bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

go_batch: A snakemake pipeline to assemble mitochondrial and ribosomal sequences from genome skims.

Oliver White[1], Andie Hall[1], Matt Clark[1], Suzanne T. Williams[1]

Corresponding authors Matt Clark m.clark@nhm.ac.uk and Suzanne Williams s.williams@nhm.ac.uk

Authors addresses

[1] The Natural History Museum, Cromwell Road, London, SW7 5BD

Orchid IDs

Oliver White	0000-0001-6444-0310
Andie Hall	0000-0001-5546-7255
Matt Clark	0000-0002-8049-5423
Suzanne T. Williams	0000-0003-2995-5823

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Low coverage “genome-skims” are often used to assemble organelle genomes and ribosomal gene sequences for cost effective phylogenetic and barcoding studies. Natural history collections hold invaluable biological information, yet degraded DNA often hinders PCR based analysis. However, with improvements to molecular techniques and sequencing technology, it is possible to use ancient DNA methods to generate libraries and sequence short fragments from degraded DNA to generate genome skims from museum collections.

Here we introduce "*go_batch*", a bioinformatic pipeline written in snakemake designed to unlock the genomic potential of historical museum specimens using genome skimming. Specifically, *go_batch* allows the batch assembly and annotation of mitochondrial genomes and nuclear ribosomal genes from low-coverage skims. The utility of the pipeline is demonstrated by analysing a novel genome skimming dataset from both recent and historical sollariellid gastropod samples.

We demonstrate that *go_batch* can recover previously unattainable mitochondrial genomes and ribosomal genes from sollariellid gastropods. In addition, phylogenetic analysis of these gene sequences helped resolve complex taxonomic relationships.

The generation of bioinformatic pipelines that facilitate processing large quantities of sequence data from the vast repository of specimens held in natural history museum collections will greatly aid species discovery and exploration of biodiversity over time, ultimately aiding conservation efforts in the face of a changing planet.

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Natural history collections are home to more than one billion expertly verified specimens worldwide (Bartolozzi et al., 2023) as well as large numbers of unsorted and unidentified bulk samples, and as such represent a vast repository of biological data that remains largely untapped. Challenges associated with such material include poor preservation, the use of unknown preservatives, the age of material, DNA degradation and contamination. Advances in novel laboratory techniques (Ruane & Austin, 2017; Straube et al., 2021) and next generation sequencing (NGS) technology overcomes many of these obstacles and make it possible to obtain DNA sequences from many historical specimens, unlocking the potential for wide-ranging genomic analyses. Using natural history collections provides the opportunity to sample species that are rarely collected or even extinct and from areas of the world that are poorly sampled. It also avoids the need for fieldwork which can be costly, time consuming and in some cases, dangerous and may involve complicated regulatory issues.

Genome skimming has gained increasing popularity as an approach for barcoding specimens from historical museum collections. The term “genome skimming” refers to the generation of low coverage NGS data and was first coined by Straub et al (2012). Although genome skimming does not generate data with sufficient coverage to assemble entire nuclear genome sequences, there are sufficient reads to assemble sequences that are present in the genome in multiple copies and are therefore still well represented in the sequence data. Common targets for genome skimming studies include organelle genomes (a typical cell has one nucleus but many organelles) and nuclear ribosomal genes (there are 100s of rRNA nuclear genes, typically arranged in arrays). For many years partial gene sequences from one or two organelle genes (usually mitochondria: *cox1*, chloroplasts: *matK* and *rbcL*) have been used as barcodes in DNA based taxonomy given their high copy number and availability of “universal” primers that work on a wide range of species, but increasingly whole organelle genomes are becoming the focus of barcoding studies, or even the entire genome skim dataset as a ”DNA-mark” (Bohmann et al., 2020).

When working with historical specimens in particular, genome skimming offers many advantages over polymerase chain reaction (PCR) amplification and sequencing. Optimally, high yields of high molecular weight genomic DNA are required for PCR, but degraded and low yield DNA are also suitable for short read NGS (such as Illumina). The wet lab work is relatively straightforward, only requiring DNA extraction and library methods optimised for degraded DNA. Genome skimming also has additional benefits over targeted PCR since multiple loci can be recovered at the same time without development and optimisation of multiple PCR primers. With advances in bioinformatic tools, it is likely that low coverage genome skimming datasets will have even greater utility in the future. For example, recent kmer based approaches have been developed for genome skims to investigate phylogenetic relationships (Sarmashghi et al., 2017) and genome properties (Sarmashghi et al.,

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2021). Finally, genome skimming is increasingly cost effective as the cost of NGS sequencing continues to decrease. In the light of these advantages, genome skimming is seen as a hugely scalable process that is suitable for batch recovery of barcode genes from museum collections.

However, few bioinformatic pipelines are available to assist with the assembly of large numbers of organelle and nuclear ribosomal sequences from batches of genome skimming data. Notable exceptions include MitoZ (Meng et al., 2019) and NOVOWrap (Wu et al., 2021) for the assembly and annotation of mitochondrial genomes. In addition, plastaumatic (W. Chen et al., 2022) is available for chloroplast assembly and annotation and PhyloHerb (Cai et al., 2022) can be used for the assembly of chloroplast and nuclear ribosomal repeats without annotation. These tools were not designed with historical and/or degraded samples in mind and do not account for issues such as contamination and the undesirable assembly of non-target sequences. In addition, these tools do not implement phylogenetic analysis of the annotated genes identified. Other targeted assembly approaches are available including Orthoskim (Pouchon et al., 2022), but this is not available as part of a pipeline that can be scaled across many samples.

This study introduces *go_batch*, a pipeline written in snakemake for batch assembly and annotation of mitochondrial genomes and nuclear ribosomal genes, and phylogenetic analysis from genome skimming data. The pipeline wraps 12 published bioinformatic tools as well as custom python and R scripts into a single user-friendly pipeline designed to cope with poor quality data from historical collections, permitting large scale genome skimming studies from museum specimens. *go_batch* (1) runs on a single machine or in parallel on a High Performance Computing cluster, (2) can be utilised to process a single sample or batches of samples, (3) can be used to assemble both mitochondrial and nuclear ribosomal sequences (3) uses GetOrganelle which an independent review found to be the best performing assembly tool (Freudenthal et al., 2020; Jin et al., 2020), (4) performs basic assembly checking for contamination and non-target sequences commonly found in historical samples and (5) generates phylogenetic gene trees based on from annotated genes.

