nature communications

Article

https://doi.org/10.1038/s41467-022-34630-w

Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny

Received: 10 February 2022

Accepted: 1 November 2022

Published online: 15 November 2022

Check for updates

Robert C. Edgar[1]

Multiple sequence alignments are widely used to infer evolutionary relationships, enabling inferences of structure, function, and phylogeny. Standard practice is to construct one alignment by some preferred method and use it in further analysis; however, undetected alignment bias can be problematic. I describe Muscle5, a novel algorithm which constructs an ensemble of high-accuracy alignment with diverse biases by perturbing a hidden Markov model and permuting its guide tree. Confidence in an inference is assessed as the fraction of the ensemble which supports it. Applied to phylogenetic tree estimation, I show that ensembles can confidently resolve topologies with low bootstrap according to standard methods, and conversely that some topologies with high bootstraps are incorrect. Applied to the phylogeny of RNA viruses, ensemble analysis shows that recently adopted taxonomic phyla are probably polyphyletic. Ensemble analysis can improve confidence assessment in any inference from an alignment.

Multiple sequence alignment (MSA) algorithms are ubiquitous in molecular biology, with popular software such as `Clustal-Omega`[1], `MAFFT`[2] and `MUSCLE`[3] receiving hundreds of citations per year. Despite decades of research into automated alignment, current algorithms predict > 30% columns incorrectly on structure-based benchmarks[4,5]. Most alignment algorithms are based on highly simplified models of evolution parameterised by substitution scores and gap penalties. Default values for model parameters are somewhat arbitrary as they are trained on data of varying relevance to a particular set of sequences in practice. Systematic changes, and hence the opportunity for systematic errors, may be induced by changing parameters. For example, reducing gap penalties tends to increase the number of gaps. Most algorithms, including Clustal-Omega, MAFFT and MUSCLE, use progressive alignment according to a guide tree[6] which may cause bias towards this tree e.g. in an estimated phylogeny[7]. However, standard practice is to construct a single MSA using some preferred method and proceed on the assumption that bias (henceforth understood to include alignment errors of any kind which may affect downstream inference) can be neglected.

The `Muscle5` algorithm constructs a collection (*H-ensemble*) of high-accuracy MSAs (*replicates*) such that no particular MSA from the collection (or by any other method) is preferred a priori (see Supplementary Table S1 for summary of terminology). An MSA is built following the strategy pioneered by `ProbCons`[8]: posterior probabilities for aligning all letter pairs are computed using a hidden Markov model (HMM), a consistency transformation[9] is applied, and the final MSA is constructed by maximum expected accuracy pair-wise alignments[10] by progressive alignment according to a guide tree. HMM parameters and its guide tree are fixed in one replicate and varied between replicates to maximise differences in bias between replicates without degrading accuracy. If each replicate has different bias, then averaging results over replicates can correct for bias, and comparing results from different replicates can assess whether bias is important in a particular downstream analysis. Variations are introduced by multiplying HMM probabilities by a random number in the range − 0.25 … + 0.25, the largest range found to maintain accuracy on structural benchmarks. The guide tree, an important potential source of bias, is also varied by permuting the joining order of large subgroups close to the root. A divide-and-conquer strategy enables scaling to tens of thousands of sequences (details and survey of related prior work in Methods and Supplementary Methods). Compared to earlier versions of MUSCLE, `Muscle5` has substantially better accuracy (primarily achieved by

[1]Independent Researcher, https://www.drive5.com. e-mail: robert@drive5.com

maximum expected accuracy alignments and the consistency transformation) and scales to much larger datasets; compared to Prob-Cons, Muscle5 has marginally better accuracy, scales to much larger datasets, and adds support for nucleotide alignments.

The *H-ensemble confidence* (HEC) of inference from an MSA is the fraction of replicates which supports it. For example, if all replicates support the same inference, then HEC = 1. HEC is calculated using a *diversified* H-ensemble, which is designed to generate the greatest possible variety in the alignments, especially in systematic errors, so that averaging over the ensemble mitigates MSA bias. This approach can be applied to the alignments themselves: the Column Confidence (CC) is the HEC of an alignment column, i.e. the fraction of replicates where the column is reproduced. Unlike typical conservation-based metrics, a column with many gaps or with biochemically dissimilar amino acids will be assigned high CC if it is consistently reproduced. Alignment Confidence (AC) is the mean column confidence of a replicate, and MAC is the mean AC over the ensemble. If MAC = 1, all replicates are identical and the alignment is robust; if MAC is smaller, then the alignment is more sensitive to small parameter adjustments. Differences between alignments necessarily reflect errors, and lower MAC values, therefore, necessarily indicate higher error rates in a typical MSA from the ensemble. The robustness of a phylogenetic tree against MSA bias can be assessed by comparing replicate trees, i.e. trees estimated from different alignment replicates (Fig. 1). Edge Confidence (EC) is the HEC of a tree edge, i.e. the fraction of replicate trees where the edge is reproduced. Topology Confidence (TC) is the fraction of replicates supporting the branching order of designated subgroups such as taxonomic clades, and Ensemble Monophyly (EM) is the mean monophylicity for a subgroup. These phylogenetic tests are independent of bootstrapping as there is no re-sampling of columns.

