MASTER THESIS

Mikal Meltvik

Candidate number: 3

BI309F Master Thesis in Marine Ecology

A PCR-free approach to meta-mitogenomics of benthic polychaetes

Date: May 15, 2017

Total number of pages: 55

NORD University

www.nord.no

Preface and acknowledgements

This thesis is part of the Master of Science in Marine Ecology with the Faculty of Biosciences and Aquaculture at Nord University.

I wish to extend my deepest gratitude to all of my supervisors. Melissa M. Brandner for countless hours of assistance and advice, both in wet and dry lab. Truls B. Moum for spoon-feeding me relevant literature and keeping me on the right track. Henning Reiss for your tremendous contribution in terms of identifying most of our captured polychaete specimens. I'd like to thank all three of you for your extensive assistance during field work, all the advice and continuous support throughout this project.

A great thanks goes out to Tor Erik Jørgensen and Martina E. L. Kopp for showing me the ropes around the laboratories as well as for keeping the place up and running.

To all my friends, in Bodø and around the world, you have kept me relatively sane through many long days and nights of pipetting and writing. For that I owe you the most sincere thanks. I wish to thank my family for supporting me and showing interest in my somewhat narrow field of study. To every single member of this faculty, I am sure that in one way or another, you have contributed to this work. Thank you for that.

Lastly, I'd like to thank you, the reader, for spending your precious time reading this. I sincerely hope that the following chapters won't bore you tremendously. Quite contrarily, I hope that you will find it a somewhat fascinating read.

Bodø, May 2017
Mikal Meltvik

A PCR-free approach to meta-mitogenomics of benthic polychaetes

Mikal Meltvik
Nord University

May 15, 2017

Abstract

Assessing biodiversity in marine benthic environments has proved difficult, and such ventures may benefit greatly from molecular approaches. The use of DNA barcoding and metabarcoding is gaining traction and may soon be the most effective way to identify large amounts of benthic organisms from bulk samples.

A multilocus metabarcoding approach will allow for utilizing greater amounts of sequence data and identify species even with low coverage sequence data. Building a more complete reference database is imperative to future metabarcoding using multiple loci.

Benthic polychaetes collected with sediment samplers in Saltfjorden, Norway had their DNA extracted individually into separate sequencing libraries for sequencing on a single MiSeq chip. Reads from this single run were subsequently pooled to simulate a single sequencing library made from equiconcentrated pooled DNA. This lets us compare assemblies from individually assembled samples to a bulk assembly.

When pooling as many as 41 polychaete species, using the MiSeq platform, varying fractions of mitogenomes—some close to complete—could be reconstructed although no complete mitogenomes were assembled. Using exclusively the CO1 gene, 19 out of 41 species could be identified. This finding suggests great potential for a multilocus metabarcoding approach, even with low coverage sequence data.

Keywords— multiplex, mitochondrial, Polychaeta, genomics, mtDNA, Saltfjorden, benthic, aquaculture, barcoding, MiSeq

Contents

1 Introduction

1.1 Background

Organismal communities of aquatic habitats, and particularly benthic environments, can be difficult to study as they aren't easily accessed and house many previously unrecorded species. Conducting biodiversity analyses on benthic species requires extensive taxonomic expertise as well as comprehensive knowledge of the study area. An increasingly relevant method relying on nucleotide sequences for species identification, known as DNA barcoding, may aid biodiversity analyses greatly.

The use of DNA barcoding allows biologists to identify and describe species using a select few molecular markers—for instance the mitochondrial gene Cytochrome C oxidase subunit 1 (CO1) widely used for animal delimitation—as unique barcodes that are distinguishable between species but bordering on identical within species (Hebert et al., 2003).

Metabarcoding takes this approach one step further and identifies multiple specimens, usually by amplifying and sequencing a specific gene, from a bulk or pooled sample to identify all the specimens present (Yu et al., 2012). The ability to rapidly identify several species from a pooled sample using metabarcoding may make biodiversity assessments of benthic communities more streamlined, cost-effective and reproducible.

Today DNA barcoding largely relies on PCR, which introduces amplification errors and taxonomic PCR bias, proving especially problematic in metabarcoding of eDNA and other pooled samples (Taberlet et al., 2012). An approach utilizing shotgun sequencing of pooled, equiconcentrated DNA from multiple specimens may alleviate the current reliance on PCR.

Amplicon sequencing is also known for producing chimeric sequences (Judo et al., 1998; Meyerhans et al., 1990), sometimes resulting in the false identification of a new species. Shotgun-sequencing without PCR is expected to produce lower amounts of chimeric sequences.

Individual assembly of each pooled sample is therefore imperative to verify a new pipeline's efficiency at producing non-chimeric assemblies from pooled specimen samples. Additionally, multiple tools have been developed with chimera-detection in mind (Edgar et al., 2011; Huber et al., 2004), which may help with avoiding composite sequences when working with bulk DNA samples.

In the future, nuclear markers such as—18s and 28s ribosomal RNAs (Floyd et al., 2005; Grant and Linse, 2009; Zimmermann et al., 2011)—are likely to be

used in conjuction with mitochondrial markers for better resolution. While single-gene DNA barcoding may be losing its relevance due to advances in sequencing (Taylor and Harris, 2012), Gillett et al. (2014) found that complete mitochondrial genomes (mitogenomes) may successfully be used to reconstruct phylogenies. A multilocus approach, relying on multiple mitochondrial markers, appears to be more fruitful when metabarcoding with low coverage sequencing data. This doesn't restrict analysis to a single gene—that may or may not be present in sequence data—while disregarding all the other potential genes that could aid in species identification (Tang et al., 2014). Mitochondrial markers have relatively high mutation rates, and recombination is restricted, as they are mostly uniparentally and clonally inherited (Birky, 1995). This clonal mode of inheritance however means that, despite investigating multiple mitochondrial genes, one is effectively dealing with a single locus. Despite this, in the present study we define *multilocus* as pertaining to multiple genes within the mitochondrial genome.

1.2 Polychaeta as focus taxon

Polychaeta represents a diverse class of segmented, bristled worms that are excessively abundant in marine soft-bottom sediment habitats (Read and Fauchald, 2017). Their ubiquitous nature make them a relevant taxon to study for environmental analyses and biodiversity assessments in marine environments.

Environmental impact analyses use measures of biodiversity to determine level of ecosystem deterioration in the soft bottom sediment, e.g. certain stages of eutrophication is characterized by a high abundance of opportunistic polychaetes (i.e. *Capitella capitata, Heteromastus filiformis* and *Paramphinome jeffreysii*), but lower overall biodiversity (Bannister et al., 2014; Taranger et al., 2015). Sequencing and assembling the full mitogenomes of common benthic species will help build a reference database for metabarcoding, thus reducing the need for specific taxonomic expertise to determine the presence of the species of interest.

Certain groups of polychaetes, namely the family Serpulidae, have proven difficult to amplify for barcoding as the available CO1 primers appear to be unfit for this group (Halt et al., 2009; Sun et al., 2012). The CO1 sequence has however turned out to be a good marker for species identification for polychaetes (Carr et al., 2011), given that one can get around the PCR-related difficulties. Sequencing the full mitogenomes allows for species identification using any desired mitochondrial marker regardless of its inclination to amplify during PCR.

As of the 17th of April 2017, there were 88 mitochondrial genomes of polychaetes on NCBI's GenBank using the search terms txid6341[Organism:exp] mi-

tochondrion AND complete. Expanding this coverage is crucial to make a good reference database for metabarcoding this group.

1.3 Main aim and design

The main aim of the present study was to explore the possibility of assembling multiple mitogenomes from shotgun-sequenced, pooled total DNA of several marine polychaete species. With the aim of achieving cost-effective mitogenome assembly, Illumina's MiSeq platform was chosen, based on its relatively long read lengths and a low total cost of sequencing.

Additionally, we wanted to explore how the PCR-free metabarcoding approach would perform on the present selection of polychaete species, given the data obtained in the present study and the current database of CO1 barcode sequences.

By barcoding each of the species specific DNA libraries with unique index primers and pooling all of the sequences, we were able to simulate a pooled DNA sequencing approach. At the same time, we were able to analyze each specimen separately for verification purposes.

Three hypotheses were put forth to test whether one can obtain multiple mitogenomes in a rapid, cost-effective fashion with pooled sequencing and assembly. The following null hypotheses were formulated:

- There is no difference in bulk-assembled and individually assembled mitochondrial genomes.
- Mitochondrial genomes assembled with different pipelines are not different.
- Mitochondrial genomes assembled with the described pipelines are not different from known conspecific reference genomes.

2 Methods

2.1 Sample collection

Sediment samples were collected using a 15 L Van Veen Grab Sampler at six different stations in Saltfjorden (see Table 1), Norway on the 22nd and 23rd of February 2016. For some stations two replicates were taken to increase sample size. Sediment was rinsed with seawater over a large sieve with 1 mm mesh size. All polychaetes were rinsed out from the sieve and kept alive in jars of seawater for transportation to the laboratory in Mørkvedbukta research station, Bodø, Norway.

Table 1: The sampling stations and their respective locations, depths and sampling times.

Station	Replicate	Latitude	Longitude	Depth (m)	Date and Time
Valosen	1	67.17.178	14.38.428	19.2	23.02.2016 10:00
Valosen	2	67.17.162	14.38.403	21.2	23.02.2016 10:07
Deep	1	67.15.522	14.35.763	374	23.02.2016 09:24
INTD	1	67.15.159	14.38.543	220	23.02.2016 08:57
INTS	1	67.14.867	14.39.713	100	22.02.2016 10:11
INTS	2	67.14.853	14.39.794	98	22.02.2016 10:25
Shallow	1	67.14.474	14.39.84	60	22.02.2016 09:42
Shallow	2	67.14.474	14.39.84	57	22.02.2016 10:00

2.2 Species identification and preservation

All polychaete specimens were morphologically identified down to family or species level by experts. Specimens that weren't successfully identified to species level were sent to the University of Bergen for identification by experts. Identified specimens were kept in separate, labeled 2 mL tubes on 96 % ethanol. All species' names validity was verified with The World Register of Marine Species (Horton et al., 2016) and changed to accepted species names where applicable.

2.3 DNA extraction

2.3.1 Mitochondrial isolation

Following the procedures for mitochondrial isolation described in Tschischka et al. (2000), with multiple centrifugation steps, minute amounts of mitochondria were isolated from one *Hediste diversicolor* individual. DNA was then extracted from this precipitate using a QIAprep Spin Miniprep Kit (Qiagen). Due to low DNA yields, this approach was abandoned and extracts were not used in subsequent analysis.

2.3.2 Tissue pre-treatment - Removal of preservative

All tissue used in extraction was submerged in 1 mL of phospate-buffered saline solution for 1 minute. This was done twice in separate baths to rinse off any residue of the 96 % ethanol preservative.

2.3.3 Tissue lysis and extraction

DNA was extracted using a DNeasy Blood & Tissue Kit (Qiagen) according to protocol. Tissue was lysed in the kit's provided ATL buffer and Proteinase K for 2–3 hours at 54–56 °C, briefly vortexing the solution every 30 minutes. Tissue disruption was omitted altogether as lysis alone was sufficient. Several specimens were lysed in their entirety to maximize DNA yield, while larger specimens had 4–5 segments cut out from the anterior end post-peristomium. For an overview of starting material mass and DNA yields see Table A4 in Appendix.

DNA was eluted in 100 µL buffer EB (Qiagen) in two steps using 50 µL each. Elution buffer was left to incubate in the spin column at 25 °C for 2 minutes before being spun through. This was done for both elution steps.

2.3.4 DNA quantity and quality scoring

Concentrations of DNA were quantified using a Qubit (ThermoFisher) broad range assay for dsDNA. DNA purity was determined with Nanodrop-1000 (Thermo Scientific) spectrophotometry according to the user manual instructions.

Using a 2200 TapeStation System (Agilent Technologies) with Genomic ScreenTapes, all samples were scanned for degradation and level of fragmentation prior to library preparation.

2.4 Library preparation

Prior to library preparation, all genomic DNA samples were quantified and quality-checked using Genomic DNA ScreenTapes with the 2200 TapeStation System (Agilent Technologies).

In 50 µL AFA Fiber microtubes (Covaris), 250 ng of genomic DNA was eluted in elution buffer EB to a total volume of 50 µL. Using the parameters in Table 2, DNA was fragmented with a Covaris S2 (Covaris) Focused-ultrasonicator aiming at a insert size of 400bp. Fragmented DNA was run on D1000 High Sensitivity tapes with the 2200 TapeStation System (Agilent Technologies) to confirm that the fragments were of desired size.

The fragmented genomic DNA underwent end repair, adapter ligation and size selection using the NEBNext Ultra II Library Prep Kit for Illumina (New England Biolabs Inc.) according to protocol. Size selection and PCR cleanup was carried out using Mag-Bind RxnPure Plus (Omega Bio-tek) beads as per the library preparation protocol. Each library was assigned a uniqued index primer from the NEBNext Multiplex Oligos for Illumina (New England Biolabs Inc.) kit.

Table 2: Fragmentation parameters used with the S2 ultrasonicator.

Parameter	Value
Intensity	5
Duty cycle	5 %
Cycles per burst	200
Treatment duration (s)	55
Temperature (°C)	7
Water level	12

Quantitation of prepared libraries was carried out using NEBNext Library Quant Kit for Illumina (New England Biolabs Inc.) with a StepOnePlus Real-Time PCR System (ThermoFisher Scientific) according to protocol.

2.5 Sequencing on the Illumina MiSeq

DNA libraries were normalized at 3 nM and subsequently pooled by adding 10 µL from each diluted library to a new tube. After thorougly mixing the pooled libraries, 5 µL was taken out and mixed with 5 µL 0.2 N NaOH by pipetting the entire volume up and down repeatedly. This mixture was left to incubate at room temperature for 5 minutes before 990 µL of HT1 buffer (Illumina) was added, resulting in a 15 pM library. In a new tube, 240 µL of the 15 pM library was mixed with 360 µL of HT1 to dilute to a final 600 µL of 6 pM library for sequencing.

2.6 Read trimming and quality assessment

The resulting reads from the sequencing run were demultiplexed using Illumina's package bcl2fastq prior to quality assessment and trimming. Adapter readthrough was trimmed from sequence data using cutadapt (Martin, 2011) in paired mode (A.6.1.1), while discarding very short reads ($<$ 200 bp). Quality assessment was done using the fastx-toolkit package without undertaking any quality trimming before assembly. A histogram displaying read length distribution for each specimen was plotted using the script provided in Appendix A.6.1.3.

For bulk assembly, raw demultiplexed reads were pooled (see shell script in Appendix A.6.1.2) and then trimmed for adapter readthrough using cutadapt in paired mode, discarding short reads ($<$ 200 bp).

2.7 Assembly

Assembly was carried out using a relevant pipeline described in contemporary literature, additionally a shorter pipeline described in this study. The following pipelines were run on the same sequence data after adapter trimming and short read removal. All assembly-related operations were run with the unix command *time* to record required CPU time.

2.7.1 Assembly of filtered mitochondrial reads

With a database consisting of 88 polychaete mitogenomes (available on the 17th of April) from GenBank, a BLAST (Morgulis et al., 2008; Zhang et al., 2000) search with > 30 % identity, E-value $\leq 10^{-5}$ was used to find mitochondrial-like sequences in the reads (Crampton-Platt et al., 2015; Tang et al., 2014). These reads were then extracted using a shell script (see Appendix A.6.2.1). This script also calls two perl scripts described in Tang et al. (2014), where all candidate mitochondrial (from all included libaries) reads extracted with the BLAST search are used as a database for a second 51-mer search to extract additional reads that resemble these first mitochondrial-like reads. Note that as one of the described scripts didn't work as intended, a simple workaround was made to extract the reads following the 51-mer search (see Appendix A.6.2.1).

Filtered reads were assembled using soapdenovo2 (Luo et al., 2012) with the arguments *-K 61 -u -R -k 45*, soapdenovo-trans (Luo et al., 2012) with k-mers of 71, and IDBA-UD (Peng et al., 2012) with standard run parameters. See Appendix A.6.2.3 for all options. Soapdenovo's configuration files for individual libraries were tailored to each library's specific insert size based on D1000 tape measurements (generated with script A.6.2.2). For bulk assembly the mean insert size across all libraries was used.

Contigs from the three different assemblers were pooled and short contigs ($<$ 1000 bp) were discarded.

2.7.2 Assembly with minia sans filtering

Using minia (Chikhi and Rizk, 2012) all libraries were individually assembled using a k-mer size of 61 and default parameters (see Appendix A.6.3.1 for script). Bulk assembly was run with the same parameters on the pooled reads.