To demonstrate the utility of *go_batch*, we used the pipeline to analyse a novel genome skimming dataset for the gastropod family Solariellidae (hereafter solariellid gastropods). This group was selected as it represents many of the challenges associated with genome skimming museum collections. Solariellids are small marine snails found predominantly in deep-water. Many species are rare and as a family they are poorly represented in museum collections worldwide, with few live-collected specimens: many species are known only from a single, dry and often damaged shell (Williams et al., 2020). Although solariellid gastropods have been the focus of previous phylogenetic studies (Sumner-Rooney et al., 2016; Williams et al., 2013, 2022), these studies have relied on partial sequence from only four genes, which have not fully resolved relationships among genera. As such, our understanding of solariellid evolution would greatly benefit from increased gene sampling, but there are no published reference genomes for the group and limited genomic data available on public databases.

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Where universal primers exist, attempts to include key taxa in previous studies has not always been possible as PCRs have failed, likely due to degraded fragment size. Given their rarity, small size and frequently poor preservation, solariellids are an excellent test case for the utility of genome skim data and pipelines designed for historical specimens.

Material and methods

Solariellid sample selection and sequencing

A total of 25 samples were selected, with representatives from 18 genera, encompassing the diversity of the solariellid family, including several species with dubious generic assignments (Table 1). Samples differ in several ways that likely affected DNA quality and yield (Supplementary Table 1), for example, time since collection (1967-2015) and preservation method (dry shell with dehydrated body tissues or live-collected snail preserved in 70- 99% ethanol). In addition, some shells were cracked, allowing the rapid penetration of ethanol, which is particularly important as snails can seal their bodies inside their shells by closing their operculum effectively excluding ethanol. Samples also differ in time kept in storage (initially at 4º C and then at -20º C) since DNA was extracted (2010–2020; Supplementary Table 1).

DNA was isolated using Qiagen DNeasy blood and tissue kit and quantified using a Qubit fluorimeter and High Sensitivity assay kit. A Tapestation 2200 was also used to assess DNA integrity prior to library preparation. Polymerase Chain Reaction (PCR) amplification and Sanger sequencing of mitochondrial (*cox1*, 16S and 12S) and ribosomal genes (28S) were attempted for each sample to compare with our genome skimming approach. Illumina Libraries were prepared using a SparQ DNA Frag and Library Prep kit (QuantaBio, Beverly USA) and sparQ PureMag Beads (QuantaBio), with Sparq Adaptor Barcode sets A and B (QuantaBio), with bespoke modifications (See Supporting Information Methods). Libraries were normalised and pooled equally before being sent to Novogene (Cambridge, UK) for sequencing. The single indexed libraries were sequenced on an Illumina Novaseq on an S4 300 cycle flowcell using 150bp paired reads.

Additional sequence data for '*Solariella*' *varicosa* was provided by Andrea Waeschenbach (Natural History Museum London, UK). Raw sequence data for two outgroups from the family Turbinidae were also analysed, including: *Turbo cornutus* (Kim et al., 2022; SRR15496837) and unpublished raw data for *Lunella* aff. *cinerea* (Williams et al., 2014). These outgroup sequences provide the possibility of comparing published assembled and manually curated organelle genomes with the results from our pipeline using the same raw sequence data.

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1 – Sample details for 25 solariellid gastropods and two outgroup species used in this study with museum registration numbers or NCBI sequence read archive number for sequence data (*Turbo cornutus* only), ocean of origin, latitude and longitude of collection location and depth. Abbreviations: AMS: Australian Museum; MNHN: Muséum national d'Histoire naturelle; SMNH: Swedish Museum of Natural History; MNSA: KwaZulu-Natal Museum; NMNZ: Museum of New Zealand Te Papa Tongarewa; NHMUK: Natural History Museum, London. Names correspond to those used in previous studies (Williams et al. 2020, 2022). Inverted commas around generic names indicates uncertainty about generic assignment based on this or previous studies. Previously published data for *Turbo cornutus* (Kim et al., 2022) and *Lunella aff. cinerea* (Williams et al., 2014) were also included in this study.

Species	Specimen voucher	Ocean	Latitude	Longitude	Depth (m)
Archiminolia oleacea	AMS C.133269	Indo-West Pacific	-24.375	153.285	192-229
Arxellia herosae	MNHN-IM-2009-28739	Indo-West Pacific	-24.717	168.167	298-324
Bathymophila gravida	NMNZ M.299691	Indo-West Pacific	-36.146	178.202	712-924
'Bathymophila' sp. 18	MNHN-IM-2009-23080	Indo-West Pacific	-22.317	171.333	925
Bathymophila-Like sp. 12	MNHN-IM-2009-28741	Indo-West Pacific	-19.667	-178.167	314-377
Chonospeira nuda	SMNH 127100	North East Pacific	36.367	-122.417	999
Clade D sp. d	MNHN-IM-2013-59648	Indo-West Pacific	22.050	119.067	1306-1756
Elaphriella wareni	MNHN-IM-2013-45837	Indo-West Pacific	-8.617	151.783	705-817
Ilanga whitechurchi	NMSA W9631	South West Indian Ocean	-33.167	28.033	90
Lamellitrochus sp. 6	MNHN-IM-2013-60491	Caribbean	16.350	-60.900	111-162
'Lamellitrochus' carinatus	MNHN-IM-2009-31169	Caribbean	16.360	-61.579	29
Microgaza rotella	MNHN-IM-2013-8023	Caribbean	16.400	-61.550	130
Phragmomphalina tenuiseptum	NMNZ M299700	Indo-West Pacific	-31.867	172.433	780-790
Solariella amabilis	NHMUK 20180166	North Atlantic	62.191	5.567	150-200
Solariella sp. 7	MNHN-IM-2019-12000	Indo-West Pacific	-24.800	168.150	250-270
'Solariella' carvalhoi	MNHN-IM-2013-61297	Caribbean	15.800	-61.467	379-428
'Solariella' obscura	NHMUK 20230529	North Atlantic	69.803	30.693	04-Dec
'Solariella' varicosa	NHMUK 20120235	North Atlantic	70.067	29.200	10-174
Spectamen bellulum	NHMUK 20110452	Indo-West Pacific	-26.943	153.404	31
'Spectamen' franciscanum	NMSA V1091	South West Indian Ocean	-34.783	23.983	171
Suavotrochus lubricus	MNHN-IM-2013-61096	Caribbean	16.033	-61.233	266-388
'Suavotrochus' sp. 2	MNHN-IM-2013-61502	Caribbean	15.783	-61.200	550-562
'Zetela' alphonsi	SMNH 10387	South East Pacific	-36.361	-73.725	865
Zetela kopua	NMNZ M.131532	Indo-West Pacific	-45.403	173.980	1386
Zetela textilis	NMNZ M.035478	Indo-West Pacific	-42.637	176.283	256-311
OUTGROUPS					
Lunella aff. *cinerea*	NHMUK 20100448	Indo-West Pacific	-12.554	130.876	NA
Turbo cornutus	SRR15496837	Indo-West Pacific	33.454	126.949	NA

