This form should be used for all taxonomic proposals. Please complete all those modules that are applicable (and then delete the unwanted sections). For guidance, see the notes written in blue and the separate document "Help with completing a taxonomic proposal"

Please try to keep related proposals within a single document; you can copy the modules to create more than one genus within a new family, for example.

MODULE 1: **TITLE, AUTHORS, etc**

Code assigned:	***2010.002aG***	(to be completed by ICTV officers)

Short title: Study Group involvement in the use of PASC (pairwise sequence comparison) as a tool for classification
(e.g. 6 new species in the genus *Zetavirus*)

Modules attached (modules 1 and 9 are required)					
	1 ☒	**2** ☐	**3** ☐	**4** ☐	**5** ☐
	6 ☐	**7** ☐	**8** ☒	**9** ☒	

Author(s) with e-mail address(es) of the proposer:

Yiming Bao, bao@ncbi.nlm.nih.gov

List the ICTV study group(s) that have seen this proposal:

A list of study groups and contacts is provided at http://www.ictvonline.org/subcommittees.asp . If in doubt, contact the appropriate subcommittee chair (fungal, invertebrate, plant, prokaryote or vertebrate viruses)	

ICTV-EC or Study Group comments and response of the proposer:

Date first submitted to ICTV:
Date of this revision (if different to above):

MODULE 8: **NON-STANDARD**

Template for any proposal not covered by modules 2-7. This includes proposals to change the name of existing taxa (but note that stability of nomenclature is encouraged wherever possible).

non-standard proposal

Code	***2010.002aG***	(assigned by ICTV officers)

Title of proposal: ***The importance of Study Groups' involvement in developing and improving PASC for virus classification***

Text of proposal:

Pairwise sequence comparison (PASC) is a molecular classification tool for viruses (Bao et al., 2008). It calculates the pairwise identities of virus sequences within a virus family and displays their distributions, and can help determine demarcations at different taxonomic levels such as strain, species, genus and subfamily.

In an attempt to a). provide an online tool to use PASC for virus families to which this method already has been applied (e.g. the families of *Geminiviridae, Papillomoviridae, Picornaviridae* and *Potyviridae*), and b). investigate the possibilities of applying PASC to a wider range of viral groups, the National Center for Biotechnology Information (NCBI) created a PASC resource (http://www.ncbi.nlm.nih.gov/sutils/pasc) which currently covers the following viral groups (see Fig. 1 for an example):

Alphaflexiviridae
Anelloviridae
Arteriviridae
Astroviridae
Betaflexiviridae
Betasatellite
Caliciviridae
Caulimoviridae
Circoviridae
Comoviridae
Coronaviridae
Dicistroviridae
Filoviridae
Flaviviridae
Geminiviridae
Hepadnaviridae
Iflaviridae
Inoviridae
Iridoviridae
Lentivirus
Leviviridae
Lipothrixviridae
Luteoviridae
Microviridae
Narnaviridae
Papillomaviridae
Paramyxoviridae

Parvoviridae
Picornaviridae
Polyomaviridae
Potyviridae
Rhabdoviridae
Sobemovirus
Tectiviridae
Tobamovirus
Togaviridae
Tombusviridae
Tymoviridae
Umbravirus

The PASC tool is linked to the NCBI's viral genome collection (Bao et al. 2004) and taxonomy database, with new viral genomes added and taxonomy status of viruses updated every day. It runs very fast and results can usually be obtained within minutes. The tool is online, so there is no software to download/install, no parameters to set, and everybody uses the same algorithm and same dataset. This allows consistency of the results by different users. We believe that PASC can be a great aid to ICTV Study Groups by a) providing a much more objective criteria for making taxonomic assignments based on sequence comparisons; and b) cleaning up problematic entries in the current NCBI taxonomy database.

Although the NCBI PASC tool has been used in several studies (Domínguez et al. 2009, Huang et al. 2010, Lam et al. 2009, Mubina et al. 2009, Vaira et al. 2008, Yan et al. 2010), there are issues need to be addressed.

