strain of BCTIV.

The species and strain demarcation thresholds assigned to this genus are 80% and 94%, respectively (Varsani *et al.,* 2014b; Figure 9), defined based on a data set of 29 complete genome sequences. With the five recently added sequences, it was possible to recalculate these demarcation thresholds. In the new distribution of pairwise identities (Figure 10) it is noticeable that there are valleys between 74 and 76%, between 84 and 85%, and between 92 and 93%. We propose new species and strain demarcation thresholds of 85% and 93%, respectively. This result demonstrates that even the addition of a few new members to a genus is capable of changing the demarcation values, which may be a consequence of the small number of members in this genus to begin with. This finding reinforces the importance of constant updates to these demarcation values.

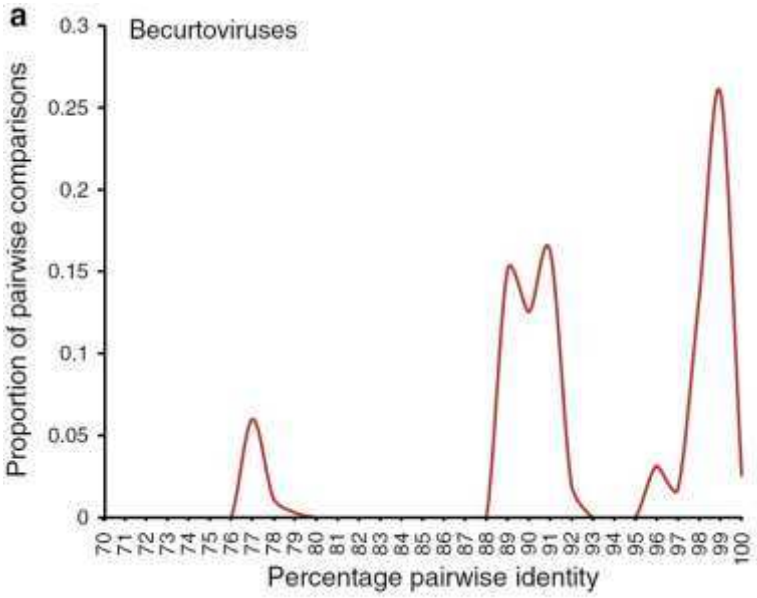


Figure 9. Previous distribution of pairwise identities for the genus *Becurtovirus* (Varsani *et al.,* 2014b).

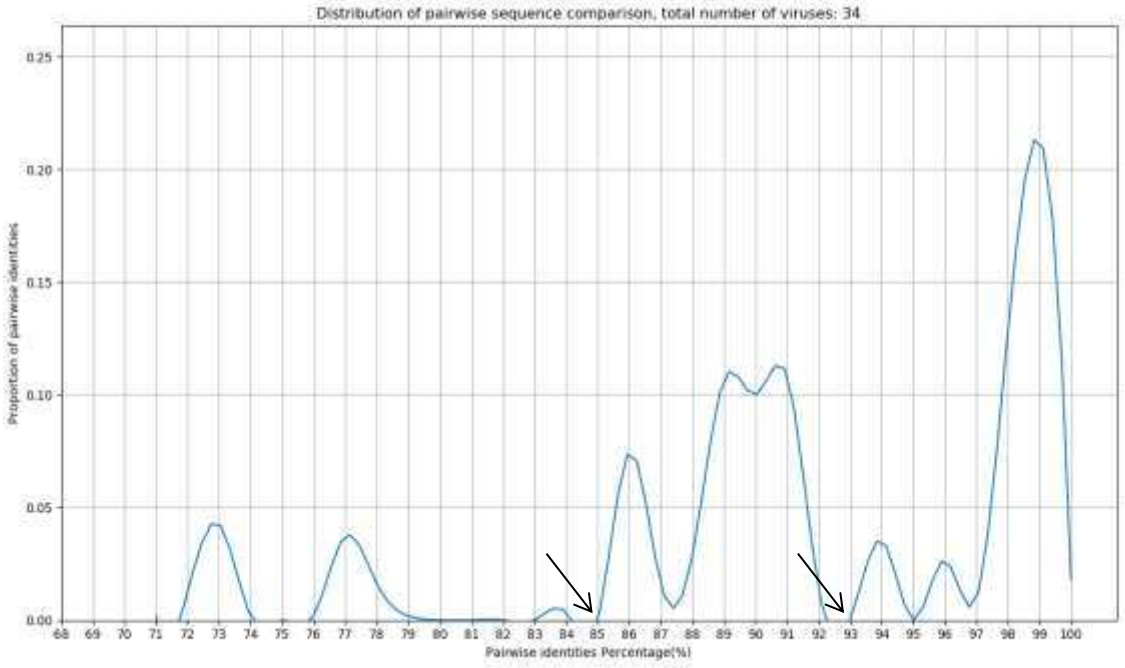


Figure 10. New distribution of pairwise identities for the genus *Becurtovirus* based on the new members (species and strains) of the genus.

4.4.2. Genus *Begomovirus*

The genus Begomovirus is the largest of all virology, with a huge quantity of begomoviruses already sequenced. Thus, even with the addition of 695 new members (strains or species), the demarcation thresholds suggested by Brown et al. (2015) held up, suggesting that this genus may have achieved stable species and strain demarcation thresholds.

4.4.3. Genus *Capulavirus*

Capulaviruses comprise a new genus (created in 2017) with only four species. The algorithm described here was able to identify three additional species to the genus, and increased the number of sequences corresponding to members of this genus from 47 to 51. The species and strain demarcation thresholds suggested by Varsani *et al.* (2017) (Figure 11) are 78% and 94% respectively. These values were obtained using a data set with 47 sequences, and the addition of only four sequences affected the distribution or pairwise identifies (Figure 12). The new distribution reveals pairwise identity valleys between 68 and 69%, between 74 and 75%, between 80 and 81%, between 87 and 90% and between 92 and 93%. We propose new species and strain demarcation thresholds of 81% and 93%, respectively.

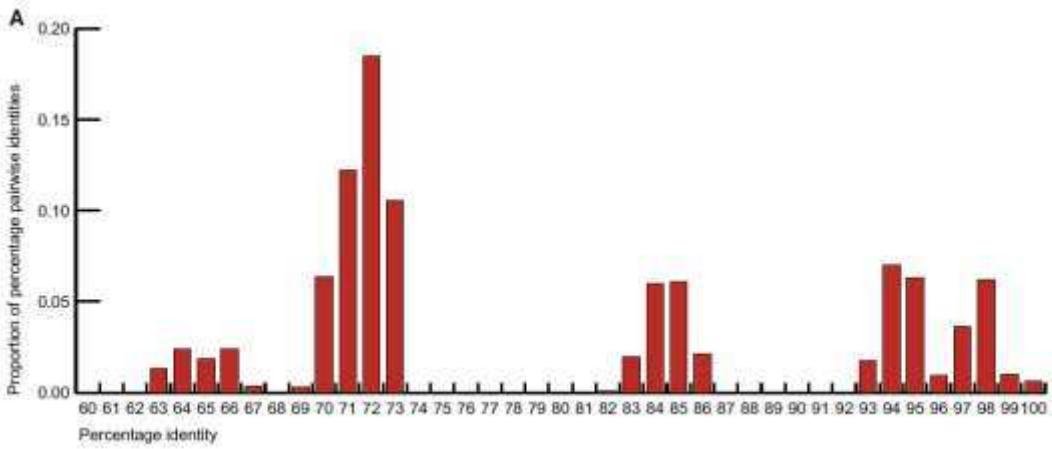


Figure 11. Previous distribution of pairwise identities for the genus *Capulavirus* (Varsani *et al.*, 2017).

