Journal of Clinical Virology 134 (2021) 104718

Contents lists available at ScienceDirect

Journal of Clinical Virology

journal homepage: www.elsevier.com/locate/jcv

Human Calicivirus Typing tool: A web-based tool for genotyping human norovirus and sapovirus sequences

Roman L. Tatusov [a,b], Preeti Chhabra [b], Marta Diez-Valcarce [b,c], Leslie Barclay [b], Jennifer L. Cannon [d], Jan Vinjé [b,*]

[a] Cherokee Nation Assurance, Arlington, VA, 22202, USA
[b] Division of Viral Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA
[c] Rollins School of Public Health, Emory University, Atlanta, GA, USA
[d] National Foundation for the Centers for Disease Control and Prevention Inc., Atlanta, GA, USA

ARTICLE INFO

Keywords:
Norovirus
Sapovirus
Calicivirus
Typing tool
Genotyping
Polymerase
k-mer
Web-based

ABSTRACT

Background: The family *Caliciviridae* consists of a genetically diverse group of RNA viruses that infect a wide range of host species including noroviruses and sapoviruses which cause acute gastroenteritis in humans. Typing of these viruses relies on sequence-based approaches, and therefore there is a need for rapid and accurate web-based typing tools.

Objective: To develop and evaluate a web-based tool for rapid and accurate genotyping of noroviruses and sapoviruses.

Methods: The Human Calicivirus Typing (HuCaT) tool uses a set of curated reference sequences that are compared to query sequences using a k-mer (DNA substring) based algorithm. Outputs include alignments and phylogenetic trees of the 12 top matching reference sequences for each query.

Results: The HuCaT tool was validated with a set of 1310 norovirus and 239 sapovirus sequences covering all known human norovirus and sapovirus genotypes. HuCaT tool assigned genotypes to all queries with 100 % accuracy and was much faster (17 s) than BLAST (150 s) or phylogenetic analyses approaches.

Conclusions: The web-based HuCaT tool supports rapid and accurate genotyping of human noroviruses and sapoviruses.

1. Introduction

The family *Caliciviridae* consists of a genetically diverse group of single-stranded RNA viruses that can be divided into 10 genera [1]. Of these, viruses belonging to the genus *Norovirus* and *Sapovirus* cause acute gastroenteritis (AGE) in humans. Noroviruses are associated with an estimated 70,000–200,000 deaths annually [2,3] while sapoviruses primarily cause sporadic AGE in young children although outbreaks in all age groups have been reported [4,5].

Noroviruses are genetically divided into 10 genogroups and 48 genotypes [6] with viruses in GI, GII, GVIII and GIX infecting humans. Sapoviruses can be classified into up to 19 genogroups (GI–GXIX) [7,8] of which viruses from GI, GII, GIV and GV infect humans [4]. Recombination events within the norovirus genome are well-documented forces that drive norovirus evolution [9,10] and most frequently occur at the junction of the RdRp (polymerase) and VP1 (capsid) encoding regions [11]. Therefore, dual typing of partial regions of both genes is increasingly used for genotyping of norovirus strains [6,12]. For routine typing of norovirus, short nucleotide regions at the 5'-end of the capsid gene and at the 3'-end of the polymerase gene and for sapovirus the 5'-end of the capsid gene are used [13–15]. Traditionally, sequences are compared to an established set of reference sequences to determine genotypes and polymerase types based on phylogenetic clustering, or by employing NCBI BLAST web service. The web-based Norovirus Typing Tool (https://www.rivm.nl/mpf/norovirus/typingtool), which uses a BLAST algorithm against a set of reference sequences followed by phylogenetic analysis to assign norovirus genotypes and polymerase (P)-types and sapovirus genotypes, has been widely used for norovirus and sapovirus typing since 2011 [16].

In this paper, we describe the development and evaluation of an alternative web-based tool, Human Calicivirus Typing (HuCaT) tool for typing of human norovirus and sapovirus sequences. Like the Norovirus

* Corresponding author at: Division of Viral Diseases, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA, 30329, USA.
E-mail address: jvinje@cdc.gov (J. Vinjé).

https://doi.org/10.1016/j.jcv.2020.104718
Received 28 August 2020; Received in revised form 7 December 2020; Accepted 9 December 2020
Available online 13 December 2020
1386-6532/Published by Elsevier B.V.