I. No borderline values in sequence identity percentages are provided in our PASC system that can be used to separate species and genera. This is because for some families, these borderlines are not well separated (e.g. the family *Potyviridae*, Fig. 2). Even for families whose peaks are well separated (e.g. the family *Flaviviridae*, Fig. 1), we felt that this should be left to the ICTV Study Group.

II. In the PASC analysis, we expect bars with the same color group together (as shown in Fig. 1). In the case of *Circoviridae* family, however, there are green bars in the yellow-dominant regions and vice versa (see Fig. 3). These are caused by the following two reasons.

A). A problem in circoviruses, like any other viruses with circular genomes, is the inconsistency in the designation of the first nucleotide in GenBank sequences. For example, Y09921 is about 765 nucleotides off from the majority of other Porcine circovirus 1 sequences, with the later start from TAGTATTA in the stem-loop region. This inconsistency messes up PASC analysis which is based on global genome alignment. Table 1 shows some of the genome pairs that contribute to the bar at 65% in Fig. 3. The two genomes in the third pair (AF055391 and AY484407) belong to the same species (therefore is green in the bar), but their sequence similarity is only 65.5% after global alignment (see http://tinyurl.com/2coeaoz). The similarity would be much higher (~97%) if the two sequences start from the same position in the genome.

B). Not all GenBank sequences are in the most appropriate taxonomy node. Table 2 shows some of the genome pairs that contribute to the bar at 86.5% in Fig. 3. One of the genome in the third pair, FJ655419, is currently in the "unclassified Circovirus" node in NCBI's taxonomy database, which is different from the species *Porcine circovirus 2*, to which the other genome in

this pair (AY484407) belong. This gives this pair the yellow color in the bar. Based on their sequence identity, FJ655419 should most definitely be placed in the species *Porcine circovirus 2*.

III. The NCBI PASC tool is currently based on complete genomes for non-segmented viruses and one of the genome segments for segmented viruses (e.g. DNA A of geminiviruses). We have not investigated whether this represents a better choice than using one or several genes, or a different segment in a genome. Many families with segmented genomes are not present in PASC, again because we believed it should be Study Group's call in terms of which segment to use.

These issues can be mostly resolved if ICTV Study Groups are involved more in the development and improvement of PASC. We therefore propose that a working relationship is established between NCBI and ICTV Study Groups. Specifically, we suggest that:

1. All Study Groups visit NCBI's PASC website at http://www.ncbi.nlm.nih.gov/sutils/pasc. If viral families that are the subjects of your Study Group can be found there, go to #2 to 5. Otherwise, go to #4.

2. For existing viral families, determine whether the PASC system involving complete genome or genome segment currently being used is, or will potentially be, applicable. If the answer is yes, then

3. To investigate if borderline values in sequence identity percentages can be determined in the PASC system to separate species, genera and other taxonomy levels. If yes, these would become official PASC demarcations and be documented in, for example, the 9th Report of ICTV. We will then label these borderline values in the NCBI PASC system, so users can use them as references when they test their new sequences in the NCBI PASC tool.

4. If the current PASC system is not optimized for certain families, the Study Group should advice us alternative ways to perform PASC, e.g. using sequences of one or several genes rather than complete genomes. These will be tested in the NCBI system and the Study Group will look at the result before the best approach is chosen. This also applies to viral families who are currently not present in the NCBI PASC tool.

5. Issue II described above is a good way to identify viral sequences with incorrect taxonomy classification in the NCBI database. Although we always try our best to catch such sequences and fix them in GenBank, it will be more efficient for the Study Groups to examine the PASC histograms and find the problematic sequences.

We understand that PASC alone is not the solution to classification of many viral families, and other biological properties have to be taken into consideration. By working together with ICTV Study Groups, we believe we can explore the potential of PASC, and maximize its application for as many viruses as possible. Any suggestions and comments are always welcome and can be sent to genomes@ncbi.nlm.nih.gov.