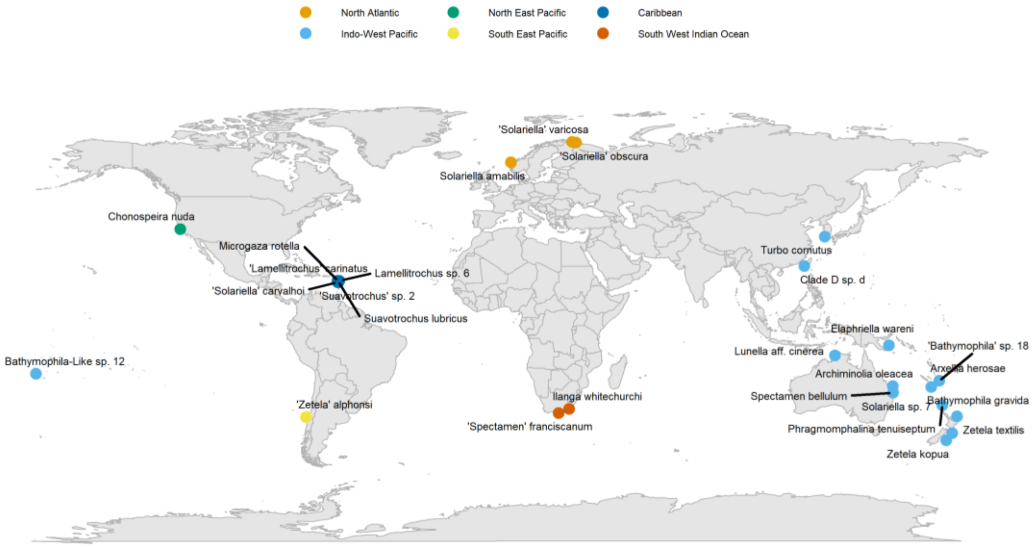

Figure 1 – Map showing collection localities for solariellid gastropods samples used in this study.

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pipeline description

As input, the pipeline requires two main inputs: (1) a config.yml file and a (2) samples.csv file. The config file outlines the main parameters including the target sequence type (mitochondrial or ribosomal), paths to reference databases (blast database, NCBI taxdump, MITOS) and number of threads to use. The samples.csv file is a list of the samples including in the analysis with paths to forward and reverse reads, and paths to the gene and seed databases required by GetOrganelle. The pipeline accepts NGS data from short read platforms (e.g., Illumina) in paired fastq format.

The pipeline starts by processing the data from each sample, using fastp (S. Chen et al., 2018) to detect and remove adapter sequences with quality filtering disabled, as the de Bruijn graph assembly method used by GetOrganelle (SPAdes; Prjibelski et al., 2020) accounts for sequencing errors in reads. GetOrganelle (Jin et al., 2020) is then used to assemble the target sequence of interest. If the target sequence is an organelle genome, GetOrganelle is implemented with the following parameters: `--reduce-reads-for-coverage inf --max-reads inf -R 20`. If the target sequence ribosomal, the following parameters are used following the authors suggestions: `-F anonym --reduce-reads-for-coverage inf --max-reads inf -R 10 --max-extending-len 100 -P 0`. Sequences assembled by GetOrganelle are typically named based on the output of SPAdes (Prjibelski et al., 2020), which can produce long sequence names. Therefore, sequences are renamed to <sample_name>_contig<n> if there are multiple contigs or <sample_name>_circular if a single circular sequence is found. Note that GetOrganelle can produce more than one assembled sequence where there are different possible paths e.g., mitochondrial genomes containing repeats. However, the pipeline simply selects the first assembled sequence for downstream analyses as the main outputs are the annotated gene sequences and the correct orientation of repeat regions is not necessary. Basic assembly statistics are summarised using SeqKit (Shen et al., 2016). Next, the assembly quality is evaluated using a blastn search (Camacho et al., 2009) against a database specified in the config.yaml file and mapping input reads to the assembled sequence using minamp2 (Li, 2018). This information is summarised using blobtools (Laetsch et al., 2017) and the likely taxonomy of the assembled sequence is define using the taxrule “bestsumorder”. Following the assembly quality check, assembled sequences are annotated using MITOS2 (Bernt et al., 2013), or barrnap (https://github.com/tseemann/barrnap) for organelle or ribosomal sequences respectively. Following assembly and annotation, a plot is created to visualise the location of annotated genes, coverage and proportion of mismatches in mapped reads.

Once the sequences are assembled and annotated, the checkpoint function of snakemake is used to recover all protein coding genes assembled across samples. For each protein coding gene recovered, mafft (Katoh & Standley, 2013) is used to align sequences with the following parameters: `--maxiterate 1000 --globalpair --adjustdirection`. The alignments are trimmed using either Gblocks (Castresana, 2000)

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

or Clipkit (Steenwyk et al., 2020) as specified in the specified in the config.yaml file. Phylogenetic analysis is then implemented in with IQ-TREE2 (Minh et al., 2020) and consensus trees are plotted in R using the ggtree package (R Core team, 2020; Yu et al., 2017).

