Trimitomics: An efficient pipeline for mitochondrial assembly from transcriptomic reads in non-model species

Bruna Plese[1,2,*] 0000-0001-6690-3876 brunaplese@googlemail.com
Maria Eleonora Rossi[1] 0000-0002-4076-5601 m.eleonora.rossi@gmail.com
Nathan James Kenny[1] 0000-0003-4816-4103 nathanjameskenny@gmail.com
Sergi Taboada[1] 0000-0003-1669-1152 sergiotab@gmail.com
Vasiliki Koutsouveli[1] 0000-0001-9117-0598 v.koutsouveli@nhm.ac.uk
Ana Riesgo[1,*] 0000-0002-7993-1523 a.riesgo@nhm.ac.uk

[1] Life Sciences Department, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

[2] Division of Molecular Biology, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia

*corresponding authors: brunaplese@googlemail.com and A.Riesgo@nhm.ac.uk

Abstract

Mitochondrial resources are of known utility to many fields of phylogenetic, population and molecular biology. Their combination of faster and slower-evolving regions and high copy number enables them to be used in many situations where other loci are unsuitable, with degraded samples and after recent speciation events.

The advent of next-generation sequencing technologies (and notably the Illumina platform) has led to an explosion in the number of samples that can be studied at transcriptomic level, at relatively low cost. Here we describe a robust pipeline for the recovery of mitochondrial genomes from these RNA-seq resources. This pipeline can be used on sequencing of a variety of depths, and reliably recovers the protein coding and ribosomal gene complements of mitochondria from almost any transcriptomic sequencing experiment. The complete sequence of the mitochondrial genome can also be recovered when sequencing is performed in sufficient depth. We evidence the efficacy of our pipeline using data from eight non-model invertebrates of six disparate phyla. Interestingly, among our poriferan data, where microbiological symbionts are known empirically to make mitochondrial assembly difficult, this pipeline proved especially useful.

Our pipeline will allow the recovery of mitochondrial data from a variety of previously-sequenced samples, and add an additional angle of enquiry to future RNA-seq efforts, simplifying the process of mitochondrial genome assembly for even the most recalcitrant clades and adding this data to the scientific record for a range of future uses.

Key words: Mitochondrial genome, transcriptomics, assembly, invertebrates

Introduction:

Mitochondria, due to their vital function within the cell, have been the subject of research from a diversity of angles for as long as we have known of their existence. In particular, combination of fast and slow-evolving characters and severe functional constraints on their operation (da Fonseca et al., 2008), render them excellent for molecular phylogenetic investigations. Mitochondrial sequence data are exceptionally useful for tracing the inter-relationships of non-model organisms. They can also, among other examples, be used to understand population structure (Avise et al., 1987), for conservation biology (Rubinoff, 2006), for forensics (Melton et al., 2012), in medicine (Picard et al., 2016), and in the study of evolutionary pressures (Romero et al., 2016). As a result of this flexibility, the sequencing of mitochondrial genomes is popular for a diverse range of uses.

While there are a number of means by which mitochondrial genomes can be sequenced, including Sanger and genomic DNA sequencing, it has been noted previously that RNA-seq data could provide a novel source of mitochondrial sequence data (Smith, 2013; Tian & Smith, 2016). Despite treatments aimed at enriching nuclear genome mRNA levels, the high copy number of mitochondria within cells, coupled to generally high expression levels, mean that some level of mitochondrial data will inevitably be present in any transcriptomic dataset (Raz et al., 2011; Smith, 2013; Tian & Smith, 2016). This can be leveraged to extract useful sequence information from transcriptomic assembly (Tian & Smith, 2016). RNA-seq has been applied to almost every major clade in the tree of life, and has proven its utility repeatedly for solving myriad questions in life's evolution. The extraction of mRNA from mixed or single tissues, followed by conversion to cDNA and the construction of libraries from this for sequencing, has become a standard technique in the study of many realms of molecular biology.

The basic technique of RNA-seq is often supplemented by a step which removes ribosomal and transfer RNA from the RNA sample, leaving mRNA in higher relative abundance. These poly-A enrichment techniques make use of the polyadenylation of RNA polymerase 2 products in eukaryotic cells (Hirose & Manley, 1998), although it is important to note that some prokaryote mRNAs are also polyadenylated (Régnier & Marujo, 2013). Polyadenylated mRNAs are bound to ligands (often with a poly-T "bait"), and subsequently separated from the remaining rRNA and tRNA fraction, which can comprise around 95% of the RNA in a cell. Mitochondrial RNA is polyadenylated in some branches of the tree of life but not others (Chang & Tong, 2012; Bratic et al., 2016). In some clades, a polyadenylation signal marks RNA for degradation (Chang & Tong, 2012). Mitochondrial RNA will therefore be variably recovered after poly-A enrichment methods, depending on the clade from which the RNA was extracted, and the role of polyadenylation within that group. Other methods of rRNA removal, for instance "Ribozero" approaches, are also used, although less frequently.

A plethora of mitochondrial assembly tools are available, but these generally rely on gDNA or Sanger-derived reads, rather than those sourced from RNA-seq experiments. Programmes such as Norgal (Al-Nakeeb et al., 2017), MitoBIM (Hahn et al., 2013), and NOVOPlasty (Dierckxsens et al., 2016) have proven useful at recovering mitogenome assemblies, particularly when the sequence of a closely related species is available for use as a map for assembly. Increasingly sophisticated tools are available for using gDNA for assembly (e.g. Schomaker-Bastos & Prosdocimi, 2008) However, previous approaches, and particularly those reliant on mining the results from transcriptomic assembly programs, are not always reliable when using mRNA sequence data as the basis for assembly, especially when using reads that have been subject to rRNA removal or mRNA enrichment (Tian & Smith, 2016).

Here we describe a novel pipeline for the reconstruction of mitochondrial gene cassettes and whole coding sequences from RNAseq reads, based on existing, freely available programmes. This pipeline uses a sequential approach and established tools, so that RNA-seq reads of good quality and uniform coverage across the mitogenome will quickly be assembled into workable data, while those datasets posing problems can still yield results, albeit with additional effort. We have benchmarked this process using reads derived from a number of Illumina platforms, read lengths and sequencing depths, and show that reliable mitochondrial genome assemblies, and particularly the sequence of protein coding genes, can be reliably recovered from even poly-A selected samples. This method will allow the recovery of mitochondrial genome data from both historical and novel RNA-seq experiments, and provide a variety of novel resources to the scientific community for use in the myriad of applications for which mitogenomes have proven their utility.

