*Bioinformatics Advances*, 2023, **00**, vbad083
https://doi.org/10.1093/bioadv/vbad083
Advance access publication 12 July 2023
**Application Note**

OXFORD

Software

# TraceTrack, an open-source software for batch processing, alignment and visualization of sanger sequencing chromatograms

**Kveta Brazdilova[1,2], David Prihoda[1,2], Quynh Ton[3], Heath Klock[3] and Danny A. Bitton [1,*]**

[1]Discovery Informatics, MSD Czech Republic s.r.o, Prague 150 00, Czech Republic
[2]Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology, Prague 160 00, Czech Republic
[3]Protein Sciences, MRL, Merck & Co., Inc., Boston, MA 02115, USA

*To whom correspondence should be addressed.

Associate Editor: Alex Bateman

## Abstract

**Motivation:** Despite the advent of next-generation sequencing technology and its widespread applications, Sanger sequencing remains instrumental for molecular biology subcloning work in biological and medical research and indispensable for drug discovery campaigns. Although Sanger sequencing technology has been long established, existing software for processing and visualization of trace file chromatograms is limited in terms of functionality, scalability and availability for commercial use.

**Results:** To fill this gap, we developed TraceTrack, an open-source web application tool for batch alignment, analysis and visualization of Sanger trace files. TraceTrack offers high-throughput matching of trace files to reference sequences, rapid identification of mutations and an intuitive chromatogram analysis. Comparative analysis between TraceTrack and existing software tools highlights the advantages of TraceTrack with regards to batch processing, visualization and export functionalities.

**Availability and implementation:** TraceTrack is available at https://github.com/MSDLLCpapers/TraceTrack and as a web application at https://tracetrack.dichlab.org. TraceTrack is a web application for batch processing and visualization of Sanger trace file chromatograms that meets the increasing demand of industrial sequence validation workflows in pharmaceutical settings.

**Contact:** danny.bitton@merck.com

**Supplementary information:** Supplementary data are available at *Bioinformatics Advances* online.

## 1 Introduction

Although next-generation sequencing (NGS) technologies have revolutionized biomedical research and high-throughput methodologies in genomics and transcriptomics (Gwinn *et al.*, 2019), Sanger sequencing still plays a pivotal role in sequence validation primarily due to its simplicity, accuracy and cost-effectiveness. Even where NGS is widely used, Sanger sequencing remains important for validation of variants, completing hard to sequence regions (Engel *et al.*, 2014), short tandem repeat (STR) analysis (Fu *et al.*, 2018) and other tasks. In pharmaceutical research, Sanger sequencing is routinely and extensively used for primer walking (Gwinn *et al.*, 2019), cloning junction verifications and point mutation detection and supporting a wide range of high-throughput pipelines for protein design, biocatalysis and antibody discovery, to name just a few. Manual interpretation of Sanger sequencing chromatograms and their comparison to reference sequences may take an experienced researcher approximately 5 min to complete for a single sample. However, when processing multiple batches of sequencing data in 96-well plates, chromatogram matching and analysis becomes a tedious and error-prone process. Thus, the heavy workload of sequence analysis and validation in drug discovery workflows demands the development of automated, high-throughput Sanger-based validation tools that can dramatically reduce manual curation of sequences, and consequently save cost and time. In this regard, many applications for automated matching and alignment of Sanger trace files to reference sequences have been reported to date (Chao *et al.*, 2021; Rausch *et al.*, 2020; Stucky, 2012). Nevertheless, many of the published software tools are no longer being supported (Stucky, 2012), not freely available for widespread commercial use or cannot handle batch processing, therefore, are unable to support large-scale chromatogram analysis in pharmaceutical settings. A notable example is Tracy (Rausch *et al.*, 2020), a recent open-source solution that represents a step forward for automating chromatogram analysis. Tracy features advanced functionalities for genome assembly, base-calling and trace file alignment through an intuitive user interface, yet it remains unsuitable for high-throughput analysis, since trace files can only be processed in bulk using a command-line interface, which limits its use by non-tech-savvy users. Similarly, the recently published 'R' packages 'sangeranalyseR' and 'SangeR' (Chao

**Received:** May 4, 2023. **Revised:** June 23, 2023. **Editorial Decision:** June 24, 2023. **Accepted:** July 11, 2023

et al., 2021; Schmid et al., 2022) offer an advanced toolbox to streamline and speed up chromatogram analysis, nevertheless it requires significant knowledge of the R programming language.

