doi: 10.1093/nargab/lqaa024

A minimum reporting standard for multiple sequence alignments

Thomas K.F. Wong[1,2,†], Subha Kalyaanamoorthy[1,3,†], Karen Meusemann[4,5,6], David K. Yeates[4], Bernhard Misof[5] and Lars S. Jermiin[1,2,7,8,*]

[1]Land & Water, CSIRO, Canberra, ACT 2601, Australia, [2]Research School of Biology, Australian National University, Canberra, ACT 2600, Australia, [3]Department of Chemistry, University of Waterloo, Waterloo, ON N2L 3G1, Canada, [4]Australian National Insect Collection, CSIRO National Research Collections Australia, Canberra, ACT 2601, Australia, [5]Zoologisches Forschungsmuseum Alexander Koenig, 53113 Bonn, Germany, [6]Evolutionsbiologie & Ökologie, Institut für Biologie I, Albert-Ludwigs-Universität Freiburg, 79085 Freiburg im Breisgau, Germany, [7]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland and [8]Earth Institute, University College Dublin, Belfield, Dublin 4 Ireland

Received January 25, 2020; Revised March 12, 2020; Editorial Decision March 28, 2020; Accepted March 30, 2020

ABSTRACT

Multiple sequence alignments (MSAs) play a pivotal role in studies of molecular sequence data, but nobody has developed a minimum reporting standard (MRS) to quantify the completeness of MSAs in terms of completely specified nucleotides or amino acids. We present an MRS that relies on four simple completeness metrics. The metrics are implemented in AliStat, a program developed to support the MRS. A survey of published MSAs illustrates the benefits and unprecedented transparency offered by the MRS.

INTRODUCTION

Multiple sequence alignments (MSAs) are widely used during annotation and comparison of molecular sequence data, allowing us to identify medically important substitutions (1), infer the evolution of species (2), detect lineage- and site-specific changes in the evolutionary processes (3) and engineer new enzymes (4). There is a wide range of computational tools for obtaining MSAs, and two of these (i.e. Clustal W (5) and Clustal X (6)) are now among the 100 most cited papers in science (7).

In addition to the completely specified nucleotides (i.e. A, C, G, T/U) or amino acids (i.e. A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y), MSAs may contain ambiguous characters (i.e. incompletely specified nucleotides or amino acids). Frequently, they also contain alignment gaps (i.e. '–') inserted between the nucleotides or amino acids of some of the sequences. Alignment gaps are inserted to maximize the homology of residues from different sequences (alignment gaps should only be used to improve alignment whereas N and X should only be used to signal missing data). A correct MSA is necessary for accurate genome annotation, phylogenetic inference and ancestral sequence reconstruction. However, deciding where to put the alignment gaps may be more art than science. This is because homology is defined as similarity due to historical relationships by descent (8). Most of these relationships belong to the unobservable distant past, so it is impossible to measure the accuracy of most MSAs inferred from real sequence data.

Without this ability, reporting the completeness of MSAs may be the best that can be achieved. So far, the only metric sometimes used is the *percent missing data* for a sequence (9) or an alignment (10), but neither is sufficiently transparent and informative. Recently, a guideline for systematic reporting of sequence alignments has been suggested (11), but it did not include completeness of MSAs—instead, it focused on quality indicators of alignment, but it did not define any of these or point to relevant literature. To rectify this, we developed a minimum reporting standard (MRS) for MSAs.

