Peer-reviewed version available at *Life* **2018**, *8*, 20; doi:10.3390/life8020020

Technical Note

phylotaR: An automated pipeline for retrieving orthologous DNA sequences from GenBank in R

Dominic J. Bennett [1,2*], Hannes Hettling [3], Daniele Silvestro [1,2], Alexander Zizka [1,2], Christine D. Bacon [1,2], Søren Faurby [1,2], Rutger A. Vos [3] and Alexandre Antonelli [1,2,4,5]

[1] Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Gothenburg, Sweden
[2] Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, SE-405 30 Gothenburg, Sweden
[3] Naturalis Biodiversity Center, P.O. Box 9517, 2300 RA Leiden, the Netherlands
[4] Gothenburg Botanical Garden, Carl Skottsbergs gata 22A, SE-413 19, Gothenburg, Sweden
[5] Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford St., Cambridge, MA 02138 USA
* Correspondence: dominic.john.bennett@gmail.com

Abstract: The exceptional increase in molecular DNA sequence data in open repositories is mirrored by an ever-growing interest among evolutionary biologists to harvest and use those data for phylogenetic inference. Many quality issues, however, are known and the sheer amount and complexity of data available can pose considerable barriers to their usefulness. A key issue in this domain is the high frequency of sequence mislabelling encountered when searching for suitable sequences for phylogenetic analysis. These issues include the incorrect identification of sequenced species, non-standardised and ambiguous sequence annotation, and the inadvertent addition of paralogous sequences by users, among others. Taken together, these issues likely add considerable noise, error or bias to phylogenetic inference, a risk that is likely to increase with the size of phylogenies or the molecular datasets used to generate them. Here we present a software package, phylotaR, that bypasses the above issues by using instead an alignment search tool to identify orthologous sequences. Our package builds on the framework of its predecessor, PhyLoTa, by providing a modular pipeline for identifying overlapping sequence clusters using up-to-date GenBank data and providing new features, improvements and tools. We demonstrate our pipeline's effectiveness by presenting trees generated from phylotaR clusters for two large taxonomic clades: palms and primates. Given the versatility of this package, we hope that it will become a standard tool for any research aiming to use GenBank data for phylogenetic analysis.

Keywords: BLAST; DNA, open source; phylogenetics; R; sequence orthology.

1. Introduction

The first step in any nucleotide-based phylogenetic analysis is the identification of sequence homology. Without establishing homology, much like comparing apples and oranges, multiple sequence alignment is meaningless. More precisely, in the context of species trees, sequences chosen for phylogenetic analysis must represent the same ancestral region resulting from a speciation event, i.e. they must be orthologous [1-2]. Sequence orthology is often determined through name-based searches via large sequence databases, most commonly GenBank [3]. This approach, however, can be problematic due to the possibility of sequences being mislabelled and differences in naming conventions. For example, gene names can differ between working groups (e.g. COI, CO1, COX1 and COXI); different sections of a gene or region may be deposited under the same sequence name [4]; and deposited sequences may represent multiple genes in what are termed chimeric sequences [5]. In the best case scenario, these issues may lead to the failure to identify all relevant orthologous sequences. Worst case, one or more of the downloaded sequences will represent different ancestral

Peer-reviewed version available at *Life* **2018**, *8*, 20; doi:10.3390/life8020020

regions, causing poor alignment and/or incorrect inference of phylogenetic trees. Without resolving the problem of orthology in a programtic fashion, any large-scale attempt at self-updating, automated pipelines and initiatives for constructing phylogenies, e.g. [6-7], are bound to fail.

In an early attempt to address these issues, Sanderson et al. [4] developed a pipeline, PhyLoTa, that uses the Basic Local Alignment Search Tool (BLAST [8]) to identify orthologous sequences without the need for gene name matching. For a given taxonomic group, PhyLoTa searches through available sequences on GenBank and identifies orthologous sequence clusters. Users are then able to survey the clusters via a web-interface [9] and select the ones that best suit their phylogenetic analysis needs, e.g. by selecting the clusters that maximise coverage of their taxonomic groups of interest. A downside of PhyLoTA is that the searching and clustering is performed via all-versus-all BLASTing, the combinatorics of which become prohibitive above a certain taxonomic level – an ever increasing barrier as public sequence databases grow. One solution is to perform the BLASTingwithin taxonomic groups, leading to potentially shared clusters among taxonomic groups remaining undiscovered by PhyLoTa.

