SequenceBouncer: A method to remove outlier entries from a multiple sequence alignment

Cory D. Dunn

Institute of Biotechnology, Helsinki Institute of Life Science, University of Helsinki, Helsinki, 00014, Finland

Email: cory.dunn@helsinki.fi
Phone: +358 50 311 9307

Abstract:

Phylogenetic analyses can take advantage of multiple sequence alignments as input. These alignments typically consist of homologous nucleic acid or protein sequences, and the inclusion of outlier or aberrant sequences can compromise downstream analyses. Here, I describe a program, SequenceBouncer, that uses the Shannon entropy values of alignment columns to identify outlier alignment sequences in a manner responsive to overall alignment context. I demonstrate the utility of this software using alignments of available mammalian mitochondrial genomes, bird cytochrome *c* oxidase-derived DNA barcodes, and COVID-19 sequences.

Background:

Outlier sequences within a dataset, which can include mis-annotated entries, mis-assembled sequences, and sequences containing rare structural rearrangements, can compromise our ability to perform robust phylogenetic analyses (1–3). Avoidance of aberrant sequence inclusion within alignments becomes more challenging with the increasing rate of sequence acquisition.

To address this problem, I sought a method that could remove outlier sequences from a multiple sequence alignment in a manner responsive to overall sequence context and at moderate computational cost. I provide a program, SequenceBouncer (https://github.com/corydunnlab/SequenceBouncer), which functions within a Python framework and takes advantage of Shannon entropy calculations to remove potentially undesirable alignment entries.

Workflow and examples:

Pairwise sequence comparisons take advantage of alignment column Shannon entropy values to calculate divergence

I first developed an approach in which all of the sequences within an input multiple sequence alignment are compared to one another to facilitate outlier identification, hereafter designated a 'full analysis'. As the initial step in this analysis, an input FASTA file is examined to determine the percentage of each alignment column that consists of gaps (Figure 1). Correct choice of a gap threshold (-g flag) by the user focuses downstream analysis upon generally well-aligned positions. A gap threshold that is excessively low leads to removal of information that might be helpful in identifying outlier sequences. A gap threshold that is too high will include alignment positions that may be non-informative at best, or will confound the analysis at worst. A gap threshold that is too high will also needlessly increase the computation time. The gap threshold default for SequenceBouncer is set at 2%.

Next, the Shannon entropy values associated with each alignment column (SE_i values), expressed in bits, are evaluated. Shannon entropy is linked to the amount of 'surprise' that one might encounter within a message (4, 5), here defined as an alignment column (6). Therefore, alignment column Shannon entropy values can be an appropriate basis for detecting unexpected divergence of a sequence from the bulk of alignment entries. The maximum SE_i value obtained when considering positions under the selected gap threshold (SE_{max}) is saved for use in later calculations.

Following the calculation of SE_i for each alignment column, as well as the SE_{max} value, pairwise comparisons of alignment sequences are initiated. If the SE_i value of an alignment column at which characters match is comparatively high, then the characters may be

considered to be surprisingly identical, providing evidence for a greater level of similarity between the compared sequences. In contrast, if the SE_i value of a column at which characters match is low or zero, a character match is less noteworthy, given the high conservation at that position and the assumption that the user has attempted to align homologous sequences. The SE_i values obtained from matched positions are added together.

Positions divergent between the two compared sequences are then considered. Here, a lack of a match at each position is penalized to a degree determined by the variability observed in that alignment column. Toward this goal, the SE_{max} value was subtracted from the SE_i value at each non-matching position to generate an 'inverse' (SE_{inv}) score. SE_{inv} scores from non-matching positions are added to the sum of SE_i values obtained at matching positions to generate a total comparison score (CS). A higher CS score emerging from a pairwise comparison would be associated with a more robust alignment and greater sequence relatedness.

Pairwise analyses of sequences continue, and CSs are generated, until all sequences within the sample have been compared to all other sequences. The resulting CSs are then converted to normalized comparison scores (nCSs) linking higher value to greater divergence. To generate nCSs, the maximum CS obtained across all pairwise sequence comparisons is subtracted from all tabulated CSs, and then the resulting set of values is multiplied by -1.

Abnormally high nCS scores associated with a given entry across a majority of pairwise sequence comparisons signals outlier status. A cut-off (CO) value is calculated from the nCSs dataset for use in discriminating alignment outliers. For CO value determination, the median nCS obtained for each sequence across all of its pairwise comparisons is determined, and the interquartile range is calculated across these median nCSs. The interquartile range is then multiplied by the coefficient k (-k flag), which can either be supplied by the user or left at the default value of 1. The result is then added to the nCS found at the 75th percentile to produce the CO score. Those sequences for which median nCSs obtained across its comparisons with other sequences falls above the CO value 'fail' testing by SequenceBouncer and will be marked for potential removal from the alignment.

When the full analysis is completed, results are reported within a comma-separated values (CSV) file that lists each input sequence accession and whether the median nCS for that sequence when calculated across pairwise comparisons is above the outlier threshold. In addition, a 'cleaned' FASTA file lacking outlier sequences, as well as a FASTA file containing rejected outlier sequences, are provided as output. Finally, when all pairwise comparisons are performed, a full table of nCSs emerging from pairwise comparisons is supplied to the user.

Example 1: Mammalian mitochondrial DNA sequences

Sequences found within the RefSeq database (7) are subject to validation and curation. However, it is clear that some organellar sequences available within this database contain errors or harbor rearrangements, insertions, and duplications that would require further treatment before a serviceable alignment can be generated (8). Accordingly, I expected that the set of mammalian mitochondrial DNA sequences retrieved from RefSeq might provide a useful test dataset when developing my approach to outlier sequence detection.

After downloading mammalian mtDNA sequences obtained from RefSeq and removing duplicate sequences, I aligned the resulting dataset using MAFFT (9). Next, I performed a full analysis using SequenceBouncer to detect and remove outliers. Positions gapped at 2% of sequences or less were analyzed during calculation of nCSs, and all possible pairwise sequence comparisons were completed (Figure 2A). Analyses were performed using a MacBook Pro with a 2.8 GHz Quad-Core Intel Core i7 processor and 16 GB of 1600 MHz DDR memory, and the total time required to complete a run of SequenceBouncer using the mammalian mtDNA input alignment was less than 6 minutes. When plotting the median nCS associated with each sequence (Figure 2B) a subset of entries was clearly associated with excessive median nCSs (nCS > 8500), and inspection of these sequences indeed indicated poor alignment with the vast majority of other sequences. A few of these sequences appear to be inverted with respect to the majority of other downloaded RefSeq mammalian mtDNA entries. Input files and SequenceBouncer output related to the examples provided here can be found at (doi: 10.5281/zenodo.4285789).

Moreover, the 37 marsupial mtDNAs and three monotreme mtDNAs included in this alignment of 1303 mammalian mtDNAs were also associated with median nCSs that were elevated relative to the bulk of entries (7800 < nCS < 8500), as might be expected when considering a sequence set heavily biased toward placental mammals.

