life

MDPI

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; daniele.silvestro@bioenv.gu.se (D.S.); alexander.zizka@bioenv.gu.se (A.Z.); christinedbacon@gmail.com (C.D.B.); soren.faurby@bioenv.gu.se (S.F.); alexandre.antonelli@bioenv.gu.se (A.A.)
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; hannes.hettling@naturalis.nl (H.H.); Rutger.Vos@naturalis.nl (R.A.V.)
4 Gothenburg Botanical Garden, Carl Skottsbergsgata 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

Received: 28 March 2018; Accepted: 1 June 2018; Published: 5 June 2018

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 mislabeling encountered when searching for suitable sequences for phylogenetic analysis. These issues include, among others, the incorrect identification of sequenced species, non-standardized and ambiguous sequence annotation, and the inadvertent addition of paralogous sequences by users. 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 and test 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 mislabeled 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 regions, causing poor alignment and/or incorrect inference of phylogenetic trees. Without resolving the problem of orthology in a programmatic fashion, any large-scale attempt at self-updating, automated pipelines and initiatives for constructing phylogenies, e.g., [6,7], are bound to fail [8].

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 [9]) 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 [10] and select the ones that best suit their phylogenetic analysis needs, e.g., by selecting the clusters that maximize 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 BLASTing within 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 in February 2013 (representing 162,886,727 sequences, Release 194 [11]). Since then over 44 million new sequences have been deposited in GenBank (Release 224, representing 207,040,555 sequences, [11]). 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 [12], 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. More recently, new databases of orthologous sequences have become available [13,14]. These databases, however, are not based on GenBank but instead on assembled genomes—massively limiting their taxonomic coverage.

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 and test 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 (see Table 1 for a comparison of phylotaR and PhyLoTa). 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, Table S1 for a description of all the parameters and Table S2 for benchmarking of how different parameters impact run-times and results.

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 filter the sequences across the clusters to a given taxonomic rank, or select sequences with clusters using a range of 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 figures below). 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 [15], which can be readily aligned by different software.

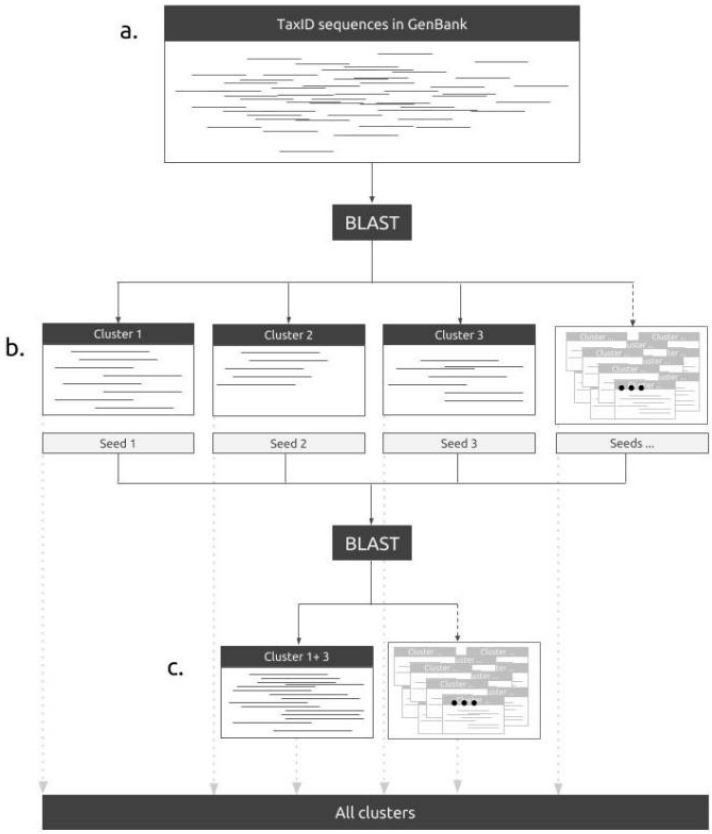

Figure 1. The phylotaR pipeline identifies all sequences in GenBank associated with a user-specified taxonomic identity (**a**). The pipeline then performs all-vs.-all BLAST across all the sequences to identify orthologous clusters (**b**). These searches are constrained to run within taxonomic groups 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 non-merged, are then reported for inspection by the user. For more details on the pipeline, see Appendix A.