Materials and Methods:

In this study we used RNA-seq data from eight representative species encompassing six phyla. We analysed published RNA-seq datasets from five phyla, with SRA IDs indicated in Table 1, and also generated new data for one sponge and one nemertean.

The species *Corticium candelabrum*, *Biomphalaria glabrata*, *Platynereis dumerilii, Stenostomum sthenum* and *Thermobia domestica* all possess previously published mitochondrial genomes (mt genomes), and these were used to estimate the accuracy and efficiency of the proposed workflow.

As a test of the pathway on novel data, for the nemertean *Antarctonemertes valida* and the poriferan *Spongia officinalis* new RNA-seq data were generated and deposited in the SRA under the following accession numbers: SRP157324 and SRP150632. The newly assembled mt genomes for these two species, and that of *I. fasciculata,* have been submitted to GenBank under accession numbers MH768970-MH768972.

RNA extraction, library preparation, Illumina sequencing

For *S. officinalis* (Koutsouveli et al., unpublished data), total RNA was extracted from 20 tissue pieces (4 different individuals and 5 replicates per individual) with TRIzol (Ambion) and mRNA purified with Ribo-Zero™ (Illumina). Library preparation of all 20 samples was performed with TruSeq Stranded Total RNA Library Prep Kit (Illumina) and further sequenced with an Illumina HiSeq 2000 using a paired-end (150 bp length) sequencing strategy.

A single individual divided into three portions (anterior part, posterior part, and proboscis) was used for *A. valida.* Total RNA extraction was performed using TRIzol (Ambion) and mRNA purification with Dynabeads mRNA DIRECT Purification kit (ThermoFisher

Scientific) following established protocols (see Riesgo et al., 2014). Library preparation for *A. valida* was performed with the ScriptSeq v2 RNA-seq library preparation kit (Epicentre) and the three libraries were sequenced on an Illumina NextSeq500 to a length of 150 bp (paired-end).

Details on the RNA extraction, library preparation and sequencing for *C. candelabrum* and *I. fasciculata* can be found in Riesgo et al. (2014), while those for *B. glabrata* are detailed in Kenny et al. (2016) and for *P. dumerilii* in Achim et al. (2018).

Recovering mt genomes

The quality assessment of all reads was performed with FastQC (Andrews, 2010) and the mean quality value across each base position in the read was obtained with MultiQC (Ewels at al., 2016). As outlined in the workflow for Trimitomics (Fig.1), our pipeline is comprised of three sequential steps using different software that are used stepwise, depending on the success of mt genome assembly in the preceding step. The first step consists of the use of the NOVOPlasty version 2.7.1 organelle assembler (Dierckxsens et al., 2016), with a range of *k*-mer distributions, using the raw reads as input. If the full mt genome is not recovered, and no or only a partial mt genome is obtained, the raw reads are then cleaned using Trimmomatic 0.33 (Bolger et al., 2014) with the following settings: ILLUMINACLIP:./Adapters.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:30, with the Adapters.fa file adjusted to include the adapter sequences specific to each read pair. RNA-seq reads are then mapped to their respective "reference" genome (the closest mt genome published) with Bowtie2 (Langmead et al., 2012), using default settings. Mapped reads are then assembled with genome guided Trinity version v2.8.4 (Grabherr et al., 2011; Haas et al., 2013).

If neither of the above-mentioned methods successfully retrieved mt genomes, then the whole transcriptome was assembled using Velvet 1.2.10 (Zerbino & Birney 2008) with a range of *k*-mer sizes (31, 51, 71). Mitochondrial contigs were then mined from *de novo* transcriptome assemblies with BlastN (Altschul et al., 1990) using the "reference" genome for the analysed species. Custom database with mitogenomes of interest can be used at this point as well. In cases where existing mitogenomes are genetically divergent BLASTX could be beneficial. A comprehensive guideline script for the pipeline Trimitomics is deposited at github https://github.com/bplese/Trimitomics.

If the complete mt genome is not recovered by any of the above three methods, the results are combined as a meta-assembly in order to obtain the best results. When the retrieved mt genomes are obtained in several contigs, this meta-assembly could be performed in Geneious v. 10.2.4, but can also be done manually by comparison of contig ends using freely available alignment softwares such as MAFFT (Katoh et al., 2002) or CLUSTAL (Larkin et al., 2007). When the complete or almost complete mt genome is obtained, the assembly data is checked for homology with BlastN against the NCBI nr database. Further annotation is then performed on the web server MITOS2 (Bernt et al., 2013) using the appropriate translation table.

In the cases of *C. candelabrum*, *B. glabrata*, *P. dumerilii*, *S. sthenum* and *T. domestica* the assembled and annotated mt genomes acquired with our pipeline were aligned with their respective published mt genomes using MAFFT (Katoh et al., 2002) to estimate the accuracy of the mt genome retrieval.

For all assembled mt genomes, data alignment statistics were obtained with SAMTOOLS (Li et al., 2009) after the RNA-seq reads were mapped to their assembled mt genome with Bowtie2. Potential PCR duplicates were marked and subsequently removed with the Picard tool

(http://broadinstitute.github.io/picard). Uniquely mapped mitochondrial reads in the transcriptomes were determined with the RSeQC package with a minimum mapping quality 30 (Wang et al., 2012). Alignment statistics and coverage graphs were plotted using the ggplot R package in Rstudio (Wickham, 2016). Read coverage of each of the recovered mt genomes was generated with deepTools Galaxy (Fidel et al., 2016) with a bin size of 50. Coverage of intergenic regions, tRNAs, protein coding and ribosomal genes was calculated with bedtools v2.27.0 (Quinlan & Hall 2010). Summary runtime metrics comparison about system memory and processes, using *Biomphalaria glabrata* as example dataset, was done with vmstat. For this purpose we used paired-end reads, single-end reads and two sets of reads subsampled by reads number. Physical coverage for the whole mt genome was calculated as follows: (number of reads x average read length)/genome length. We computed linear regressions between the number of reads and the percentage of mt genome recovery for each of the methods to understand whether there existed a relationship between those variables, using the R package *lattice* (CRAN.R-project.org, Sarkar, 2008).

Results

Quality of RNA-seq data

The quality control checks on raw sequence data for all analyzed transcriptomes are summarized in Supplementary File 1. Most reads from all datasets showed average quality scores (Phred score) in the range 30–40, with the sponge *C. candelabrum* as the sole exception. In this species, a Phred score plateau of 20 was observed at the beginning and at the end of the reads. The majority of the analyzed raw reads were up to 100 bp in length, except for *S. officinalis* and *A. valida*, which were up to 150 bp (Supplementary File 1).