To address this unmet need we developed TraceTrack, an open-source web application for batch alignment of trace files, mutation detection and chromatogram visualization that can significantly reduce the time of sequence validation (Fig. 1). Through an intuitive user-interface TraceTrack enables simultaneous matching and alignment of multiple trace files to multiple reference sequences. Each resultant alignment is displayed separately with highlighted sequence variations. Original chromatograms can also be interrogated directly. TraceTrack is an extensible software tool not only enabling sequence validation at speed and scale, but also aiming to encourage scientists to add additional features that can further reduce manual work and consequently accelerate drug discovery campaigns.

## 2 Implementation

TraceTrack features a user-friendly interface and an extensible computational backend which are described in detail in the following sections. Versions of all packages used at the time of deployment are listed in Supplementary Table S1.

### 2.1 Application overview

The web application is composed of three main components: (i) an html and JavaScript user interface with a Flask backend, (ii) a worker with a Celery asynchronous task queue, where computation takes place and (iii) a Redis in-memory database that is used to temporarily store information about the task and its respective results. Each of these components can be run separately in a terminal or all three parts can be run together in a single Docker container. TraceTrack offers an easy and versatile deployment of all components on a local machine or on a remote server. Alternatively, deployment can be further simplified by running the application with synchronous tasks, thereby excluding the task queue and the database components.

TraceTrack employs BioPython (Cock et al., 2009) built-in functions for sequence manipulations and for extracting information from ab1 trace files. Yet, it offers numerous new custom classes and functions for storing trace and reference sequences as well as for sequence alignment. To perform multiple sequence alignment (MSA) of traces and references sequences, TraceTrack utilizes the Clustal Omega algorithm (Sievers et al., 2011) via the Bio.Align package (Cock et al., 2009). In cases where only two sequences are being aligned the same algorithm is used for consistency. The tool architecture and basic workflows are illustrated in Supplementary Figure S1.

### 2.2 File input

TraceTrack's input page displays buttons for uploading trace and reference sequence files (Fig. 2A). Trace files can be uploaded in .ab1 format or as a zipped archive with multiple .ab1 files. TraceTrack accepts a single file with reference sequences in one of the following formats: .xlsx, .csv, .fasta or .gb (GenBank). Hundreds of traces and reference sequences can be uploaded and processed simultaneously, and users can optionally define distinct sub-groups of traces to be aligned separately.

### 2.3 File processing

Once user uploads the trace and reference sequence files, TraceTrack generates a sequence list and a reference database in preparation for the subsequent matching and MSAs. When