More importantly, however, the current PhyLoTa release is outdated as it was built on a GenBank release (representing 162,886,727 sequences, Release 194 [10]) in February 2013. Since then over 44 million new sequences have been deposited in GenBank (Release 224,representing 207,040,555 sequences, [10]). Additionally, NCBI's taxonomic database has been updated as new sequences are added, species are discovered and groupings are revised. Between March 2013 and February 2018, 170,000 new nodes of the database's taxonomic tree were added [11], representing 30% of all current nodes. Clearly, more frequent and regular updates to the pipeline are needed for phylogeneticists to make use of newly acquired and improved data.

To date, there have been just two alternatives for those who wish to discover orthologous sequences from GenBank: rely on error-prone gene names, or make do with outdated information. Here we present phylotaR, an R package comprising a pipeline for identifying and retrieving orthologous sequence clusters directly from the latest GenBank release. Our pipeline recreates the original PhyLoTa output for specific clades of interest to the user in a series of independent stages. Additionally, a user has the option of a secondary cluster stage (cluster^2) to identify and merge clusters at higher taxonomic levels than is available with PhyLoTa. We demonstrate the capacity of phylotaR by generating phylogenetic trees for two model clades, widely studied in evolutionary biology: palms and primates.

2. Implementation

2.1. The pipeline

The phylotaR pipeline consists of four automated, independent stages: **taxise** (identify all descendant taxonomic nodes), **download** (hierarchically retrieve all sequences from across the taxonomic tree), **cluster** (identify clusters from the downloaded sequences within nodes) and **cluster^2** (merge clusters identified within separate taxonomic nodes to identify clusters at higher taxonomic levels) (see Figure S1 for a conceptual outline). At a minimum, all a user needs to do is provide a taxonomic identity (name or NCBI ID at any taxonomic level), for which they would like to generate sequence clusters, and then run the phylotaR pipeline. The pipeline mimics the original PhyLoTa but with the following improvements: i) it makes use of sequence feature information to break up large sequences which may have otherwise been discarded for being too long, ii) it can generate paraphyletic clusters from nodes which are too small in themselves and iii) it has the additional stage for matching sister clusters, cluster^2, which makes our method scalable to larger groups of taxa with many sequences available. For more details on the pipeline see Figure 1 for an outline of the process, refer to appendix A for a detailed description of each stage and Table S1 for a description of all the parameters.

After the phylotaR pipeline stages have completed, the user can interrogate the identified clusters using a series of functions supplied within the phylotaR package. For example, a user can

Peer-reviewed version available at *Life* **2018**, *8*, 20; doi:10.3390/life8020020

filter the sequences across the clusters to a given taxonomic rank, or select sequences with clusters using a range factors: sequence lengths, GC-ratios, sequence definitions, proportion of ambiguous nucleotides and/or maximum alignment density (MAD score, [4]). Additionally, plotting functions allow the user to see which taxa are covered by which clusters (for examples, see the Figures 3, S2 and S3). After exploring, modifying and/or manipulating the clusters, the user can export them in tabular format as per the PhyLoTa database schema or as sequence files in FASTA format [12], which can be readily aligned by different software.