Typing Tool, it uses the most recent norovirus and sapovirus reference sequences and nomenclature [4,6,8]. HuCaT uses an algorithm based on matching nucleic acid k-mers (short substrings of DNA [n = 9 nt in HuCaT]) to build alignments between query and reference sequences. With this new approach, we demonstrated the HuCaT algorithm offers speed improvement without losing accuracy for norovirus and sapovirus typing. The HuCaT output includes taxonomic classification (genotype and P-type), percent nucleotide identity, an alignment and phylogenetic tree of query sequence with the closest 12 reference sequences, summary reports, and visualization of query sequences submitted.

2. Methods

2.1. Reference sequences

The HuCaT tool uses a set of reference sequences as defined in the most recent classification papers for norovirus and sapovirus [4,6]. All reference sequences are publicly available on the HuCaT website hosted by CDC (Atlanta, USA) (https://norovirus.ng.philab.cdc.gov/," Reference sequences" tab).

2.2. Genotype and polymerase typing regions

Specific nucleotide percent identity cut-off values are used to designate norovirus and sapovirus genogroups and genotypes [4,17] and P-types for norovirus (Table 1). Norovirus P-types are based on a 172 nt (173 nt for GIV) region at the 3' end of the polymerase encoding region and norovirus genotypes are based on a 264 nt region for GI (267 nt for GI.9), 252 nt region for GII, and 255 nt region for GIX at the 5'end of the VP1 (major capsid protein) encoding region. Sapovirus genotypes are based on a 420 nt typing region encoding VP1 [14,15].

2.3. HuCaT algorithm and workflow

Users can upload nucleotide sequences in FASTA format individually or multiple sequences as a batch. The HuCaT workflow processes query sequences into k-mers of 9 nucleotides in length (called 9-mers) to compare query sequences to the set of reference sequences representing all known human norovirus and sapovirus strains. It does this through a hash table or index of all possible 9-mers stored as keys and values including a reference sequence identifier and the nucleotide position of the 9-mer within the reference sequences (Figs. S1 and S2). Although k-mers with a k value in the 9–14 range can be chosen without compromising accuracy and speed, k-mers of k = 9 (9-mers) were deemed appropriate for optimal speed and accuracy (Fig. S1). After all query 9-mers are scanned, the stored nucleotide position information is used to construct an alignment of each query sequence with the top matching reference sequences. This is done by calculating the relative difference in the nucleotide positions (ntpos) of each matching pair ($ntpos_{reference} - ntpos_{query}$) which can be visualized using a dot plot (Fig. S2) with matching pairs of coordinates ($ntpos_{query}$, $ntpos_{reference}$) represented by colored dots and the longest spanning diagonal representing the optimal arrangement of the pair. The reference sequences with the highest number of matching queries, k-mers are used to calculate the actual number of nucleotide sequence matches for type assignment using the specific nucleotide percent identity cut-off values (Table 1). Based on this nucleotide identity the tool assigns a genogroup, a genotype and/or polymerase type, and a variant type for GII.4 noroviruses.

Table 1
Nucleotide percent identity cut-off values to designate genogroups and genotypes for norovirus and sapovirus and polymerase types for norovirus used by the HuCaT tool.

Virus	Genomic region	Genogroup	Genotype	GII.4 variant
Sapovirus[1]	Capsid	51 %	83 %	NA
Norovirus[2]	Capsid	70 %	90 %	98 %
	Polymerase	70 % (GI) 70 % (GII)	87 % (GI) 93 % (GII)	NA 98 % (GII)

NA = not applicable.
[1] [4].
[2] [17].