Short contigs ($<$ 1000 bp) were discarded from each separate assembly as well as from the bulk contig pool.

2.7.3 Read mapping with bowtie2

Using bowtie2 (Langmead and Salzberg, 2012), demultiplexed sequence reads from *Owenia fusiformis* was mapped to an appropriate reference genome (GenBank accession: NC_028712.1). The resulting SAM file was imported into ugene (Okonechnikov et al., 2012) and a consensus sequence was exported.

2.8 Assembly comparison

2.8.1 Visual comparison

For a simple visual comparison of the different assembly approaches, contigs from each assembly pipeline were mapped onto selected reference genomes and plotted using BRIG (Alikhan et al., 2011). Most specimens were mapped to reference species in the same taxonomic order. In some cases where closer references were available, confamilials were grouped together and mapped to a fitting reference genome from the same family.

2.9 CO1 metabarcoding

With a polychaete CO1 reference database consisting of 383 published records from GenBank (obtained 2nd of May 2017), a BLASTn search was done using sequencing reads as query with relaxed critera (> 30 % identity, E-value $\leq 10^{-5}$). CO1 reads were extracted following this search and then assembled with minia using a k-mer size of 61.

The contigs file from this assembly was used for a regular BLAST search (megablast), to assess the quality of the assembled CO1 sequences. Any significant hits with an identity ≥ 97 % (Smith et al., 2005) were recorded as *detected* in our pooled "ambiguous" sample.

For a more visual approach to species detection, 23 reference CO1 sequences, 2 additional polychaete sequences (not sequenced in the present study), and one earthworm sequence were downloaded from GenBank and used as a reference in a species plot made with BRIG (Alikhan et al., 2011). The two polychaete species *Riftia pachyptila, Paralvinella palmiformis* and the earthworm *Lumbricus terrestris* served as a negative control "outgroup" to sieve out false positives. For this map, both the filtered CO1 reads and the assembled contigs from this dataset were mapped separately for comparison. A FASTA-file with all the reference sequences was set as reference in BRIG and the header for each sequence was used to label *gene features* as separate species throughout the circle.

3 Results

Across 41 libraries, a total of 14,477,628 read pairs were produced from the multiplex MiSeq run. Where 13,061,395 (90.21 %) passed the trimming criteria and were used in subsequent analyis. Out of these reads, 87.71 % were of max length (301 bp), while the rest was fairly evenly spread between lengths of 200 and 300 bp (post trimming). Mitochondrial reads made up 5.65 % of the total read count post-trimming (738,090 out of 13,061,395). Out of these mitochondrial reads, 182,516 were retained from the initial BLAST search whilst the rest were extracted with the 51-mer search. Total reads and mitoreads were weakly, albeit not significantly correlated ($r_s = 0.2655$, p = 0.0935, see Figure 1).

3.1 Assemblies

Assembly with the lightweight assembler, minia, was considerably faster (Table 3) than the somewhat longer pipeline relying on BLAST for filtering and three different assemblers. The contigs produced with the filtering pipeline were however generally longer and mapped more successfully to the reference genomes. With filtering for mitochondrial reads and the use of multiple assemblers, this pipeline, yielded an appreciable increase in overall coverage over assembly sans filtering.

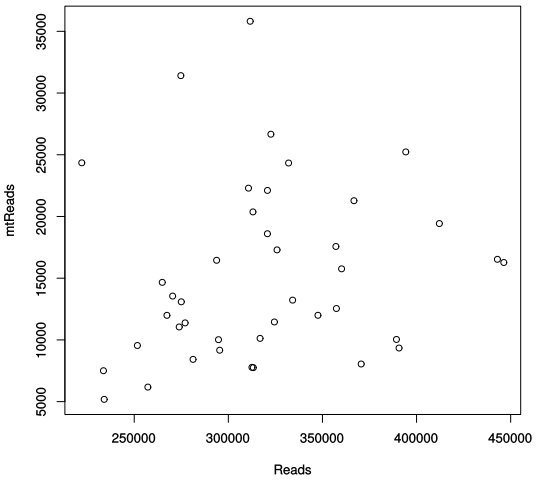

Figure 1: Relationship between total quantity of sequence reads post-trimming and filtered mitoreads. Each point represents one of the 41 libraries.

Contigs in figures 2 and 3 were mapped to conspecifics with more stringent criteria (min. 80 % sequence identity) than what is seen in Figures 4–7. Here more relaxed criteria (min. 50 % sequence identity) were used to ensure successful mapping to distant relatives. *Paraphinome jeffreysii* was not mapped as no references in the same taxonomic order were available. See Appendix A.6.5 for all the full-page mitogenome coverage figures.

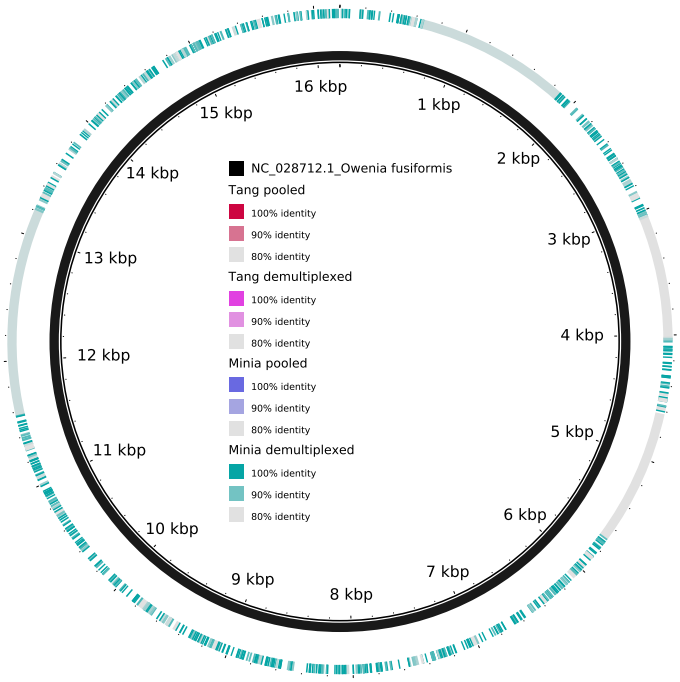

Figure 2: *Owenia fusiformis* mitogenome.

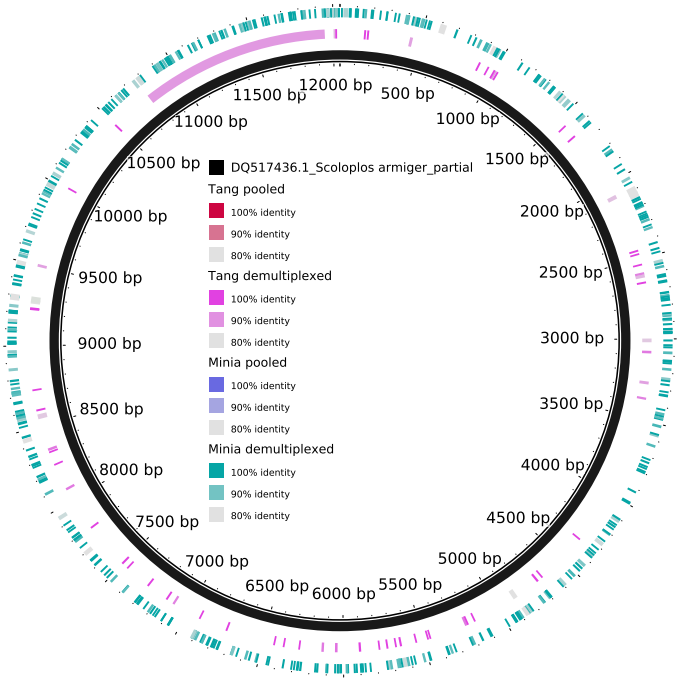

Figure 3: Partial *Scoloplos cf. armiger* mitogenome.

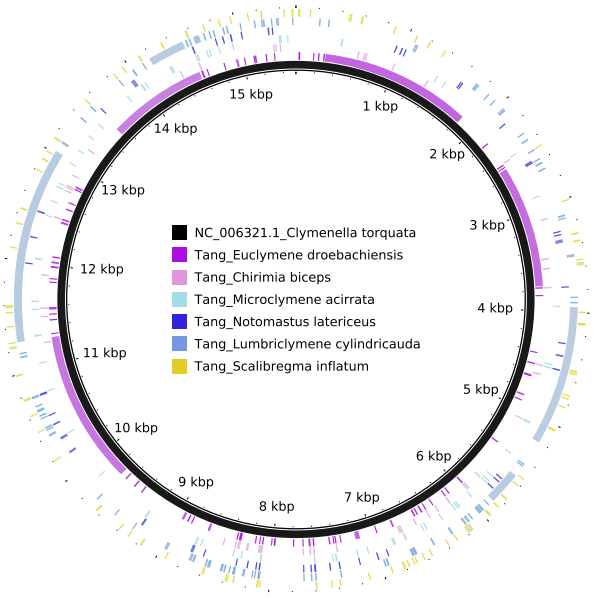

Figure 4: Sampled members of the order Capitellida, assembled from filtered mitoreads, mapped to *Clymenella torquata*.

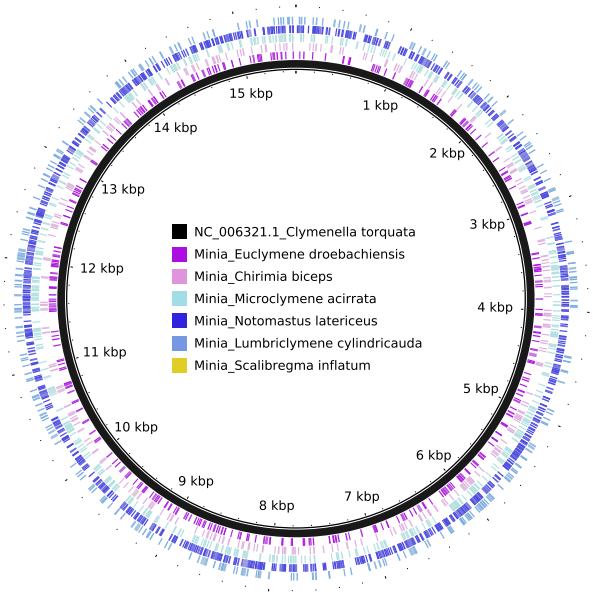

Figure 5: Sampled members of the order Capitellida, assembled with minia pipeline, mapped to *Clymenella torquata*.

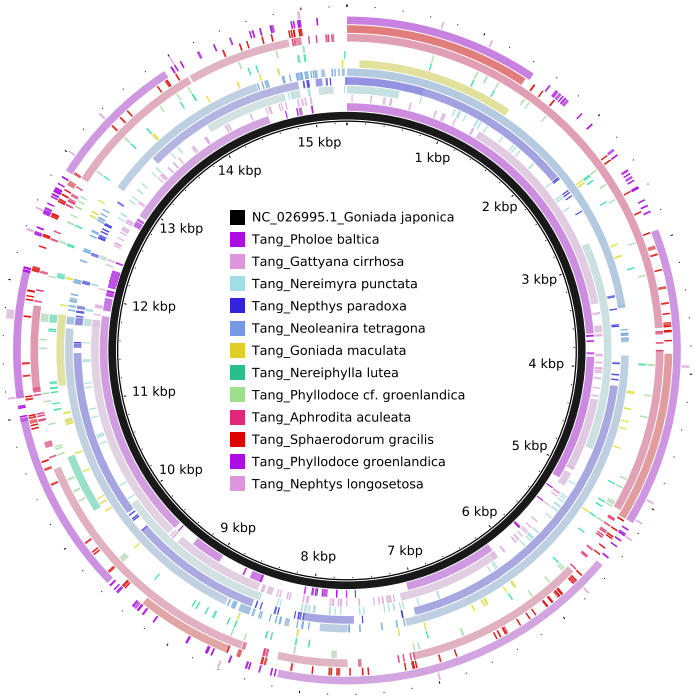

Figure 6: Sampled members of the order Phyllodocida, assembled from filtered mitoreads, mapped to *Goniada japonica*.

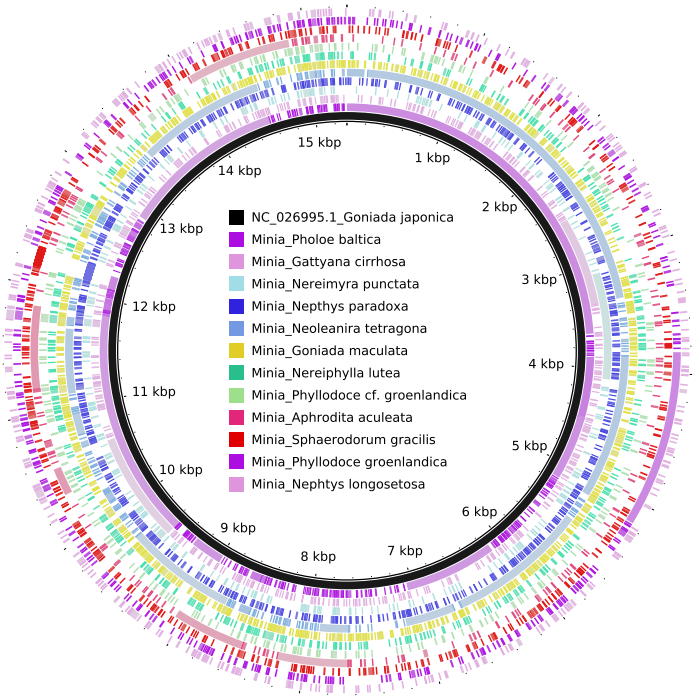

Figure 7: Sampled members of the order Phyllodocida, assembled with minia pipeline, mapped to *Goniada japonica*.

Table 3: Assembly times, including filtering time where applicable. All times given in real time spent actively processing reads with filtering and assembly. Post-assembly processing is not included in these times.

Pipeline	Voucher	Bulk
Filtered	5h 49m 40.3s	5h 15m 31.0s
Minia	41m 24.4s	48m 12.2s

3.2 CO1 metabarcoding

Detected polychaete species from 'BLASTing' all CO1-contigs were *Eclysippe vanelli, Paraphinome jeffreysii, Pholoe baltica, Melinna albicincta, Sphaerodorum gracilis, Owenia fusiformis, Scoloplos armiger, Neoleanira tetragona, Nepthys paradoxa, Hyalinoecia tubicola, Heteromastus filiformis, Amphitrite cirrata, Prionospio dubia, Nereimyra punctata, Spirobranchus triqueter, Spiophanes kroyeri, Phyllodoce groenlandica, Spirobranchus giganteus, Hydroides norvegicus, Laonice sp., Nothria conchylega* and 61 contigs with significant hits where identity was $<$ 97 %. Out of the detected polychaetes *Heteromastus filiformis, Amphitrite cirrata, Prionospio dubia* were not identified during sampling. There were also 71 contigs with no significant hits in BLAST.

Out of the 41 specimens initially sampled, 19 were identified with a BLAST search using the CO1 contigs as a query.

In Figure 8, 13 species were successfully mapped with sequence identity $\geq$ 97 %. The filtered CO1 reads were more inclined to map to the references than the contigs resulting from the assembly. With the exception of a relatively short fragment (10-20 bp) at the beginning of the *Riftia pachyptila* sequence, no sequences mapped to the negative controls.

3.3 Short read mapping

See Appendix A.6.6 for nucleotide sequence: *Owenia fusiformis*, partial mitochondrial genome.