Next, I explored the outcome of changing k, the coefficient used to calculate CO values. At a k value of 0.75, the most poorly aligned sequences were removed, as well as all monotreme and marsupial sequences (Figure 2C). In contrast, at k values of 1.75 and above, all monotreme and marsupial sequences were retained, and only clearly aberrant sequences were removed. These tests demonstrate that SequenceBouncer output can be easily tailored, by use of the k variable, to the user's desired definition of 'outlier'.

I used simple metrics to investigate whether the output alignments might have been improved by SequenceBouncer. When aberrant sequences are included in an alignment, the proportion of highly gapped alignment columns is often raised due to the accommodation of these outlier entries (10). Consequently, removal of outlier sequences should lead, upon a repeat of the alignment process, to a more 'compact' alignment with a reduced number of

highly gapped alignment columns and a lower number of total columns. When comparing my initial input alignment to the re-aligned output of SequenceBouncer run using a k setting of 1.75, many abundantly gapped columns were removed (Figure 2D), and the overall length of the aligned sequences was reduced (Figure 2E), suggesting the generation of a dataset characterized by greater overall sequence similarity.

Analysis of large alignments by sequence sampling

When alignment depth and length are moderate, as in the case of the mammalian mtDNA sequence set, all sequences can be compared to all other sequences within minutes by use of a laptop or desktop computer. However, more expansive alignments, which may encompass tens of thousands of sequences containing tens of thousands of characters, begin to tax the computational resources and patience available to researchers who would like to quickly and effectively remove non-homologous entries. Consequently, I incorporated into SequenceBouncer the possibility of 'sampling-based' approaches that permit outlier detection, yet do not require an extraordinary amount of computational effort.

Here, within each 'trial', samples are selected from the alignment (Figure 3A), and the size of each sample selected from the source alignment is chosen by the user (-n flag). Whether the median nCS calculated for a sequence during pairwise comparisons within the sample is higher than the CO value is calculated as if the chosen sample represented the entire input alignment (Figure 1). Sequences are included in samples, without replacement to the starting dataset, until no more sequences remain to be selected and analyzed from the larger, source alignment.

A single trial of each sequence is unlikely to be sufficient under most use cases for confident identification of outliers. Therefore, the user may request multiple trials (-t flag). The outcomes of these trials, along with the stringency (-s flag) setting, are used to determine whether a sequence should be removed from the alignment. When SequenceBouncer is set to minimal stringency (-s 1), then a sequence must fail all trials to be removed from the alignment. When the software is set to moderate stringency (-s 2), a sequence must fall above the CO value in more than 50% of the trials in order to be selected for removal from the alignment. At maximum stringency (-s 3), a sequence is excluded from the cleaned alignment if it is flagged as an outlier in any of the trials. When a sampling-based approach is applied, the output CSV file lists accessions, the number and fraction of trials during which a sequence was identified as an outlier, and whether, based upon the stringency setting, a sequence was retained in a cleaned FASTA output alignment. I note that the final outputs from sampling-based analyses are not deterministic, due to the random shuffling of sequences in the source alignment that occurs before samples are extracted.

Below, I demonstrate how the SequenceBouncer program can take advantage of a sampling-based approach to remove outliers from larger alignments.

Example 2: Bird cytochrome c *oxidase-derived barcodes*

A portion of the well-conserved gene encoding cytochrome *c* oxidase subunit I is often deployed as a sequence barcode during ecological studies (11–13). These barcodes, as well as details of sample acquisition, are readily available from the Barcode of Life Data (BOLD) Systems database (14). However, not all of the barcode sequences found within this database are of high quality and properly annotated. I tested whether SequenceBouncer might successfully remove outlier barcodes, and I specifically focused upon bird COI-5P entries, which I found to be of utility during a recent study seeking hummingbird-specific changes to mtDNA-encoded proteins (15).

Toward this goal, I downloaded barcode data for class Aves from the BOLD database, deleted duplicate and non-COI-5P barcodes, and aligned the remaining sequences using MAFFT (9). Next, I performed a full analysis that carried out all pairwise sequence comparisons (Figure 4A). The gap threshold was set at 5%, and the k coefficient was set to 1. This full analysis required nearly 5 hours on the above-mentioned test system. Upon re-alignment of the cleaned dataset, the fraction of highly gapped columns (95-100% gapped) was reduced, and the fraction of seldom gapped columns (0-5% gapped) was increased (Figure 4B), indicating successful removal of aberrant sequences. Moreover, the length of the sequences within the cleaned and realigned dataset was reduced (Figure 4C), also suggesting removal of outliers potentially compromising the initial alignment. Finally, the distribution of the ungapped sequence lengths of entries within the cleaned alignment clearly demonstrated removal of short sequences of a length inconsistent with the length of a useful COI-5P barcode (Figure 4D).

Because an analysis taking advantage of all pairwise sequence comparisons was computationally costly, I explored whether my sampling-based approach could also successfully identify bird barcode outliers, but at a fraction of computational effort. First, I tested which sample size, given the input alignment and 10 trials at moderate stringency, would permit a close approximation of results obtained when performing a full analysis of all pairwise sequence comparisons. I found that the percent of outlier calls that matched when comparing a sampling-based approach with the full analysis appeared to plateau at ~ 99.7% at a sample size of ~ 100 sequences (Figure 4E). Of course, results achieved using any single input alignment may not necessarily generalize to every possible dataset, and so each SequenceBouncer user may assess the stability of results obtained on their own input dataset at different parameter settings.

Next, I examined the link between the number of trials and the correspondence between sampling-based and full analyses. Gap threshold and k values were maintained at 5% and 1, respectively. The sample size was set to 100, stringency was set to ‘moderate’, and trial numbers between 1 and 20 were tested. Indeed, as expected, a greater number of

sampling-based trials was associated with improved correlation with the full pairwise analysis. A plateau of ~ 99.7% correspondence between these two analyses was obtained at a trial number of ~ 10 (Figure 4F). Again, the most appropriate parameters would depend upon the specific input alignment, yet an increase in trial number can clearly lead to improved results. As expected, those entries for which the sampling-based and full analysis output did not match were sequences for which the median nCS across all pairwise comparisons was close to the full analysis CO value (Figure 4G). Excitingly, the time to complete ten trials of the input bird COI-5P alignment at a sample size of 100 and a gap threshold of 5% was only ~ 12 minutes on the test system, a greater than 20-fold decrease in required computational time.

Example 3: COVID-19 sequences

The analyses described above demonstrate that mammalian full-length mtDNA sequence alignments of moderate depth, as well as high-depth alignments consisting of short barcodes, can be effectively and rapidly processed by SequenceBouncer. Next, I turned my attention to a rapidly growing set of COVID-19 sequences available for download from the RefSeq database. The COVID-19 RNA is approximately 30,000 bases in length, and my input alignment consisted of nearly 37,500 entries, most of which are nearly identical or identical to other sequences within the dataset. I again used a sampling-based approach to determine which COVID-19 sequences might be marked as outliers and removed from downstream analysis.