Assembling mitochondrial genomes from RNA-seq data

In this study, we aimed to retrieve mt genomes from transcriptomic data sourced from six phyla: Porifera, Nemertea, Mollusca, Annelida, Platyhelminthes and Arthropoda. Efficiency of the proposed Trimitomics workflow is summarized in Table 1 and discussed in detail below. It is important to note that there was no correlation between the number of input reads and the percentage of recovery of the mt genome for any of the three steps of Trimitomics (Supplementary File 2).

Porifera

For *C. candelabrum,* NOVOPlasty with *k*-mer size 39 resulted in one contig of 12,773 bp in size, thus recovering 69.4% of the expected genome size. This assembly of the mt genome showed 90.6% sequence similarity with that published previously. When using the Bowtie2/Trinity combination, we obtained 7 contigs with mitochondrial similarity (size: 215–6,335 bp). The final meta-assembly of these contigs in Geneious allowed us to recover 17,556 bp

(95.4% of the total genome size) in two contigs. The similarity percentage with the published mt genome in this case was 95%. Finally, our Velvet assembly with *k*-mer size 71 initially gave 10 contigs (size: 339–4,293 bp), that were later assembled in a contig that recovered 57% of the mt genome with 94.5% accuracy. When we combined the results from NOVOPlasty and Bowtie2/Trinity, we obtained an almost complete mitochondrial genome (99.4% mt genome recovery) with only 120 bp missing within an intergenic region (Fig. 2).

In the case of *S. officinalis*, the complete mt genome was recovered by NOVOPlasty alone (Fig. 2). The circularized 16,115 bp long mt genome is the first representative published for the genus *Spongia*. The mt genome of the closest related species *Hippospongia lachne* (family Spongiidae) was used as a reference, and the MAFFT alignment of both species, *S. officinalis* and *H. lachne*, showed 57% identity. The Bowtie2/Trinity combination initially resulted in 18 contigs (size: 215–8,649 bp). The meta-assembly of such contigs gave 14,303 bp in 2 contigs. Finally, Velvet using a *k*-mer size of 71 gave 26 contigs with lengths 71–1,893 bp, that were further assembled into 3 contigs (13,945 bp), which represented 96% of the mt genome (Table 1).

For *I. fasciculata* the expected mt genome size was roughly estimated from the median size of already published mt genomes from the congeneric *Ircinia strobilina* (NC_013662.1) and *Ircinia* sp. (KC510273.1) to be approximately 16,000 bp. NOVOPlasty (*k*-mer size 45) assembly resulted in 3 contigs with length ranging from 314–9,639 bp. The meta-assembly of these 3 contigs resulted in one contig of 11,669 bp. The Bowtie2/Trinity combination using *Ircinia strobilina* as reference mt genome gave 7 contigs that could be assembled into one contig with 11,543 bp in total. Velvet (*k*-mer 71) gave 14 contigs (size: 225–5,553 bp) that were assembled for the final mt genome of 13,945 bp in 3 contigs. When the results from NOVOPlasty and

Velvet were combined we attained 98.3% of mt genome recovery with a sequence length of 15,954 bp and all protein coding genes recovered (*nad2* partial at the 5' end), as well as rrnS and rrnL. Some parts of one intergenic region and 2 tRNAs were missing (Fig. 2).

Nemertea

For *Antarctonemertes valida*, the closest related species with already published mt genome is *Gononemertes parasita,* from the same suborder *Eumonostilifera*. Thus, this nemertean species, with mt genome size of 14,742 bp was used as reference genome.

With NOVOPlasty at a range of *k*-mer sizes only 3,088 bp were obtained. With Bowtie2/Trinity initially gave 13 contigs of sizes 321–3,876 bp. Velvet *de novo* assembly, using all 3 *k*-mer sizes, a total of 8,884 bp in 12 contigs were recovered. Final meta-assembly recovered a partial mt genome with 13,285 bp in 10 contigs representing 90.12% of the expected mt genome, with only *nad2* completely missing. In addition, *nad1*, *nad4, nad5* and *cox2* were partially obtained and 4 tRNAs were missing.

Mollusca

For *B. glabrata*, 97% of the genome was recovered from the raw sequencing reads with NOVOPlasty (*k*-mer size 39) in one contig, with accuracy of 99.6% when compared to the published mt genome, leaving only 616 bp unaccounted for. Bowtie2/Trinity resulted in 5 contigs with length ranging from 447–8,584 bp, thus recovering 99.4% of the mt genome. The meta-assembly of these 5 contigs resulted in 13,590 bp in one contig with 99.8% accuracy. Finally, assembly from Velvet gave 19 contigs with length from 141–1,329 bp with 84.6% accuracy. Final meta-assembly resulted in complete mt genome recovery with 99.9% accuracy.

Annelida

For *P. dumerilii,* NOVOPlasty assemblies with *k*-mer size 25, 39 and 45 gave poor results (<600bp), although *k*-mer size 51 gave results of 1,573 bp in one contig with 99.3% accuracy. Velvet assembly resulted in 3,353 bp in 23 contigs (with size ranging from 141–251 bp) with 89.7% accuracy. The combination of Bowtie2/Trinity resulted in 11 contigs (size: 324–3,929 bp). Final meta-assembly combining all of the above data sources resulted in 10 contigs encompassing 99.5% of the genome with 98.9% accuracy. One tRNA was missing and *nad5* partial at the 5' end.

Platyhelminthes

In the case of *S. sthenum,* the NOVOPlasty assembly (*k*-mer size 39) recovered 92.8% of the complete mt genome in 11 contigs (with size ranging from 123-4,658 bp) with 99.97% accuracy. Bowtie2/Trinity assembly retrived 14,651 bp in 4 contigs (size: 764-8,525 bp) with 99.2% identity. Velvet assembly (*k*-mer 71) gave 594 contigs with length from 127-600 bp and 99.2% accuracy. Final meta-assembly recovered an almost complete mt genome with only 3 nucleotides missing, and accuracy 99.9% (Fig. 2).

Arthropoda

For *T. domestica,* the NOVOPlasty assembly resulted in less than 50% recovery of the mt genome. Three contigs (size: 98 – 7,221 bp) recovered only 7,260 bp with 91.2% accuracy. 99% of the complete mt genome was recovered with Bowtie2/Trinity assembly. Two contigs gave 15,020 bp mt genome with 90.5% accuracy. Similar results were obtained with Velvet assembly (*k*-mer 71) where 611 contigs ranging from 113-1,392 bp recovered 99.5% of complete mt

genome. Final meta-assembly combining all of the above data sources resulted in an almost complete mt genome with only 13 nucleotides missing, scattered throughout the mt genome.

