

**PROTOCOL FOR HEPATITIS C VIRUS GENOTYPING/SUBTYPING TOOL**

May 4, 2018

1. Background

Hepatitis C Viruses (HCV) have diversified into seven major genotypes (1-7) over time. Each major genotype is further classified into genotype/subtypes, e.g., 1a, 1b, 1c, etc. A list of current genotype and subtype assignments is maintained by the Flaviviridae Study Group of the International Committee on Taxonomy of Viruses (ICTV) (https://talk.ictvonline.org/ictv_wikis/flaviviridae/w/sg_flavi/56/hcv-classification). As of June 2017, the number of confirmed genotypes/subtypes has increased to 86 (ICTV, 2017). In order to assist researchers in designating appropriate assignments for new HCV sequences using current genotype/subtype assignments, the ViPR team has developed an HCV Genotyping/Subtyping Tool. This document describes the HCV genotyping/subtyping tool in ViPR.

2. Method Description

An automated pipeline was developed for assigning genotype/subtype to un-genotyped HCV sequences, whereby:

2.1 A reference alignment is constructed following the steps below:

2.1.1 The reference alignment published by the ICTV on June 8, 2017 (Updated alignment (FASTA) of HCV genotypes and subtypes 1.6.17.FST;

https://talk.ictvonline.org/ictv_wikis/flaviviridae/w/sg_flavi/57/hcv-reference-sequence-alignments) was trimmed to the CDS region only.

2.1.2 Additional sequences with confirmed subtype (provided by Dr. Donald Smith) were added to the above alignment using MAFFT (mafft --addfragments).

2.1.3 Manually adjusted one insertion introduced by the new sequences to keep the reference alignment intact.

2.1.4 The resulting alignment contains 231 HCV reference sequences. The reference alignment can be downloaded from the ViPR site:

<https://www.viprbrc.org/brc/workbenchSequenceSearch.spg?uploadedFileId=20272&decorator=flavi&method=SubmitForm>

2.2 A reference tree is computed following the steps below:

2.2.1 The multiple sequence alignment described above was input to RAxML (version 7.2.6) with the GTR model of nucleotide substitution and a discrete gamma model with 4 categories.

2.2.2 The output best tree (RAxML_bestTree) is then midpoint rooted using Archaeopteryx.

2.2.3 The resulting midpoint-rooted tree is used as the reference tree in the HCV typing tool. It can be viewed or downloaded from the ViPR site:

<https://www.viprbrc.org/brc/uploadedFileDetail.spg?method=SharedFileDetail&uploadedFileId=20275&decorator=flavi>

2.3 A query sequence is checked with regard to its sequence type and sequence length. Minimum length requirement is 400 bp.

2.4 A query sequence is aligned against the reference alignment using MAFFT (mafft --keeplength --add).

2.5 The query sequence is placed into the reference tree using pplacer (Matsen, 2010), with the reference tree serves as a “scaffold” onto which the query sequence is placed.

2.6 The pplacer output is parsed by guppy.

2.7 The guppy output is analyzed by cladinator

(<https://sites.google.com/site/cmzmasek/home/software/forester/cladinator>).

2.7.1 Background of cladinator logics

cladinator assigns a genotype/subtype for the query sequence based on its placement in the phylogeny:

- When a query sequence is placed unequivocally within the bounds of a single defined type, this type name is assigned to the query (Figure 1A).
- When a query sequence is bracketed by two different types and these two types share a common parent type in their type names, the parent type name is assigned to the query (Figure 1B).
- When a query sequence is bracketed by two different types with no common type name in the type names, the query sequence is of unknown type (Figure 1C).