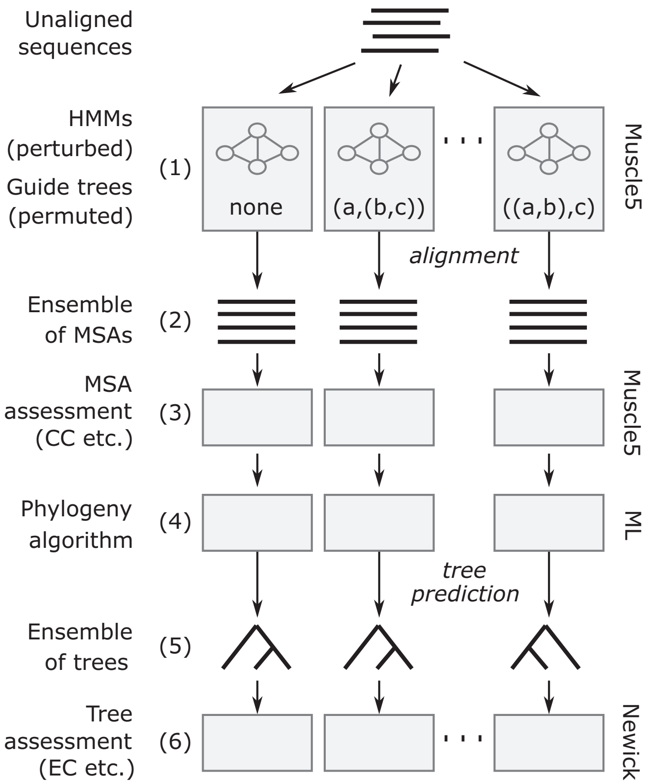


Fig. 1 | Typical ensemble workflow for alignment and phylogeny assessment. An ensemble of MSAs is generated and assessed for accuracy using Muscle5. Gray rectangles are processing steps made by an algorithm or software package. First, Muscle5 (step 1) generates an ensemble of MSAs (step 2), each alignment is generated by a different combination of a perturbed HMM and permuted guide tree. The accuracy of the MSAs can be assessed by Muscle5 (step 3) using accuracy metrics such as Column Confidence (CC). A phylogeny algorithm (step 4), e.g. maximum likelihood (ML), is used to predict a tree from each MSA (step 5). Finally, accuracy metrics, e.g. Ensemble Confidence (EC), are calculated from the resulting ensemble of trees (step 6). The Newick package (https://github.com/rcedgar/newick) was used to calculate the novel metrics described in this paper.

Guide tree bias can be assessed by comparing results on replicates where the guide tree is held fixed. If inferences differ and correlate with the choice of guide tree, then guide tree bias is present by definition. In general, a *stratified ensemble* has subsets (*strata*) where some parameters are held fixed while others are varied. Comparing results from different strata enables the detection of particular types of bias in any inference from an MSA.

In this work, I show that high-accuracy alignment ensembles enable novel, unbiased metrics of confidence in alignments and inferences therefrom.

Results

Alignment accuracy

As shown in Fig. 2 and Supplementary Table S2, the accuracy of Muscle5 on structure-based benchmarks (Balibase[4] for proteins and Bralibase[5] for RNA) is higher than state-of-the-art represented by Clustal-Omega and MAFFT. There is a negligible difference in average accuracy between parameter variants, showing that all replicates are equally plausible a priori. On a benchmark with 10,000 protein sequences per set, Muscle5 aligns 59% of columns correctly, which is a 13% improvement over Clustal-Omega (52% columns correct) and 26% over MAFFT (47% correct) (Supplementary Material). Figure 2 also shows that CC and AC provide predictive unsupervised estimates of accuracy (i.e., the estimates are independent of a trusted reference alignment). This is illustrated in Fig. 3, which shows replicate alignments of four proteins, focusing on a region with a well-conserved sequence and secondary structure and a more variable surface loop where neither sequence nor structure aligns well. This transition in secondary structure is reflected in the column confidence values, where the first 15 columns have $CC = 1.0$ and later values drop to $CC \approx 0.5$ in the loop, thereby identifying a segment where the alignment is error-prone.

RNA virus phylogeny

I investigated whether reported phylogenies of RNA viruses from the recent literature are reproducible and supportable, focusing on the topology of the four *Coronaviridae* genera and five *Riboviria* phyla inferred from alignments of the RNA-dependent RNA polymerase (RdRp) gene, which is widely used for phylogenetic and taxonomic analysis of viruses[11] (see Fig. 1 for workflow). Coronavirus genera have well-conserved RdRp alignments with amino acid (a.a.) identities ~ 60%. *Riboviria* overall are highly diverged with RdRp identities often as low as 5% to 10%, representing a very challenging case. I created diversified ensembles using Muscle5, finding MAC was 0.91 for genus, indicating generally high confidence with some variability in the alignments, but only 0.18 for phylum, indicating a high error rate. Trees were estimated by six different methods: RAxML[12], PhyML[13], IQ-Tree[14], FastTree[15], and neighbour-joining (NJ) and minimum-evolution (ME) using MEGA[16]. Trees were rooted using outgroups *Torovirus* for genus and reverse transcriptases for phylum. As shown in Fig. 4, four of the tree methods report a strong consensus (((A,B),G),D) for the genus topology (A = *Alphacoronavirus*, B = *Betacoronavirus*, G = *Gammacoronavirus* and D = *Deltacoronavirus*), while the faster but more approximate methods MEGA-NJ and FastTree reported more variants. The combined ensemble topology confidence of the consensus is 98.4% (82.5%) including (excluding) the latter two methods. A conventional analysis using the default Muscle5 MSA gave low bootstrap confidence to most edges (Fig. 5). Here, ensemble analysis confidently resolves topology while a single MSA with bootstrapping does not. For phylum, there is no consensus; all tree methods report varying

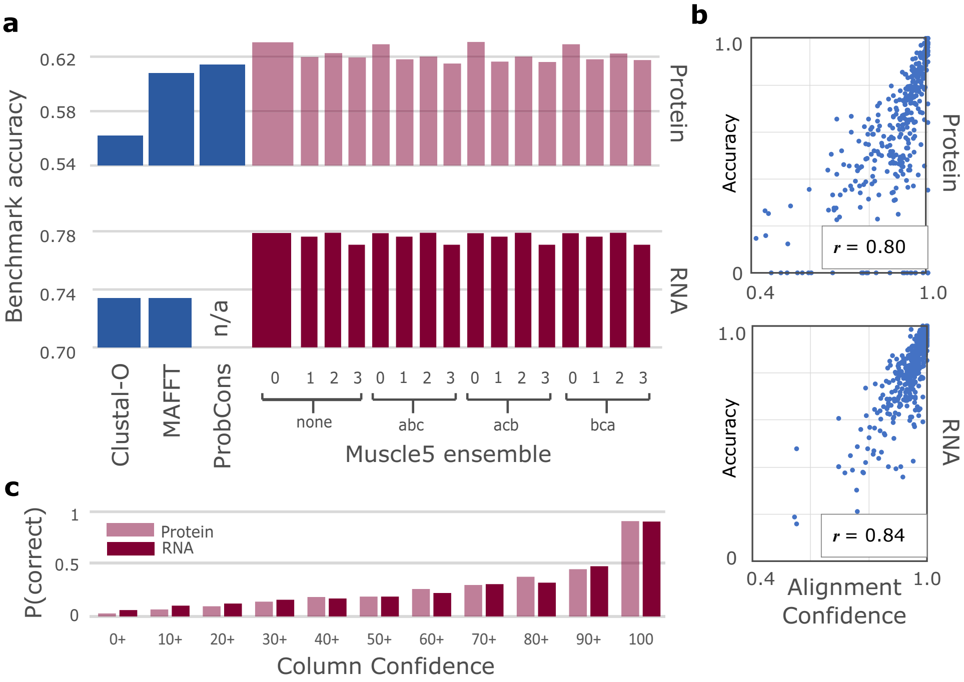


Fig. 2 | Accuracy of Muscle5 on structure-based benchmarks. a Average accuracy of Muscle5 ensemble replicates compared to `Clustal-Omega`, `ProbCons` and `MAFFT` on benchmarks of protein[4] (top) and RNA[5] (bottom) alignments; the default variant *none.0* is the wider bar. **b** Correlation between AC and accuracy (fraction correct columns) for protein (top) and RNA (bottom); Pearson's $r = 0.80, 0.84$, respectively. **c** Probability that a column is correct after binning into CC percentage intervals: 0 + is $0\% \le CC < 10\%$, 10+ is $10\% \le CC < 20\%$ etc.; the last bar is $CC = 100\%$. Thus CC is predictive of correctness and AC is predictive of accuracy.

