Yet Another Quick Assembly, Analysis and Trimming Tool (YAQAAT): A Server for the Automated Assembly and Analysis of Sanger Sequencing Data

Darius Wen-Shuo Koh,[1,*] *Kwok-Fong Chan,*[2,*] *Weiling Wu,*[2] *and Samuel Ken-En Gan*[1,2,†]

[1]*APD SKEG Pte Ltd, Singapore 439444, Singapore; and* [2]*Antibody and Product Development Lab, Agency for Science, Technology and Research (A*STAR), Singapore 138672, Singapore*

Even with the ubiquity of Sanger sequencing, automated assembly software are predominantly stand-alone software packages for desktop/laptop use with very few online equivalents, thus geospatially constraining sequence analysis and assembly. With increased data output worldwide, there is also a need for automated quality checks and trimming prior to large assemblies, along with automated detection of mutations. Through web servers with expanded automation and functionalities, even smartphones/phablets can be used to perform complex analysis previously limited to desktops, especially if they can upload files from cloud storage. To facilitate such online accessible sequence assembly and analysis, we created Yet Another Quick Assembly, Analysis and Trimming Tool web server for the automated assembly of multiple .ab1 and .FASTQ sequencing reads *de novo* with automated trimming and scanning of the assembled sequences for single nucleotide polymorphisms and insertions or deletions without installation of software, allowing it to be accessed from anywhere with Internet access and with minimal dependency on other software and web tools.

Key Words: batch assembly, AB1, FASTQ, SNP, INDEL

INTRODUCTION

The Sanger method, first developed by Frederick Sanger, [1] is widely used to generate accurate reads from homogenous samples economically. Although the capillary-based Sanger method is dated, it is still the gold standard given its superior accuracy and reliability over the second- and third-generation sequencing technologies (*i.e.*, next-generation sequencing, single-molecule real-time sequencing, *etc.*). Sanger sequencing is used in primer walking, mass sequencing of error-prone PCR mutant libraries for protein engineering, [2, 3] *in vitro* reverse-transcriptase fidelity assays, [4–6] and prediction of host deaminases RNA editing sites. [7] It is also routinely performed in general molecular methods in genetic engineering to generate .ab1/.FASTQ files. To facilitate Sanger sequence analysis, many tools have been developed over the years, even as smartphone apps. [8, 9]

Many of these software programs for Sanger sequencing require stand-alone installations on the desktop/laptop, typically restricting such analysis to office spaces. Some still require manual assembly, making the analysis of multiple sequencing samples tedious and inefficient. To increase automation and allow analysis from any location with Internet access, Yet Another Quick Assembly, Analysis and Trimming Tool (YAQAAT) server was built to perform automated batch trimming and contig assembly not only on ab1 but also on FASTQ files. In addition, it can automatically detect single nucleotide polymorphisms/insertions or deletions when provided a reference template sequence (**Fig. 1*A***) to support quality control as well as detection of mutations. To allow online direct verification, translated sequences in 3 reading frames together with chromatogram peaks of mutant bases are also shown directly in the web server (Fig. 1*B*) without the need for using additional web tools or software, allowing a seamless online sequence analysis experience.

MATERIALS AND METHODS

YAQAAT is written in Python 3.7 using Flask (https://palletsprojects.com/p/flask/) and Bootstrap 4.4.1 (https://getbootstrap.com/) web application frameworks with Biopython [10] and jQuery 3.4.1 (https://jquery.com/) libraries. The ab1 traces are plotted using the D3 graphing library (https://d3js.org/).

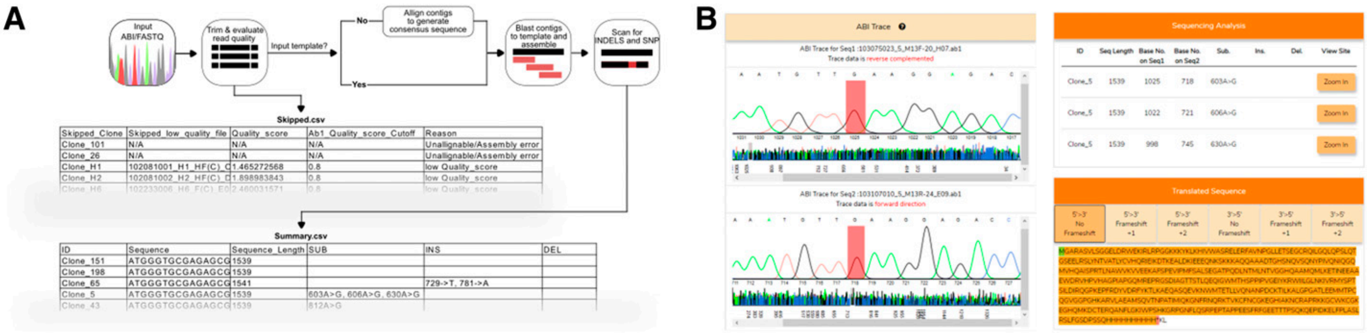
An overview of the contig assembly process is shown in Fig. 1*A*. First, read quality (described in the Supplemental

*These authors contributed equally to this work.