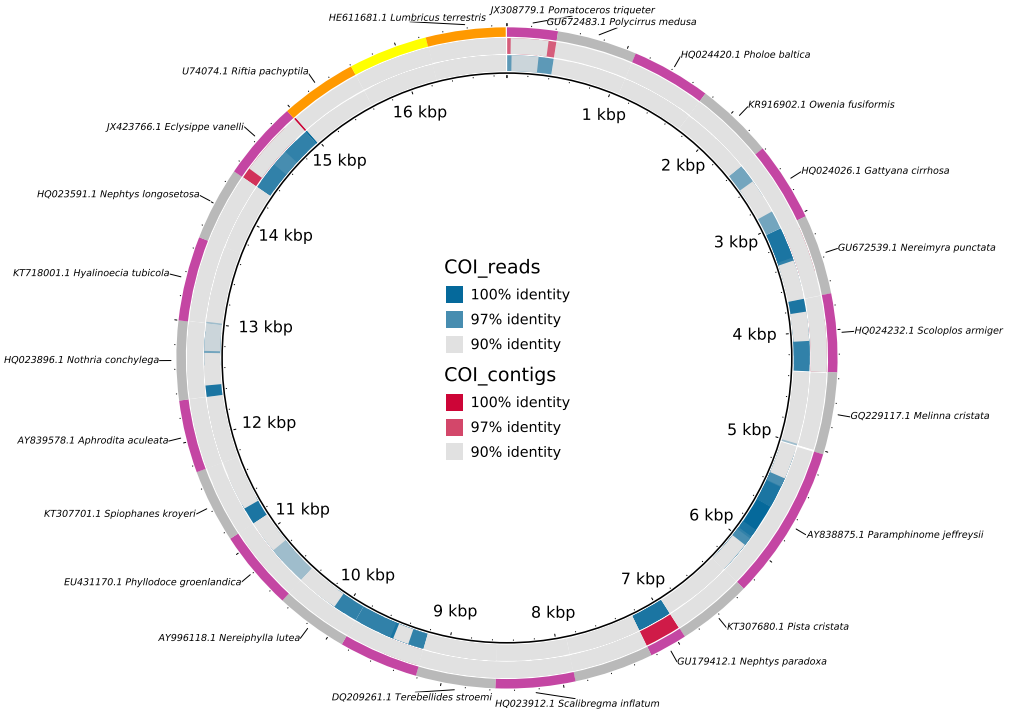

Figure 8: Filtered CO1 reads mapped in the innermost circle, assembled CO1 sequences in the second circle. References on the outermost circle, in alternating colors, are 23 Genbank sequences of sampled polychaetes. Negative control (in orange and yellow), are 2 abritrarily chosen polychaetes not sampled in this experiment as well as one earthworm CO1.

4 Discussion

4.1 Assembly of multiplex-sequenced mitogenomes

Several long contigs were produced, spanning long stretches of their respective reference genomes. Coverage was however too low to reconstruct any complete mitogenomes from the polychaete sequence data. With approximately 25 million sequence reads as an upper limit, the MiSeq (Illumina) platform appears to be suboptimal for sequencing such a high number of specimens with sufficient depth to assemble complete mitochondrial genomes.

Mapping of contigs from bulk assembly yielded—in most cases—lower coverage and consistently lower sequence identity than assembly of demultiplexed reads

(see Figures 2 and 3 for bulk vs. demultiplexed comparisons). With increased sequencing depth, this approach might still be viable. Overall, contigs assembled from extracted mitoreads mapped more completely to reference genomes than the contigs produced from the total DNA.

The first hypothesis—claiming there's no difference in bulk-assembled and individually assembled mitogenomes—can to some extent be rejected as the bulk-assembled mitogenomes didn't map as well to references. Demultiplexed, individually assembled mitogenomes performed better across the board in terms of successful contig mapping to references. The second hypothesis is also put into question as the comparisons clearly show more fragmented mapping with the non-filtered, minia-assembled contigs as opposed to the contigs made with the filtered mitoreads. The last hypothesis cannot be tested without increasing sequencing depth and getting the necessary coverage to produce complete mitogenomes.

In Tang et al. (2014), a total of 230 million reads were produced on the HiSeq 2000 (Illumina) for 49 specimens. This is more than a fifteen-fold increase over the 14.5 million reads aqcuired in this study from 41 specimens. Despite the large discrepancy in total reads, this study recovered over 5 % candidate mitochondrial reads. Close to the same read count as the 0.5 % Tang et al. (2014) extracted. The high number of non-mitochondrial contigs—found by megablast searches—however indicates that a 51-mer search with the same algorithms may be too relaxed for use with this taxonomic group or the longer read length. The relative abundance of mitochondrial reads recovered by a BLAST search prior to the 51-mer search was more in line with the expected ratio.

Every library was prepared with the same protocol, using the same amount of DNA. The final concentration of each library was normalized before pooling to ensure an equal amount of DNA from each sample. Despite this, read count was not equal across all libraries (see Figure 1).

Such inequalities may be caused by interspecific variations in mitochondria count, type and amount of tissue used, adapter-ligation bias (leading to PCR bias down the line in library prep), or minor inaccuracies in normalization of libraries to mention some possibilities.

With that said, the amount of mitochondrial reads in each library correlated weakly with the total amount of reads in the same library (Figure 1). While this is far from conclusive on its own, it is however safe to assume that one may recover more mitochondrial reads by increasing the total read count.

It is worth noting that bulk assembly is a fairly novel approach that appears to be mainly employed when dealing with various coleopterans (Crampton-Platt

et al., 2015; Gillett et al., 2014; Tang et al., 2014). How well such methods perform with different taxonomic groups cannot be confidently estimated. The results of this study indicate that it might work well with polychaetes given a sufficiently large dataset.

In the event that one should reconstruct complete mitochondrial genomes with the help of next-generation sequencing, some Sanger sequencing may still be useful for verifying single-nucleotide polymorphisms (SNPs) and stretches of homopolymers (Baudhuin et al., 2015; Mu et al., 2016), as well as verifying gene order and linking specimens to sequences (Gillett et al., 2014; Tang et al., 2014).

4.2 CO1 metabarcoding

Nineteen out of 41 specimens were identified after filtering CO1 reads from a pooled sample. The fact that this many species were correctly identified with such a small database is promising for future applications of such methods with more complete databases.

For instance, a BLAST search could be made towards a database containing all the $>$ 13,000 polychaete CO1 records found in BOLD's servers. Smaller databases may also be tailored for situations where researchers are looking for a specific set of species endemic to a study area.

Although *Heteromastus filiformis, Amphitrite cirrata, Prionospio dubia* were not identified during sampling, these specimens may have been present and misidentified due to close resemblance to sister taxa.

Figure 8 shows a plot with 26 different reference CO1 sequences. Despite choosing specifically species that were identified during sampling, only 13 out of these 23 actually mapped well enough to be confidently identified. With the low sequencing depth characterizing the raw data this doesn't come as a great surprise. Additionally, it is clear that mapping sequencing reads to CO1 is a better approach than an attempt at assembling these. This is likely due to most assemblers' tendency to produce long contigs and discard reads that weren't used for contigs (Haridas et al., 2011).

In this instance 23 polychaete species were hand-picked as they had been identified during sampling. Given a situation where the sequenced specimens haven't been identified one must know what species are endemic to the sampling area to build an appropriate reference database.

4.3 On mitochondrial isolation

An approach relying on mitochondrial isolation prior to DNA extraction was initially attempted. Due to low yields of mtDNA and the method's labour-intensive and time-consuming nature, this venture was abandoned to extract and sequence total DNA instead and to filter mtDNA bioinformatically.

One can make an educated guess about what implications a successful mitochondrial isolation would entail for sequencing on the given platform. The only available close to complete, nuclear polychaete genome on GenBank as of May 5th 2017 is *Capitella teleta* (Genbank accession no.: KB291798.1) at 1,620,044 bp. The exclusion of such large amounts of nuclear DNA—assuming the nuclear genome sizes of other polychaetes to be similar—would very likely increase mtDNA coverage considerably.

4.4 Conclusions

Relevant and current methodologies for multiplex mitogenome assembly was tested and crudely compared to a less sophisticated pipeline without any read screening aside from adapter trimming. While the former pipeline shows a lot of promise, it performed better when applied to demultiplexed single species than with bulk assembly. With adequate sequence coverage the filtering pipeline is likely to successfully assemble multiple high quality mitogenomes from bulk samples.

Bulk-sequencing over 40 polychaete species on the MiSeq platform proved to not yield high enough coverage to assemble full mitogenomes. For metabarcoding purposes however, it seems to perform well in terms of identifying species, even from a mixed pool of reads. Given a more complete database as well as a multilocus approach to metabarcoding it is likely that one should be able to identify confidently to species level based on pooled mitochondrial reads alone.

5 References

Alikhan, N.-F., Petty, N. K., Zakour, N. L. B., and Beatson, S. A. (2011). Blast ring image generator (brig): simple prokaryote genome comparisons. *BMC genomics*, 12(1):402.

Bannister, R. J., Valdemarsen, T., Hansen, P. K., Holmer, M., and Ervik, A. (2014). Changes in benthic sediment conditions under an atlantic salmon farm at a deep, well-flushed coastal site. *Aquaculture Environment Interactions*, 5(1):29–47.

Baudhuin, L. M., Lagerstedt, S. A., Klee, E. W., Fadra, N., Oglesbee, D., and Ferber, M. J. (2015). Confirming variants in next-generation sequencing panel testing by sanger sequencing. *The Journal of Molecular Diagnostics*, 17(4):456–461.

Birky, C. W. (1995). Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proceedings of the National Academy of Sciences*, 92(25):11331–11338.

Carr, C. M., Hardy, S. M., Brown, T. M., Macdonald, T. A., and Hebert, P. D. (2011). A tri-oceanic perspective: Dna barcoding reveals geographic structure and cryptic diversity in canadian polychaetes. *PLoS One*, 6(7):e22232.

Chikhi, R. and Rizk, G. (2012). Space-efficient and exact de bruijn graph representation based on a bloom filter. In *WABI*, volume 7534 of *Lecture Notes in Computer Science*, pages 236–248. Springer.

Crampton-Platt, A., Timmermans, M. J., Gimmel, M. L., Kutty, S. N., Cockerill, T. D., Khen, C. V., and Vogler, A. P. (2015). Soup to tree: the phylogeny of beetles inferred by mitochondrial metagenomics of a bornean rainforest sample. *Molecular biology and evolution*, 32(9):2302–2316.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200.

Floyd, R. M., Rogers, A. D., Lambshead, P., and Smith, C. R. (2005). Nematode-specific pcr primers for the 18s small subunit rrna gene. *Molecular Ecology Resources*, 5(3):611–612.

Gillett, C. P. D. T., Crampton-Platt, A., Timmermans, M. J. T. N., Jordal, B. H., Emerson, B. C., and Vogler, A. P. (2014). Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, 31(8):2223–2237.

Grant, R. A. and Linse, K. (2009). Barcoding antarctic biodiversity: current status and the caml initiative, a case study of marine invertebrates. *Polar Biology*, 32(11):1629.

Halt, M. N., Kupriyanova, E. K., Cooper, S. J., and Rouse, G. W. (2009). Naming species with no morphological indicators: species status of galeolaria caespitosa

(annelida: Serpulidae) inferred from nuclear and mitochondrial gene sequences and morphology. *Invertebrate Systematics*, 23(3):205–222.

Haridas, S., Breuill, C., Bohlmann, J., and Hsiang, T. (2011). A biologist's guide to de novo genome assembly using next-generation sequence data: a test with fungal genomes. *Journal of microbiological methods*, 86(3):368–375.

Hebert, P. D., Cywinska, A., Ball, S. L., et al. (2003). Biological identifications through dna barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1512):313–321.

Horton, T., Kroh, A., Bailly, N., Boury-Esnault, N., Nunes Brando, S., Costello, M., ..., and Zeidler, W. (2016). World Register of Marine Species (WoRMS). Available from http://www.marinespecies.org [Accessed: 2016-11-07].

Huber, T., Faulkner, G., and Hugenholtz, P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*, 20(14):2317–2319.

Judo, M. S., Wedel, A. B., and Wilson, C. (1998). Stimulation and suppression of pcr-mediated recombination. *Nucleic acids research*, 26(7):1819–1825.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1):18.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10.

Meyerhans, A., Vartanian, J.-P., and Wain-Hobson, S. (1990). Dna recombination during pcr. *Nucleic acids research*, 18(7):1687–1691.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., and Schäffer, A. A. (2008). Database indexing for production megablast searches. *Bioinformatics*, 24(16):1757–1764.

Mu, W., Lu, H.-M., Chen, J., Li, S., and Elliott, A. M. (2016). Sanger confirmation is required to achieve optimal sensitivity and specificity in next-generation sequencing panel testing. *The Journal of Molecular Diagnostics*, 18(6):923–932.

Okonechnikov, K., Golosova, O., Fursov, M., et al. (2012). Unipro ugene: a unified bioinformatics toolkit. *Bioinformatics*, 28(8):1166–1167.

Peng, Y., Leung, H. C., Yiu, S.-M., and Chin, F. Y. (2012). Idba-ud: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428.

Read, G. and Fauchald, K. (2017). World Polychaeta Database. Available from http://www.marinespecies.org/polychaeta/ [Accessed: 2017-02-26].

Smith, M. A., Fisher, B. L., and Hebert, P. D. (2005). Dna barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of madagascar. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1462):1825–1834.

Sun, Y., Kupriyanova, E., and Qiu, J. (2012). Coi barcoding of hydroides: a road from impossible to difficult. *Invertebrate Systematics*, 26(6):539–547.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. (2012). Towards next-generation biodiversity assessment using dna metabarcoding. *Molecular ecology*, 21(8):2045–2050.

Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S., Song, W., Li, Y., Wu, Q., Zhang, A., and Zhou, X. (2014). Multiplex sequencing of pooled mitochondrial genomes - A crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, 42(22):1–13.

Taranger, G. L., Karlsen, Ø., Bannister, R. J., Glover, K. A., Husa, V., Karlsbakk, E., Kvamme, B. O., Boxaspen, K. K., Bjørn, P. A., Finstad, B., et al. (2015). Risk assessment of the environmental impact of norwegian atlantic salmon farming. *ICES Journal of Marine Science: Journal du Conseil*, 72(3):997–1021.

Taylor, H. and Harris, W. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of dna barcoding. *Molecular Ecology Resources*, 12(3):377–388.

Tschischka, K., Abele, D., and Portner, H. (2000). Mitochondrial oxyconformity and cold adaptation in the polychaete nereis pelagica and the bivalve arctica islandica from the baltic and white seas. *Journal of Experimental Biology*, 203(21):3355–3368.

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(4):613–623.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning dna sequences. *Journal of Computational biology*, 7(1-2):203–214.

Zimmermann, J., Jahn, R., and Gemeinholzer, B. (2011). Barcoding diatoms: evaluation of the v4 subregion on the 18s rrna gene, including new primers and protocols. *Organisms Diversity & Evolution*, 11(3):173.

A.6 Appendix

A.6.1 Scripts

Most of the following scripts utilize loops to give all demultiplexed libraries equal treatment in filtering, converting, assembly etc. For bulk samples the exact same commands were run on the pooled fastq and fasta files directly from the command line interface. The looping scripts take a list of sample names—with one sample name per line—as an argument and runs the script within the loop once for each sample.