Read mapping and coverage for the mt genome

Our analyzed transcriptomic data exhibited diverse physical coverage (calculation in Materials and Methods) of mitochondrial-derived reads, ranging from an average of 549.54 in the sponge *C. candelabrum* to 177,755.57 in the annelid *P. dumerilii* (Table 1). Regardless of this diversity, in the majority of samples spanning six different phyla more than 98% of the mt genome was recovered. The only exception is the nemertean *A. valida* where 90% of mt genome was recovered. The alignment statistics of exclusively mitochondrial reads are presented in Supplementary File 3. Average percentage of duplicate reads in the studied RNA-seq datasets was 6-9% for most species (Supplementary File 3), with the exception of *S. officinalis, C. candelabrum* and *B. glabrata* (up to 40, 24, and 19%, respectively). The number of removed reads and those uniquely mapped to the analyzed mt genome was quite similar among the different taxa used in our study (Supplementary File 3).

It is not quite clear whether these reads are artefactual duplicates that occurred during library construction or indeed represent independent mitochondrial reads in the transcriptomic data. Mt genomes are relatively small and are present in high copy number in cells and these percentages could reflect current situation in the analyzed species. Nevertheless, to ensure unbiased analysis for further studies these reads were removed.

Mapping of RNA-seq data onto the retrieved mt genomes gave a first insight into the transcription of the mitochondrial genome in our target invertebrates (Fig. 2 and Supplementary File 4). Even though exhibiting lower coverage in comparision with protein coding or ribosomal

genes, intergenic regions were recovered well, ensuring complete mt genome recovery. Overall mitochondrial expression patterns were quite diverse among the analyzed species.

In Porifera both small and large ribosomal genes (*rrnS* and *rrnL*) were more expressed than the protein coding genes, although the number of reads mapped to those ribosomal genes varied substantially among datasets (Supplementary table 1 and Fig. 3). In *S. sthenum* (Platyhelminthes), *T. domestica* (Arthropoda), *B. glabrata* (Mollusca), *P. dumerilii* (Annelida) and *A. valida* (Nemertea), a similar pattern was observed, with *rrnL* as the most expressed gene followed by *cox1*.

Overall, in Porifera the second-most covered genes by RNA-seq data were protein coding genes of respiratory complex I, while in Nemertea, Mollusca, Annelida, Platyhelminthes and Arthropoda proteins of respiratory complex IV were the next most well-expressed (Fig. 3). *C. candelabrum* (Porifera) and *A. valida* (Nemertea) showed similar expression levels of these two mitochondrial complexes. These values can be used to approximate expression internally within samples gained with a consistent library construction technique, but should not be taken as definitive proof of expression values due to the poly-A selection steps incorporated into library construction, which will favour transcripts proximal to polyadenylation signals, whether incorporated by RNA polymerases or natively present within the raw mitochondrial sequence.

Discussion

The 'omics' revolution and the development of novel bioinformatic tools that are now routinely available at little cost to all laboratories has provided a wealth of data from almost all animal groups. Nowadays, genome skimming and RNA-seq projects are feasible for non-model organisms and the SRA database grows larger on a daily basis. Despite awareness that RNA-seq data contain high numbers of mt genome reads (Smith, 2013), the majority of available methods for mt genome assembly are customized for recovering these from genome sequencing projects (either whole genome sequencing or genome skimming). Nevertheless, the mining of mt genomes from RNA-seq data has been successfully performed in a handful of eukaryotes, including algae (Tian & Smith, 2016), reptiles (Lyra et al., 2017), sea urchins (Dilly et al., 2015) and bony fishes (Moreira et al., 2015), yet using disparate software approaches, including Bowtie2/Trinity, Trinity assembly, MIRA v4.0/MITObim v1.8 (Lyra et al., 2017), and blast against the mt genomes of closely related species (Dilly et al., 2015). In the case of incomplete mt genome recovery from transcriptomic data, gaps were filled by combining with the genomic data (Nan Song et al., 2016; Fabre et al., 2013). Here we propose a tentative workflow for recovering mt genomes from RNA-seq data across Metazoa. Using data from a number of non-model invertebrates spanning 6 phyla we recovered with our pipeline more than 98% of the mt genome in the majority of cases (Table 1). Trimitomics combine three different methods in order to recover maximum posssible mt genome. This is a first study with detailed representation and summary of efficiency of the proposed workflow.

In Trimitomics (Fig. 1) NOVOPlasty is the first tool implemented, and was empirically superior in runtime and memory consumption when compared with other existing tools for mitochondrial assembly (Supplementary Table 2). This is expected, as it is a newly published

pipeline which incorporates advancements in bioinformatics tools targeting the mt genome specifically (Dierckxsens et al., 2016).

The Bowtie2/Trinity approach has been used previously to recover mt genomes (Tian & Smith, 2016), adding the coverage and expression levels of each of the genes in the mitochondrial chromosomes (Tian & Smith, 2016). Although, memory and time consuming (Supplementary Table 2), it has proven useful in this analysis as well. For the poriferan *C. candelabrum*, nemertean *A. valida*, mollusca *B. glabrata* and the annelid *P. dumerilii*, Bowtie2/Trinity yielded the best results. However, the initial identification of reads as being of mitochondrial origin depends on inference of homology using Bowtie2 mapping, and thus requires a relatively closely-related sequence to act as "bait". Without this inference of homology, reads are simply excluded from the following Trinity assembly. We therefore recommend that if "unspannable" gaps are found when using the Bowtie2/Trinity approach, a *de novo* assembly with Velvet is assayed and contigs from these are tested to see if they can be used to span such problematic areas. Velvet can be memory and time consuming approach (Supplementary Table 2), as this assembler was designed primarily for genomic read data, it is optimised to assemble reads of relatively uniform coverage. It can therefore be optimised further for assembly of specific regions of the mitochondrion using the –min_cov, –max_cov and –exp_cov settings once general coverage levels are known, but this can be a time-consuming approach, and is beyond the scope of this study. However, in many cases it yields significant results. For example, in the case of *I. fasciculata* Velvet increased the final mt genome assembly by 27%.