†Address correspondence to: Samuel Ken-En Gan, APD SKEG Pte Ltd, Singapore 439444, Singapore (Tel: +65 6407 0584; E-mail: samgan@apdskeg.com; samuel_gan@eddc.a-star.edu.sg).

This article includes supplemental data. Please visit http://jbt.abrf.org/ to obtain this information.

https://doi.org/10.7171/jbt.2021-3202-003



Skipped.csv

Skipped_Clone	Skipped_low_quality_file	Quality_score	Ab1_Quality_score_Cutoff	Reason
Clone_101	N/A	N/A	N/A	Unallignable/Assembly error
Clone_26	N/A	N/A	N/A	Unallignable/Assembly error
Clone_H1	102081001_H1_HF(C)_C	1.465272568	0.8	low Quality_score
Clone_H2	102081002_H2_HF(C)_D	1.898983843	0.8	low Quality_score
Clone_H6	102233006_H6_F(C)_E0	2.460031571	0.8	low Quality_score

Summary.csv

ID	Sequence	Sequence_Length	SUB	INS	DEL
Clone_151	ATGGGTGCGAGAGCG	1539			
Clone_198	ATGGGTGCGAGAGCG	1539			
Clone_65	ATGGGTGCGAGAGCG	1541		729->T, 781->A	
Clone_5	ATGGGTGCGAGAGCG	1539	603A>G, 606A>G, 630A>G		
Clone_43	ATGGGTGCGAGAGCG	1539	812A>G		

ABI Trace for Seq1 :103075023_5_M13F-20_H07.ab1
Trace data is reverse complemented

ABI Trace for Seq2 :103107010_5_M13R-24_E09.ab1
Trace data is forward direction

ID	Seq Length	Base No. on Seq1	Base No. on Seq2	Sub.	Ins.	Del.	View Site
Clone_5	1539	1025	718	603A>G			Zoom In
Clone_5	1539	1022	721	606A>G			Zoom In
Clone_5	1539	998	745	630A>G			Zoom In

MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERFAVNPGLLETSEGCRQILGQLQPSLQTGSEELRSLYNTVATLYCVHQRIEIKDTKEALDKIEEEQNKSKKKAQQAAADTGHSNQVSQNYPIVQNIQGQMVHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPVHAGPIAPGQMREPRGSDIAGTTSTLQEQIGWMTHSPPIPVGEIYKRWIILGLNKIVRMYSPTSILDIRQGPKEPFRDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGVGGPGHKARVLAEAMSQVTNPATIMIQKGNFRNQRKTVKCFNCGKEGHIAKNCRAPRKKGCWKCGKEGHQMKDCTERQANFLGKIWPSHKGRPGNFLQSRPEPTAPPEESFRFGEETTTPSQKQEPIDKELFPLASLRSLFGSDPSSQHHHHHHHHHH*KL

FIGURE 1

A) Schematic of the process from contig assembly to output files: Summary.csv and Skipped.csv files. *B*) The YAQAAT web server allows the user to download the results and conveniently analyze single nucleotide polymorphism (SNP) mutations by zooming in to mutant chromatogram peaks online. Results shown are sequencing data from HIV1 Gag genes generated *via in vitro* fidelity assays. [6] csv, comma separated values; INDEL, insertion or deletion; INS, insertion; Del, Deletion; SUB, Substitution; Seq, Sequence.

Data) and trimming are performed. An arbitrary noise threshold ($noise_{threshold}$) of 0.8 is used for assembly. Prior to contig assembly, Basic Local Alignment Search Tool Plus (BLAST+) 2.90 [11] is used to locally map the input reads to the template sequence. In the absence of a template, a consensus of all the input reads is generated to serve as a template reference to guide assembly. A forked implementation of lamassemble [12, 13] (https://gitlab.com/mcfrith/lamassemble) was found suitable for assembling multiple long sequencing reads, [14] meeting the requirements of such sequence analysis. lamassemble uses Local Alignment Search Tool version 1047 (LAST 1047) [15] and Multiple Alignment using Fast Fourier Transform (MAFFT) version 7.45 [16] for pairwise and multiple sequence alignment, respectively, thus allowing efficient assembly and comparisons. In homologous regions where bases are different between aligned contigs, the base with a higher Phred quality value is taken to generate the final sequence.