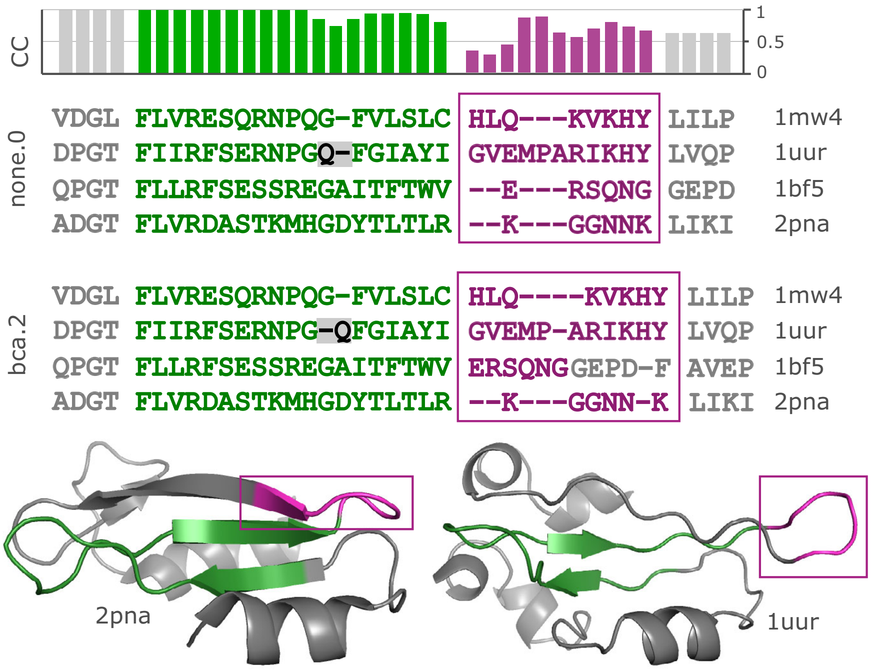
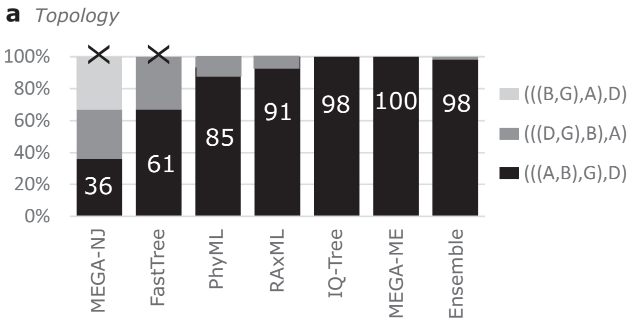

CC
1
0.5
0
none.0
VDGL FLVRESQRNPQG-FVLSLC HLQ---KVKHY LILP 1mw4
DPGT FIIRFSERNPGQ-FGIAYI GVEMPARIKHY LVQP 1uur
QPGT FLLRFSESSREGAITFTWV --E---RSQNG GEPD 1bf5
ADGT FLVRDASTKMHGDYTLTLR --K---GGNNK LIKI 2pna
bca.2
VDGL FLVRESQRNPQG-FVLSLC HLQ----KVKHY LILP 1mw4
DPGT FIIRFSERNPG-QFGIAYI GVEMP-ARIKHY LVQP 1uur
QPGT FLLRFSESSREGAITFTWV ERSQNGGEPD-F AVEP 1bf5
ADGT FLVRDASTKMHGDYTLTLR --K---GGNN-K LIKI 2pna
2pna
1uur

Fig. 3 | Replicate alignments of BBS11008. Two replicate alignments of a segment in `Balibase` set `BBS11008` are shown together with ribbon diagrams of two of its four structures (`2pna` and `1uur`). This region comprises a well-conserved anti-parallel beta-sheet (green) which transitions into a variable exposed loop (magenta, outlined by rectangles). Sequence homology in the beta-sheet is unambiguous except for one gap (grey background), while both sequence and structure similarity is unclear in the loop, which is reflected by lower CC values (top histogram; CC was calculated from a diversified ensemble of 100 replicates).

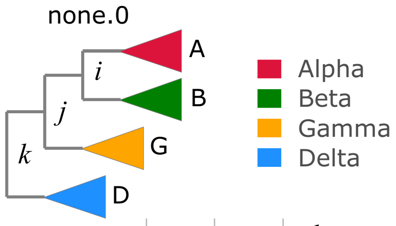
topologies across the ensemble (Fig. 6). Figure 7 shows trees and bootstraps obtained on MSAs with two different guide trees and other parameters fixed. All six tree methods agree with each other on the topology according to one of these MSAs, but the topologies conflict, and therefore, one or both must be wrong. Further, the bootstraps of most methods are high, with values ≥84 for all edges from all maximum-likelihood methods except for FastTree on edge *k*. Thus, for phylum, high bootstraps for at least one incorrect topology are necessarily induced by MSA bias, and ensemble analysis shows that the topology cannot be reliably resolved. Coronavirus trees reported in



b *Monophyly*

	RAxML	PhyML	FastTree	MEGA-ME	MEGA-NJ
Beta	97.6	96.4	98.4	100.0	100.0
Alpha	100.0	99.0	99.9	100.0	100.0
Gamma	100.0	100.0	100.0	100.0	100.0
Delta	82.0	100.0	100.0	100.0	100.0

Fig. 4 | Ensemble confidence of coronavirus genus topologies and monophyly. **a** Relative frequencies of tree topologies for coronavirus genera from a diversified ensemble using six different tree estimation methods. The rightmost bar shows the combined ensemble average with MEGA-NJ and FastTree excluded. A = *Alphacoronavirus*, B = *Betacoronavirus*, G = *Gammacoronavirus* and D = *Deltacoronavirus* **b** Ensemble Monophyly (EM) of coronavirus genera. All six tree estimation methods give confidence > 96% to monophyly of all genera except for *EM* = 82.0% for Deltacoronavirus by RaxML.



	i	*j*	*k*
RaxML	*53*	**86**	*40*
PhyML	*43*	*33*	**95**
IQ-Tree	*71*	*59*	*47*
MEGA-ME	**91**	*39*	*65*
MEGA-NJ	-	-	-
FastTree	-	-	-

Fig. 5 | Bootstraps for coronavirus consensus genus topology. The coronavirus genus topology is `(((A,B),G),D)` with high ensemble confidence (Fig. 4). Using the default Muscle5 MSA (none.0), this topology was reported by four of the six tree estimation methods with bootstraps as shown in the figure, where bootstrap values are mostly low.

recent literature are shown in Fig. 8, which are seen to have conflicting genus topologies with high bootstraps, suggesting that systematic alignment errors induced overconfidence.