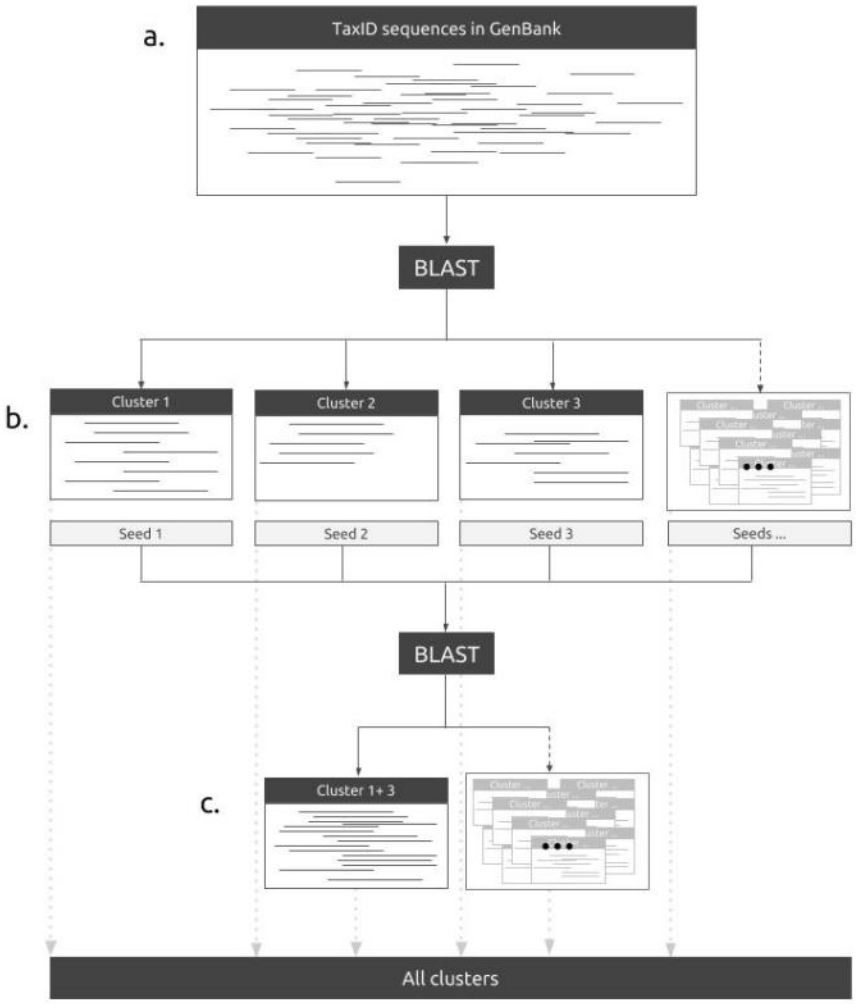

Figure 1. The phylotaR pipeline identifies all sequences in GenBank associated with a user-specified taxonomic identity (a.). The pipeline then performs all-vs-all BLAST across all the sequences to identify orthologous clusters (b.). These searches are constrained to run within taxonomic groups up to a user-determined limit (default 50,000 sequences and 100,000 nodes). To generate higher taxonomic level clusters, an additional BLAST search is performed of the most connected sequences within clusters (i.e. the seed sequences) from the lower-level clusters. The clusters of overlapping seed sequences are then merged into larger clusters (c.). All clusters, merged and unmerged, are then reported for inspection by the user. For more details on the pipeline, see appendix A.

2.2. Installation, features and future developments

The development version of the phylotaR package is currently available via GitHub (github.com/AntonelliLab/phylotaR) and can be installed through the R package devtools [13]. It will soon be available via CRAN [14]. The package depends on R (v 3+) and on a range of R packages [15-22], and requires stand-alone BLAST. Instructions for installing the BLAST+ suite can be found

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2018 doi:10.20944/preprints201804.0047.v1

Peer-reviewed version available at Life 2018, 8, 20; doi:10.3390/life8020020

via NCBI [23].

The entire pipeline can be run from an R console with just a few lines of code (Figure 2). The package comes with tools that enable a user to kill the pipeline process at any point. All downloaded and BLAST results are automatically cached, allowing a user to restart the pipeline after halting without loss of data. A log file records any user interventions and all pipeline progress, thus, increasing reproducibility of results and facilitating the methodological description in scientific publications. Finally, the code is developed to be modular, allowing users to contribute additional modules, functions and/or improvements. All internal functions and classes are documented to this end. To maximise the transparency and stability of the phylotaR package we have submitted the package to ROpenSci [24] – a community for ensuring good R coding practices in reproducible research – for which it is currently under review.

```
library(phylotaR)
wd <- 'PATH TO WORKING DIRECTORY'
ncbi_dr <- 'PATH TO NCBI BLAST TOOLS'
# NCBI txid, e.g. primates
txid <- 9443
# Set-up working directory
setUp(wd=wd, txid=txid, ncbi_dr=ncbi_dr)
# Run pipeline
run(wd=wd)
```

Figure 2. Initiating the phylotaR pipeline in R for primates (TaxID: 9443)

For future versions of phylotaR we envisage a range of additional features. For example, the ability to identify clusters across disparate taxonomic groups using the cluster² stage; allow users to incorporate their own unpublished sequences or specify their own taxonomy that is independent of GenBank's; allow a user to download FTP copies of large sections of GenBank to avoid slow-querying via Entrez [25]; and enable BLAST queries via NCBI's online BLAST server [26] and other servers. Users can request new features and highlight bugs via the phylotaR GitHub page (github.com/AntonelliLab/phylotaR).