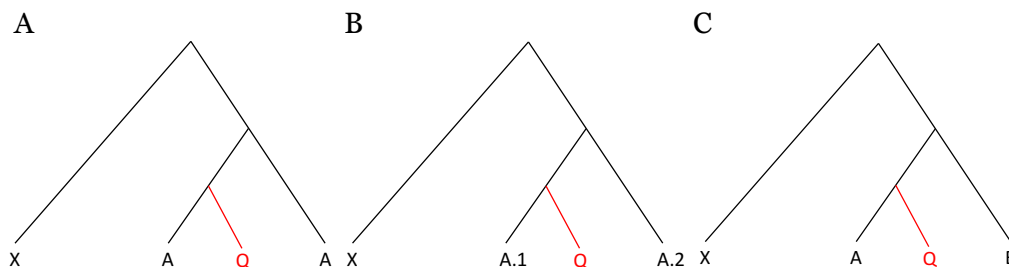


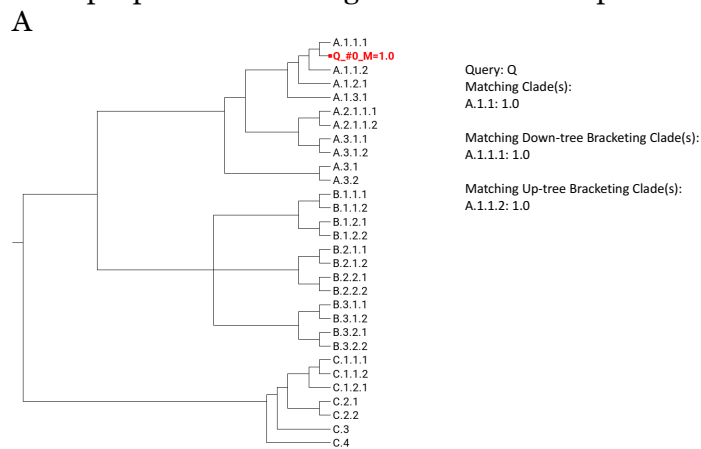
Figure 1. cladinator analysis of query placements in hierarchically annotated artificial trees. (A) Query is A-type (bracketed by A and A). (B) Query is A-type (bracketed by A.1 and A.2). (C) Query is of unknown type (bracketed by A and B). In reference to Q, A is called “down-tree”, while B is called “up-tree.” Naïvely, it looks like Q might be of A-type, but we do **not** know at which point along the branch going from AB-ancestor to A, the type changes from AB-ancestor-type to A-type. Therefore, Q is of unknown type.

2.7.2 cladinator analysis of a single query placement

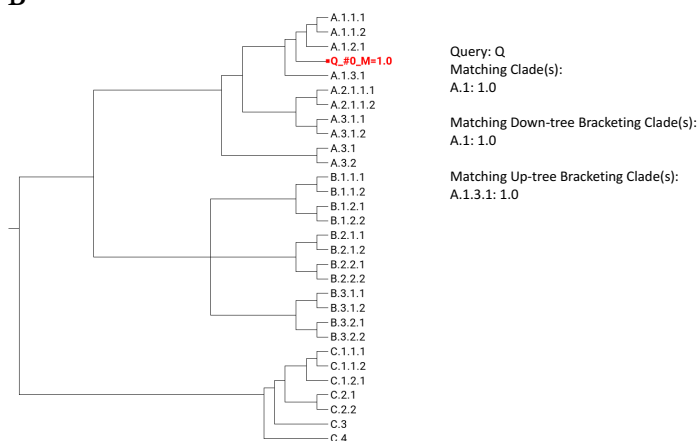
For each query placement, cladinator reports the typing assignment in the following fields:

- Matching Clade
- Matching Down-tree Bracketing Clade
- Matching Up-tree Bracketing Clade

Example placements along with cladinator reports are provided in Figure 2.