Table 1. Comparing phylotaR and PhyLoTa.

	phylotaR	**PhyLoTa**
Features		
Direct clades	Yes	Yes
Subtree clades	Yes	Yes
Paraphyletic clades	Yes	No
Merged clades	Yes	No
Outputs	Clusters	Clusters, alignments, trees
Implementation		
Language	R	Perl
Open source	Yes	No
Execution	Local computer	Web-interface
Modular design	Yes	No
Resources		
GenBank release	Latest	2013
Search-tool	BLAST, user-choice *	BLAST
Taxonomy	NCBI, user-choice *	NCBI
Sequence features	Yes	No
Non-NCBI sequences	Yes *	No

* Yet to be implemented features.

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 [16]. It will soon be available via CRAN [17] and later Bioconda [18]. The package depends on R (v 3+) and on a range of R packages [19–26], and requires stand-alone BLAST. Instructions for installing the BLAST+ suite can be found via NCBI [27].

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 halt 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 maximize the transparency and stability of the phylotaR package we have submitted the package to ROpenSci [28]—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 cluster^2 stage; the incorporation of a user's own unpublished sequences or the specification of their own taxonomy that are independent of NCBI; the incorporation of alternative search tools other than BLAST that may be faster or sequence specific (e.g., DIAMMOND [29], usearch [30], BLAT [31], PLAST [32]); a download feature that would allow a

user to make use of FTP copies of large sections of GenBank to avoid slow-querying via Entrez [33]; and the ability to send BLAST queries via NCBI's online BLAST [34] 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 and test the phylotaR pipeline's ability to identify orthologous sequences by constructing phylogenetic trees from clusters generated for palms (*Arecaceae*, TaxID: 4710) and primates (*Primates*, TaxID: 9443) using default parameters, see Table S1. For palms and primates, respectively, we identified 449 and 1021 clusters, each containing over ten unique taxonomic identifiers. Taken together, these clusters included 1238 and 653 unique taxonomic identifiers 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.) The runtimes for palms and primates took in total 1.03 and 4.11 h, respectively. The download stage took by far the longest to complete, taking 0.89 and 3.47 h for palms and primates, respectively, representing 86% and 84% of total runtime.

Because phylotaR makes use of a more recent GenBank release than PhyLoTa, more sequences, taxa and clusters can be identified by phylotaR than PhyLoTa. We can compare the phylotaR results presented here to the equivalent results generated by PhyLoTa. According to the PhyLoTa browser [10] for palms there are 98 'phylogenetically informative clusters'—i.e., clusters with more than four unique taxonomic identifiers. In the phylotaR results, there are 1011 phylogenetically informative clusters. Comparing the clusters with the most taxonomic identifiers between the two methodologies, the PhyLoTa browser identifies 648 taxonomic identifiers, 926 sequences and 164 genera. The top cluster in the phylotaR results for palms identifies 720 taxonomic identifiers, 1346 sequences and 160 genera. The differences are even greater for the primates. 3727 clusters with more than four unique taxa were identified by phylotaR, whereas the equivalent figure for PhyLoTa is 1103. Comparing the clusters with the most taxonomic identifiers, PhyLoTa has 129 taxa, 543 sequences and 49 genera while phylotaR has respectively 491, 4871 and 75. (See, respectively, Tables S2 and S3 for comparisons of clusters identified for palms and primates by PhyLoTa [S3a and S4a] and phylotaR [S3b and S4b].).