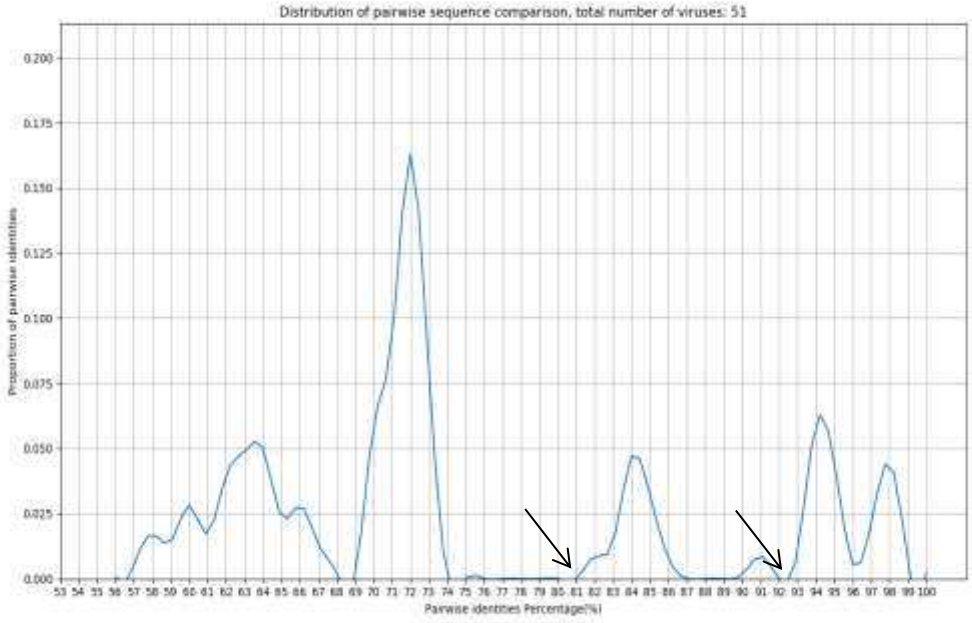


Figure 12. New distribution of pairwise identities for the genus *Capulavirus* based on the new members (species and strains) of the genus.

4.4.4. Genus *Curtovirus*

The species and strain demarcation thresholds for the genus *Curtovirus* were last revised by Varsani *et al.* (2014a) (Figure 13), and were set at 77% and 94%, respectively. However, the analyses were performed using only the 19 complete sequences available at that time. Nowadays, 110 curtovirus sequences are deposited in GenBank according to AutoTaxa's output. This allows the definition of more precise demarcation thresholds. In the new distribution of pairwise identities (Figure 14) it is possible to identify valleys between 60 and 66%, between 69 and 70%, at 77, between 86 and 87% and at 92%. As the 77% value is the same proposed before, we propose only the adjustment of the strain demarcation threshold to 92%, since 94% exhibits higher proportion values.

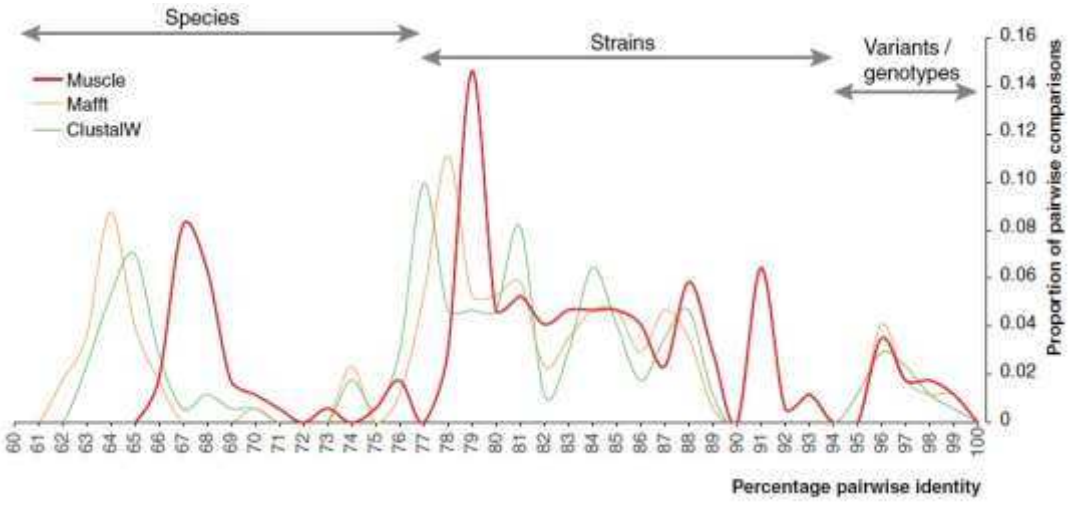


Figure 13. Previous distribution of pairwise identities for the genus Curtovirus (Varsani et al., 2014a).

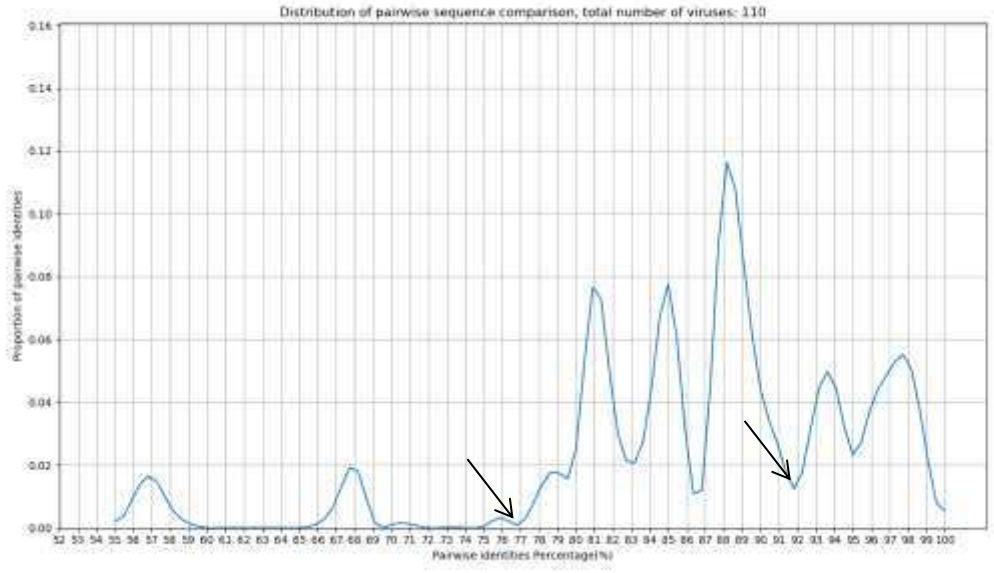


Figure 14. New distribution of pairwise identities for the genus *Curtovirus* based on the new members (species and strains) of the genus.

4.4.5. Genus *Eragrovirus*

As the genus *Eragrovirus* consists only of one species and no new member was classified by AutoTaxa, the pairwise identity distribution obtained for the genus (Figure 15) presented the same valleys identified by Varsani *et al.* (2014b). Therefore, the strain demarcation value suggested remains up to date.