B



C

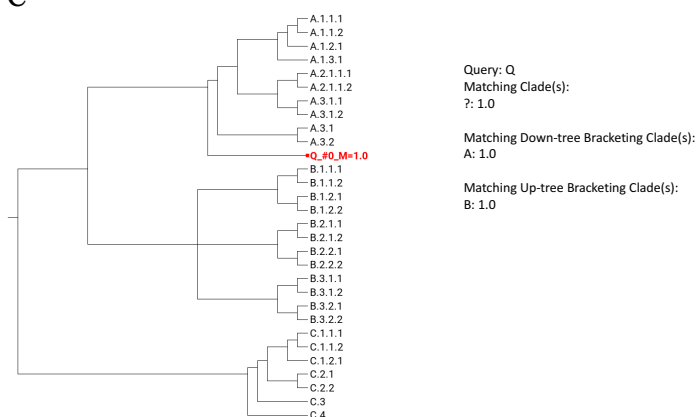
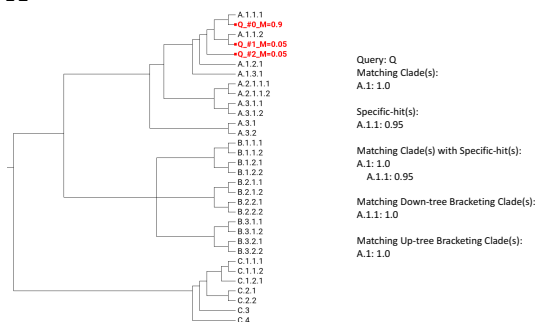


Figure 2. cladinator analysis of single query placements in a hierarchically annotated artificial tree. Query placements are in red. cladinator output is to the right of the tree.

2.7.3 cladinator analysis of multiple query placements

When a query has multiple placements, cladinator summarizes the results if possible. Specifically, when two or more placements are assigned the same type (e.g., A.1.1 in Figure 3), the Specific-hit field reports that the probability score for the shared type (e.g., A.1.1 in Figure 3) is the sum of individual placement's probability score.

A



B

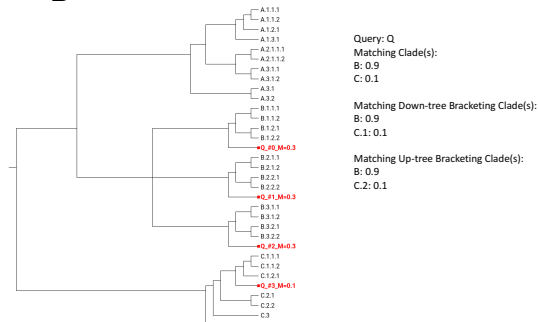


Figure 3. cladinator analysis of multiple query placements in a hierarchically annotated artificial tree. Query placements are in red. cladinator output is to the right of the tree.

3. Access of the tool

The HCV genotyping/subtyping tool is accessible from **Virus Pathogen Resource > Hepatitis C Virus or Flaviviridae > Analyze & Visualize > Genotype-Recombination Detection**

(https://www.viprbrc.org/brc/genotypeRecombination.spg?method=ShowCleanInputPage&decorator=flavi_hcv). Input sequences can be provided by choosing a working set saved in the Workbench, uploading a file, pasting in FASTA-formatted sequences, or a sequence file uploaded to the Workbench (Figure 4). The analysis report provides the full report from cladinator and the alignment and tree used to type the input sequence (Figure 5).