Efficiency in the recovery of the complete mt genomes, as expected, depends on RNA-seq library preparation, quality of read data and the depth of sequencing. In this paper, it was not

our goal to compare the suitability of available methods of library preparation, data quality or sequencing depth in the recovery of the mt genome but rather to demonstrate that the complete mt genome of understudied species could be obtained with our pipeline using a diverse set of RNA-seq data. Complete mt genome recovery is data and phylum specific. Neither of the methods worked perfectly thus the advantge of Trimitomics is in the combined aproach.

The recovery of complete (rather than CDS-containing only), mt genomes from RNA-seq datasets could be due to either genomic contamination, incomplete DNA digestion during RNA library preparation or could be an indication that the majority of the mt genome is actually transcribed (Tian & Smith, 2016), even if only transiently before RNA editing. Even the intergenic regions were recovered in RNA-seq datasets, which enabled complete mt genome recovery.

Thus, given that in many different laboratories RNA-seq experiments are performed more and more routinely, our Trimitomics pipeline might become a powerful approach to exploit this neglected goldmine of novel mitochondrial genomes within RNA-seq data and ensure their future usage to address various biological and evolutionary questions. In any case, the recovery of even partial mt genomes with our pipeline in non-model invertebrates represents a significant contribution, as in these taxa mitochondrial genome data is scarce. We have shown that even when the mitochondrial physical coverage is low, as in the cases of *C. candelabrum* and *S. officinalis,* almost complete mt genomes can be obtained (Table 1).

mRNA read data generated using standard methods are not sufficient to fully characterize mitochondrial expression patterns, as most library preparation methods target transcripts with a poly-A tail, which is not universally the case in mitochondrial genes. However, it can provide information regarding general transcriptional trends, especially on a within-species basis, that

can be explored further. Little has been published about the baseline gene expression levels of mitochondrial genes in invertebrates (e.g., Wang et al., 2013; Perera et al., 2016).

As expected, and previously reported for other eukaryotic groups (e.g., Tian & Smith 2016; Wang et al., 2013; Perera et al., 2016), coverage in our datasets was lowest in tRNAs and respective intergenic regions and highest in areas encoding either rRNAs or genes within respiratory complex I and IV (Supplementary table 1.). Interestingly, our results for the transcriptional landscape across invertebrate phyla indicates that the transcriptional profiles are diverse, even among the three analyzed poriferan species. Besides the ribosomal genes (especially *rrnL*), the cytochromes (*cox1, cox2,* and *cox3*) and *atp6* are usually the most transcribed in animal groups (e.g., Wang et al., 2013; Perera et al., 2016). This was true for the sponge *C. candelabrum* (although only for *cox3* and *atp6*) and the mollusc *B. glabrata* (Supplementary table 1). In the latter mentioned *cob* was slightly more expressed then *atp6*. In the annelid *P. dumerilii* and platyhelminth *S. sthenum, cox1, nad5* and *cob* showed high levels of expression. Contrastingly, in the arthropod *T. domestica nad4* and *cob* were the most expressed mitochondrial gene besides respiratory complex IV. In the dictyoceratid sponge *S. officinalis,* high levels of expression were seen in the proteins of respiratory complex I as well as in *cox1* (Fig. 3 and Supplementary table 1). In *I. fasciculata cob*, *atp6, cox3* and *nad5* showed significant expression (Supplementary table 1). This indicates that mitochondrial genes are expressed variable according to the animal species sampled, and, presumably, the current energy requirements of the sample taken.

By way of example, although ribosomal genes are usually the most expressed genes within the mitochondria in the lepidopteran *Helicoverpa zea*, in its embryos *cox1* is by far the most highly expressed gene (Perera et al., 2016), while in the vole *Clethrionomys glareolus cox1,*

cox2, cox3 and *atp6* are expressed more than ribosomal genes (Markova et al., 2015). Our pathway might provide a means to determine this information in a more systematic manner in any previously published RNA-seq experiment, provided library construction was performed in an internally-consistent manner, and thus provides a pathway to understanding the mitochondrial transcriptional landscape in any interesting biological framework, either by leveraging extant resources, or through novel investigations in the future.

Conclusions:

Mitochondrial data are of diverse utility, and can be particularly crucial for investigations into the biology and evolution of non-model organisms. Here we have demonstrated a means of rapid and relatively inexpensive derivation of coding and full length mitochondrial sequences from a range of non-model species. This method will allow new information to be leveraged from extant datasets, with minimal or no cost, which will open doors for investigations in even the most problematic clades.

Acknowledgements:

The authors thank Dr Cristina Diez for her helpful discussion and support and Dr Peter Foster for IT support. Carlos Leiva and Dr Juan Junoy are greatly acknowledged for collecting *A. valida* samples used in this study. This work was supported by a grant from the Villum Foundation (grant number 9278). NJK was supported by a H2020 MSCA grant during manuscript preparation [IF750937].

References:

Achim, K., Eling, N., Vergara, H. M., Bertucci, P.Y., Musser, J., Vopalensky, P., Brunet, T., Collier, P., Benes, V., Marioni, J.C. & Arendt, D. (2018). Whole-Body Single-Cell Sequencing Reveals Transcriptional Domains in the Annelid Larval Body, *Molecular Biology and Evolution*, 35(5), 1047–1062.

Al-Nakeeb, K., Petersen, T.N. & Sicheritz-Pontén, T. (2017). Norgal: extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data. *BMC bioinformatics*, *18*(1), 510.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. (http://www.bioinformatics.babraham.ac.uk/projects/fastqc)

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A. & Saunders, N.C. (1987). Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual review of ecology and systematics*, *18*(1), 489-522.

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.

Bolger, A.M., Lohse, M. & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, btu170.

Bratic, A., Clemente, P., Calvo-Garrido, J., Maffezzini, C., Felser, A., Wibom, R., … Wredenberg, A. (2016). Mitochondrial Polyadenylation Is a One-Step Process Required for mRNA Integrity and tRNA Maturation. *PLoS Genet* 12(5): e1006028.

Chang, J.H. & Tong, L. (2012). Mitochondrial poly (A) polymerase and polyadenylation. *Biochimica et Biophysica Acta* (BBA)-Gene Regulatory Mechanisms 1819(9), 992-7.
da Fonseca, R.R., Johnson, W.E., O'Brien, S.J., Ramos, M.J. & Antunes, A. (2008). The adaptive evolution of the mammalian mitochondrial genome. *BMC Genomics* 9(1), 119.

da Fonseca, R. R., Johnson, W. E., O'Brien, S. J., Ramos, M. J., & Antunes, A. (2008). The adaptive evolution of the mammalian mitochondrial genome. *BMC genomics*, *9*(1), 119.