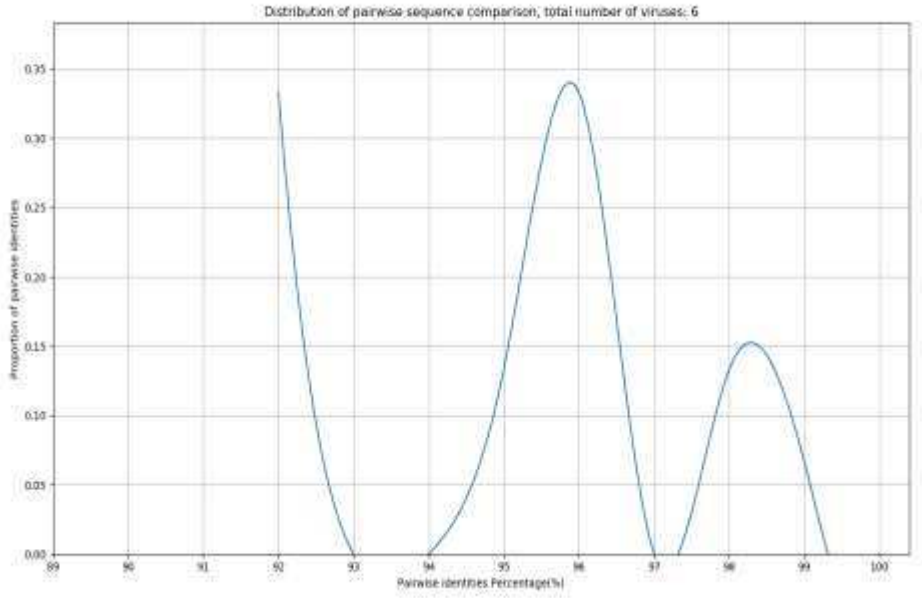


Figure 15. Distribution of pairwise identities for the genus *Eragrovirus* obtained by AutoTaxa.

4.4.6. Genus *Grablovirus*

The genus *Grablovirus* is also new (created in 2017), and consists of only one species. Furthermore, as there were only 27 complete genome sequences of grabloviruses, the species and strain demarcation thresholds could not be calculated properly. Varsani *et al.* (2017) (Figure 16) proposed the usage of demarcation thresholds similar to those of other geminivirus genera, with a species demarcation value of 80%. With the new grabloviruses added by AutoTaxa it was possible to identify valleys between 60 and 67%, between 70 and 72%, between 75 and 91% and between 93 and 95%. Based on these results, we propose new species and strain demarcation thresholds of 91% and 95%, respectively.

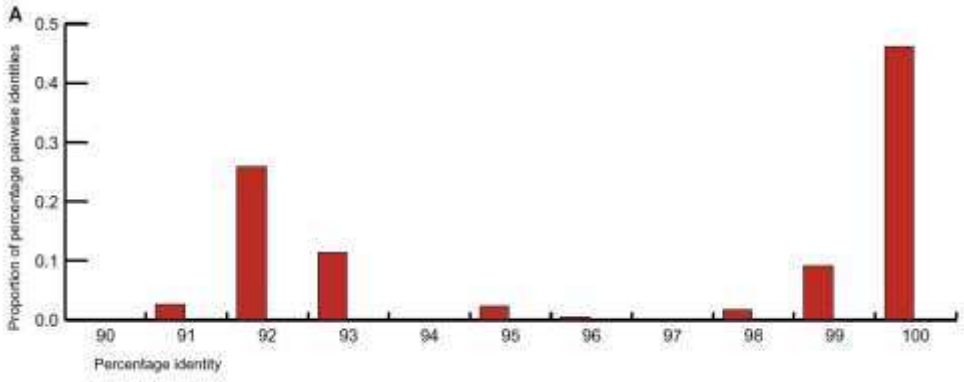


Figure 16. Previousistribution of pairwise identities for the genus *Grablovirus* (Varsani *et al.*, 2017).

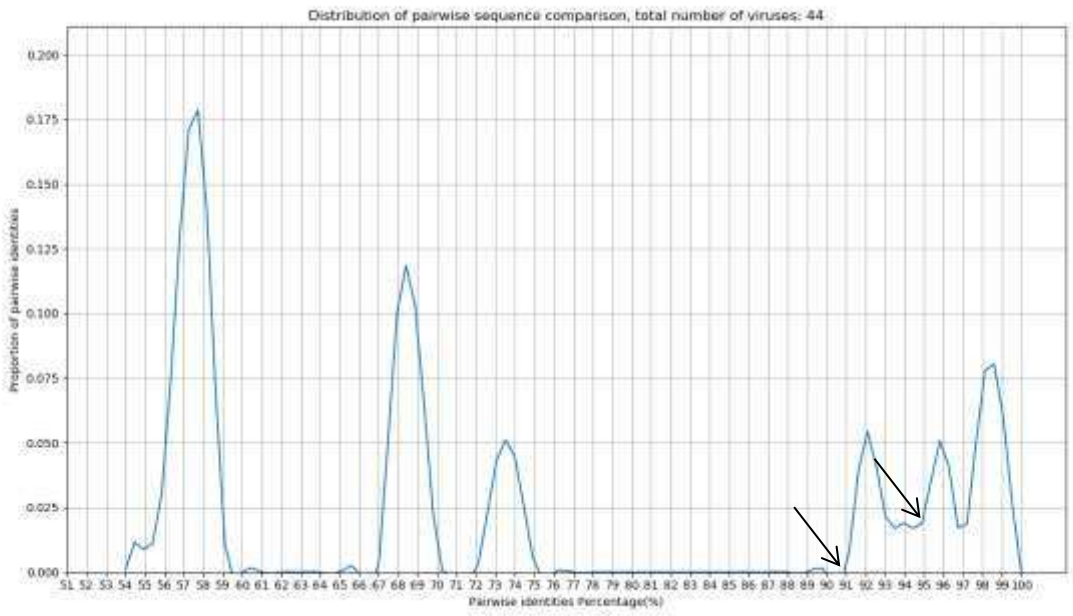


Figure 17. New distribution of pairwise identities for the genus *Grablovirus* based on the new members (species and strains) of the genus.

4.3.7. Genus *Mastrevirus*

The genus *Mastrevirus* is the second largest genus in the family *Geminiviridae*. This genus' demarcation thresholds (78% and 94% for species and strains, respectively) were proposed by Muhire *et al.* (2014) (Figure 18), and have not been updated since then.

With the addition of a large number of new sequences, a data set consisting of 1626 members was assembled. Based on the new distribution of pairwise identities (Figure 19) it was possible to identify valleys between 63 and 66%, between 71 and 79%, at 87%, and between 91 and 94%. Based in the previously defined values, we propose that only the species demarcation thresholds could be updated to 79%.

