Nucleic Acids Research, 2025, 1–3
https://doi.org/10.1093/nar/gkaf325
Web Server issue

ClipKIT in the browser: fast online trimming of multiple sequence alignments for phylogenetics

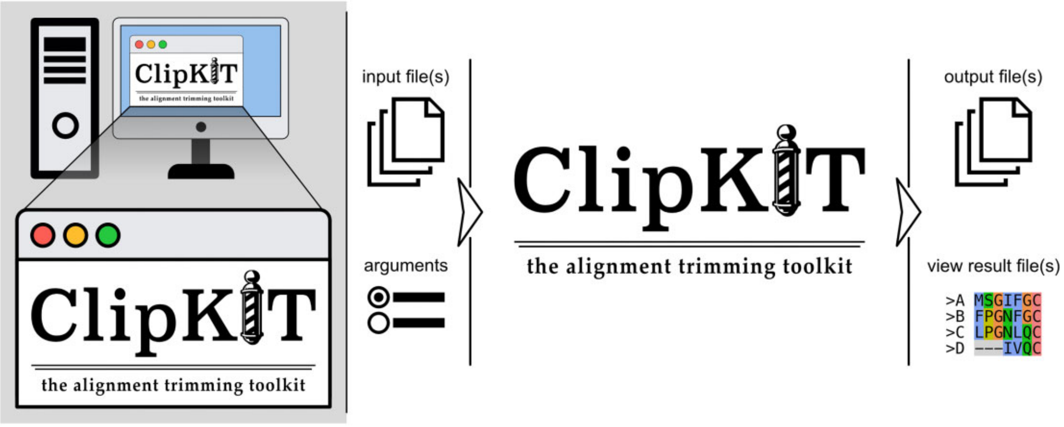
Jacob L. Steenwyk [1,*], Jeffrey T. Loucks [2], Thomas J. Buida [3]

[1] Howard Hughes Medical Institute and the Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, United States
[2] Independent Researcher, Durham, NC 27707, United States
[3] Independent Researcher, Nashville, TN 37209, United States
[*] To whom correspondence should be addressed. Email: jlsteenwyk@berkeley.edu

Abstract

Multiple sequence alignment trimming can help improve phylogenetic signal and reduce computational load. ClipKIT trims multiple sequence alignments by retaining phylogenetically informative sites and removing all others. Here, we present a web browser application for ClipKIT, which supports DNA, protein, and codon data types in FASTA format. The web browser application can process one or many multiple sequence alignment files, which users can subsequently download. Users can also view the trimmed multiple sequence alignment using a web-based multiple sequence alignment viewer. ClipKIT is available at https://clipkit.genomelybio.com, is free and open to all users, and there is no login requirement. ClipKIT in the browser aims to broaden the accessibility of web-based tools for phylogenetics research.

Graphical abstract



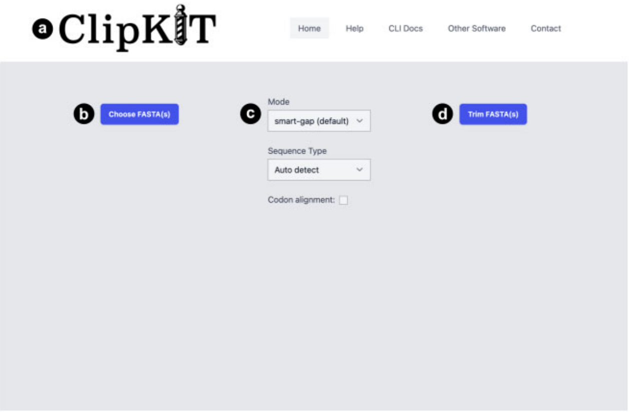
Introduction

ClipKIT is an efficient software that conducts multiple sequence alignment trimming for phylogenomics [1]. While most algorithms aim to identify and remove highly divergent sites in multiple sequence alignments [2], ClipKIT identifies phylogenetic informative sites and removes all others. Benchmarking revealed that ClipKIT outperformed other multiple sequence alignment trimming tools [1–5]. ClipKIT is flexible, featuring numerous modes for multiple sequence alignment trimming.

Although ClipKIT has been adopted by numerous researchers, it is only available as a command-line tool and is, therefore, difficult for nonexpert bioinformaticians to use. Moreover, there is a dearth of tools that enable multiple sequence alignment trimming in the browser [6], underscoring the broad inaccessibility of trimming multiple sequence alignments to nonexperts.

