A Commandline Norovirus Typing Tool

1. Introduction

1.1 Taxonomic classification of caliciviridae

The family Caliciviridae includes viruses with a linear positive-sense RNA genome of 6.4–8.5 kb with the non-structural and structural proteins encoded by different ORFs. Virions are non-enveloped particles ranging from 27–40 nm in diameter with an icosahedral symmetry.[1] Thirteen species are placed in this family, divided among eleven genera (Figure 1).[2]In addition, unclassified caliciviruses have been detected in many animals.[3]

1.2 Discovery and genus classification of norovirus

Noroviruses are non-enveloped, single-stranded RNA viruses, belonging to the genus Norovirus in the family Caliciviridae. Most noroviruses contain three open reading frames (ORFs), except for murine noroviruses, which contain a fourth ORF (Figure 2). ORF1 codes for the polymerase, ORF2 codes for the capsid protein(VP1) and ORF3 codes for minor structural protein VP2.[4] Like most viruses, recombination in noroviruses greatly increases their genetic variation and affect their phylogenetic clusterings.[5]

Noroviruses were first isolated and identified in 1972 using electron microscopy[6], and were originally named the "Norwalk agent" after the place where an outbreak of acute gastroenteritis occurred 4 years ago.

1.3 Norovirus Epidemiology

Noroviruse is a highly contagious virus that is a common cause of acute gastroenteritis, often referred as the stomach flu or winter vomiting bug. With no specific antiviral treatment, it causes an estimated 200,000 deaths per year, including 50,000 child deaths.[7]

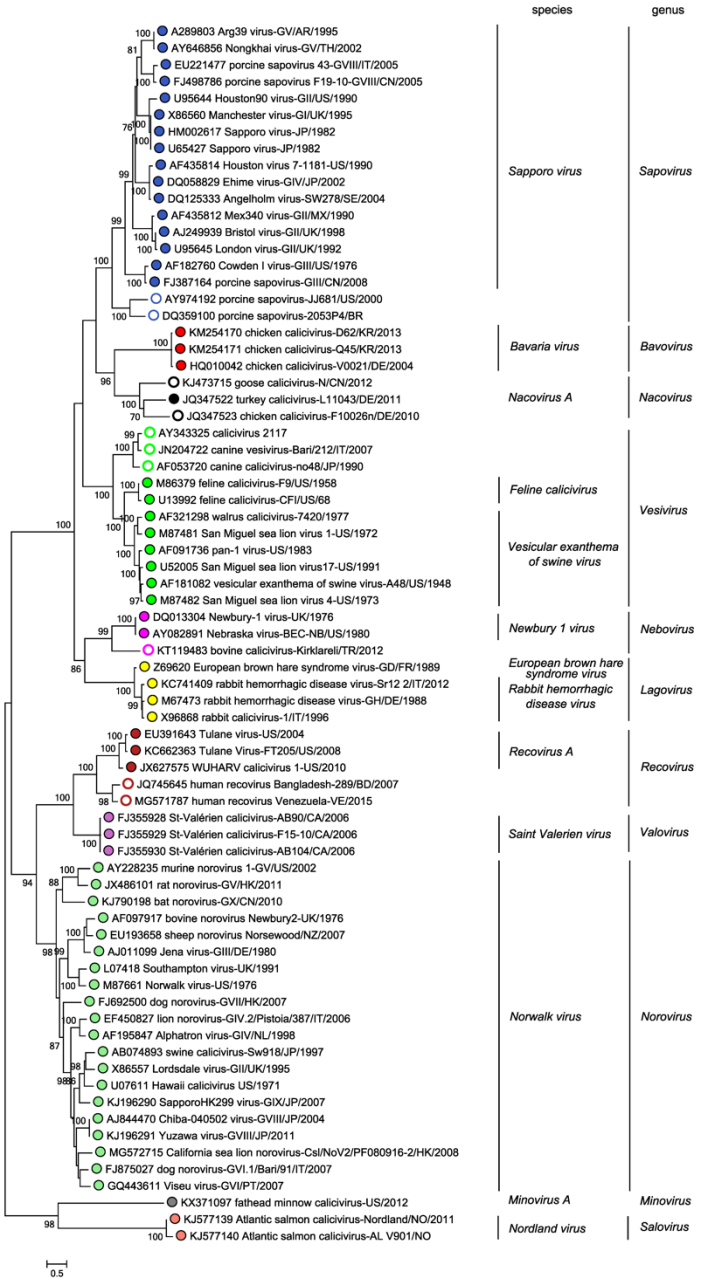

Figure 1 Caliciviridae[2]