A.6.1.1 Trim adapter readthrough and short reads

```
#!/usr/bin/env bash

# Pass text file with one sample name per line to iterate through all
    samplenames.

FASTQFOLDER="/media/mikal/Polyseq_Mikal/raw_reads/"
TRIMFOLDER="/media/mikal/Polyseq_Mikal/trimmed_reads_fastq/"

while read p; do
  echo "Trimming adapter sequences for $p";
  # Clip adapter sequence and write report to trim.report
  echo "---TRIM REPORT FOR SAMPLE $p---" >> ${TRIMFOLDER}trim.report;
  cutadapt -m 200 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -A
      AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT -o ${
      TRIMFOLDER}${p}R1.fastq -p ${TRIMFOLDER}${p}R2.fastq ${
      FASTQFOLDER}${p}R1.fastq ${FASTQFOLDER}${p}R2.fastq >> ${
      TRIMFOLDER}trim.report;
done <$1
```

A.6.1.2 Shell script - Pool demultiplexed libraries

```
#!/usr/bin/env bash

# Pass text file with one sample name per line to iterate through all
    samples.

ROOTFOLDER="/media/mikal/Polyseq_Mikal/"

```

```
while read p; do
  echo "Adding $p to the pool...";
  cat ${ROOTFOLDER}raw_reads/${p}R1.fastq >> ${ROOTFOLDER}pooled_reads
      /pooled_R1.fastq;

  cat ${ROOTFOLDER}raw_reads/${p}R2.fastq >> ${ROOTFOLDER}pooled_reads
      /pooled_R2.fastq;
done <$1
```

A.6.1.3 R script - Plot read length distribution

```
# Change "path" and pattern of "filenames".

path <- "/media/mikal/Polyseq_Mikal/trimmed_reads/stats"
filenames <- list.files(path=path, pattern="*readlendist*")

for(i in filenames){
  readlen <- read.table(file = paste(path, i, sep = "/"), header = F,
      sep = " ")

  svg(paste(path, "/", substr(i, 1, nchar(i)-4), ".svg", sep = ""))

  # Title (main) uses filename (i) and trims off from the right end
      of the filename. Use nchar(i)-x, x being amount of letters to
      trim from right end (use 4 to trim off file extension and ".").

  plot(readlen$V1, readlen$V2, type = "h", log="y", main =
      paste(substr(i, 1, nchar(i)-15), "Read length
      distribution"),xlab = "Fragment length (bp)", ylab = "Frequency")

  dev.off()
}
```

A.6.2 Scripts for mitofiltering pipeline

A.6.2.1 Shell script - BLAST and extract mitochondrial reads

```
#!/bin/bash

ROOTFOLDER="/media/mikal/Polyseq_Mikal/trimmed_reads_fastq/"
FAFOLDER="${ROOTFOLDER}tangfiltered_new/"
```

```
SFOLDER="~/Documents/thesis/literature/supplementary/tang2014/scripts/"

mkdir -p ${FAFOLDER}

while read p; do
  mkdir -p ${FAFOLDER}${p}

  echo "Converting $p to fasta...";
  # Report sample name and read numbers to fq2fa.report
  echo "---$p R1---" >> ${FAFOLDER}fq2fa.report;
  fastq_to_fasta -v -n -i ${ROOTFOLDER}${p}R1.fastq -o ${FAFOLDER}${p}/${p}R1.fasta >> ${FAFOLDER}fq2fa.report;
  echo "---$p R2---" >> ${FAFOLDER}fq2fa.report;
  fastq_to_fasta -v -n -i ${ROOTFOLDER}${p}R2.fastq -o ${FAFOLDER}${p}/${p}R2.fasta >> ${FAFOLDER}fq2fa.report;

  echo "Merging fasta reads of $p to single readfile..."
  cat ${FAFOLDER}${p}/${p}R1.fasta ${FAFOLDER}${p}/${p}R2.fasta > ${FAFOLDER}${p}/${p}_comb.fasta

  # BLAST search all reads in database
  echo "BLASTing $p in polymitos database...";
  blastall -p blastn -i ${FAFOLDER}${p}/${p}R1.fasta -F F -d /media/mikal/Polyseq_Mikal/blast_dbs/polymitos -o ${FAFOLDER}${p}/${p}_R1_BLAST -m 8 -e 1e-5

  blastall -p blastn -i ${FAFOLDER}${p}/${p}R2.fasta -F F -d /media/mikal/Polyseq_Mikal/blast_dbs/polymitos -o ${FAFOLDER}${p}/${p}_R2_BLAST -m 8 -e 1e-5

  # Extract read names
  echo "Extracting mitochondrial-like reads from $p based on BLAST results...";
  cut -f1 ${FAFOLDER}${p}/${p}_R1_BLAST > ${FAFOLDER}${p}/${p}_R1_mtreadnames.txt;

  cut -f1 ${FAFOLDER}${p}/${p}_R2_BLAST > ${FAFOLDER}${p}/${p}_R2_mtreadnames.txt;

```

```
  # Extract all unique lines
  sort ${FAFOLDER}${p}/${p}_R1_mtreadnames.txt | uniq > ${FAFOLDER}${p}/${p}_R1_unique_mitoreads.txt;

  sort ${FAFOLDER}${p}/${p}_R2_mtreadnames.txt | uniq > ${FAFOLDER}${p}/${p}_R2_unique_mitoreads.txt;

  # Extract reads matching the unique reads from raw data
  grep -A 1 -F -f ${FAFOLDER}${p}/${p}_R1_unique_mitoreads.txt ${FAFOLDER}${p}/${p}R1.fasta | grep -v '^--' >> ${FAFOLDER}${p}/${p}_mitoreads.fasta;

  grep -A 1 -F -f ${FAFOLDER}${p}/${p}_R2_unique_mitoreads.txt ${FAFOLDER}${p}/${p}R2.fasta | grep -v '^--' >> ${FAFOLDER}${p}/${p}_mitoreads.fasta;

  # Combine all mitoreads for 51-mer set generation
  cat ${FAFOLDER}${p}/${p}_mitoreads.fasta >> ${FAFOLDER}all_mitoreads.fasta;
done <$1

# New loop for perl script and second extraction, as a complete "all_mitoreads.fasta" is required
while read p; do
  echo "reBLASTing $p reads in 51-mer set of all mitoreads";
  cd ${FAFOLDER}${p};

  perl ${SFOLDER}onerun.pl ${FAFOLDER}all_mitoreads.fasta ${p}_comb.fasta 51;

  echo "Extracting $p reads based on second BLAST...";
  # final.pl didn't work as intended. Manual extraction below.
  #perl ${SFOLDER}final.pl ${ROOTFOLDER}${p}R1.fastq ${ROOTFOLDER}${p}R2.fastq idout;

  grep -A 1 -F -f idout ${p}R1.fasta | grep -v '^--' > ${p}_R1_mitoreads.fasta;
  grep -A 1 -F -f idout ${p}R2.fasta | grep -v '^--' > ${p}_R2_mitoreads.fasta;
done <$1
```

A.6.2.2 Shell script - Generate soapdenovo config files

```
#!/usr/bin/env bash

# Pass text file with one sample name per line to iterate through all
    samples. Pass text file with corresponding average insert sizes as
    second argument.

ROOTFOLDER="/media/mikal/Polyseq_Mikal/trimmed_reads_fastq/"
FAFOLDER="${ROOTFOLDER}tangfiltered_new/"

while read p; do
  if ! read -u 3 p2
  then
    break
  fi

  echo "max_rd_len=301
[LIB]
avg_ins=$p2
reverse_seq=0
asm_flags=3
rank=1
f1=${FAFOLDER}${p}/${p}_R1_mitoreads.fasta
f2=${FAFOLDER}${p}/${p}_R2_mitoreads.fasta
" > ${FAFOLDER}${p}/soap_asm.conf
done <$1 3<$2
```

A.6.2.3 Shell script - Soapdenovo and IDBA assembly

Please note that IDBA was compiled from source with a change in config file *src/sequence/short_sequence.h* to allow for reads longer than 128 bp. This was done by changing *static const uint32_t kMaxShortSequence = 128* to *static const uint32_t kMaxShortSequence = 301*.

```
#!/usr/bin/env bash

# Pass text file with one sample name per line to iterate through all
    samples.

ROOTFOLDER="/media/mikal/Polyseq_Mikal/trimmed_reads_fastq/"
```

```
FAFOLDER="${ROOTFOLDER}tangfiltered_new/"
SFOLDER="/home/mikal/hax/soapdenovo-trans/"
IDFOLDER="/home/mikal/hax/idba/bin/"

while read p; do
  mkdir -p ${FAFOLDER}${p}/asm/soaptrans ${FAFOLDER}${p}/asm/IDBA;
  cd ${FAFOLDER}${p}/asm;

  echo "Assembling $p with soapdenovo2..."
  soapdenovo2-63mer all -s ${FAFOLDER}${p}/soap_asm.conf -K 61 -u -R -
      k 45 -o ${p} &> ${p}_assembly.report;

  less ${p}.scafSeq | perl -e 'while(<>){chomp; if(/>/){my @a=split /\
      s+/; print"$a[0]\*$a[1]\_D\n";}else{print"$_\n";}}' > denovo_${p
      }.scafSeq;

  cd soaptrans;
  echo "Assembling $p with soapdenovo-trans..."
  ${SFOLDER}SOAPdenovo-Trans-127mer pregraph -s ${FAFOLDER}${p}/soap_
      asm.conf -K 71 -o ${p}_outgraph;

  ${SFOLDER}SOAPdenovo-Trans-127mer contig -g ${p}_outgraph -M 3;

  ${SFOLDER}SOAPdenovo-Trans-127mer map -s ${FAFOLDER}${p}/soap_asm.
      conf -g ${p}_outgraph -r -f;

  ${SFOLDER}SOAPdenovo-Trans-127mer scaff -g ${p}_outgraph -F -L 100 -
      t 1 -r;

  less ${p}_outgraph.scafSeq | perl -e 'while(<>){chomp; if(/>/){my @a
      =split /\s+/; if(/Locus/){print"$a[0]\*$a[2]\_T\n";}else{print"$a
      [0]\*$a[1]\_T\n";}}else{print"$_\n";}}' > ${p}_trans71.scafSeq;

  cd ${FAFOLDER}${p}/asm/IDBA/
  echo "Merging mitoreads for $p..."
  cat ${FAFOLDER}${p}/${p}_R1_mitoreads.fasta ${FAFOLDER}${p}/${p}_R2_
      mitoreads.fasta > ${FAFOLDER}$p/${p}_all_mitoreads.fa

  echo "Assembling $p with IDBA-UD"
```

```
  ${IDFOLDER}idba_ud -r ${FAFOLDER}$p/${p}_all_mitoreads.fa -o ${
      FAFOLDER}${p}/asm/IDBA/
done <$1
```

A.6.3 Minia pipeline

A.6.3.1 Shell script - Minia assembly

```
#!/usr/bin/env bash

# Pass text file with one sample name per line to iterate through all
    samples. Pass K-mer size as second argument.

FASTAFOLDER="/media/mikal/Polyseq_Mikal/trimmed_reads/"
ASMFOLDER="/media/mikal/Polyseq_Mikal/minia-assembled/"
MINIAFOLDER="/home/mikal/hax/minia/minia/build/bin/"

while read p; do
  mkdir ${ASMFOLDER}${p} -p;

  echo ${FASTAFOLDER}${p}_R1.fasta >> "${ASMFOLDER}${p}.paths";
  echo ${FASTAFOLDER}${p}_R2.fasta >> "${ASMFOLDER}${p}.paths";

  echo "Assembling $p with k-mer size of $2, writing log to ${
      ASMFOLDER}${p}/${p}_asm.report";

  ${MINIAFOLDER}minia -in "${ASMFOLDER}${p}.paths" -kmer-size $2 -out
      ${ASMFOLDER}${p}/${p}_assembled &>> "${ASMFOLDER}${p}/${p}_asm.
      report";
done <$1
```

A.6.4 Qubit and Nanodrop measurements

Table A4: DNA concentrations in ng/μL measured with Qubit fluorometer and purity measurements from Nanodrop 1000 photospectrometer. An x in the sample ID denotes that the specimen has been photographed.

Sample ID	Species	260/280	260/230	ng/μL	Tissue (mg)
A1x	*Spirobranchus triqueter*	1.84	1.53	64.4	46.6
A4x	*Polycirrus medusa*	2.05	2.11	48.8	20.9
A6x	*Pholoe baltica*	2.01	1.71	28.6	8.9
A7x	*Owenia fusiformis*	2.20	1.01	8.2	2.0
B1x	*Gattyana cirrhosa*	1.88	1.11	19.5	18.6
B2x	*Nereimyra punctata*	1.89	2.23	200	5.5
B3x	*Eupolymnia nesidensis*	2.06	2.01	260	40.7
B4x	*Scoloplos armiger*	1.84	1.44	21.6	24.0
C4x	*Euclymene droebachiensis*	1.97	1.53	16.3	1.4
C7x	*Chirimia biceps*	1.86	1.91	77.6	16.3
D2x	*Melinna cristata*	1.99	2.02	60.0	46.3
D5x	*Microclymene acirrata*	1.97	1.59	25.8	1.3
D7x	*Amphictene auricoma*	2.04	1.94	19.3	24.4
E1x	*Paraphinome jeffreysii*	2.00	1.09	29.8	0.9
E2x	*Pista cristata*	1.93	1.27	82.0	32.7
E3x	*Nephtys paradoxa*	1.93	1.80	258	42.3
E5x	*Laonice sarsi*	1.94	1.86	75.8	7.4
E6x	*Notomastus latericeus*	2.10	1.94	30.6	7.2
E7x	*Neoleanira tetragona*	2.08	2.14	58.4	25.7
E8x	*Aurospio banyulensis*	1.91	1.52	24.2	5.8
E9x	*Streblosoma intestinale*	1.93	1.59	43.0	3.0
F4x	*Polycirrus plumosus*	1.95	1.42	33.0	4.9
G1x	*Goniada maculata*	1.97	1.97	165	19.5
G5x	*Lumbriclymene cylindricauda*	1.97	2.07	129	8.5
G8x	*Scalibregma inflatum*	1.80	1.47	15.2	5.2
G9x	*Terebellides stroemii*	1.96	1.64	23.8	2.2
H1x	*Melinna elisabethae*	1.94	1.81	37	11.0
H2	*Melinna albicincta*	1.97	1.88	137	26.0
I5x	*Nereiphylla lutea*	2.08	2.00	89.6	37.8
I6x	*Phyllodoce cf. groenlandica*	1.94	1.99	142	31.2
J1x	*Spiophanes kroyeri*	2.00	1.81	12.9	2.4
J9	*Hydroides norvegica*	2.64	0.82	4.50	2.7
L7x	*Aphrodita aculeata*	1.53	0.65	16.1	37.0
M4x	*Aricidea catherinae*	1.82	1.37	22.2	2.7
O5x	*Melinna cristata elisabethae*	1.82	1.09	43.6	4.5
Q1x	*Nothria conchylega*	1.96	1.17	17.3	1.2
Q2x	*Hyalinoecia tubicola*	1.89	1.67	153	11.1
Q3x	*Sphaerodorum gracilis*	2.01	2.04	39.6	3.0
Q5x	*Phyllodoce groenlandica*	1.95	2.04	97.8	7.2
Q8x	*Nephtys longosetosa*	1.93	1.40	30.2	5.4
R2x	*Eclysippe vanelli*	1.98	1.85	36.6	2.3

A.6.5 Coverage maps

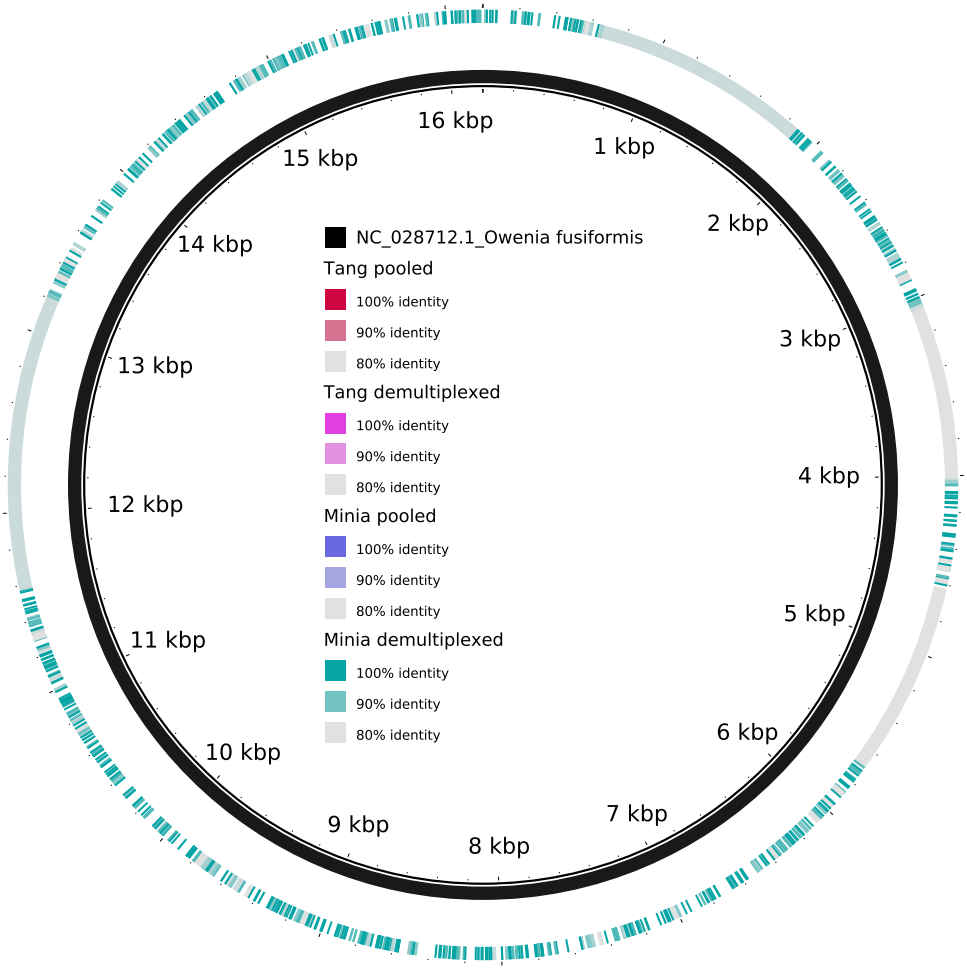

Figure A9: *Owenia fusiformis* mitogenome.

