GigaScience

Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale

--Manuscript Draft--

Manuscript Number:	GIGA-D-18-00245R1
Full Title:	Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale
Article Type:	Research
Funding Information:	
Abstract:	Background In light of the current biodiversity crisis, DNA barcoding is developing into an essential tool to quantify state shifts in global ecosystems. Current barcoding protocols often rely on short amplicon sequences, which yield accurate identification of biological entities in a community, but provide limited phylogenetic resolution across broad taxonomic scales. However, the phylogenetic structure of communities is an essential component of biodiversity. Consequently, a barcoding approach is required that unites robust taxonomic assignment power and high phylogenetic utility. A possible solution is offered by sequencing long ribosomal DNA (rDNA) amplicons on the MinION platform (Oxford Nanopore Technologies). Findings Using a dataset of various animal and plant species, with a focus on arthropods, we assemble a pipeline for long rDNA barcode analysis and introduce a new software (MiniBar) to demultiplex dual indexed nanopore reads. We find excellent phylogenetic and taxonomic resolution offered by long rDNA sequences across broad taxonomic scales. We highlight the simplicity of our approach by field barcoding with a miniaturized, mobile laboratory in a remote rainforest. We also test the utility of long rDNA amplicons for analysis of community diversity through metabarcoding and find that they recover highly skewed diversity estimates. Conclusions Sequencing dual indexed, long rDNA amplicons on the MinION platform is a straightforward, cost effective, portable and universal approach for eukaryote DNA barcoding. Although bulk community analyses using long-amplicon approaches may introduce biases, the long rDNA amplicons approach signifies a powerful tool for enabling the accurate recovery of taxonomic and phylogenetic diversity across biological communities.
Corresponding Author:	Henrik Krehenwinkel UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Henrik Krehenwinkel
First Author Secondary Information:	
Order of Authors:	Henrik Krehenwinkel
	Aaron Pomerantz

	James B. Henderson
	Susan R. Kennedy
	Jun Ying Lim
	Varun Swamy
	Juan Diego Shoobridge
	Natalie Graham
	Nipam H. Patel
	Rosemary G. Gillespie
	Stefan Prost
Order of Authors Secondary Information:	

Response to Reviewers:

Dear Editors
We thank the two reviewers for their constructive and helpful comments. We have now incorporated their suggested changes to our manuscript. We have provided additional detail to the introduction, methods, results and discussion and added two Supplementary Figures to better illustrate our findings.
We hope that the manuscript will now be deemed acceptable for publication in Gigascience.
Sincerely,
Henrik Krehenwinkel

Reviewer reports:
Reviewer #1: The authors present us an rDNA-based barcoding and phylogeny study using a MinION sequencing platform. It is an instructive trial and I suggest the editor make it published after addressing several issues as follows:

1. The authors should be cautious of scientific writing and provide evidences to what you have written. For example, the authors stated that one of the pitfalls of mitochondrial genes is the risk of homoplasy of divergent lineages because of saturation. However, a short standard COXI barcode of length ca. 600 bp can hold a variety of 4^600, 4^200 even only take into the third position into account, which is far more than the species number on earth. In addition, nowadays mitochondrial genes are well known of its limitation in phylogeny works due to reasons mentioned by the authors in lines 80-90, but I image that most of these limitations should affect much on demographic history inferences for single species or phylogenetic work of closely related species, rather than biodiversity oriented and alpha or beta diversity based ecological works. I encourage the authors to pay more attentions on their writing to avoid biased texts which may mislead readers.
-We fully acknowledge the almost unlimited number of informative sites in a COI barcode. COI is certainly well suited to distinguish species and this is not affected by homoplasy. We simply meant that its utility to resolve phylogenetic divergence is limited by homoplastic sites. At deep phylogenetic divergence, the sequence saturates with mutations, making it hard to properly reconstruct relationships. We have made this clearer in the introduction.

2. Same to 1, at line 116, in opposite to what the authors stated, ITS2 is proposed to be the optimal barcode marker for plants and fungi.
-We personally have found considerable drop out of arthropod specimens during PCR using common universal ITS primers, but agree that it is a widely used and well-suited taxonomic marker for many other lineages. We have rephrased the according section. We now particularly focus on the difficulty of aligning the extremely variable ITS sequences across divergent lineages.

3. Although the authors mentioned the Pacbio sequencer as an alternative method to explore community compositions in lines 123-127, I think it needs more words to make it clear that the CCS (circular consensus sequencing) tech of Pacbio sequencing

platform may be more suitable for amplicons-based barcoding and biodiversity work. However, comparing to Nanopore tech, it can hardly be conducted in a real-time way and in the field.
-We have rewritten the relevant sections. We highlight the utility and advantage of the CSS sequencing. We then focus on the advantage of the MinION of being a portable and easily accessible device.

4. I agree that an empirical experiment is necessary to test how Nanopore tech works on the estimation of metazoan community diversity. However, what impedes MinION from amplicons-based diversity study is its lower per base accuracy. The authors should understand that the alpha diversity inflation is still one of the major concerns even using the widely applied HiSeq sequencing platform which holds much higher sequencing accuracy. I believe the MinION-based study, at current stage, is far from being worry about such problems. I am afraid that researchers in this field are still skeptical of its applicability in metabarcoding at current stage. As I see in the authors' work, you manually mixed phylogenetically divergent species - species from different orders - to avoid taxonomic assignment issues. But the authors should also be aware that such a design has less practical guiding significances.
-We agree that the MinION is not yet ready for routine community analysis, as also shown by our data. Expecting difficulties with this system, we have used highly simplified community samples, to explore its potential utility for community analysis. We acknowledge that our mock communities are not directly comparable to natural communities and have revised the methods and discussion to highlight this. Finding highly biased results in these simplified communities already highlights the possible difficulties of this system in real communities.

5. For the consensus sequences of plants or fungi mentioned in lines 408 - 410, if they are food chain derived, have you ever tried to cluster reads at first, then call consensus for each cluster? Or as you mentioned in lines 650 -652, check taxonomic composition by blasting a reference library before assembly.
-We have tried this and found that for these samples, the majority or even all assigned sequences belonged to the non-targeted species; the host was almost undetectable in these cases. For example, we did not find a single Zophobas beetle sequence in an extract of Zophobas larvae, but highly abundant rye DNA sequences. We have added an explanatory sentence to the results.

6. The authors mentioned that coverage larger than 300 can lead to a decrease of consensus accuracy. It deserves further scrutiny to get reasonable explanations. In addition, read number increased a lot per sample with a minibar setting of edit distance of 4, which, however, generated less accurate consensuses. Are there any correlations between these two observations?
-This difference is visible in the plot, but it is not significant. We have added this information in the results. It is also visible that there is an overlap of consensus accuracy at coverages > 300 and < 300. Hence, only part of the samples showed a lower consensus accuracy at high coverage. We assume that this is due to some samples randomly getting assigned more wrongly demultiplexed samples at high coverages. There always seems to be a small carryover between indexes. These wrongly assigned sequences may affect consensus building. At an edit distance of four, we indeed found a considerable increase of wrongly assigned sequences, e.g. cross contamination between samples. This affected consensus building and led to inaccurate consensus sequences. We have added this information to the results.

7. How do you annotate the rDNA to separate the different segments - 18S, 5.8S, ITS, et al.
-We used annotated reference sequences from Genbank. We have now included this information in methods

8. Is there any data that support what you mentioned in lines 661 - 662: "indices of 20 or 30 bp attached to primers doesn't strongly affect CPR efficiency"?
It is common practice to use tailed primers in Illumina amplicon sequencing, which are longer than 50 bp and work efficiently. Our indexes are considerably shorter. Yet, on the other hand, the targeted amplicon is also much longer. To our knowledge, there is

no proof for our assumption. We have thus rephrased the according sentence in the discussion and removed the statement
-
9. Please make sure correct citations, e.g. I don't think reference number 48 talked about anything related to what you stated there at line 666.
-The reference was corrected

10. Others:

Supplemental figure 1. Please add the unite of your Y axis, should be in percent, isn't it?
-The unit added to the figure is percent

Line 255, is it minimap2?
-We have used minimap, not minimap2. We have, however recently tested minimap2 and it did not yield better results.

Line 285, do you mean crossover?
-We mean samples being wrongly assigned due to indexes being misidentified due to sequencing error. We have reworded this to crossover.

Reviewer #2: The manuscript entitled 'Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale' by Krehenwinkel and collaborators aims to evaluate the use of Nanopore to produce high quality consensus sequences for a long fragment spanning the rDNA region. The authors evaluate the usage of the different genes and two gene interspace regions contained within the amplified fragment and that present different levels of nucleotide variability, for a comparative analysis at different taxonomic levels, most notably within arthropods and specifically within a group of spiders.

The manuscript is well written, and mostly clear in the ideas. The methods and experiments presented seem to complement each other and are ultimately showing the potentiality of using the methodology described in the manuscript for barcoding a long and highly informative region of the rDNA 'operon' for any given arthropod (or potentially eukaryote) species in the location where it is sampled, independently of the local laboratory infrastructure. The use of such 'portable approaches' for the study of biodiversity are highly desirable in times in which biodiversity is fast declining and samples exporting from regions representing biodiversity hotspots are facing more severe regulations. Certainly, performing methods locally but with infrastructure that can be easily transported from abroad if needed is a great advantage.

My main criticism to the text is the confusion made between biases produced by long and short-read technologies and those produced by the different types of amplicons generated. The authors should differentiate between PCR efficiency and sequencing technologies across the paper. For example, instead of 'long-read metabarcoding' please call it 'long-amplicon metabarcoding'. This makes it clear that the problems found are due to the PCR, potentially due to its long-range nature. However, it should also be made clear that optimization of PCR conditions is needed for both short and long-range when new primers are developed. Note that even though it is a natural expectation that long reads will be used for sequencing long amplicons and short reads for short amplicons, this is not a rule. Illumina can be used for shotgun sequencing of long amplicons and Nanopore could potentially be used for short amplicons or even concatenated short amplicons.
Abstract

I suggest to change in line 48 long-read by long-amplicon or by 'long-amplicon approaches combined with long-sequencing technologies'.
-Was done throughout

Background

Line 81 - the authors are mostly talking about the COI gene and not mitochondrial DNA in general. Mitogenomes also have a combination of genes with more or less expected levels of divergence between species. Some genes, such as the non-coding 16S and 12S have very conserved regions across taxa. If one could potentially amplify different mitochondrial genes across taxa in one single amplicon, the power would be probably at least similar to the rDNA operon, but apart from the issues already described by the authors regarding the peculiarities of the mitogenome such as maternal inheritance and the possibility for introgressive hybridization, mitogenomes might vary a lot in synteny, content and number of gene copies in some phyla (e.g. Fungi) and are therefore not exactly useful for amplifying a number of homologous regions consistently across eukaryotes.
-We have also considered using mitochondrial DNA, which has many advantages as well. 16S and 12S do indeed have fairly conserved sequence stretches allowing the design of primers, which efficiently amplify a wide range of taxa. However, they are not nearly as conserved as nuclear rDNA. In our experience, universal 12S or 16S primers may allow us to amplify all taxa across an order or phylum, but not a whole domain as nuclear rDNA does.
In the long run, we aim to develop a combined approach utilizing nuclear and mitochondrial long amplicon information. We have added additional information on this in the discussion.

Having said that, I never looked in more detailed into this possibility, so there might be certain genes that always occur in synteny in mitogenomes. But I agree that mitochondrial DNA is not always representative of phylogenies. This brings us to the general questions that should be posted after line 111. Are the peculiarities of the rDNA operon a potential bias for some phylogenetic inferences? For example, the variable (and unknown) number of copies across species that may or may not be all identical. I would appreciate some acknowledgement of the potential uncertainties on phylogenies based on rDNA already in the introduction.
-We now acknowledge the limitations of single rDNA sequences in the introduction and discussion. E.g. nuclear rDNA can also be prone to paralogs and possibly pseudogenization. We also highlight the combination of long mitochondrial and nuclear amplicons as an ideal solution for future barcoding applications in the discussion.

Line 116 - it is true the ITS regions are too variable for designing universal primers, but they are flanked by conserved regions, and to the best of my knowledge ITS2 is not as variable in length as ITS1. So, instead of describing the variability of ITS regions as impeditive to short-amplicon primers design, I would rather discuss the fact that it cannot be aligned among unrelated taxa, and are not suitable for deeper phylogenies. Besides, it can only be used for taxonomical assignment if a somehow related group is represented in the database.
-We acknowledge this and have rewritten the according section in the introduction.

Line 133 - I would add consensus sequences 'from single individuals'. I was confused at first thinking that Nanopore could maybe do some sort of 'circular consensus', but if the consensus sequences are produced by homologous sequences from a single individual this should be made clear.
-Has been made clear in the introduction

Line 141 - I would rephrase 'universal eukaryote'. Even though the primers could potentially work for all eukaryotes, there was no representative collection tested, and the authors stated themselves that there was a focus in animals.
-Was rephrased

Data Description and Analyses

Line 201 - following the idea above of exploring the universality of the primers, I would like to see some sort of figure or graph showing the representativeness of the different groups of eukaryotes in the 1000 sequences used for the primers design.
-We have added this graph as a supplementary figure

Line 214 - How was the quantification on an agarose gel performed? I would suggest a description how this was done and an evaluation of the pooling method in the Results/Discussion as fluctuations on samples sequence numbers may highly influence the efficiency and costs of the method.
-We have added the details for this approach in the methodology. We acknowledge that it may introduce some biases and have added a discussion for this

Line 221 - Please inform the concentration of AMPure beads utilized
-We used 0.75 X beads on 100 ul, e.g. 75 ul of beads. The volume was added to the manuscript

Results

Line 383 and Fig.2 - the authors state that at a distance of 4, samples had an increase in wrongly assigned sequences and a significantly lower accuracy in the consensus generated. However, what is shown in Fig. 2 is a box plot of pairwise distances of Nanopore sequences assigned to the sample against the Illumina consensus. How do the authors know that the sequences were wrongly assigned? Could they be assigned to other samples based on sequence distance? Is there a real change in the consensus sequence generated by the sequences assigned to a sample at a distance of 4? If so, why is that? Due to more indels and/or more mismatches? What are the features of the newly assigned sequences that decrease the accuracy of the consensus? Could the higher distance at the barcode also incorporate sequences with more errors (i.e. is the number of errors in barcodes correlated to lower quality/more errors)? Are the errors distributed throughout the sequences? In my view it's important to understand the causes of lower accuracy, because absolute numbers, such as 2, 3 or 4 mismatches, might not represent the same issues when different barcode length, sequences or combinations are used.
-We have blasted the raw reads to explore potential carry over between indexes. At an edit distance of four, we indeed found a considerable increase of wrongly assigned sequences, e.g. cross contamination between samples. This affected consensus building and led to inaccurate consensus sequences. We have added this information to the results.

Line 420 - please show examples of alignments with errors clustered in indel regions in the supplementary material. It is important for the reader to understand the patterns of errors found.
-We added a supplementary figure detailing the increased error at homopolymers.

Line 429 - what could be the reason for a decrease in accuracy in higher coverages? Is this increase stochastic and no significant? Or is the incorporation of sequences with more (and maybe slightly repetitive) errors causing differences in the consensus? It would be very interesting to understand if the consensus creation is very sensitive to accumulation of identical errors, even if in small rates.
-As stated above, this difference was not significant. We have added this information in the results. It is also visible that there is an overlap of consensus accuracy at coverages > 300 and < 300. Hence, only part of the samples showed a lower consensus accuracy at high coverage. We assume that this is due to some samples randomly getting assigned more wrongly demultiplexed samples at high coverages. There always seems to be a small carryover between indexes. These wrongly assigned sequences may affect consensus building.