RESULTS AND DISCUSSION

Born out of our own research needs, YAQAAT was created for the efficient assembly of both ab1 and FASTQ files *via* an online platform without the need to install stand-alone software. It was successfully applied for the automated analysis of mutant viral libraries generated though error-prone Reverse-Transcriptase Polymerase Chain Reaction (RT-PCR), Topoisomerase based (TOPO) cloning, and Sanger sequencing [6] to perform batch assembly of multiple gene sequences online, a feature not available by other online Sanger assembly servers, *e.g.*, the GEAR-GENOMICS server using Tracy software. [17] In comparison with available options, YAQAAT also allows fast verification, with an estimated time of ~5 s for processing 1 pair and ~133 s for 18 pairs.

For assembly, multiple ab1/FASTQ files may be uploaded, and a template sequence may be prepared by the user. In the absence of a template, the sequences would be assembled *de novo* without scanning for mutations. Files to be assembled may be grouped automatically according to an identifier after a user-specified separator (*e.g.*, a string of characters such as an underscore or a dash) and the position of the identifier. For more familiar users, the pythonic implementation of the Regular Expression (RegEx) expression (see Supplemental Data) may be used. Alternatively, a downloadable comma separated values spreadsheet may be used to manually specify the files to be assembled.

Prior to submission, the user may adjust the default parameters (**Table 1**) to fine-tune the process of assembly, trimming, and quality control. Furthermore, the user can further trim the sequences should the reads be deemed low quality.

The process of contig assembly is shown in Fig. 1*A*. For automated trimming at the beginning and end of the reads, Phred values (log-transformed probabilities of a misassigned base [18]) encoded in both the ab1 and FASTQ sequencing file formats are used because the poor-quality ends typically have fluctuating low Phred values (**Fig. 2**). [19] Base-wise Phred values were limited by a user-defined cutoff ($cutoff_{Trim}$), and the instantaneous change in base Phred values in a given $sliding - window_{phred}$ (by applying a least squares fit linear regression) is calculated from both ends of the raw sequencing reads to determine windows of high-quality sequences. These high-quality sequences would be expected to have Phred values closer or above the $cutoff_{Trim}$ and $cutoff_{QC}$. The region between 2 high-quality windows of sequences is used for subsequent assembly and quality assessments. When reads cannot be consecutively aligned to each other even after trimming, the sliding window would

TABLE 1

List of parameters for trimming, evaluation of read quality, and alignment of reads to the template

Parameter	Description	Adjustment
Trim	Trim sequences based on Phred quality.	Default = "*N*" to avoid unnecessary truncation of assembled sequences. "Y" option should be used when sequencing files are judged to be noisy reads at the ends.
$sliding - window_{phred}$	Length of the sliding window used to calculate instantaneous gradient change of Phred values in order to determine the window of regions at the ends of the reads that are of poor quality to be trimmed out from evaluation of the overall sequence quality of assemble reads.	Default = 30; higher values will result in overtrimming. Values range between 0 and half the length of the sequence submitted.
$cutoff_{Trim}$, $cutoff_{QC}$	Cutoff used to evaluate the quality and to trim ab1/FASTQ reads. Base Phred values of 20 and 40 correspond to base call accuracy of 99% and 99.99%, respectively.	Default = 30; higher values result in selective assembly of only sequences with generally high Phred quality scores or stringent trimming. If a value of 50 is chosen, for example, for QC, most sequences chosen for assembly generally will be expected to have Phred scores of around or above 50.
$noise_{threshold}$	The threshold is the amount of noise tolerated for reads to be assembled. Noise scores above the threshold are considered of poor quality and would not be processed.	Default = 0.8; this should be decreased to allow the software to assemble reads that may have been skipped because of their poor quality.
$cutoff_{nBase}$	Bases below this cutoff are assigned the base "*n*."	Default = 1; increase to assign bases below the cutoff as "*n*."
$E - Value$	E-Value used by Blast+ to perform alignment and mapping of reads to template.	Default = 0.5; this may be increased if contigs cannot be aligned or if they share too few homologous regions.
$Trials_{Max}$	Number of times the sliding window (Trimming_Phred_window) and quality (Quality_evaluation_Phred_window) are extended by a multiple of 1.	Default = 3; reduce to increase processing speed and increase to possibly increase the degree of trimming low-quality bases.