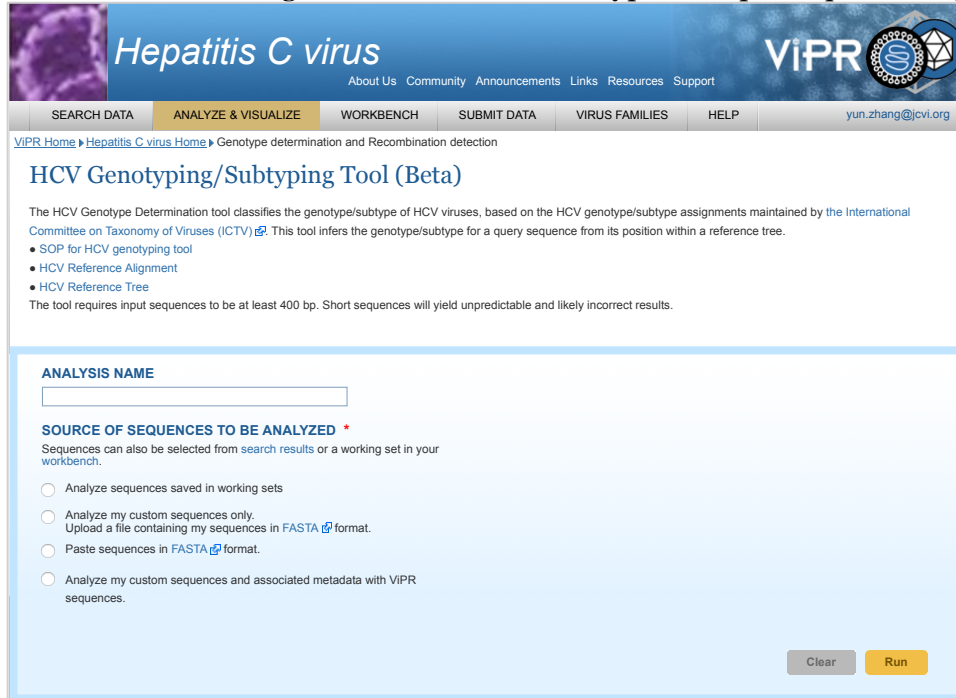


Figure 4. The HCV genotyping tool landing page

HCV Genotyping/Subtyping Report (Beta) (SOP)

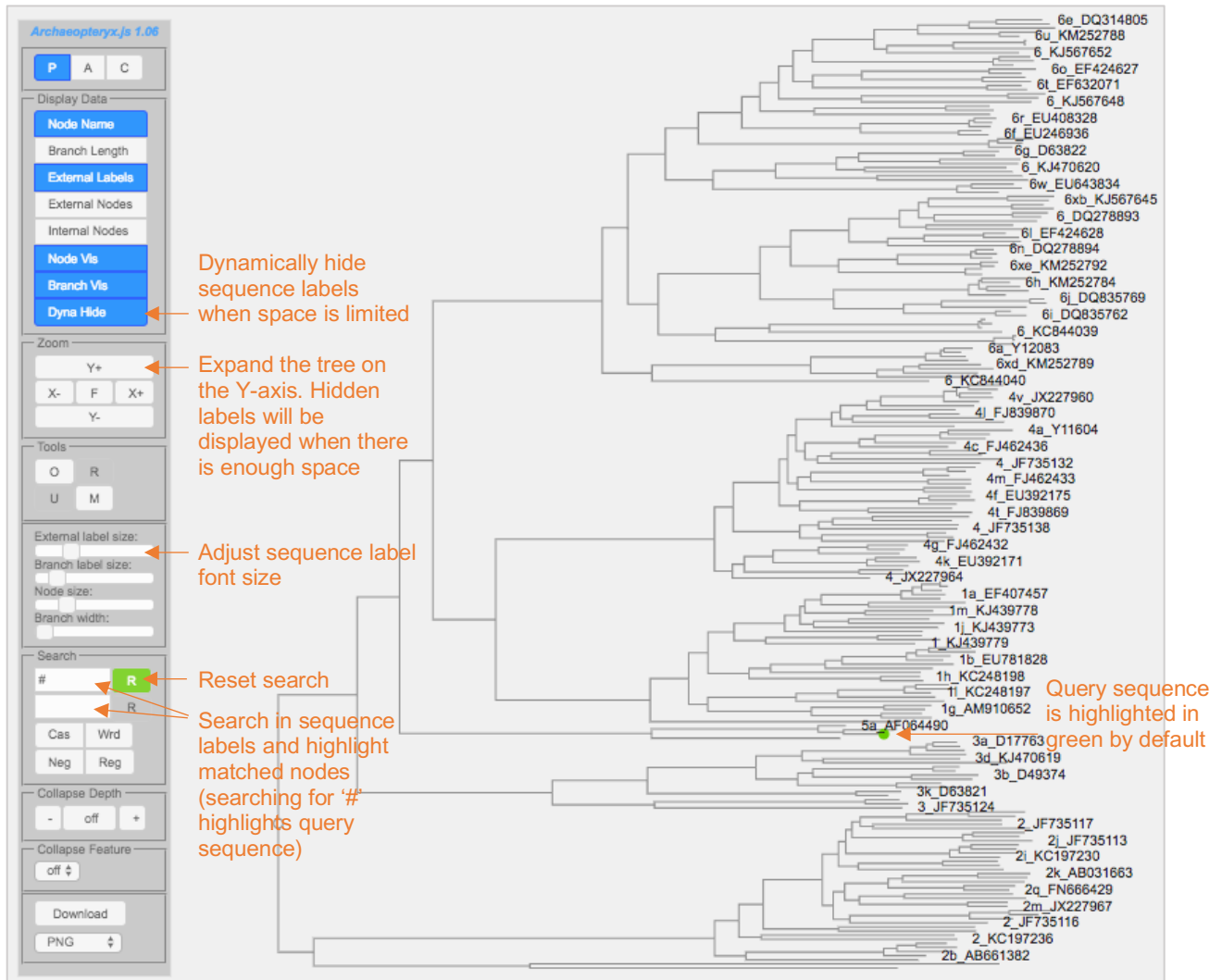
Save Analysis Download results in tsv file