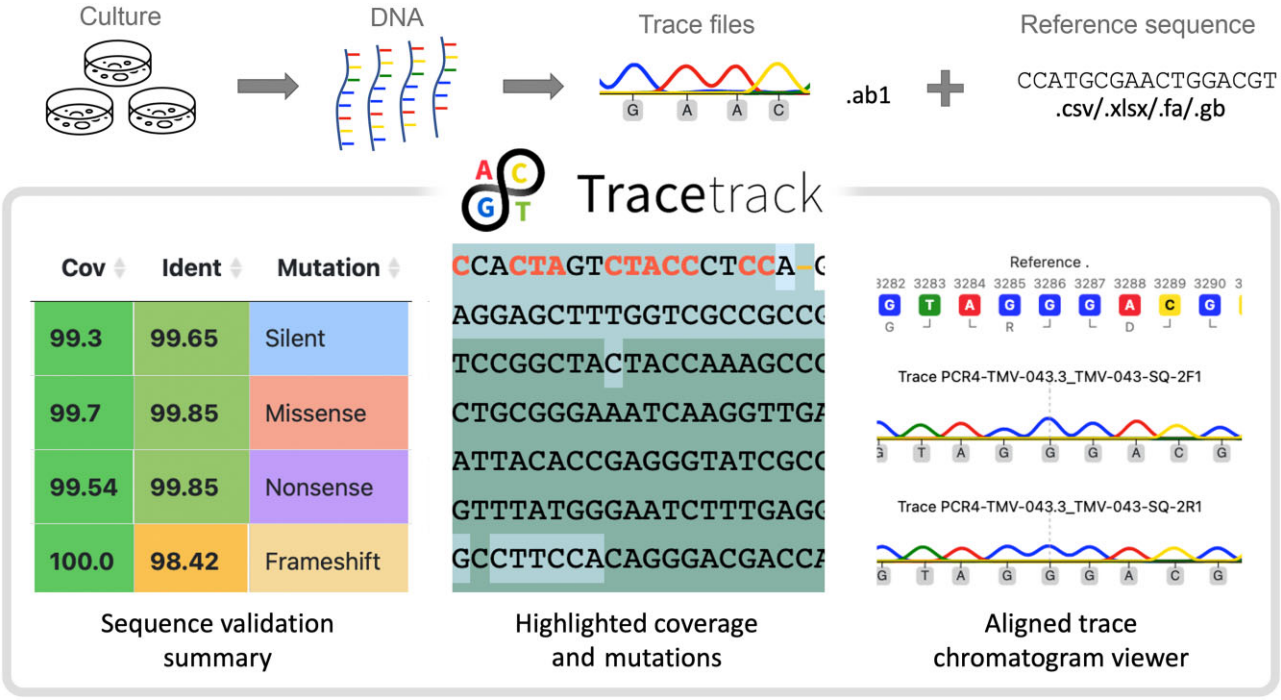


**Figure 1.** TraceTrack enables high-throughput matching of Sanger chromatograms files and reference sequences. (Top) TraceTrack is routinely and extensively used to validate sequences in subcloning work. Sanger sequencing trace files (.ab1 file format) are matched with reference sequences (.csv, .xslx, .fa and .gb file formats) via systematic MSAs. (Bottom) For each resultant alignment, TraceTrack reports consensus sequence alongside percentage coverage, identity and the type of detected mutations as well as enables users to interrogate trace files via a dedicated trace viewer