MODULE 9: **APPENDIX**: supporting material

additional material in support of this proposal

References:

Bao Y, Federhen S, Leipe D, Pham V, Resenchuk S, Rozanov M, Tatusov R, Tatusova T. (2004). National center for biotechnology information viral genomes project. J Virol. 78(14):7291-8.

Bao Y, Kapustin Y, Tatusova T. 2008. Virus Classification by Pairwise Sequence Comparison (PASC). Encyclopedia of Virology, 5 vols. (B.W.J. Mahy and M.H.V. Van Regenmortel, Editors). Oxford: Elsevier. Vol. 5, 342-348.

Domínguez M, Ramos PL, Sánchez Y, Crespo J, Andino V, Pujol M, Borroto C. 2009. Tobacco mottle leaf curl virus, a new begomovirus infecting tobacco in Cuba. Plant Pathology. 58(4): 786.

Huang YW, Ni YY, Dryman BA, Meng XJ. 2010. Multiple infection of porcine Torque teno virus in a single pig and characterization of the full-length genomic sequences of four U.S. prototype PTTV strains: implication for genotyping of PTTV. Virology. 396(2):289-97.

Lam N, Creamer R, Rascon J, Belfon R. 2009. Characterization of a new curtovirus, pepper yellow dwarf virus, from chile pepper and distribution in weed hosts in New Mexico. Arch Virol. 154(3):429-36.

Mubin M, Briddon RW, Mansoor S. 2009. Diverse and recombinant DNA betasatellites are associated with a begomovirus disease complex of Digera arvensis, a weed host. Virus Research. 142(1-2):208-212.

Vaira AM, Maroon-Lango CJ, Hammond J. 2008. Molecular characterization of Lolium latent virus, proposed type member of a new genus in the family Flexiviridae. Arch Virol. 153(7):1263-70.

Yan ZL, Song LM, Zhou T, Zhang YJ, Li MF, Li HF, Fan ZF. 2010. Identification and molecular characterization of a new potyvirus from Panax notoginseng. Arch Virol. In press.

Annex:

Include as much information as necessary to support the proposal, including diagrams comparing the old and new taxonomic orders. The use of Figures and Tables is strongly recommended but direct pasting of content from publications will require permission from the copyright holder together with appropriate acknowledgement as this proposal will be placed on a public web site. For phylogenetic analysis, try to provide a tree where branch length is related to genetic distance.