Dierckxsens, N., Mardulyn, P. & Smits, G. (2016). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic acids research*, *45*(4), 18-18.

Dilly, G.F., Gaitán-Espitia, J.D. & Hofmann, G.E. (2015). Characterization of the Antarctic sea urchin (*Sterechinus neumayeri*) transcriptome and mitogenome: a molecular resource for

phylogenetics, ecophysiology and global change biology. *Molecular Ecology Resources* 15(2), 425-436.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19),3047–3048.

Fabre, PH., Jønsson, KA., Douzery, EJP. (2013). Jumping and gliding rodents: mitogenomic affinities of Pedetidae and Anomaluridae deduced from an RNA-Seq approach. Gene 531, 388–397.

Ramírez, F., Ryan, DP., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, AS., Heyne, S., Dündar, F., and Manke, T.(2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research, 44* (W1), pp. W160–W165.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*(7), 644–652. http://doi.org/10.1038/nbt.1883

Hahn, C., Bachmann, L. & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic acids research*, *41*(13), e129.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., …. Regrev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8, 1494–1512.

Hirose, Y. & Manley, J.L. (1998). RNA polymerase II is an essential mRNA polyadenylation factor. *Nature*, 395(6697), 93.

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.

Kenny, N. J., Truchado-García, M., & Grande, C. (2016). Deep, multi-stage transcriptome of the schistosomiasis vector *Biomphalaria glabrata* provides platform for understanding molluscan disease-related pathways. *BMC Infectious Diseases*, *16*, 618. http://doi.org/10.1186/s12879-016-1944-x

Langmead, B. & Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23, 2947-2948.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Lyra, M.L., Haddad, C.F.B. & Azeredo-Espin, A.M.L. (2017). Meeting the challenge of DNA barcoding Neotropical amphibians: polymerase chain reaction optimization and new COI primers. *Molecular Ecology Resources*, 17, 966–980.

Marková, S., Filipi, K., Searle, J. B., & Kotlík, P. (2015). Mapping 3′ transcript ends in the bank vole (*Clethrionomys glareolus*) mitochondrial genome with RNA-Seq. *BMC genomics*, *16*(1), 870.

Melton, T., Holland, C. & Holland, M. (2012). Forensic Mitochondria DNA Analysis: Current Practice and Future Potential. *Forensic science review*, *24*(2), 101.

Mercer, TR., Neph, S., Dinger, ME., Crawford, J., Smith, MA., Shearwood, AMJ., Haugen, E., Bracken, CP., Rackham, O., Stamatoyannopoulos, JA., Filipovska, A., Mattick1, JS. (2011). The Human Mitochondrial Transcriptome. Cell 146, 645–658.

Moreira, DA., Furtado, C., Parente, TE (2015). The use of transcriptomic next-generation sequencing data to assemble mitochondrial genomes of Ancistrus spp. (Loricariidae). Gene 573(1), 171–175.

Perera, O.P., Walsh, T.K., & Luttrell, R.G. (2016). Complete Mitochondrial Genome of *Helicoverpa zea* (Lepidoptera: Noctuidae) and Expression Profiles of Mitochondrial-Encoded Genes in Early and Late Embryos. *Journal of Insect Science*, *16*(1), 40. http://doi.org/10.1093/jisesa/iew023

Picard, M., Wallace, D.C. & Burelle, Y. (2016). The rise of mitochondria in medicine. *Mitochondrion*, *30*, 105-116.

Quinlan, AR., Hall, IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Raz, T., Kapranov, P., Lipson, D., Letovsky, S., Milos, P.M. & Thompson J.F. (2011). Protocol dependence of sequencing-based gene expression measurements. *PLoS ONE* 6, e19287.

Régnier, P. & Marujo, P.E. (2013). Polyadenylation and Degradation of RNA in Prokaryotes.

Riesgo, A., Farrar, N., Windsor, P.J., Giribet, G. & Leys, S. P. (2014). The Analysis of Eight Transcriptomes from All Poriferan Classes Reveals Surprising Genetic Complexity in Sponges, *Molecular Biology and Evolution*, 31(5), 1102–1120, https://doi.org/10.1093/molbev/msu057

Romero, P.E., Weigand, A.M. & Pfenninger, M. (2016). Positive selection on panpulmonate mitogenomes provide new clues on adaptations to terrestrial life. *BMC evolutionary biology*, *16*(1), 164.

Rubinoff, D. (2006). Utility of mitochondrial DNA barcodes in species conservation. *Conservation Biology*, *20*(4), 1026-1033.

Sarkar, D. (2008). *Lattice*: Multivariate Data Visualization with *R*. Springer, New York.

Schomaker-Bastos, A., & Prosdocimi, F. (2018). mitoMaker: A Pipeline for Automatic Assembly and Annotation of Animal Mitochondria Using Raw NGS Data. *Preprints*, 2018080423 (doi: 10.20944/preprints201808.0423.v1).

Song, N., An, S., Yin, x., Cai, w., Li, H.. (2016). Application of RNA-seq for mitogenome reconstruction, and reconsideration of long-branch artifacts in Hemiptera phylogeny. Scientific Reports 6, 33465. DOI: 10.1038/srep3346

Smith, D.R. (2013). RNA-Seq data: a goldmine for organelle research. *Brief. Funct. Genom.* 12, 454–456.

Tian, Y. & Smith, DR. (2016). Recovering complete mitochondrial genome sequences from RNA-Seq: A case study of Polytomella non-photosynthetic green algae. *Molecular Phylogenetics and Evolution* 98, 57-62.

Wang, H. L., Yang, J., Boykin, L. M., Zhao, Q. Y., Li, Q., Wang, X. W. & Liu, S. S. (2013). The characteristics and expression profiles of the mitochondrial genome for the Mediterranean species of the *Bemisia tabaci* complex. *BMC genomics*, *14*(1), 401.

Wang, L., Wang, S. & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 28(16), 2184–2185,https://doi.org/10.1093/bioinformatics/bts356

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Zerbino, D. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*,18, 821–829.

Data Accessibility Statement:

Raw reads for *Antarctonemertes valida* and *Spongia officinalis* were deposited at SRA with accession numbers SRP157324 and SRP150632. The newly assembled mt genomes for *Spongia officinalis*, *Antarctonemertes valida* and *Ircinia fasciculata* were deposited at Genbank with the following accessions: MH768970-MH768972.