Lines 431 to 449 and Suppl. Figures 3 and 4 - even though I understand the value of presenting and summarizing results, the authors should not treat the data as representative neither for animals nor for plants. Please refer to the main groups analyzed (Arachnids, Insects and Magnoliopsida) and if there is an interest to compare to animals and plants in general, pick representative sequences from both groups (animals and plants) from public databases and present a comparison. For the data presented here, my suggestion would be one single boxplot graph for length difference presenting Arachnids, Insects and if wanted Magnoliopsida including both lengths

excluding and including ITS regions, for a better understanding of the differences between full versus coding-only lengths. Another graph (or figure number) can summarize the same way the GC content.
-We have remade the plots as suggested by the reviewer. And have rewritten the according text section in the results.

Figures 3, 4 and 5 are presented before they are mentioned in the text.
-Has been corrected

Lines 471 to 483 and figure 3 - I wonder here what the value is in building such a phylogenetic tree including non-representative but yet arbitrarily picked species from three different kingdoms. Is the intent to show that the sequences produced by Nanopore are as accurate as sequences produced by other technologies and that differences/errors do not affect phylogenetic reconstruction? In that case, I would compare the tree with the data from same/similar species from databases for each group. Or is the intent showing that rDNA can be used to reconstruct true phylogenies for the groups? This doesn't seem to be part of the goals, but would demand other tests, again including many more sequences from databases. If the authors are presenting novelties and would like to place some group phylogenetically, there is nothing wrong (or better, it's the right thing to go) in picking representative sequences from databases for showing the correct phylogenetic placement of a group.
-We aimed to show a widely applicable method here, which is why we used different other taxa besides arthropods, even though our focus is clearly on arthropods. We did not aim to reconstruct the tree of life for these groups, but merely show that it is possible to amplify, sequence and align rDNA sequences across different domains of the eukaryote tree of life. Our primary goal is to allow amplification of all organisms from a given biological community; importantly, this provides a means to generate metrics of similarity between communities based on quantitative phylogenetic data. If Figure 3 is not important, we are happy to move it to supplement. Our sampling is particularly focused on arthropods, for which we present a wide range of taxa. Starting at the phylum level, we move into the order spiders and show that the recovered phylogeny is well comparable to recent work based on whole transcriptomes. We then even move to the genus level and present a detailed analysis in a genus of Hawaiian spiders. Reconstructing the tree of life with additional database sequences for all eukaryote groups would extend beyond the scope of our study, which already is extensive and has multiple facets.

Lines 485 to 499 - Spiders are surely much better represented than other groups. But if there are other sequences in databases not represented here, I would include them. The question always goes back to the intent. Is it to show that the authors are contributing with valuable and correct rDNA sequences to populate databases? Then the improvement in phylogeny reconstruction should involve all sequences available and for which taxonomy can be trusted (I'm accounting here for possible errors or uncertainties in databases). This would reinforce the value of the approach in creating new references for an important and informative marker (or better saying, cluster of markers with different levels of divergence among different taxonomical groups), the rDNA region.
-Whole rDNA clusters are still not very well represented in public databases. This holds particularly true for spiders, for which very few whole rDNA sequences are present in the databases. It was not our aim to reconstruct a complete spider tree of life, but rather to show that the rDNA cluster allows to recover the known phylogenetic divergence for the limited set of taxa we have used here. Our data already covers a considerable portion of the araneaomorph spider tree of life and we present the resolution of rDNA sequence across multiple taxonomic levels, from family down to species. The spider tree of life is well resolved by RNA seq data. We simply show that the rDNA cluster alone resolves a very congruent phylogeny at multiple levels.

Line 531 and 532, Fig.6 and Discussion - the overlap between inter and intra-specific distance in rDNA might seem small but could have serious consequences if not interpreted well. Please show if in the dataset the distance would be impeditive of taxonomical assignment based on distance for some lineages. One simple way would be to highlight the circles (e.g. make them darker) in Fig. 6 and suppl. Fig 5 with high

intra-specific variability in rDNA for both intra and inter-specific distances. If the highlighted circles have high distance for inter-specific comparisons as well, this would not affect taxonomical classification, at least for those set of species/specimens presented.
-The overlap between distances for the rDNA cluster is caused by only very few species. In fact, it is only a single case of high intraspecific distance basing the overlap. This case possibly involves a cryptic species pair. Which also shows a high distance in COI. We are currently examining material of the species morphologically, to explore this further. Also, the species shows a considerably higher interspecific distance to other Tetragnatha species, than its intraspecific distance. A pair of very closely related species from Maui show the lowest interspecific distance. We have added some more details on this in the results section

I wonder if the inter X intra-specific distances gap in COI is a natural one, or if in some cases COI was itself used for re-defining species boundaries, which would make the analysis redundant. In any case, it seems that in general it could be a good approach, even if it increases the level of Complexity, to sequence both rDNA and COI, as mentioned in the Discussion. I just wonder that, if COI could also be sequenced in the field with a similar approach (i.e. using Nanopore), as some samples might never reach the lab in the original country where the research is being conducted.
-The observed barcode gap in COI is in fact natural. We have now explained this in the text. All species, which we used for this study were identified morphologically before we performed barcoding analysis. Also, we are currently exploring the possibility of using a combined approach of long mitochondrial and nuclear rDNA amplicons for taxonomic analyses, which we believe would be an ideal solution.

Lines 569-586 - The Illumina metabarcoding seem highly accurate when compared to Nanopore in Suppl. Fig 8 (please correct in line 574 the figure number), but this does not confirm that Illumina is quantitatively accurate, as the long amplicons generated huge biases. In fact, some taxa seem to have more than 2-fold difference in the Illumina results compared to their original frequency in the mock community based on Fig. 7. Was the adjustment per taxon performed as in reference [34]?
-Illumina sequencing was not perfectly accurate quantitatively and we did not correct read abundances. However, Illumina sequences recovered abundance trends very well. We have toned down the relevant sections in the results and discussion and added additional explanations. However, the bias observed by MinION sequencing was considerably higher than that found by Illumina.

Discussion
Line 616 - I would like to see a bit deeper discussion on the feasibility for developing universal primers for sequencing full or partial mitogenomes across taxa, especially considering gene synteny and content within eukaryotes, apart from nucleotide variation.
-We have added this in the discussion. We fully agree that adding long mitochondrial amplicons would be a great complement to our work. Ideally, one would rely on both markers. We have added additional details in the discussion highlighting the need for multi locus approaches and a combination of mitochondrial and nuclear data.

Lines 625-626 - please rephrase to something similar to: long rDNA amplicons can potentially be amplified across diverse eukaryote taxa, here largely demonstrated in arthropods and arachnids, and in very small scale in fungi and plants.
-Was rephrased

Line 654 - rDNA is not only diploid but present in multiple (and unknown) number of copies, that might not be identical
-We have added additional sentences in the introduction and discussion, considering the possible problem with multi copy rDNA markers.

Line 661 - it is not clear that even longer tails would not affect PCR, please either add a reference to confirm the statement or change it.

	-As stated above, long primer tails are commonly used for Illumina sequencing. But to our knowledge their effect was not exhaustively tested yet. Also, our amplicon is much longer than typical Illumina sequenced amplicons. As we did not test it, we have removed the statement. Line 666 - reference number is wrong, maybe [47]? -Was corrected Nanopore metabarcoding - I think explanations are quite reasonable for justifying why certain taxonomical groups, especially those with shorter rDNA regions, would be preferentially amplified. However, given that the whole study focused on spiders, could the conditions be better optimized for spiders rather than other arthropods? -The protocol could probably work better for groups like spiders, which show small variation in amplicon length. We are currently testing and optimizing our protocol in that regard. We have added additional explanations and details in the methods and discussion. References [22] is now published in Mol Ecol Resources -Was changed
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.	Yes

Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale

Henrik Krehenwinkel[1,4], Aaron Pomerantz[2], James B. Henderson[3,4], Susan R. Kennedy[1], Jun Ying Lim[1,2], Varun Swamy[5], Juan Diego Shoobridge[6], Natalie Graham[1], Nipam H. Patel[2,7], Rosemary G. Gillespie[1], Stefan Prost[2,8]

[1] Department of Environmental Science, Policy and Management, University of California, Berkeley, USA

[2] Department of Integrative Biology, University of California, Berkeley, USA

[3] Institute for Biodiversity Science and Sustainability, California Academy of Sciences, San Francisco, USA

[4] Center for Comparative Genomics, California Academy of Sciences, San Francisco, USA

[5] San Diego Zoo Institute for Conservation Research, Escondido, USA

[6] Applied Botany Laboratory, Research and development Laboratories, Cayetano Heredia University, Lima, Perú

[7] Department of Molecular and Cell Biology, University of California, Berkeley, USA

[8] Research Institute of Wildlife Ecology, Department of Integrative Biology and Evolution, University of Veterinary Medicine, Vienna, Austria

Corresponding authors: Henrik Krehenwinkel (krehenwinkel@berkeley.edu) and Stefan Prost (stefan.prost@berkeley.edu)

Abstract

Background: In light of the current biodiversity crisis, DNA barcoding is developing into an essential tool to quantify state shifts in global ecosystems. Current barcoding protocols often rely on short amplicon sequences, which yield accurate identification of biological entities in a community, but provide limited phylogenetic resolution across broad taxonomic scales. However, the phylogenetic structure of communities is an essential component of biodiversity. Consequently, a barcoding approach is required that unites robust taxonomic assignment power and high phylogenetic utility. A possible solution is offered by sequencing long ribosomal DNA (rDNA) amplicons on the MinION platform (Oxford Nanopore Technologies).

Findings: Using a dataset of various animal and plant species, with a focus on arthropods, we assemble a pipeline for long rDNA barcode analysis and introduce a new software (MiniBar) to demultiplex dual indexed nanopore reads. We find excellent phylogenetic and taxonomic resolution offered by long rDNA sequences across broad taxonomic scales. We highlight the simplicity of our approach by field barcoding with a miniaturized, mobile laboratory in a remote rainforest. We also test the utility of long rDNA amplicons for analysis of community diversity through metabarcoding and find that they recover highly skewed diversity estimates.

Conclusions: Sequencing dual indexed, long rDNA amplicons on the MinION platform is a straightforward, cost effective, portable and universal approach for eukaryote DNA barcoding. Although bulk community analyses using long-amplicon approaches may introduce biases, the long rDNA amplicons approach signifies a powerful tool for enabling the accurate recovery of taxonomic and phylogenetic diversity across biological communities.

Keywords

Biodiversity, ribosomal, eukaryotes, long DNA barcodes, Oxford Nanopore Technologies, MinION, metabarcoding

Background

The world is changing at an unprecedented rate, threatening the integrity of biological communities [1, 2]. To understand the impacts of change, whether a system is close to a regime shift, and how to mitigate the impacts of a given environmental stressor, it is important to consider the biological community as a whole. In recognition of this need, there has been a shift in emphasis from studies that focus on single indicator taxa, to comparative studies across multiple taxa and metrics that consider the properties of entire communities [3]. Such efforts require accurate information on the identity of the different biological entities within a community, as well as the phylogenetic diversity that they represent.

Comparative ecological studies across multiple taxa have been greatly simplified by molecular barcoding [4], where species identifications are based on short PCR amplicon “barcode” sequences. Different barcode marker genes have been established across the tree of life [5, 6], with mitochondrial cytochrome oxidase subunit I (COI) commonly used for animal barcoding [4]. The availability of large sequence reference databases and universal primers, together with its uniparental inheritance and fast evolutionary rate, make COI a useful marker to distinguish even recently diverged taxa. In recent years, DNA barcoding has greatly profited from the emergence of next generation sequencing (NGS) technology. Current NGS platforms enable the parallel generation of barcodes for hundreds of specimens at a fraction of the cost of Sanger sequencing [7]. Furthermore, NGS technology has enabled metabarcoding, the sequencing of bulk community samples, which allows scoring the diversity of entire ecosystems [8].

However, despite their undeniable advantages, barcoding approaches using short, mitochondrial markers have several drawbacks. The phylogenetic resolution offered by short barcodes is very limited, as they contain only a restricted number of informative sites. This problem is exacerbated by the fast evolutionary rate of mitochondrial DNA, which leads to a quick saturation with mutations, increasing the probability of homoplasy. While this does not affect the taxonomic utility of COI, it causes problems in phylogenetic analyses of divergent lineages. The accurate estimation of phylogenetic diversity across wide taxonomic scales, however, is an important component of biodiversity research [9]. Moreover, mitochondrial DNA is not always the best marker to reflect species differentiation, as different factors are known to inflate mitochondrial differentiation in relation to the nuclear genomic background. For example, male biased gene flow [10] or infections with reproductive parasites [11] (e.g. *Wolbachia*) can lead to highly divergent mitochondrial lineages in the absence of nuclear differentiation. In contrast, introgressive hybridization can cause the complete replacement of mitochondrial genomes (see e.g. [12, 13]), resulting in shared mitochondrial variation between species.

Considering this background, it would be desirable to complement mitochondrial DNA based barcoding with additional information from the nuclear genome. An ideal nuclear barcoding marker should possess sufficient variation to distinguish young species pairs, but also provide support for phylogenetic hypotheses between divergent lineages. Moreover, the marker should be present across a wide range of taxa and amplification should be possible using universal primers. A marker that fulfils all the above requirements is the nuclear ribosomal DNA (rDNA). As an essential component of the ribosomal machinery, rDNA is a common feature across the tree of life from microbes to higher eukaryotes [14]. All eukaryotes share homologous transcription units of the 18S, 5.8S and 28S-rDNA genes, which include two internal transcribed spacers (ITS1 and ITS2) [15]. Due to varying evolutionary constraints acting on different parts of the rDNA, it consists of regions of extreme sequence conservation, which are interrupted by highly variable stretches [16]. While some rDNA gene regions are entirely conserved across all eukaryotes, the two ITS sequences are distinguished by such rapid evolutionary change that they separate even lineages within species [5, 17]. rDNA markers thus offer taxonomic and phylogenetic resolution at a very broad taxonomic scale. As an essential component of the translation machinery, nuclear rDNA is required in large quantities in each cell. It is thus present in multiple copies across the genome [15] and is readily accessible for PCR amplification. Due to the above advantages, rDNA already is a popular and widely used marker for molecular taxonomy and phylogenetics in many groups of organisms [5, 6, 15, 17, 18]. However, its presence in multiple copies across the genome may also make rDNA susceptible to the emergence of paralogs and pseudogenization, which could affect taxonomic and phylogenetic utility.

Spanning about 8 kb, the ribosomal cluster is fairly large, and current barcoding protocols, e.g. using Sanger sequencing or Illumina amplicon sequencing, can only target short sequence stretches of 150 – 1,000 bp. Such short stretches of 28S and 18S are often too conserved to identify young species pairs [19]. The ITS regions, on the other hand, are so variable that they cannot be properly aligned across divergent lineages. Moreover, ITS sequences can show considerable length variation between taxa, and are often too long for short amplicon-based barcoding [20].

Consequently, it would be ideal to amplify and sequence a large part of the ribosomal cluster in one fragment. A solution to sequence the resulting long amplicons is offered by recent developments in third generation sequencing platforms, which now enable researchers to generate ultra-long reads, of up to 800 kb [21]. Recently, amplicons of several kilobases of the rDNA cluster were sequenced using Pacific Bioscience (PacBio) technology, to explore fungal community composition [22, 23]. With its circular consensus sequencing technology, PacBio allows the generation of very accurate consensus reads. But while PacBio sequencing is well suited for long amplicon sequencing, it is currently not readily available to every laboratory due to the high cost and limited distribution of sequencing machines. PacBio sequencers are also bulky and cannot be used outside of conventional laboratory settings.