Here, we present ClipKIT in the browser, a user-friendly application for multiple sequence alignment trimming using cloud-based resources. Currently, ClipKIT runs using resources from Amazon Web Services (https://aws.amazon.com/). Since first launch, ClipKIT in the browser has processed ~250 files per month. When used with other tooling—such as the MAFFT online service [7] and the IQTREE web server [8]—ClipKIT in the browser enables rigorous phylogenetic analysis without needing command-line expertise,

Received: January 21, 2025. **Revised:** March 12, 2025. **Editorial Decision:** April 8, 2025. **Accepted:** April 14, 2025

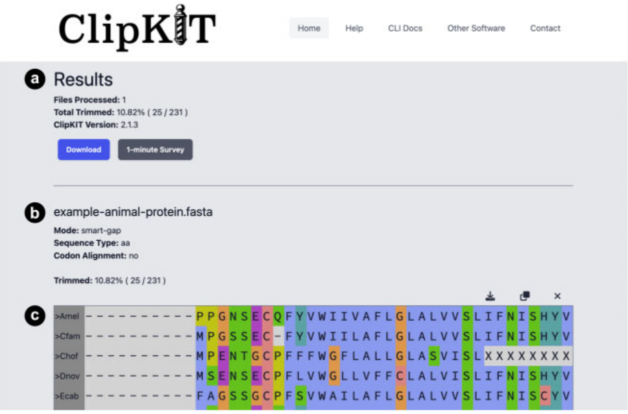


Mode: smart-gap (default)

Sequence Type: Auto detect

Codon alignment: □

Figure 1. ClipKIT in the browser landing page. The landing page directly takes users to where files are uploaded. (**a**) The header bar provides key links, including the "Home" page (depicted here), the "Help" page that contains additional information about using ClipKIT in the browser and exemplary files to test running the software. Other tabs include documentation for the command-line interface (CLI) tool, other software our team has developed—such as PhyKIT [9, 10], BioKIT [11], OrthoHMM [12], OrthoSNAP [13], and other algorithms—that may be of interest to users, and, lastly, contact information in case users have feature requests, comments, or questions. (**c**) Users can specify what trimming mode to use, and sequence type (amino acid, nucleotide, or the default, auto-detect). Users can also specify if the input data is a codon alignment; if so, the sequence type is ignored and assumed to be nucleotides. (**d**) The "Trim FASTA(s)" button catalyzes the file processing using cloud-based computing resources.



Results

Files Processed: 1
Total Trimmed: 10.82% (25 / 231)
ClipKIT Version: 2.1.3

example-animal-protein.fasta

Mode: smart-gap
Sequence Type: aa
Codon Alignment: no

Trimmed: 10.82% (25 / 231)

```
>Amel -----------PPGNSECQFYVWIIVAFLGLALVVSLIFNISHYV
>Cfam -----------MPGSSEC-FYVWIILAFLGLALVVSLIFNISHYV
>Chof -----------MPENTGCPFFFWGFLALLGLASVISLXXXXXXXX
>Dnov -----------MSENSECPFLVWGLLVFFCLALVISLIFNISHYV
>Ecab -----------FAGSSGCPFSVWAILAFLGLALVVSLIFNISCYV
```

Figure 2. ClipKIT in the browser results and output. (**a**) Summary information about all processed files is provided. This includes how many files were processed and the total percentage and number of trimmed sites as well as the total number of sites examined. The version of ClipKIT is also specified. (**b**) Thereafter, information about individually processed files is provided, including the trimming mode used, the sequence type, whether the file represents a codon alignment, and the percentage of sites trimmed. (**c**) A multiple sequence alignment viewer enables users to examine the resulting trimmed multiple sequence alignment file easily. Buttons are included above the viewer for downloading, copying the results to the clipboard, and closing the file information.

broadening the accessibility of commonplace analyses in evolutionary biology.

ClipKIT web-application

ClipKIT in the browser works on the latest versions of major browsers like Chrome, Firefox, Edge, and Safari. The web interface provides a "Help" section, which includes exemplary files and other helpful information for using the toolkit (Fig. 1a). Minimally, users upload a multiple sequence alignment file. Then, the input file and default argument specifications are sent and processed by cloud resources, alleviating the user from providing any computational resources. Elements of the web interface are discussed below.

Input data and arguments

ClipKIT in the browser takes one or more FASTA files as input (Fig. 1b). The user can then specify the trimming mode, sequence type (default is to auto-detect sequence type), and whether the multiple sequence alignment is a codon-based multiple sequence alignment (Fig. 1c). The various ClilpKIT modes are kpic (keep only parsimony informative and constant sites), kpi (keep only parsimony informative sites), gappy (keep sites with few gap characters based on a hard threshold), and smart-gap (dynamic determination of gappyness threshold); combinations of kpic/kpi and gappy/smart-gap (such as kpic-gappy) can also be used. ClipKIT also supports codon-based trimming. If one site in a codon is trimmed, the whole codon will be removed. ClipKIT also has a c3 mode for trimming, eliminating the third codon position from an alignment. Thereafter, users can catalyze file processing (Fig. 1d). After doing so, the "Trim FASTA (s)" button will update with a loading spinner, and the text will read as "Trimming," indicating to the user that processing is underway.