Author Contributions:

AR and BP conceived the study and gathered specimens and datasets. BP, ST, VK, MER and NJK performed RNA extraction and library preparation. BP and NJK designed the pipeline. BP and MER performed assembly and bioinformatic analysis. AR provided the funding necessary to perform experiments and analyses. BP wrote a first draft of the manuscript and all authors made major contributions to the writing. All authors reviewed the final version of the manuscript.

Figure legends

Figure 1. Proposed Trimitomics pipeline workflow. NOVOPlasty organelle assembler with range of *k*-mer distributions is applied to raw reads (1). Preferably a complete circularized mt genome is acquired, or several contigs are obtained. In the latter case, the resulting contigs are merged into supercontigs by mapping to the reference mt genome to recover the complete mt genome. If mt genome is still partial we quality trim the raw reads with Trimmomatic and use them in reference-guided Bowtie2 alignment (2). Thereafter, mapped reads corresponding to the mt genome are extracted and further assembled with genome-guided Trinity assembly. If the mt genome is still partial we use Velvet assembly (3) with range of *k*-mer values. From the resulting assembly, mt genome contigs are extracted, baiting with a reference mt genome. If possible, resulted contigs are assembled into supercontigs. If none of the proposed methods acquired a complete mt genome, all resulting contigs were merged into supercontigs in order to get the best results. In the final steps, supercontigs are checked for homology with BlastN against the NCBI nr database and annotated using Mitos2.

Figure 2. Gene order and read coverage for the obtained mt genomes from the eight species analysed in this study. Scale bar represents the size of retrieved mt genomes. Associated numbers on the left side represent minumum and maximum read coverage. Mean read coverage is as follows from down to up: 37,344, 27,564.50, 59,826.50, 27,303.50, 11,428.00, 1,844.50, 609.50 and 1,878.50. Below is annotation with genes, rRNAs, tRNAs and non coding regions (white space). Mitochondrial expression patterns are shown as read coverage graphs above.

Figure 3. Percentage of mt reads aligned to each of the ribosomal genes (rrnl and rrns), respiratory complex I-V, intergenic regions (IGR) and tRNAs.

Supplementary File 1. Quality assessment of all analysed raw reads expressed as mean quality value across each base position in the read. Each species, labelled with different colour as outlined in the legend, has two data sets for forward and reverse paired reads.

Supplementary File 2. Correlation between number of reads used as input and the percentage of mt genome recovery for each of the three steps of Trimitomics.

Supplementary File 3. Alignment statistics for the number of reads obtained for the eight species analysed in this study. Total mitochondrial reads count in analysed transcriptomes (reads count, in red). After filtering potential PCR duplicates (removed, in blue) and those that are uniquely mapped (unique, in green).

Supplementary File 4. Read coverage (log scale) for the obtained mt genomes from the eight species analysed in this study.

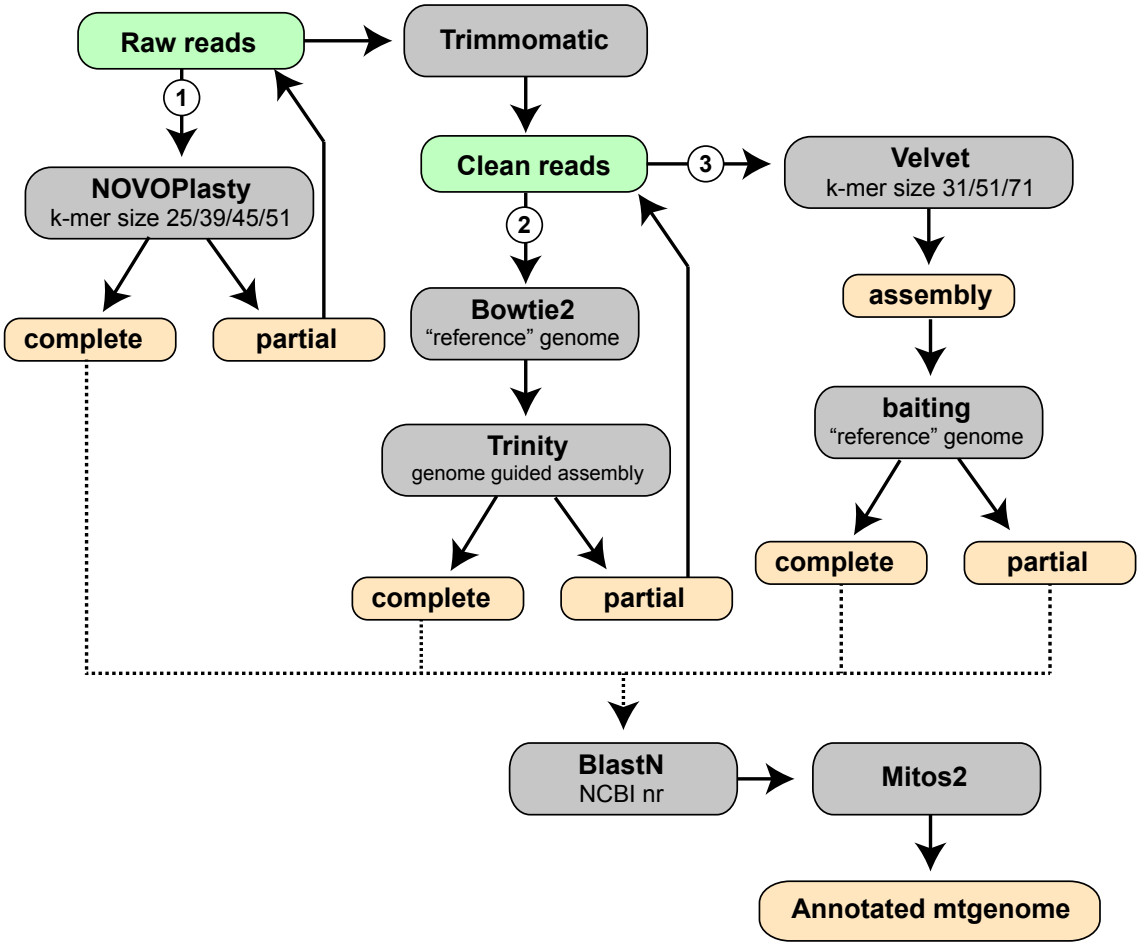
Raw reads
Trimmomatic
1
NOVOPlasty
k-mer size 25/39/45/51
complete
partial
Clean reads
3
Velvet
k-mer size 31/51/71
2
Bowtie2
“reference” genome
assembly
baiting
“reference” genome
Trinity
genome guided assembly
complete
partial
complete
partial
BlastN
NCBI nr
Mitos2
Annotated mtgenome