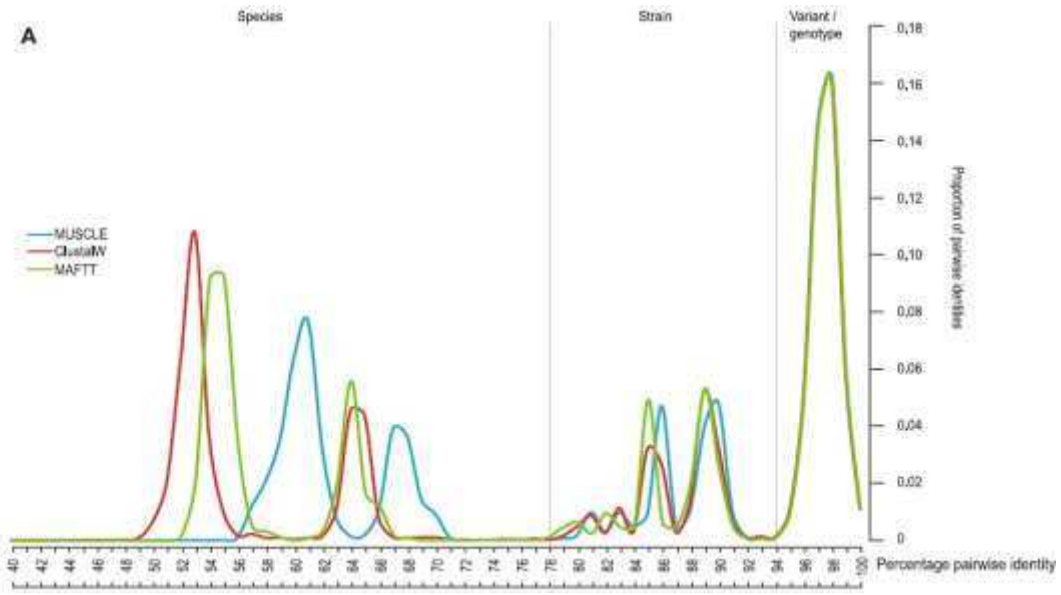


Figure 18. Previous distribution of pairwise identities for the genus *Mastrevirus* (Muhire *et al.*, 2014)

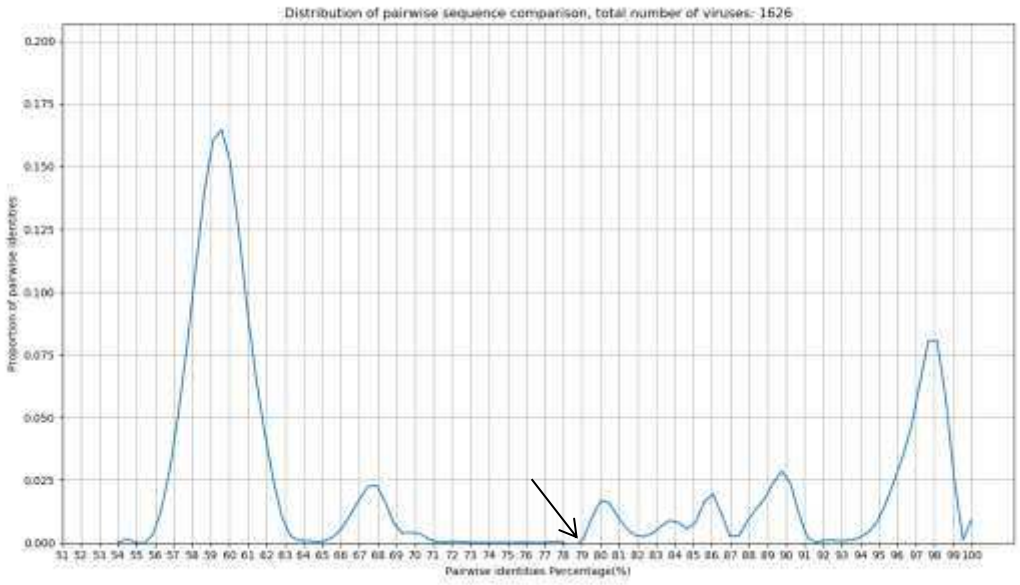


Figure 19. New distribution of pairwise identities for the genus *Mastrevirus* based on the new members (species and strains) of the genus.

4.4.8. Genus *Topocuvirus*

This genus consists of only one species with a single sequence deposited in GenBank. AutoTaxa did not identify any new sequences corresponding to species or strains in this genus. Therefore, it was not possible to generate pairwise alignments to plot a distribution graph.

4.4.9. Genus *Turncurtovirus*

The genus *Turncurtovirus* consists of one species, with 41 sequences deposited in GenBank. A strain demarcation threshold of 95% was proposed by Varsani *et al.* (2014b) (Figure 20). Even with the addition of new sequences to the data set, the distribution of pairwise identities (Figure 21) exhibited valleys which are very similar to those identified by Varsani *et al.* (2014b). Therefore, we propose to maintain the strain demarcation value of 95%.

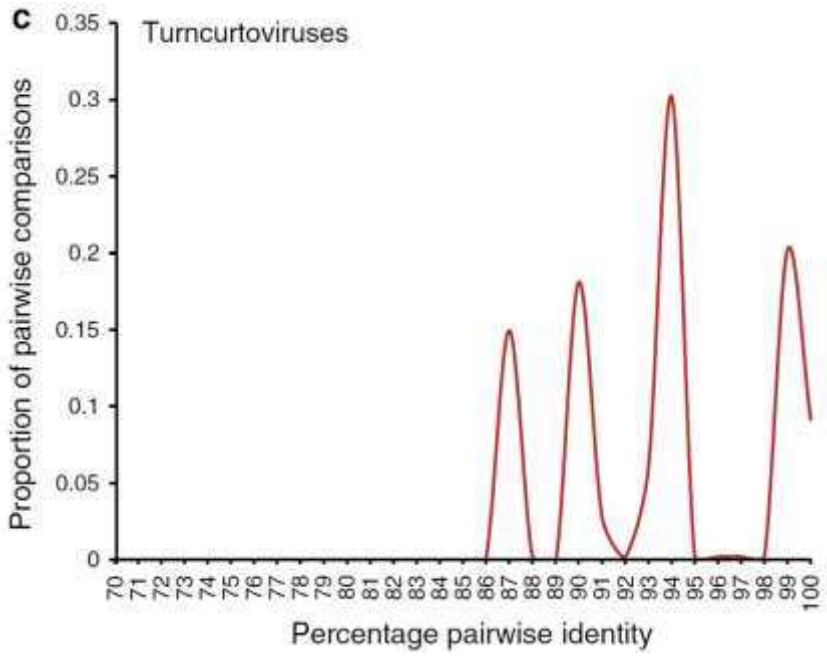


Figure 20. Previous distribution of pairwise identities for the genus *Turncurtovirus* (Varsani *et al.*, 2014b).

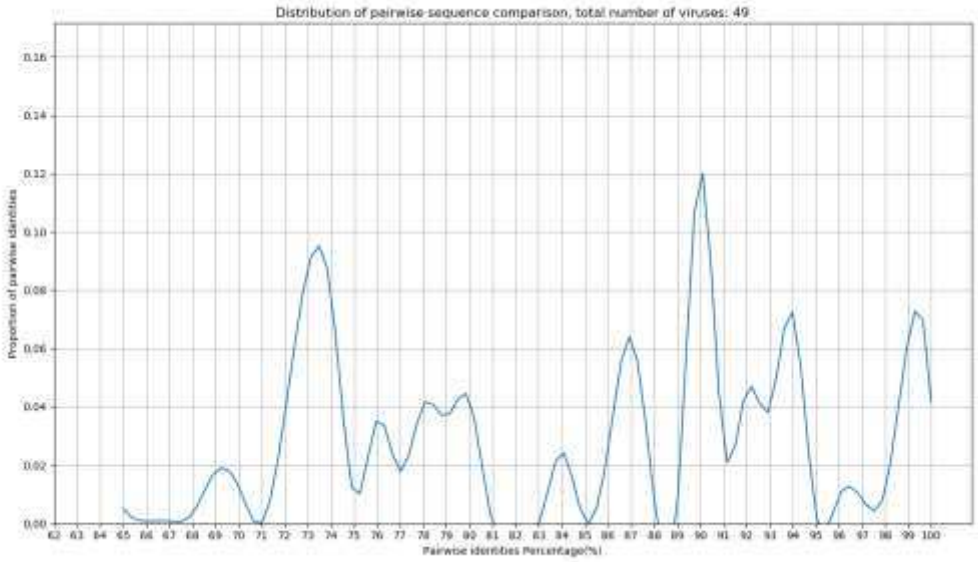


Figure 21. New distribution of pairwise identities for the genus *Turncurtovirus* based on the new members (species and strains) of the genus.