2.4. HuCaT performance optimization and validation

A total of 52,176 norovirus and sapovirus sequences were downloaded from GenBank on April 8, 2020 and used to optimize the speed and performance of the HuCaT algorithm as a function of k-mer size. Sequences were submitted in a single batch to HuCaT and by stand-alone BLAST tool against the same set of reference sequences (BLAST-ref) to compare the speed of the algorithms. Next, a total of 1538 unique sequences (1299 norovirus and 239 sapovirus) representing all genotypes circulating in humans were selected to validate the typing results of HuCaT (Table S1a and S1b). Norovirus sequences were selected from databases at CDC [13] and from GenBank. The selected sequences included all norovirus and sapovirus genotypes and norovirus polymerase types. All sequences were of high quality (no indels or ambiguous bases within the typing regions). Initial genotypes and/or polymerase types for these selected sequences were determined using traditional phylogenetic methods using reference sequences [4,6,8]. In addition, the sequences were subjected to typing by HuCaT, BLAST-ref and the Norovirus Typing Tool.

3. Results

To optimize the performance of HuCaT, we calculated the time it took to process 52,176 norovirus and sapovirus sequences downloaded from GenBank. Of these, 6155 norovirus sequences that contained both polymerase and capsid typing regions and 320 sapovirus sequences could be typed by HuCaT. All sequences were queried using different size of k-mers. We settled on a k-mer size of 9 (Fig. 1, Fig S1,b), which resulted in typing of all norovirus and sapovirus viruses in 17 s. In contrast, the BLAST-ref algorithm took 150 s to type all sequences.

Validation of HuCaT using 1299 unique norovirus and 239 sapovirus sequences (Table S1a,b) showed that, except for one GI.P3 sequence (2019-SP-0093), HuCaT accurately assigned all sequences into C-types (genotypes), GII.4 variants and P-types while the Norovirus Typing Tool correctly typed 99.9 % of norovirus genotypes (except GIV.NA1) and 99.0 % of norovirus P-types (Table 2) except one GI.P3, two GI.P11, four GII.P4, one GII.P7, one GII.P17, one GII.PNA6 and one GIV.PNA1 sequence (Table 3). For typing of sapovirus, the Norovirus Typing Tool typed 10 phylogenetically confirmed sapovirus GII.8 sequences as GII.7 and two GII.NA1 sequences as GII.5 (Table 3).

In addition to typing norovirus and sapovirus sequences, HuCaT provides reports on individual (Fig. S3a) or batch submission (Fig. S3b) of query sequences. Single queries generate a report providing the query sequence name and a visual representation of the typing region location (orange [P region] and green [C region] superimposed rectangles) relative to the query sequence (teal rectangle). Additional information includes strand sense (plus/minus), genus, genotype and/or polymerase type (providing the dual typing nomenclature when available), and the percent nucleotide identity with the top matching reference sequences for each typing region. Finally, visualizations are provided for the optimal arrangements (alignments). The 12 reference sequences with the highest similarity are used to illustrate the phylogenetic relationship with reference sequences using UPGMA trees. A close-up view shows the nucleotide differences between the query and reference pairs. Multiple queries are reported initially in a table and include a visual representation of the location of the typing regions similar to the output for single queries. A link is also provided to submit query sequences directly to NCBI BLAST together with highlighted ambiguous residues. Both HuCaT and the Norovirus Typing Tool use the same set of reference sequences.