Fig 1

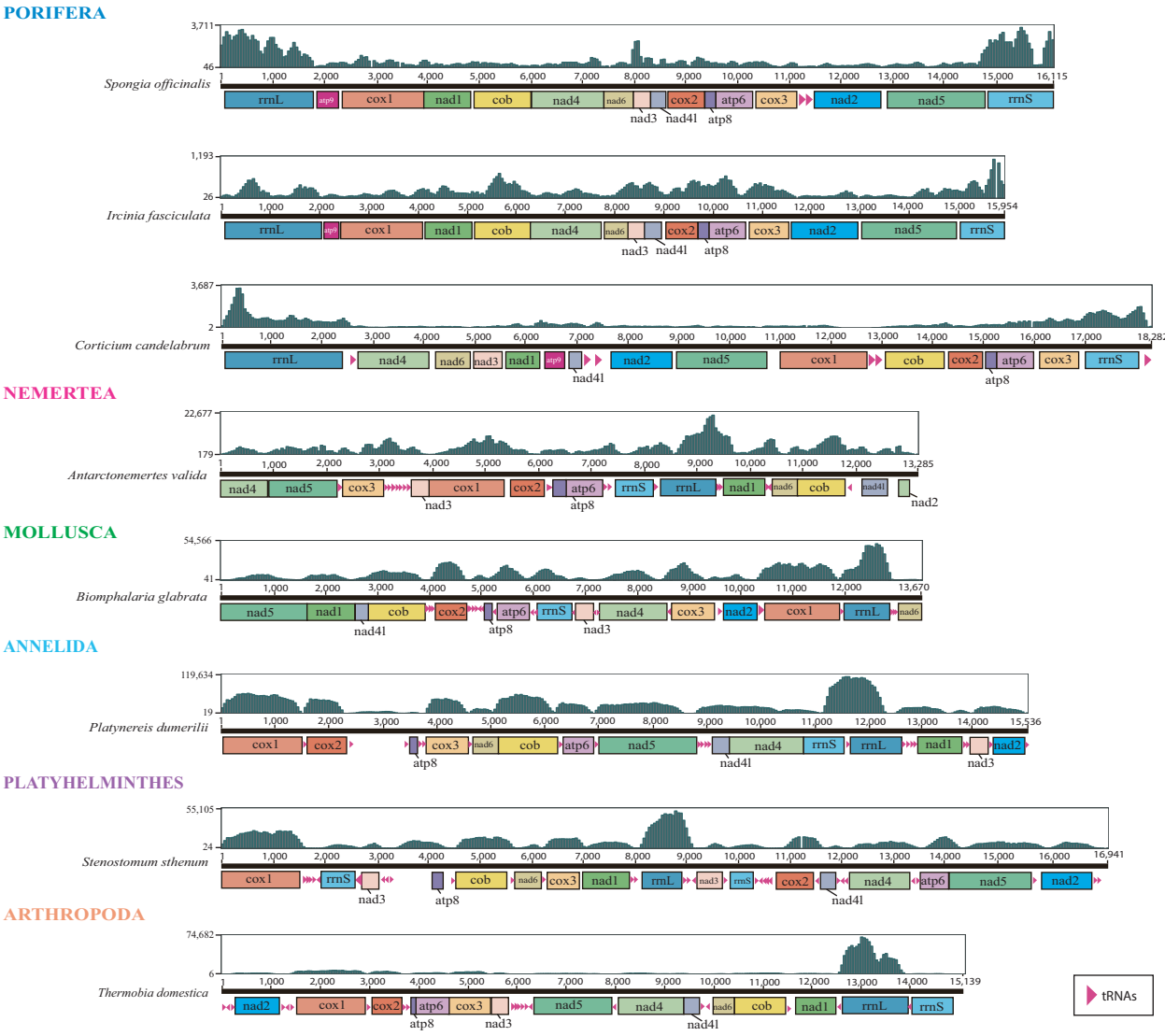
PORIFERA
Spongia officinalis
rrnL atp9 cox1 nad1 cob nad4 nad6 cox2 atp6 cox3 nad2 nad5 rrnS
nad3 nad41 atp8
Ircinia fasciculata
rrnL atp9 cox1 nad1 cob nad4 nad6 cox2 atp6 cox3 nad2 nad5 rrnS
nad3 nad41 atp8
Corticium candelabrum
rrnL nad4 nad6 nad3 nad1 atp9 nad2 nad5 cox1 cob cox2 atp6 cox3 rrnS
nad41 atp8
NEMERTEA
Antarctonemertes valida
nad4 nad5 cox3 cox1 cox2 atp6 rrnS rrnL nad1 nad6 cob nad41 nad2
nad3 atp8
MOLLUSCA
Biomphalaria glabrata
nad5 nad1 cob cox2 atp6 rrnS nad4 cox3 nad2 cox1 rrnL nad6
nad41 atp8 nad3
ANNELIDA
Platynereis dumerilii
cox1 cox2 cox3 nad6 cob atp6 nad5 nad4 rrnS rrnL nad1 nad2
atp8 nad41 nad3
PLATYHELMINTHES
Stenostomum sthenum
cox1 rrnS cob nad6 cox3 nad1 rrnL nad3 rrnS cox2 nad4 atp6 nad5 nad2
nad3 atp8 nad41
ARTHROPODA
Thermobia domestica
nad2 cox1 cox2 atp6 cox3 nad5 nad4 nad6 cob nad1 rrnL rrnS
atp8 nad3 nad41
tRNAs

Fig 2

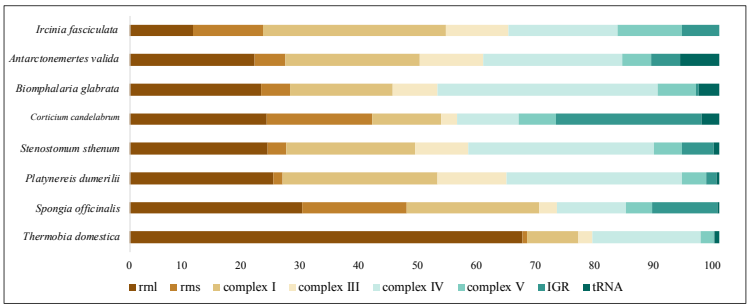
Ircinia fasciculata
Antarctonemertes valida
Biomphalaria glabrata
Corticium candelabrum
Stenostomum sthenum
Platynereis dumerilii
Spongia officinalis
Thermobia domestica
0
10
20
30
40
50
60
70
80
90
100
rrnl
rrns
complex I
complex III
complex IV
complex V
IGR
tRNA

Fig 3