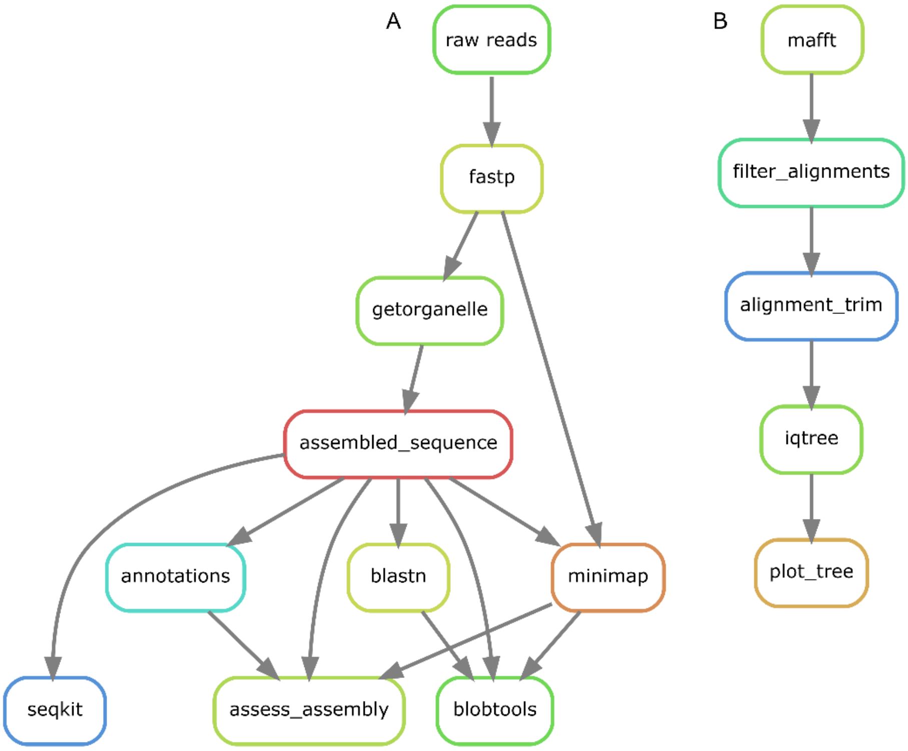

Figure 2 – Schematic diagram of pipeline workflow. Workflow A is applied to all samples provided in the samples.csv. Workflow B is applied to all annotated gene sequences found across assembled sequences.

The pipeline output was manually checked to identify possible contamination in assembled sequences for a given sample or individual genes. Specifically, the blobtools output and was checked for sequences with unusual blast hits and the gene alignments and gene trees were reviewed by taxonomic experts to identify incongruent relationships.

After the removal of putative contaminant sequences, individual gene alignments were reanalysed with IQ-TREE with 1000 ultrafast bootstraps. The gene trees generated were used to infer phylogenetic relationships using astral which uses individual gene trees as input (Zhang et al., 2018). In addition, individual gene alignments were combined into a partitioned alignment using a custom python script and a partitioned phylogenetic analysis was implemented using IQ-TREE and ultrafast bootstraps with 1000 replicates.

Results

Amplification of four genes (28S, COI, 12S and 16S) was attempted soon after DNA was first extracted, and the results are listed in Supplementary Table 1. In some cases, faint bands were observed when PCR products were visualised on agarose gels, but clean sequence could not be obtained, because of low yield and noisy

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

background. Often, only 12S, the smallest PCR fragment could be amplified and sequenced, suggesting that DNA was degraded. This was confirmed by recording the DNA Integrity Number (DIN) for samples. DIN is automatically assigned by the instrument following an algorithm based on the signal distribution across the size range. A DIN of 10 indicates highly intact DNA fragments, whilst a DIN of 1 indicates a highly degraded DNA sample (Supplementary Table 1). DNA quality for the samples used in this study ranged from not detectable for the poorest samples to 6.5 for the best.

Approximately 870 M of raw sequence reads were generated for all samples, with an average of 32 M raw reads per sample. Approximately 77% of reads were retained following adapter removal with fastp. Overall, *go_batch* successfully recovered mitochondrial genome sequences from 25/28 samples with an average assembly size of 13,539 bp. A circular mitochondrial genome was assembled for a single sample (*Zetela kopua*; **Error! Reference source not found.**). However, no mitochondrial sequences could be assembled for *Bathymophila gravida* or *Zetela textilis*. Of the 15 mitochondria genes annotated by MITOS2 (13 protein coding genes and two mitochondrial ribosomal subunit), an average of 12 genes were annotated across samples, with 12 samples having all protein coding and rRNA genes annotated.