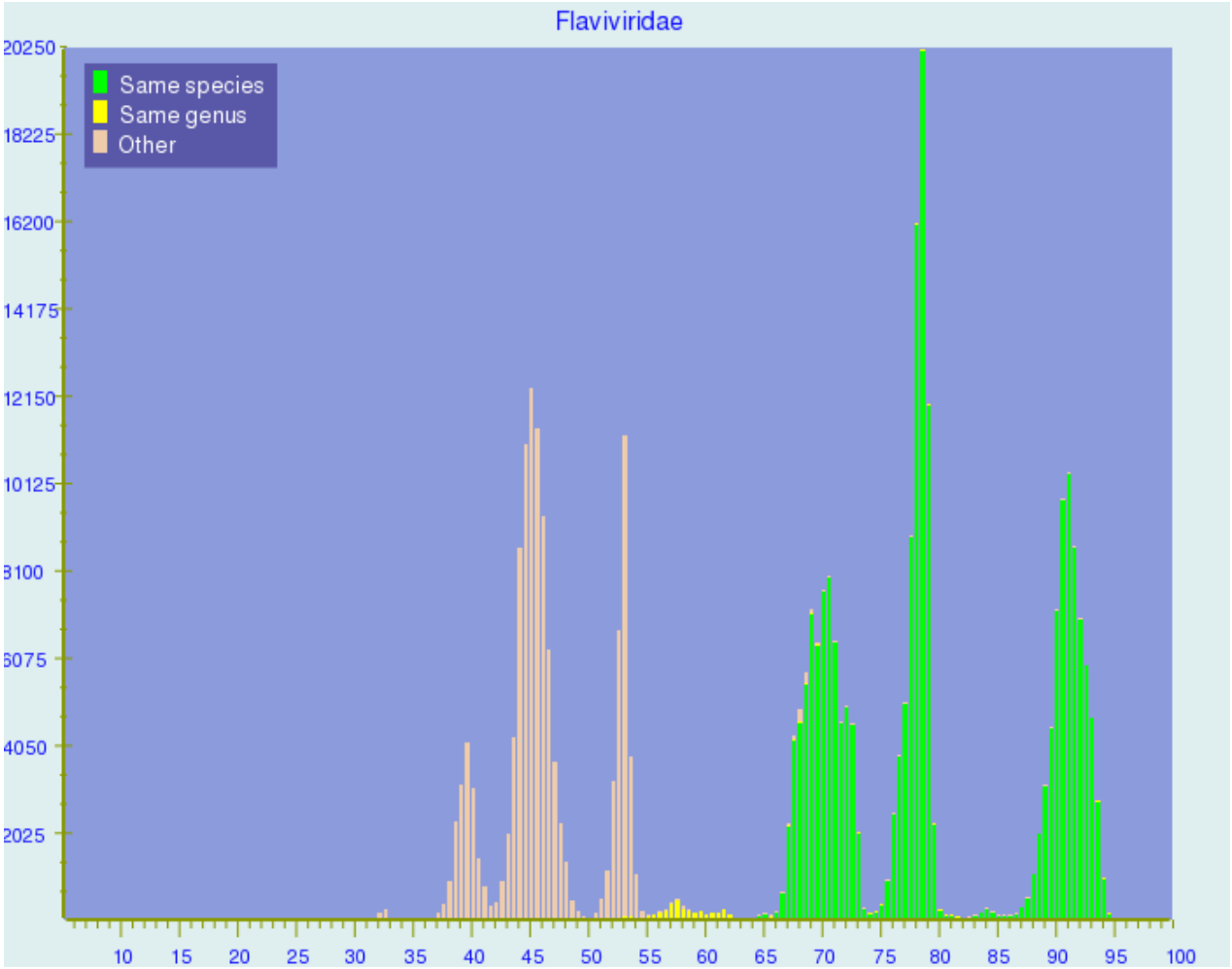

Figure 1. Frequency distribution of pairwise identities from the complete nucleotide sequence comparison in the family *Flaviviridae*.