5. LITERATURE CITED

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410.

Bao Y, Chetvernin V, Tatusova T (2012) PAirwise Sequence Comparison (PASC) and its application in the classification of filoviruses. Viruses 4:1318-1327.

Bao Y, Chetvernin V, Tatusova T (2014) Improvements to pairwise sequence comparison (PASC): a genome-based web tool for virus classification. Arch Virol 159:3293-3304.

Briddon RW, Patil BL, Bagewadi B, Nawaz-ul-Rehman MS, Fauquet CM (2010) Distinct evolutionary histories of the DNA-A and DNA-B components of bipartite begomoviruses. BMC Evol Biol 10:1.

Brown JK, Zerbini FM, Navas-Castillo J, Moriones E, Ramos-Sobrinho R, et al. (2015) Revision of Begomovirus taxonomy based on pairwise sequence comparisons. Arch Virol 160:1593-1619.

Coenye T, Gevers D, Van de Peer Y, Vandamme P, Swings J (2005) Towards a prokaryotic genomic taxonomy. FEMS Microbiol Rev 29:147-167.

Datta S, Budhauliya R, Das B, Chatterjee S, Vanlalhmuaka, Veer V (2015) Next-generation sequencing in clinical virology: discovery of new viruses. World J Virol 4: 265-276

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Fauquet CM, Briddon RW, Brown JK, Moriones E, Stanley J, Zerbini FM, Zhou X (2008) Geminivirus strain demarcation and nomenclature. Arch Virol 153:783-821

Hanley-Bowdoin L, Bejarano ER, Robertson D, Mansoor S (2013) Geminiviruses: masters at redirecting and reprogramming plant processes. Nat Rev Microbiol 11:777-88.

Inoue-Nagata AK, Albuquerque LC, Rocha WB, Nagata T (2004) A simple method for cloning the complete begomovirus genome using the bacteriophage phi29 DNA polymerase. J Virol Methods 116:209-211.

Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, Sicheritz-Pontén T, Aarestrup FM, Ussery DW, Lund O (2014) Benchmarking of methods for genomic taxonomy. J Clin Microbiol 52:1529-39.

Lauber C, Gorbalenya AE (2012) Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses. J Virol 86:3890-3904.

Luo C, Rodriguez-R LM, Konstantinidis KT (2014) MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. Nucleic Acids Res 42:e73.

Muhire B, Martin DP, Brown JK, Navas-Castillo J, Moriones E, Zerbini FM, Rivera-Bustamante R, Malathi V, Briddon RW, Varsani A (2013) A genome-wide pairwise-

identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). Arch Virol 158:1411-24.

Muhire BM, Varsani A, Martin DP (2014) SDT: A virus classification tool based on pairwise sequence alignment and identity calculation. PLoS ONE 9:e108277.

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443-453.

Rojas MR, Hagen C, Lucas WJ, Gilbertson RL (2005) Exploiting chinks in the plant's armor: Evolution and emergence of geminiviruses. Annu Rev Phytopathol 43:361-394.

Silva JCF, Carvalho TFM, Fontes EPB, Cerqueira FR (2017) Fangorn Forest (F2): a machine learning approach to classify genes and genera in the family Geminiviridae. BMC Bioinformatics 18:431.

Simmonds P (2015) Methods for virus classification and the challenge of incorporating metagenomic sequence data. J Gen Virol 96:1193-1206.

Thompson CC, Emmel VE, Fonseca EL, Marin MA, Vicente ACP (2013) Streptococcal taxonomy based on genome sequence analyses. F1000Research 2:67

Varsani A, Martin DP, Navas-Castillo J, Moriones E, Hernandez-Zepeda C, Idris A, Zerbini FM, Brown JK (2014a) Revisiting the classification of curtoviruses based on genome-wide pairwise identity. Arch Virol 159:1873-1882.

Varsani A, Navas-Castillo J, Moriones E, Hernández-Zepeda C, Idris A, Brown JK, Zerbini FM, Martin DP (2014b) Establishment of three new genera in the family Geminiviridae: Becurtovirus, Eragrovirus and Turncurtovirus. Arch Virol 159:2193-2203.

Varsani A, Roumagnac P, Fuchs M, Navas-Castillo J, Moriones E, Idris A, Briddon RW, Rivera-Bustamante R, Zerbini FM, Martin DP (2017) Capulavirus and Grablovirus: Two new genera in the Family Geminiviridae. Arch Virol 162:1819-1831.

Varsani A, Krupovic M (2017) Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. Virus Evol 3:vew037.

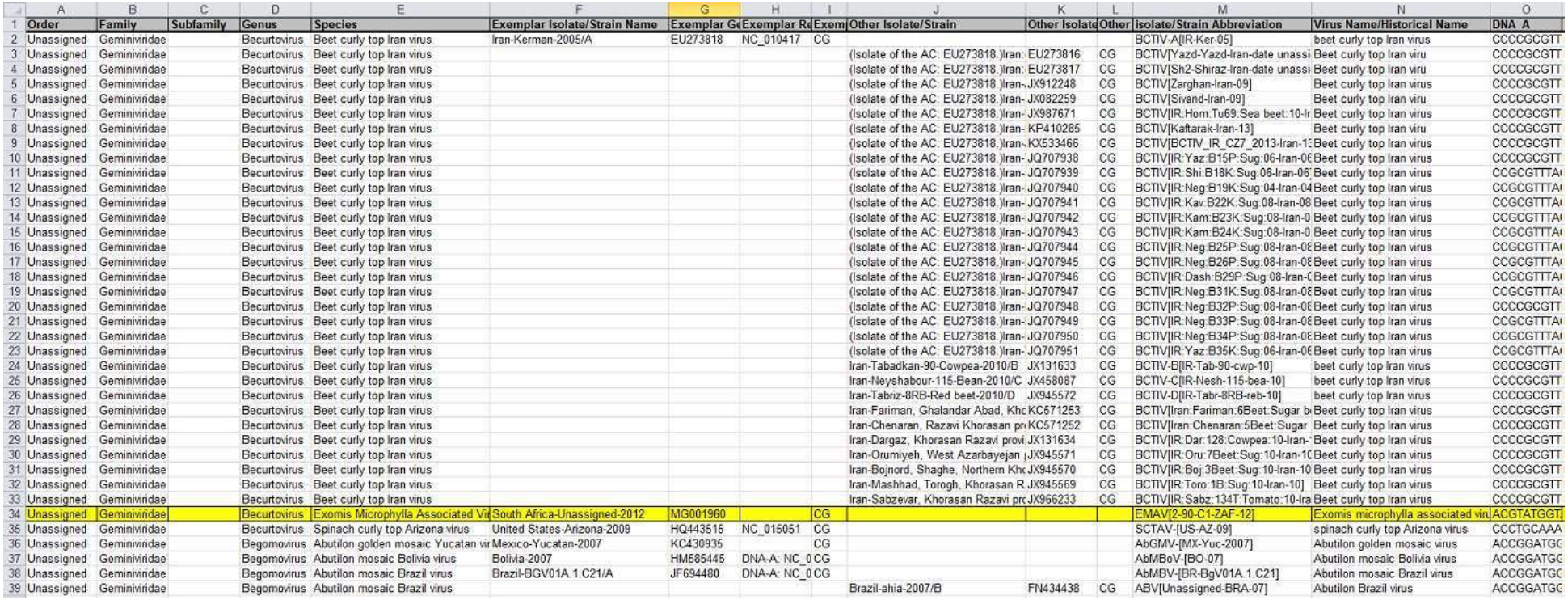
Walter JM, Coutinho FH, Dutilh BE, Thompson FL, Thompson CC (2017) Proposal of a new genome-based taxonomy for Cyanobacteria. PeerJ e2676v1.