be extended (determined by the $Trials_{Max}$ parameter) by the current $sliding - window_{phred}$ length to reduce the effects of small changes in windows of poor sequencing quality, which could overestimate or underestimate the true quality of the windows of sequences, leading to either over- or under-trimming. After the job completion, the results can be accessed through the web portal utilizing the provided job identifier or *via* an e-mail notification to inform the users. Three reading frames in both the forward and reverse directions are displayed in the results page of the chromatogram. If a template is provided, a list of mutations with reference to the template sequence are also shown with the capability to zoom into mutant peaks on the chromatogram. YAQAAT provides the results of the sequence assembly, mutation scan, and skipped reads as downloadable comma separated values text files (Fig. 1*A* and **Table 2**).

For large sequencing projects that require rapid batch analysis, various offline software suites that perform chromatogram display, trimming, and contig assembly separately, such as 4peaks (Nucleobytes, The Netherlands), SnapGene (GSL Biotech LLC, Chicago, IL, USA), AutoSeqMan,[20] DNASTAR (https://www.dnastar.com/), and Sequencher (http://www.genecodes.com/), *etc.*, are available. YAQAAT has an additional advantage in also allowing an automated online pipeline for batch processing of sequencing samples through a user-friendly interface, without the need to install external bioinformatics programs. Compared with existing online tools,[17] few can

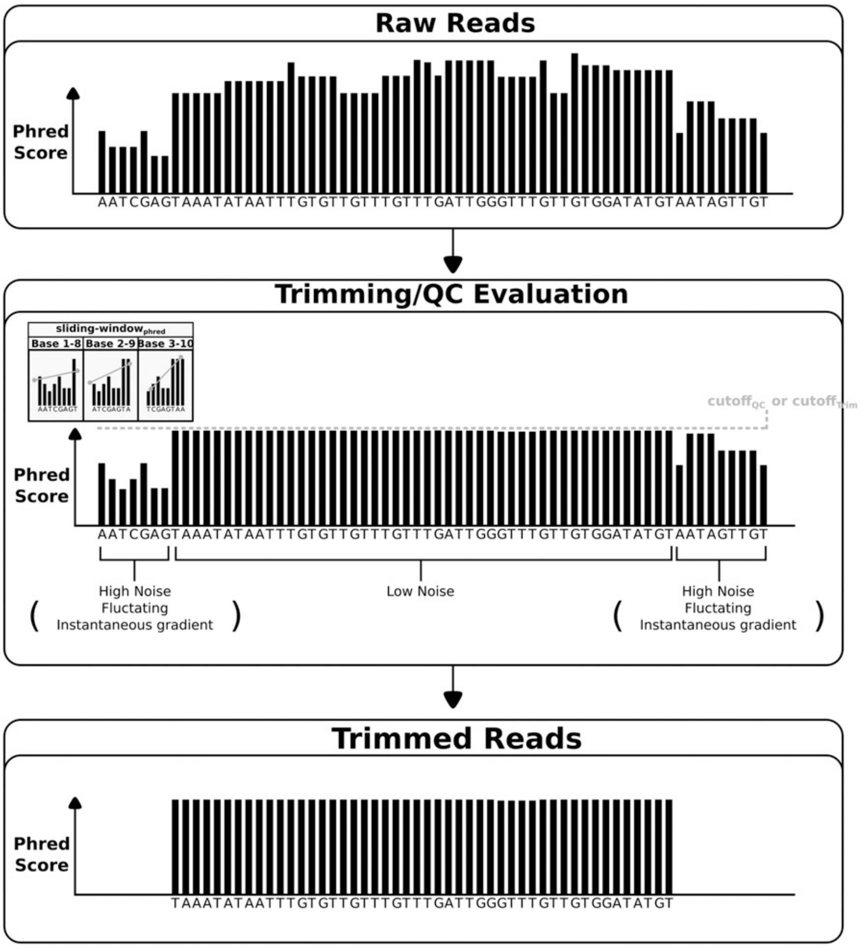


FIGURE 2

Sequence trimming and Quality Control (QC) evaluation. The *cutoff$_{Trim}$* and the *cutoff*$_{QC}$ parameters are based on the Phred scores. A sliding window is applied to detect good-quality reads, with the noise of reads estimated *via* a linear regression procedure applied to a sliding window to determine the instantaneous gradient change of Phred values throughout the sequence. Sequences are trimmed up to where there was no instantaneous gradient change in Phred values (region marked as low noise) and evaluated by summing up the instantaneous gradient change.

selectively filter out unreliable .FASTQ/.AB1 files, perform batch assembly, highlight display mutant peaks, and detect mutations simultaneously. When processing a large number of sequence files, automated differentiation of poor- and high-quality reads (*i.e.*, overlapping peaks and low Phred scores throughout the whole sequence) is important and is thus a feature included in YAQAAT.