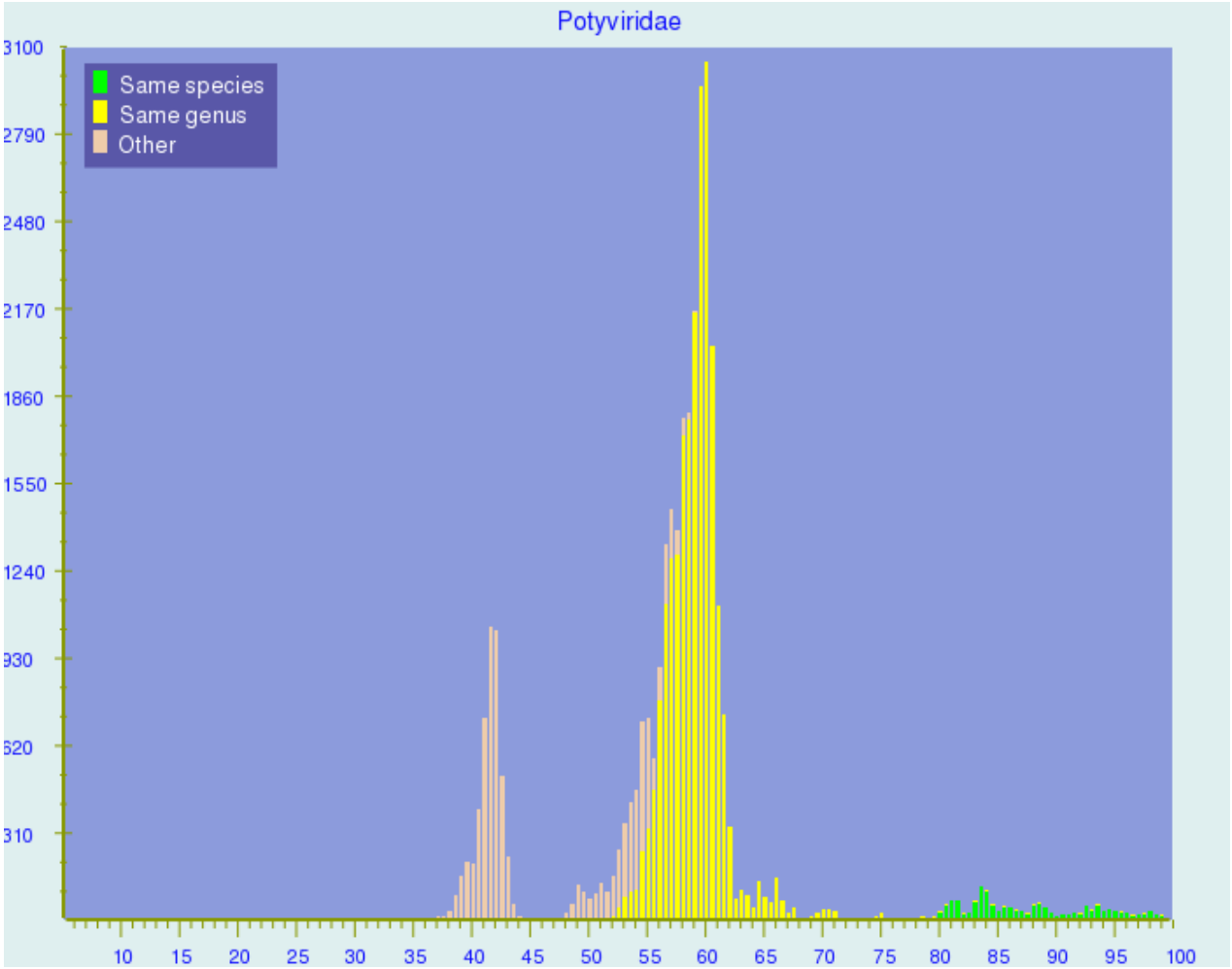

Figure 2. Frequency distribution of pairwise identities from the complete nucleotide sequence comparison in the family *Potyviridae*.