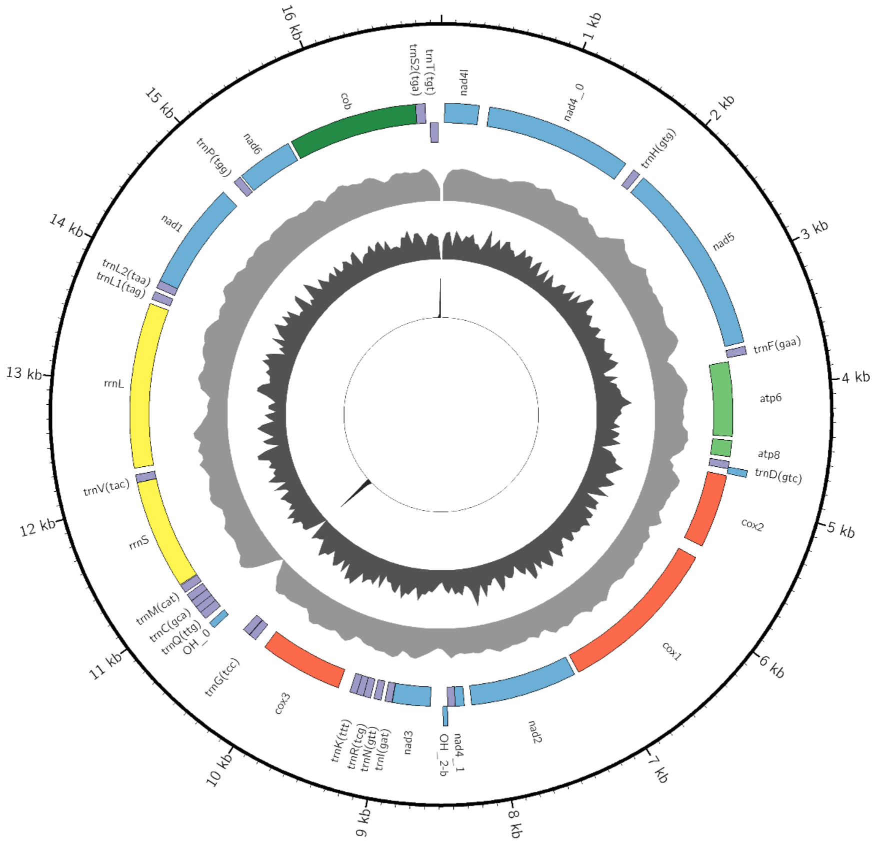

Figure 3 - Assembled circular sequence for *Zetela kopua* with the following attributes from outside to inside: sequence position, annotation names, annotations on the + strand, annotations on the - strand, coverage (max=2779), GC content (max=0.6) and repeat content (max=1.0). This image was created using a custom organelle visualisation tool available on GitHub (https://github.com/o-william-white/circos_plot_organelle; accessed 08/2023).

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Nuclear ribosomal gene sequences were assembled from 25/28 samples with an average size of 1,500 bp. No ribosomal sequences could be assembled for *Bathymophila gravida* or *'Spectamen' franciscanum*. The 28S rRNA sequence could be annotated for all assembled sequences except samples with the shortest partially assembled sequences including *Phragmomphalina tenuiseptum* (709 bp), *Zetela kopua* (728 bp) and *Zetela textilis* (382 bp).

After manual checking of the alignments and phylogenetic trees generated by go_batch, it was determined that all sequence data from *Spectamen cf. bellulum* and *Archiminolia oleacea* were likely contaminants (gastropods), likely originating from lab contamination during DNA extraction and handling the specimen during tissue harvesting and were therefore removed from further analyses. In addition, a single duplicated sequence of nad3 identified from *Solariella amabilis* was removed from downstream analyses.

After the removal of contaminant sequences, IQ-TREE was repeated for both individual and concatenated gene alignments using ultrafast bootstraps with 1000 replicates. The partitioned alignment and partition file formatted for IQ-TREE were created using custom python script. Phylogenetic analysis of the partitioned alignment in IQ-TREE (Figure 4) recovered a tree with support values ranging from poor to optimal to poor (35–89%).

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

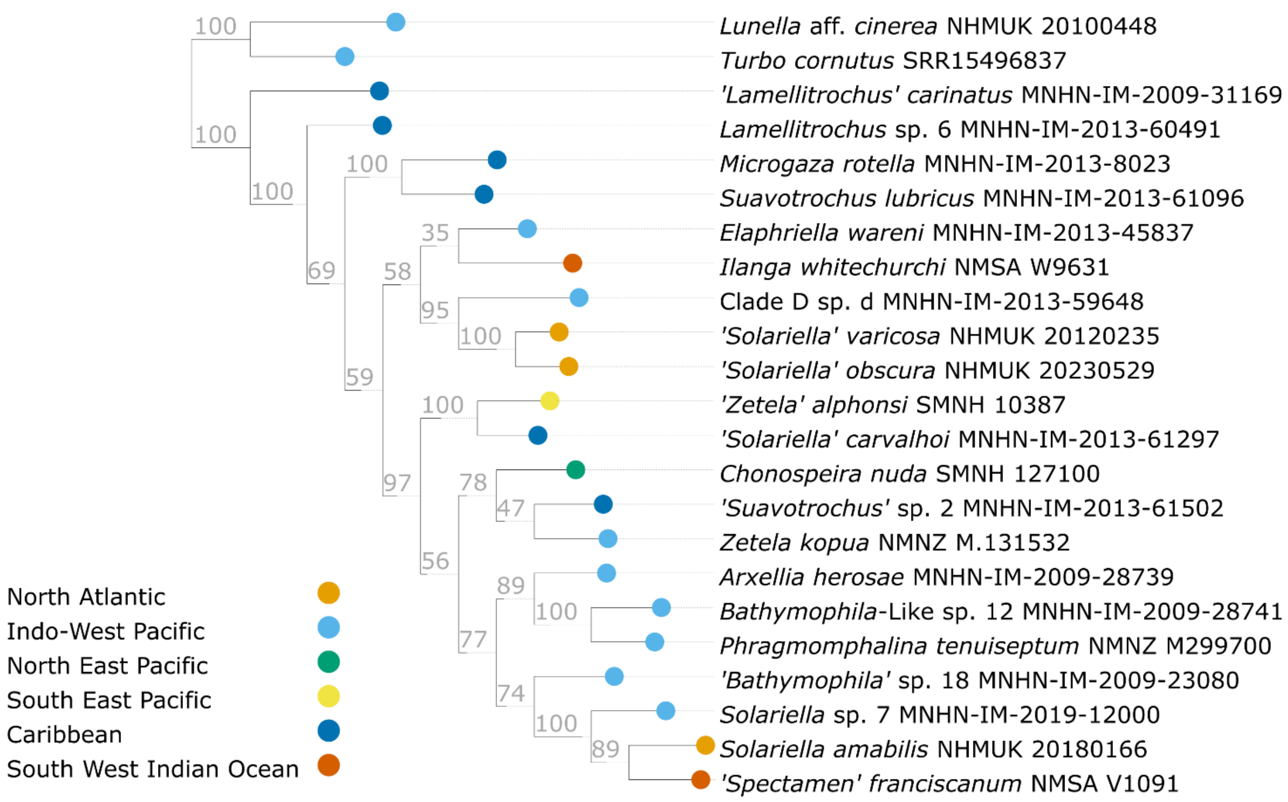
100
100
100
100
69
58
35
95
100
59
100
97
78
47
56
89
100
77
74
100
89
Lunella aff. *cinerea* NHMUK 20100448
Turbo cornutus SRR15496837
'Lamellitrochus' carinatus MNHN-IM-2009-31169
Lamellitrochus sp. 6 MNHN-IM-2013-60491
Microgaza rotella MNHN-IM-2013-8023
Suavotrochus lubricus MNHN-IM-2013-61096
Elaphriella wareni MNHN-IM-2013-45837
Ilanga whitechurchi NMSA W9631
Clade D sp. d MNHN-IM-2013-59648
'Solariella' varicosa NHMUK 20120235
'Solariella' obscura NHMUK 20230529
'Zetela' alphonsi SMNH 10387
'Solariella' carvalhoi MNHN-IM-2013-61297
Chonospeira nuda SMNH 127100
'Suavotrochus' sp. 2 MNHN-IM-2013-61502
Zetela kopua NMNZ M.131532
Arxellia herosae MNHN-IM-2009-28739
Bathymophila-Like sp. 12 MNHN-IM-2009-28741
Phragmomphalina tenuiseptum NMNZ M299700
'Bathymophila' sp. 18 MNHN-IM-2009-23080
Solariella sp. 7 MNHN-IM-2019-12000
Solariella amabilis NHMUK 20180166
'Spectamen' franciscanum NMSA V1091
North Atlantic ●
Indo-West Pacific ●
North East Pacific ●
South East Pacific ●
Caribbean ●
South West Indian Ocean ●