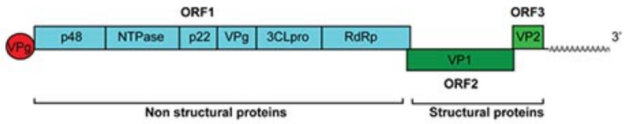

Figure 2 Norovirus[4]

GII.4 is the most common etiology in sporadic and epidemic acute gastroenteritis. Genotype GII.17 (GII.P17-GII.17) was reported in China to be responsible for gastroenteritis linked with norovirus.[8] Noroviruses can be transmitted from person to

person directly or indirectly through contaminated food, water, or environments. Foodborne transmission accounts for 10% of outbreaks caused by GII.4, 27% by all other single genotypes, and 37% by mixtures of GII.4 and others.[9]

1.4 History of viral taxonomy

Virus taxonomy addresses the classification of viruses into categories called taxa and the nomenclature for taxa.[10] Several systematic classification of virus have been proposed since the discovery of virus.

Baltimore Classification system was published in 1971. It categorizes viruses into seven groups (I-VII) based on their genome type and replication strategy.[11] While this classification system was valuable for understanding viral replication and genome organization, it has become less commonly used in modern virology, as our understanding of virus taxonomy has greatly expanded, with the advent of molecular biology and sequencing technologies.

The current official virus taxonomy is maintained by the International Committee on Taxonomy of Viruses (ICTV). The ever-expanding catalogue of new viral sequences called for a reflection on how to best adapt the current viral taxonomy to properly classify viruses discovered through metagenomics.[12]

1.5 History of norovirus taxonomy

It was recognized in the mid-1990s that norovirus strains should be organized into different genetic groups. Initially, noroviruses were classified into genogroups and genotypes using partial RdRp (RNA-dependent RNA polymerase) sequences.[13] As more sequences became available, genotyping methods using cut-off values for percentage pair-wise similarity started to be used for norovirus taxonomy,[14,15] but the high time complexity of this method made it difficult to handle large numbers of sequences. In addition, as prototype norovirus strains are often quite different from more recent strains due to accumulation of mutations, it is difficult to assign types using pairwise similarity cut-offs.

In 2013 the Norovirus Classification Working Group (NCWG) proposed a standardized nomenclature and typing system using phylogenetic clustering of the complete VP1 amino acid sequences.[16] In addition, dual typing (ORF1-RdRp=P type, ORF2=genotype) was proposed to include diversity of the partial RdRp sequences to address the increasing norovirus diversity.[16] Using 2×standard deviation(sd) criteria to group sequences into separate clusters, the number of genogroups of noroviruses were expanded to 10 (GI-GX) and the number of genotypes were expanded to 49. Based on nucleotide diversity in the RdRp region, noroviruses were divided into more than 60 P-types.[17]

It is important to note that while dual typing is widely used, many norovirus sequences available on GenBank have not been updated to reflect this nomenclature system.

1.6 Challenges within current taxonomy framework for high throughput sequencing

During the last decade, High Throughput Sequencing(HTS), also refered to as next generation sequencing, has greatly advanced our knowledge of viruses. However, it is important to consider the challenges HTS faces in norovirus taxonomy. RNA viruses like noroviruses are difficult to sequence and characterize using HTS. Firstly, they are short in genome length, approximately 7.5k. Secondly, noroviruses exhibit notable genetic diversity. Thirdly, they lack universally conserved markers and genome plasticity, which contributes challenges for common PCR-based approaches.[18]

1.7 Current norovirus taxonomy tools

Analyzing high throughput data of norovirus is crucial for gaining insights into its mechanisms of infection and transmission. Two gold-standard norovirus typing tools are available, the NoroNet typing tool[19] and the human calicivirus typing tool (HuCaT)[20].