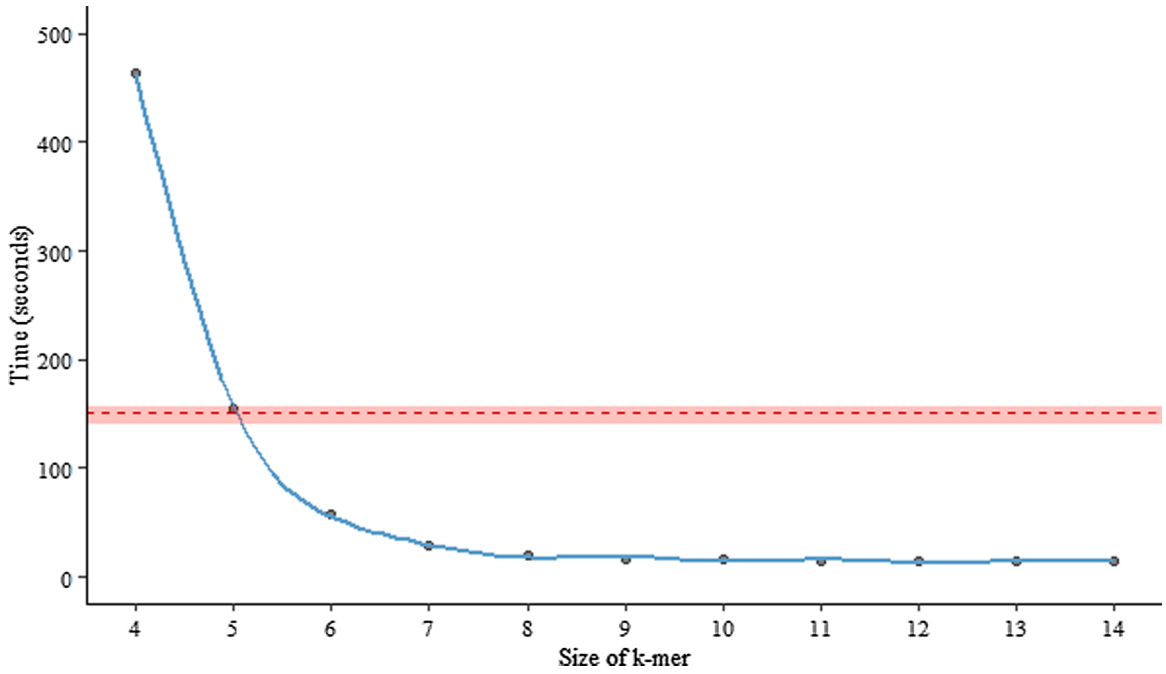

Fig. 1. Typing speed of the HuCaT tool related to size of k-mer. The blue line shows the time required to complete typing using HuCaT and the dashed line shows the median time required for typing using the BLAST-ref algorithm with reference sequences and the interquartile range is shaded in red. Each analysis was performed with the same number of query sequences (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 2
Performance of HuCaT, BLAST-ref and Norovirus Typing Tool compared to classification based on phylogenetic clustering as gold standard [6]. Query sequences included 1299 norovirus capsid sequences (query 1-Q1), 1299 norovirus polymerase sequences (query 2-Q2) and 239 sapovirus full capsid sequences (query 3-Q3). FASTA files are available upon request.

query	Average Accuracy/concordance (%)					
	HuCaT		BLAST-ref		Norovirus Typing Tool	
	genogroup	genotype	genogroup	genotype	genogroup	genotype
Q1: norovirus capsid	100	100	100	100	100	99.9
Q2: norovirus polymerase	100	99.9	100	100	NA	99.0
Q3: sapovirus capsid	100	100	100	100	100	69[1]

[1] Current version of the Norovirus Typing Tool does not type sapovirus GIV and GV at the genotype level.

Future versions of the HuCaT tool will include additional sequences of animal noro- and sapoviruses as well as options for exporting reports into different file formats.

4. Conclusions

The HuCaT tool is a fast web-based tool for accurately typing human norovirus and sapovirus sequences (https://norovirus.ng.philab.cdc.gov). It uses the same set of reference sequences as the Norovirus Typing Tool (https://www.rivm.nl/mpf/norovirus/typingtool) [16] which has been extensively used globally for norovirus and sapovirus typing but uses a BLAST-ref algorithm with reference genomes for typing. HuCaT is much faster than BLAST-ref or phylogenetic clustering approaches without compromising typing accuracy (Table 4). HuCaT is hosted on servers at CDC and is updated when new reference sequences are identified. HuCaT simplifies norovirus and sapovirus typing and generates detailed reports including alignment scores, visualizations and phylogenetic trees.

HuCaT has several limitations that will be addressed in the next version of the tool. First, only query sequences that include the entire typing region can be typed accurately, and shorter sequences and those with indels are not typed. Also, sequences with >13 % nucleotide difference compared to references sequences cannot be typed. Hence, if there are significant regions with poor sequence quality, the sequence may be untypeable. Future updates to the algorithm will account for sequence quality problems or error messages will describe why a query is untypeable. Including error messages would also remove the ambiguity when a strain is deemed “untypeable”, reserving the term for unique strains rather those that more likely contain sequencing errors. Inherent limitations of any sequence-based typing tool that uses relatively short sequences for typing, include the possibility of mistyping a novel recombinant virus or a unique strain that has differences outside the typing region.