MATERIALS AND METHODS

Metrics for measuring completeness of MSAs

The MRS uses four metrics to quantify the completeness of different attributes of MSAs. Given an MSA with m sequences and n sites, we may compute four metrics: $\mathcal{C}_a = x_a/(m \times n)$, $\mathcal{C}_r = x_r/n$, $\mathcal{C}_c = x_c/m$ and $\mathcal{C}_{ij} = x_{ij}/n$, where x_a is the number of completely specified characters (12) in the MSA, x_r is the number of completely specified characters in the r-th sequence of the MSA, x_c is the number of completely specified characters in the c-th column of the MSA and x_{ij} is the number of homologous sites with completely specified characters in both sequences (i and j). In summary, $\mathcal{C}_a$, $\mathcal{C}_r$, $\mathcal{C}_c$ and $\mathcal{C}_{ij}$ measure the completeness of the

[*]To whom correspondence should be addressed. Email: lars.jermiin@anu.edu.au
[†]Joint first authors.

alignment, the r-th sequence, the c-th site, and the i-th and j-th sequences, respectively.

The first of these metrics ($\mathcal{C}_a$) is related to the *percent missing data* used previously, but it is also, as shown in Figure 1A, the least useful completeness metric considered here: alignments A and B differ greatly, but they have the same $\mathcal{C}_a$ value (i.e. 0.7). The $\mathcal{C}_r$, $\mathcal{C}_c$ and $\mathcal{C}_{ij}$ metrics, on the other hand, are able to detect these differences. For example, the $\mathcal{C}_r$ values range from 0.3 to 1.0 for alignment A and from 0.4 to 1.0 for alignment B, raising greater concern, from a sequence-centric perspective, about alignment A than about alignment B. If we were to omit any sequence from alignment A, then it would be sensible to omit the one with the smallest $\mathcal{C}_r$ value. The $\mathcal{C}_c$ values range from 0.2 to 1.0 for alignment A and from 0.5 to 0.8 for alignment B. Again, there is greater concern about alignment A than about alignment B (due to the lower $\mathcal{C}_c$ scores and the greater range of values). The $\mathcal{C}_{ij}$ values range from 0.3 to 1.0 for alignment A and from 0.0 to 0.9 for alignment B. There is cause for great concern if $\mathcal{C}_{ij} = 0.0$ is detected because it means that sequences i and j have no shared homologous sites with completely specified characters in both sequences. Evolutionary distances between such sequences cannot be estimated unless the MSA contains at least one other sequence that overlaps both i and j. When such a case occurs, the evolutionary distance between sequences i and j is *inferred by proxy*. Currently, the prevalence of this problem is unknown.

Figures 1B and 1C reveal the distributions of $\mathcal{C}_r$ and $\mathcal{C}_c$ for alignments A and B, offering additional insight into the alignments' completeness. Conveniently, the $\mathcal{C}_c$ scores may be used to selectively omit the least complete sites. This *masking of sites* in MSAs is popular in phylogenetics and many methods (13–21) are now available. Additional information can be obtained by analyzing heat maps generated from the $\mathcal{C}_{ij}$ values. Figure 1D shows the heat maps obtained from alignments A and B. The most obvious things to note are that in alignment A *Tagliatelle* stands out as being the least complete sequence whereas *Capellini* and *Spaghetti* share no homologous sites with completely specified nucleotides in both sequences in alignment B. Although this was easy to detect in Figure 1A, it will be more difficult to do if n and/or m were larger, as is typically the case in phylogenomic data.

The benefits offered by the new completeness metrics are clear, but embedding figures like those in Figure 1 in publications may be impractical. Alternatively, the essential details may be reported in a table (Table 1), or in one line (e.g. alignment B: $m = 10, n = 100, \mathcal{C}_a = 0.7, \mathcal{C}_r = [0.4, 1.0]$, $\mathcal{C}_c = [0.4, 0.8]$ and $\mathcal{C}_{ij} = [0.0, 0.9]$). The closer to 1.0 the four $\mathcal{C}$ scores are, the more complete an alignment is. If, on the other hand, the values are closer to 0.0 than to 1.0, users may consider masking some of the sequences and/or sites before starting a phylogenetic analysis of the data.

Given their potential to inform researchers across a wide range of scientific disciplines, we argue that m, n, $\mathcal{C}_a$, $\mathcal{C}_r$, $\mathcal{C}_c$ and $\mathcal{C}_{ij}$ should be combined into what we henceforth call an MRS for MSAs, and that publications that report all of these values be labeled *compliant with the MRS for MSAs*. To our knowledge, this has never been done beforehand, leading to widespread ignorance about the MSAs that are relied upon in ground-breaking biomedical research.