3. Empirical Demonstration: Palms and Primates

Here we demonstrate the phylotaR pipeline on palms (Arecaceae, TaxID: 4710) and primates (Primates, TaxID: 9443) using default parameters, see Table S1. For palms and primates, respectively, we identified 449 and 1,021 clusters, each containing over ten unique taxonomic nodes. Taken together, these clusters included 1,238 and 653 unique taxonomic nodes and 13,344 and 56,112 sequences, respectively. See Figures S2 and S3 for visual representations of the relative distributions of sequences/taxa/clusters among the different clusters and taxa.

To illustrate the ability of the pipeline to correctly identify orthologous sequence clusters and to keep the demonstration fast, we generated small, representative trees for both palms and primates at the tribe and genus levels, respectively. We reduced the number of sequences within each cluster to just representatives within these taxonomic levels by filtering the sequences by proportion of ambiguous nucleotides and length of sequence. We then selected the 'best' clusters for both palms and primates from these reduced clusters, by dropping all clusters with sequence maximum alignment densities (MAD, [4]) less than 0.75 and then selecting the top ten with the greatest number of tribes/genera (see Tables S2a and S2b for detailed information on each of the selected clusters). These top clusters were found to be representative of palms and primates as a whole (Figure 3). With these clusters, we then generated alignments with MAFFT [27] and constructed phylogenetic trees with RAxML [28]. We found the generated trees to be similar to those published for these groups (Figures 4, S3 and S4) with few differences in topology. For further details on tree construction and comparison methods, see appendix B. The methods described above and in the appendix can be reproduced via scripts available from GitHub (github.com/AntonelliLab/phylotaR_demo).

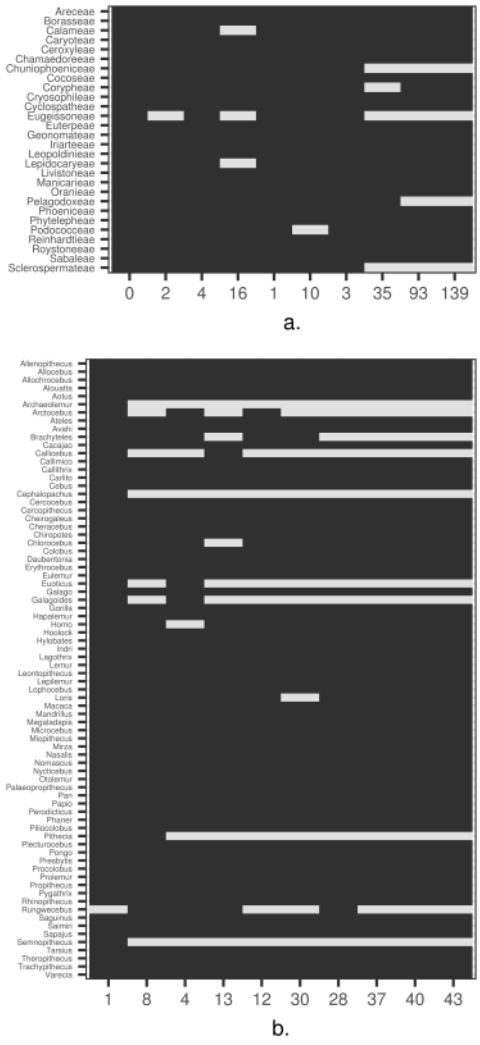

Figure 3. Presence/absence of tribes and genera for palms (**a**) and primates (**b**), respectively, across the top ten best clusters. X-axis numbers are unique cluster Ids. For more details on each of these clusters, see Tables S2a and S2b.