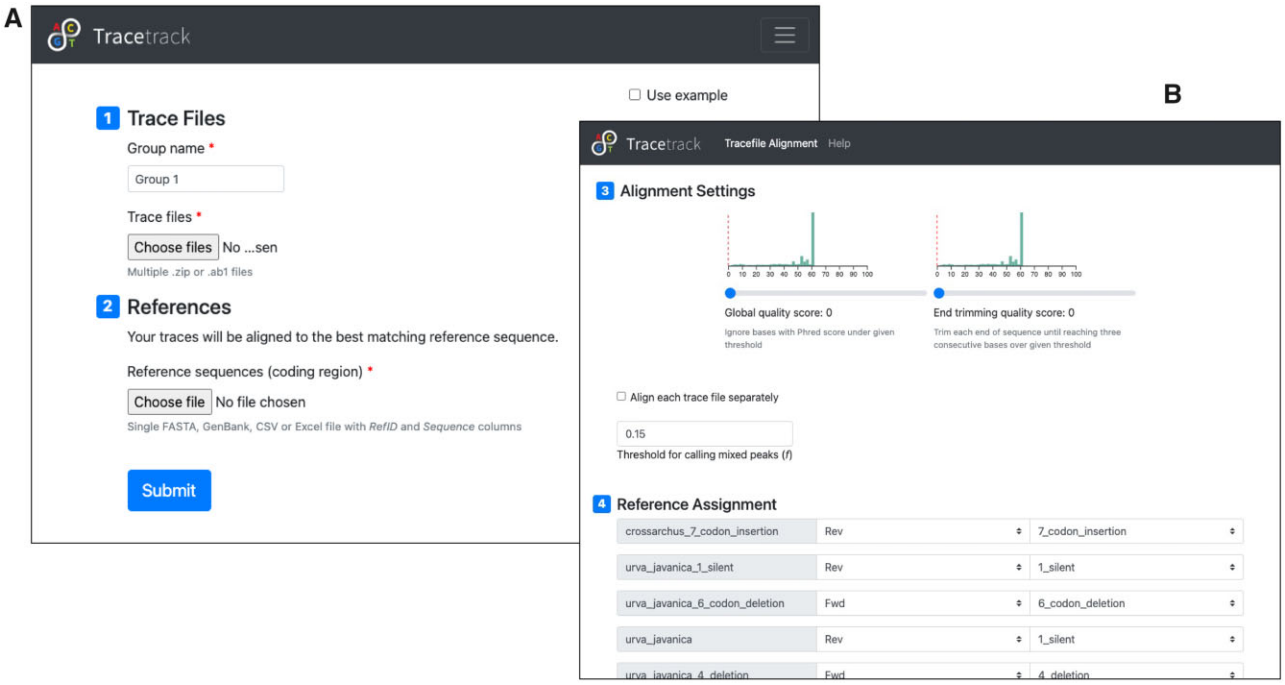


**Figure 2.** TraceTrack input page. (A) Users can upload hundreds of Sanger trace files and reference sequences for matching. (B) Settings page with options for choosing references and read directions. The data analysed in these screenshots are sequencing reads from *Urva javanica* and *Crossarchus obscurus* obtained from the Barcode of Life Database (Sujeevan and Hebert, 2007)

the reference is provided in GenBank format, TraceTrack simply extracts the coding sequence (CDS) features under default settings. In any other case, the entire sequence is considered as the coding sequence and translation begins at the first codon in the sequence and ends with the last triplet of the sequence.

The software inspects the trace chromatograms from each file for mixed peaks. Each chromatogram is composed of four traces that correspond to each of the DNA nucleotides (A, T, G and C). In principle, only a single peak in one of the traces at a given position should be identified. In cases where more than one peak is present, TraceTrack detects it, but it reports it only when the peaks are detected in high signal-to-noise regions (based on a data-derived threshold) as opposed to mixed peaks in low-complexity regions that are ignored throughout. To define a mixed peak, the area of the secondary peak must be at least '$f$' times the area of the main peak and the secondary peak must be at least '$f$' times the height of the main peak, where '$f$' is a number between 0 and 1 (as derived from sample data; default 0.15) and the secondary trace is concave around the centre of the main peak. The default value for '$f$', 0.15, was determined empirically to be as sensitive as possible, without picking up too much noise. Influence of the value of '$f$' on the detected mixed peaks is shown in Supplementary Figure S2.

Finally, TraceTrack pre-assigns all trace files to their respective reference sequence, either by sequence similarity or by matching filenames.

## 2.4 Reference assignment and alignment settings

TraceTrack settings page displays a list of trace files and their pre-assigned reference sequences as well as several settings options (Fig. 2B). Each uploaded trace file is assigned to the best matching reference sequence, as follows. If the reference ID is contained in the trace file name, the trace file is assigned to it. Otherwise, the trace file is aligned to all references in both directions and the match with the highest score is chosen. Sequencing direction is determined automatically by evaluating read matches in both directions. These automatic reference assignments can be refined by the user using dropdown lists on the settings page. Trace sequences may also be filtered or trimmed according to user-defined quality or trimming thresholds. When a base quality threshold is set, the software discards all positions with a lower quality. When an end trimming threshold is set, the tool trims the ends of each trace sequence until three consecutive bases with quality higher than the threshold are encountered. This is intended to remove ends of reads with low quality, even when they contain some bases passing the quality threshold. TraceTrack ignores ambiguous base calls ('N's) throughout.