Figure 4 - Partitioned maximum likelihood tree of 17 mitochondrial protein-coding genes, two mitochondrial ribosomal genes and one nuclear ribosomal gene (28S) generated using IQ-TREE and visualised using ete3. The tree is rooted on the outgroup taxa and values on branches are ultrafast bootstrap values.

Discussion

This study demonstrates the utility of *go_batch*, a snakemake pipeline for the assembly and annotation of organelle and ribosomal genes from genome skimming datasets, using a novel dataset from solariellid gastropods. The mitochondrial genomes generated here are the first for the family Solariellidae. Complete or partial mitochondrial genomes were obtained for 25 out of 28 specimens, including samples collected more than 50 years ago, DNA extracted more than ten years ago from dehydrated tissue samples (2/3 successful) and specimens preserved in low percentage (70-80%) ethanol with uncracked shells (10/12 successful) and with highly degraded DNA (DIN<2). Previous phylogenetic analyses of solariellid gastropods have highlighted complex and unresolved phylogenetic relationships. Although the tree in this study has reasonable support values for most nodes, some of the generic assignments require further taxonomic assignment, ideally using increased taxon sampling.

Although this methodology will work for any sample type, go_batch was written specifically to account for many of the issues associated with historical museum samples, for example, DNA degradation and contamination. By default, go_batch implements GetOrganelle using all reads as input (`--reduce-reads-for-coverage inf --max-reads inf`) and an increased number of rounds of (`-R 20`) of target read

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

selection. For museum samples that are likely to be degraded, this maximises the inclusion of sequencing reads. In addition, the user can specify a custom reference database for GetOrganelle using sequences from closely related taxa. This is necessary because benchmarking of GetOrganelle using simulated datasets (pers. com. Oliver White 2023), highlighted that a reference dataset containing closely related sequences increases the likelihood of successful assembly. Conversely, a broad reference dataset can increase the likelihood of non-target sequence assembly. Taxonomic assignment of assembled sequences using blobtools also provides the opportunity to identify non-target sequences. Finally, the phylogenetic analysis implemented for all annotated genes can also help to identify contamination based on incongruent relationships.

Although *go_batch* simplifies the bioinformatic analyses significantly, allowing for the analysis of many samples simultaneously, there is an opportunity for trade-off with accuracy with increasing scale. Indeed, our study highlighted that it was important to manually check the assembled sequences for contamination or poorly annotated sequences using the outputs of blobtools and by examining the individual gene alignments and phylogenetic analyses. In addition, taxonomic expertise may be necessary to identify incongruent phylogenetic relationships that can result from cross contamination from closely related taxa, highlighting the need for particular care when extracting DNA from historical specimens.

In conclusion, this study demonstrates that *go_batch* pipeline can cope with poor quality data from historical collections, facilitating large scale genome skimming studies from museum specimens. Given the current biodiversity crisis and lack of taxonomic expertise, it has become more important than ever to document biodiversity before it is lost. By sequencing natural history collections at scale using bioinformatic tools such as *go_batch*, researchers can increase the rate of phylogenetic and barcoding studies, and ultimately species discovery.

Acknowledgements

We thank Andrea Waeschenbach for sequence data for *'Solariella' varicosa;* QuantaBio for providing the sparQ DNA Library Prep kit and sparQ PureMag Beads; Tore Høisaeter for providing specimens of '*Solariella' obscura* and '*Solariella' varicosa*; Anders Warén for providing specimens of *Solariella amabilis* and loaning specimens of *Chonospeira nuda* and '*Zetela' alphonsi*; Dai Herbert for loaning specimens of *Ilanga whitechurchi* and '*Spectamen' franciscanum*, Ian Loch for loaning specimens of *Archiminolia oleacea*; Bruce Marshall for loaning *Bathymophila gravida*, *Zetela textilis, Z. kopua* and the holotype of *Phragmomphalina tenuiseptum* and Barbara Buge for loaning the remaining specimens from MNHN. Thanks also to Jon Kongsrud for specimen locality data and to Harry Taylor for photos of NHMUK and MNHN specimens, Bruce Marshall for the photo of *Zetela kopua* and Sadie Mills for the photo of *Phragmomphalina tenuiseptum*. MNHN Specimens were obtained during research cruises and expeditions organized by the MNHN with Institut de Recherche pour le

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Développement and National Taiwan University, as part of the Tropical Deep-Sea Benthos program: ZhongSha 2015 in the South China Sea; MADEEP (dx.doi.org/10.17600/14004000) in Papua New Guinea; BERYX 11 (dx.doi.org/10.17600/92005011), EXBODI (dx.doi.org/10.17600/11100080) and NORFOLK 1 (dx.doi.org/10.17600/1100050) in New Caledonia; BORDAU 1 (dx.doi.org/10.17600/99100020) in Fiji; KARUBENTHOS 2012 and KARUBENTHOS 2 (dx.doi.org/10.17600/15005400) in Guadeloupe (more information can be found at expeditions.mnhn.fr). These expeditions operated under the regulations then in force in the countries in question and satisfy the conditions set by the Nagoya Protocol for access to genetic resources. Finally, we would like to thank Rutger Vos and Dick Groenenberg for their initial feedback on the go_batch pipeline.