Zhao G, Wu G, Lim ES, Droit L, Krishnamurthy S, Barouch DH, Virgin HW, Wang D (2017) VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. Virology 503:21-30.

Zerbini FM, Briddon RW, Idris A, Martin DP, Moriones E, Navas-Castillo J, Rivera-Bustamante R, Roumagnac P, Varsani A, ICTV Report Consortium (2017) ICTV Virus Taxonomy Profile: Geminiviridae. J Gen Virol 98:131-133.

6. APPENDIX

Supplementary figure S1. Example of output log for a virus, including the information of changes in its sequence, the identities obtained, the classification and the alignment with the virus with higher sequence identity.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Order	Family	Subfamily	Genus	Species	Exemplar Isolate/Strain Name	Exemplar Ge	Exemplar Re	Exem	Other Isolate/Strain	Other Isolate	Other	isolate/Strain Abbreviation	Virus Name/Historical Name	DNA_A
2	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus	Iran-Kerman-2005/A	EU273818	NC_010417	CG				BCTIV-A[IR-Ker-05]	beet curly top Iran virus	CCCCGCGTT
3	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran:	EU273816	CG	BCTIV[Yazd-Yazd-Iran-date unassi	Beet curly top Iran viru	CCCCGCGTT
4	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran:	EU273817	CG	BCTIV[Sh2-Shiraz-Iran-date unassi	Beet curly top Iran viru	CCCCGCGTT
5	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JX912248	CG	BCTIV[Zarghan-Iran-09]	Beet curly top Iran virus	CCCCGCGTT
6	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JX082259	CG	BCTIV[Sivand-Iran-09]	Beet curly top Iran viru	CCCCGCGTT
7	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JX987671	CG	BCTIV[IR:Hom:Tu69:Sea beet:10-Ir	Beet curly top Iran virus	CCCCGCGTT
8	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	KP410285	CG	BCTIV[Kaftarak-Iran-13]	Beet curly top Iran viru	CCCCGCGTT
9	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	KX533466	CG	BCTIV[BCTIV_IR_CZ7_2013-Iran-1	Beet curly top Iran virus	CCCCGCGTT
10	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707938	CG	BCTIV[IR:Yaz:B15P:Sug:06-Iran-0	Beet curly top Iran virus	CCCCGCGTT
11	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707939	CG	BCTIV[IR:Shi:B18K:Sug:06-Iran-06	Beet curly top Iran virus	CCGCGTTTA
12	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707940	CG	BCTIV[IR:Neg:B19K:Sug:04-Iran-04	Beet curly top Iran virus	CCGCGTTTA
13	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707941	CG	BCTIV[IR:Kav:B22K:Sug:08-Iran-08	Beet curly top Iran virus	CCGCGTTTA
14	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707942	CG	BCTIV[IR:Kam:B23K:Sug:08-Iran-0	Beet curly top Iran virus	CCGCGTTTA
15	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707943	CG	BCTIV[IR:Kam:B24K:Sug:08-Iran-0	Beet curly top Iran virus	CCGCGTTTA
16	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707944	CG	BCTIV[IR:Neg:B25P:Sug:08-Iran-08	Beet curly top Iran virus	CCGCGTTTA
17	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707945	CG	BCTIV[IR:Neg:B26P:Sug:08-Iran-08	Beet curly top Iran virus	CCGCGTTTA
18	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707946	CG	BCTIV[IR:Dash:B29P:Sug:08-Iran-(	Beet curly top Iran virus	CCGCGTTTA
19	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707947	CG	BCTIV[IR:Neg:B31K:Sug:08-Iran-08	Beet curly top Iran virus	CCGCGTTTA
20	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707948	CG	BCTIV[IR:Neg:B32P:Sug:08-Iran-08	Beet curly top Iran virus	CCCCGCGTT
21	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707949	CG	BCTIV[IR:Neg:B33P:Sug:08-Iran-08	Beet curly top Iran virus	CCGCGTTTA
22	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707950	CG	BCTIV[IR:Neg:B34P:Sug:08-Iran-08	Beet curly top Iran virus	CCGCGTTTA
23	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					(Isolate of the AC: EU273818.)Iran-	JQ707951	CG	BCTIV[IR:Yaz:B35K:Sug:06-Iran-06	Beet curly top Iran virus	CCGCGTTTA
24	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					Iran-Tabadkan-90-Cowpea-2010/B	JX131633	CG	BCTIV-B[IR-Tab-90-cwp-10]	beet curly top Iran virus	CCCCGCGTT
25	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					Iran-Neyshabour-115-Bean-2010/C	JX458087	CG	BCTIV-C[IR-Nesh-115-bea-10]	beet curly top Iran virus	CCCCGCGTT
26	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					Iran-Tabriz-8RB-Red beet-2010/D	JX945572	CG	BCTIV-D[IR-Tabr-8RB-reb-10]	beet curly top Iran virus	CCCCGCGTT
27	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					Iran-Fariman, Ghalandar Abad, Khc	KC571253	CG	BCTIV[Iran:Fariman:6Beet:Sugar b	Beet curly top Iran virus	CCCCGCGTT
28	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					Iran-Chenaran, Razavi Khorasan pr	KC571252	CG	BCTIV[Iran:Chenaran:5Beet:Sugar	Beet curly top Iran virus	CCCCGCGTT
29	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					Iran-Dargaz, Khorasan Razavi provi	JX131634	CG	BCTIV[IR:Dar:128:Cowpea:10-Iran-	Beet curly top Iran virus	CCCCGCGTT
30	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					Iran-Orumiyeh, West Azarbayejan p	JX945571	CG	BCTIV[IR:Oru:7Beet:Sug:10-Iran-1(	Beet curly top Iran virus	CCCCGCGTT
31	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					Iran-Bojnord, Shaghe, Northern Kho	JX945570	CG	BCTIV[IR:Boj:3Beet:Sug:10-Iran-10	Beet curly top Iran virus	CCCCGCGTT
32	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					Iran-Mashhad, Torogh, Khorasan R	JX945569	CG	BCTIV[IR:Toro:1B:Sug:10-Iran-10]	Beet curly top Iran virus	CCCCGCGTT
33	Unassigned	Geminiviridae		Becurtovirus	Beet curly top Iran virus					Iran-Sabzevar, Khorasan Razavi pro	JX966233	CG	BCTIV[IR:Sabz:134T:Tomato:10-Ira	Beet curly top Iran virus	CCCCGCGTT
34	Unassigned	Geminiviridae		Becurtovirus	Exomis Microphylla Associated Vir	South Africa-Unassigned-2012	MG001960		CG				EMAV[2-90-C1-ZAF-12]	Exomis microphylla associated viru	ACGTATGGT
35	Unassigned	Geminiviridae		Becurtovirus	Spinach curly top Arizona virus	United States-Arizona-2009	HQ443515	NC_015051	CG				SCTAV-[US-AZ-09]	spinach curly top Arizona virus	CCCTGCAAA
36	Unassigned	Geminiviridae		Begomovirus	Abutilon golden mosaic Yucatan vir	Mexico-Yucatan-2007	KC430935		CG				AbGMV-[MX-Yuc-2007]	Abutilon golden mosaic virus	ACCGGATGC
37	Unassigned	Geminiviridae		Begomovirus	Abutilon mosaic Bolivia virus	Bolivia-2007	HM585445	DNA-A: NC_0	CG				AbMBoV-[BO-07]	Abutilon mosaic Bolivia virus	ACCGGATGC
38	Unassigned	Geminiviridae		Begomovirus	Abutilon mosaic Brazil virus	Brazil-BGV01A.1.C21/A	JF694480	DNA-A: NC_0	CG				AbMBV-[BR-BgV01A.1.C21]	Abutilon mosaic Brazil virus	ACCGGATGC
39	Unassigned	Geminiviridae		Begomovirus	Abutilon mosaic Brazil virus					Brazil-ahia-2007/B	FN434438	CG	ABV[Unassigned-BRA-07]	Abutilon Brazil virus	ACCGGATGC