A cost-efficient and readily available alternative is provided by nanopore sequencing technology. The MinION sequencer (Oxford Nanopore Technologies) is small in size, lightweight, allows for sequencing of several Gb's of DNA with average read lengths over 10 kb on a single flow cell [24] and is available starting at $1,000. Despite a raw read error rate of about 12-22 % [21], highly accurate consensus sequences can be called from nanopore data [25, 26], by assembling multiple sequences for individual specimens. The MinION is well suited for amplicon sequencing, and a simple dual indexing strategy can be used to demultiplex amplicon samples [27]. This technology offers tremendous potential for long-amplicon barcoding applications, as recently shown in an analysis in fungi [26]. Oxford Nanopore Technologies' MinION is a portable sequencer, and Nanopore based DNA barcoding can be applied with mobile laboratories in remote sites outside of conventional labs (see e.g. [25, 28, 29]). However, current analyses are still exploratory or limited in taxonomic focus, and streamlined analysis pipelines to establish the method across the eukaryote tree of life are still missing.

Considering this background, we explore the feasibility of nanopore sequencing of long rDNA amplicons as a simple, cost efficient DNA barcoding approach for animals and other eukaryote taxa. We compile a workflow from PCR amplification, to library preparation, to demultiplexing and consensus calling (see Fig. 1 for an overview). We explore the error profile of nanopore consensus sequences and introduce MiniBar, a new software to demultiplex dual indexed nanopore amplicon sequences. We test the utility of the ribosomal cluster for molecular taxonomy and phylogenetics across divergent plant and animal taxa. A particular focus of our analysis are arthropods, the most diverse group in the animal kingdom [30], which are highly threatened by current mass extinctions [31]. Using a dataset of spiders, we compare the taxonomic resolution of the ribosomal cluster with that offered by molecular barcoding using mitochondrial COI, the currently preferred barcode marker for arthropods. Oxford Nanopore Technologies' MinION is a portable sequencer, and Nanopore based DNA barcoding has been applied in remote sites outside of conventional labs (see eg. [25, 30, 31]). As mentioned above, the MinION is portable and can be used for DNA barcoding in field settings. Such field-based applications confront researchers with additional complexities and challenges. To highlight the simplicity of our approach, we tested it under field conditions and generated long rDNA barcode sequences using a miniaturized mobile laboratory in a Peruvian rainforest.

We also tested the efficacy of long-amplicon rDNA sequencing for metabarcoding of bulk community samples. A study of bacterial communities [32] suggests Nanopore long-amplicon sequencing as a powerful tool for community characterization, but also found pronounced biases in the recovered taxon abundance. Currently, little is known about the utility of long-amplicon sequencing for animal community analysis. Metabarcoding protocols for community samples need to be carefully optimized, as they can suffer from pronounced taxonomic biases, e.g. due to primer binding or polymerase efficiency [33]. Well established Illumina based short amplicon metabarcoding protocols can account for these biases and allow for a relatively good qualitative and even quantitative recovery of taxa in communities [34]. However, additional, yet unexplored, biases may affect long-amplicon metabarcoding. We thus also test the utility of long-amplicon rDNA barcoding to recover taxonomic diversity from arthropod mock communities. We compare the qualitative (species richness) and quantitative (species abundance) recovery of taxa in simple mock communities by long-amplicon sequencing with that based on short read Illumina amplicon sequencing of the 18SrDNA.

Overall, we demonstrate that long rDNA amplification and sequencing on the MinION platform is a straightforward, cost effective, and universal approach for eukaryote DNA barcoding. It combines robust taxonomic assignment power with high phylogenetic resolution and will enable future analyses of taxonomic and phylogenetic diversity across wide taxonomic scales.

Data Description and Analyses

DNA extraction, PCR and library preparation

We analyzed 114 specimens of eukaryotes including 17 insect and 42 spider species, two annelid and nine plant species (Supplementary Table 1). Some feeder insects and the annelids were purchased at a pet store. The remaining specimens were collected in oak forest on the University of California Berkeley's campus or in native rainforests of the Hawaiian Archipelago (under the Hawaii DLNR permit: FHM14-349). We particularly focused our arthropod sampling on spiders, which are ubiquitous and essential predators in all terrestrial ecosystems. Recent phylogenomic work [35] provided us with a solid baseline to test the efficiency of rDNA amplicons for phylogenetic and taxonomic purposes. We included a taxonomically diverse collection of 16 spider families from the Araneoidea, the RTA clade and a haplogyne outgroup species. Within spiders, we additionally focused on the genus *Tetragnatha*, which has undergone a striking adaptive radiation on Hawaii.

DNA was extracted from each sample using the Qiagen Archivepure kit (Qiagen, Valencia, CA, USA) according to the manufacturer's protocols. The DNA integrity was checked on an agarose gel. Only samples with high DNA integrity were used for the following PCRs. All DNA extracts were quantified using a Qubit fluorometer using the high sensitivity dsDNA assay (Thermo Fisher, Waltham, MA, USA) and diluted to concentrations of 20 ng/µl. We designed a primer pair of each 27 bases to amplify a ~4,000 bp fragment of the ribosomal DNA, including partial 18S and 28S as well as full ITS1, 5.8S and ITS2 sequences (18S_F4 GGCTACCACATCYAARGAAGGCAGCAG and 28S_R8 TCGGCAGGTGAGTYGTTRCACAYTCCT). The primers were designed using alignments of partial 18S and 28S sequences of ~1,000 species of eukaryotes, with a focus on animals

(Supplementary Fig. 1). The primers targeted highly conserved regions across all analyzed taxa. Degenerate sites were incorporated to account for variation. We aimed for high annealing temperatures (65-70°C) to impose stringent amplification. These were calculated using the NEB Tm Calculator (https://tmcalculator.neb.com/#!/main).

To index every PCR amplicon separately, we used a dual indexing strategy with each primer carrying a unique 15 bp index sequence at its 5'-tail. Index sequences were designed using Barcode Generator (http://comailab.genomecenter.ucdavis.edu/index.php/Barcode_generator) with a minimum distance of 10 bases between each index. A total of 15 forward and 16 reverse indexes were designed. Every sample was amplified separately using the Q5 Hot Start High-Fidelity 2X Master Mix (NEB, Ipswitch, MA, USA) in 15 µl reactions, at 68°C annealing temperature, with 35 PCR cycles and using 50 ng of template DNA per PCR. All PCR products were quantified on an agarose gel, based on band intensity on the gel, using the Gel Doc XR System with the Quantity One software (Bio-Rad, CA, USA) and then pooled.

100 µl of the final pool were cleaned from residual primers by 0.75 X AMpure Beads XP (Beckman Coulter, Brea, CA, USA). DNA library preparation was carried out according to the 1D PCR barcoding amplicons SQK- LSK108 protocol (Oxford Nanopore Technologies, Oxford, UK). Barcoded DNA products were pooled with 5 µl of DNA CS (a positive control provided by ONT) and an end-repair was performed (NEB-Next Ultra II End-prep reaction buffer and enzyme mix), then purified using AMPure XP beads. Adapter ligation and tethering was carried out with 20 µl Adapter Mix and 50 µl of NEB Blunt/TA ligation Master Mix. The adapter-ligated DNA library was then purified with AMPure beads XP, followed by the addition of Adapter Bead binding buffer, and finally eluted in 15 µl of Elution Buffer. Each R9 flow cell was primed with 1000 µl of a mixture of Fuel Mix and nuclease-free water. Twelve µl of the amplicon library were diluted in 75 µL of running buffer with 35 µL RBF, 25.5 uL LLB, and 2.5 µL nuclease-free water and then added to

the flow cell via the SpotON sample port. The “NC_48Hr_sequencing_FLO-MIN107_SQK-LSK108_plus_Basecaller.py” protocol was initiated using the MinION control software, MinKNOW.

Field trial in the Amazon rainforest

A field trial using the protocol described above was conducted in Tambopata, Peru, at the Refugio Amazonas lodge (-12.874797, -69.409669) using two butterflies, a grasshopper, one mosquito, unidentified insect eggs and two plant specimens. Collection permits in Peru were issued by the Servicio Nacional Forestal y de Fauna Silvestre, 403-2016-SERFOR-DGGSPFFS, 019-2017-SERFOR-DGGSPFFS. DNA extractions, PCR and library preparation were performed in the field using a highly miniaturized laboratory consisting of portable equipment. Equipment used for sequencing under remote tropical conditions is described in further detail in Pomerantz, et al. [25]. DNA extractions were carried out with the Quick-DNA Miniprep Plus Kit (Zymo Research, Irvine, CA, USA) according to manufacturer's protocol. PCRs were performed using the Q5 Hot Start High-Fidelity 2X Master Mix and the same primers as described above. A battery operated portable miniPCR device (Amplyus, Cambridge, MA, USA) was used to run PCRs. The sequencing on the MinION was carried out as described above.

Bioinformatics

Raw data processing and consensus calling

The fastq files generated by the ONT software MinKNOW were de-multiplexed using MiniBar (see description below), with index edit distances of 2, 3, and 4 and a primer edit distance of 11. Next, the reads were filtered for quality (>13) and size (>3kb) using Nanofilt [36](https://github.com/wdecoster/nanofilt). Individual consensus sequences were created using Allele Wrangler (https://github.com/transplantation-immunology/allele-wrangler/) for demultiplexed fastq files with a minimum coverage of 30. Error correction was performed using RACON [37] (https://github.com/isovic/racon). To do so, we first mapped all the reads back to the consensus using minimap (https://github.com/lh3/minimap2). We performed two cycles of running minimap and RACON. Final consensus sequences were compared against the NCBI database using BLASTn to check if the taxonomic assignment was correct.

We performed multiple tests to validate and optimize the consensus accuracy of long-amplicon barcode sequences. To comparatively assess the accuracy, we used consensus sequences of short 18S and 28SrDNA amplicons, which were previously generated using Illumina amplicon sequencing for the 47 analyzed Hawaiian *Tetragnatha* specimens (Kennedy unpublished data). These sequences were aligned with the respective stretches of our nanopore consensus sequences using ClustalW in MEGA [38]. All alignments were then visually inspected and edited manually, where necessary. Pairwise distances between Illumina and nanopore consensus were calculated in MEGA.

To measure consensus accuracy over the whole ribosomal amplicon, we utilized genome skimming data [39] for six Hawaiian *Peperomia* plant species (Lim et al unpublished data). 150 bp paired-end TruSeq gDNA shotgun libraries for the six *Peperomia* samples were sequenced on a single HiSeq v4000 lane (Illumina, San Diego, CA, USA). The resulting paired-end reads were trimmed and filtered using Trimmomatic v0.36 [40] and mapped to their respective nanopore consensus sequences using bowtie2 [41] under default parameter values and allowing for minimum and maximum fragment size of 200 and 700 bases respectively. Mapping coverage of Illumina reads to nanopore consensus sequences ranged between 150 - 600 X with a mean of ~ 300 X across all six samples. We called Illumina read based consensus sequences for each *Peperomia* species using bcftools [42], and aligned them with the previously generated nanopore consensus sequences. Pairwise genetic distances were then calculated in MEGA as described above. We performed two independent distance calculations: 1) excluding indels, i.e. only using nucleotide substitutions to estimate genetic distances, and 2) including indels as additional characters.

Our demultiplexing software allows flexible edit distances to identify forward and reverse indexes from Nanopore reads. Due to the high raw read error rate, too large edit distances could lead to crossover between samples during demultiplexing. This crossover could possibly affect the accuracy of the called consensus sequence. On the other hand, too stringent edit distances may result in very large read dropout. Assuming an average error rate of 12-22 %, 3 bp of our 15 bp indexes should maximize sequence recovery. We thus tested index edit distances of 2, 3, and 4 bp in MiniBar for the six *Peperomia* specimens for which we had generated Illumina based consensus sequences. We counted the number of recovered reads and estimated the accuracy of the resulting consensus sequence based on the relevant edit distances as described above.

A recent study [25] showed that accurate consensus sequences from nanopore data can be generated using only 30x coverage. We tested 18 different assembly coverages from 10 to 800 sequences for a *Peperomia* species, to explore optimal assembly coverage. We randomly subsampled the quality filtered and demultiplexed fastq file for the relevant specimen 10 times for each tested assembly coverage. Consensus sequences were then assembled and genetic distances to the Illumina consensus calculated as described above.

Phylogenetic and taxonomic analysis

We carried out phylogenetic analyses on two hierarchical levels. First, we built a phylogeny for all higher eukaryote taxa in our dataset, which included plants, animals and fungi. Second, we took a closer look into the phylogeny of spiders. The resulting quality checked consensus sequences of all taxa were aligned using ClustalW in MEGA. The alignments were visually inspected and manually edited. The exact position of gene sequences was identified by downloading full length 18S, 5.8S and 28S sequences from GenBank and then aligning them against the amplicons. Due to the deep divergence in the eukaryote data set, the highly variable ITS sequences could not be aligned and were excluded. For the analyses of spiders, we retained both ITS sequences and aligned the whole rDNA amplicon. Appropriate models of sequence evolution for each gene fragment of the rDNA cluster were identified using PartitionFinder [43]. Phylogenies were built using MrBayes [44], with 4 heated chains, a chain length of 1,100,000, subsampling every 200 generations and a burnin length of 100,000.

Focusing on the endemic Hawaiian *Tetragnatha* species, we also tested the utility of the ribosomal cluster for taxonomic identification, as we also had COI barcodes available for these species. Our dataset contained ribosomal DNA sequences for 47 specimens in 16 species, which had been identified morphologically before barcoding. We calculated pairwise genetic distances between and within all species for the whole ribosomal cluster and for each separate gene region of the rDNA cluster using MEGA. As the 18S and 5.8S did not yield any species level resolution within Hawaiian *Tetragnatha*, they were not analyzed separately. To compare the taxonomic resolution of the ribosomal cluster with that of the commonly used mitochondrial COI, we calculated inter- and intraspecific distances for an alignment of 418 bp of the COI barcode region for the same spider specimens (Kennedy et al. unpublished data). We performed a Mantel test using the R package ade4 [45] to test for a significant correlation between COI and ribosomal DNA based distances. A comparison of intraspecific and interspecific distances for mitochondrial COI and ribosomal DNA also allowed us to test for the presence of a barcode gap.

Nanopore based arthropod metabarcoding

To test for the possibility of estimating arthropod community composition from Nanopore sequencing, we prepared four mock communities of different amounts of DNA extracts from 9 species of arthropods from different orders (see Supplementary Table 2). It should be noted that with representatives of nine different orders, these community samples were highly simplified and arenot necessarily representive of a natural arthropod community. Due to the high error rate of individual reads, we did not know if, and how, the MinION's high error rate would affect taxonomic assignment, hence we decided to limit our current analysis to these simplified communities.

The samples were amplified using the Q5 High Fidelity Mastermix as described above at 68 °C annealing temperature and 35 PCR cycles. We additionally tested two variations of PCR conditions: 1) we either reduced the annealing temperature to 63 °C or, 2) reduced the PCR cycle number to 25.

In order to compare our results with those from an optimized Illumina short read protocol, we amplified all samples for a ~300 bp fragment of the 18S rDNA using the primer pair 18S2F/18S4R [46]. Amplification and library preparation were performed as described in [47] using Qiagen Multiplex PCR kits. The 18S amplicon pools were sequenced on an Illumina MiSeq using V3 chemistry and 2 x 300 bp reads. Sequence quality filtering, read merging and primer trimming were performed as described in [34].

A library of 18S sequences for all included arthropod species (from [34]) was used as a reference database to identify the recovered sequences using BLASTn [48], with a minimum e-value of 10^{-4} and a minimum overlap of 95 %. Despite the high raw error rate of nanopore reads, taxonomic status of sequences could be assigned using BLAST, as our pools contained members of highly divergent orders. We compared the qualitative (number of species) and quantitative (abundance of species) recovery of taxa from the communities by nanopore long-amplicon and Illumina short read data. To estimate the recovery of taxon abundances, we calculated a fold change between input DNA amount and recovered reads for each taxon and mock community. A fold change of zero corresponded to a 1:1 association of taxon abundance and read count, while positive or negative values indicated higher or lower read counts than the taxon's actual abundance.