References

Bartolozzi, L., Bettison-Varga, L., & Chernetsov, N. (2023). A global approach for natural history museum collections. *Science*, *379*(6638), 1192–1194. https://www.researchgate.net/publication/369480306

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., & Stadler, P. F. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, *69*(2), 313–319. https://doi.org/10.1016/j.ympev.2012.08.023

Bohmann, K., Mirarab, S., Bafna, V., & Gilbert, M. T. P. (2020). Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Molecular Ecology*. https://doi.org/10.1111/mec.15507

Cai, L., Zhang, H., & Davis, C. C. (2022). PhyloHerb: A high-throughput phylogenomic pipeline for processing genome skimming data. *Applications in Plant Sciences*, *November 2021*, 1–9. https://doi.org/10.1002/aps3.11475

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1–9. https://doi.org/10.1186/1471-2105-10-421

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol*, *17*(4), 540–552. https://academic.oup.com/mbe/article/17/4/540/1127654

Chen, S., Zhou, Y., Chen Y, & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, W., Achakkagari, S. R., & Strömvik, M. (2022). Plastaumatic: Automating plastome assembly and annotation. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.1011948

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Freudenthal, J. A., Pfaff, S., Terhoeven, N., Korte, A., Ankenbrand, M. J., & Förster, F. (2020). A systematic comparison of chloroplast genome assembly tools. *Genome Biology*, *21*(1), 1–21. https://doi.org/10.1186/s13059-020-02153-6

Jin, J. J., Yu, W. Bin, Yang, J. B., Song, Y., DePamphilis, C. W., Yi, T. S., & Li, D. Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(241). https://doi.org/10.1101/256479

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kim, E., Kim, J., Lee, Y., Nah, G., & Kim, H. Y. (2022). The complete mitochondrial genome of Turbo cornutus (Trochida: Turbinidae) and its phylogeny analysis. *Mitochondrial DNA Part B: Resources*, *7*(4), 637–639. https://doi.org/10.1080/23802359.2022.2060764

Laetsch, D. R., Blaxter, M. L., & Leggett, R. M. (2017). BlobTools : Interrogation of genome assemblies [version 1 ; referees : 2 approved with reservations]. *F1000Research 2017*, *6*(1287), 1–16.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Meng, G., Li, Y., Yang, C., & Liu, S. (2019). MitoZ: A toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Research*, *47*(11). https://doi.org/10.1093/nar/gkz173

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Pouchon, C., Boyer, F., Roquet, C., Denoeud, F., Chave, J., Coissac, E., Alsos, I. G., Lavergne, S., Smyčka, J., Boleda, M., Thuiller, W., Gielly, L., Taberlet, P., Rioux, D., Hombiat, A., Bzeznick, B., Alberti, A., Wincker, P., Orvain, C., … Wincker, P. (2022). ORTHOSKIM: In silico sequence capture from genomic and transcriptomic libraries for phylogenomic and barcoding applications. *Molecular Ecology Resources*, *22*(5), 2018–2037. https://doi.org/10.1111/1755-0998.13584

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, *70*(1). https://doi.org/10.1002/cpbi.102

R Core team. (2020). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing* (3.6.3). R Foundation for Statistical Computing.

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ruane, S., & Austin, C. C. (2017). Phylogenomics using formalin-fixed and 100+ year-old intractable natural history specimens. *Molecular Ecology Resources*, *17*(5), 1003–1008. https://doi.org/10.1111/1755-0998.12655

Sarmashghi, S., Balaban, M., Rachtman, E., Touri, B., Mirarab, S., & Bafna, V. (2021). Estimating repeat spectra and genome length from low-coverage genome skims with RESPECT. *PLoS Computational Biology*, *17*(11). https://doi.org/10.1371/journal.pcbi.1009449

Sarmashghi, S., Bohmann, K., Thomas P Gilbert, M., Bafna, V., & Mirarab, S. (2017). Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology*, *23*(34), 1–20. https://doi.org/10.1101/230409

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE*, *11*(10). https://doi.org/10.1371/journal.pone.0163962

Steenwyk, J. L., Buida, T. J., Li, Y., Shen, X. X., & Rokas, A. (2020). ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biology*, *18*(12). https://doi.org/10.1371/journal.pbio.3001007

Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., & Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*, *99*(2), 349–364. https://doi.org/10.3732/ajb.1100335

Straube, N., Lyra, M. L., Paijmans, J. L. A., Preick, M., Basler, N., Penner, J., Rödel, M. O., Westbury, M. V., Haddad, C. F. B., Barlow, A., & Hofreiter, M. (2021). Successful application of ancient DNA extraction and library construction protocols to museum wet collection specimens. *Molecular Ecology Resources*, *21*(7), 2299–2315. https://doi.org/10.1111/1755-0998.13433

Sumner-Rooney, L., Sigwart, J. D., McAfee, J., Smith, L., & Williams, S. T. (2016). Repeated eye reduction events reveal multiple pathways to degeneration in a family of marine snails. *Evolution*, *70*(10), 2268–2295. https://doi.org/10.1111/evo.13022

Williams, S. T., Foster, P. G., & Littlewood, D. T. J. (2014). The complete mitochondrial genome of a turbinid vetigastropod from MiSeq Illumina sequencing of genomic DNA and steps towards a resolved gastropod phylogeny. *Gene*, *533*(1), 38–47. https://doi.org/10.1016/j.gene.2013.10.005

Williams, S. T., Kano, Y., Warén, A., & Herbert, D. G. (2020). Marrying molecules and morphology: First steps towards a reevaluation of solariellid genera (Gastropoda: Trochoidea) in the light of molecular phylogenetic studies. *Journal of Molluscan Studies*, *86*(1), 1–26. https://doi.org/10.1093/mollus/eyz038