Table 1. Example of the MRS for the alignments in Figure 1A

Feature	Alignment A	Alignment B
Sequences	10	10
Sites	100	100
Alphabet	Nucleotides	Nucleotides
$\mathcal{C}_a$	0.7	0.7
$\mathcal{C}_r$ [min–max]	0.3–1.0	0.4–1.0
$\mathcal{C}_c$ [min–max]	0.1–1.0	0.4–0.8
$\mathcal{C}_{ij}$ [min–max]	0.3–1.0	0.0–0.9

AliStat: a program supporting the MRS for MSAs

To enable compliance with the MRS for MSAs, we developed AliStat, which is written in C++. To our knowledge, it is the first program to compute the four completeness scores presented above.

AliStat reads a text file with sequences of single nucleotides (i.e. a 4-state alphabet), di-nucleotides (i.e. a 16-state alphabet), codons (a 64-state alphabet) and amino acids (a 20-state alphabet), which are aligned and saved in the FASTA format. If the sequences comprise single nucleotides, then the characters may be 'lumped' to form six 3-state alphabets (i.e. CRT, AGY, ACK, GMT, AST and CGW) and seven 2-state alphabets (i.e. RY, KM, SW, AB, CD, GH and TV)—here R = A or G, Y = C or T, K = A or C, M = G or T, B = C or G or T, D = A or G or T, H = A or C or T, and V = A or C or G. If the 3- and 2-state alphabets are used, the letters R, Y, K, M, S, W, B, D, H and V are considered completely specified characters, unlike normal practice (12).

AliStat can be run in two modes: Brief mode or Full mode. Execution in brief mode is done using the following command:

```
alistat < infile > < data type > −b
```

and results in the following output format being printed to the terminal:

File name, #seqs, #sites, $\mathcal{C}_a$, max $\mathcal{C}_r$, min $\mathcal{C}_r$, max $\mathcal{C}_c$, min $\mathcal{C}_c$, max $\mathcal{C}_{ij}$, min $\mathcal{C}_{ij}$