MiniBar

We created a de-multiplexing software, called MiniBar. It allows customization of search parameters to account for the high read error rates and has built-in awareness of the dual barcode and primer pairs flanking the sequences. MiniBar takes as input a tab-delimited barcode file and a sequence file in either fasta or fastq format. The barcode file contains, at a minimum, sample name, forward barcode, forward primer, reverse barcode, and reverse primer for each of the samples potentially in the sequence file. The software searches for barcodes and for a primer, each permitting a user defined number of errors, an error being a mismatch or indel. Error count to determine a match can either be a percentage of each of their lengths or can be separately specified for barcode and primer as a maximum edit distance [49]. Output options permit saving each sample in its own file or all samples in a single file, with the sample names in the fasta or fastq headers. The found barcode primer pairs can be trimmed from the sequence or can remain in the sequence distinguished by case or color. MiniBar, written in Python 2.7, can also run in Python 3 and has the single dependency of the Edlib library module for edit distance measured approximate search [50]. MiniBar can be found at https://github.com/calacademy-research/minibar along with test data.

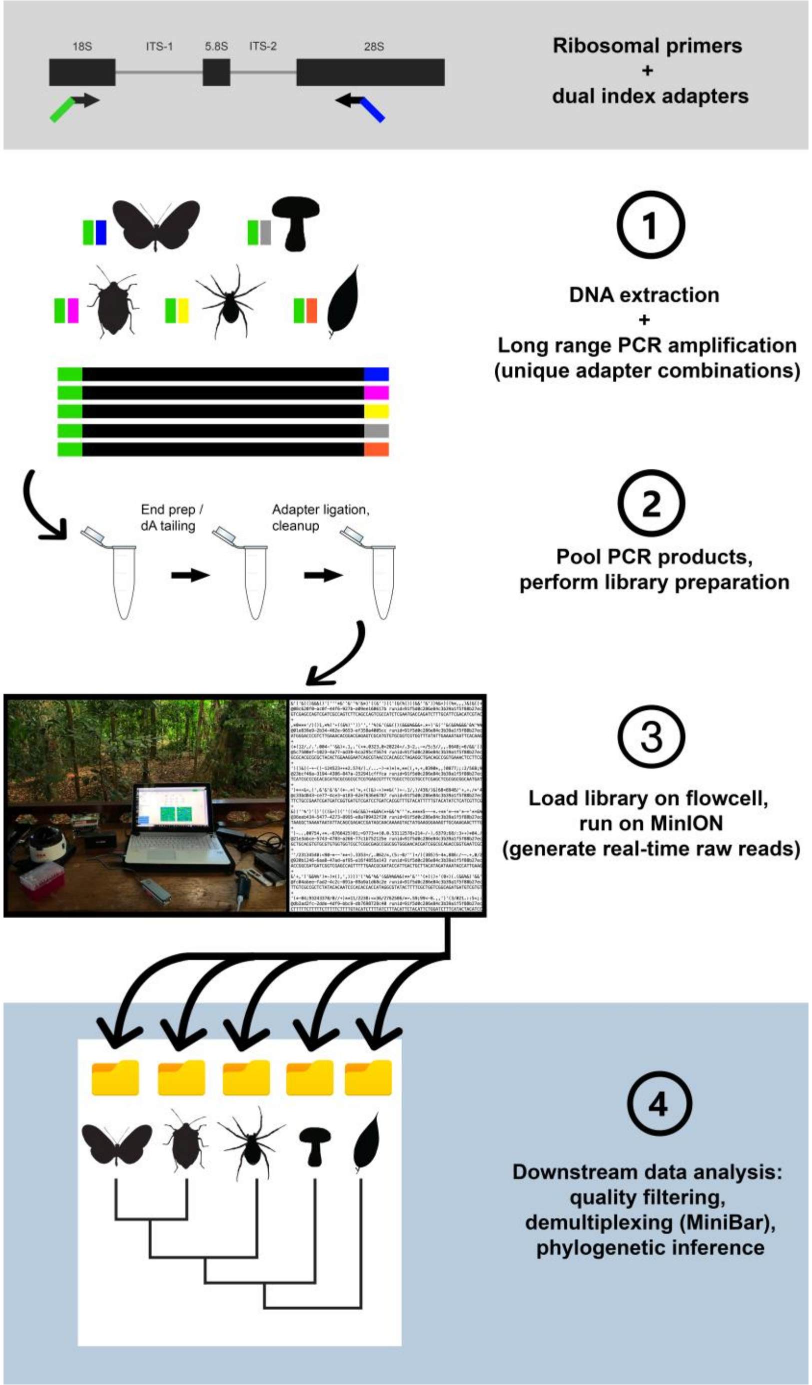

Figure 1. Workflow for the design, amplification, and sequencing of the ribosomal DNA cluster.

Results

Sequencing, specimen recovery and consensus quality

After quality filtering and trimming, our nanopore run yielded 245,433 reads. We tested edit distances of two, three and four bases in MiniBar to demultiplex samples. Increasing edit distances led to a significant increase in read numbers assigned to index combinations (Pairwise Wilcoxon Test, FDR-corrected *P*-value < 0.05). On average, we found 355 reads per specimen for an edit distance of two, 647 for a distance of three and 1,051 for a distance of four. However, at an edit distance of four, we found a considerable increase of wrongly assigned samples. A relatively high number of index combinations were incorrectly assigned at the highest edit distance. Demultiplexed samples were then mixtures of different taxa, which probably affected consensus accuracy. Using Illumina shotgun sequencing-derived consensus sequences of rDNA from six *Peperomia* plants, we tested the accuracy of the nanopore consensus assemblies based on the three edit distances (Fig. 2). While a distance of four yielded the highest number of assigned reads (1,785 on average), it also led to slightly more inaccurate consensus assemblies, with an average distance of 2.072 % to Illumina based consensus sequences. We found a significant increase of consensus accuracy (Pairwise Wilcoxon Test, FDR corrected *P* < 0.05) for edit distances of two (0.165 % average distance) and three (0.187 % average distance). Despite significant differences in assigned reads (1,091 vs. 637 reads on average), there was not a significant difference in consensus accuracy of edit distances of two versus three bases (Pairwise Wilcoxon Test, FDR corrected *P* > 0.05).

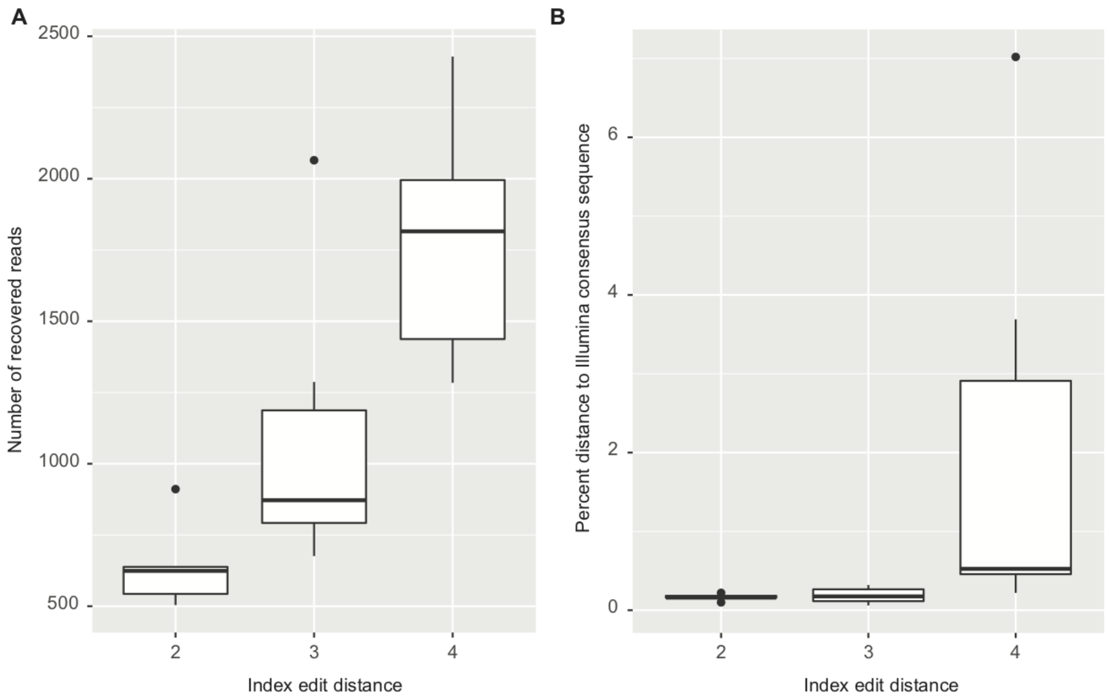

Figure 2: Comparison of recovered sequences and consensus accuracy for different index edit distances in Minibar. A) Number of recovered reads for six *Peperomia* species at index edit distances of two, three and four. B) Pairwise sequence divergence between Illumina and Nanopore based consensus sequences of the same six *Peperomia* specimens at the same index edit distances.

We chose a minimum coverage of 30 (see below) and an edit distance of two (which showed the smallest final consensus error rate) for all subsequent analyses. BLAST analyses suggested a correct taxonomic assignment for the majority of these consensus sequences. However, we found some notable exceptions. For two insect specimens, we amplified mite rDNA sequences. One of these specimens was *Drosophila hydei*, with the mite taxon being a well known phoretic associated with arthropods. A different mite taxon was assembled from an unidentified termite species. A species of isopod and a neuropteran yielded fungal sequences after assembly. The larva of a butterfly and a feeder mealworm (*Zophobas morio*) generated consensus sequences for plants. In most of these samples, the targeted arthropod species was either extremely underrepresented among the read populations or completely absent.

A comparison of our consensus sequences for 47 Hawaiian specimens of the spider genus *Tetragnatha* with short Illumina amplicon sequencing-derived 18S and 28S rDNA sequences suggests a very high consensus accuracy. Except for a single specimen, with a single substitution error, all nanopore based consensus sequences were completely identical to the Illumina based consensus. However, the corresponding 18S and 28S fragments did not contain long stretches of homopolymer sequences, where nanopore raw read errors are known to accumulate [51]. Despite containing several homopolymers, the nanopore derived *Peperomia* consensus sequences were highly accurate (Supplementary Fig. 2). Including gaps in the alignment, an average distance of 0.165 % to Illumina based consensus sequences was found. Errors were clustered in indel regions (Supplementary Fig. 3). After excluding gaps, the average distance dropped to 0.102 %.

We found only a small effect of sequence coverage on consensus assembly accuracy (Supplementary Fig. 4). Even at 10-fold coverage, a low average distance of 0.257% to Illumina consensus sequences was observed. However, at 20-fold coverage, the average distance significantly decreased to 0.128 % (Pairwise Wilcoxon Test, FDR corrected $P < 0.05$). A slight, but not significant, decrease of distance was observed with increasing coverage, with optimal consensus accuracy at 300-fold coverage (0.031 % distance). At coverages larger than 300, the consensus accuracy slightly decreased (average distance of 0.103 % at 800 X coverage), but this change was not significant.

The length of the rDNA amplicon was quite variable between taxa. Arachnids, hexapods and magnoliopsid plant specimens all showed a significantly different amplicon lengths (Pairwise Wilcoxon Test, FDR corrected $P < 0.05$). The length difference was found for the actual gene sequences (18S, 5.8S, 28S: plants: 2781 ± 4.96; hexapods: 3154 ± 50.35; arachnids: 3047 ± 10.77 %; Supplementary Fig. 5A) as well as including the ITS sequences (plants: 3243 ± 11.78; hexapods: 4192 ± 498.05; arachnids: 3644 ± 129.07, Supplementary Fig. 5B). While most spiders showed very stable length distributions for the rDNA amplicon length (average length ± standard deviation across all Araneae: 3,629 bp ± 81), several hexapod orders had rDNA sequences of more variable length (Coleoptera: 4,488 bp ± 352; Lepidoptera: 4363 bp ± 603).

In contrast to the variable length of the rDNA cluster, we found a very stable GC content across the whole taxonomic spectrum (46.75 ± 2.67 % across all taxa). GC content of magnoliopsid plants, hexapods and arachnids was highly similar (plants: 46.01 ± 1.66 %; hexapods: 46.67 ± 3.73 %; arachnids: 46.93 ± 2.47 %) (Supplementary Fig 5c).

Phylogenetic reconstruction

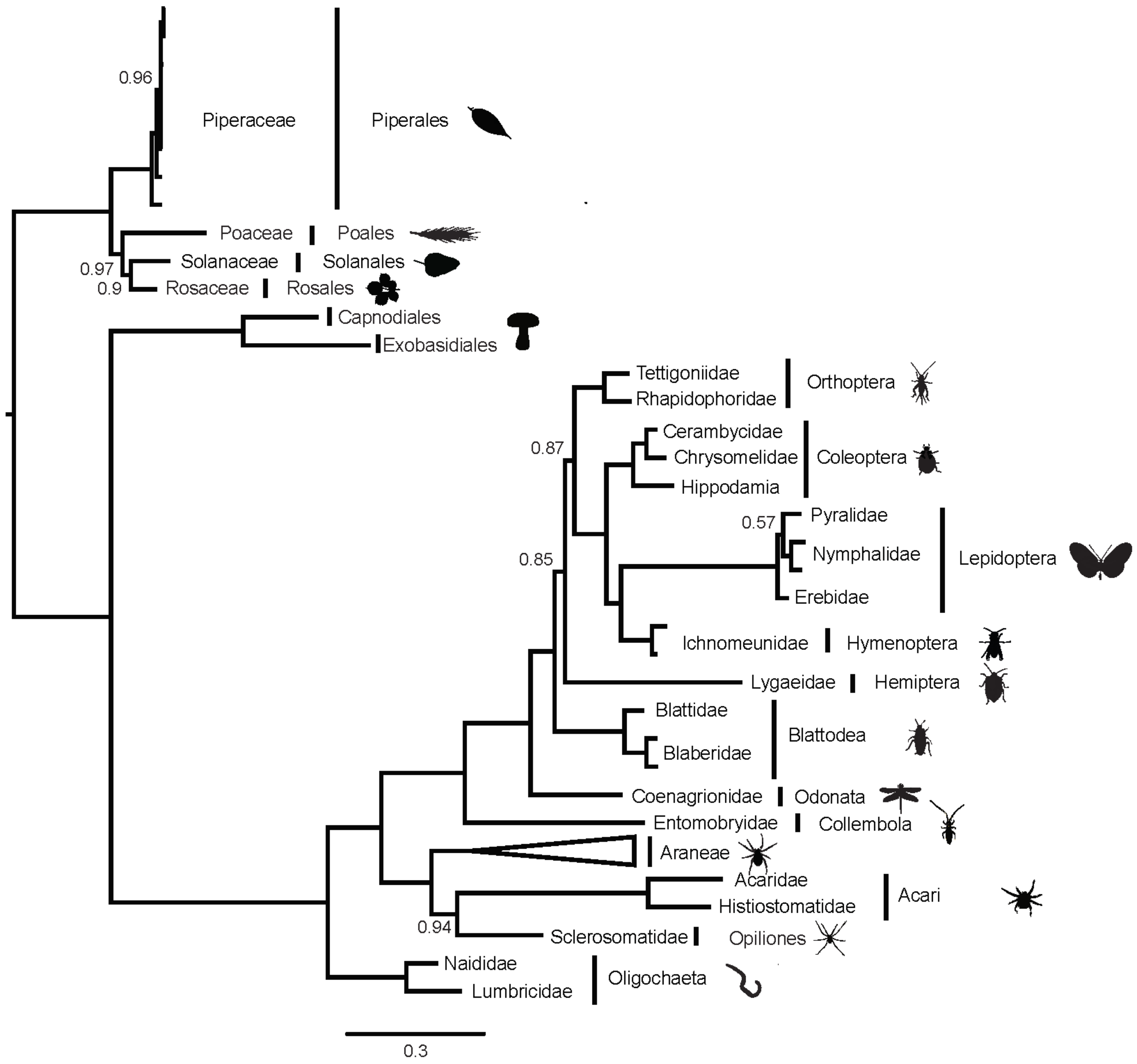

Figure 3 Bayesian consensus phylogeny based on a 3,656 bp alignment of 18S, 5.8S and 28S sequences of 117 animal, fungal and plant taxa. The phylogeny is rooted using plants as outgroup. Branches are annotated with family and order level taxonomy. The Araneae clade of 83 specimens is collapsed. Only posterior probability values below 1 are displayed.