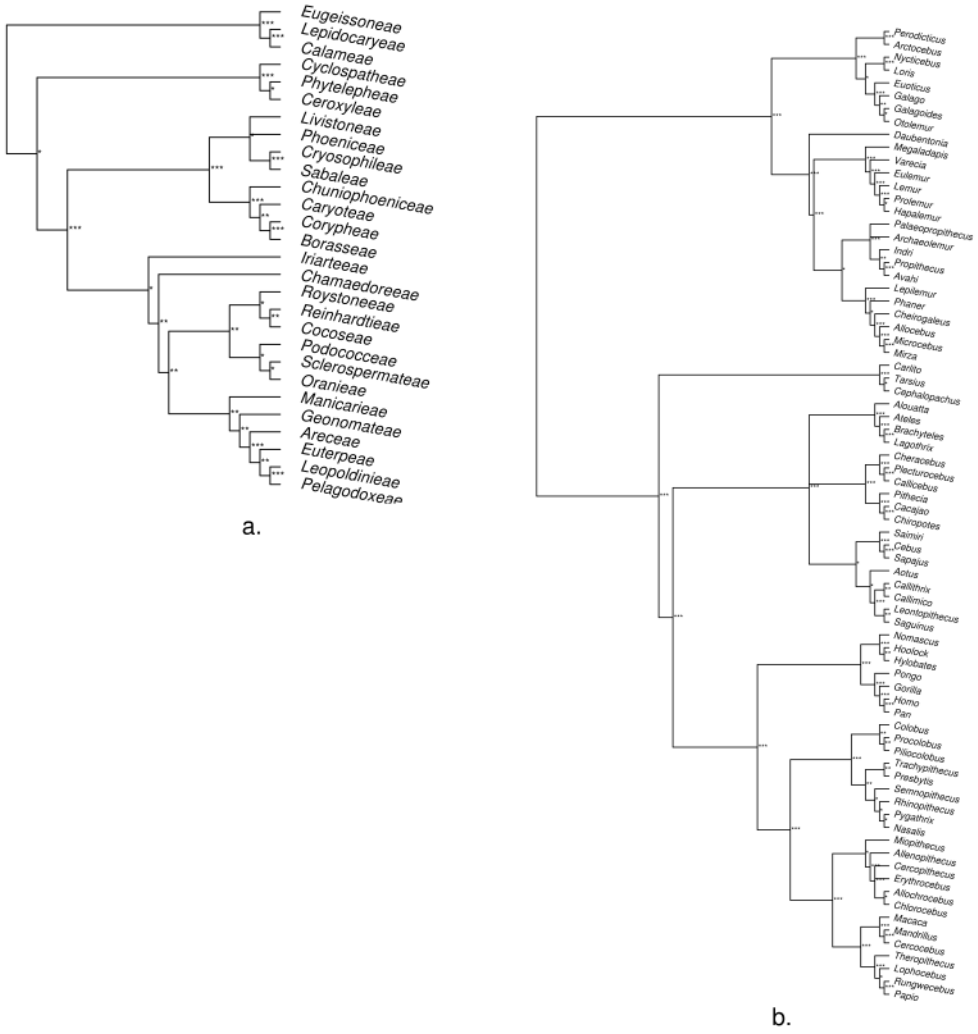

Figure 4. Tribe- and genus-level trees for palms (**a**) and primates (**b**). Roots were determined manually by rooting with Strepsirrhini and Calamoideae for primates and palms respectively. Branch lengths have been removed. Support calculated from 100 rapid bootstraps: *** > .95, ** > .75 and * > .50. Complete tree construction methods are in appendix B. For tree comparisons with published trees for palms [29] and primates [30], see Figures S4 and S5.

5. Conclusions

The phylotaR package offers a user-friendly pipeline to obtain orthologous gene-sequences for phylogenetic inferences from Genbank in R. Building on PhyLoTa, the pipeline makes use of sequence feature information, is able to generate small paraphyletic clusters to prevent dropping taxa and uses a subsequent cluster[2] to identify shared clusters across large taxonomic groups. The phylotaR pipeline is modular, can be easily integrated into R workflow pipelines that, for example, replicate the functionality of the SUPERSMART pipeline [6], and yields reliable results for phylogenies of large taxonomic groups. The package is currently available via GitHub (github.com/AntonelliLab/phylotaR) and comes with detailed vignettes containing documentation and tutorials.

Supplementary Materials: The following are available online at www.mdpi.com/link, Figure S1: conceptual outline of the phylotaR pipeline stages, Figure S2: relative number of sequences and clusters for each genus and tribe, Figure S3: relative number of sequences and taxa for each cluster. Figures S4 and S5:comparison between phylotaR-based trees to published trees, Table S1: default phylotaR pipeline parameters, Table S2(a)(b): details on top clusters used for phylogeny constructions for palms and primates, primates.tre: resulting primate Newick tree, palms.tre: resulting palms Newick tree.

Acknowledgments: We would like to thank Michael Sanderson for initiating the PhyLoTa project and providing early feedback and advice. We would also like to thank ROpenSci for their feedback and support regarding the development of the phylotaR package. In particular Scott Chamberlain for initial feedback and Zebulun Arendsee and Naupaka Zimmerman for taking the time to review our code in thorough detail. D.S. received funding from the Swedish Research Council (2015-04748).AA is supported by the Swedish Research Council (B0569601), the Swedish Foundation for Strategic Research, a Wallenberg Academy Fellowship, the Faculty of Sciences at the University of Gothenburg, the Wenner-Gren Foundations, and the David Rockefeller Center for Latin American Studies at Harvard University.