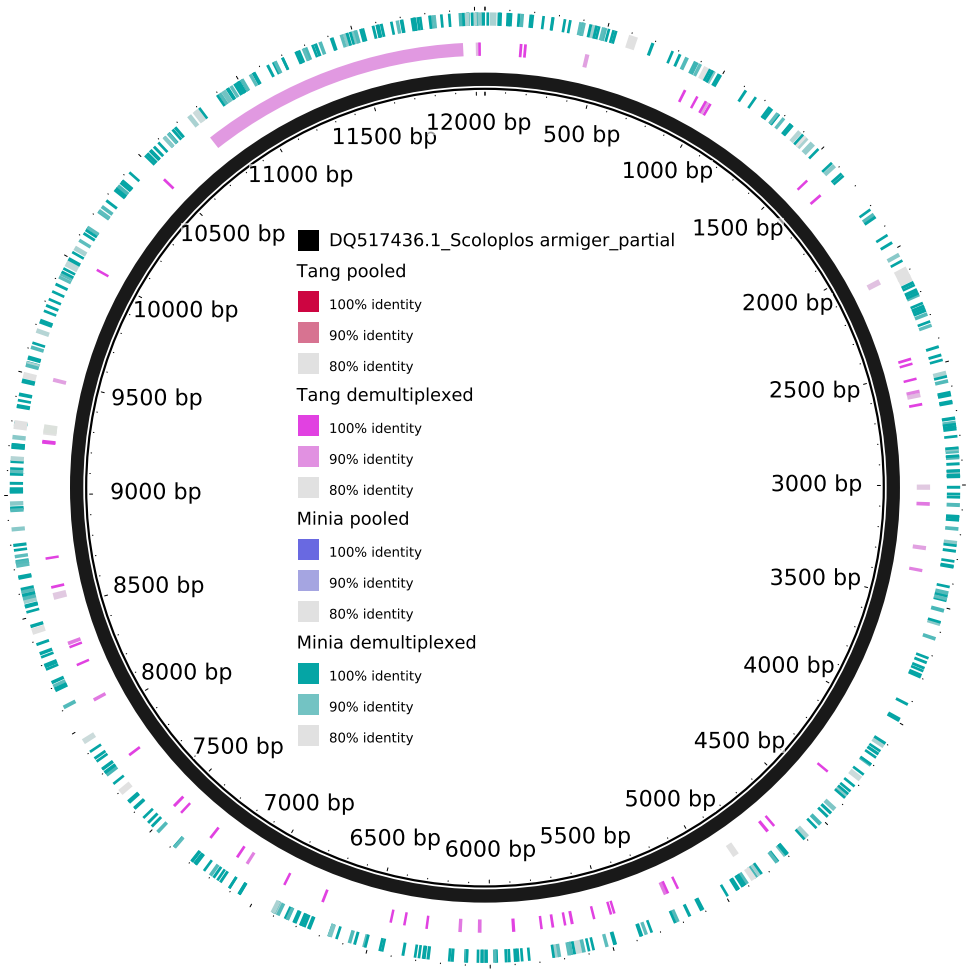

Figure A10: Partial *Scoloplos cf. armiger* mitogenome.

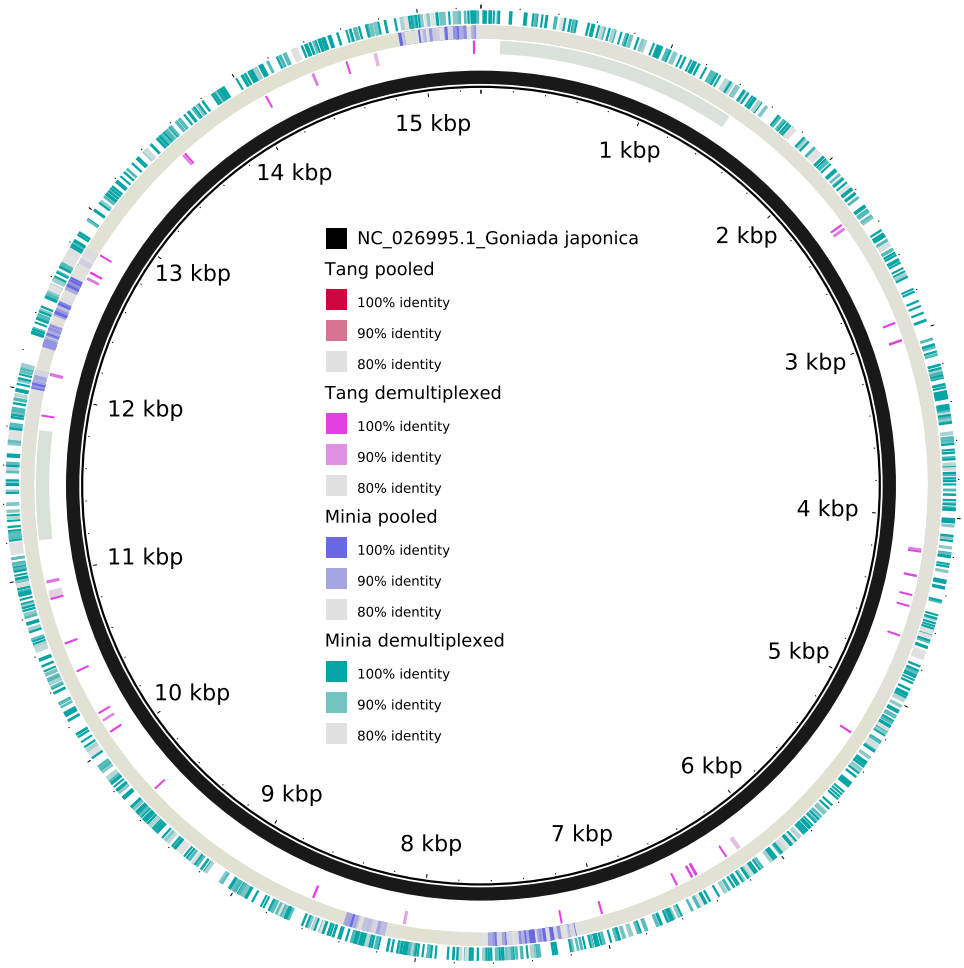

Figure A11: *Goniada maculata* mapped to *Goniada japonica*.

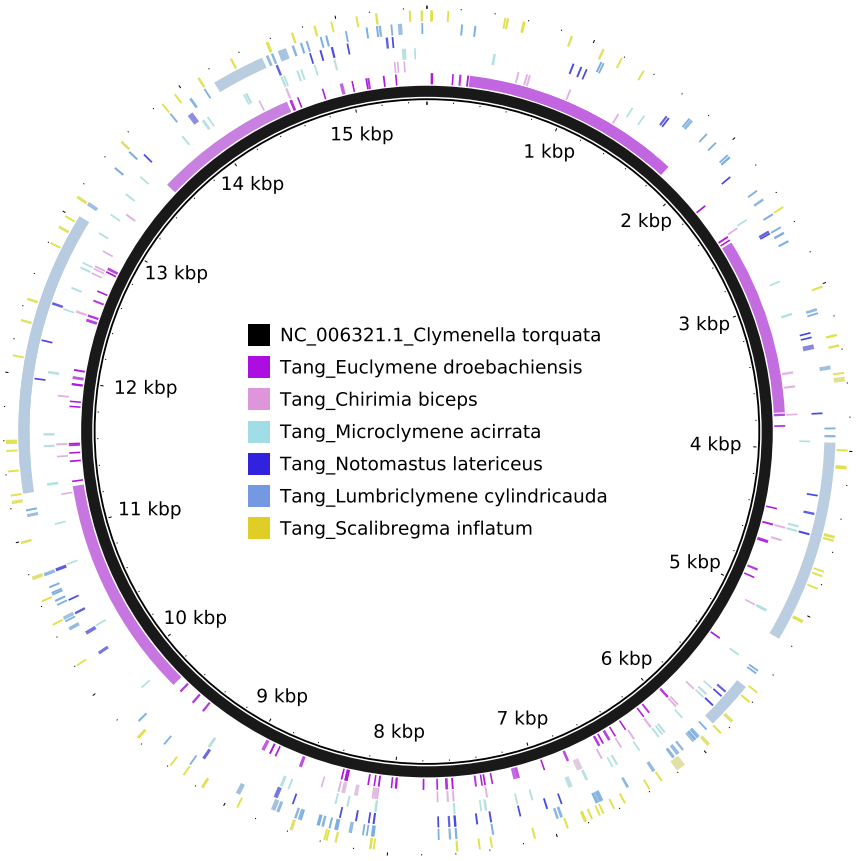

Figure A12: Sampled members of the order Capitellida, assembled from filtered mitoreads, mapped to *Clymenella torquata*.

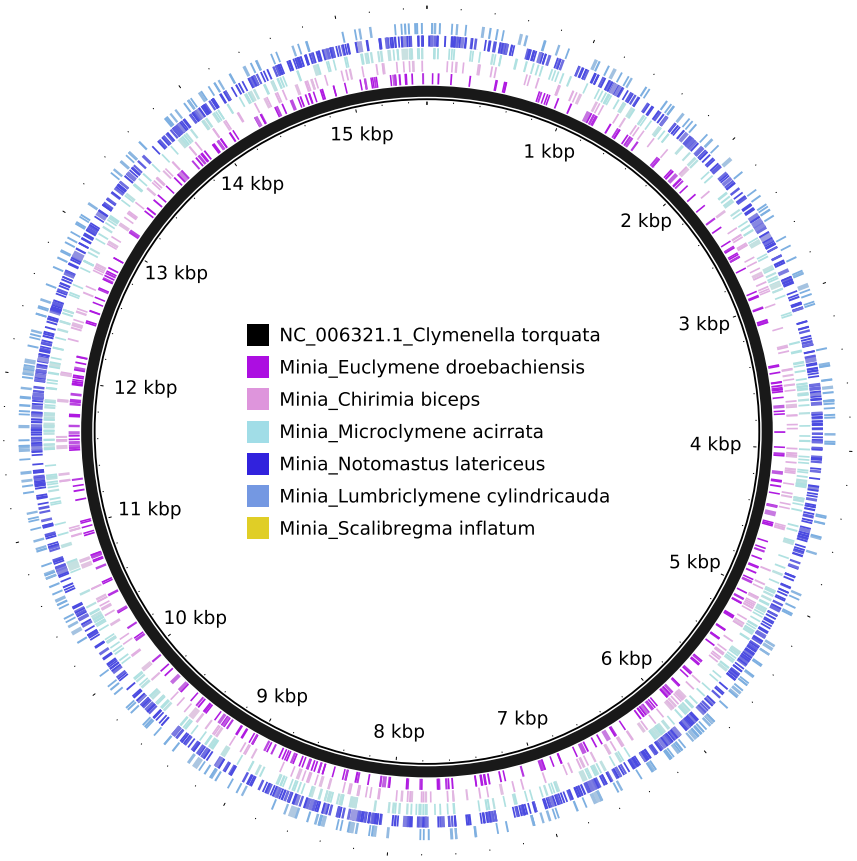

Figure A13: Sampled members of the order Capitellida, assembled with minia pipeline, mapped to *Clymenella torquata*.

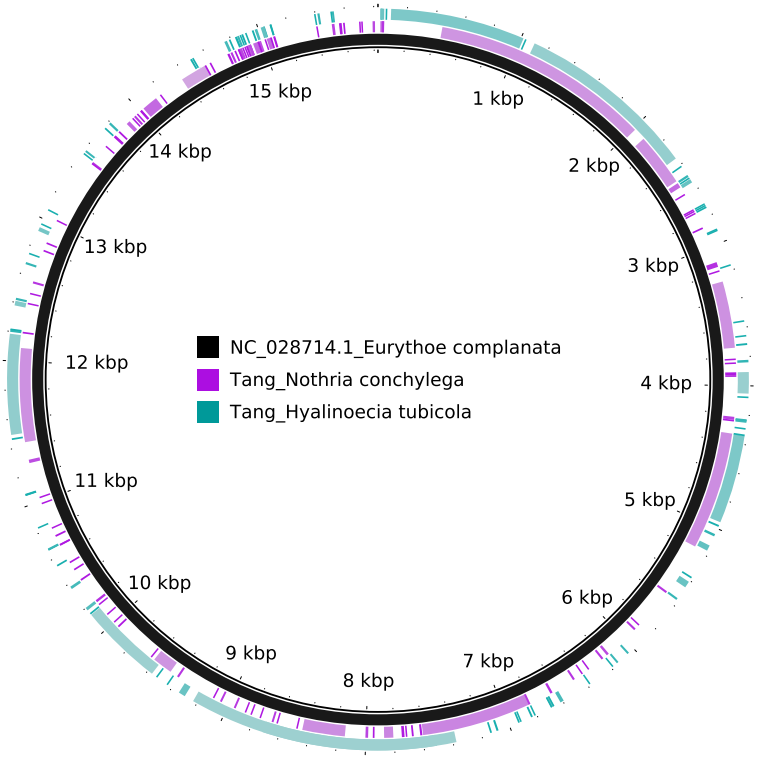

Figure A14: Sampled members of the order Eunicida, assembled from filtered mitoreads, mapped to *Eurythoe complanata*.

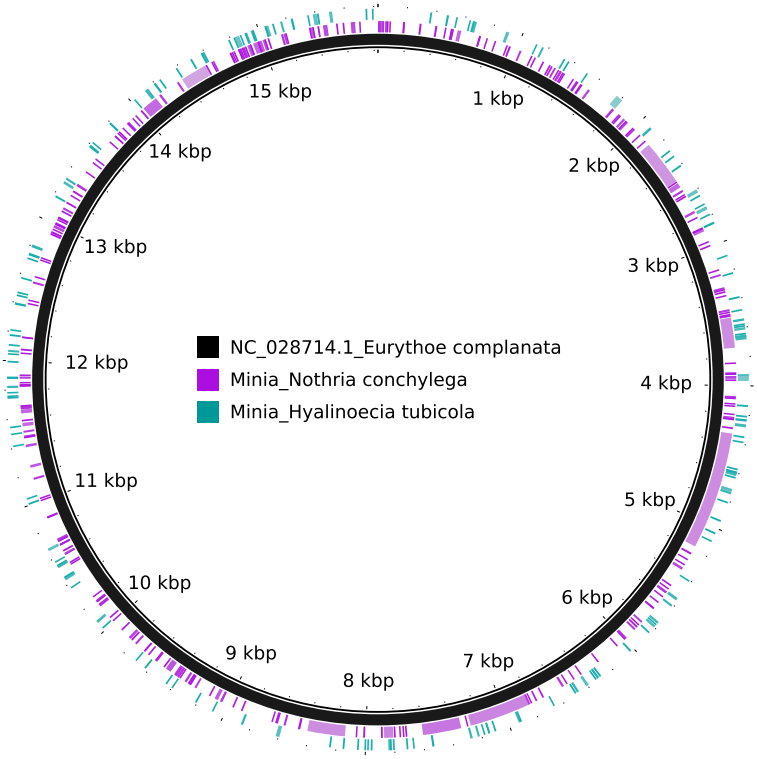

Figure A15: Sampled members of the order Eunicida, assembled with minia pipeline, mapped to *Eurythoe complanata*.