While analyzing COVID-19 sequences, I explored the outcomes of different SequenceBouncer stringency settings. Gap threshold was set to 2%, k value was set to 1, sample size was set to 100, and trial number was set to 10. As discussed above, the stringency setting directs exclusion of entries from output alignments based upon the proportion of trials for which an entry is determined to be an outlier. At minimal stringency, a substantial number of truncated sequences were removed from the alignment (Figure 5A-B). The range of ungapped sequence lengths was, as expected, further narrowed by performing the analysis of COVID-19 sequences as moderate and maximum stringency. The COVID-19 alignment mostly consists of very similar sequences, and so comparatively few columns are almost totally gapped, in contrast to the mammalian mtDNA and bird barcode alignments. Nonetheless, realignment after SequenceBouncer processing clearly led to a reduction in gapped alignment positions that accommodate outlier sequences (Figure 5C), particularly at moderate and maximal stringency settings. Of note, a set of corrupted COVID-19 sequences, LR897977.1 - LR898047.1 (https://mafft.cbrc.jp/alignment/software/) were identified as outliers when SequenceBouncer was run at moderate and maximum stringency, further supporting the utility of my approach. Improvements in alignment quality as stringency was progressively increased are also clear when alignments are graphically represented (Figure 6). On my test system, analyses took only ~ 140 minutes. Clearly, SequenceBouncer output can remove aberrant entries from very sizable datasets of highly homologous, virus-derived sequences.

Conclusion:

I report the development of a new program, SequenceBouncer, which allows detection and removal of outlier entries from a multiple sequence alignment. My approach takes advantage of Shannon entropy computations to allow a context-dependent calculation of which sequences might not be homologous with other entries. Some outliers are mis-annotated, mis-assembled, contain many sequencing errors, or are not oriented appropriately when considering the bulk of other analyzed sequences. Sequences rejected by SequenceBouncer might be further analyzed to reveal important structural rearrangements not found within the bulk of other alignment entries (16).

This software, written in Python 3, is uncomplicated to install and run. Moreover, key parameters can be tailored to the specific needs of the user. Depending upon the input alignment size, analyses either can be performed by completion of all pairwise sequence comparisons or, if computational resources are limiting, can be executed by a rapid and reliable sampling-based approach. Even a large COVID-19 input alignment of 37,470 sequences, each consisting of 31,012 nucleotide positions, could be effectively processed within a few hours using a laptop computer.

I note that alternative approaches are available that allow outlier detection in multiple sequence alignments or COI barcodes (17–23), although a direct comparison of these methods against SequenceBouncer has not yet been completed. Future efforts on the development of SequenceBouncer will explore the possibility of further acceleration by use of parallelization and just-in-time compilation.

Methodology:

Sequence acquisition

Mammalian mtDNA RefSeq accessions (RefSeq release 202) were downloaded on September 29, 2020 from the Organelle Genome Resources page of NCBI (https://www.ncbi.nlm.nih.gov/genome/organelle/), and the 1316 associated full length mtDNA sequences were obtained by use of the NCBI Entrez utilities (https://www.ncbi.nlm.nih.gov/books/NBK25501/). Duplicate sequences were removed by use of seqkit v0.13.2 (24) to generate a file containing 1303 mammalian mtDNA entries.

Bird COI-5P barcodes were downloaded from the BOLD database (14) on October 21, 2020 using the query ‘Aves’. 46,397 entries were filtered for those containing COI-5P in their description line, then duplicate sequences were deleted by seqkit v0.13.2 (24) to generate a file for initial alignment and subsequent SequenceBouncer testing that contained 31,087 barcodes.

Full length COVID-19 sequences were downloaded from the severe acute respiratory syndrome coronavirus 2 data hub (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049) on November 9, 2020. Here, duplicate sequences were retained during downstream analysis.

Sequence alignment

For full-length mammalian mtDNA sequences and bird COI-5P barcodes, initial alignment and alignment after SequenceBouncer processing were carried out by MAFFT v7.471 (9) using the FFT-NS-2 approach. For bird COI-5P barcodes, the 'anysymbol' flag was activated.

For full-length COVID-19 sequences, initial alignment and realignment following SequenceBouncer processing was performed using an online MAFFT server tailored to the analysis of datasets consisting of viral sequences with a high level of identity (https://mafft.cbrc.jp/alignment/software/closelyrelatedviralgenomes.html) (25). Sequences with greater than 5% ambiguous letters were removed, a '6merpair' strategy was deployed, 'anysymbol' and 'memsave' modes were activated, and the scoring matrix selected was '1PAM/kappa=2'. The COVID-19 reference sequence, NC_045512.2, was used as the alignment to which all other sequences were added.

Shannon entropy calculations

Code for calculating alignment column Shannon entropy was modified from the 'Shannon' script (https://gist.github.com/jrjhealey/) under its GPLv3 license.

Gap quantification and removal

De-gapping of sequences and sequence length calculations was performed by use of seqkit v0.14.0 (24). Calculation of alignment column gap frequencies for analysis of SequenceBouncer results was carried out using trimal v1.4.rev15 (26).

Data visualization

Graphs were produced by use of Prism 9.0.0 (https://www.graphpad.com), and COVID-19 alignment visualization was achieved using AliView 1.24 (27).

Acknowledgements:

This work was funded by a Senior Researcher grant from the Sigrid Jusélius Foundation to C.D.D. I thank Gülayşe İnce Dunn and Svetlana Konovalova for helpful comments on the manuscript. I also thank Dave Whipp and Vuokko Heikinheimo for allowing me to attend their Geo-Python course.

Figure Legends:

Figure 1: Alignment column Shannon entropy is deployed toward outlier sequence identification. A gap threshold for column analysis is selected, and the Shannon entropy for each alignment column (SE_i) is calculated. All possible pairwise comparisons are performed, and where the sequences match, the SE_i values are added to the comparison score (CS). For mismatched positions, the maximum Shannon entropy (SE_{max}) calculated across all positions below the gap threshold is subtracted from the column SE_i value to generate 'inverse' Shannon entropy (SE_{inv}) values, and these results are also added to the CS. Each CS is normalized (nCS), and the median nCS values across the comparisons to each sequence are used to determine a cut-off (CO) value above which a sequence is determined to be an outlier.

Figure 2: Identification of outlier sequences found within an alignment of mammalian mtDNA sequences. SequenceBouncer was run on an input alignment of mammalian mtDNA sequences (1303 entries) using a gap threshold of 2%, and all pairwise sequence comparisons were performed. (**A**) SequenceBouncer output showing the nCS associated with each pairwise sequence comparison. Blue reflects higher identity, and yellow signals higher divergence. (**B**) A subset of input mtDNA sequences are clearly outliers from the alignment. Mammalian mtDNA sequences are ordered by increasing median nCSs, and infraclasses are reflected by dot color. (**C**) k settings shift the CO value for outlier detection. The CO value obtained using the mammalian mtDNA dataset is indicated at k values ranging from 0.75 to 2 in steps of 0.25. (**D**) Removal of outlier sequences leads to a decrease in heavily gapped columns and an increase in minimally gapped columns upon realignment. The cleaned output alignment obtained at a k value of 1.75 (1283 sequences) was realigned, and the gap frequency was calculated for input alignment and cleaned, re-aligned output. (**E**) Removal of outlier sequences allows for a reduction in alignment sequence length. SequenceBouncer analysis was performed as in (**D**), and the length of cleaned, realigned entries was assessed.