Your analysis contains 3 records

Query Identifier	Query Length	Type	Consensus Assignment	Support	Phylogenetic Tree	Report
KX107885	1772	Matching Clades Matching Down-tree Bracketing Clades Matching Up-tree Bracketing Clades	Sa Sa Sa	1.0 1.0 1.0	View	Input alignment (FASTA) Output tree (Newick) Subtype assignment (text)
KX107872	1773	Matching Clades Matching Down-tree Bracketing Clades Matching Up-tree Bracketing Clades	Sa Sa Sa	1.0 1.0 1.0	View	Input alignment (FASTA) Output tree (Newick) Subtype assignment (text)
KX107874	1773	Matching Clades Matching Down-tree Bracketing Clades Matching Up-tree Bracketing Clades	Sa Sa Sa	1.0 1.0 1.0	View	Input alignment (FASTA) Output tree (Newick) Subtype assignment (text)

Figure 5. An example of the HCV genotyping report. On this page, users can download: (a) the input alignment which is an alignment of the query sequence with the reference alignment, (b) the output tree with the query sequence placed in the tree, and (c) the subtype assignment file which includes the table being displayed and additional information from the typing tool. The phylogenetic tree – View hyperlinks link to the output tree displayed in the Archaeopteryx.js tree viewer as shown in Figure 6.

A



B

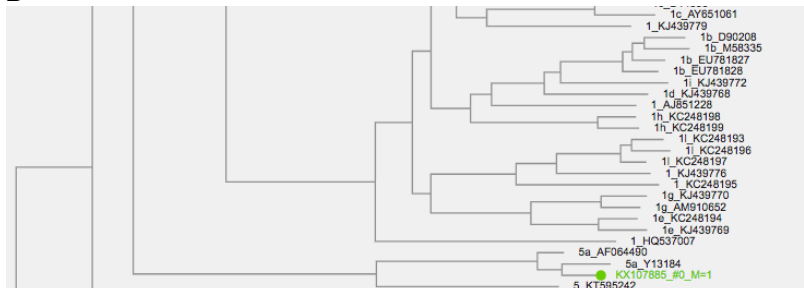


Figure 6. (A) An example HCV typing output tree visualized in Archaeopteryx.js tree viewer. The query sequence is highlighted in green by default. Users can adjust the look of the tree by using various visualization options in the left panel. (B) The same tree shown in (A) is zoomed in on the Y-axis. Hidden sequence labels in (A) are displayed.

References

International Committee on Taxonomy of Viruses (ICTV). HCV Classification.
https://talk.ictvonline.org/ictv_wikis/flaviviridae/w/sg_flavi/56/hcv-classification

Matsen FA, et al. pplacer: linear time maximumlikelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Biomathematics. 2010, 11:538. PMID: 21034504.

Zmasek CM. cladinator.
<https://sites.google.com/site/cmzmasek/home/software/forester/cladinator>

Zmasek CM. Archaeopteryx.js.
<https://sites.google.com/site/cmzmasek/home/software/archaeopteryx-js>