The Norovirus Typing Tool (https://www.rivm.nl/mpf/norovirus/typingtool) uses a BLAST algorithm against a set of reference sequences followed by phylogenetic analysis to assign norovirus genotypes and P-types.[19]

Hucat(HuCaT) 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.[20]

However, both tools are web-based, creating a break in the processing of sequencing data. There is a pressing need for an efficient, locally-operated, command-line-based taxonomic classification tool for norovirus.

2 Methods

2.1 Dataset Preparation

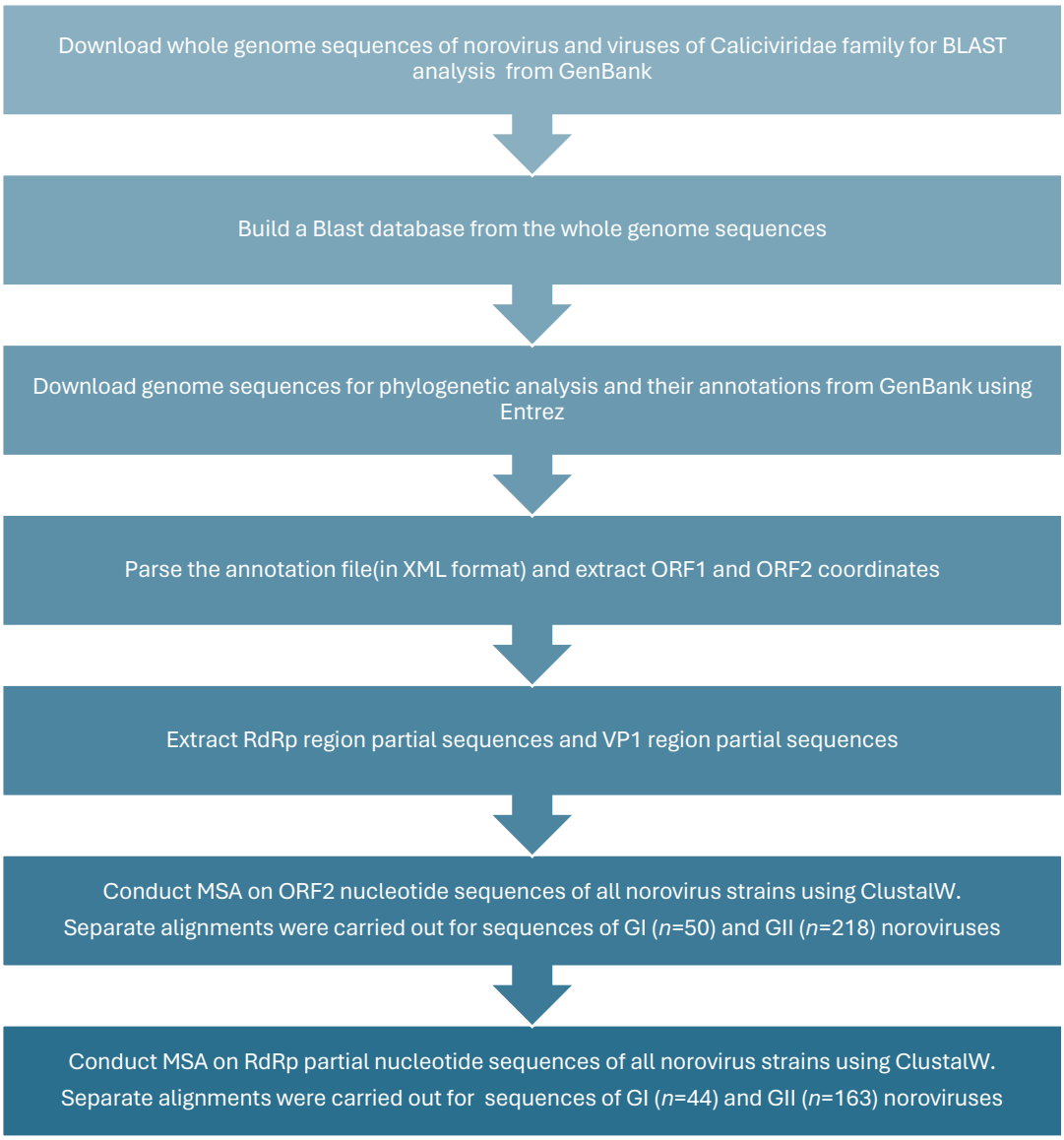

Figure 3 Workflow of Data Preparation

30 norovirus whole genome sequences from RefSeq dataset[21] and GenBank were downloaded (Figure 3, Figure 4). In addtion, representatives of other genera of the Caliciviridae family were also downloaded. The Blastn database was built with these whole genome sequences aforementioned.