Given that the job submission procedure follows an online queue system, users can submit batches of sequences for online analysis (without the stand-alone software). In fact, users can even leverage smartphone browser access and upload files from cloud servers in preparation for later analysis even on the go (see example of using YAQAAT and uploading of files from Dropbox: https://youtu.be/BiqyfGQZYog).

In conclusion, we describe the YAQAAT web server, which can allow online sequence analysis with automated features allowing seamless on the go, making such procedures more convenient and accessible.

TABLE 2

List of downloadable outputs

File	Contents
Summary.csv	Final sequence, sequence length, SNP/INDELS
Skipped.csv	Lists skipped sequences and the *read$_{quality}$*
Results.INF	Raw data, alignment, and assembly details

csv, comma separated values; INDEL, insertion or deletion; INF, Indexed Nucleotide Format; SNP, single nucleotide polymorphism.

ACKNOWLEDGMENTS

The authors thank Drs. Su Tran-To Chinh and Li-Xiang for useful input. The authors also thank Ghin-Ray Goh and Joshua Yeo Yi for help in testing the server. This work is self-funded and supported by APD SKEG Pte Ltd. The authors declare no conflicts of interest.

REFERENCES

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*. 1977;74: 5463–5467.
2. Herwig L, Rice AJ, Bedbrook CN, et al. Directed evolution of a bright near-infrared fluorescent rhodopsin using a synthetic chromophore. *Cell Chem Biol.* 2017;24:415–425.

3. Taylor ND, Garruss AS, Moretti R, et al. Engineering an allosteric transcription factor to respond to new ligands. *Nat Methods*. 2016;13:177–183.
4. Abram ME, Ferris AL, Das K, et al. Mutations in HIV-1 reverse transcriptase affect the errors made in a single cycle of viral replication. *J Virol*. 2014;88:7589–7601.
5. Sebastián-Martín A, Barrioluengo V, Menéndez-Arias L. Transcriptional inaccuracy threshold attenuates differences in RNA-dependent DNA synthesis fidelity between retroviral reverse transcriptases. *Sci Rep*. 2018;8:627.
6. Yeo J. Y., Koh D. W.- S., Yap P., Goh G.-R., & Gan S. K.-E. Spontaneous Mutations in HIV-1 Gag, Protease, RT p66 in the First Replication Cycle and How They Appear: Insights from an In Vitro Assay on Mutation Rates and Types. International Journal of Molecular Sciences, 2021; 22: 370.
7. Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. *Nat Commun*. 2011;2:319.
8. Nguyen P-V, Verma CS, Gan SK-E. DNAApp: a mobile application for sequencing data analysis. *Bioinformatics*. 2014; 30:3270–3271.
9. Sim JZ, Nguyen PV, Zang Y, Gan SKE. DNA2App: mobile sequence analyser. *Scientific Phone Apps and Mobile Devices*. 2016;2:2.
10. Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25:1422–1423.
11. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10: 421.
12. Mitsuhashi S, Ohori S, Katoh K, Frith MC, Matsumoto N. A method for complete characterization of complex germline rearrangements from long DNA reads. *medRxiv*. 2019: 19006379.
13. Lei M, Liang D, Yang Y, et al. Long-read DNA sequencing fully characterized chromothripsis in a patient with Langer-Giedion syndrome and Cornelia de Lange syndrome-4. *J Hum Genet*. 2020;65:667–674.
14. Mitsuhashi S, Ohori S, Katoh K, Frith MC, Matsumoto N. A pipeline for complete characterization of complex germline rearrangements from long DNA reads. *Genome Med*. 2020;12: 67.
15. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. *Genome Res*. 2011; 21:487–493.
16. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30:772–780.
17. Rausch T, Fritz MH-Y, Untergasser A, Benes V. Tracy: basecalling, alignment, assembly and deconvolution of sanger chromatogram trace files. *BMC Genomics*. 2020;21:230.
18. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. 1998;8: 186–194.
19. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 1998;8:175–185.
20. Jin JQ, Sun YB. AutoSeqMan: batch assembly of contigs for Sanger sequences. *Zool Res*. 2018;39:123–126.