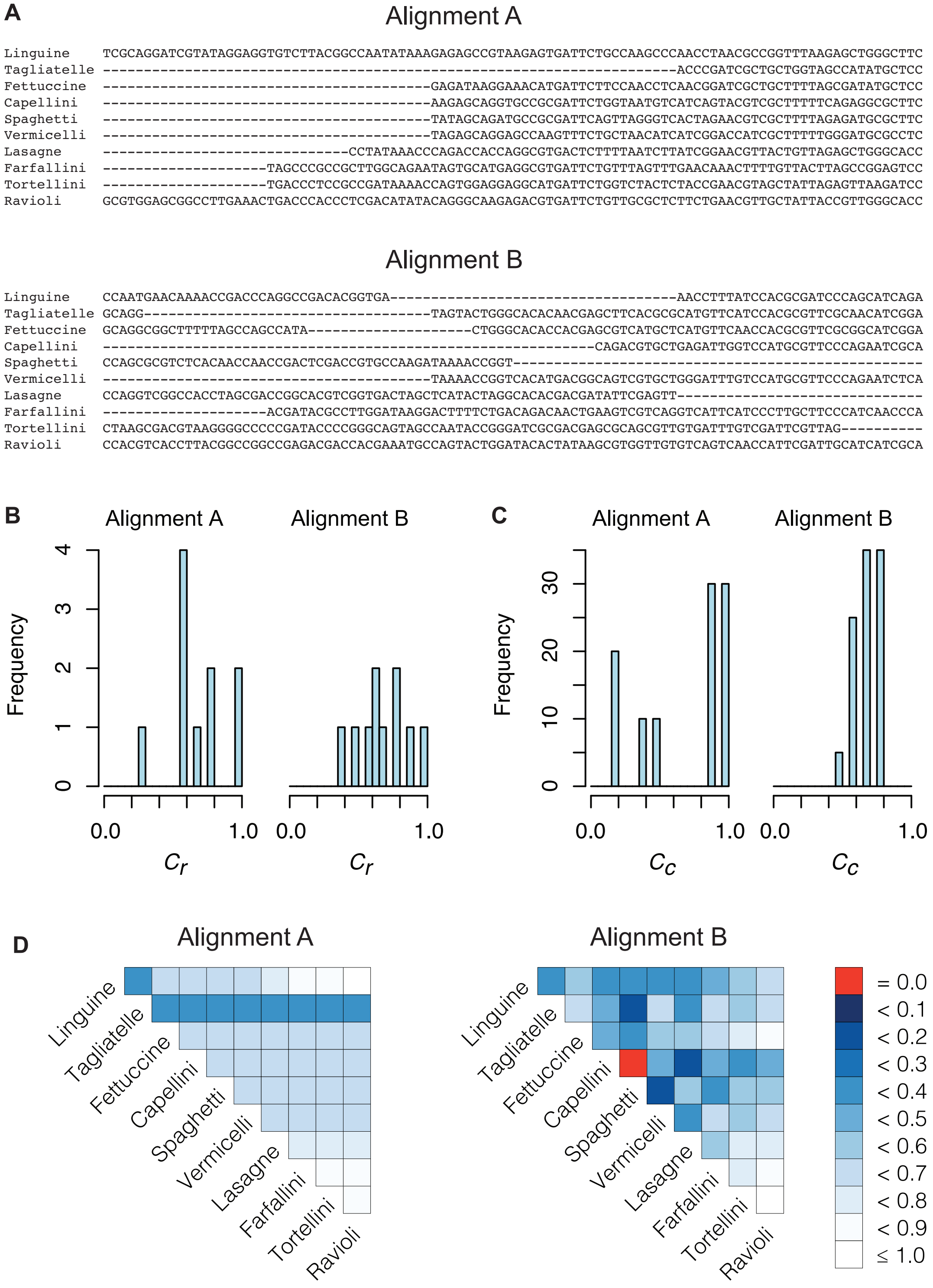
The brief-mode execution was included to allow users to quickly obtain the essential values from a great number of alignments (e.g. when comparing genomes phylogenetically).

The full-mode execution (default option) allows other options to be used and is intended when a more detailed examination of an MSA is required. For example, the –t option is used to indicate what types of $\mathcal{C}$ scores should be printed in output files, the –m option is used to set a threshold for masking sites and the –i option is used to indicate that a heat map is needed. Other options and how all of the options may be used are described in the AliStat manual. The same information can be obtained by typing

```
alistat − h
```

in the command-line.

The output files appear in the .txt, .csv, .R, .dis, .svg and .fst formats, which can be processed by other software packages. The .txt file summarizes the results. The .csv files present the $\mathcal{C}$ scores and may be examined using R. For ex-