Author Contributions:HH, RAV, DS, AZ and AA initiated, devised and developed the project. HH wrote the initial pipeline code. DJB developed upon the pipeline, created the R package and wrote the manuscript. CDB and SF reviewed the palm and primate trees. All authors contributed to the writing of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Further pipeline details

The first stage, **taxise**, looks up taxonomic information on all the descendant nodes from NCBI taxonomy [31]. At the end of this stage a modifiable taxonomic dictionary is created containing names, IDs, lineages and ranks of all descendant IDs. The taxonomic dictionary contains a taxonomic tree (TreeMan class [20]), allowing fast querying of number of descendants and parent taxa. In addition, the taxonomic nodes for which clusters can be generated are identified by counting the number of possible sequences. Clades that contain too many sequences or too many descendants, as defined by the user, are broken down into their subclades and these subclades are analysed separately. A limit is required to prevent all-vs-all BLAST searches, at the clusters stage, becoming too large.

In the second stage, **download**, sequences are hierarchically downloaded for each node identified during the taxise stage. For 'model organisms' (sensu Sanderson et al. [4]: taxa for which there are large numbers of sequence data available) a random subset of the available sequences are downloaded. To prevent mega-clusters covering large numbers of different genes, all downloaded sequences are broken down into their constituent annotated features. This stage queries GenBank [3] through the rentrez package [16].

In the third stage, **cluster**, all-vs-all BLAST searches are performed within clades of user-determined size using all the downloaded sequences. Clusters are identified for all nodes of the taxonomy from sequences strictly associated with the node – direct – and all descendant sequences – subtree. BLAST searches are performed externally to R through NCBI's BLAST+ suite [23]. All sequences that have BLAST e-values and coverages, respectively, less than and greater than the

user-determined maximum E-value and minimum coverage are considered part of a single cluster. Entire clusters of sequences are then inferred from these BLAST results by identifying single-linkage clusters, as per the original PhyLoTa pipeline [4], with the iGraph package [15]. For large clades that have been broken down into many subclades for all-vs-all BLAST, an option in the phylotaR parameters allows this stage to be run in parallel. Because the pipeline clusters hierarchically within taxonomic clades, no clusters may be identified for small clades that are sister to very large clades due to too few sequences. To prevent taxa from being excluded, paraphyletic clusters are generated by non-hierarchically searching for clusters across all clades where no clusters were identified.

Because of the computational need to partition the cluster stage into subclades for very large clades, a final, fourth phylotaR stage, **clusters²**, is run to combine the subclade clusters into higher-level clade clusters. For every cluster identified in the previous stage a 'seed sequence' is determined as the sequence with the greatest number of BLAST hits with other sequences in its cluster. This is slightly different from Sanderson et al.'s [4] conception of a 'seed sequence', which was simply the starting point for single-linkage clustering. Because single-linkage clustering has a tendency to wander – leading to stretched out clusters where the starting point may be far off from the centre – we have opted to take the sequence with the highest connectivity instead. An all-vs-all BLAST search is performed with these seed sequences. All subclade clusters where a valid BLAST hit has occurred for their seed sequences are then merged into higher-level clusters.

At the end of the pipeline, a user will have identified clusters of the four different types described above: direct, subtree, paraphyletic and merged.

Appendix B

Further details on tree construction and assessment

For each of the clusters, the sequences were written out in FASTA format and alignments were constructed using mafft (v7.271, [27]) with its *--auto* argument. Supermatrices were then constructed using the ten sets of alignments. RAxML (v8.2.4, [28]) was then used to construct the trees. We used the GTRGAMMA model and partitioned the supermatrices based on the identified clusters. We ran a rapid bootstrap analysis of 100 iterations.

To monitor for errors at any point in the demonstration pipeline two sequences were selected instead of just one when reducing the clusters down to tribe/genus level. If in the resulting trees, large numbers of representative sequences of the same taxon were not uniquely clustered, data and methods were re-checked and the pipeline was run again. For palms and primates, sequences in the same tribe/genus were sisters in, respectively, 100% and 94% of cases.