Supplementary figure S2. Third Excel sheet: complete taxonomic list with the sequences corresponding to new species/strains marked in yellow.

	A	B	C	D	E	F
1	Genus	Species	Exemplar Isolate/Strain Name	Exemplar Gen	Total number of strains/Isolate	Isolates
2	Becurtovirus	Beet curly top Iran virus	Iran-Kerman-2005/A	EU273818	31	The related AC: EU273818 have this isolates: [EU273816], [EU273817], [JX912248], [JX082259], [JX987671], [KP41
3	Becurtovirus	Exomis Microphylla Associated Vir	South Africa-Unassigned-2012	MG001960	1	
4	Becurtovirus	Spinach curly top Arizona virus	United States-Arizona-2009	HQ443515	1	
5	Begomovirus	Abutilon golden mosaic Yucatan vir	Mexico-Yucatan-2007	KC430935	1	
6	Begomovirus	Abutilon mosaic Bolivia virus	Bolivia-2007	HM585445	1	
7	Begomovirus	Abutilon mosaic Brazil virus	Brazil-BGV01A.1.C21/A	JF694480	4	The related AC: JF694480 have this isolates: [JF694481], [JF694482].
8	Begomovirus	Abutilon mosaic virus	Germany	X15983	7	The related AC: X15983 have this isolates: [LN611623], [HQ588899], [HQ588900], [HQ588901], [U51137].The relate
9	Begomovirus	African Cassava Mosaic Burkina Fa	Burkina Faso-uagadougou-2008	HE616777	1	
10	Begomovirus	African cassava mosaic virus	Cameroon-1998	J02057	179	The related AC: J02057 have this isolates: [KJ888082], [HG530110], [HG530111], [KJ887753], [KJ887754], [KJ8877
11	Begomovirus	Ageratum enation virus	Nepal-1999/Nepal	AJ437618	20	The related AC: AJ437618 have this isolates: [EU867513], [KJ488990], [KM262823].The related AC: EU867513 have
12	Begomovirus	Ageratum Leaf Curl Sichuan Virus	China-Sichuan-2016	MG917698	2	The related AC: MG917698 have this isolates: [MG917697].
13	Begomovirus	Ageratum leaf curl virus	China-Guangxi 52-2003	AJ851005	5	The related AC: AJ851005 have this isolates: [KU376491].The related AC: KU954387 have this isolates: [KJ016239]
14	Begomovirus	Ageratum yellow vein Hualian virus	Taiwan-Hualian 4-2000/A	DQ866132	18	The related AC: EF544600 have this isolates: [DQ866134], [EF458639], [X74516], [KC282641].The related AC: DQ8
15	Begomovirus	Ageratum yellow vein Sri Lanka viru	Sri Lanka-1999	AF314144	1	
16	Begomovirus	Ageratum yellow vein virus	China-Guangxi 129-2005/Guanxi	AJ495813	34	The related AC: AJ495813 have this isolates: [KF999980], [KF999981], [KU601622].The related AC: KF999981 have
17	Begomovirus	Allamanda leaf curl virus	China-Guandong 10-2006	EF602306	1	
18	Begomovirus	Allamanda Leaf Mottle Distortion V	India-Kalyani, West Bengal-2012	KC202818	1	
19	Begomovirus	Alternanthera yellow vein virus	China-Guangxi-G38-2005	AJ965540	22	The related AC: AJ965540 have this isolates: [KX710155], [AM050736], [DQ375456], [DQ641703], [EF544603], [EU
20	Begomovirus	Andrographis Yellow Vein Leaf Curl	India-Barabanki-2012	KM359406	1	
21	Begomovirus	Apple Geminivirus P	China-Unassigned-2013	KM386645	1	
22	Begomovirus	Asystasia Mosaic Madagascar Viru	Madagascar-Unassigned-2011	KP663485	2	The related AC: KP663485 have this isolates: [KP663483].
23	Begomovirus	Bean calico mosaic virus	Mexico-Sonora-1986	AF110189	1	
24	Begomovirus	Bean chlorosis virus	Venezuela-LaBarinesa459-2006	JN848770	1	
25	Begomovirus	Bean Chlorotic Mosaic Virus	Venezuela-Bhatinda, Punjab-2007	JN848772	1	
26	Begomovirus	Bean dwarf mosaic virus	Colombia-1987	M88179	1	
27	Begomovirus	Bean golden mosaic virus	Brazil-Campinas 1-1978	M88686	148	The related AC: M88686 have this isolates: [KJ939710], [KJ939711], [KJ939720], [KJ939766], [KJ939767], [KJ9397
28	Begomovirus	Bean golden yellow mosaic virus	country unassigned:date unassigne	L01635	9	The related AC: D00201 have this isolates: [AF173555], [AJ544531], [DQ119824], [M10070], [M91604].The related A
29	Begomovirus	Bean Leaf Crumple Virus	Colombia-Unassigned-2015	KX857725	1	
30	Begomovirus	Bean yellow mosaic Mexico virus	Mexico-06.05.11-2011	FJ944023	1	
31	Begomovirus	Bhendi yellow vein Bhubhaneswar v	India-Orissa-2003	FJ589571	1	
32	Begomovirus	Bhendi Yellow Vein Delhi Virus	India-New Delhi, Palem-2004	FJ515747	1	
33	Begomovirus	Bhendi Yellow Vein Haryana Virus	India-Karnal, Haryana-2006	KJ628309	1	
34	Begomovirus	Bhendi yellow vein mosaic virus	India-Madurai/India	AF241479	10	The related AC: AJ002451 have this isolates: [KC501923].The related AC: EU589392 have this isolates: [KJ462081]

Supplementary figure S3. Forth Excel sheet: list containing only the species representatives, with the number of members in each species and the isolates of some members in the species.