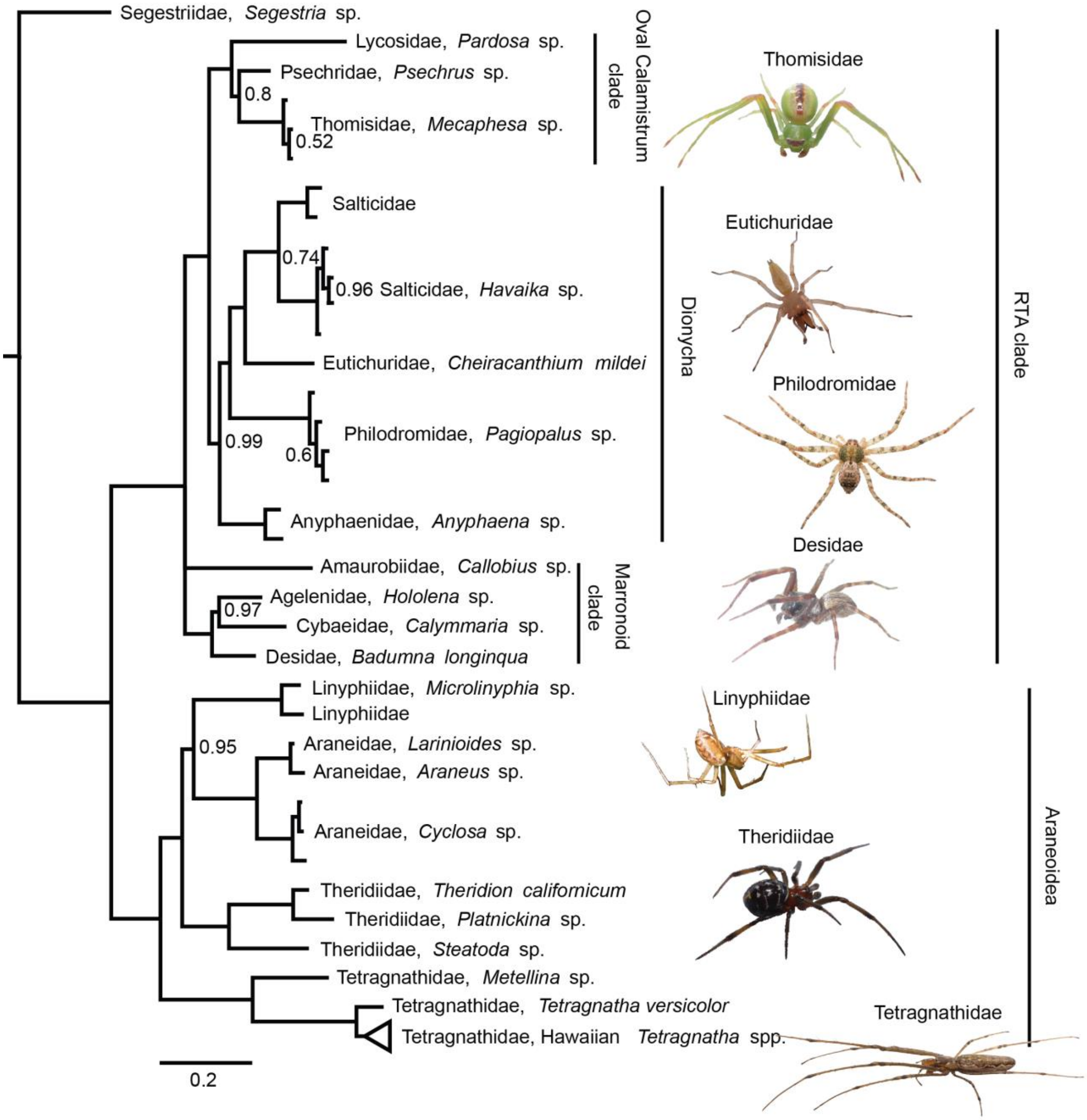
Segestriidae, *Segestria* sp.
Lycosidae, *Pardosa* sp.
Psechridae, *Psechrus* sp.
0.8
Thomisidae, *Mecaphesa* sp.
0.52
Oval Calamistrum clade
Thomisidae
Salticidae
0.74
0.96 Salticidae, *Havaika* sp.
Eutichuridae, *Cheiracanthium mildei*
0.99
Philodromidae, *Pagiopalus* sp.
0.6
Anyphaenidae, *Anyphaena* sp.
Dionycha
Eutichuridae
Philodromidae
Amaurobiidae, *Callobius* sp.
Agelenidae, *Hololena* sp.
0.97
Cybaeidae, *Calymmaria* sp.
Desidae, *Badumna longinqua*
Marronoid clade
Desidae
RTA clade
Linyphiidae, *Microlinyphia* sp.
Linyphiidae
0.95
Araneidae, *Larinioides* sp.
Araneidae, *Araneus* sp.
Araneidae, *Cyclosa* sp.
Theridiidae, *Theridion californicum*
Theridiidae, *Platnickina* sp.
Theridiidae, *Steatoda* sp.
Tetragnathidae, *Metellina* sp.
Tetragnathidae, *Tetragnatha versicolor*
Tetragnathidae, Hawaiian *Tetragnatha* spp.
0.2
Linyphiidae
Theridiidae
Tetragnathidae
Araneoidea

Figure 4. Bayesian consensus phylogeny of 83 spiders in 16 families, based on a 4,214 bp alignment of 18S, ITS1, 5.8S, ITS2 and 28S. The phylogeny is rooted using the basal haplogyne *Segestria* sp. The clade containing Hawaiian members of the genus *Tetragnatha* is collapsed (the uncolapsed clade is shown in Fig. 5). Only posterior probability values below 1 are displayed.

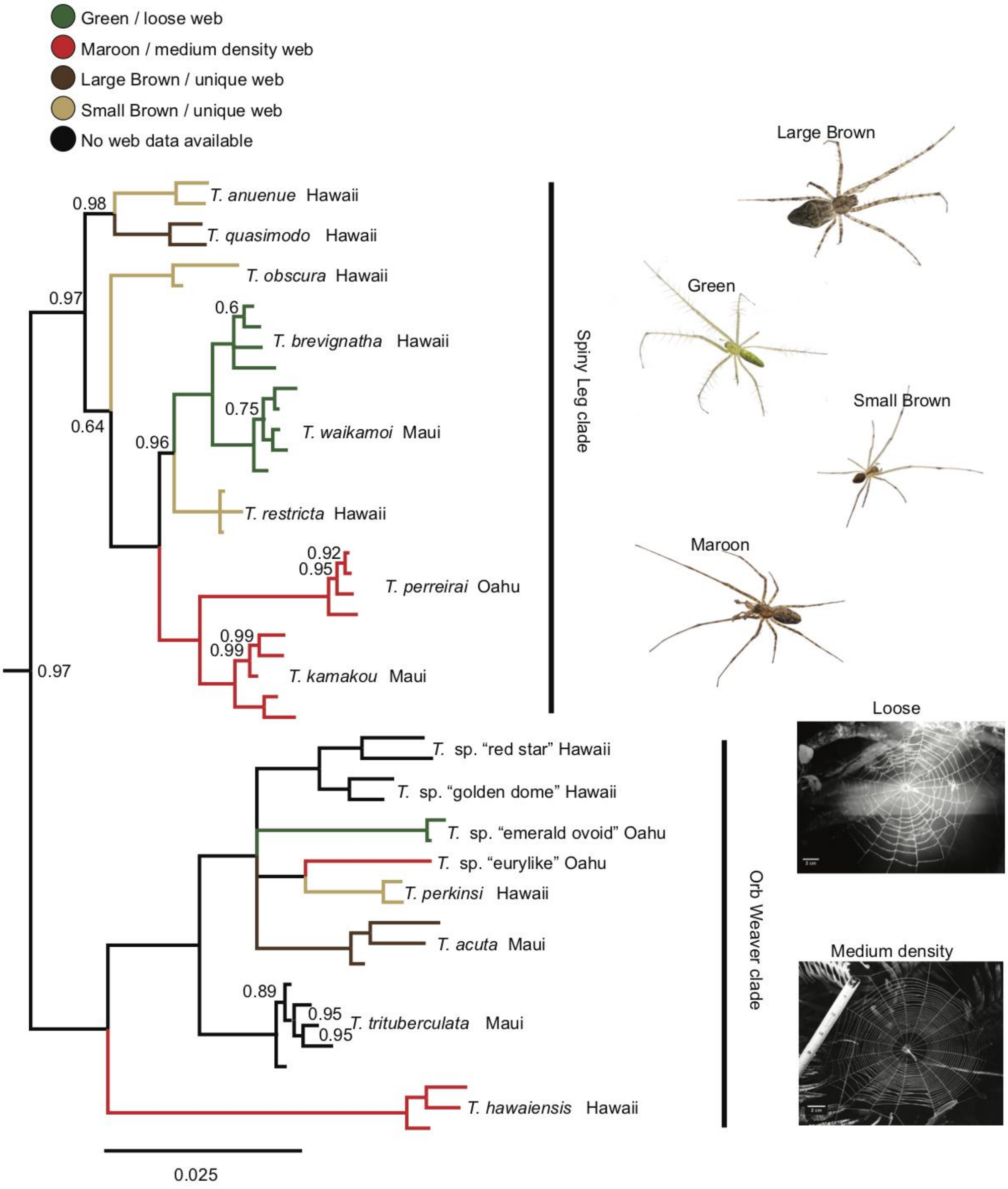

Figure 5. Section of the same phylogeny as Fig. 4, with expansion of the clade of 16

Hawaiian *Tetragnatha* species. Different “Spiny Leg” ecomorphs and web architectures are indicated by branch coloration. Only posterior probability values below 1 are displayed.

We generated an alignment of 3,656 bp for 117 concatenated 18S, 5.8S and 28S sequences of plants, fungi, annelids and arthropods. Our phylogeny was well supported (most posterior support values equal one; Fig. 3). A basal split separated plants from fungi and animals. Within plants, the genus *Peperomia* was recovered as monophyletic. Fungi formed the sister group of animals. Within animals, annelids formed a separate clade from arthropods. Arthropods separated into arachnids and hexapods. Each separate arthropod order formed well supported groups. The hexapod phylogeny generally resembled that found in latest phylogenomic work [52]. The Collembola species *Salina* sp. formed the base to the insect tree, followed by the odonate *Argia* sp. A higher branch led to Blattodea, Hemiptera and Orthoptera. However, the support values for the relationships between these three orders were comparatively low (~ 0.85). Finally, holometabolan insects (Hymenoptera, Coleoptera and Lepidoptera) were recovered as monophyletic. The two Acari species, together with Opiliones, formed the sister clade to the monophyletic Araneae clade.

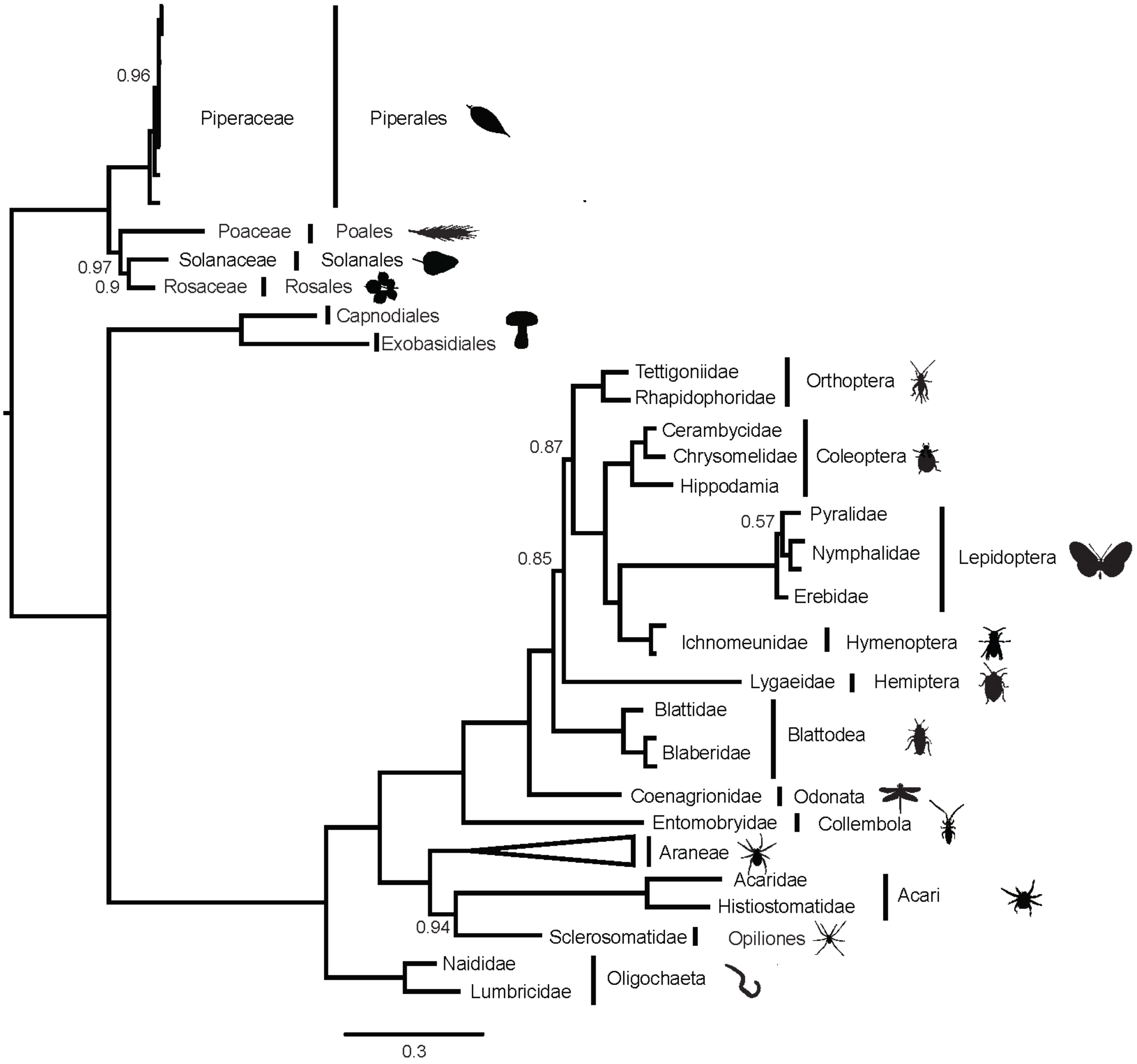

Figure 3 Bayesian consensus phylogeny based on a 3,656 bp alignment of 18S, 5.8S and 28S sequences of 117 animal, fungal and plant taxa. The phylogeny is rooted using plants as outgroup. Branches are annotated with family and order level taxonomy. The Araneae clade of 83 specimens is collapsed. Only posterior probability values below 1 are displayed.