Figure 3: A sampling-based analysis can permit high prediction precision at reduced computational cost. (**A**) A subset of sequences, with the sample size determined by the user, is selected from the full alignment, and whether the nCS falls above the CO score is determined as in Figure 1. (**B**) Multiple trials, with the number determined by the user, can be performed to increase the certainty of outlier selection. Stringency, related to the number of trials during which a sequence fell above the CO value, can be set based upon the goals of the user.

Figure 4: Identification of outlier sequences within an alignment of bird COI-5P barcodes. (**A**) An alignment of bird COI-5P barcodes (31,087 sequences) was analyzed with SequenceBouncer at a gap threshold of 5% and k setting of 1. A full analysis consisting of all pairwise sequence comparisons was performed. The graphical output file is provided, with

blue linked to higher identity between two sequences and yellow signaling higher divergence. The cleaned output alignment (28,141 sequences) and the input alignment were realigned and processed (**B**) as in Figure 2D or (**C**) as in Figure 2E, as fewer highly gapped alignment columns and shorter aligned sequences can indicate improved homology between alignment entries. (**D**) Short sequences unlikely to encode a full-length COI-5P barcode are depleted from the input alignment using SequenceBouncer. Input and cleaned output alignments were de-gapped, and the distribution of resulting sequence lengths is shown. (**E**) An exploration of sample size during a sampling-based approach. The alignment of bird COI-5P barcodes was used as input for SequenceBouncer with k setting of 1, a gap threshold of 5%, a stringency setting of 'moderate', and a trial number of 10. The percent of outlier determinations from the sampling-based approach that match with the output of a full analysis that takes advantage of all pairwise sequence comparisons is shown. (**F**) An exploration of trial number during a sampling-based approach. The alignment of bird COI-5P barcodes was used as input with k setting of 1, a gap threshold of 5%, a stringency setting of 'moderate', and a sample size of 100. The percent of outlier determinations matching with the output of a full analysis after the indicated number of trials is reported. (**G**) Entries for which the outcome of a sampling-based analysis does not match the full analysis lie near the CO value found during the full analysis. Bird COI-5P barcodes are ordered by increasing median nCSs. Samples for which the outcome of a full analysis matches or does not match the sampling-based decision on sequence retention are labelled in blue and red, respectively. The CO value obtained during a full alignment analysis with all pairwise sequence comparisons is designated by an orange line.

Figure 5: Identification of outlier sequences within an alignment of COVID-19 sequences. SequenceBouncer was used to analyze an input alignment by a sampling-based approach over 10 trials at a gap threshold of 2%, a k setting of 1, and a sample size of 100. The analyses were performed at minimal, moderate, and maximum stringency. (**A**) Apparently truncated sequences shorter than the reference sequence are removed from the input alignment (37,470 sequences) at minimal stringency (33,351 sequences remain), moderate stringency (29,821 sequences remain), or maximum stringency (27,147 sequences remain). (**B**) as in (**A**), but showing a range with minimum and maximum approaching the COVID-19 reference sequence length. (**C**) Removal of COVID-19 outlier sequences leads to a decrease in heavily gapped columns and an increase in minimally gapped columns upon realignment. Cleaned alignments were realigned, and the fraction of heavily gapped columns (95-100% of samples) and seldom-gapped columns (0-5% of samples) was calculated.

Figure 6: COVID-19 alignment quality appears to be augmented as SequenceBouncer stringency is increased. The COVID-19 input alignment (**A**), as well as the realignments of files cleaned by a sampling-based methodology at minimal (**B**), moderate (**C**), and maximum (**D**) stringency are visualized using AliView (27) at a comparable level of magnification.

References:

1. P. Kapli, Z. Yang, M. J. Telford, Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* **21**, 428–444 (2020).

2. W. Fletcher, Z. Yang, The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* **27**, 2257–2267 (2010).

3. T. H. Ogden, M. S. Rosenberg, Multiple sequence alignment accuracy and phylogenetic inference. *Syst. Biol.* **55**, 314–328 (2006).

4. C. E. Shannon, A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423 (1948).

5. T. D. Schneider, Information Theory Primer: With an Appendix on Logarithms (2007).

6. P. S. Shenkin, B. Erman, L. D. Mastrandrea, Information-theoretical entropy as a measure of sequence variability. *Proteins: Structure, Function, and Genetics* **11**, 297–313 (1991).

7. N. A. O'Leary, *et al.*, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).

8. D. R. Smith, Revisiting published genomes with fresh eyes and new data: Revising old sequencing data can yield unexpected insights and identify errors. *EMBO Rep.* **20**, e49482 (2019).

9. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

10. D. G. Higgins, P. M. Sharp, CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244 (1988).

11. G. E. Hill, Mitonuclear Ecology. *Mol. Biol. Evol.* **32**, 1917–1927 (2015).

12. M. Pentinsaari, H. Salmela, M. Mutanen, T. Roslin, Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. *Sci. Rep.* **6**, 35275 (2016).

13. P. D. N. Hebert, A. Cywinska, S. L. Ball, J. R. deWaard, Biological identifications through DNA barcodes. *Proc. Biol. Sci.* **270**, 313–321 (2003).

14. S. Ratnasingham, P. D. N. Hebert, BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes* **7**, 355–364 (2007).

15. C. D. Dunn, B. A. Akpınar, V. Sharma, An Unusual Amino Acid Substitution Within Hummingbird Cytochrome c Oxidase Alters a Key Proton-Conducting Channel. *G3* **10**, 2477–2485 (2020).

16. S. Zheng, P. Poczai, J. Hyvönen, J. Tang, A. Amiryousefi, Chloroplot: An Online Program for the Versatile Plotting of Organelle Genomes. *Front. Genet.* **11**, 576124 (2020).

17. P. Jehl, F. Sievers, D. G. Higgins, OD-seq: outlier detection in multiple sequence alignments. *BMC Bioinformatics* **16**, 269 (2015).

18. M. Borowiec, Spruceup: fast and flexible identification, visualization, and removal of outliers from large multiple sequence alignments. *JOSS* **4**, 1635 (2019).

19. A. Płuciennik, *et al.*, BALCONY: an R package for MSA and functional compartments of protein variability analysis. *BMC Bioinformatics* **19**, 300 (2018).

20. A. Chiner-Oms, F. González-Candelas, EvalMSA: A Program to Evaluate Multiple Sequence Alignments and Detect Outliers. *Evol. Bioinform. Online* **12**, 277–284 (2016).

21. V. Ranwez, N. Chantret, Strengths and limits of multiple sequence alignment and filtering methods (2020).

22. M. L. Zepeda Mendoza, S. Nygaard, R. R. da Fonseca, DivA: detection of non-homologous and very divergent regions in protein sequence alignments. *BMC Res. Notes* **7**, 806 (2014).

23. C. M. Nugent, T. A. Elliott, S. Ratnasingham, S. J. Adamowicz, coil: an R package for cytochrome c oxidase I (COI) DNA barcode data cleaning, translation, and error evaluation. *Genome* **63**, 291–305 (2020).