	A	B	D	E	F	G	I	M	N	O	P	Q	R
1	Order	Family	Genus	Species	Exemplar Isolate/Strain Name	Exemplar G	Exem	isolate/Strain Abl	Virus Name/Historical Name	DNA_A	DNA_B	parsed? (1 =	yes; 0 = n
2	Unassigned	Geminiviridae	Begomovirus	Malvastrum Yellow Vein Lahore Vir	Pakistan-Lahore-2013	MF683828	CG	MYVLV[J47-PAK-1	Malvastrum yellow vein Lahore virus	ACCGGATGC	DNA_B	The AC MF683828 is a n	
3	Unassigned	Geminiviridae	Begomovirus	Ageratum Leaf Curl Sichuan Virus	China-Sichuan-2016	MG917698	CG	ALCSV[SC782-CHI	Ageratum leaf curl Sichuan virus	ACCGGATGC	DNA_B	The AC MG917698 is a r	
4	Unassigned	Geminiviridae	Begomovirus	Tomato Leaf Curl Virus	India-Unassigned-2017	MF737344	CG	TLCV[CS1-IND-17]	Tomato leaf curl virus	ACCGGATGC	DNA_B	The AC MF737344 is a n	
5	Unassigned	Geminiviridae	Begomovirus	Hibiscus Vein Enation Virus	Taiwan-Changhua-2014	MF140455	CG	HVEV[Unassigned	Hibiscus vein enation virus	ACCGGATGC	DNA_B	The AC MF140455 is a n	
6	Unassigned	Geminiviridae	Begomovirus	Boerhavia Golden Mosaic Virus	Dominican Republic-Unassigned-2(	KY971539	CG	BGMV[DO/Azua/B	Boerhavia golden mosaic virus	ACCGGATGC	DNA_B	The AC KY971539 is a n	
7	Unassigned	Geminiviridae	Grablovirus	Prunus Geminivirus A	USA-Unassigned-2016	MF579394	CG	PGA[HTS-USA-16]	Prunus geminivirus A	ACCGGCCC(	DNA_B	The AC MF579394 is a n	
8	Unassigned	Geminiviridae	Capulavirus	Limeum Africanum Associated Viru	South Africa-Unassigned-2012	MG001961	CG	LAAV[2-12-F-ZAF-'	Limeum africanum associated virus	ACGGTTTTG(	DNA_B	The AC MG001961 is a r	
9	Unassigned	Geminiviridae	Becurtovirus	Exomis Microphylla Associated Vi	South Africa-Unassigned-2012	MG001960	CG	EMAV[2-90-C1-ZAI	Exomis microphylla associated viru	ACGTATGGT.	DNA_B	The AC MG001960 is a r	
10	Unassigned	Geminiviridae	Begomovirus	Polygala Garcinii Associated Virus	South Africa-Unassigned-2012	MG001959	CG	PGAV[1-1-ZAF-12]	Polygala garcinii associated virus	ACCGGATGC	DNA_B	The AC MG001959 is a r	
11	Unassigned	Geminiviridae	Begomovirus	Juncus Maritimus Associated Virus	France-Unassigned-2012	MG001958	CG	JMAV[13-FMN-1-FI	Juncus maritimus associated virus	ACCGGATGC	DNA_B	The AC MG001958 is a r	
12	Unassigned	Geminiviridae	Capulavirus	Tomato Apical Leaf Curl Virus	Argentina-Unassigned-2008	MG491197	CG	TALCV[AR:Yuto:Tc	Tomato apical leaf curl virus	ACGGCTTTG	DNA_B	The AC MG491197 is a r	
13	Unassigned	Geminiviridae	Begomovirus	Coccinia Mosaic Virudhunagar Viru	India-Unassigned-2016	KY860899	CG	CMVV[Unassigned	Coccinia mosaic Virudhunagar virus	ACCGGATGC	DNA_B	The AC KY860899 is a n	
14	Unassigned	Geminiviridae	Mastrevirus	Rice Latent Virus 2	Australia-Unassigned-2015	KY962381	CG	RLV2[AU-NA24-20	Rice latent virus 2	ACAGCGATC	DNA_B	The AC KY962381 is a n	
15	Unassigned	Geminiviridae	Mastrevirus	Rice Latent Virus 1	Australia-Unassigned-2015	KY962380	CG	RLV1[AU-NA70-20	Rice latent virus 1	ACCAGTGCC	DNA_B	The AC KY962380 is a n	
16	Unassigned	Geminiviridae	Begomovirus	Datura Leaf Curl Virus	Sudan-Unassigned-2016	MF402919	CG	DLCV[Sudan-Datur	Datura leaf curl virus	ACCGGATGC	DNA_B	The AC MF402919 is a n	
17	Unassigned	Geminiviridae	Mastrevirus	Maize Striate Mosaic Virus	Brazil-Unassigned-2016	MF167307	CG	MSMV[MSMV_BR	Maize striate mosaic virus	ACGCGCAC(	DNA_B	The AC MF167307 is a n	
18	Unassigned	Geminiviridae	Capulavirus	Tomato Associated Geminivirus 1	Brazil-Unassigned-2015	MF072689	CG	TAG1[Cleome_BR-	Tomato associated geminivirus 1	ACGGATGTC	DNA_B	The AC MF072689 is a n	
19	Unassigned	Geminiviridae	Begomovirus	Macroptilium Golden Yellow Mosai	Dominican Republic-Unassigned-2(	KY196219	CG	MGYMV[DR:M45:1	Macroptilium golden yellow mosaic	ACCGGATGC	DNA_B	The AC KY196219 is a n	
20	Unassigned	Geminiviridae	Begomovirus	Tomato Leaf Curl Purple Vein Virus	Brazil-Unassigned-2015	KY196221	CG	TLCPVV[BR:PD6:1	Tomato leaf curl purple vein virus	ACCGGATGC	DNA_B	The AC KY196221 is a n	
21	Unassigned	Geminiviridae	Grablovirus	Wild Vitis Virus 1	USA-Unassigned-2015	MF185010	CG	WVV1[WVV1-NY1	Wild vitis virus 1	ACCGGCCC(	DNA_B	The AC MF185010 is a n	
22	Unassigned	Geminiviridae	Begomovirus	Sida Yellow Mosaic Virus	India-Gujarat, Gandhinagar-2016	KX513859	CG	SYMV[India-Gandh	Sida yellow mosaic virus	ACCGGATGC	DNA_B	The AC KX513859 is a ne	
23	Unassigned	Geminiviridae	Begomovirus	Emilia Sonchifolia Yellow Vein Thai	Thailand-Surat Thani-2015	KY373213	CG	ESYVTV[TH4872-6	Emilia sonchifolia yellow vein Thaila	ACCGGATGC	DNA_B	The AC KY373213 is a n	
24	Unassigned	Geminiviridae	Begomovirus	Tomato Chlorotic Leaf Curl Virus	Venezuela-Unassigned-2008	KY449277	CG	TCLCV[Zulia-1084a	Tomato chlorotic leaf curl virus	ACCGGATGC	DNA_B	The AC KY449277 is a n	
25	Unassigned	Geminiviridae	Begomovirus	Tomato Mosaic Trujillo Virus	Venezuela-Unassigned-2006	KY449275	CG	TMTV[Trujillo-427a-	Tomato mosaic Trujillo virus	ACCGGATGC	DNA_B	The AC KY449275 is a n	

Supplementary figure S4. Sixth Excel sheet, containing only the new species representatives found by the algorithm.