Despite these differences in numbers, many of the same sequences identified by PhyLoTa were identified by phylotaR. For example, we were able to identify equivalent clusters between PhyLoTa and phylotaR by matching sequence identifiers. We found that across the clusters for palms identified by PhyLoTa, a mean 92.02% of the sequences in any of the clusters could be matched to sequences in one of the clusters identified by phylotaR. This figure was much smaller for primates. After controlling for differences in taxonomy, sequence length, and model organisms (for which sequences are selected randomly), only 55.12% of PhyLoTa sequences could be matched to sequences in phylotaR clusters. These 'missing' sequences in the phylotaR clusters were found to have either new sequence identifiers (10%), to be clustered in clusters with fewer than four sequences (30%) or to have not been in any clusters (60%).

To illustrate the ability of the pipeline to correctly identify orthologous sequence clusters while keeping 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 S5a and S5b 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).

For each of the clusters, the sequences were written out in FASTA format and alignments were constructed using MAFFT (v7.271, [35]) with its *–auto* argument. Supermatrices were then constructed using the ten sets of alignments. RAxML (v8.2.4, [36]) 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.

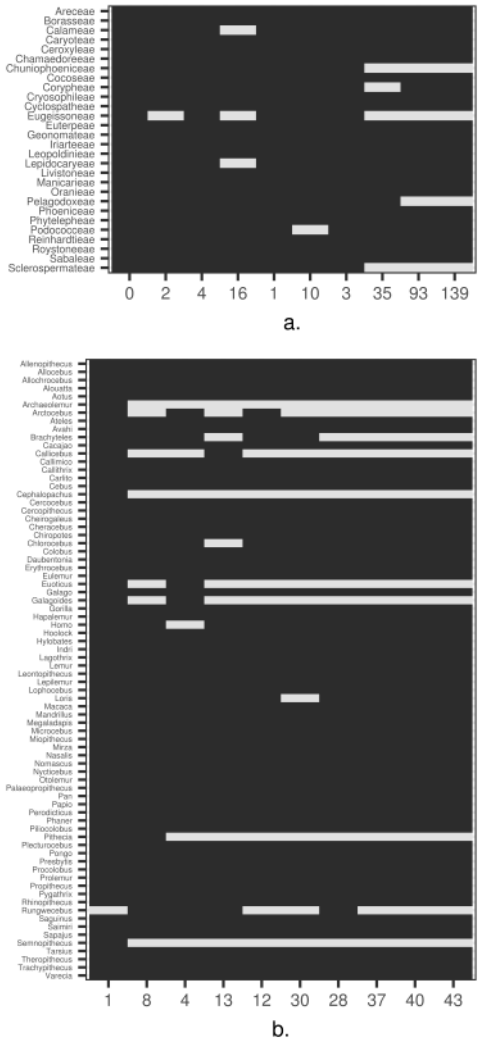

Figure 3. Presence/absence of tribes and genera for palms (**a**) and primates (**b**), respectively, across the top ten best clusters. X-axis numbers are unique cluster Ids. For more details on each of these clusters, see Tables S4a and S4b.

We found the generated trees to be similar to those published for these groups (Figure 4, Figures S3 and S4) with few differences in topology (for a detailed assessment of the trees, see Appendix B). Although it is likely that our post-phylotaR pipeline for constructing the trees could be improved to account for these topological differences, by successfully recreating the overall phylogenetic relationships of these groups we have demonstrated that phylotaR is able to successfully identify orthologous sequences. Additionally, we would expect no major topological differences to equivalent results generated from PhyLoTa data as, respectively for palms and primates, a mean 99.1% and 99.9% of sequences across the best-matching PhyLoTa clusters were found in the phylotaR clusters used for generating the trees.

The methods described above and in the appendix can be reproduced via scripts available from GitHub (github.com/AntonelliLab/phylotaR_demo).