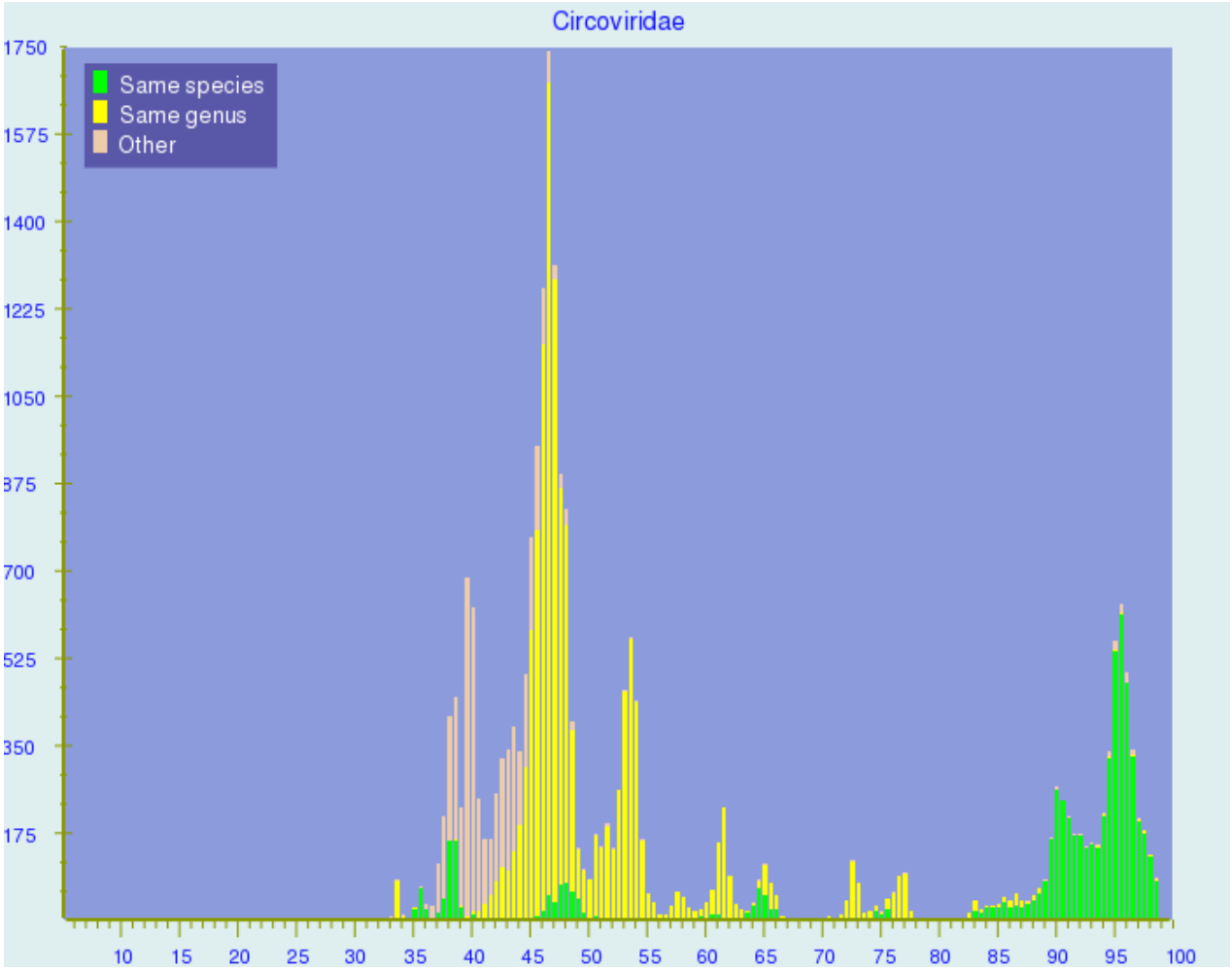

Figure 3. Frequency distribution of pairwise identities from the complete nucleotide sequence comparison in the family *Circoviridae*.

Table 1. Pairwise comparison of complete nucleotide sequences in the family *Circoviridae*.

Identity	Same genus?	Same species?	Genome 1	Genome 2
0.654974	Yes	No	gi|15148076|ref|NC_003054.1| Circovirus|Goose circovirus	gi|257815103|gb|GQ423742.1| Circovirus|unclassified Circovirus|Muscovy duck circovirus
0.654846	Yes	Yes	gi|3598796|gb|AF055391.1| Circovirus|Porcine circovirus 2	gi|62082437|gb|AY874166.1| Circovirus|Porcine circovirus 2
0.654846	Yes	Yes	gi|3598796|gb|AF055391.1| Circovirus|Porcine circovirus 2	gi|45476733|gb|AY484407.1| Circovirus|Porcine circovirus 2

Table 2. Pairwise comparison of complete nucleotide sequences in the family *Circoviridae*.

Identity	Same genus?	Same species?	Genome 1	Genome 2
0.869995	Yes	Yes	gi|11119220|gb|AF311301.1| Circovirus|Beak and feather disease virus	gi|225904192|gb|FJ685978.1| Circovirus|Beak and feather disease virus
0.8698	Yes	Yes	gi|157418992|gb|EU136711.1| Circovirus|Porcine circovirus 2	gi|217323271|gb|EU909688.1| Circovirus|Porcine circovirus 2
0.869614	Yes	No	gi|45476733|gb|AY484407.1| Circovirus|Porcine circovirus 2	gi|257838637|gb|FJ655419.1| Circovirus|unclassified Circovirus|Porcine circovirus Canada-2010a