24. W. Shen, S. Le, Y. Li, F. Hu, SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).

25. K. Katoh, J. Rozewicki, K. D. Yamada, MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2019).

26. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

27. A. Larsson, AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).

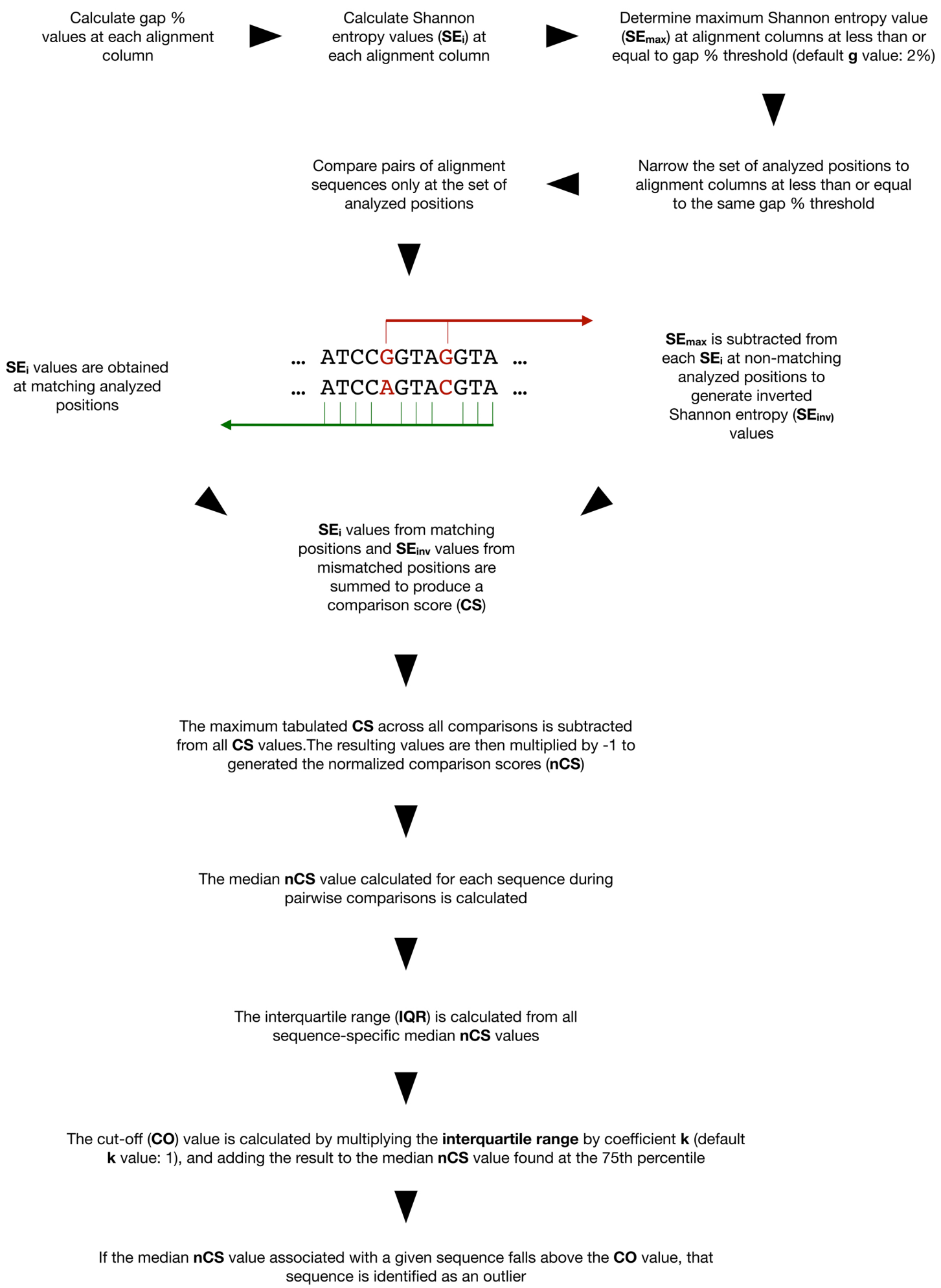
Calculate gap % values at each alignment column
Calculate Shannon entropy values (SEi) at each alignment column
Determine maximum Shannon entropy value (SEmax) at alignment columns at less than or equal to gap % threshold (default g value: 2%)
Narrow the set of analyzed positions to alignment columns at less than or equal to the same gap % threshold
Compare pairs of alignment sequences only at the set of analyzed positions
... ATCCGGTAGGTA ...
... ATCCAGTACGTA ...
SEi values are obtained at matching analyzed positions
SEmax is subtracted from each SEi at non-matching analyzed positions to generate inverted Shannon entropy (SEinv) values
SEi values from matching positions and SEinv values from mismatched positions are summed to produce a comparison score (CS)
The maximum tabulated CS across all comparisons is subtracted from all CS values.The resulting values are then multiplied by -1 to generated the normalized comparison scores (nCS)
The median nCS value calculated for each sequence during pairwise comparisons is calculated
The interquartile range (IQR) is calculated from all sequence-specific median nCS values
The cut-off (CO) value is calculated by multiplying the interquartile range by coefficient k (default k value: 1), and adding the result to the median nCS value found at the 75th percentile
If the median nCS value associated with a given sequence falls above the CO value, that sequence is identified as an outlier