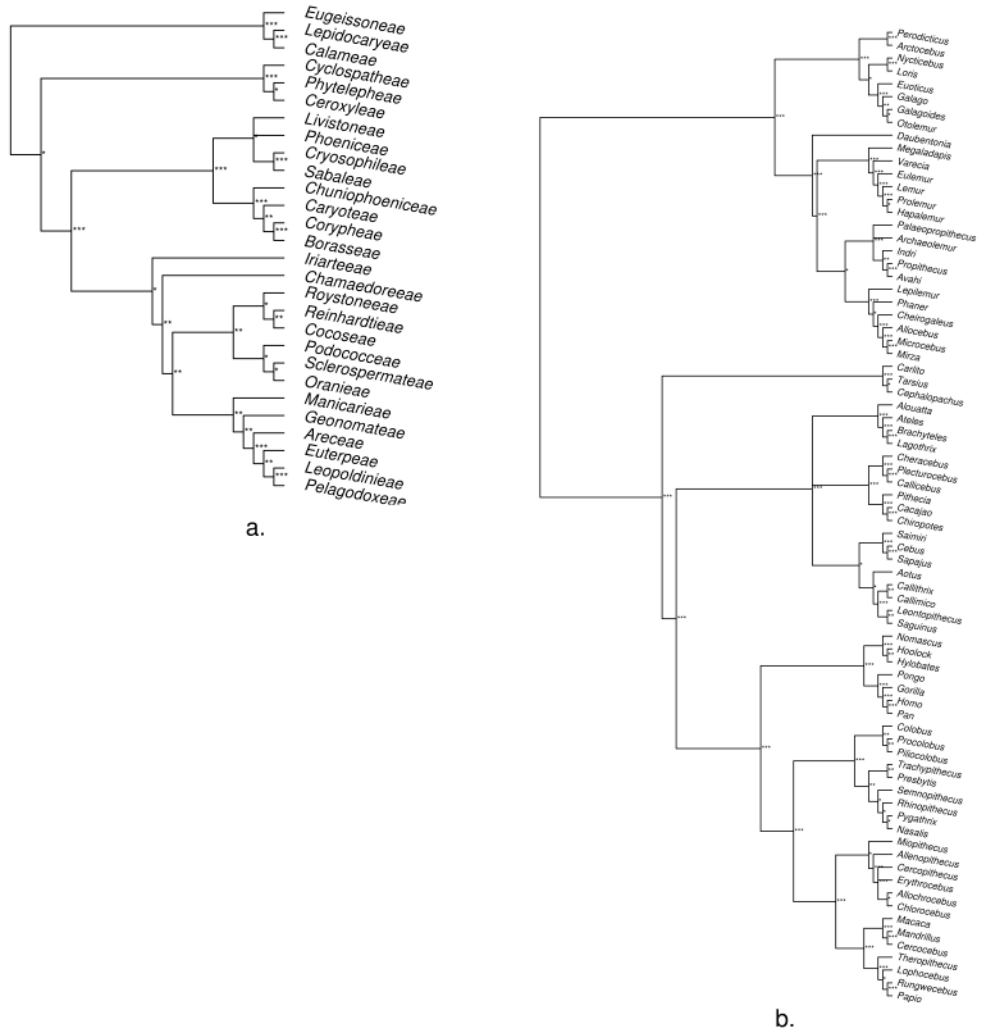

Figure 4. Tribe- and genus-level trees for palms (**a**) and primates (**b**). Roots were determined manually by rooting with *Strepsirrhini* and *Calamoideae* for primates and palms respectively. Branch lengths have been removed. Support calculated from 100 rapid bootstraps: *** >0.95, ** >0.75 and * >0.50. Complete tree construction methods are in Appendix B. For tree comparisons with published trees for palms [37] and primates [38], see Figures S4 and S5.

4. Conclusions

The phylotaR package offers a user-friendly pipeline to obtain orthologous sequences for phylogenetic inferences from GenBank in R. The phylotaR pipeline extends upon the PhyLoTa framework by, in addition, making use of sequence feature information, generating small paraphyletic clusters to prevent dropping taxa and implementing a subsequent cluster^2 to identify shared clusters across large taxonomic groups. The phylotaR pipeline, as shown here, can be used to yield reliable results for phylogenies of large taxonomic groups. The pipeline is modular and can be easily integrated into R workflows. We envisage phylotaR to be an important first step and module as part of an ecosystem of current and future automated, phylogeny-generating platforms [8]. 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 http://www.mdpi.com/2075-1729/8/2/20/s1, 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: benchmarking results for phylotaR using different parameters, Table S3a,b: information on all phylogenetically informative clusters identified for palms by PhyLoTa and phylotaR. Table S4a,b: information on all phylogenetically informative clusters identified for primates by PhyLoTa and phylotaR, Table S5a,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.