Next, we generated a separate alignment of rDNA sequences for 83 spiders, including both ITS regions (totaling 4,214 bp). The spider phylogeny was also strongly supported (Fig. 4). Overall, our phylogenetic tree topology agreed with the most recent phylogenetic work of [53] and [35]. With the haplogyne *Segestria* sp. (family Segestriidae) forming the root, we recovered the so-called RTA clade (represented in our dataset by families Agelenidae, Amaurobiidae, Anyphaenidae, Cybaeidae, Desidae, Eutichuridae, Lycosidae, Philodromidae, Psechridae, Salticidae and Thomisidae) and the Araneoidea (Araneidae, Linyphiidae, Tetragnathidae, Theridiidae) as two well supported monophyla. Within these clades, all families and genera formed well supported monophyletic groups. Similar to recent studies, we found the Marronoid clade as basal to the rest of the RTA clade; more derived clades were the Oval Calamistrum and the Dionycha clade. Inter-family relationships also closely matched those found in recent work: Lycosidae was basal to the clade formed by Psechridae and Thomisidae; Salticidae was closest to Eutichuridae and Philodromidae, with Anyphaenidae falling basal within Dionycha. Within Araneoidea, our results differed slightly from recent studies in that we recovered Tetragnathidae, rather than Theridiidae, as basal.

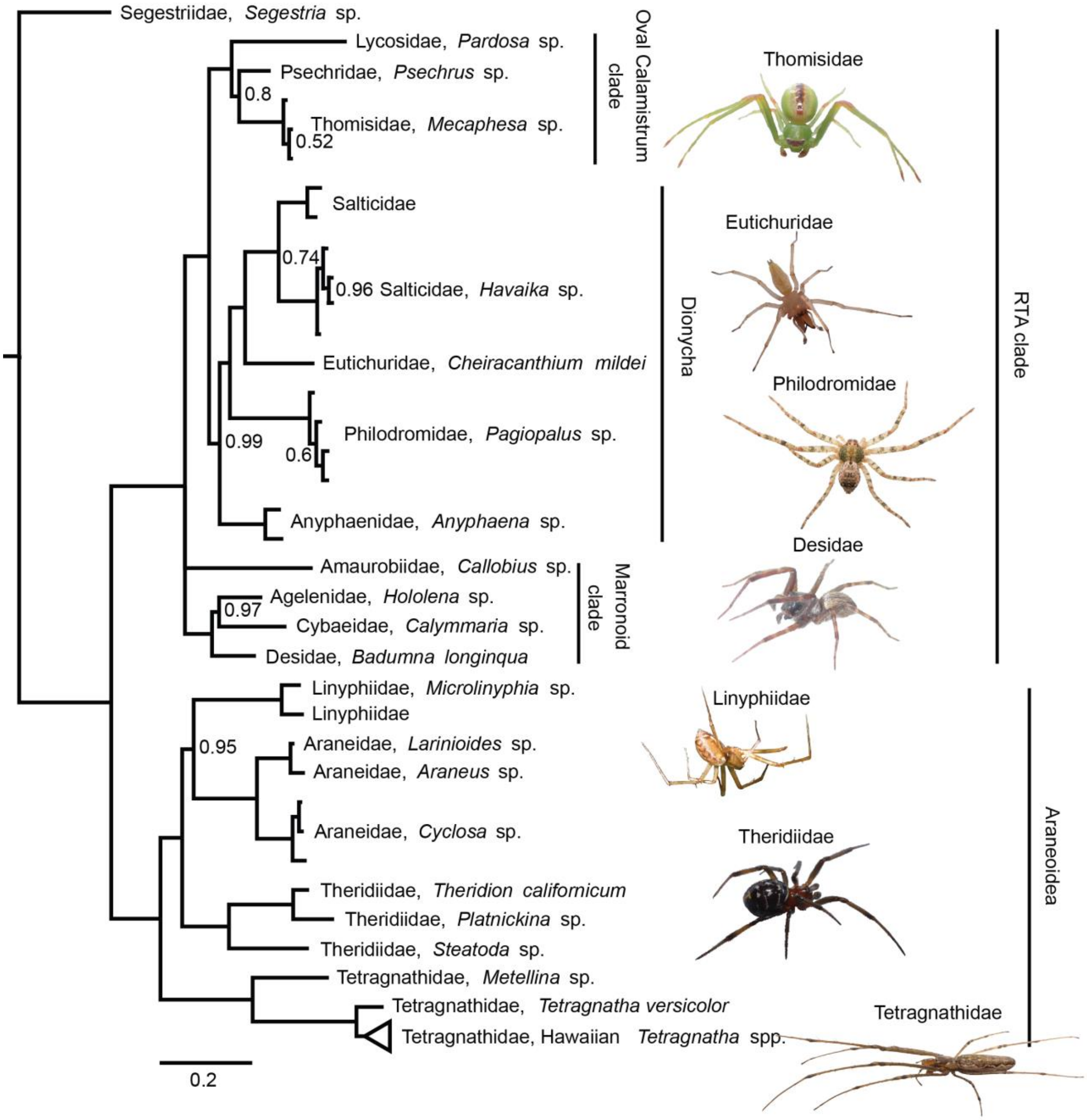
Segestriidae, *Segestria* sp.
Lycosidae, *Pardosa* sp.
Psechridae, *Psechrus* sp.
0.8
Thomisidae, *Mecaphesa* sp.
0.52
Oval Calamistrum clade
Thomisidae
Salticidae
0.74
0.96 Salticidae, *Havaika* sp.
Eutichuridae
Eutichuridae, *Cheiracanthium mildei*
Dionycha
Philodromidae
0.99
0.6
Philodromidae, *Pagiopalus* sp.
Anyphaenidae, *Anyphaena* sp.
RTA clade
Desidae
Amaurobiidae, *Callobius* sp.
Agelenidae, *Hololena* sp.
0.97
Cybaeidae, *Calymmaria* sp.
Desidae, *Badumna longinqua*
Marronoid clade
Linyphiidae, *Microlinyphia* sp.
Linyphiidae
Linyphiidae
0.95
Araneidae, *Larinioides* sp.
Araneidae, *Araneus* sp.
Araneidae, *Cyclosa* sp.
Theridiidae
Theridiidae, *Theridion californicum*
Theridiidae, *Platnickina* sp.
Theridiidae, *Steatoda* sp.
Araneoidea
Tetragnathidae, *Metellina* sp.
Tetragnathidae, *Tetragnatha versicolor*
Tetragnathidae, Hawaiian *Tetragnatha* spp.
Tetragnathidae
0.2

Figure 4. Bayesian consensus phylogeny of 83 spiders in 16 families, based on a 4,214 bp alignment of 18S, ITS1, 5.8S, ITS2 and 28S. The phylogeny is rooted using the basal haplogyne *Segestria* sp. The clade containing Hawaiian members of the genus *Tetragnatha* is collapsed (the uncollapsed clade is shown in Fig. 5). Only posterior probability values below 1 are displayed.

We recovered Hawaiian *Tetragnatha* as a well supported monophyletic clade within the Tetragnathidae. We found two main clades of Hawaiian *Tetragnatha* (Fig. 5), both of which have been supported by earlier work [54-57]: the orb weaving clade and the "Spiny Leg clade" of actively hunting species. All *Tetragnatha* species formed monophyletic groups, and the relationships among different species were mostly well supported. Within the Spiny Leg clade, species fell into one of four ecomorphs, each of which is associated with a particular substrate type [58]: "large brown" (*T. quasimodo*) with tree bark, "small brown" (*T. anuenue*, *T. obscura* and *T. restricta*) with twigs, "green" (*T. brevignatha* and *T. waikamoi*) with green leaves, and "maroon" (*T. perreirai* and *T. kamakou*) with lichen. While green and maroon ecomorphs clustered phylogenetically, small brown species appeared in three separate clades on the tree. Within the orb weaving clade, *T. hawaiensis*, a generalist species which occurs on all of the Hawaiian Islands, fell basal. The characteristic web structures of some of these species have been documented [59, 60]. We found a pattern of apparent convergence in web structure for some species. *T.* sp. "emerald ovoid" spins a loose web with widely spaced rows of capture silk. *T. hawaiensis* and *T.* sp. "eurylike," which are distant relatives within the Hawaiian *Tetragnatha* clade, both spin webs of medium silk density, i.e. with more rows of capture silk per unit area than *T.* sp. "emerald ovoid." *T. perkinsi* and *T. acuta* each spin a web structure that is not comparable in its silk density or size to any other known *Tetragnatha* species in this group [60], and are thus classified as "unique".

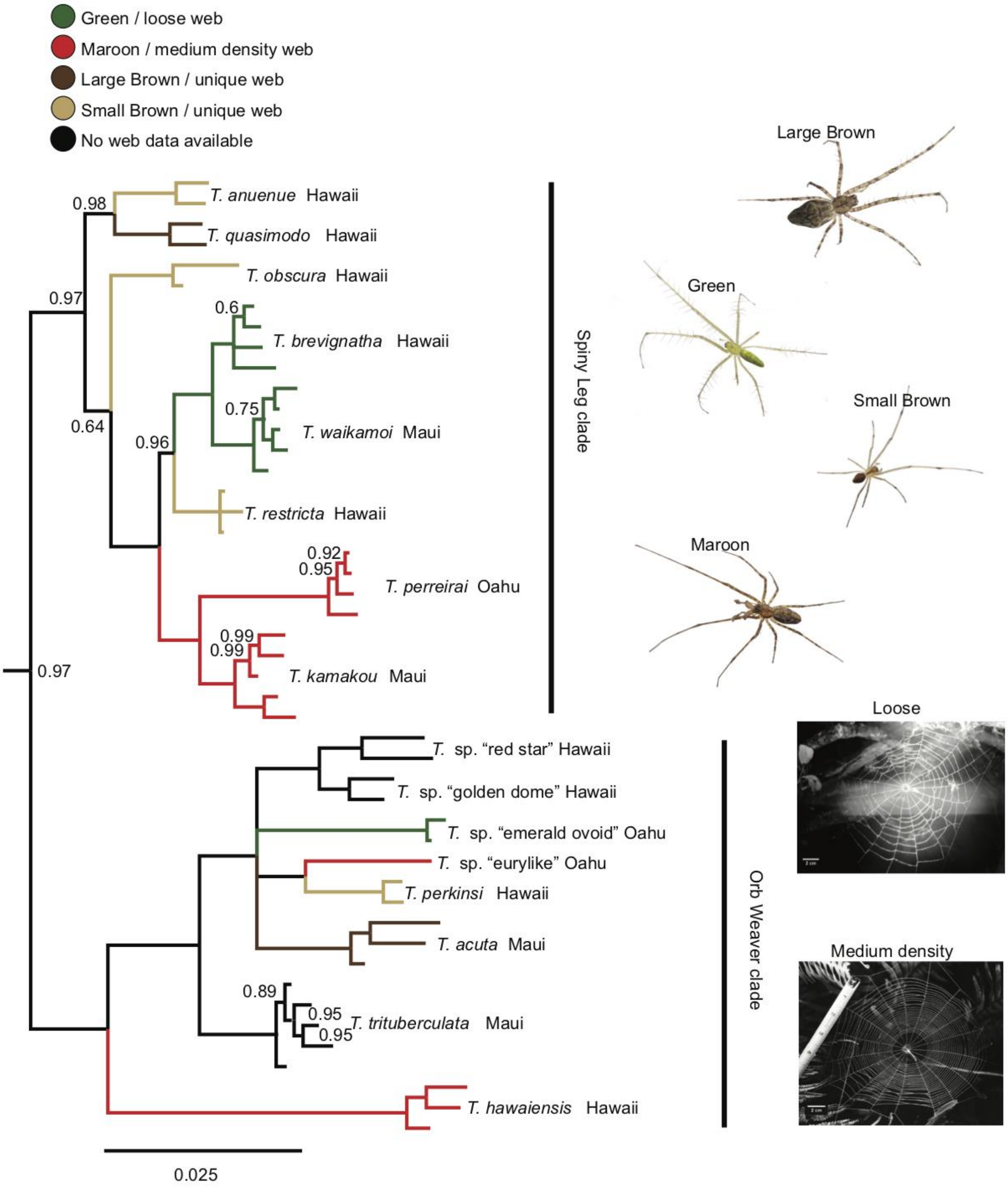

Figure 5. Section of the same phylogeny as Fig. 4, with expansion of the clade of 16

Hawaiian *Tetragnatha* species. Different “Spiny Leg” ecomorphs and web architectures are indicated by branch coloration. Only posterior probability values below 1 are displayed.

Taxonomic resolution

Our inferred genetic distances for rDNA sequences within and between Hawaiian *Tetragnatha* species were significantly correlated to those found for COI sequences of the same taxa (R^2 = 0.70, $P < 0.001$) (Fig. 6a). A Mantel test also suggested highly significant correlation of mitochondrial COI and nuclear rDNA-based distances (Mantel test, 9999 replicates, $P < 0.001$). Hence, the rDNA cluster supported a very similar pattern of genetic differentiation to COI. However, the faster evolutionary rate of COI was reflected in lower distances for the whole rDNA than for COI. Interspecific distances were significantly higher than intraspecific ones for COI and rDNA (Fig 6b,c). No overlap of intra- and interspecific distances was evident for COI, suggesting the presence of a barcode gap. A small overlap of intra- and interspecific distances was evident for the rDNA (Supplementary Table 3). However, this overlap was caused only by a single undescribed species (*T.* sp. “golden dome”) with unclear status, which showed a high intraspecific divergence in rDNA. Further morphological analyses will be necessary to rule out that the included samples do not actually comprise two species. At the same time, the interspecific rDNA distance of the relevant species was higher than its intraspecific distance. The lowest interspecific distance was found for a complex of closely related species from Maui. Like the combined rDNA cluster, genetic distances for different parts of the rDNA cluster all showed significant correlation with COI based distances, when analyzed separately (R^2 28S = 0.57, R^2 ITS1 = 0.68, R^2 ITS2 = 0.56, $P <$ 0.001) (Supplementary Fig. 6). While the 28SrDNA showed considerably lower distances than COI, those for ITS1 and ITS2 were more comparable to COI (Supplementary Fig. 6b-d). Yet, interspecific and intraspecific distances for COI were significantly different from those for any part of the rDNA cluster (Pairwise Wilcoxon Test, FDR corrected $P < 0.05$).