GI	NC_001959 NC_039897.1 NC_044853 NC_044854.1 NC_044856.1
GII	NC_029646 NC_039475.1 NC_039476.1 NC_039477.1 NC_040876 NC_044045.1 NC_044046.1 NC_044932
GIII	NC_029645
GIV	NC_029647 NC_044855.1
GV	NC_008311 MW174170.1
GVI	NC_044047 MW662289.1 MN908340.1 MW945229.1
GVII	FJ692500 OL757872.1
GVIII	AB985418.2
GIX	OR050586.1 OR050585.1 OR050584.1
GX	KJ790198.1 MF373609.1

Figure 4 Reference Sequences for Blast Analysis

305 nucleotide sequences for VP1 region phylogenetic analyses, and 232 nucleotide sequences for RdRp region phylogenetic analyses, representing the genetic diversity of all norovirus genogroups and genotypes were downloaded from GenBank, as described previously[16](last download date: March 16 2024). 305 Complete ORF2 sequences (ranging from 1560 to 1767 nucleotides in length) were extracted from the former using Biopython[22]. 232 partial nucleotide sequences (762 nucleotides) of the RdRp region at the 3′-end of ORF1 were extracted from the latter (Figure 5).

Group	number of ORF1	min length of ORF1	max length of ORF1	number of ORF2	min length of ORF2	max length of ORF2
GI	41	777	5400	34	1593	1653
GII	121	442	5358	96	1596	1674
GIII	2	5043	5055	3	1560	1569
GIV	4	685	5064	12	1665	1737
GV	2	5064	5099	2	1626	1767
GVI	8	701	5241	8	1719	1749
GVII	2	5076	5076	2	1587	1587
GVIII	0	0	0	2	1668	1668
GNA1	1	4434	4434	1	1632	1632
GNA2	2	5284	5284	2	1668	1668
GIX	0	0	0	2	1668	1668
GX	2	4104	4869	2	1602	1623

Figure 5 Reference Sequences for Phylogenetic Analysis

For the RdRp region, an alignment of all 232 norovirus sequences was generated, as well as separate alignments for GI (n=44) and GII (n=163) sequences. ORF2 nucleotide sequences of all strains were aligned using ClustalW. Separate alignments were carried out for VP1 nucleotide sequences of GI (n=50) and GII (n=218) noroviruses. These two alignments are used as reference dataset for phylogenetic analysis in the next step.

2.2 Assignment of Genogroup

In the first step, the query sequence is analyzed against the Blast database. If the expectation (E)-value of the top hit is less than 10^{-5}, a genus (or species, genogroup) will be assigned. Matching length and genome localization are also determined.

2.3 Phylogenetic analysis

In the second step, phylogenetic analysis is performed only if the genogroup of the query is identified and the matching region has a specified minimal length in specified region of the genome. Profile alignments of the query sequence are performed using ClustalW against the alignment of the reference sequences generated in 2.1.

Phylogenetic trees are constructed using the neighbor-joining method with a HKY substitution, (phangorn[23]). Supporting values are written into the phylogenetic tree as branch labels.

2.4 Future work: Assignment of Genotype

The phylogenetic tree should be parsed. A genotype is assigned only if the bootstrap value of the query to its nearest neighbor exceeds 70%.