Alignment A

```
Linguine    TCGCAGGATCGTATAGGAGGTGTCTTACGGCCAATATAAAGAGAGCCGTAAGAGTGATTCTGCCAAGCCCAACCTAACGCCGGTTTAAGAGCTGGGCTTC
Tagliatelle ----------------------------------------------------------------------ACCCGATCGCTGCTGGTAGCCATATGCTCC
Fettuccine  ----------------------------------------GAGATAAGGAAACATGATTCTTCCAACCTCAACGGATCGCTGCTTTTAGCGATATGCTCC
Capellini   ----------------------------------------AAGAGCAGGTGCCGCGATTCTGGTAATGTCATCAGTACGTCGCTTTTTCAGAGGCGCTTC
Spaghetti   ----------------------------------------TATAGCAGATGCCGCGATTCAGTTAGGGTCACTAGAACGTCGCTTTTAGAGATGCGCTTC
Vermicelli  ----------------------------------------TAGAGCAGGAGCCAAGTTTCTGCTAACATCATCGGACCATCGCTTTTTGGGATGCGCCTC
Lasagne     ------------------------------CCTATAAACCCAGACCACCAGGCGTGACTCTTTTAATCTTATCGGAACGTTACTGTTAGAGCTGGGCACC
Farfallini  --------------------TAGCCCGCCGCTTGGCAGAATAGTGCATGAGGCGTGATTCTGTTTAGTTTGAACAAACTTTTGTTACTTAGCCGGAGTCC
Tortellini  --------------------TGACCCTCCGCCGATAAAACCAGTGGAGGAGGCATGATTCTGGTCTACTCTACCGAACGTAGCTATTAGAGTTAAGATCC
Ravioli     GCGTGGAGCGGCCTTGAAACTGACCCACCCTCGACATATACAGGGCAAGAGACGTGATTCTGTTGCGCTCTTCTGAACGTTGCTATTACCGTTGGGCACC
```

Alignment B

```
Linguine    CCAATGAACAAAACCGACCCAGGCCGACACGGTGA-----------------------------------AACCTTTATCCACGCGATCCCAGCATCAGA
Tagliatelle GCAGG-----------------------------------TAGTACTGGGCACACAACGAGCTTCACGCGCATGTTCATCCACGCGTTCGCAACATCGGA
Fettuccine  GCAGGCGGCTTTTTAGCCAGCCATA--------------------CTGGGCACACCACGAGCGTCATGCTCATGTTCAACCACGCGTTCGCGGCATCGGA
Capellini   ------------------------------------------------------------CAGACGTGCTGAGATTGGTCCATGCGTTCCCAGAATCGCA
Spaghetti   CCAGCGCGTCTCACAACCAACCGACTCGACCGTGCCAAGATAAAACCGGT--------------------------------------------------
Vermicelli  ----------------------------------------TAAAACCGGTCACATGACGGCAGTCGTGCTGGGATTTGTCCATGCGTTCCCAGAATCTCA
Lasagne     CCAGGTCGGCCACCTAGCGACCGGCACGTCGGTGACTAGCTCATACTAGGCACACGACGATATTCGAGTT------------------------------
Farfallini  --------------------ACGATACGCCTTGGATAAGGACTTTTCTGACAGACAACTGAAGTCGTCAGGTCATTCATCCCTTGCTTCCCATCAACCCA
Tortellini  CTAAGCGACGTAAGGGGCCCCCGATACCCCGGGCAGTAGCCAATACCGGGATCGCGACGAGCGCAGCGTTGTGATTTGTCGATTCGTTAG----------
Ravioli     CCACGTCACCTTACGGCCGGCCGAGACGACCACGAAATGCCAGTACTGGATACACTATAAGCGTGGTTGTGTCAGTCAACCATTCGATTGCATCATCGCA
```

Figure 1. Example, based on two multiple sequences alignments (**A**), illustrating the corresponding distributions of completeness scores for rows (**B**), columns (**C**) and pairs of sequences (**D**).

Table 2. Example of the MRS for two published MSAs

Feature	Carboxyl/colineesterase (23)	Lepidoptera (24)
Sequences	364	203
Sites	2645	749 791
Alphabet	Amino acids	Amino acids
$\mathcal{C}_a$	0.2262	0.6422
$\mathcal{C}_r$ [min–max]	0.0106–0.5550	0.0609–0.9738
$\mathcal{C}_c$ [min–max]	0.0027–0.9972	0.0000–0.9655
$\mathcal{C}_{ij}$ [min–max]	0.0000–0.5550	0.0084–0.9672

ample, if a user wishes to generate a histogram of the $\mathcal{C}_c$ scores, the Table_2.csv file may be analyzed using the Histogram_Cr.R file. In some cases, users may want to infer a tree or network based on the $\mathcal{C}_{ij}$ score (or the $\mathcal{I}_{ij}$ score, where $\mathcal{I}_{ij} = 1.0 - \mathcal{C}_{ij}$). In such cases, .dis files may be analyzed by, for example, SplitsTree (22). The heat map, which may be triangular or square, is stored in the .svg file and may be opened using Adobe Illustrator™. If the –m option is used, the original MSA is split into two, with all sites having a $\mathcal{C}_c$ score larger than a user-specified threshold saved in a file called Mask.fst and the other sites saved in a file called Disc.fst. The two .fst files may be analyzed separately by other means (e.g. phylogenetic programs).