Figure 1

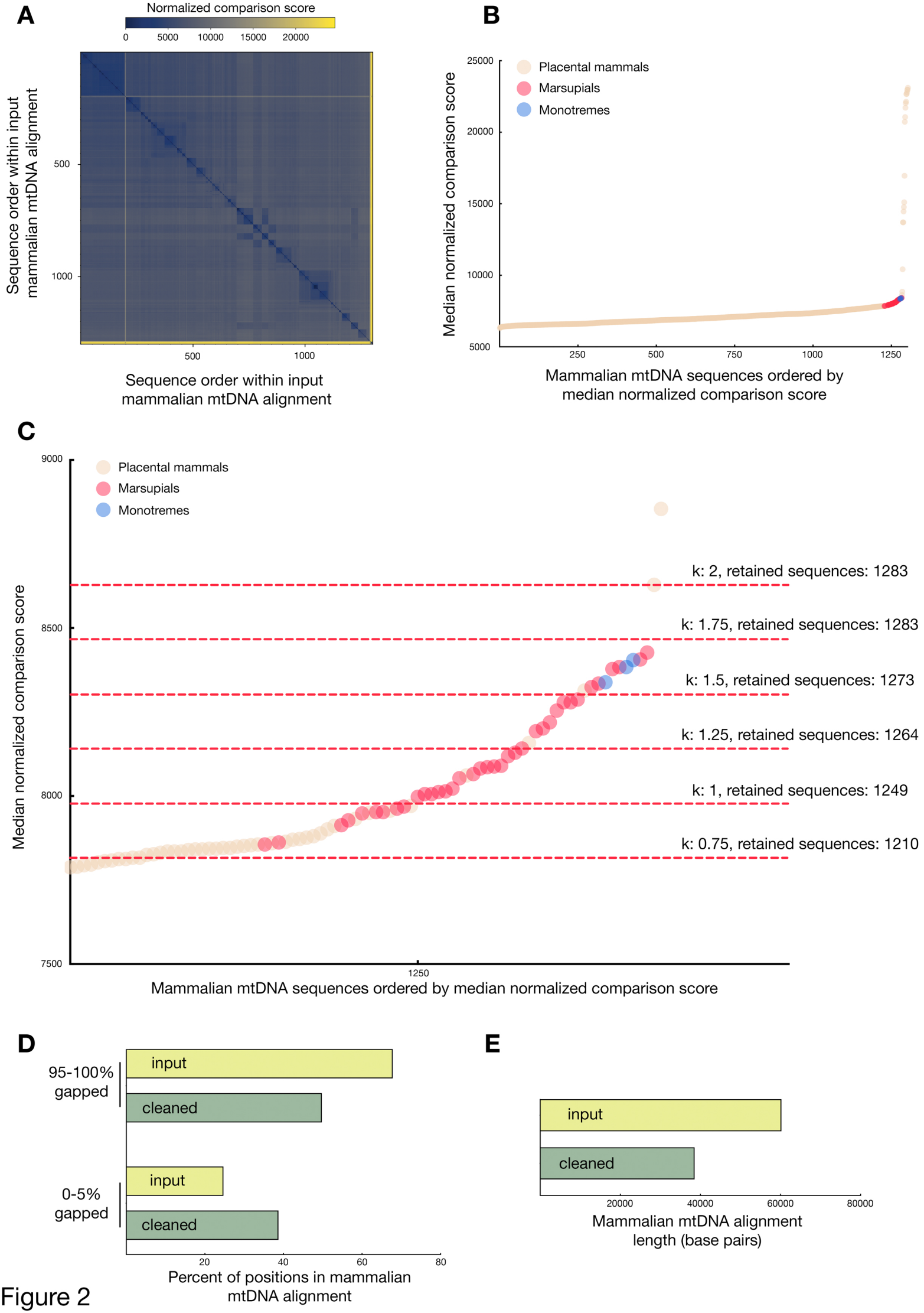
A
Normalized comparison score
0
5000
10000
15000
20000
Sequence order within input mammalian mtDNA alignment
500
1000
Sequence order within input mammalian mtDNA alignment
B
Placental mammals
Marsupials
Monotremes
Median normalized comparison score
25000
20000
15000
10000
5000
250
500
750
1000
1250
Mammalian mtDNA sequences ordered by median normalized comparison score
C
Placental mammals
Marsupials
Monotremes
Median normalized comparison score
9000
8500
8000
7500
1250
k: 2, retained sequences: 1283
k: 1.75, retained sequences: 1283
k: 1.5, retained sequences: 1273
k: 1.25, retained sequences: 1264
k: 1, retained sequences: 1249
k: 0.75, retained sequences: 1210
Mammalian mtDNA sequences ordered by median normalized comparison score
D
95-100% gapped
input
cleaned
0-5% gapped
input
cleaned
20
40
60
80
Percent of positions in mammalian mtDNA alignment
E
input
cleaned
20000
40000
60000
80000
Mammalian mtDNA alignment length (base pairs)

Figure 2

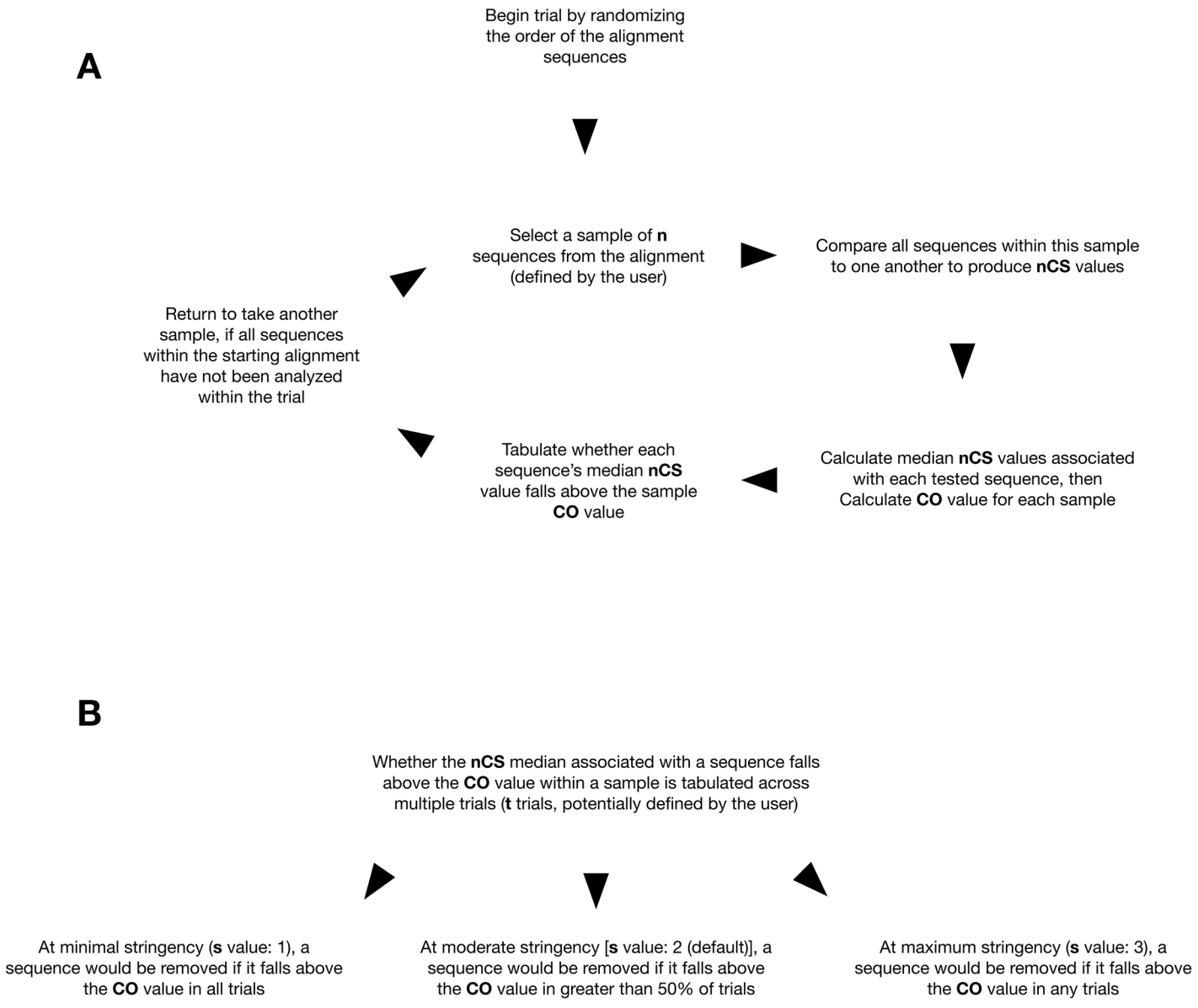
A
Begin trial by randomizing the order of the alignment sequences
Select a sample of **n** sequences from the alignment (defined by the user)
Compare all sequences within this sample to one another to produce **nCS** values
Calculate median **nCS** values associated with each tested sequence, then Calculate **CO** value for each sample
Tabulate whether each sequence's median **nCS** value falls above the sample **CO** value
Return to take another sample, if all sequences within the starting alignment have not been analyzed within the trial
B
Whether the **nCS** median associated with a sequence falls above the **CO** value within a sample is tabulated across multiple trials (**t** trials, potentially defined by the user)
At minimal stringency (**s** value: 1), a sequence would be removed if it falls above the **CO** value in all trials
At moderate stringency [**s** value: 2 (default)], a sequence would be removed if it falls above the **CO** value in greater than 50% of trials
At maximum stringency (**s** value: 3), a sequence would be removed if it falls above the **CO** value in any trials