From the bootstrapped trees we calculated majority-rule consensus trees. These final trees consisted of 28 and 80 tips for palms and primates, respectively representing 1.00 and 0.95 of all known tribes and genera according to NCBI taxonomy (Federhen, 2012). Both trees were very similar to published phylogenetic trees (Figures S3 and S4). We compared the final trees to already-published ones (palms [29] and primates [30]) using the Robinson-Foulds distance [32] and the triplet distance [33] through the R package treeman [20]. We found normalised Robinson-Foulds and triplet distances between our trees and the published trees, respectively, of 0.083 and 0.002 for palms and 0.189 and 0.016 for primates. For the primate tree, there were three key differences between our tree and those published in the literature: the paraphyletic separation of the family Lorisidae [(Perodicticus, Arctocebus), (Nyctiecbus, Loris)], although this has been suggested before [34]; a different branch ordering of the gibbons [35]; and the misplacement of the genus Semenopithecus which has been suggested to be grouped with Trachypithecus [36]. For the palms tree, the greatest problem was the misplacement of the subfamily Arecoideae (Iriarteae - Pelagodoxeae) with the Coryphoideae (Livistoneae - Borasseae) rather than the Ceroxyloideae (Cyclospatheae - Ceroxyleae), with which it is most often grouped [30, 37]. Additionally, there were some potential errors and inconsistencies regarding the taxonomy: there is no representative of the well-studied Nypa genus because there is no tribe name present in the NCBI taxonomy; also the current accepted name for the Livistoneae is Trachycarpeae [38].

References

1. de Pinna, M.C.C. Concepts and tests of homoloogy in the cladistics paradigm. *Cladistics*, **1991**, *7*(4), 367-394. DOI: 10.1111/j.1096-0031.1991.tb00045.x
2. Salemi, M.; Vandamme, A.-M.; Lemey, P. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press, Cambridge, UK, 2009.
3. Benson, D.A.; Karsch-Mizrachi, I.; Clark, K.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. Nucleic Acids Research, **2012**, *40*(Database issue), D48-53. DOI: 10.1093/nar/gkr1202
4. Sanderson, M.J.; Boss, D.; Chen, D.; Cranston, K.A.; Wehe, A. The PhyLoTA Browser: Processing GenBank for molecular phylogenetics research. *Systematic Biology*, **2008**, 57(3), 335-346. DOI:10.1080/10635150802158688
5. Ashelford, K.E.; Chuzhanova, N.A.; Fry, J.C.; Jones, A.J.; Weightman, A.J. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, **2005**, *71*(12), 7724-7736. DOI:10.1128/AEM.71.12.7724-7736.2005
6. Antonelli, A.; Hettling, H.; Condamine, F.L.; Vos, K.; Nilsson, R.H.; Sanderson, M.J.; Sauquet, H.; Scharn, R.; Silvestro, D.; Töpel, M.; Bacon, C.D.; Oxelman, B.; Vos, R.A. Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of Taxa. *Systematic Biology*, **2017**, *66*(2), 153-166. DOI: 10.1093/sysbio/syw066
7. Pearse, W.D.; Purvis, A. phyloGenerator: An automated phylogeny generation tool for ecologists. *Methods in Ecology and Evolution*, **2013**, *4*(7), 692-698. DOI: 10.1111/2041-210X.12055
8. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *Journal of Molecular Biology*, **1990**, 215(3), 403-410. DOI: 10.1016/S0022-2836(05)80360-2
9. PhyLoTa Browser. Available online: phylota.net (accessed on 28/03/18)
10. GenBank and WGS Statistics. Available online: www.ncbi.nlm.nih.gov/genbank/statistics (accessed on 28/03/18)
11. Taxonomy Browser: Taxonomy Statistics. Available online: www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html (accessed on 28/3/18)
12. Pearson, W.R.; Lipman, D.J. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, **1988**, *85*(8), 2444-2448. DOI: 10.1073/pnas.85.8.2444
13. Wickham, H.; Hester, J.; Chang W.; Rstudio and R Core team. devtools: Tools to make developing R packages easier. 2018. Available online: CRAN.R-project.org/package=devtools (accessed on 28/3/18)
14. The comprehensive R archive network. Available online: cran.r-project.org (accessed online 28/3/18)
15. Csardi G.; Nepusz T. The igraph software package for complex network research. *InterJournal, Complex Systems*. **2018**, *1695*.
16. Winter, D. rentrez: Entrez in R. R package version 1.1.0. 2017. Available online: CRAN.R-project.org/package=rentrez (accessed on 28/3/18)
17. Lang, D.T.; the CRAN Team. XML: Tools for parsing and generating XML within R and S-Plus. 2018. Available online: CRAN.R-project.org/package=XML (accessed on 28/3/18)
18. Wickham. H. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag, New York, USA, 2009.