RESULTS AND DISCUSSON

The MRS may help to identify dubious MSAs. These alignments occur regularly in biomedical research and may also be present in large phylogenomic research, due to problems that might have arisen during the assembly, orthology assignment and alignment procedures.

Typically, MSAs comprise more sequences and sites than those in Figure 1A, so to facilitate using the MRS, we implemented AliStat, a fast, flexible and user-friendly program for surveying MSAs. AliStat computes the $\mathcal{C}_a$, $\mathcal{C}_r$, $\mathcal{C}_c$ and $\mathcal{C}_{ij}$ values from MSAs of nucleotides, di-nucleotides, codons and amino acids., AliStat lists the results on the command-line or in files that can be accessed by other programs.

The benefit of the MRS for MSAs is underlined in two surveys of large MSAs (Table 2). In the first case, surveying an MSA of the enzyme carboxyl/cholinesterase (23) revealed that some of the $\mathcal{C}_r$ and $\mathcal{C}_c$ scores are closer to 0.0 than 1.0, and that at least two sequences have no homologous sites in common with completely specified characters in both sequences. Further inspection of the output files revealed large proportions of low $\mathcal{C}_r$, $\mathcal{C}_c$ and $\mathcal{C}_{ij}$ scores (Supplementary Figures, S1–3), so it might be wise to mask some of the sequences or sites before phylogenetic analysis of these data. Given the main objective of the original analysis of these data (to annotate the genes in two major crop pests), masking sites with completeness scores below $\mathcal{C}_c = 0.5$ had a big impact on the $\mathcal{C}_a$ score (it increased from 0.2262 to 0.9562) and, hence, also on the maximum scores of $\mathcal{C}_r$ and $\mathcal{C}_{ij}$ (Supplementary Table S1).

In the second case, surveying a massive concatenation of MSAs of nuclear genes (24) revealed a more complete alignment but also low $\mathcal{C}_r$, $\mathcal{C}_c$ and $\mathcal{C}_{ij}$ values. The presence of these values shows that additional masking of this MSA might have been wise (Supplementary Figures S4–6). For example, omitting the two most incomplete sequences (i.e. the genera *Leucoptera* and *Pseudopostega*) could have been considered (Supplementary Figures S4 and 6).

The MRS for MSAs is a robust and sensible solution to a large and so-far-neglected problem: how do we report, as transparently and informatively as possible, the completeness of the MSAs used in biomedical research? Better transparency about the completeness of MSAs is clearly needed, because MSAs represent a foundational cornerstone in many biomedical research projects and, as revealed by the example in Figure 1, MSAs may look different but have the same percentage of missing data. So far, information on the completeness of MSAs used in biomedical research has been largely absent, leaving readers unable to critically evaluate the merits of scientific discoveries made on the basis of MSAs. It is critical to recognize, and acknowledge, that many MSAs are the result of scientific procedures. Therefore, it is necessary to present the results of these procedures more transparently and comprehensively. Many scientific papers now include links to the MSAs used, but the MSAs are often so large that it is impossible to form a comprehensive picture about the completeness of these MSAs.

Our MRS enables a radical change in scientific behavior, allowing authors to report their results more transparently and readers the ability to critically assess discoveries made from analyses of sequence data stored in MSAs.