## 2.5 MSA, consensus sequence and mutation calls

As mentioned earlier, TraceTrack employs Clustal Omega to perform MSA. For each reference within a given group, the MSA is created separately with all its corresponding trace files. The resultant aligned sequences are then used to generate a consensus sequence, according to the following principles: (i) TraceTrack calls point mutations, insertions and deletions with respect to the reference sequence only if all reads agree, (ii) in case only some traces contain an insertion, an ambiguous insertion character is displayed ('?') and the base is not considered as a viable sequence position and (iii) in all other cases when the reads do not agree, the reference base is kept and the number of ignored reads is shown as reduced read coverage.

Once consensus sequence is defined, TraceTrack translates the CDS and classifies mutations as silent (same amino acid), missense (different amino acid), nonsense (produces a stop codon) or frameshift (caused by insertions and deletions).

Since TraceTrack translates the coding sequence continuously from the beginning, an insertion or deletion of a number

of bases not divisible by 3 leads to a frameshift. All the shifted positions are displayed in a different colour and the translation continues in the new frame.

## 2.6 Alignment and results

After all trace files are assigned to their corresponding reference sequences, TraceTrack performs the alignments, generates consensus sequences, calls the mutations and displays the results in a sortable table containing one alignment per row (Fig. 3A). The rows are labelled with a reference ID and contain the following information: percentage of sequence coverage and percentage identity, numbers of different mutation types in the consensus sequence and the number and names of aligned trace files.

Each row can be expanded to show the consensus sequence with colour-coded coverage and respective mutations (Fig. 3B). Background colour reflects coverage and text colour represents mutation type. Additional information about each position is provided by the tooltips widget, including the position and the base for each aligned trace sequence at that position.

## 2.7 Chromatogram visualization

When the user clicks any position in the alignment, a trace viewer appears. The reference sequence is displayed at the top and all trace chromatograms and sequences are shown underneath (Fig. 3C). Amino acid translations are shown for the reference sequence and differences relative to the consensus sequence are highlighted. An interactive navigation bar can be used to navigate to specific mutations. The viewer can also be navigated using arrow buttons or by clicking any position in the reference or trace sequences.

## 2.8 Exporting results

A summary report can be downloaded as an interactive spreadsheet. The first sheet contains the overview table with


**Figure 3.** TraceTrack results page. (A) Table of trace and reference sequence alignments. (B) Alignment detail with colour-coded coverage and mutations. (C) Trace viewer showing the reference sequence and traces aligned to it

links to subsequent sheets, which contain individual alignments. Both the reference and consensus sequences are displayed, as well as each trace sequence, along with the corresponding amino acid translations. A second sheet is also provided for each alignment with a list of all mismatching positions and regions with zero coverage. The sequences can be easily navigated by clicking mutation positions. If desired, each alignment can also be downloaded separately using a download button in the corresponding table row.

## 3 Results

Taken together, TraceTrack enables batch processing of trace files and their alignment to multiple reference sequences as well as streamlines the inspection of chromatograms via a user-friendly web application. A comparative analysis to existing commercial or freely available tools highlights TraceTrack's advantages and limitations (Supplementary Table S2). The tool is flexible, it can run both on a web server as well as locally on a personal computer, and for quick access, TraceTrack is also available as a web application at https://tracetrack.dichlab.org. The main advantage of TraceTrack compared with other available open-source tools is the ability to process large numbers of both trace files and reference sequences simultaneously. In terms of computational performance, matching and aligning of a hundred of trace files to their respective reference sequences can take less than minute on a laptop computer or on the server (Supplementary Table S3). All the while the user is guided by a convenient graphical interface with no need for command line use or any knowledge of scripting.