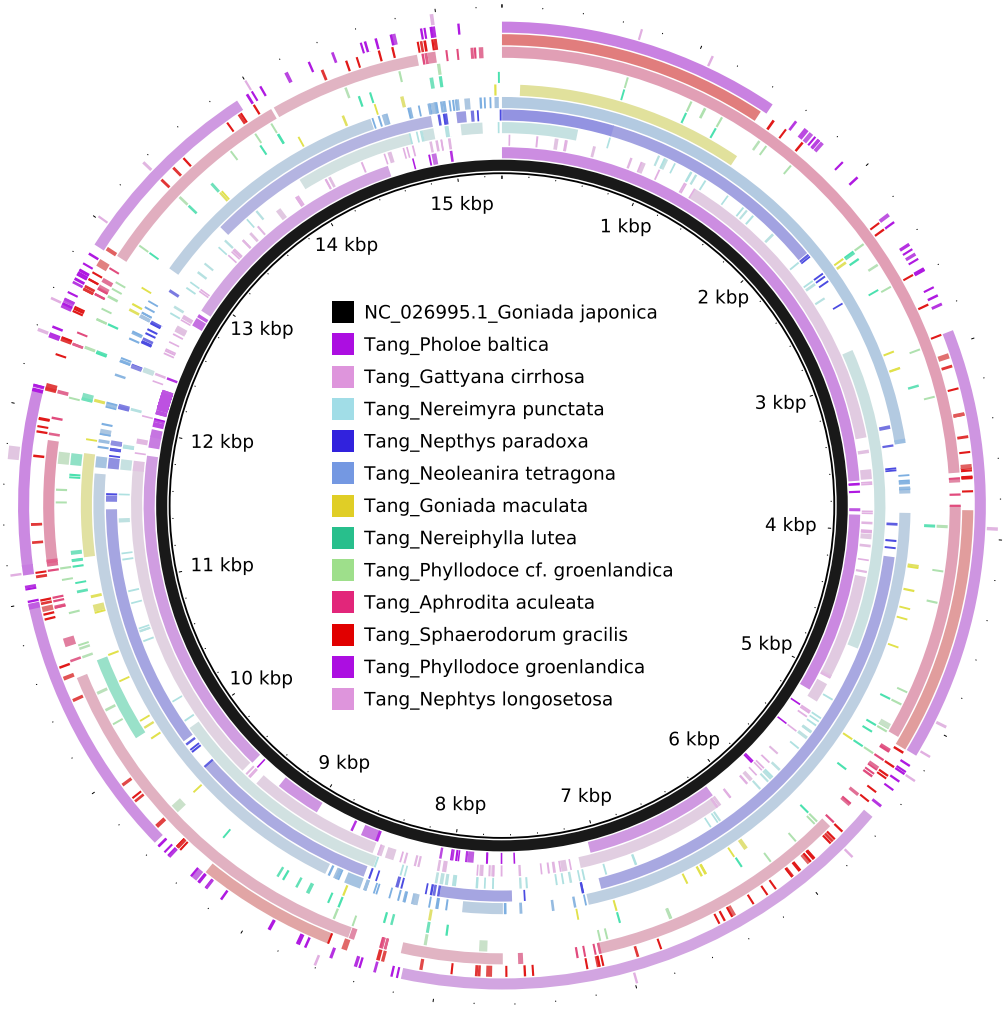

Figure A16: Sampled members of the order Phyllodocida, assembled from filtered mitoreads, mapped to *Goniada japonica*.

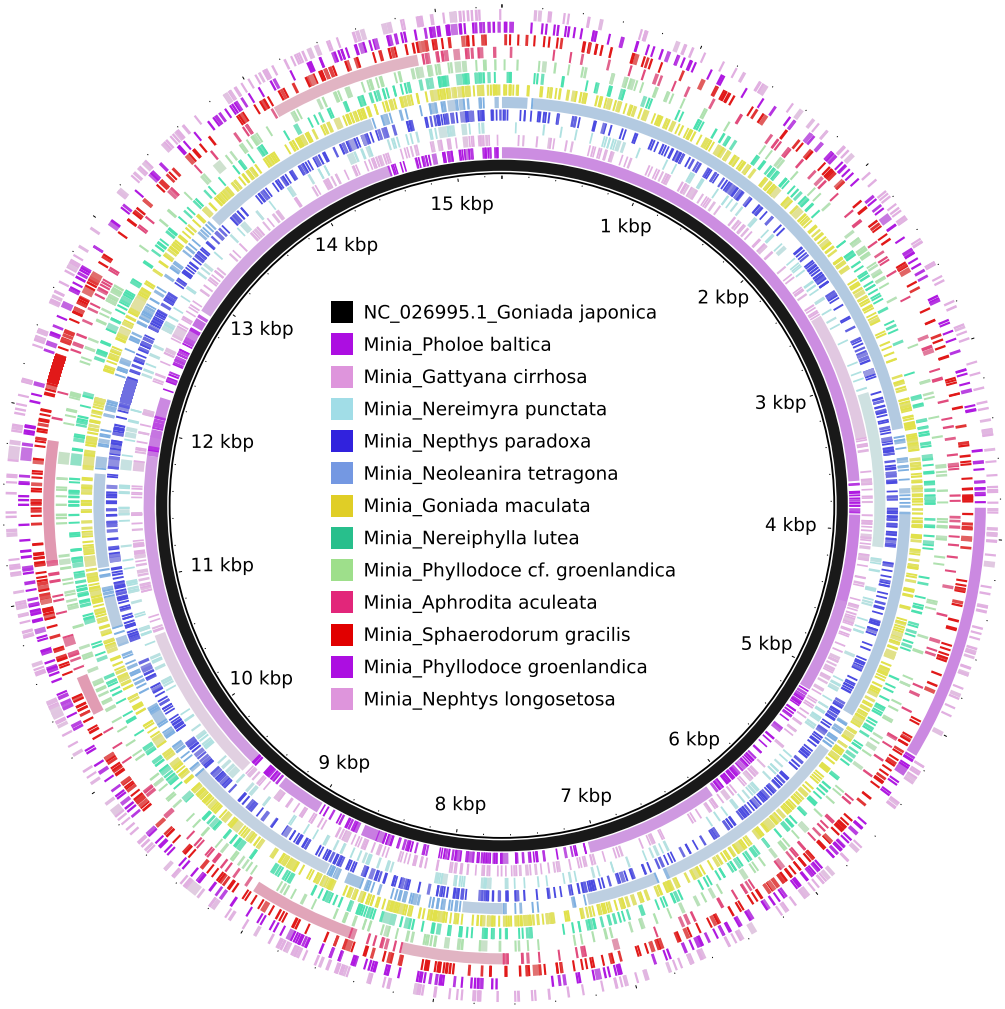

Figure A17: Sampled members of the order Phyllodocida, assembled with minia pipeline, mapped to *Goniada japonica*.

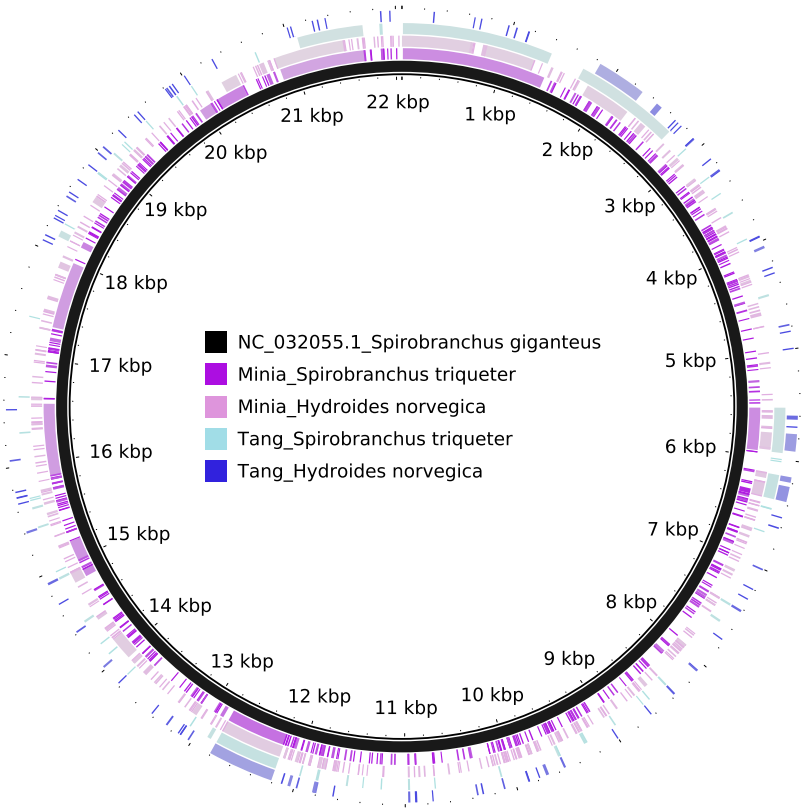

Figure A18: Members of the family Serpulidae mapped to a reference *Spirobranchus giganteus* mitogenome.

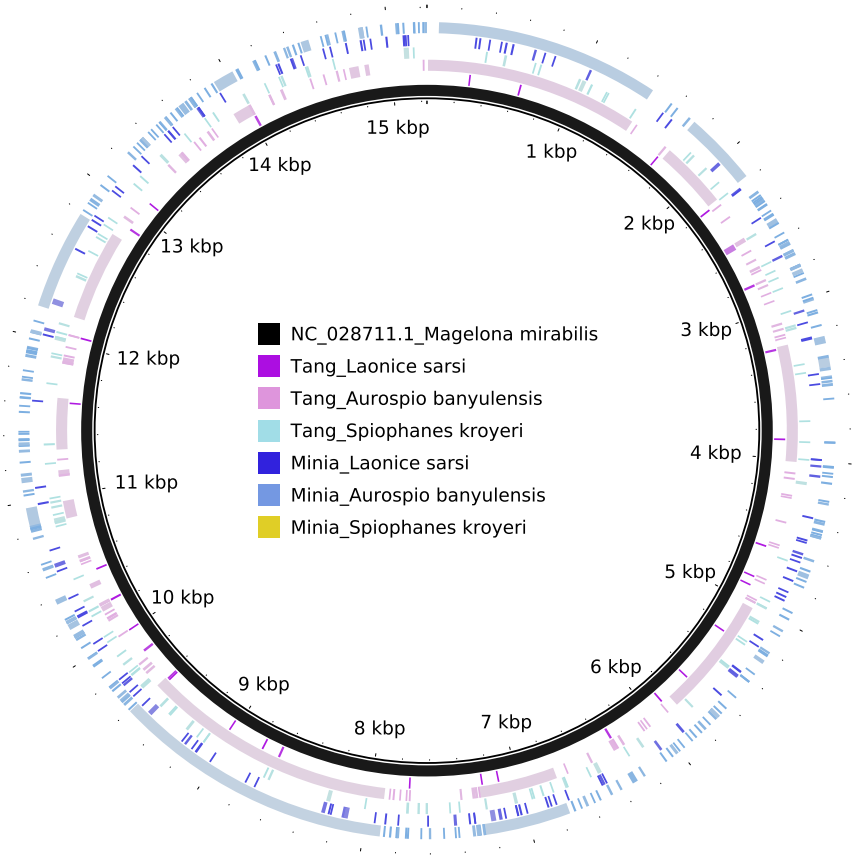

Figure A19: Members of the order Spionida mapped to *Magelona mirabilis*.

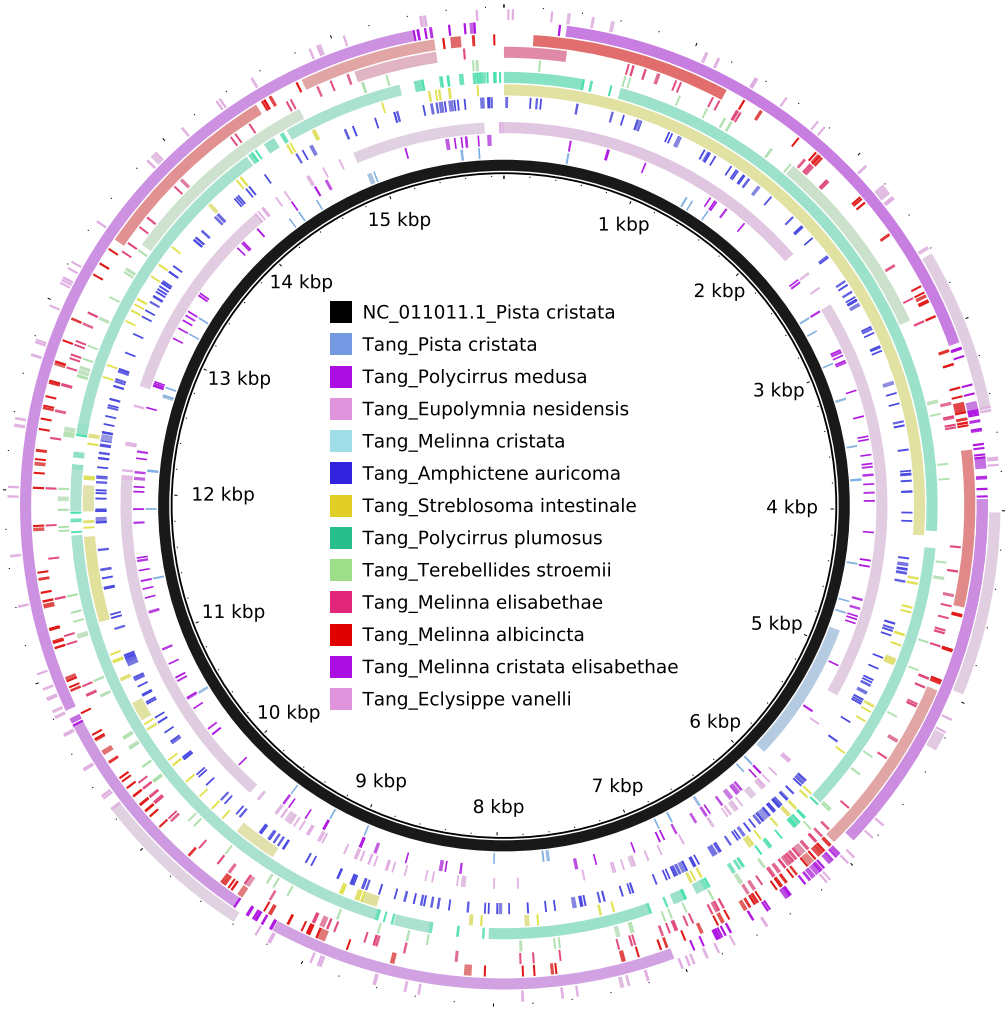

Figure A20: Sampled members of the order Terebellida, assembled from filtered mitoreads, mapped to *Pista cristata*.

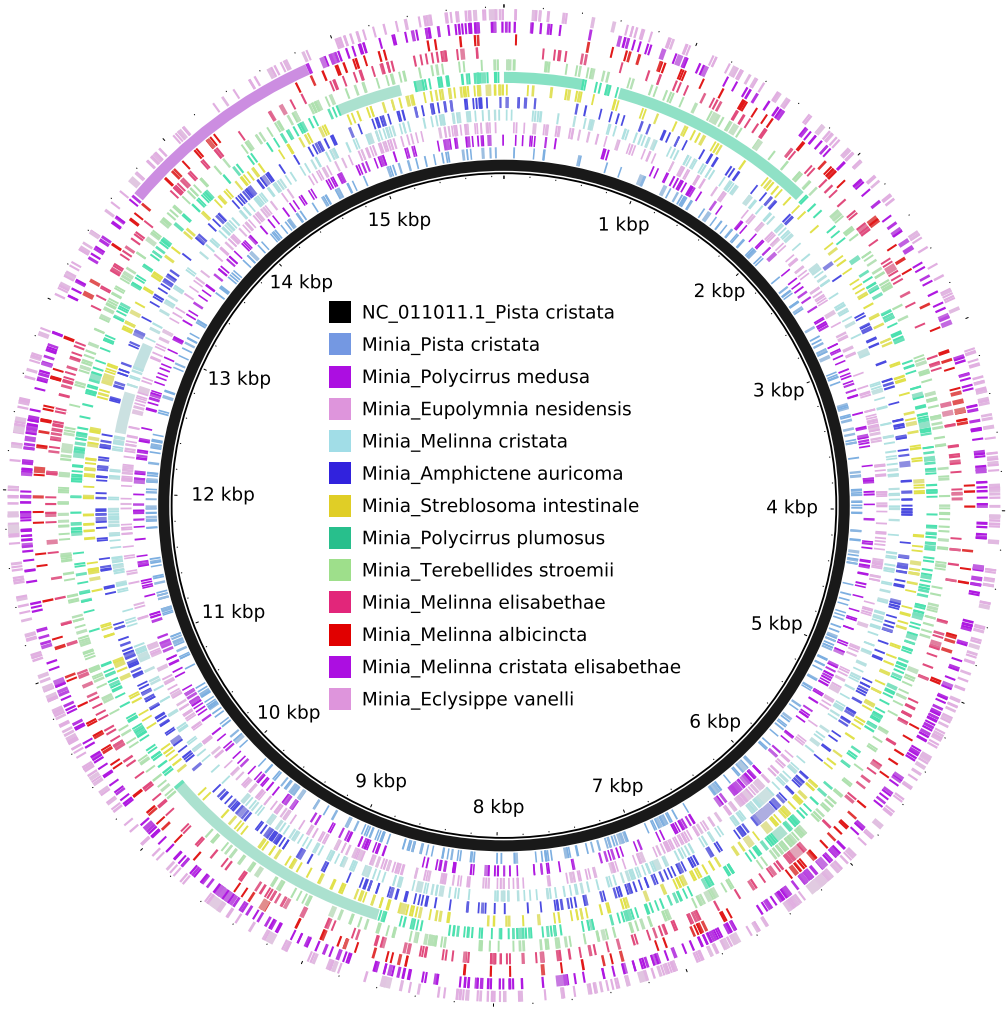

Figure A21: Sampled members of the order Terebellida, assembled with minia pipeline, mapped to *Pista cristata*.