Robust typing tools are needed for standardized genotyping of highly diverse RNA viruses such as noroviruses and sapoviruses. There are a tremendous number of bioinformatic tools that use k-mer vocabulary, including alignment-free classification tools that use k-mer frequency and distance calculations to infer phylogeny [14]. Such methods are much faster than alignment-based k-mer methods, such as BLAST, but positional information on the similarity/dissimilarity of query and reference sequences is lost without a sequence alignment. The HuCaT algorithm starts with an alignment free approach, but simultaneously builds an alignment of query-reference pairs for calculating nucleotide percent identity. As a result, HuCaT overcomes the loss of information of other k-mer based tools and accurately types input sequences.

In conclusion, genotyping of norovirus and sapoviruses is important

Table 3
Sequences with discrepant typing results when comparing the HuCaT tool, BLAST-ref and Norovirus Typing Tool.

GenBank Accession number	HuCaT		BLAST-ref		Norovirus Typing Tool	
	capsid	polymerase	capsid	polymerase	capsid	polymerase
norovirus						
MT928712	GI.3	untypeable	GI.3	GI.P3	GI.3	untypeable
MT928707	GII.4	GII.P4	GII.4	GII.P4	GII.4	untypeable
MT928708	GII.4	GII.P4	GII.4	GII.P4	GII.4	untypeable
MT928709	GII.4	GII.P4	GII.4	GII.P4	GII.4	untypeable
MT928710	GII.4	GII.P4	GII.4	GII.P4	GII.4	untypeable
MT928714	GII.4	GII.P31	GII.4	GII.P31	GII.4	untypeable
MN226991	GI.6	GI.P11	GI.6	GI.P11	GI.6	untypeable
MT928711	GI.6	GI.P11	GI.6	GI.P11	GI.6	untypeable
MT928718	GII.6	GII.P7	GII.6	GII.P7	GII.6	untypeable
MT928716	GII.17	GII.P17	GII.17	GII.P17	GII.17	untypeable
MT928715	GII.17	GII.PNA6	GII.17	GII.PNA6	GII.17	untypeable
NC_044855	GIV.NA1	GIV.PNA1	GIV.NA1	GIV.PNA1	untypeable	untypeable
sapovirus						
KM092511	GII.8		GII.8		GII.7	not applicable
KT306742	GII.8		GII.8		GII.7	not applicable
KX894314	GII.8		GII.8		GII.7	not applicable
KX894315	GII.8		GII.8		GII.7	not applicable
MF462287	GII.8		GII.8		GII.7	not applicable
MF462288	GII.8		GII.8		GII.7	not applicable
MG012452	GII.8		GII.8		GII.7	not applicable
MG012453	GII.8		GII.8		GII.7	not applicable
MG674583	GII.8		GII.8		GII.7	not applicable
MG674584	GII.8		GII.8		GII.7	not applicable
MH922771	GII.NA1[1]		GII.NA1[1]		GII.5	not applicable
MH922772	GII.NA1[1]		GII.NA1[1]		GII.5	not applicable

[1] NA = not assigned [6].

Table 4
Comparison of BLAST (NCBI or against reference (ref) sequences), Norovirus Typing Tool and HuCaT for typing of norovirus and sapovirus.

	NCBI BLAST	BLAST-ref	Norovirus Typing Tool	HuCaT
Speed	Medium	Fast	Medium	Fastest
Accuracy of typing	Low	High	High	High
Curated references	No	Yes	Yes	Yes
Genotype report	No	No	Yes	Yes
Built-in genotyping criteria	No	No	Yes	Yes
Alignment and dendrogram	No	No	Brief	Detailed
Updates	Daily	User defined	Irregular	Frequent

to monitor epidemiological trends of outbreaks and sporadic illnesses. HuCaT adds to the availability of tools for typing the highly divergent human noroviruses and sapoviruses. Together with the Norovirus Typing Tool, HuCaT includes the latest updates in new norovirus genogroups, genotypes and P-types [6].

Disclaimer

The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Funding

This work was supported by an intramural grant from CDCs food safety program.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jcv.2020.104718.