Chromatograms annotated by the sequence and translation can be directly viewed within the application, providing an easy way to compare trace files to each other, or to the reference, and viewing changes at the amino acid level. A report can be downloaded for the entire analysis or for separate alignments.

Despite the advantages listed above, it is important to note TraceTrack's limitations. The use of an out-of-the-box MSA algorithm might be rigid and less flexible when it comes to handling indels and other edge cases in the alignment. The behaviour is different to using a read mapping algorithm, such as BWA (Burrows-Wheeler Aligner) , which might better reflect the nature of mapping sequenced reads to a reference sequence. TraceTrack could further be extended by adding functionality to adjust trimming of trace sequences manually and edit base calls. Currently, trace sequences can only be trimmed based on quality thresholds and the sequence is immutable.

## 4 Conclusion

As Sanger sequencing is still widely used for biological and biomedical research, there is a need to find automated solutions to common tasks. The available open-source tools for aligning and inspecting trace files were scarce and lacked functionality for batch processing. Therefore, we developed TraceTrack, an application to bridge this gap and facilitate many areas of research, including protein design, biocatalysis and therapeutic antibody discovery. Since the tool is fully open-source, there is potential for updates and extensions not only by the authors, but also the greater scientific community.

## Acknowledgements

We thank Jens Christensen, Vincent Antonucci and Carol A. Rohl for supporting this work. We are immensely grateful to Arthur Fridman, Charles Tilford and Nick Mukhitov for their advice and stimulating discussions. We also thank Anja Muzdalo for reviewing the manuscript and Martin Spale for preparing the tool for open source.

## Author contributions

Kveta Brazdilova (Formal analysis [lead], Methodology [equal], Software [lead], Writing—original draft [equal], Writing—review & editing [supporting]), David Prihoda (Methodology [supporting], Software [supporting], Supervision [supporting], Writing—review & editing [supporting]), Quynh Ton (Data curation [lead], Formal analysis [supporting], Validation [lead], Writing—review & editing [supporting]), Heath Klock (Conceptualization [lead], Methodology [equal], Validation [supporting], Writing—review & editing [supporting]) and Danny A. Bitton (Conceptualization [lead], Funding acquisition [lead], Methodology [lead], Project administration [lead], Resources [lead], Supervision [lead], Writing—original draft [lead], Writing—review & editing [lead])

## Funding

This work was supported by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA.

## Conflict of Interest

All authors are/were employees of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA and may hold stocks and/or stock options in Merck & Co., Inc., Rahway, NJ, USA.

## References

Chao,K.H. *et al.* (2021) sangeranalyseR: simple and interactive processing of sanger sequencing data in R. *Genome Biol. Evol.*, **13**, evab028.

Cock,P.J.A. *et al.* (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Engel,S.R. *et al.* (2014) The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 (Bethesda)*, **4**, 389–398.

Fu,J. *et al.* (2018) Evaluation genotypes of cancer cell lines HCC1954 and SiHa by short tandem repeat (STR) analysis and DNA sequencing. *Mol. Biol. Rep.*, **45**, 2689–2695.

Gwinn,M. *et al.* (2019) Next-generation sequencing of infectious pathogens. *JAMA*, **321**, 893–894.

Rausch,T. *et al.* (2020) Tracy: basecalling, alignment, assembly and deconvolution of sanger chromatogram trace files. *BMC Genomics*, **21**, 230.

Schmid,K. *et al.* (2022) SangeR: the high-throughput sanger sequencing analysis pipeline. *Bioinformatics Adv.*, **2**, vbac009.

Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.*, **7**, 539.

Sujeevan,R. and Hebert,P.A. (2007) BOLD: the barcode of life data system. *Mol. Ecol. Notes*, **7**, 355–364.

Stucky,B.J. (2012) SeqTrace: a graphical tool for rapidly processing DNA sequencing chromatograms. *J. Biomol. Tech.*, **23**, 90–93.