Monophylicity of phyla in *Riboviria*

A deep RNA virus phylogeny was recently reported in[11] (hereinafter Wolf2018) and subsequently used as the basis for introducing new taxonomic ranks including phylum[17]. I measured the monophylicity of the new phyla on a diversified ensemble. For each MSA, a tree was estimated using RAxML, and the best-fit subtree identified for each phylum. To investigate whether the Wolf2018 MSA might be more accurate than Muscle5, I checked the alignment of essential catalytic residues, finding that Muscle5 is better by this measure (Fig. 9, Methods). Results are shown in Fig. 10. Panel **a** is the Wolf2018 tree topology showing the high reported bootstrap values which supported the

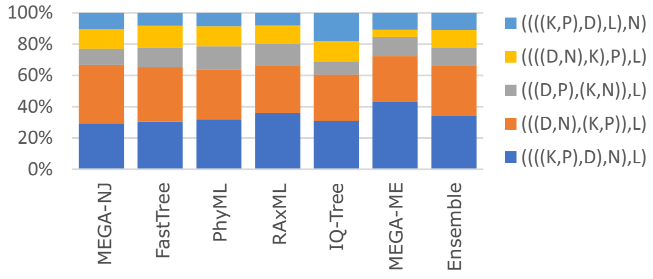
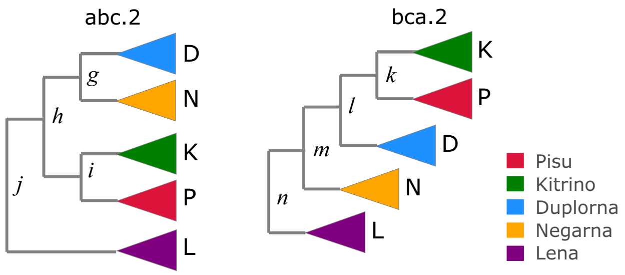


Fig. 6 | Ensemble frequencies of phylum topologies. Relative frequencies of tree topologies of Ribovirus phyla from a diversified ensemble using six different tree estimation methods. The rightmost bar shows the combined ensemble average. D = *Duplornaviricota*, K = *Kitrinoviricota*, L = *Lenarviricota*, N = *Negarnaviricota*, P = *Pisuviricota*.



	g	*h*	*i*	*j*	*k*	*l*	*m*	*n*
RaxML	**88**	**90**	**100**	**100**	**80**	**91**	**86**	**96**
PhyML	**84**	**88**	**100**	**95**	**100**	**91**	**98**	**95**
IQ-Tree	**98**	**100**	**96**	**97**	**95**	**98**	**99**	**100**
FastTree	**100**	**93**	**90**	**90**	*4*	**90**	**96**	**97**
MEGA-ME	*49*	*67*	**84**	**92**	**85**	**80**	*57*	**86**
MEGA-NJ	*40*	*37*	*58*	*75*	*59*	*58*	*65*	*68*

Fig. 7 | Phylum topologies estimated from replicates abc.2 and bca.2. HMM parameters are held fixed for making the MSAs (both have random seed 2) while the guide tree topology is permuted. All six tree estimation methods agree with each other on the topology on a given MSA, but the topologies are different so one or both topologies must be wrong and the reproducible wrong tree must be induced by guide tree bias as the MSA is otherwise unchanged. Most methods give high bootstraps (shown in table below the trees) for most or all of the edges.

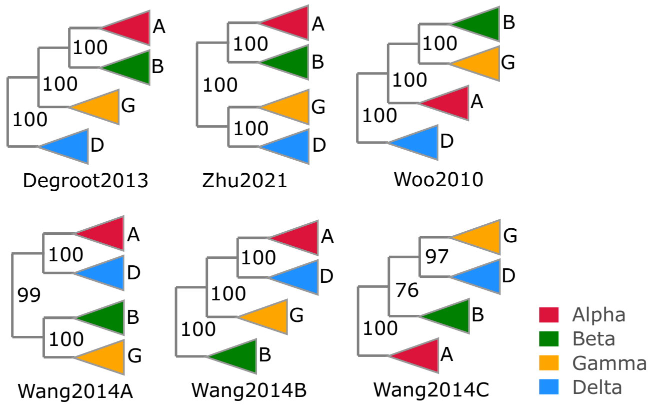
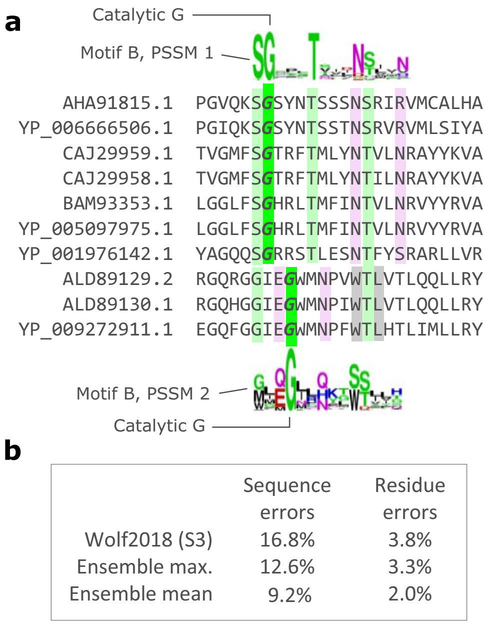


Fig. 8 | Conflicting coronavirus genus topologies in the literature. Trees from four published papers reporting high bootstraps for conflicting genus topologies: Degroot2013[31], Wang2014[32], Woo2010[33] and Wang2014[32]. Wang2014 estimated trees from three different alignments: (A) whole genomes, (B) spike protein and (C) nucleocapsid protein.