Figure 3

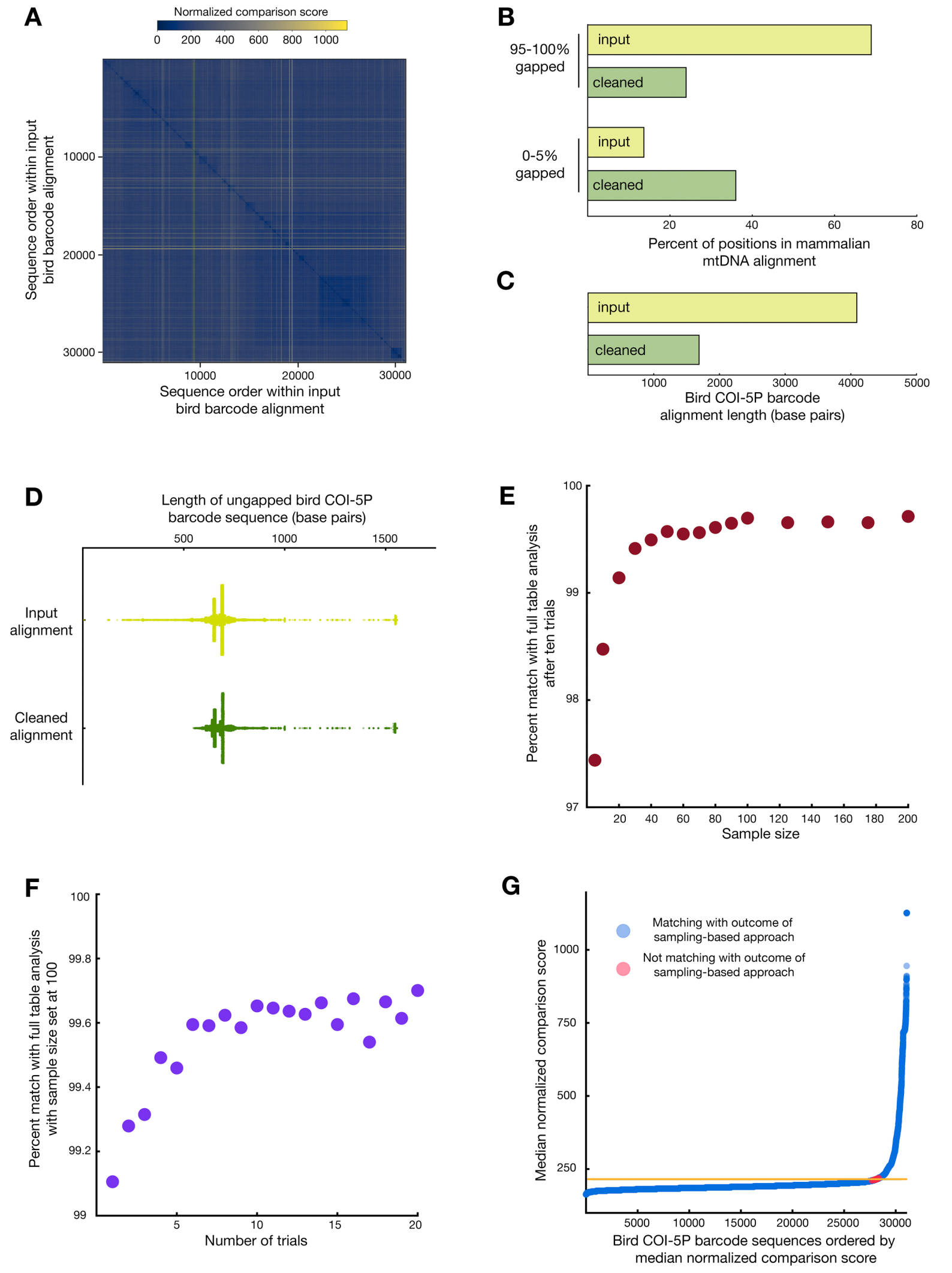
A
Normalized comparison score
0 200 400 600 800 1000
Sequence order within input bird barcode alignment
10000
20000
30000
Sequence order within input bird barcode alignment
B
95-100% gapped
input
cleaned
0-5% gapped
input
cleaned
20 40 60 80
Percent of positions in mammalian mtDNA alignment
C
input
cleaned
1000 2000 3000 4000 5000
Bird COI-5P barcode alignment length (base pairs)
D
Length of ungapped bird COI-5P barcode sequence (base pairs)
500 1000 1500
Input alignment
Cleaned alignment
E
Percent match with full table analysis after ten trials
100
99
98
97
20 40 60 80 100 120 140 160 180 200
Sample size
F
Percent match with full table analysis with sample size set at 100
100
99.8
99.6
99.4
99.2
99
5 10 15 20
Number of trials
G
Matching with outcome of sampling-based approach
Not matching with outcome of sampling-based approach
Median normalized comparison score
1000
750
500
250
5000 10000 15000 20000 25000 30000
Bird COI-5P barcode sequences ordered by median normalized comparison score