19. Ooms, J. sys: Portable system utilities. Available online: CRAN.R-project.org/package=sys (accessed on 28/3/18)
20. Bennett, D.J.; Sutton, M.D.; Turvey, S.T. treeman: an R package for efficient and intuitive manipulation of phylogenetic trees. *BMC Research Notes*, **2017**, *10*(1), 30.
21. Wilkins D. treemapify: Draw treemaps in 'ggplot2'. Available online: CRAN.R-project.org/package=treemapify (accessed on 28/3/18)
22. Bengtsson, H. R.utils: Various Programming Utilities. Available online: CRAN.R-project.org/package=R.utils (accessed on 28/3/18)
23. BLAST® Command Line Applications User Manual. Available online: www.ncbi.nlm.nih.gov/books/NBK279690 (accessed on 28/3/18)
24. Transforming science through open data and software. Available online: ropensci.org (accessed on 28/3/18)
25. Entrez molecular sequence database system. Available online: www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html (accessed on 28/3/18)
26. Basic local alignment search tool. Available online: blast.ncbi.nlm.nih.gov/Blast.cgi (accessed on 28/3/18)
27. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **2013**, *30*(4), 772-780. DOI:10.1093/molbev/mst010
28. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **2014**, *30*(9), 1312-1313. DOI: 10.1093/bioinformatics/btu033
29. Baker, W.J.; Savolainen, V.; Asmussen-Lange, C.B.; Chase, M.W.; Dransfield, J.; Forest, F.; Harley, M.M.; Uhl, N.W.; Wilkinson, M. Complete generic-level phylogenetic analyses of palms (Arecaceae) with comparisons of supertree and supermatrix approaches. *Systematic Biology*, **2009**, *58*(2): 240-256. DOI: 10.1093/sysbio/syp021
30. Perelman, P.; Johnson, W.E.; Roos, C.; Seuánez, H.N.; Horvath, J.E.; Moreira, M.A.M.; Kessing, B.; Pontius, J.; Roelke, M.; Rumpler, Y.; Schneider, M.P.C.; Silva, A.; O'Brien, S.J.; Pecon-Slattery, J. A molecular phylogeny of living primates. *PLoS Genetics*, **2011**, 7(3), 1-17. DOI: 10.1371/journal.pgen.1001342
31. Federhen, S. The NCBI taxonomy database. *Nucleic Acids Research*, **2012**, *40*(Database issue), D136-43. DOI: 10.1093/nar/gkr1178
32. Robinson, D.F.; Foulds, L.R. Comparison of phylogenetic trees. *Mathematical Biosciences*, **1981**, *53*(1-2), 131-147. DOI: 10.1016/0025-5564(81)90043-2
33. Critchlow D.E.; Pearl D.K.; Qian C. The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology*, **1996**, *45*, 323-34.
34. Masters, J.C.; Anthony, N.M.; De Wit, M.J.; Mitchell, A. Reconstructing the evolutionary history of the Lorisidae using morphological, molecular, and geological data. *American Journal of Physical Anthropology*, **2005**, *127*(4), 465-480. DOI: 10.1002/ajpa.20149

35. Shi, C.M.; Yang, Z. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Molecular Biology and Evolution*, **2018**, *35*(1), 159-179. DOI: 10.1093/molbev/msx277

36. Osterholz, M.; Walter, L.; Roos, C. Phylogenetic position of the langur genera Semnopithecus and Trachypithecus among Asian colobines, and genus affiliations of their species groups. *BMC Evolutionary Biology*, **2008**, *8*(1), 1-12. DOI: 10.1186/1471-2148-8-58

37. Couvreur, T.L.P.; Forest, F.; Baker, W.J. Origin and global diversification patterns of tropical rain forests: Inferences from a complete genus-level phylogeny of palms. *BMC Biology*, **2011**, *9*. DOI: 10.1186/1741-7007-9-44

38. Dransfield, J.; Uhl, N.W.; Asmussen-Lange, C.B.; Baker, W.J.; Harley, M.M.; Lewis, C.E. A new phylogenetic classification of the palm family, Arecaceae. *Kew Bulletin*, **2005**, *60*(September), 559-569. DOI: 10.2307/25070242