Results and output information

After processing the file(s) using cloud resources, the browser will automatically update with the results and descriptive statistics. Specifically, there are summary statistics about the number of files processed and the total percentage and number of trimmed sites (Fig. 2a). The specific version of ClipKIT being used in the browser application is also noted; thus, there will be congruence in the results outputted by ClipKIT in the browser and the corresponding version implemented in the command-line. Users can also download individual result files or all results at this time.

Thereafter, information regarding individually processed files is presented (Fig. 2b). This includes what trimming mode was used, the sequence type, if the alignment is a codon, and the percentage of the alignment trimmed, including the number of sites trimmed out of the total number of possible sites. A multiple sequence alignment viewer also displays the trimmed alignment (Fig. 2c), enabling users to inspect the resulting output file quickly. Above the numerous sequence alignments are buttons to download, copy results to the clipboard, or close an individually processed file.

Processed files are maintained for the duration of the user's session. The input file is not kept during the session. We note that users are responsible for ensuring that any data shared do not violate privacy laws or contain personally identifiable information unless explicitly authorized and that all data are handled following the General Data Protection Regulation.

Acknowledgements

We thank the King lab for the helpful discussion and comments.

Author contributions: Jacob L. Steenwyk (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Jeffrey T. Loucks (Software [equal], Visualization [equal]), and Thomas J. Buida (Conceptualization [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Project administration [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—review & editing [equal])

Conflict of interest

J.L.S. is an advisor to ForensisGroup Inc. J.L.S. is a scientific consultant to FutureHouse Inc. During this project, J.L.S. was a Bioinformatics Visiting Scholar at MantleBio Inc. Entities stated herein had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Funding

J.L.S. is a Howard Hughes Medical Institute Awardee of the Life Sciences Research Foundation. Funding to pay the Open Access publication charges for this article was provided by Howard Hughes Medical Institute Award of the Life Sciences Research Foundation.

Data availability

ClipKIT is freely accessible in the browser at https://clipkit.genomelybio.com. Users are responsible for ensuring that any data shared does not violate privacy laws or contain personally identifiable information unless explicitly authorized and that all data are handled in accordance with the General Data Protection Regulation (GDPR). The full terms of use are available at https://clipkit.genomelybio.com/#/terms. The user interface was developed using Vue and JavaScript. The backend was written in Python. Tutorials and documentation are available on the "Help" page https://clipkit.genomelybio.com/#/help.

References

1. Steenwyk JL, Buida TJ, Li Y *et al.* ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol* 2020;**18**:e3001007. https://doi.org/10.1371/journal.pbio.3001007
2. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 2007;**56**:564–77. https://doi.org/10.1080/10635150701472164
3. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 2010;**10**:210. https://doi.org/10.1186/1471-2148-10-210
4. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;**25**:1972–3. https://doi.org/10.1093/bioinformatics/btp348
5. Dress AW, Flamm C, Fritzsch G *et al.* Noisy: identification of problematic columns in multiple sequence alignments. *Algorithms Mol Biol* 2008;**3**:7. https://doi.org/10.1186/1748-7188-3-7
6. Dereeper A, Guignon V, Blanc G *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* 2008;**36**:W465–9. https://doi.org/10.1093/nar/gkn180
7. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinf* 2019;**20**:1160–6. https://doi.org/10.1093/bib/bbx108
8. Trifinopoulos J, Nguyen L-T, von Haeseler A *et al.* W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res* 2016;**44**:W232–5. https://doi.org/10.1093/nar/gkw256
9. Steenwyk JL, Buida TJ, Labella AL *et al.* PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics* 2021;**37**:2325–31. https://doi.org/10.1093/bioinformatics/btab096
10. Steenwyk JL, Martínez-Redondo GI, Buida TJ *et al.* PhyKIT: a multitool for phylogenomics. *Curr Protoc* 2024;**4**:e70016. https://doi.org/10.1002/cpz1.70016
11. Steenwyk JL, Buida TJ, Gonçalves C *et al.* BioKIT: a versatile toolkit for processing and analyzing diverse types of sequence data. *Genetics* 2022;**221**:iyac079. https://doi.org/10.1093/genetics/iyac079
12. Steenwyk JL, Buida TJ, Rokas A *et al.* OrthoHMM: Improved inference of ortholog groups using hidden markov models. bioRxiv, https://doi.org/10.1101/2024.12.07.627370, 21 December 2024, preprint: not peer reviewed.
13. Steenwyk JL, Goltz DC, Buida TJ *et al.* OrthoSNAP: a tree splitting and pruning algorithm for retrieving single-copy orthologs from gene family trees. *PLoS Biol* 2022;**20**:e3001827. https://doi.org/10.1371/journal.pbio.3001827

Received: January 21, 2025. **Revised:** March 12, 2025. **Editorial Decision:** April 8, 2025. **Accepted:** April 14, 2025