DATA AVAILABILITY

AliStat is available from http://github.com/thomaskf/AliStat/ under an CSIRO Open Source Software License Agreement (variation of the BSD / MIT License).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank staff at the Australian National University and University College Dublin for feedback on the color scheme used in the heat map; many of the respondents are color-blind. At last, we wish to thank three reviewers for their constructive comments.

FUNDING

CSIRO.
Conflict of interest statement. None declared.

REFERENCES

1. Higgs,D.R. and Wood,W.G. (2008) Genetic complexity in sickle cell disease. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 11595–11596.
2. Misof,B., Liu,S.L., Meusemann,K., Peters,R.S., Donath,A., Mayer,C., Frandsen,P.B., Ware,J., Flouri,T., Beutel,R.G. *et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science*, **346**, 763–767.
3. Jayaswal,V., Wong,T.K.F., Robinson,J., Poladian,L. and Jermiin,L.S. (2014) Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Syst. Biol.*, **63**, 726–742.

4. Wilding,M., Peat,T.S., Kalyaanamoorthy,S., Newman,J., Scott,C. and Jermiin,L.S. (2017) Reverse engineering: transaminase biocatalyst development using ancestral sequence reconstruction. *Green Chem.*, **19**, 5375–5380.
5. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
6. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
7. Van Noorden,R., Maher,B. and Nuzzo,R. (2014) The top 100 papers. *Nature*, **514**, 550–553.
8. Morrison,D.A. (2015) Is sequence alignment an art or a science? *Syst. Bot.*, **40**, 14–26.
9. Wiens,J.J. (2003) Missing data, incomplete taxa, and phylogenetic accuracy. *Syst. Biol.*, **52**, 528–538.
10. Driskell,A.C., Ane,C., Burleigh,J.G., McMahon,M.M., O'Meara,B.C. and Sanderson,M.J. (2004) Prospects for building the tree of life from large sequence databases. *Science*, **306**, 1172–1174.
11. Vihinen,M. (2020) Guidelines for systematic reporting of sequence alignments. *Biol. Methods Protoc.*, **5**, 1–3.
12. Cornish-Bowden,A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.*, **13**, 3021–3030.
13. Castresana,J. (2000) Selection of conservative blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
14. Talavera,G. and Castresana,J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, **56**, 564–577.
15. Dress,A.W.M., Flamm,C., Fritzsch,G., Grunewald,S., Kruspe,M., Prohaska,S.J. and Stadler,P.F. (2008) Noisy: identification of problematic columns in multiple sequence alignments. *Algorith. Mol. Biol.*, **3**, 7.
16. Hartmann,S. and Vision,T.J. (2008) Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evol. Biol.*, **8**, 95.
17. Capella-Gutierrez,S., Silla-Martinez,J.M. and Gabaldon,T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
18. Misof,B. and Misof,K. (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst. Biol.*, **58**, 21–34.
19. Kück,P., Meusemann,K., Dambach,J., Thormann,B., von Reumont,B.M., Wägele,J.W. and Misof,B. (2010) Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front. Zool.*, **7**, 10.
20. Criscuolo,A. and Gribaldo,S. (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.*, **10**, 210.
21. Wu,M.T., Chatterji,S. and Eisen,J.A. (2012) Accounting for alignment uncertainty in phylogenomics. *PLoS One*, **7**, e30288.
22. Huson,D.H. and Bryant,D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, **23**, 254–267.
23. Pearce,S.L., Clarke,D.F., East,P.D., Elfekih,S., Gordon,K.H.J., Jermiin,L.S., McGaughran,A., Oakeshott,J.G., Papanikolaou,A., Perera,O.P. *et al.* (2017) Genomic innovations, transcriptional plasticity and gene loss underlying the evolution and divergence of two highly polyphagous and invasive Helicoverpa pest species. *BMC Biol.*, **15**, 63.
24. Kawahara,A.Y., Plotkin,D., Espeland,M., Meusemann,K., Toussaint,E.F.A., Donath,A., Gimnich,F., Frandsen,P.B., Zwick,A., dos Reis,M. *et al.* (2019) Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 22657–22663.