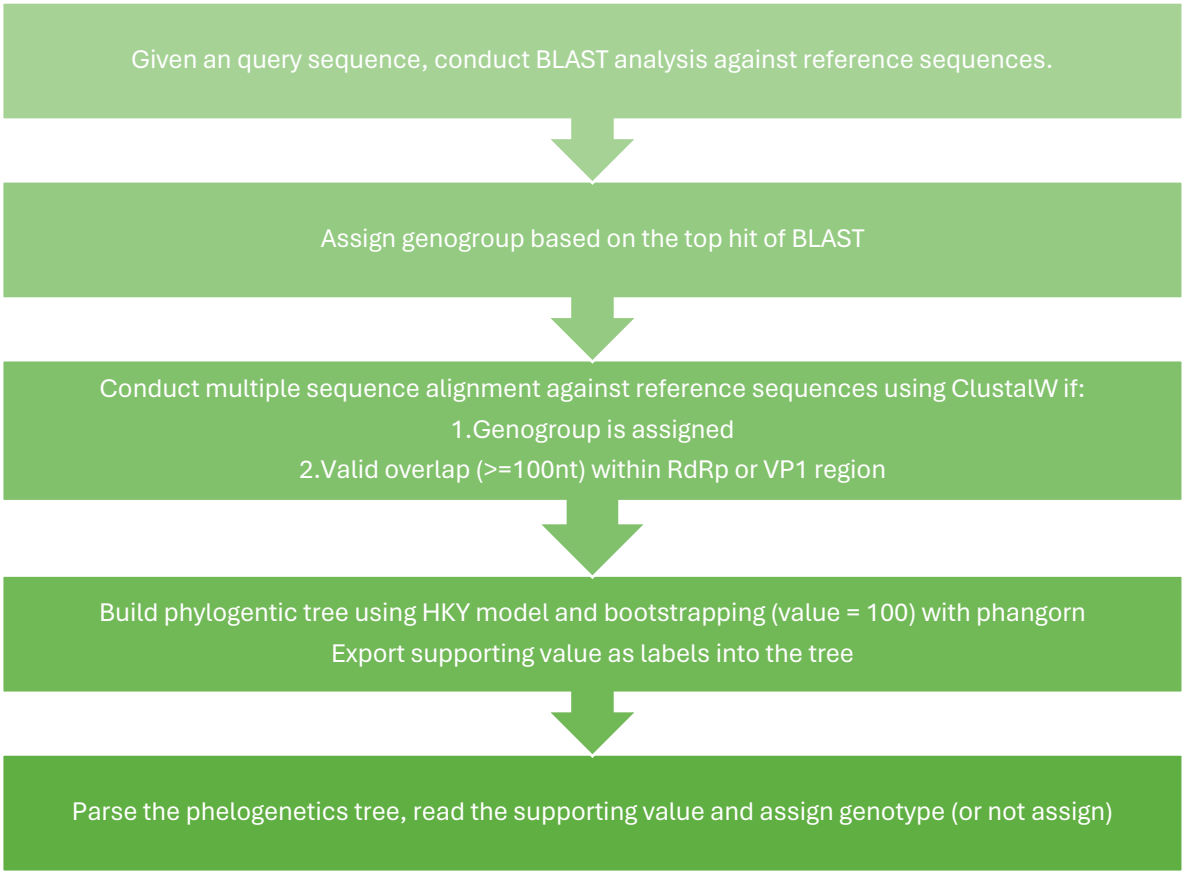

Figure 6 Genotype assignment

3 Results

3.1 Genogroup Assignment

The test dataset contains 5986 sequences of the Caliciviridae family, including noroviruses. The results generated by the RIVM online typing tool served as the ground truth. The accuracy of our group assignment is 98.66% (Figure 7). Most of the false negative and false positive cases fall under norovirus genogroup GIV, GVI, GIX and GX. Because GVI, GVIII, GIX and GX noroviruses are not included in the reference dataset for BLAST analysis on RIVM, if we exclude these test cases, the accuracy of our group assignment would be higher. 14 sequences in the test dataset fail to be assigned in our typing tool but are correctly assigned as *other positive-strand ssRNA virus* on RIVM. There might be two possibilities: the database of RIVM contain whole genome sequences of other virus families, or RIVM conducts a BLAST analysis against a database of another virus families first.