Figure 4

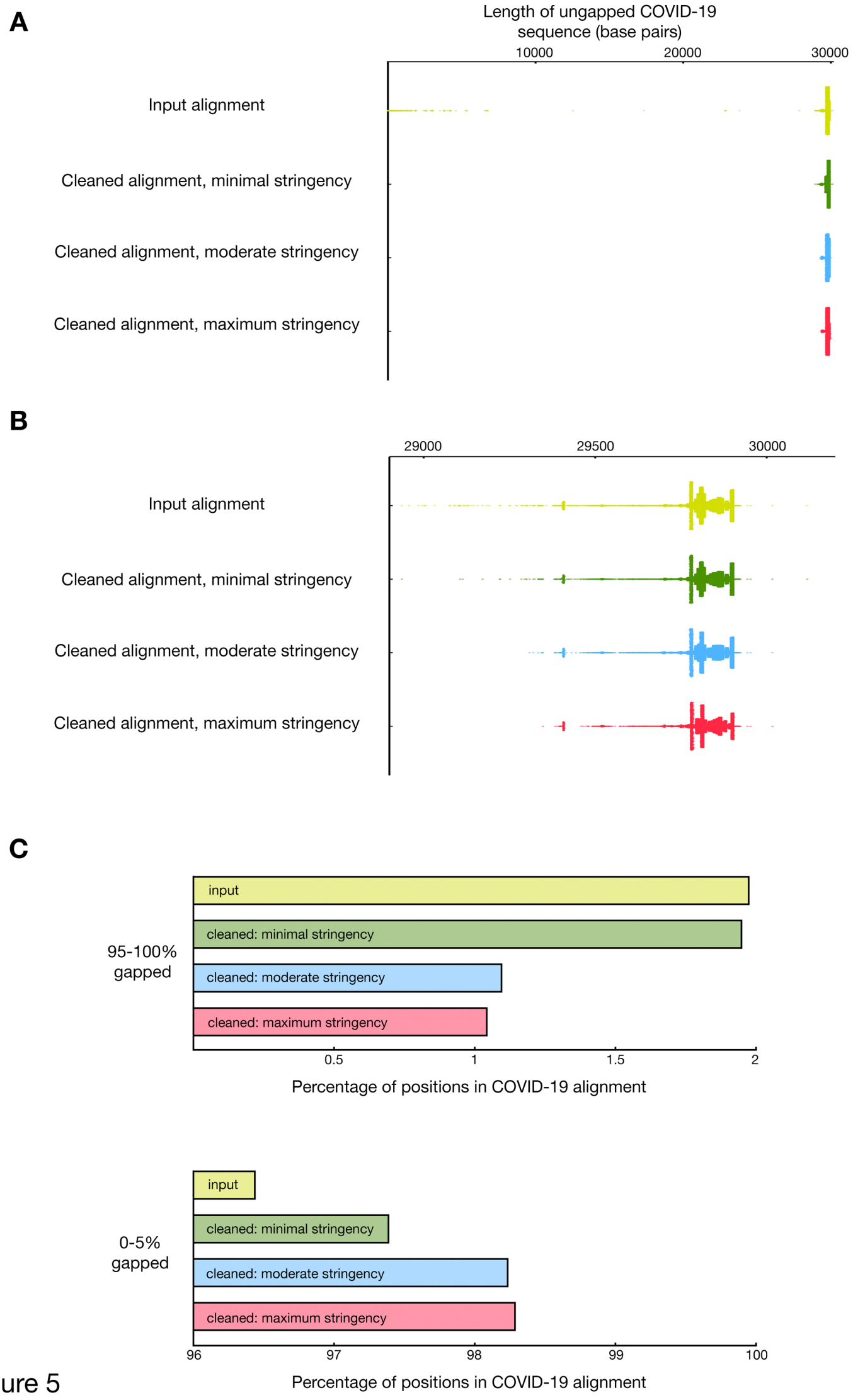
A
Length of ungapped COVID-19 sequence (base pairs)
10000
20000
30000
Input alignment
Cleaned alignment, minimal stringency
Cleaned alignment, moderate stringency
Cleaned alignment, maximum stringency
B
29000
29500
30000
Input alignment
Cleaned alignment, minimal stringency
Cleaned alignment, moderate stringency
Cleaned alignment, maximum stringency
C
95-100% gapped
input
cleaned: minimal stringency
cleaned: moderate stringency
cleaned: maximum stringency
0.5
1
1.5
2
Percentage of positions in COVID-19 alignment
0-5% gapped
input
cleaned: minimal stringency
cleaned: moderate stringency
cleaned: maximum stringency
96
97
98
99
100
Percentage of positions in COVID-19 alignment

Figure 5

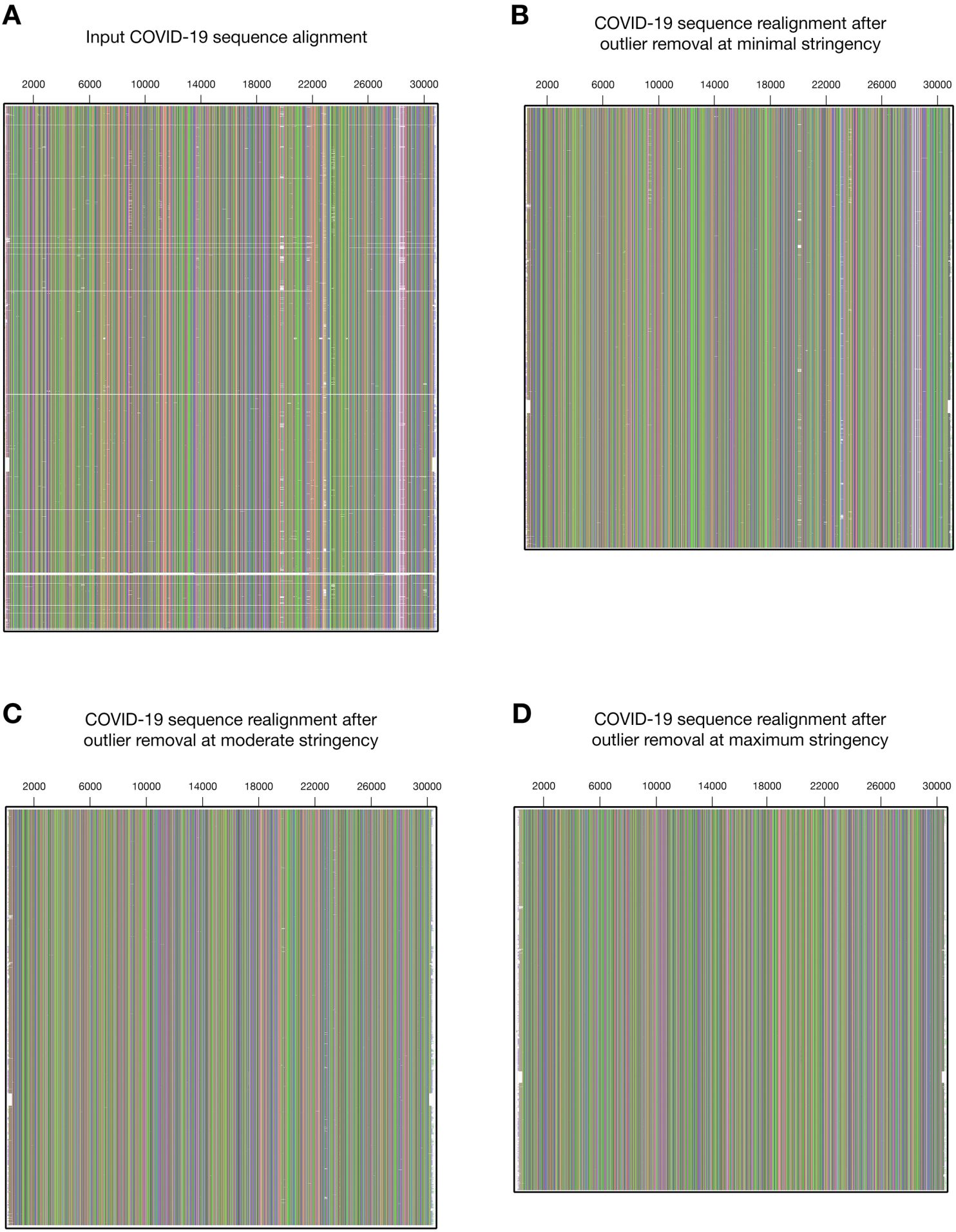
A
Input COVID-19 sequence alignment
2000
6000
10000
14000
18000
22000
26000
30000
B
COVID-19 sequence realignment after outlier removal at minimal stringency
2000
6000
10000
14000
18000
22000
26000
30000
C
COVID-19 sequence realignment after outlier removal at moderate stringency
2000
6000
10000
14000
18000
22000
26000
30000
D
COVID-19 sequence realignment after outlier removal at maximum stringency
2000
6000
10000
14000
18000
22000
26000
30000

Figure 6