A.6.6 Sequence data

Owenia fusiformis, partial mitochondrial genome.

```
   1 ---------- ---------- ---------- ---------- ---------- ----------
  61 ---------- ---------- ---------- ---------- ---------- TAAGATTGTT
 121 AATCCGGGCG GAATTAGGAC AGCCCGGTAG ATTGTTGGGA GATGATCAGT TGTATAATAC
 181 TATTGTTACT GCTCATGCTT TTGTTATAAT TTTTTTCTTT GTTATGCCAA TTATGATTGG
 241 GGGGTTCGGG AATTGATTGT TGCCGTTGAT ATTAGGTGCC CCGGATATGG CTTTTCCTCG
 301 AATAAATAAT ATAAGATTCT GATTGTTACC GCCTGCTTTG TTATT----- ----------
 361 ---------- ---------- ---------- ---------- ---------- ----------
 421 ---------- ---------- ---------- ---------- ---------- ----------
 481 ---------- ---------- ---------- ---------- ---------- ----------
 541 ---------- ---------- ---------- ---------- ---------- ----------
 601 ---------- ---------- ---------- ---------- ---------- ----------
 661 ---------- ---------- ---------- ---------- ---------- ----------
 721 ---------- ---------- ---------- ---------- ---------- ----------
 781 ---------- ---------- ---------- ---------- ---------- ----------
 841 ---------- ---------- ---GATTTAT TGTTTGAGCG CATCATATGT TTACGGTGGG
 901 CATAGACGTG GATACTCGTG CCTATTTTAC TGCTGCAACT ATAATTATTG CAGTGCCGAC
 961 GGGAATTCAA GTTTTTAGGT GATTAATAAC AATTTATGGG GCTAAGGTAT AGTATGAGAC
1021 CCCCATGTTA TGGGCGTTGG GGTTTATCTT TTTGTTTACT GTAGGTGGTT TGACGGGAAT
1081 TCTTCTTTCT AACTCTTCTA TTGATATTGT GCTCCATGAT ACTTATTATG TGGTTGCTCA
1141 TTTTCACTAT GTTTTGTCTA TGGGGGCTGT ATTTGCTTTG TTTGGGGGGT TTAATTACTG
1201 GTACCCGTTG TTGACAGGGA TTACATTGAA TACGCGTTGG GCAAAGGCTC ATTTTTTTAT
1261 AATGTTTTTT GGGGTAAATG TGACTTTTTT TCCTCAGCAT TTTTTAGGGT TAGGAGGAAT
1321 GCCTCGACGA TATTCTGACT ATCCAGATGT ATTTATGAAA TGGAATGTCA TTTCTTCTAT
1381 GGGTTCTTTA GTTTCATTTG TTGCAGTGTT GTTTTTTATA TTCATTATTT GAGAAAGATT
1441 ATTATCTCAA CGTGTGGTTA TGTGAAGTTC TCATTTGTCT GTAATATTGG AGTGAGATAA
1501 TCGTTTGCCT GTAGACTTTC ATAATTCTTC TGAGAGAGGG TTAGTTGTGG TTTAA----T
1561 TTAATGACTT CTTGAGGAAG ATTAGGATTT GTGGAGGCTG GGTCTCCTTT AATGGAGCAA
1621 TTGATTTTTT TTCATGATTA CGCAATATGT ATTTTATTAA TAATTATTTT TTTTGTGGGG
1681 GTTGCAGGTG TATTCATAAG AAAAAGAAGA TATACGGACC GGTATGTTTT AGAAGAGAAC
1741 ATGGTGGAAG TAATTTGAAC GATAGTGCCT GTGTTTATTT TAGTAGGTTT AGCTTTGCCT
1801 TCGTTACGTT TGTTATATTT ATTGGATGAG TCTTCTTCTC CTGGGTTGAC ACTTAAGGTT
1861 GGAGGGCATC AATGATATTG GAGGTATGAA TATTCTGATT TTAAAGATTT AGAATTTGAT
1921 TCTTATATAA TTCCCACAGA AGATTTAGAT AGGGGAAGAT ACCGTCTACT TGAAGTTGAT
1981 CACCGAACTG TATTGCCGGT TGGTGTAGAA GTTCGTGTTT TGGTTAGTGC TGCGGATGTT
2041 TTACATTCTT GAACTGTTCC TGCGCTTGGG GTGAAGGCTG ATGCTGTGCC AGGGCGTTTA
```

```
2101 AATCAGTTGA GGTTTGTTCT TATGCGAAGA GGAGTTTGAT ATGGGCAATG GTCTGAGATT
2161 TGTGGTGCTA ATCATTCTTT TATGCCCATT GTTGTTGAAG GGATTAGGTG TGATAGGTTT
2221 ATAAATTGAG TGGGGAGACT GGTGAGAGTG TAAGTTGTAG TGAAAAGTTA GTTTAATATA
2281 TAAAATTTTA GTTTGTCGTT CTATAGATAC TT-TTATTAG TACTTTTTTA TGCCACATTT
2341 AGCTCCTTTG AGTTGAAGAA TAAGAATAGT TTTTTTTTGG GCCTTGTTAA TTGGAGTGTG
2401 TTCATTAGTG TGATGAATGA AGGGCGTTGA CGTTAAGGTT ---------- ----------
2461 ---------- ---------- ---------- ---------- ---------- -------AAT
2521 ATGAAGAAGT GTAATGTGAG CATAGGAAAA TTTGAATTTC CAGGAAGATA TTTTGTCTCT
2581 TTATAAGGGC ATTTGTGAAG AACATGGATC AAGTAAATTT GTTTAAGATT TCGGCTCTTA
2641 ATGAAGTGAT AGAGTTCACT TATGTTTTAT GGTTTCGTTG TGAATGTGAG TAGCTTTAGC
2701 TATTGGTATT AGAGGGGGTA TTTGTGGTGC TGCTTTTGTG TTGTCTTCTA AGGTAGAATT
2761 AGATAGAGAA AAAAGCACAC CGTTTGAGTG TGGGTTTGAT CCGAAGGGGG GGGCACGAAT
2821 CCCGTTTTCT ATGCGATTCT TTTT------ ---------- ---------- ----------
2881 ---------- ---------- ---------- ---------- ---------- ----------
2941 ---------- ---------- ---------- ---------- ---------- ----------
3001 ---------- ---------- ---------- ---------- ---------- ----------
3061 ---------- ---------- ---------- ---------- ---------- ----------
3121 ---------- ---------- ---------- ---------- ---------- ----------
3181 ---------- ---------- ---------- ---------- ---------- ----------
3241 --------TT TAATATATGA GGAAGAAGGG TAATCTTACC CGTTATTTTG GATAAAATTG
3301 GTGTTCTTTT TGCAATGCTG GTTTTGCTTA TTTCTTCTTG TGTTTTTATA TTTTCTAGGA
3361 GATACATAGA AGACGAGATT TACCTTAGTC GGTTCATAAT TCTGGTTTTG ATGTTTGTTG
3421 TTTCTATGTG CCTTCTAATT TTTATTCCGA ATATAATTGC GTTACTAATT GGGTGGGATG
3481 GGTTGGGGTT GGTTTCTTTT TTATTAGTAA TTTATTATCA AAGATTTTAT TCTTTAAGAG
3541 CTGGATTGTT GACTGTTTTA ATTAATCGAG TGGGGGATGT GATACTATTA TTAAGAATTG
3601 GTTGGTGTGT TAATCAGGGG CATTGGTCTG CTTTAAGAGA TTATTCTTTT GGCTATGATT
3661 CTTGTGTGTT TTTGGTCGTG TTGGTTGCAT GTATAACTAA AAGAGCGCAG ATTCCTTTTA
3721 GTAGGTGATT ACCATCGGCT ATGGCGGATC CTACCCCTGT TTCCGCCCTG GTTCATTCTT
3781 CTACCTTAGT TACTGCTGGA ATTTTTTTAA TTATTCGGTT TTATAGATTC TTAAGTGAAT
3841 TTAATTGATT TTTACCGTTA ACTTTGTTAA TTGGGGTAAG GACTATGTTC ATGGCGGTAT
3901 T--------- ---------- ---------- ---------- ---------- ----------
3961 ---------- ---------- ---------- ---------- ---------- ----------
4021 ---------- ---------- ---------- ---------- ---------- ----------
4081 ---------- ---------- ---------- ---------- ---------- ----------
4141 ---------- ---------- ---------- ---------- ---------- ----------
4201 ---------- ---------- ---------- ---------- ---------- ----------
4261 ---------- ---------- ---------- ---------- ---------- ----------
```

```
4321 ---------- ---------- ---------- ---------- ---------- ----------
4381 ---------- ---------- ---------- ---------- ---------- ----------
4441 ---------- ---------- ---------- ---------- ---------- ----------
4501 ---------- ---------- ---------- ---------- ---------- ----------
4561 ---------- ---------- ---------- ---------- ---------- ----------
4621 ---------- ---------- ---------- ---------- ---------- ----------
4681 ---------- ---------- ---------- ---------- ---------- ----------
4741 ---------- ---------- ---------- ---------- ---------- ----------
4801 ---------- ---------- ---------- ---------- ---------- ----------
4861 ---------- ---------- ------TAAA AAGAGTAGGT TTGGAGGCAA AGTAGGGCTG
4921 CTAACTTTAT TTTGAGCAGT TCGATTCTGT TTTTACTTTT ATGGTTGTTA TTTTTCCTTT
4981 TAGAGTTTTA TTTTTTTTTG TAGTTATTAT TAGTACTATT TTTTCCCTAT CTAGGGGGCA
5041 TATTTTAGGT GTTTGGTTGG GGTTGGAGAT AAATATGTTA AGTTTTATTA GATTTAGGAT
5101 TCAAAGAAAA AGTTTTAATG AAGTGGAGGC TGGTTTAAAA TATTTTTTAG TGCAGGCTCT
5161 GGGGTCTGGG TTTTGATTAT TGGGATCTTT TT-------- ---------- ----------
5221 ---------- ---------- ---------- ---------- ---------- ----------
5281 ---------- ---------- -----TTGCT TACGAAGTTG GGTGCTTTTC CGACTTATTT
5341 TTGAGTGCCC AGGGTTATAA ATGAAATAAG TTGATTTAGA TGCTTTATTC TAGCTACATG
5401 ACAAAAATTC ATTCCCCTAA TTTTATT--- ---------- ---------- ----------
5461 ---------- ---------- ---------- ---------- ---------- ----------
5521 ---------- ---------- ---------- ---------- ---------- ----------
5581 ---------- ---------- ---------- ---------- ---------- ----------
5641 ---------- ---------- ---------- ---------- ---------- ----------
5701 ---------- ---AAAAGAA TATTAATTAT TTTTGGATTG AATTTTATTT CATTGGCCGG
5761 GTTACCCCCT TTCTTAGGAT TTATAAGGAA ATGAGTGGCT TTACAAGCTA TTGTTAGAAA
5821 TGGTAGATAT TTTTTAGGAT TTGGTCTTTT GATCGGAGCG TTAATAAATT TGTAATATTA
5881 TATGGTTGTA GTTAGAAATA TAAGAGTTTG GGGGGGGGTG TGAACTGTTA ATGACCGTTC
5941 TTTATTTAAT TTTAAATTGG AGTATTTATT CATATAAATA GTTAGAATAA TTT-------
6001 ---------- ---------- ---------- ---------- ---------- ----------
6061 -----TTAAC CTTTGTAGTC TAAAA-TAAG ATAAAGCGTT GAAGGTGCTG TGATAGAATT
6121 GTTCTCGTGG GTGGCGCGTA GGTATGATGT TGAAGTATTT TAATTTAAAA TATCATATTG
6181 TGGTTATGAA GAAGCAAAAA TTTTTGCCTT TGACA----- ---------- ----------
6241 --GCCTTGTA ATTGAATAAC AATATTAAAT TGCAAATTTA AATGTGCTTT ATATAAGCTT
6301 AGGCTTCTTT TTTATATTGG ATGATAGAAA TTATTTTAAG TTTGTTGTTT TTTTTAATTA
6361 CTATTTTGTG TGTCTTAATT AGAATGGCTT TTTTTACTTT AATAGAGCGT AAGTTATTAG
6421 GGTATATACA AATACGAAAG GGGCCCAATA AGGTTGGATT CATAGGGTTG CCTCAGCCAT
6481 TGGCGGATGC CGGAAAATTG TTTTTTAAAG AGCA------ ---------- ----------
```

```
6541 ---------- ---------- ---------- ---------- ---------- ----------
6601 ---------- ---------- ---------- ---------- ---------- ----------
6661 ---------- ---------- ---------- ---------- ---------- ----------
6721 ---------- ---------- ------CAAA CTATTTCCTA CGACGTCCGT TTAGTGTTAA
6781 TTTTACTAGG GTTAGTTGTT ATTTTTCTAA GGTATGATTT TATATATTGT TCTTATACTT
6841 TTTTA--TTA TTGAGTTGGG TTCATTTGTT TTCCGTTAAG GGTGGTTTGA GTACTTAGAA
6901 GAATTGCAGA AACTATTCGT GCGCCTTTCG ATTTTGCAGC GGGGGAGTCT GAATTAGTTT
6961 CAGGGTTTTA TGTTGAGTAT AGCGCGGGGG GGTTTGCGTT AATTTTTATT GCAGAGTTTC
7021 CTATAATTTT ACTTTTAAGT TTAATT---- ---------- ---------- ----------
7081 ---------- ---------- ---------- ---------- -TTATTTGCT GTATTATTTG
7141 TGTGGGTTCG CGGGGCATAC CCGCGGTATC GTTAGATTTG TTAATGGGAT TGGCGATGAA
7201 AAAGATTTCT ACCTTTTGAG ATTGTTTTAT TTGTTTGTTG A-ATTGATTT TTATTAT-TA
7261 TTTATTCAGA ATATAAAAA- AATCTTTATT TTCTAACTCC CAAAGTTAGT GTTTTACATA
7321 AACTATATTC TGTATTTACT TATCTTAGGA GAAAATATGT TTTGAAAACA TTAAAAGGAC
7381 GTTCGAATCG TTCAGTAAAT TTATGTTAGC CATTATTTGT AGATTGCTGT TGTCTGTAAG
7441 ATTTTTATTT CCGTTAATAC TTCAACCCTT AAGTTTAGGG GTAATGTTAC TAGTGAGAAC
7501 CGTGTTAGTA AGGTTAATTC TAGGATTTAT TATGAGAAGG TGATTTGGGT ATATTTTGTT
7561 TTTAGTATAT GTGGGGGGGT TATTAGTTAT -ATTCGGGTA TGTTGCCGCA CTAATGCCGA
7621 ACACAAATTT TTTTTAAAGG TGTTGGTTTT TAAATGGGAG TTTTTTTTTT TTTTTTTTTT
7681 TTTTTTTGTA TATTTATATG TTTATTGGGC AAAGAATGAG AGGGGCAGTG GAAGT-AATA
7741 TTACTTCTGG GTACTATGGT GTGAGTATCT TCTTATATTC CTGAAAAGGT ATTTATATTG
7801 TTGTGGGACT GGGGTTGGTA TTGTTACTAG CTTTGTTATG CGTTATTAAG GTGTGTTATT
7861 TTCACGGAGG TCCATTGCGT CCCTTCTTCT AATT------ ---------- ------TATG
7921 CATAGTCCTT TTCGTAAGGT TCATCCTGTT ATCCGAATCG TAAATGGTTC TTTGGTAGAT
7981 TTACCCGCTC CGATTAATTT GTCAGCTTGG TGAAATTTTG GTTCTTTGTT AGGTTTATGT
8041 TTAGTGGTTC AAATTTTAAC TGGTTTGTTT TTAGCTATGC ACTATACGCC GCATGTAGAT
8101 ATGGCTTTTT CTTCTGTCTC TCATATCGTT CGTGACGTAA ATAGGGGTTG GTTTCTTCGT