True Positive 4222	False Negative 75
False Positive 5	True Negative 1684

Figure 7 Genogroup Assignment Result

[1] Vinje J, et al.; ICTV Report Consortium. 2019. ICTV virus taxonomy profile: caliciviridae. J Gen Virol. 100(11):1469–1470.
[2] Vinjé et al., (2019):ICTV Virus Taxonomy Profile: Caliciviridae, Journal of General Virology, 100, 1469–1470.
[3] Wille M, Eden JS, Shi M et al. Virus-virus interactions and host ecology are associated with RNA virome structure in wild birds. Mol. Ecol. 2018, DOI: 10.1111/mec.14918.
[4] COMPARE Europe (2015). Reference genomes - Norovirus. Available at: https://www.compare-europe.eu/library/reference-genomes/norovirus (Accessed: March 16, 2024).
[5] Worobey M, Holmes EC. Evolutionary aspects of recombination in RNA viruses. J Gen Virol. 1999;80:2535–43.
[6] Kapikian AZ, Wyatt RG, Dolin R, Thornhill TS, Kalica AR, Chanock RM. 1972. Visualization by immune electron microscopy of a 27-nm particle associated with acute infectious nonbacterial gastroenteritis. J Virol 10:1075–1081.
[7] World Health Organization. (2022). Norovirus. Retrieved from https://www.who.int/teams/immunization-vaccines-and-biologicals/diseases/norovirus.
[8] Zhou, X., Kong, D. G., Li, J., Pang, B. B., Zhao, Y., Zhou, J. B., et al. (2019). An outbreak of gastroenteritis associated with GII. 17 Norovirus-contaminated secondary water supply system in Wuhan, China, 2017. Food Environ. Virol. 11, 126–137. doi: 10.1007/s12560-019-09371-7
[9] Verhoef L, Hewitt J, Barclay L, Ahmed SM, Lake R, Hall AJ, et al. Norovirus Genotype Profiles Associated with Foodborne transmission. Emerg Infect Dis. 2015;21:592–599. pmid:25811368
[10] Kuhn JH. Virus taxonomy. In: Bamford DH and Zuckerman M (eds). Encyclopedia of Virology, 4th edn. Oxford: Academic Press; 2021. pp. 28–37.
[11] Baltimore D. 1971. Viral genetic systems. Trans N Y Acad Sci 33:327–332. 10.1111/j.2164-0947.1971.tb02600.x.
[12] Simmonds, P., Adams, M., Benkő, M. et al. Virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15, 161–168 (2017). https://doi.org/10.1038/nrmicro.2016.177
[13] Vinjé J, Koopmans MP. Molecular detection and epidemiology of small round-structured viruses in outbreaks of gastroenteritis in the Netherlands. J Infect Dis. 1996;174:610–615. doi: 10.1093/infdis/174.3.610.
[14] Vinje J,Hamidjaja RA,Sobsey MD. Development and application of acapsid VP1(region D) based reverse transcription PCR assay for genotyping of genogroup I and II noroviruses. J Virol Methods 2004;116(March(2)):109–17.
[15] Zheng DP, Ando T, Fankhauser RL, Beard RS, Glass RI, Monroe SS. Norovirus classification and proposed strain nomenclature. Virology 2006;346(March(2)):312–23.
[16] Kroneman A, Vega E, Vennema H, Vinjé J, White PA, et al. Proposal for a unified norovirus Nomenclature and genotyping. Arch Virol. 2013;158:2059–2068. doi: 10.1007/s00705-013-1708-5.
[17] Chhabra, P. et al. Updated classification of norovirus genogroups and genotypes. J. Gen. Virol. 100, 1393–1406 (2019).

[18] Fitzpatrick, A.H.; Rupnik, A.; O'Shea, H.; Crispie, F.; Keaveney, S.; Cotter, P. High Throughput Sequencing for the Detection and Characterization of RNA Viruses. Front. Microbiol. 2021, 12, 190.
[19] Kroneman A, Vennema H, Deforche K, v d Avoort H, Penaranda S, Oberste MS, et al., An automated genotyping tool for enteroviruses and noroviruses, J. Clin. Virol 51 (2011) 121–125.
[20] Tatusov RL, Chhabra P, Diez-Valcarce M, Barclay L, Cannon JL, Vinjé J. 2021. Human calicivirus typing tool: a web-based tool for genotyping human norovirus and sapovirus sequences. J Clin Virol 134:104718. https://doi.org/10.1016/j.jcv.2020.104718.
[21] Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. Nucleic Acids Res. 2015 Jan;43(Database issue):D571-7 PubMed PubMedCentral
[22] Cock, P. J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423 (2009).
[23] Schliep K (2011). "phangorn: phylogenetic analysis in R." Bioinformatics, 27(4), 592–593. doi:10.1093/bioinformatics/btq706.