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Williams, S. T., Noone, E. S., Smith, L. M., & Sumner-Rooney, L. (2022). Evolutionary loss of shell pigmentation, pattern, and eye structure in deep-sea snails in the dysphotic zone. *Evolution*, *76*(12), 3026–3040. https://doi.org/10.1111/evo.14647

Williams, S. T., Smith, L. M., Herbert, D. G., Marshall, B. A., Warén, A., Kiel, S., Dyal, P., Linse, K., Vilvens, C., & Kano, Y. (2013). Cenozoic climate change and diversification on the continental shelf and slope: Evolution of gastropod diversity in the family Solariellidae (Trochoidea). *Ecology and Evolution*, *3*(4), 887–917. https://doi.org/10.1002/ece3.513

Wu, P., Xu, C., Chen, H., Yang, J., Zhang, X., & Zhou, S. (2021). NOVOWrap: An automated solution for plastid genome assembly and structure standardization. *Molecular Ecology Resources*, *21*(6), 2177–2186. https://doi.org/10.1111/1755-0998.13410

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*. https://doi.org/10.1186/s12859-018-2129-y

Supporting Information

Tables

Supplementary Table 1 - Summary of sample quality for extracted DNA and details of factors affecting DNA including the year of sample collection, preservative (ethanol or dry shell), if shell was cracked to allow penetration of ethanol, year DNA was extracted, DNA Integrity Number (DIN) and the amplification success for four partial gene sequences (nuclear 28S rRNA, and mitochondrial genes: COI, 16S rRNA and 12S rRNA). PCR success is summarised as GenBank number if PCR were successful and published, "SEQ" if the PCR was successful and sequenced but unpublished, or "PCR only" if a band was observed when the PCR amplicon was run on a gel but attempts to sequence the amplicon were unsuccessful.

Specimen	Year collected	Preservative	Shell cracked?	Year DNA extracted	DIN	PCR 28S	PCR COI	PCR 16S	PCR 12S
Archinimolia oleacea	1977	70% ethanol	N	2011	none	–	–	–	–
Arxellia herosae	2001	dry	–	2011	1.8	–	–	–	HF585844
Bathymophila gravida	2001	80% ethanol	N	2010	1	PCR only	PCR only	–	–
‘Bathymophila' sp. 18	2011	95% ethanol	N	2013	6.2	LT575957	–	LT575910	LT575928
Bathymophila-Like sp. 12	1999	dry	–	2010	1	–	–	–	HF585775
Chonospeira nuda	2009	95% ethanol	N	2013	1.9	–	SEQ	SEQ	SEQ
Clade D sp. d	2015	95% ethanol	Y	2020	1.3	–	SEQ	–	–
Elaphriella wareni	2014	95% ethanol	Y	2019	2.8	SEQ	–	–	SEQ
Ilanga whitechurchi	2013	99%? ethanol	N	2014	none	–	–	OK393755	–
Lamellitrochus sp. 6	2015	95% ethanol	Y	2020	5.8	OK393809	OK392062	OK393760	OK393784
‘Lamellitrochus' carinatus	2012	95% ethanol	Y	2013	5.7	SEQ	SEQ	SEQ	SEQ
Microgaza rotella	2012	95% ethanol	Y	2013	6.1	LT575964	LT575902	LT575920	LT575947
Phragmomphalina tenuiseptum	1988	80% ethanol	N	2010	1.6	PCR only	PCR only	–	PCR only
Solariella amabilis	1970	70% ethanol	N	2011	1	–	–	–	HF585871
Solariella sp. 7	1992	dry	–	2011	none	–	–	–	HF585874
‘Solariella' carvalhoi	2015	95% ethanol	Y	2020	6.5	OK393814	OK392068	OK393764	OK393789
‘Solariella' obscura	1967	70% ethanol	N	2011	1	–	–	–	PCR only
'Solariella' varicosa	1967	70% ethanol	N	2011	–	–	–	–	HF585720
Spectamen cf. bellulum	2005	99% ethanol	N	2010	3.3	SEQ	–	PCR only	HE800677
‘Spectamen’ franciscanum	1995	75% ethanol	N	2010	1	–	PCR only	–	PCR only

Suavotrochus lubricus	2015	95% ethanol	Y	2020	3.7	SEQ	SEQ	SEQ	SEQ
‘Suavotrochus’ sp. 2	2015	95% ethanol	Y	2020	5.7	SEQ	SEQ	SEQ	SEQ
‘Zetela’ alphonsi	2006	70% ethanol	N	2010	none	PCR only	–	–	PCR only
Zetela kopua	1979	80% ethanol	N	2010	1.8	PCR only	PCR only	–	–
Zetela textilis	1974	80% ethanol	N	2010	none	–	PCR only	–	–

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.11.552985; this version posted September 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

The least degraded samples had a DIN of 3-6.2 and sufficient DNA was available to add 10 ng to each reaction. A fragmentation time of 16 mins was found to be sufficient to create libraries of 150-215 bp, the adaptor was diluted 1 in 5, and the library was amplified with 10 PCR cycles.

Degraded samples (DIN<3) were treated individually with trial and error at each step of library preparation. Libraries were prepared a few at a time with adjustments made to subsequent library preps based on QC results. Many of these samples were also of low concentration, meaning it was not possible to add the recommended 10ng DNA per reaction. To avoid further damage to the DNA, it was not concentrated; the maximum available volume of dilute sample was used, and the protocol adjusted for low input, as detailed in the user protocol provided with the kit. After library preparation all libraries were analysed with a Tapestation 2200 D1000 kit (Agilent).

A fragmentation time of 4 minutes was initially trialled for a subset of particularly poor samples (DIN<2), but following a comparison of 4 and 10 mins, there was little discernible effect on library quality. For ease of processing, all subsequent libraries were made with 10 min fragmentation time.

Many libraries from degraded samples showed high concentrations of adaptor-dimer and so were cleaned using SparQ PureMag beads (QuantaBio) at 1.8x. A repeat PCR was performed on 2 of the libraries which showed extremely high quantities of adaptor-dimer, low library concentration and possible bubble product.