a
Catalytic G
Motif B, PSSM 1
AHA91815.1 PGVQKSGSYNTSSSNSRIRVMCALHA
YP_006666506.1 PGIQKSGSYNTSSTNSRVRVMLSIYA
CAJ29959.1 TVGMFSGTRFTMLYNTVLNRAYYKVA
CAJ29958.1 TVGMFSGTRFTMLYNTILNRAYYKVA
BAM93353.1 LGGLFSGHRLTMFINTVLNRVYYRVA
YP_005097975.1 LGGLFSGHRLTMFINTVLNRVYYRVA
YP_001976142.1 YAGQQSGRRSTLESNTFYSRARLLVR
ALD89129.2 RGQRGGIEGWMNPVWTLVTLQQLLRY
ALD89130.1 RGQHGGIEGWMNPIWTLVTLQQLLRY
YP_009272911.1 EGQFGGIEGWMNPFWTLHTLIMLLRY
Motif B, PSSM 2
Catalytic G
b

	Sequence errors	Residue errors
Wolf2018 (S3)	16.8%	3.8%
Ensemble max.	12.6%	3.3%
Ensemble mean	9.2%	2.0%

Fig. 9 | Misaligned catalytic residues in RdRp MSAs. Misalignments of essential catalytic residues were identified using Palmscan. **a** Ten representative sequences from the Wolf2018 alignment are shown. The top seven sequences place the catalytic glycine (G) in motif B in a different column than the bottom three. Sequence logos for the relevant Palmscan PSSMs are shown above and below the alignment. **b** Percentages of sequences with at least one catalytic residue misalignment and the total number of residue misalignments for Wolf2018 alignment S3 and the maximum and mean values on the corresponding Muscle5 ensemble. S3 is a subset alignment used by Wolf2018 to estimate the top-level (near-root) branching order of their tree, it contains 238 sequences. The equivalent Muscle5 ensemble has 249 sequences per MSA, selecting different subsets in addition to different alignment parameters to construct replicates. Note that all Muscle5 replicates have substantially fewer errors than S3.

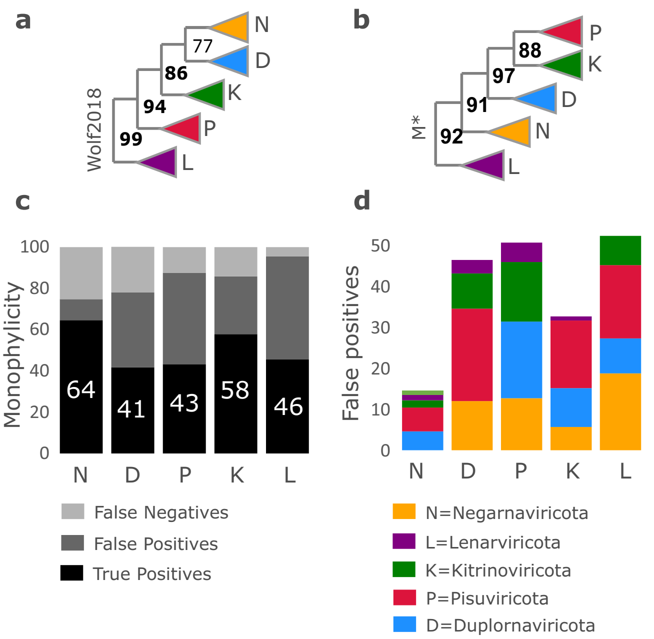


Fig. 10 | Monophylicity of RNA virus phyla. a Wolf2018 tree showing high bootstrap values as reported in their paper. **b** RAxML tree estimated from M^{*} (Muscle5 replicate with highest AC) showing high bootstrap values for a conflicting topology. **c** Mean false positive frequencies as a percentage of the best-fit subtree, averaged over a diverse ensemble. **d** Monophyly of the tree in panel of the best-fit subtree as percentage TP, FP and FN, respectively, averaged over a diverse ensemble. Note that TP is low, ranging from 43% for *Pisuviricota* to 65% for *Negarnaviricota*. Compare with Fig. 4 (**b**)shows high monophyly of coronavirus genera.

putative monophylicity of the new phyla, panel **b** shows a typical tree from the ensemble exhibiting similarly high bootstrap values for a conflicting topology. Panel **c** shows monophyly of the phyla, where EM ranges from 41% for *Duplornaviricota* to 64% for *Negarnaviricota*. Panel **d** shows the composition of false-positive leaves under the best-fit subtree for each phylum, showing substantial mixing of all pairs of phyla (note that while *Negarnaviricota* has relatively few FPs, it has 25% FNs mixed into other subtrees). Combined, these results strongly suggest that the high bootstrap values in the Wolf2018 tree are artefacts of biases in their MSA, and the newly-adopted phyla are far from monophyletic.

Discussion

High accuracy ensembles enable improved confidence estimates

It is de facto standard practice in biological sequence analysis to make a "best effort" analysis based on one preferred alignment, proceeding on the implicit assumption that this alignment is correct, or at least good enough to make the desired downstream inferences. Sometimes, columns considered to be less reliable (e.g. "gappier") may be discarded, but the possible impact of any remaining biases are almost universally neglected, presumably due to a lack of awareness that they may be present combined with a lack of convincing methods for identifying and mitigating such errors. The results presented here show that alignment bias can have a significant impact, and as an important special case that tree bootstrapping from a single MSA may give high confidence to incorrect edges caused by bias in the MSA. In contrast to previous ensemble methods, Muscle5 generates alignments with a greater diversity of systematic errors (if present) while maintaining state-of-the-art benchmark accuracy, thereby enabling detection and mitigation of incorrect inferences due to MSA bias.

Ensemble analysis complements bootstrapping

The Felsenstein bootstrap relies on several assumptions[18]: the alignment is correct, sites evolve independently according to Markov models, the best tree is successfully identified for each resampled alignment, and the best tree will converge on the true tree as more columns are sampled from the underlying distribution, i.e., that the evolutionary model is a good enough approximation to identify the correct tree. All these assumptions are surely violated in practice. Alignments are often wrong. Sites are not independent, and their evolution is not strictly Markovian. The space of trees is too large to search exhaustively for more than a few leaves. The best tree may not be the true tree, which seems almost certain in non-trivial cases because evolutionary models are highly simplified. Thus, bootstrapping may be unreliable, as illustrated by examples presented in this work. Ensemble analysis assumes that alignment errors are sampled sufficiently across the ensemble to induce variations in downstream inferences comparable in size with a typical inference error. This assumption is violated if the alignment is robust against parameter perturbations but nevertheless wrong. Thus, for assessment of phylogenies, bootstrapping and ensemble analysis are complementary. If there is no variation in the ensemble, then standard bootstrapping alone is appropriate. If the ensemble is variable, ensemble confidence may be more credible than bootstrapping, as shown by the examples of coronavirus genera and RNA virus phyla where bootstrapping and ensemble confidence imply different conclusions but the ensemble is more credible.