References

[1] J. Vinje, M.K. Estes, P. Esteves, K.Y. Green, K. Katayama, N.J. Knowles, et al., ICTV virus taxonomy profile: *Caliciviridae*, J. Gen. Virol. 100 (2019) 1469–1470.
[2] S.M. Ahmed, A.J. Hall, A.E. Robinson, L. Verhoef, P. Premkumar, U.D. Parashar, et al., Global prevalence of norovirus in cases of gastroenteritis: a systematic review and meta-analysis, Lancet Infect. Dis. 14 (2014) 725–730.
[3] K. Banyai, M.K. Estes, V. Martella, U.D. Parashar, Viral gastroenteritis, Lancet 392 (2018) 175–186.
[4] T. Oka, Q. Wang, K. Katayama, L.J. Saif, Comprehensive review of human sapoviruses, Clin. Microbiol. Rev. 28 (2015) 32–53.
[5] F. Bucardo, Y. Reyes, L. Svensson, J. Nordgren, Predominance of norovirus and sapovirus in Nicaragua after implementation of universal rotavirus vaccination, PLoS One 9 (2014), e98201.
[6] P. Chhabra, M. de Graaf, G.I. Parra, M.C. Chan, K. Green, V. Martella, et al., Updated classification of norovirus genogroups and genotypes, J. Gen. Virol. 100 (2019) 1393–1406.
[7] C.K. Yinda, N. Conceicao-Neto, M. Zeller, E. Heylen, P. Maes, S.M. Ghogomu, et al., Novel highly divergent sapoviruses detected by metagenomics analysis in straw-colored fruit bats in Cameroon, Emerg. Microbes Infect. 6 (2017) e38.
[8] T. Oka, Z. Lu, T. Phan, E.L. Delwart, L.J. Saif, Q. Wang, Genetic characterization and classification of human and animal sapoviruses, PLoS One 11 (2016), e0156373.
[9] L. Barclay, J.L. Cannon, M.E. Wikswo, A.R. Phillips, H. Browne, A.M. Montmayeur, et al., Emerging novel GII.P16 noroviruses associated with multiple capsid genotypes, Viruses (2019) 11.
[10] L.F. Ludwig-Begall, A. Mauroy, E. Thiry, Norovirus recombinants: recurrent in the field, recalcitrant in the lab - a scoping review of recombination and recombinant types of noroviruses, J. Gen. Virol. 99 (2018) 970–988.
[11] R.A. Bull, M.M. Tanaka, P.A. White, Norovirus recombination, J. Gen. Virol. 88 (2007) 3347–3359.
[12] A. Kroneman, E. Vega, H. Vennema, J. Vinje, P.A. White, G. Hansman, et al., Proposal for a unified norovirus nomenclature and genotyping, Arch. Virol. 158 (2013) 2059–2068.
[13] J.L. Cannon, L. Barclay, N.R. Collins, M.E. Wikswo, C.J. Castro, L.C. Magana, et al., Genetic and epidemiologic trends of norovirus outbreaks in the United States from 2013 to 2016 demonstrated emergence of novel GII.4 recombinant viruses, J. Clin. Microbiol. 55 (2017) 2208–2221.
[14] M. Diez-Valcarce, C.J. Castro, R.L. Marine, N. Halasa, H. Mayta, M. Saito, et al., Genetic diversity of human sapovirus across the Americas, J. Clin. Virol. 104 (2018) 65–72.
[15] X. Liu, H. Jahuira, R.H. Gilman, A. Alva, L. Cabrera, M. Okamoto, et al., Etiological role and repeated infections of sapovirus among children aged less than 2 years in a

cohort study in a peri-urban community of Peru, J. Clin. Microbiol. 54 (2016) 1598–1604.
[16] A. Kroneman, H. Vennema, K. Deforche, H. v d Avoort, S. Penaranda, M.S. Oberste, et al., An automated genotyping tool for enteroviruses and noroviruses, J. Clin. Virol. 51 (2011) 121–125.
[17] E. Vega, L. Barclay, N. Gregoricus, K. Williams, D. Lee, J. Vinjé, Novel surveillance network for norovirus gastroenteritis outbreaks, United States, Emerg. Infect. Dis. 17 (8) (2011) 1389–1395.