8161 TCTATGCACG CTAATGGTGC GTCTGCTTTT TTTGTATGTC TCTATGCCCA TGTTGGGCGA
8221 GGAATGTATT ATGGTTCTTA TTTTATATTA GAGACTTGAA ATATCGGTGT CATTTTATTA
8281 TTCGGTGTTA TAGCTACTTC CTTTCTAGGG TATGTACTAC CTTGAGGTCA GATGTCTTTT
8341 TGAGGGGCGA CTGTGATTAC AAATTTGTTC TCTGCTTTTC CCTATGTTGG TAAGATGTTA
8401 GTGGAATGAG CTTGAGGGGG GTTTT----- ---------- ---------- ----------
8461 ---------- ---------- ---------- ---------- ---------- ----------
8521 ---------- ---------- ---------- ---------- ---------- ----------
8581 ---------- ---------- ---------- ---------- ---------- ----------
8641 ---------- ---------- ---------- ---------- ---------- ----------
8701 ---------- ---------- ---------- ---------- ---------- ----------
```

```
 8761 ---------- ---------- ---------- ---------- ---------- ----------
 8821 ---------- ---------- ---------- ---------- ---------- ----------
 8881 ---------- ---------- GTTTGTAAGA ATTTTTCTCT TACTAACTTG GATTGGAGCC
 8941 TGTCCTGTGG AAGATCCGTA TATTTTTATA GGACAAGTGT TGACGGTGAG ATATTTTCTT
 9001 TATTTTGGTT TAAATATA-- -GTTTGTAAA GTTTGAGATT TTTTAATTA- -----ATTAT
 9061 TAAGGATATT AATAAAGTTT CCAG-AAAGT TTTTTCTTCT TTGGTTTACA AGACCAATGC
 9121 TTCTTTAAAG CTTTAGAAAC ATATA-A-TA TAGATGATAA CTTGATTTAT GATAAATGGT
 9181 TTTATTTTTA GGTTGGTTTT GGTTCTTAGA GGGATGATTA CTTTATCTTT ACAAAGGGTG
 9241 CACCTATTGG GGGCTTTGTT AAGGTTAGAA ATGATGGTCT TGGGGTTATT TTTCTTTTTT
 9301 TCTTTAATTT CTTCTG---T CACCGGCGGT GAATCTTACT TAGGATTGAT TATTCTAACT
 9361 TTTGGGGTCT GTGAAGCTAG ATTAGGTTTA GCTTTATTTG TGAGAATTGT GCGGTCCCAC
 9421 GGGGTAGGTT ATTTTTCTAT TTTTTTTTTA AATTA-AAGT TGTAGAAGGA GTAATGCTTT
 9481 CGTTAGGGGT TTTTGGGATT TTGGGGGTTA CAAAAAACAA TTGAATCTAT TTGATTAGGG
 9541 TCTATTTCTT TTTAAGTTTT TTTTGTATT- ---------- ---------- ----------
 9601 ---------- ---------- ---------- ---------- --------TA ATTTTTTTGG
 9661 TGTTTTGAAT TAGTGGGATA ATAATAATAG CGAGTTATTC TTTGGAAAGT AAGGGAAGAG
 9721 GTTATATGTT TTATTTCCTA ATTTACTTAT TAACAGTGAT TTTGTTTTTG TGTTTTTCAG
 9781 TTACTAACAT GTTTTATTTT TATTTAGTTT TTGAAGTATC TCTTATTCCA ACATTAATGT
 9841 TAATTTTAGG TTGAGGTTAT CAGCCAGAGC GATTGCAAGC AGGATCCTAT ATAATCATAT
 9901 ATACAGTTAG TGCTTCTTTG CCCTTATTAA TGGGAATTTT AAGTGTTGAT TGGCTATTTG
 9961 GAGGCCTTAA TTTTTTTTCT TTGGTGCA-G GTAAATCAAT TATTAGGGGA GGAGCATATA
10021 TAGGGGCCTT TATAATAACG TTGATGTTGG CGTTTTTGGT TAAGTTGCCA ATATATTTTA
10081 GCCACTTGTG GCGTCCCAAA GCTCATGTGG AGGCACCTGT AGCTGGGTCT ATGGTTCTGG
10141 CAGGGGTGTT ACTAAAACTT GGGGGGTATG GATTAATGCG TATGATAAGA TTTTTCTGAA
10201 GGGGAATTTA CATGGTTTAC GGTGTTATTG GGCTCATTAG ATTATGGGGA GCTATGATTA
10261 CTAGATTAAT TTGTTTACGA CAAAGAGATG TAAAATCACT TATTGCCTAT TCTTCAGTGG
10321 CCCACATGGG GTTGTTAATC GGGGGTTTAA TAAGGGGAGT ATATTGAGGG TGAGAAGGAA
10381 GGCTAATAAT AATGTTGGCG CATGGAGTAT GTTCTTCTGG AATGTTTGCT TTGGCAGGTT
10441 TAAGTTATGA TATTTTATCG TCACGCAGCT TGGTTGTGAA TAAAGGTATA TTAGTTTTGT
10501 TACCTTCTAT GACTTTTTGG TGGTTTCTAT TGAGGGTTTG TAATCTTGCT GGGCCTCCTA
10561 CTTTGAATTT AGTAAGGGAA TTATTCTTAA TTAGAAGGAT TTTAAGCTAT AGTGATTATT
10621 TAGGGGTGGT TATTGGTGTA TTAAGGTTTT TTGCTGCGGG GTATAGCTTA TATTTATATG
10681 TAAGAACTCA GCATGGGAAG GTATCGTTGT TTATTAATAC ATTAGTAATT AATTTGAGGG
10741 GTCGAGTGAA TAGAATGTTT TTTTTTCAT- ---------- -AATTTGATT ATTTTAAAGG
10801 TAGATCTTTT CTTAATTTAA AGTTTGCGTT TATAGTATAA TT-AATACGC TAAGGTTTCA
10861 TTTTAGAGAT GCAT--AT-A TATGTTGAAT GTT-TGTGG- TTATTGTTAA -TCTCTTATT
10921 ACATTTTTAG AAAATGTTAA GTTAAGATAA ACTGAAAACC TTCAAAGTTT TTAATGA--T
```

```
10981 CGGTTGATCA CATTTTGTGA TAAAATAAGC TAAGGTTAAG CTGTTGGGTT CATACCTCAA
11041 AAATAGATAA AATGTCTTTT TATTA----- ---AACTAAA AGACTTTGAT TTTGGGTTAA
11101 TGATGGCGTA AATTAATAAT TCACATACAA ATATTTGTTC TTAT-AAAAT AAAA------
11161 --TTTAA-GG AAAGAAAAGT TAAAGAAGTA GTGTAGAAAT TTATAAAATG ATAGAAATAT
11221 TGATTTAGGT TAATTTATGA AGGGTAGATA AAATTGTGCC AGCATCTGCG GTTATACAAT
11281 GGCCCTAAGT TATCGTGTTA CGGAAAAAAT TTTTTGAGGT TT-----GTT ATTTTGATAA
11341 TAAAGTGTGA AAATGTGGGT TAAATCTTTT TTCATATTAT AATTTTTTTA TCAGTAAACT
11401 -TCAAAGAAA AATTGATAAA AAACTGGGAT TAGATACCCC GTTATTCTTT AGGTTAAATG
11461 TATTTCCAAA GTATTATGAA GGAGACTTTA AAACTTAAAA AATTTGGCGG TATTTTATTT
11521 CTTATTAGGG GAACTTGTCC TGTAATCGAT AATCCCCGTT ATAAATTACC TTTTTTTGTT
11581 GTGTGTCAGT TTGTGTATTT TCGTCATAAA TTTATCTTTA AGAGTCAAGA AATAAAT--A
11641 ATGTCTATTA AGGCAGATGT CAGATCAATG TGCAGTTAAT -AAAAAGGAA GGATGAGTTA
11701 CATTAAATTA ATATAAAAAT GGCGCATAAA TTGAAAGATT TACGTTAAAT TGGACTTAAA
11761 AGTAAAAAAA AGTGATAAAG TTTTTTTGGA GTAGCTCTAA AATGTGCACA AATCGCCCGT
11821 CGCTCTCGTT TAAAGATGAG ATAAGTCGTA ACACAGTAAG CGTAATGGAA ATTGTGCTTG
11881 GTAGGATTTT TTAAAAGATA A---AAGTTT -GT--TGGGT TGACTATAAA AAAAAATTAT
11941 TTTTTGTTTA AAGTAGAAAA ACGGTTTTAT ATATTTGTAT TATTTATATT A-TTAGTATT
12001 GTAAGAGAAA AATATTGTCT AAAAAAAATA ATAGGTTAAT TCTAGTACCT TTTGGATCAT
12061 GATATACTTA ATTTTATAGA ATTATTTTTT TGTATCGAAG CCTTTTGAGC TATTATTTTT
12121 AAGAC--TT- GGTTTATAT- TGTGATGTGG CATAAT--CA T-GAAATAAA AGATAATAGA
12181 GGTGATAAGT TTATCGGAAA AGGTGATAGC TAATTTCAGA AGAAATGTAT TTTAGTACAG
12241 GAATATTTGA --GAA---TA TATCTGGTAA TAGTAAAGAT AAGTTTGTTA TTATATAAAA
12301 TTAAA-ATTT TTTTTAAAAG AGCACTGATT TAGGTGTAAA AGTAACTGAG TTAGAAGAAT
12361 TTATATTAAA GGT-TTGGAT TAATTTTATT TCTG-ATAAA GTTTTGTTTA TTTTATTCTG
12421 AGGAATTAAA AAATTTAATT TGTAATTGTT TTTATAGTAT AATTAGTATA TAA--AAATT
12481 TTTATAATGT TAATTATTTT TAAATATTAA CTGAAAGATT TTTGTTTAAA TAATTAATGT
12541 TAGGAACTCG GCAAAAATTT TTTCCAACTG GTTATCAAAA CCATAGTTTC TTGGTA-ACA
12601 AAAGAAATAT TTTCTGCCCG GTGAGTATTT TAACGGCCGC AGTTTTACTG TGCTAAGGTA
12661 GCATAATCAT TTGCCTTTTA ATTGGAGGCT AGAATGAATG GAAGAATGAG GGATTTACTT
12721 TCTTACATTA ACAAATTTGA ATTTTTATTT TAGGTGAAGA TACCTGAATG AGATTGTTAG
12781 ACGAGAAGAC CCTGCTGATC TTTATTGATT TTATAAATTT TAATTTGAT- GTAAAATATT
12841 GATATAAAAT TAAGTTGATT GGGGCGATCA AGGATAGAT- -TATCATCAT CCTTTACATA
12901 TAAGAA--AT ATTTTTTATA AATGATCCAT TTATTATGAT TAAAAAATTA AGTTACCACA
12961 GGGATAACAG AGTAATATAT TTTGAGAGTT CAGATCGATA AATATGTTTG CTACCTCGAT
13021 GTTGGATTAG GATCCCCAAA AGGTGTAGAA GTTTTTTTGG TTAGTCTGTT CGACTATTAA
13081 AATCCTACAT GATCTGAGTT CAGACCGGCG TGAGCCAGGT CAGTTTCTAT CTACAATTAA
13141 TGAGATATTA TTTTTTAGTA CGAAAGGAAT TAAAATAAGA GTTGTA-ACT TAAAATAAGA
```

```
13201 AAATTTTAAT TTATTATTTG AATTAGTAA- ----ATTTGA AATTTGATTG GCAGATATAT
13261 GTGATTGGTT TAGGACCATT TTATAAGATG A--AATTCTT ATCAAATAA- --T--TAAAA
13321 TAGAGCATGT GTTTTATGTG TTTCAGTTAC ATTGAAAAAA AGCTATTTAT AGTTGTTTTA
13381 A--------- ---------- ---------- ---------- ---------- ----------
13441 ---------- ---------- ---------- ---------- ---------- ----------
13501 ---------- ---------- ---------- ---------- ---------- ----------
13561 ---------- ---------- ---------- ---------- ---------- ----------
13621 ---------- ---------- ---------- ---------- ---------- ----------
13681 ---------- ---------- ---------- ---------- ---------- ----------
13741 ---------- ---------- ---------- ---------- ---------- ----------
13801 ---------- ---------- ---------- ---------- ---------- ----------
13861 ---------- ---------- ---------- ---------- ---------- ----------
13921 ---------- ---------- ---------- ---------- ---------- ----------
13981 ---------- ---------- ---------- ---------- ---------- ----------
14041 ---------- ---------- ---------- ---------- ---------- ----------
14101 ---------- ---------- ---------- -------AAT TAGTTTAAAT AATGTTTAAG
14161 TGGCAGAGTA GTGCATTAGA TTTAAGCTTT AAATATGGTG AATTCATATA CCCTTAAATA
14221 AATGATAGTC GACATTTTTT CTTCCTTTGA TA-GAAGATT TATTTTAGTG GCGGAGAAAG
14281 AGGGGTTAAA TAGGAGGGAT TTTGGTGGGT TGTTCAATTT TTTTTAATAA ATATGTTTTT
14341 TGTGAGGTAT AAATG-CAGG TTAGTCGTTT TTGGTGAGTA ACTACTTTAC CTTTGATTTT
14401 GTTTCATGGG GAAGTGGGGC GTAGG----- ---------- ---------- ----------
14461 ---------- ---------- ---------- ---------- ---------- ----------
14521 ---------- ---------- ---------- ---------- ---------- ----------
14581 ---------- ---------- ---------- ---------- ---------- ----------
14641 ---------- ---------- ---------- ---------- ---------A CTTCAGTTCG
14701 CCCTATTACT CTTTCAGCTC GATTAGCTGC TAATTTGAGA GCGGGGCATA TTGTTTTGGG
14761 CCTTATAAGA GTTTACTTAT GTAAGATAAT TTTTTCTTTT AG--GGGTAT TTTTTTTTTT
14821 ATTAATATTA GAGGTTGGAT ATTTTTTATT TGAAGTGGGG GTCTGTATAA TTCAAGGTTA
14881 CATCTTTAGG CTTTTAATCA CATTATATGC TGATGAGCAT AGGTTGTAAA ATTTTAAATA
14941 AAAACTGATA ATAGTGGTTG ATTGTTGGTC AATTTGATGT GGGGAATTTC CCATTTATTT
15001 AGGGGGTTAA GGTGGGT-TA TGTAGTAATT TAAGAAAAAT GTTAAATTGT AAATTTAAAT
15061 TTGAGAAG-T GTCTCTTACA TAAAAGGGTA TTGCGCCGGG TTGAACGGGC TATGTTGATG
15121 TTGTAGAGCA CGGAATAATT A---TTCTCA ATACT-AAGT GATTAGGTGG TGATTTACTT
15181 TAAGTAAAAG ATATAGTTTG CACCTATAAA ATAGAATTTG ATTTTCTTTT CACAAG----
15241 ---------- ---------- ---------- --CTTCCTTC CAAGAAGTAG GTTTATAGAA
15301 GATAAAATGA ATAAATGATT TCTAGACCTT TTCATTTGGT TGAGTATAGT CCTTGGCCTT
15361 TAACGGGGTC AGTGGGAGGA TTTTGCATGG TTATAGGGTT GGTTAGTTGG TTTCATGGTT
```

```
15421 ACGGGGTTTG AGGGGCAATT TTAGGGTTAT TTTTAATTAT TAGTACTATA TAT-------
15481 ---------- ---------- ---------- -------TCA TACTTTTATA GTTTGTAAGG
15541 GTCTTCGGTG GGGAATGATT TTGTTCATTA CATCAGAAGT GTTGTTTTTT TTTGCCTTTT
15601 TCTGAGCTTA TTTTCATAGG AGACTTTCTC CTTCTGTGGA GTTGGGTTCT TGTTGGCCGC
15661 CAGTCGGTAT TTATCCTTTG AATGCTTTTT CTGTGCCTCT TTTGAATACT GCTGTACTAT
15721 TGGCTTCAGG CTTTACAGTT ACTTGGGCTC ACCATAGATT CATAAGTGGG GAGTATTCTT
15781 CAGGGTTGCA AGGGTTGTTG ATTACTGTTG TTTTAGGAAT GTATTTTAGT GTGCTCCAGG
15841 CTGGAGAGTA TTGAGAGGCT CCTTTTAACC TGTCTGACGG GGTTTACGGG TCTAGGTTTT
15901 TTGTGGCTAC AGGTTTTCAC GGGCTTCATG TCCTTATTGG AAGAACTTTT CTTTTAGTGT
15961 GTTTAATTCG CGGGTATAAT TACCATTTTT CCGATGGGCA TCATTTTGGG TTCGAGGCTG
16021 CTGCGTGGTA TTGGCATTTT GTTGATGTAG TGTGATTATT TTTGTATACT TGAGTATATT
16081 GATGGGGGGC ATAAACAGGT GATAAACATT TAAGGGTTGA ACGAGTAGCT AAAA-ATTAT
16141 AGCTTAAAAT TTTTACTTTT AGGGTGAGTA ---------- ---------- ----------
16201 ----
```