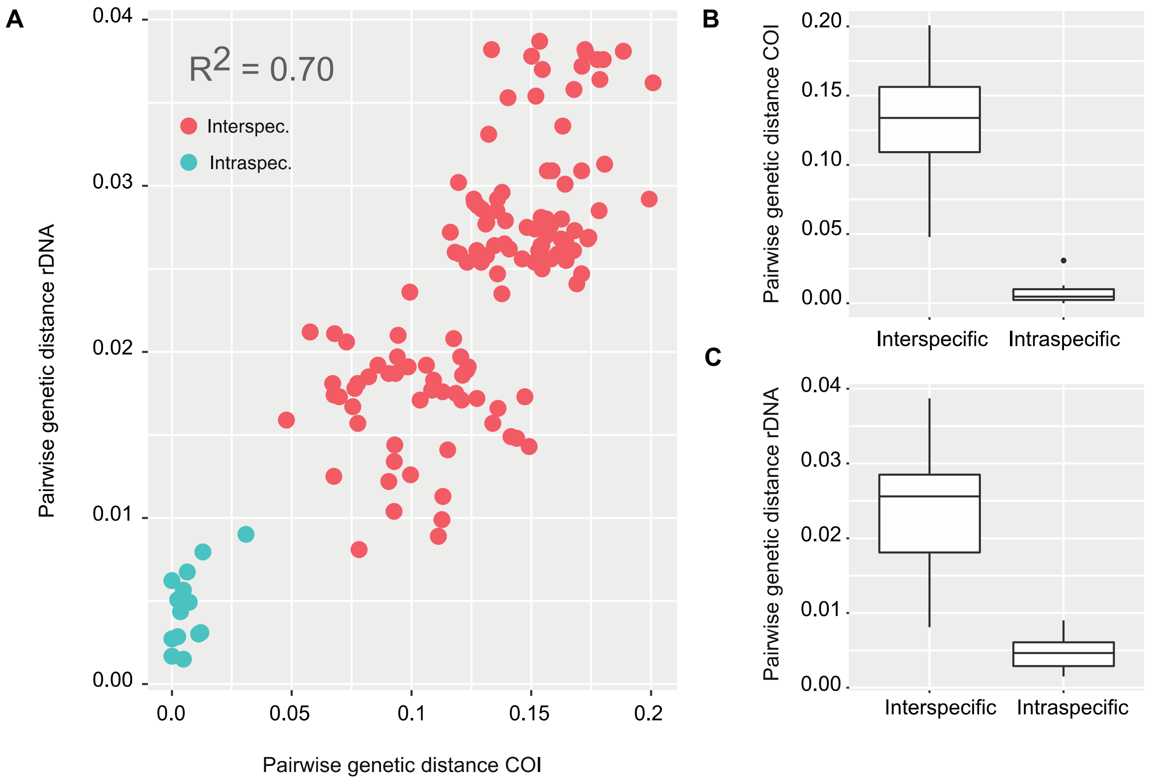

Figure 6 Inter and intraspecific genetic distances for the nuclear rDNA and mitochondrial COI for Hawaiian *Tetragnatha* spiders. A) Correlation of pairwise genetic distance between (red) and within (green) 16 Hawaiian *Tetragnatha* species based on COI and the full rDNA amplicon. B) Interspecific and intraspecific genetic distances for the same spider species based on mitochondrial COI and C) the whole rDNA amplicon.

Field trial in the Amazon rainforest

On March 26, 2018, we set out to test this method and a portable laboratory (as described in Pomerantz, et al. [25]) during an expedition to the Peruvian Amazon at the Refugio Amazonas Lodge (Supplementary Fig. 7). This field site is a “Terra firme” forest in the sector of “Condenado”, approximately two and a half hours by boat up river from the native community of Infierno on the buffer zone of the Tambopata National Reserve. We collected plant and insect material, extracted DNA, amplified the rDNA cluster, and sequenced material on the MinION platform using the MinKNOW offline software (provided by ONT). The first run generated 17,149 reads and the second one 20,167 reads. We generated consensus sequences for five out of the seven analyzed specimens. One plant sample and the grasshopper could not be assembled due to too low read coverage. Moreover, BLAST analysis of the reads assigned to the grasshopper suggested that we had sequenced a mite, instead of the grasshopper DNA. The unidentified insect eggs resulted in a butterfly consensus sequence, possibly a pierid species.

Nanopore based arthropod metabarcoding

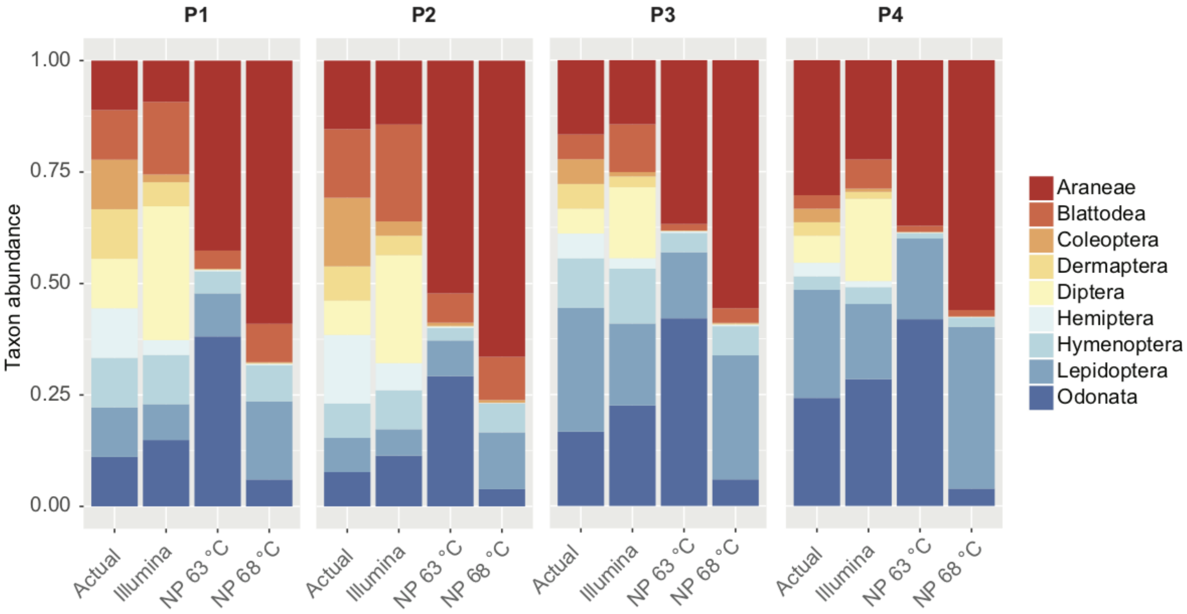

Figure 7: Relative abundances for nine arthropod species in our four mock communities (actual), compared to an Illumina amplicon sequencing protocol, and nanopore protocols at 63 °C and 68 °C annealing temperature

On average, we recovered 2,645 reads for each Illumina sequenced mock community and 1,149 for each nanopore mock community. The optimized Illumina amplicon sequencing based 18SrDNA protocol resulted in a very good taxon recovery. All nine taxa were recovered from all four mock communities (Fig. 7). Moreover, the Illumina based protocol allowed relatively accurate predictions of taxon abundances. Even though no taxon's actual abundance was predicted by Illumina amplicon data, the average fold change between input DNA and recovered read count was closely distributed around zero (Supplementary Fig 8). In contrast, the long-amplicon nanopore protocol showed very biased qualitative and quantitative taxon recovery (Fig. 7). On average, only 83.33 % of taxa were recovered per nanopore sequenced mock community. Moreover, the fold change of input DNA and recovered read count were highly biased between taxa. Some taxa were considerably over or underrepresented among the read population. This led to a significantly higher variation of fold change between input DNA and read count compared to the Illumina amplicon-based protocol (Levene's test $P < 0.05$; Supplementary Fig. 8). A reduction of PCR annealing temperature did result in a considerable increase of Odonata sequences, but overall did not have a strong effect on qualitative (77.78 % of taxa recovered) or quantitative taxon recovery (Fig. 7). The variation of fold change between different PCR annealing temperatures was not significantly different (Levene's test, $P > 0.05$). A reduction of PCR cycle number by 10 also did not yield any significant effect on qualitative (88.89 % of taxa recovered) or quantitative taxon recovery (Supplementary Fig. 9).

Discussion and Potential implications

Phylogenetic and taxonomic utility of long rDNA amplicons

Developments in long-amplicon sequencing hold great promise for molecular taxonomy and phylogenetics across very broad taxonomic scales. We recovered phylogenetic relationships across the eukaryote tree of life, which were mostly consistent with the current state of research (e.g. [52]). Separate orders of arthropods all formed well supported monophyletic groups. Our spider phylogeny was highly congruent with recent work based on whole transcriptomes [35] and multi-amplicon data [53]. Moreover, using the rDNA cluster allowed us to resolve young phylogenetic divergences: the relationships within the recent adaptive radiation of the genus *Tetragnatha* in Hawaii confirmed previous research [58, 60].

Besides their high phylogenetic utility, long rDNA amplicons showed excellent support for taxonomic hypotheses. All morphologically identified species of Hawaiian *Tetragnatha* were recovered as monophyletic groups. The divergence patterns and taxonomic classifications of spiders based on rDNA were strongly correlated to those based on mitochondrial COI, the most commonly used animal barcode marker [4]. rDNA may thus be ideal to complement mitochondrial barcoding. A universal and variable nuclear marker as a supplement to COI barcoding will be particularly useful in cases of mito-nuclear discordance due to male biased gene flow [10, 61], hybridization [12] or infections with reproductive parasites [11].

Their high phylogenetic utility across very broad taxonomic categories also provides long rDNA amplicons with a distinct advantage over short read barcoding protocols, which are not well suited to support broad scale phylogenetic hypotheses [62]. The inclusion of long amplicons would make it possible to scale up barcoding from simple taxon assignment to community wide phylogenetic inferences [9]. It should be noted that the nuclear rDNA cluster is a single locus and its divergence pattern does not necessarily reflect species divergence. Also, the multiple genomic rDNA copies do not necessarily all evolve in concert. rDNA genes may even be prone to pseudogenization.
Taxonomic and phylogenetic analyses based on rDNA may thus be affected by paralogues, and additional information from unlinked genomic regions would therefore be highly desirable to support taxonomic and phylogenetic hypotheses. The mitochondrial genome may be an ideal target for this purpose. Recently, the amplification of whole mitochondrial genomes was

suggested for animal barcoding [63]. This would increase taxonomic and phylogenetic resolution and alleviate some disadvantages of short COI amplicons. However, it is challenging to develop truly universal primers to target mitochondrial genomes across a very wide range of taxonomic groups [64]. A straightforward way to achieve highly resolved phylogenies may be the combination of long rDNA amplicon sequencing with multiplex PCRs of short mitochondrial amplicons, to amplify multiple mitochondrial DNA fragments [65]. Conserved stretches in mitochondrial rDNA may also allow the design of order- or even phylum-specific primers for long range amplification [65]. A combination of long mitochondrial and nuclear rDNA amplicons, possibly in a multiplex PCR, would be a desirable development for future DNA barcoding. With whole genome sequences of different taxa rapidly accumulating, it may also be possible to identify additional unlinked DNA barcoding markers.

Simple, accurate, universal and cost efficient long-amplicon DNA barcoding

Despite the high raw read error of nanopore data, consensus sequences were highly accurate, and library preparation and sequencing for our protocol are simple and cost efficient. Using a single pair of universal primers, long rDNA amplicons can potentially be amplified across diverse eukaryote taxa, here largely demonstrated in arthropods, and in small scale in fungi and plants. A simple dual indexing approach during PCR allows large numbers of samples to be pooled before library preparation [27]. Only a single PCR is required per specimen, while subsequent cleanup and library preparation can be performed on pooled samples. The simplicity of our approach is additionally highlighted by its effectiveness even under field conditions in a remote rainforest site. Nanopore sequencing technology is affordable and universally available to any laboratory. Our ONT MinION generated about 250,000 reads per run. Aiming for about 1,000 reads per amplified specimen, 250 long rDNA barcodes could be generated in single MinION run. Input DNA amounts for different specimens will have to be carefully balanced to maximize the recovery. The total reagent costs, including PCR, library preparation and sequencing, then amount to less than $4 for each long barcode sequence generated.

Pitfalls of nanopore based long-amplicon barcoding

While our protocol was generally straightforward and reliable, we found several drawbacks, which require further considerations and optimization. First, it needs to be noted that long rDNA amplification will not be possible with highly degraded DNA molecules, e.g. from historical specimens [66]. Moreover, amplification success of long range PCRs proved less consistent than that for amplification of short amplicons. We observed a complete failure of some PCRs when too high template DNA concentrations were loaded. The long range polymerase may be more sensitive to PCR inhibitors present in some arthropod DNA extractions [67]. PCR conditions will have to be carefully optimized for reliable and consistent amplification. We also found that highly universal eukaryote primers may result in undesired amplification, for example plants from beetle and butterfly larval guts, phoretic mites, or fungal sequences. However, as long as the DNA of the target taxon is still dominating the resulting amplicon mixture, this undesired amplification will not affect consensus calling. It may be advisable to check the taxonomic composition of amplicon samples before assembly, e.g. by blasting against a reference library. To reduce non-target amplification, PCR primers could also be redesigned to exclude certain lineages from amplification.

It should also be noted that our approach results in only a single consensus sequence for each processed specimen. As a diploid marker, the rDNA cluster can contain heterozygous positions in some specimens, in particular within the ITS regions. This information is currently lost, and a different assembly approach may be necessary to recover heterozygosity as well. Furthermore, index length and edit distance are also important considerations. We used indexes of 15 bp and with a minimum distance of 10 bp to index both sides of our amplicons. Index edit distance of only 4 bp between samples already led to considerable cross-specimen index bleeding. It may thus be better to increase the length and edit distances of indexes. For example, indexes of 20 or 30 bp could be easily attached to the 5'-tails of PCR primers. without strongly affecting PCR efficiency. We have used a relatively crude gel-based approach for pooling amplicon samples. This could have contributed to biased read abundance between some samples. Instead of gel electrophoresis, it may be advisable to use a more precise spectrophotometric quantification.

Nanopore based arthropod metabarcoding

It is well known that Illumina amplicon sequencing of short 18SrDNA fragments can yield accurate taxon recovery in metabarcoding experiments [34], a finding that is confirmed by our results. Except for some outliers (e.g. *Diptera* were overrepresented), even the approximate relative abundance of all taxa was recovered. In contrast, little is known on the performance of long-amplicon nanopore sequencing for community diversity assessments [32]. Our long barcode-based approach resulted in the dropout of several taxa and highly skewed relative taxon abundances. Skewed abundances were already found in microbial community analysis using nanopore [32]. In the simplest case, primer mismatches may be responsible for biased amplification [32, 68]. However, the targeted priming sites in our study were extremely conserved. Also, a change of PCR cycle number and annealing temperature did not have a strong effect on taxon abundances, as would be expected in the case of PCR priming bias [69]. Another possibility is the preferential amplification of template molecules with a certain GC content by the DNA polymerase [33]. However, we found the GC content of the rDNA cluster to be very stable across taxa. Yet another potential explanation for the differential recovery of taxa in community samples is taxonomic bias in DNA degradation [70], but we do not expect DNA degradation to have played a role in our experiment because we used only high quality DNA extractions (verified by gel electrophoresis) from fresh specimens. The most plausible explanation appears to be that variable rDNA lengths are found between different taxa. It is well known that shorter sequences are amplified preferentially in a PCR, especially after it reaches the plateau stage [71]. Such dominance of shorter amplicons could explain the observed biases very well. In fact, the most abundant taxon in our pools was a spider, which also had the shortest amplicon length. The dominant amplification of shorter sequences may also explain the amplification of plant DNA from a butterfly and a flour beetle larva, as plants showed considerably shorter rDNA amplicons than insects. We found a very high variation of rDNA amplicon length within many taxonomic groups, which could be a considerable problem for long read metabarcoding applications. This suggests that it may be worthwhile to focus on narrower taxonomic groups for long amplicon metabarcoding. For example, all spiders in our study share rDNA amplicons of very similar size and would probably be less affected by amplification bias. However, with more closely related taxa in a community, the high error rate of raw reads may cause problems during read clustering and taxon assignments. It should also be noted that we used highly simplified mock community samples, not reflecting actual community composition in nature. Even with those simplified communities, we encountered considerable problems in taxon recovery. Metabarcoding with MinION sequencing may thus be much less trivial than single specimen sequencing. More research into the causes and possible mitigation of these biases will be required before long-amplicon sequencing can be routinely utilized for metabarcoding applications.

Conclusion

Sequencing long dual indexed rDNA amplicons on Oxford Nanopore Technologies' MinION is a simple, cost effective, accurate and universal approach for eukaryote DNA barcoding. Long rDNA amplicons offer high phylogenetic and taxonomic resolution across broad taxonomic scales from kingdom down to species. They also prove to be an excellent complement to mitochondrial COI based barcoding in arthropods. However, despite the long-amplicon advantages in the analysis of separate specimens, we found considerable biases associated with sequencing bulk community samples. The observed taxonomic bias is possibly a result of taxon-specific length variation of the rDNA cluster and preferential amplification of species with shorter rDNA. Further research into the sources of the observed bias is required before long rDNA amplicon sequencing can be utilized as a reliable resource for the analysis of bulk samples.