Extending the ensemble approach

Ensemble confidence can straightforwardly be applied to other inferences from an alignment, such as predicted secondary structure, paralog/ortholog discrimination and so on. The topology of putative clades can be assessed by comparing results when different subsets of sequences are selected; in particular resampling of sequences with replacement can be considered bootstrapping of rows rather than columns, noting that the rows should be realigned after resampling. If phylogeny is estimated from multiple genes, results can be compared on different subsets of genes. The Felsenstein bootstrap can be improved by sampling columns from an MSA ensemble rather than a single MSA (*ensemble bootstrap*), thereby accounting for alignment uncertainty in addition to under-sampling of columns[19]. However, I would not advocate relying on the ensemble bootstrap as a complete solution to accounting for MSA bias in phylogenetic tree estimation. While ensemble bootstrap values from a `Muscle5` ensemble will surely be more reliable than conventional bootstrapping from a single MSA, there is no substitute for intelligently assessing robustness by replicating an analysis in distinctly different ways, paying particular attention to likely sources of systematic bias.

Reported RdRp phylogeny is not replicated

The results reported here show definitively that high bootstrap values for the RdRp-based phylogeny reported in Wolf2018 are artefacts of MSA bias. Typical approaches to improving phylogeny inference from sequence alignments include adding other genes and removing less-conserved columns, but neither approach is applicable in this case because even the most conserved columns are not correctly aligned (Fig. 9), and other RNA virus genes cannot be used to infer deep branching order because RdRp is the only universal gene and the only gene with recognisable sequence similarity between more diverged groups. Ancient groups and their branching order reported by Wolf2018 are not reproduced in an ensemble analysis based on MSAs that are more accurate than the Wolf2018 alignment. This, it appears that new taxa based on these groups were introduced prematurely as they are probably far from monophyletic.

Ensembles enable sequence analysis replicates

Science is suffering from a replication crisis driven by practices such as p-value hacking, harking (hypothesis after result is known) and cherry-picking[20]. Sequence analysis software offers the biologist a bewildering array of alternatives for ubiquitous routine tasks such as alignment and tree-building. MUSCLE or MAFFT? The maximum likelihood or minimum evolution? RAxML or PhyML? How many discrete gamma categories should you have in your model? Common practice is to choose one protocol for a mix of stated and unstated reasons which may be more or less defensible: my colleague does it this way, it got a good benchmark score, or it gave a result I like better on my own data. Picking a single best protocol disregards the possibility that the best may not be good enough. You can guard against this pitfall by performing your own replication study. This may be as simple as trying different software packages with a few different options. Even if alternative protocols are believed to be less accurate, a thoughtful comparison of the results provides a useful indication of whether the preferred protocol can be trusted. While automated sequence analysis methods should never be entirely trusted, including methods for constructing replicates, high-accuracy alignment ensembles enable a substantial improvement in assessment of inferences in many areas of molecular biology.

Methods

Related work

Several methods for generating collections of alternative sequence alignments have previously been described in the literature. The earliest I am aware of date from 1995[21,22]. In[21] the author assessed the robustness of arthropod phylogenies under variation in transversion-transition probabilities and gap penalties, noting that "[the] disturbing circularity of the interaction between the specification of [insertion-deletion and substitution probabilities] a priori and their inference *a posteriori* is a general and central problem in molecular phylogenetics". The `Elision` method[22] concatenates variant MSAs before estimating a phylogenetic tree. A 1997 study[23] of 18S ribosomal RNA in *Apicomplexa* assessed robustness of phylogenetic inferences using multiple alignments from different software packages, finding that "different alignments produced trees that were on average more dissimilar from each other than did the different tree-building methods used". More recent proposals along similar lines include[19,24,25]. Muscle5 improves on previous ensemble methods in two crucial respects. First, all Muscle5 replicates have high accuracy such that no other MSA is preferred a priori, while in previous methods there is a clearly preferred MSA, i.e. the alignment generated by the algorithm and parameter combination giving the best benchmark score. In my terminology, previous methods generated S-ensembles while `Muscle5` generates H-ensembles (Supplementary Table S1). Second, `Muscle5` replicates explore a substantially greater range of possible biases by introducing more consequential variations into substitution scores and guide trees. Scaling to large datasets by a divide-and-conquer strategies was implemented in an update to `MAFFT`[26], followed by `Clustal-Omega`[27] and others. Below, the `Muscle5` algorithm is briefly described; further details of the algorithm and its improvements over previous work are provided in Supplementary Material.

Hidden Markov model

Posterior probabilities for aligning every pair of letters in the input sequences are calculated using a hidden Markov model (HMM) with topology shown in Supp. Fig. S1. This is a coupled Markov model[10], extended to double-affine gaps as found in `ProbCons` source code version 1.12, which differs from the HMM described in the paper[8]. For a pair of sequences x, y, alignment columns are emitted by the match state M and by insert states I_x, I_y, J_x and J_y. M emits a column containing an aligned pair of letters. I_x and J_x emit one letter from x, similarly I_x and J_x emit one letter from y. The I states induce short gaps while the J states induce longer gaps. Alignments begin in the start state S. Match state emission probabilities are obtained from the joint-probability form of the BLOSUM62 matrix[28]; insert states emit letters according to the marginal probabilities of BLOSUM62. In `ProbCons`, transition probabilities were trained by expectation-maximisation on version 2 of the `Balibase` benchmark. For `Muscle5`, I chose somewhat arbitrary round numbers, guided by the defaults in `ProbCons` and the premise that small differences should be immaterial to alignment quality. Symmetries are enforced between x and y and between sequences and reversed sequences, giving five independent sets of probabilities for mutually exclusive alternative events: (1) transitions from the start state S (symmetric with transitions into the end state E), (2) transitions out of M, (3) transitions out of I, (4) transitions out of J, and (5) letter-pair emissions.

Muscle5 algorithm

The core component of `Muscle5` is a parallelised re-implementation of `ProbCons`. Iterative consistency transformations are applied to the posterior probabilities (two rounds by default); a greedy maximum expected accuracy guide tree is constructed; and refinement by randomised partitioning is applied (100 rounds by default). Scaling to large datasets is achieved by a divide-and-conquer strategy (the `Super5` algorithm); details in Supplementary Material. A single MSA is generated by (1) perturbing the HMM, (2) choosing a guide tree permutation, and (3) executing parallel `ProbCons`. An ensemble is generated by repeating this procedure.