Author Contributions: H.H., R.A.V., D.S., A.Z. and A.A. initiated, devised and developed the project. H.H. wrote the initial pipeline code. D.J.B. developed upon the pipeline, created the R package and wrote the manuscript. C.D.B. and S.F. reviewed the palm and primate trees. All authors contributed to the writing of the manuscript.

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). A.A. 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. Additionally, we would like to thank three anonymous reviewers whose insightful comments and suggestions have greatly improved this article.

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 [39]. 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 [24]), 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 [20].

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 [27]. 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 [19]. 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, **clusters^2**, 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. Methods and Results for Tree Assessment

From the bootstrapped trees generated for the palms and primates 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 [39]. Both trees were very similar to published phylogenetic trees (Figures S3 and S4). We compared the final trees to already-published ones (palms [37] and primates [38]) using the Robinson-Foulds distance [40] and the triplet distance [41] through the R package treeman [24]. 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. The higher values for the primate tree can in part be attributed to the greater number of tips and the attempt to resolve phylogenetic relationships at lower taxonomic levels. 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 [42]; a different branch ordering of the gibbons [43]; and the misplacement of the genus *Semenopithecus* which has been suggested to be grouped with *Trachypithecus* [44]. 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 [37,45]. 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* [46].

References

1. De Pinna, M.C.C. Concepts and tests of homoloogy in the cladistics paradigm. *Cladistics* **1991**, *7*, 367–394. [CrossRef]
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 Res.* **2012**, *40*, D48–D53. [CrossRef] [PubMed]
4. Sanderson, M.J.; Boss, D.; Chen, D.; Cranston, K.A.; Wehe, A. The PhyLoTA Browser: Processing GenBank for molecular phylogenetics research. *Syst. Biol.* **2008**, *57*, 335–346. [CrossRef] [PubMed]
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. *Appl. Environ. Microbiol.* **2005**, *71*, 7724–7736. [CrossRef] [PubMed]
6. Antonelli, A.; Hettling, H.; Condamine, F.L.; Vos, K.; Nilsson, R.H.; Sanderson, M.J.; Sauquet, H.; Scharn, R.; Silvestro, D.; Töpel, M.; et al. Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of Taxa. *Syst. Biol.* **2017**, *66*, 153–166. [CrossRef] [PubMed]
7. Pearse, W.D.; Purvis, A. phyloGenerator: An automated phylogeny generation tool for ecologists. *Methods Ecol. Evol.* **2013**, *4*, 692–698. [CrossRef]
8. Eiserhardt, W.L.; Antonelli, A.; Bennett, D.J.; Botigué, L.R.; Burleigh, J.G.; Dodsworth, S.; Enquist, B.J.; Forest, F.; Kim, J.T.; Kozlov, A.M.; et al. A roadmap for global synthesis of the plant tree of life. *Am. J. Bot.* **2018**, *105*, 1–9. [CrossRef] [PubMed]

9. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
10. PhyLoTa Browser. Available online: Phylota.net (accessed on 28 March 2018).
11. GenBank and WGS Statistics. Available online: www.ncbi.nlm.nih.gov/genbank/statistics (accessed on 28 March 2018).
12. Taxonomy Browser: Taxonomy Statistics. Available online: www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html (accessed on 28 March 2018).
13. Altenhoff, A.M.; Glover, N.M.; Train, C.M.; Kaleb, K.; Warwick Vesztrocy, A.; Dylus, D.; De Farias, T.M.; Zile, K.; Stevenson, C.; Long, J.; et al. The OMA orthology database in 2018: Retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.* **2018**, *46*, D477–D485. [CrossRef] [PubMed]
14. Waterhouse, R.M.; Zdobnov, E.M.; Tegenfeldt, F.; Li, J.; Kriventseva, E.V. OrthoDB: The hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res.* **2011**, *39* (Suppl. 1). [CrossRef] [PubMed]
15. Pearson, W.R.; Lipman, D.J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 2444–2448. [CrossRef] [PubMed]
16. Wickham, H.; Hester, J.; Chang, W.; Rstudio; R Core Team. Devtools: Tools to Make Developing R Packages Easier. 2018. Available online: CRAN.R-project.org/package=devtools (accessed on 28 March 2018).
17. The Comprehensive R Archive Network. Available online: Cran.r-project.org (accessed on 28 March 2018).
18. Bioconda. Available online: https://bioconda.github.io/ (accessed on 7 May 2018).
19. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *Int. J. Complex Syst.* **2018**, *1695*, 1–9.
20. Winter, D. Rentrez: Entrez in R. R Package Version 1.1.0. 2017. Available online: CRAN.R-project.org/package=rentrez (accessed on 28 March 2018).
21. 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 March 2018).
22. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2009.
23. Ooms, J. Sys: Portable System Utilities. Available online: CRAN.R-project.org/package=sys (accessed on 28 March 2018).
24. Bennett, D.J.; Sutton, M.D.; Turvey, S.T. Treeman: An R package for efficient and intuitive manipulation of phylogenetic trees. *BMC Res. Notes* **2017**, *10*, 30. [CrossRef] [PubMed]
25. Wilkins, D. Treemapify: Draw Treemaps in 'ggplot2'. Available online: CRAN.R-project.org/package=treemapify (accessed on 28 March 2018).
26. Bengtsson, H.R. utils: Various Programming Utilities. Available online: CRAN.R-project.org/package=R.utils (accessed on 28 March 2018).
27. BLAST® Command Line Applications User Manual. Available online: www.ncbi.nlm.nih.gov/books/NBK279690 (accessed on 28 March 2018).
28. Transforming Science Through Open Data and Software. Available online: Ropensci.org (accessed on 28 March 2018).
29. Buchfink, B.; Xie, C.; Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **2015**, *12*, 59–60. [CrossRef] [PubMed]
30. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461. [CrossRef] [PubMed]
31. Kent, W.J. BLAT—The BLAST-like alignment tool. *Genome Res.* **2002**, *12*, 656–664. [CrossRef] [PubMed]
32. Nguyen, V.H.; Lavenier, D. PLAST: Parallel local alignment search tool for database comparison. *BMC Bioinform.* **2009**, *10*. [CrossRef] [PubMed]
33. Entrez Molecular Sequence Database System. Available online: www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html (accessed on 28 March 2018).
34. Basic Local Alignment Search Tool. Available online: https://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed on 28 March 2018).
35. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef] [PubMed]
36. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef] [PubMed]

37. 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. *Syst. Biol.* **2009**, *58*, 240–256. [CrossRef] [PubMed]
38. 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.; et al. A molecular phylogeny of living primates. *PLoS Genet.* **2011**, *7*, 1–17. [CrossRef] [PubMed]
39. Federhen, S. The NCBI taxonomy database. *Nucleic Acids Res.* **2012**, *40*, D136–D143. [CrossRef] [PubMed]
40. Robinson, D.F.; Foulds, L.R. Comparison of phylogenetic trees. *Math. Biosci.* **1981**, *53*, 131–147. [CrossRef]
41. Critchlow, D.E.; Pearl, D.K.; Qian, C. The triples distance for rooted bifurcating phylogenetic trees. *Syst. Biol.* **1996**, *45*, 323–334. [CrossRef]
42. 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. *Am. J. Phys. Anthropol.* **2005**, *127*, 465–480. [CrossRef] [PubMed]
43. Shi, C.M.; Yang, Z. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Mol. Biol. Evol.* **2018**, *35*, 159–179. [CrossRef] [PubMed]
44. 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 Evolut. Biol.* **2008**, *8*, 1–12. [CrossRef] [PubMed]
45. 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 Biol.* **2011**, *9*. [CrossRef] [PubMed]
46. 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 Bull.* **2005**, *60*, 559–569. [CrossRef]