Availability of source code and requirements

1. The program Minibar can be found at https://github.com/calacademy-research/minibar

Programming language: Python 2.7 (but can be run in Python 3)

Operating systems: MacOS, Linux and Windows

Other requirements: Edlib library module (https://github.com/Martinsos/edlib)

Availbity of supporting data

The following data supporting the results of this article are available in the [will be submitted to the GIgaScience database] repository.

1. Raw fastq read files from Nanopore sequencing runs and Illumina sequencing of arthropod mock communities for short 18S amplicons

2. Fasta sequences of rDNA amplicon for all taxa, mitochondrial COI for Hawaiian *Tetragnatha* spp., as well as Illumina derived consensus sequences for Hawaiian *Peperomia* spp.

3. Newick tree files

4. Analysis tables for the mock community sequencing experiment, the comparison of genetic distances within and between Hawaiian *Tetragnatha* species for COI and rDNA and the distance between Nanopore based and Illumina based consensus sequences

Author contributions

HK and SP designed the study. HK, AP, SRK, JYL, NG, VS and JDS collected the specimens. Laboratory work was carried out by HK, AP and SRK and the data were subsequently analyzed by HK, AP, JBH, SRK and SP. The paper was writen by HK, AP, JBH, SRK, JYL, VS, JDS, NG, NHP, RGG, SP.

Abbreviations

ONT: Oxford Nanopore Technologies; PCR: polymerase chain reaction; rDNA: ribosomal DNA; COI: Cytochrome c oxidase subunit I; RTA: retrolateral tibial apophysis

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Taylor Liu for help during laboratory work, and Tara Gallant for help during specimen collection. Hitomi Asahara graciously provided access to a laboratory facility and the necessary software for our MinION sequencing run. We thank the State of Hawaii Department of Land and Natural Resources and the Servicio Nacional Forestal y de Fauna Silvestre, who provided collection permits, rainforest Expeditions and Gabriela Orihuela for providing assistance and support with fieldwork in Peru. We thank Anna Holmquist for providing the *Psechrus* sp. specimen. The specimen was collected with permits from the Indonesian Ministry of Research and Technology (KEMENRISTEK) and in collaboration with Pungki Lupiyaningdyah and Anang Achmadi from the Museum Zoologicum Bogoriense and funded by an NSF grant DEB 1457845.

References

1. Sala, O.E., Chapin, F.S., Armesto, J.J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L.F., Jackson, R.B., and Kinzig, A. (2000). Global biodiversity scenarios for the year 2100. Science *287*, 1770-1774.
2. Pimm, S.L., Jenkins, C.N., Abell, R., Brooks, T.M., Gittleman, J.L., Joppa, L.N., Raven, P.H., Roberts, C.M., and Sexton, J.O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. Science *344*, 1246752.
3. Rominger, A., Goodman, K., Lim, J., Armstrong, E., Becking, L., Bennett, G., Brewer, M., Cotoras, D., Ewing, C., and Harte, J. (2016). Community assembly on isolated islands: macroecology meets evolution. Global ecology and biogeography *25*, 769-780.
4. Hebert, P.D., Ratnasingham, S., and de Waard, J.R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London B: Biological Sciences *270*, S96-S99.
5. Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Bolchacova, E., Voigt, K., and Crous, P.W. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences *109*, 6241-6246.
6. China Plant BOL Group, Li, D.-Z., Gao, L.-M., Li, H.-T., Wang, H., Ge, X.-J., Liu, J.-Q., Chen, Z.-D., Zhou, S.-L., and Chen, S.-L. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proceedings of the National Academy of Sciences *108*, 19641-19646.
7. Shokralla, S., Porter, T.M., Gibson, J.F., Dobosz, R., Janzen, D.H., Hallwachs, W., Golding, G.B., and Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Scientific reports *5*, 9687.

8. Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C., and Ding, Z. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution *3*, 613-623.

9. Graham, C.H., and Fine, P.V. (2008). Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. Ecology Letters *11*, 1265-1277.

10. Krehenwinkel, H., Graze, M., Rödder, D., Tanaka, K., Baba, Y.G., Muster, C., and Uhl, G. (2016). A phylogeographical survey of a highly dispersive spider reveals eastern Asia as a major glacial refugium for Palaearctic fauna. Journal of Biogeography *43*, 1583-1594.

11. Hurst, G.D., and Jiggins, F.M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proceedings of the Royal Society of London B: Biological Sciences *272*, 1525-1534.

12. Bernatchez, L., Glémet, H., Wilson, C.C., and Danzmann, R.G. (1995). Introgression and fixation of Arctic char (Salvelinus alpinus) mitochondrial genome in an allopatric population of brook trout (Salvelinus fontinalis). Canadian Journal of Fisheries and Aquatic Sciences *52*, 179-185.

13. Melo-Ferreira, J., Boursot, P., Suchentrunk, F., Ferrand, N., and Alves, P. (2005). Invasion from the cold past: extensive introgression of mountain hare (Lepus timidus) mitochondrial DNA into three other hare species in northern Iberia. Molecular Ecology *14*, 2459-2464.

14. Soltis, P.S., and Soltis, D.E. (1998). Molecular evolution of 18S rDNA in angiosperms: implications for character weighting in phylogenetic analysis. In Molecular systematics of plants II. (Springer), pp. 188-210.

15. Hillis, D.M., and Dixon, M.T. (1991). Ribosomal DNA: molecular evolution and phylogenetic inference. The Quarterly review of biology *66*, 411-453.

16. Black IV, W.C., Klompen, J., and Keirans, J.E. (1997). Phylogenetic relationships among tick subfamilies (Ixodida: Ixodidae: Argasidae) based on the 18S nuclear rDNA gene. Molecular Phylogenetics and Evolution *7*, 129-144.

17. Powers, T.O., Todd, T., Burnell, A., Murray, P., Fleming, C., Szalanski, A.L., Adams, B., and Harris, T. (1997). The rDNA internal transcribed spacer region as a taxonomic marker for nematodes. Journal of Nematology *29*, 441.

18. Sonnenberg, R., Nolte, A.W., and Tautz, D. (2007). An evaluation of LSU rDNA D1-D2 sequences for their use in species identification. Frontiers in zoology *4*, 6.

19. Tang, C.Q., Leasi, F., Obertegger, U., Kieneke, A., Barraclough, T.G., and Fontaneto, D. (2012). The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. Proceedings of the National Academy of Sciences *109*, 16208-16212.

20. von der Schulenburg, J.H.G., Hancock, J.M., Pagnamenta, A., Sloggett, J.J., Majerus, M.E., and Hurst, G.D. (2001). Extreme length and length variation in the first ribosomal internal transcribed spacer of ladybird beetles (Coleoptera: Coccinellidae). Molecular Biology and Evolution *18*, 648-660.

21. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., and Fiddes, I.T. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature biotechnology *36*, 338.

22. Heeger, F., Bourne, E.C., Baschien, C., Yurkov, A., Bunk, B., Spröer, C., Overmann, J., Mazzoni, C.J., and Monaghan, M.T. (2018). Long-amplicon DNA metabarcoding of ribosomal rRNA in the analysis of fungi from aquatic environments. Molecular Ecology Resources.

23. Tedersoo, L., Tooming-Klunderud, A., and Anslan, S. (2018). PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytologist *217*, 1370-1385.

24. Giordano, F., Aigrain, L., Quail, M.A., Coupland, P., Bonfield, J.K., Davies, R.M., Tischler, G., Jackson, D.K., Keane, T.M., and Li, J. (2017). De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. Scientific reports *7*, 3935.

25. Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L.A., Barrio-Amorós, C.L., Salazar-Valenzuela, D., and Prost, S. (2018). Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. GigaScience *7*, giy033.

26. Wurzbacher, C., Larsson, E., Bengtsson-Palme, J., Van den Wyngaert, S., Svantesson, S., Kristiansson, E., Kagami, M., and Nilsson, R.H. (2018). Introducing ribosomal tandem repeat barcoding for fungi. bioRxiv, 310540.

27. Srivathsan, A., Baloğlu, B., Wang, W., Tan, W.X., Bertrand, D., Ng, A.H., Boey, E.J., Koh, J.J., Nagarajan, N., and Meier, R. (2018). A Min ION™-based pipeline for fast and cost-effective DNA barcoding. Molecular ecology resources.

28. Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., Bore, J.A., Koundouno, R., Dudas, G., and Mikhail, A. (2016). Real-time, portable genome sequencing for Ebola surveillance. Nature *530*, 228.

29. Edwards, A., Debbonaire, A.R., Sattler, B., Mur, L.A., and Hodson, A.J. (2016). Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N. bioRxiv, 073965.

30. Giribet, G., and Edgecombe, G.D. (2012). Reevaluating the arthropod tree of life. Annual review of entomology *57*, 167-186.

31. Hochkirch, A. (2016). The insect crisis we can’t ignore. Nature News *539*, 141.

32. Benítez-Páez, A., Portune, K.J., and Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. GigaScience *5*, 4.

33. Nichols, R.V., Vollmers, C., Newsom, L.A., Wang, Y., Heintzman, P.D., Leighton, M., Green, R.E., and Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. Molecular ecology resources.

34. Krehenwinkel, H., Wolf, M., Lim, J.Y., Rominger, A.J., Simison, W.B., and Gillespie, R.G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific reports *7*, 17668.

35. Fernández, R., Kallal, R.J., Dimitrov, D., Ballesteros, J.A., Arnedo, M.A., Giribet, G., and Hormiga, G. (2018). Phylogenomics, Diversification Dynamics, and Comparative Transcriptomics across the Spider Tree of Life. Current Biology *28*, 1489-1497. e1485.

36. De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long read sequencing data. bioRxiv, 237180.

37. Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research *27*, 737-746.

38. Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution *30*, 2725-2729.

39. Straub, S.C., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C., and Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. American Journal of Botany *99*, 349-364.

40. Bolger, A., and Giorgi, F. Trimmomatic: A Flexible Read Trimming Tool for Illumina NGS Data. URL http://www.usadellab.org/cms/index. php.

41. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754-1760.

42. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078-2079.

43. Lanfear, R., Calcott, B., Ho, S.Y., and Guindon, S. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution *29*, 1695-1701.

44. Huelsenbeck, J.P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics *17*, 754-755.

45. Dray, S., and Dufour, A.-B. (2007). The ade4 package: implementing the duality diagram for ecologists. Journal of statistical software *22*, 1-20.

46. Machida, R.J., and Knowlton, N. (2012). PCR primers for metazoan nuclear 18S and 28S ribosomal DNA sequences. PLoS one *7*, e46180.

47. Krehenwinkel, H., Kennedy, S., Pekár, S., and Gillespie, R.G. (2017). A cost-efficient and simple protocol to enrich prey DNA from extractions of predatory arthropods for large-scale gut content analysis by Illumina sequencing. Methods in Ecology and Evolution *8*, 126-134.

48. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.

49. Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, Volume 10. pp. 707-710.

50. Šošić, M., and Šikić, M. (2017). Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. Bioinformatics *33*, 1394-1395.

51. Loman, N.J., Quick, J., and Simpson, J.T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. bioRxiv, 015552.

52. Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., and Beutel, R.G. (2014). Phylogenomics resolves the timing and pattern of insect evolution. Science *346*, 763-767.

53. Wheeler, W.C., Coddington, J.A., Crowley, L.M., Dimitrov, D., Goloboff, P.A., Griswold, C.E., Hormiga, G., Prendini, L., Ramírez, M.J., and Sierwald, P. (2017). The spider tree of life: phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling. Cladistics *33*, 574-616.

54. Gillespie, R.G. (1991). Hawaiian spiders of the genus Tetragnatha: I. Spiny leg clade. Journal of Arachnology, 174-209.

55. Gillespie, R.G. (1999). Comparison of rates of speciation in web-building and non-web-building groups within a Hawaiian spider radiation. Journal of Arachnology, 79-85.

56. Gillespie, R.G. (2016). Island time and the interplay between ecology and evolution in species diversification. Evolutionary applications *9*, 53-73.

57. Gillespie, R.G., Croom, H.B., and Hasty, G.L. (1997). Phylogenetic relationships and adaptive shifts among major clades of Tetragnatha spiders (Araneae: Tetragnathidae) in Hawai'i.

58. Gillespie, R. (2004). Community assembly through adaptive radiation in Hawaiian spiders. Science *303*, 356-359.

59. Blackledge, T.A., Binford, G.J., and Gillespie, R.G. (2003). Resource use within a community of Hawaiian spiders (Araneae: Tetragnathidae). In Annales Zoologici Fennici. (JSTOR), pp. 293-303.

60. Blackledge, T.A., and Gillespie, R.G. (2004). Convergent evolution of behavior in an adaptive radiation of Hawaiian web-building spiders. Proceedings of the National Academy of Sciences of the United States of America *101*, 16228-16233.

61. Wilmer, J.W., Hall, L., Barratt, E., and Moritz, C. (1999). Genetic Structure and Male-Mediated Gene Flow in the Ghost Bat (Macroderma gigas). Evolution, 1582-1591.

62. Kjer, K.M., Zhou, X., Frandsen, P.B., Thomas, J.A., and Blahnik, R.J. (2014). Moving toward species-level phylogeny using ribosomal DNA and COI barcodes: an example from the diverse caddisfly genus Chimarra (Trichoptera: Philopotamidae). Arthropod Systematics & Phylogeny *72*, 345-354.

63. Deiner, K., Renshaw, M.A., Li, Y., Olds, B.P., Lodge, D.M., and Pfrender, M.E. (2017). Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. Methods in Ecology and Evolution *8*, 1888-1898.

64. Briscoe, A.G., Goodacre, S., Masta, S.E., Taylor, M.I., Arnedo, M.A., Penney, D., Kenny, J., and Creer, S. (2013). Can long-range PCR be used to amplify genetically divergent mitochondrial genomes for comparative phylogenetics? A case study within spiders (Arthropoda: Araneae). PLoS one *8*, e62404.

65. Krehenwinkel, H., Kennedy, S., Rueda, M., and Gillespie, R. (2018). Low cost molecular systematics of entire arthropod communities: Primer sets for rapid multi locus analyses by multiplex PCRs and Illumina amplicon sequencing. Methods in Ecology and Evolution.

66. Krehenwinkel, H., and Pekar, S. (2015). An analysis of factors affecting genotyping success from museum specimens reveals an increase of genetic and morphological variation during a historical range expansion of a European spider. PLoS one *10*, e0136337.

67. Margam, V.M., Gachomo, E.W., Shukle, J.H., Ariyo, O.O., Seufferheld, M.J., and Kotchoni, S.O. (2010). A simplified arthropod genomic-DNA extraction protocol for

polymerase chain reaction (PCR)-based specimen identification through barcoding. Molecular biology reports *37*, 3631-3635.

68. Sipos, R., Székely, A.J., Palatinszky, M., Révész, S., Márialigeti, K., and Nikolausz, M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiology Ecology *60*, 341-350.

69. Suzuki, M.T., and Giovannoni, S.J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Applied and environmental microbiology *62*, 625-630.

70. Krehenwinkel, H., Fong, M., Kennedy, S., Huang, E.G., Noriyuki, S., Cayetano, L., and Gillespie, R. (2018). The effect of DNA degradation bias in passive sampling devices on metabarcoding studies of arthropod communities and their associated microbiota. PLoS one *13*, e0189188.

71. Wattier, R., Engel, C., Saumitou-Laprade, P., and Valero, M. (1998). Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in Gracilaria gracilis (Rhodophyta). Molecular Ecology *7*, 1569-1573.

Supplementary Table 1