HMM perturbations

Perturbations are introduced by adjusting all probabilities according to the rule $P \rightarrow (1 + \alpha\delta)P$, where α (amplitude) is a constant and δ is a random number uniformly distributed between -1 and $+1$. All probabilities including transitions and emissions are perturbed.

Perturbations are introduced before calculating posteriors, then the HMM is held fixed. This procedure is independent of the data; there is no attempt to optimise parameters. By default, $\alpha = 0.25$, causing perturbations up to ± 25%. This value was chosen by trial and error on a subset of `Balibase`. To maintain normalisation (i.e., ensure that $0 < P < 1$ for all P and the sum over mutually exclusive events is 1), a second adjustment $P_k \to P_k/\sum_i P_i$ is applied after all probabilities have been perturbed by the first rule. The sum in the denominator is over probabilities in the same set of mutually exclusive events as k, e.g. transitions from M. A single parameter (α) sets a scale for all perturbations, reducing possible concerns about over-fitting (too many parameters) and over-tuning (choosing a value that performs well on the training set but poorly on new data) to a minimum.

Guide tree permutations

The goal of permuting the guide tree is to induce substantive variation into any systematic errors due to progressive alignment, without compromising accuracy. Optimising accuracy requires that closely-related sequences are aligned before more diverged sequences are added[3,6], which in turn requires that the guide tree joining order should be preserved close to its leaves. Substantive variations require that larger groups are joined in different orders. These constraints imply that the tree should be mostly unchanged close to the leaves and large changes should be made to the joining order of larger groups close to the root, but these goals can be conflicting in practice as guide trees are often highly unbalanced, i.e. many nodes join small groups to large groups, in which case naive re-arrangements of the tree may fail to induce substantive variations. With these considerations in mind, `Muscle5` manipulates the guide tree T as follows. An edge is identified which divides the leaves of T into subsets a and bc such that the ratio $|a|/|bc| \approx 1/2$, i.e. a has approximately one third of the leaves in T. The tree bc is then divided into subsets b and c of equal size so that $|b|/|c| \approx 1$. Regardless of the original guide tree topology, when there are many leaves this procedure successfully divides T into three subtrees a, b and c of approximately equal size where the joining order close to the leaves is mostly preserved. Progressive alignment is performed using the original guide tree and permutations $((a, b), c)$, $((a, c), b)$ and $((b, c), a)$, abbreviated to *none*, *abc*, *acb* and *bca* respectively. A replicate is identified as *perm.s*, e.g. `abc.3`, where perm is the guide tree permutation and s is the random number seed, where the special case $s = 0$ indicates that no perturbations are applied, i.e. default HMM parameters are used. See Supplementary Materials for further details and discussion.

Diversified ensemble

A diversified ensemble is designed to maximise variation among replicates by setting the random number seed $s = 0, 1...N$ where N is the desired number of replicates, while guide tree permutations cycle through the four variants *none*, *abc*, *acb* and *bca*. For the results reported in this work, $N = 100$ was used. Alternatively, convergence criteria could be set which terminate generation of further replicates when sufficient diversity has been sampled, though this feature was not implemented in the code described here. For convergence, I suggest an upper limit of m replicates where m is the median number of columns in the ensemble so far, and also testing the number of singleton distinct columns (n_1, found in exactly one replicate) compared to the number of reproduced distinct columns (n_2, found in two replicates). If $n_2 > n_1$, most of the potential diversity in the ensemble has been sampled. In practice, when sequences are closely related most of all replicates may be identical; in such cases setting convergence criteria would save substantial computing resources in high-throughput applications.

Best-fit subtree

Given a tree and categories assigned to a subset of its leaves (e.g., phylum names), the best fit for category C is identified as the node t which minimises the number of errors in its subtree. The number of true positives (TP) is the number of leaves under t which belong to C. Errors include false positives (FP, i.e. leaves under t which are not in C), and false negatives (FN, i.e. leaves in C which are not under t).

Ensemble Monophyly

Given a tree and a category C (e.g. phylum name), the best-fit subtree t is identified. Monophyly of C is then $m = TP/(TP + FP + FN)$. If C is monophyletic, $m = 1$, otherwise $m < 1$, and m is smaller with increasing errors (Supp. Fig. S7). Ensemble Monophyly is mean m over an ensemble of trees.

Root identification by an out-group

Given a tree and an out-group category C, the best-fit subtree is identified after assigning all leaves not belonging to C to a second category. The root is then placed in the edge joining the best-fit subtree of C to the rest of the tree. This procedure accommodates the trivial case where C is monophyletic in the estimated tree, and also more difficult cases where C is polyphyletic.

Condensed tree

Given a rooted tree T and category labels on a subset of its leaves (e.g., phylum names), the best-fit node is identified for each category (Supp. Fig. S7). Edges in the path from each best-fit node to the root are preserved, all other edges are deleted. This produces a tree T_b in which all leaves are best-fit nodes. Each unary node u in T_b is collapsed by replacing u and its incoming and outgoing edges u_i and u_o by a single edge e_u; this is repeated until all nodes have degree > 1. The length of e_u is the sum of the lengths of u_i and u_o; the bootstrap value assigned to e_u is the larger of the bootstraps for u_i and u_o. The resulting tree is the condensed tree of T according to the category labels. A condensed tree summarises the branching order of its categories, assuming they are monophyletic or approximately monophyletic so that best-fit nodes are estimates of most recent common ancestors.

RdRp alignment quality

The `Palmscan` algorithm[29] uses position-specific scoring matrices (PSSMs) to identify three well-conserved motifs in the RdRp palm domain, which are conventionally designated A, B and C, respectively. These motifs include six essential catalytic residues: two aspartic acids (`D`) in motif A, a glycine (`G`) in motif B, and `GDD` in motif C[30]. The quality of an RdRp alignment was assessed by using `Palmscan` to identify the position of each of these catalytic residues in all sequences, which successfully matched the PSSMs. If a catalytic residue appears in a different column from the majority of other sequences, it is considered to be misaligned (Fig. 9). The total number of misaligned catalytic residues was used as a quality metric.

Validation on simulated data

Alignment and tree inference methods are often validated by simulating sequence evolution in silico. I chose not to do so in this work. Simulations employ drastically simplified models of evolution which are similar to the drastically simplified models used by multiple alignment and tree inference algorithms. In reality, sequence evolution is an enormously complex process disrupted by historical contingencies ranging from fortuitous outcomes of DNA repair machinery failures and narrowly-won host-pathogen arms races to asteroid impacts. Therefore, simulations of deep evolutionary history are at best suggestive and at worst entirely uninformative if one is interested in real biology. Simulations also exacerbate a common sociological problem in computational biology, namely that the developers of a new method have an opportunity to fish for significance before publication. Confronted with disappointing results, authors may rationalise tweaking a simulation until improved (simulated) performance is obtained for their method. These considerations beg the question of whether simulations could convincingly support the main claims of this paper, which are (1) `Muscle5` MSA replicates have high and practically indistinguishable accuracy, and (2) the effects of alignment errors can be assessed by comparing inferences from different replicates. Claim (1) is supported by results on structure-based benchmarks. While structural similarity does not necessarily imply sequence homology, structural alignments are largely independent of sequence, and greater agreement with structural alignments therefore surely correlates strongly with more accurate alignment of homologous residues. If a simulation fails to recapitulate relative algorithm performance according to structure, the failure is more plausibly explained by a defect in the simulation than a defect in the structural benchmark. Claim (2) is self-evidently true because two different alignments of the same sequences cannot both be correct.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Code and scripts required to reproduce experimental results are available under the GPL v3 license at: https://github.com/rcedgar/rdrp_tree_experiments. Balibase v3 is available at https://www.re3data.org/repository/r3d100012946. Bralibase is available at https://projects.binf.ku.dk/pgardner/bralibase/.

Code availability

Binaries and source code are available under the GPL v3 license at: https://github.com/rcedgar/musclehttps://github.com/rcedgar/newickhttps://doi.org/10.5281/zenodo.7255768.

References

1. Sievers, F. & Higgins, D. G. Clustal omega. *Curr. Protoc. Bioinforma.* **48**, 3–13 (2014).
2. Katoh, K. & Standley, D. M. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evolution* **30**, 772–780 (2013).
3. Edgar, R. C. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
4. Thompson, J. D., Plewniak, F. & Poch, O. Balibase: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinforma. (Oxf., Engl.)* **15**, 87–88 (1999).
5. Gardner, P. P., Wilm, A. & Washietl, S. A benchmark of multiple sequence alignment programs upon structural rnas. *Nucleic Acids Res.* **33**, 2433–2439 (2005).
6. Feng, D.-F. & Doolittle, R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evolution* **25**, 351–360 (1987).
7. Lake, J. A. The order of sequence alignment can bias the selection of tree topology. *Mol. Biol. Evolution* **8**, 378–385 (1991).
8. Do, C. B., Mahabhashyam, M. S., Brudno, M. & Batzoglou, S. Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15**, 330–340 (2005).
9. Notredame, C., Higgins, D. G. & Heringa, J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
10. Holmes, I. & Durbin, R. Dynamic programming alignment accuracy. *J. Computational Biol.* **5**, 493–504 (1998).
11. Wolf, Y. I. et al. Origins and evolution of the global rna virome. *MBio* **9**, e02329–18 (2018).
12. Stamatakis, A. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
13. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).
14. Minh, B. Q. et al. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evolution* **37**, 1530–1534 (2020).
15. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490 (2010).
16. Tamura, K., Stecher, G. & Kumar, S. Mega11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evolution* **38**, 3022–3027 (2021).
17. ICTV et al. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.* **5**, 668 (2020).
18. Wróbel, B. Statistical measures of uncertainty for branches in phylogenetic trees inferred from molecular sequences by using model-based methods. *J. Appl. Genet.* **49**, 49–67 (2008).
19. Chang, J.-M. et al. Incorporating alignment uncertainty into felsenstein's phylogenetic bootstrap to improve its reliability. *Bioinformatics* **37**, 1506–1514 (2021).
20. Frias-Navarro, D., Pascual-Llobell, J., Pascual-Soler, M., Perezgonzalez, J. & Berrios-Riquelme, J. Replication crisis or an opportunity to improve scientific production? *Eur. J. Educ.* **55**, 618–631 (2020).
21. Wheeler, W. C. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. *Syst. Biol.* **44**, 321–331 (1995).
22. Wheeler, W. C., Gatesy, J. & DeSalle, R. Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. *Mol. Phylogenetics Evolution* **4**, 1–9 (1995).
23. Morrison, D. A. & Ellis, J. T. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18s rdnas of apicomplexa. *Mol. Biol. Evolution* **14**, 428–441 (1997).
24. Chatzou, M., Floden, E. W., Di Tommaso, P., Gascuel, O. & Notredame, C. Generalized bootstrap supports for phylogenetic analyses of protein sequences incorporating alignment uncertainty. *Syst. Biol.* **67**, 997–1009 (2018).
25. Sela, I., Ashkenazy, H., Katoh, K. & Pupko, T. Guidance2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* **43**, W7–W14 (2015).
26. Katoh, K. & Toh, H. Parttree: an algorithm to build an approximate tree from a large number of unaligned sequences. *Bioinformatics* **23**, 372–374 (2007).
27. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.* **7**, 539 (2011).
28. Pietrokovski, S., Henikoff, J. G. & Henikoff, S. The blocks database-a system for protein classification. *Nucleic Acids Res.* **24**, 197–200 (1996).
29. Babaian, A. & Edgar, R. C. Ribovirus classification by a polymerase barcode sequence. *PeerJ* **10**, e14055 https://doi.org/10.7717/peerj.14055.
30. te Velthuis, A. J. Common and unique features of viral rna-dependent polymerases. *Cell. Mol. life Sci.* **71**, 4403–4420 (2014).
31. De Groot, R. J. et al. Commentary: Middle east respiratory syndrome coronavirus (mers-cov): announcement of the coronavirus study group. *J. Virol.* **87**, 7790–7792 (2013).
32. Wang, L., Byrum, B. & Zhang, Y. Porcine coronavirus hku15 detected in 9 us states, 2014. *Emerg. Infect. Dis.* **20**, 1594 (2014).
33. Woo, P. C. et al. Discovery of seven novel mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. *J. Virol.* **86**, 3995–4008 (2012).

Acknowledgements

I am grateful to Alexandros Stamatakis and Artem Babaian for helpful discussions. Some of the computing resources used in this work were kindly provided by the University of British Columbia Community Health and Wellbeing Cloud Innovation Centre, powered by AWS.

Author contributions

R.C.E. designed the study, performed the experiments and wrote the manuscript.

Competing interests

The author declares no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-34630-w.

Correspondence and requests for materials should be addressed to Robert C. Edgar.

Peer review information *Nature Communications* thanks Heng Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.