*DNA Research*, 2018, 25(6), 577–586
doi: 10.1093/dnares/dsy026
Advance Access Publication Date: 6 August 2018
Full Paper

OXFORD

Full Paper

# Comparative mitochondrial genomics reveals a possible role of a recent duplication of NADH dehydrogenase subunit 5 in gene regulation

**Runsheng Li[1], Xiaoliang Ren[1], Yu Bi[1], Qiutao Ding[1], Vincy Wing Sze Ho[1] and Zhongying Zhao[1,2,*]**

[1]Department of Biology and [2]State Key Laboratory of Environmental and Biological Analysis, Hong Kong Baptist University, Hong Kong, China

*To whom correspondence should be addressed. Tel. 852-34117058. Fax. 852-34115995. Email: zyzhao@hkbu.edu.hk

Edited by Prof. Kenta Nakai

Received 14 November 2017; Editorial decision 17 July 2018; Accepted 22 July 2018

## Abstract

Mitochondrial genome (mtDNA) carries not only well-conserved protein coding, tRNA and rRNA genes, but also highly variable non-coding regions (NCRs). However, the NCRs show poor conservation across species, making their function and evolution elusive. Identification and functional characterization of NCRs across species would be critical for addressing these questions. To this end, we devised a computational pipeline and performed *de novo* assembly and annotation of mtDNA from 19 *Caenorhabditis* species using next-generation sequencing (NGS) data. The mtDNAs for 14 out of the 19 species are reported for the first time. Comparison of the 19 genomes reveals species-specific sampling of partial displacement-loop (D-loop) sequence as a novel NCR inserted into a unique tRNA cluster, suggesting an important role of the D-loop and the tRNA cluster in shaping NCR evolution. Intriguingly, RNA-Seq analysis suggests that a novel NCR resulting from a recent duplication of NADH dehydrogenase subunit 5 (*ND5*) could be utilized as a 3′ UTR for up-regulation of its upstream gene. The expression analysis shows a species- and sex-specific expression of mitochondrial genes encoded by mtDNA and nucleus, respectively. Our analyses provide important insights into the function and evolution of mitochondrial NCRs and pave the way for further studying the function and evolution of mitochondrial genome.

**Key words:** *Caenorhabditis*, non-coding region (NCR), mitochondrial genome, mitochondrial transcriptome

## 1. Introduction

A typical mitochondrial genome (mtDNA) consists of a circular DNA, which encodes ribosomal RNAs (rRNAs), transfer RNAs (tRNAs) and a subset of proteins that are exclusively used by mitochondrion.[1] Defect in these genes causes severe diseases.[2] These elements are usually located on two separate DNA strands. Unlike eukaryotic nuclear genes that contain exon(s), intron(s), untranslated region (UTR) and independent regulatory regions, mtDNA-encoded genes are usually transcribed into a polycistronic transcript before they are punctuated into individual transcripts.[3] Therefore, these genes do not contain their own regulatory sequence and UTR and carry few intergenic sequences.[4] One of the most sizable non-coding regions (NCRs) is the D-loop that is found across species and has been proposed to function as a start site for DNA replication and

transcription. In addition to the D-loop, species- or population-specific NCRs have also been identified across species.[5] Given the highly compact configuration of mitochondrial genes, determination of the origin and function of these poorly conserved NCRs is critical for understanding the evolution and regulation of mtDNA.

The nematode genus *Caenorhabditis* is an excellent model for studying the NCR biogenesis and function because species- or strain-specific NCRs have already been documented in this genus. In addition to the D-loop, all the currently available *Caenorhabditis* mtDNAs share another NCR, which is approximately 100 bp in length and is located between the NADH dehydrogenase subunit 4 (ND4) and cyclooxygenase (COX1) genes.[6–8] An NCR with similar sequence is also present in the *Pristionchus pacificus* mtDNA, albeit with a reduced length of 51 bp.[9] This NCR was postulated to be another mtDNA replication origin aside from the D-loop.[9,10] With the exception of the two relatively conserved NCRs, other NCRs have also been documented occasionally. For example, in addition to the two relatively conserved NCRs, *C. briggsae* carries another one or two NCR(s) in a strain-dependent manner.[6,7] *C. sinica* also carries an extra NCR.[6] Comparative and functional genomics study of the mitochondria from more related *Caenorhabditis* species is valuable for understanding NCR evolution and function. We referred to the available nematode mtDNA that can and cannot be rendered circular as complete and partial genome, respectively. Partial or complete mtDNA have been reported for five *Caenorhabditis* species (Table 1), including *C. elegans*,[11] *C. briggsae*,[8] *C. nigoni*,[12] *C. tropicalis*[13] and *C. sinica*.[6] Unlike mtDNAs in mammals and other species that encodes 13 proteins on two separate DNA strands,[4] all the existing *Caenorhabditis* mtDNAs encode 12 proteins on a single strand along with two rRNAs (12S and 16S) and 22 tRNAs, all of which are presumably transcribed as a single unit followed by punctuation into individual transcript as those in human.[11] *C. briggsae* mtDNA usually carries four NCRs arranged in a clockwise order, i.e. a D-loop, a pseudo ND5-1 located between ND2 and CYTB, a shared NCR mentioned above and a pseudo-ND5-2. To facilitate the comparison of NCRs across species, we referred to the four as NCR1, NCR2, NCR3 and NCR4, respectively. We named the NCR in other species in the similar way based their location in respective mtDNA regardless of their sequence homology. The NCR2 is present in *C. sinica* but not in *C. nigoni*, the sister species of *C. briggsae*,[7] indicating its fast turnover rate.

Thanks to an effective isolation method, many new *Caenorhabditis* species have recently been recovered in their wild habitats,[14] 20 of which have their genome sequenced.[15] However, complete or partial mtDNA is available only for a handful of species. The availability of genomic data and mRNA-sequencing (mRNA-Seq) data creates an opportunity to generate and annotate mtDNA in these species.

## 2. Materials and methods

### 2.1. Data sources

The NGS genomic sequencing data for the mtDNA assembly were downloaded from NCBI SRA website. The details of each data source were listed in Supplementary Table S1. Also listed in the table are the data sources for the mRNA sequencing from mix-staged worms that were used for nuclear gene annotation here (see below). The mRNA sequencing data for L4/young adult worms were produced by the following three research projects which were used for quantifying abundance of mitochondrial gene transcripts. RNA-Seq data for *C. elegans fog-2 (q71)* mutant pseudo-females/males, *C. briggsae she-1 (v47)* mutant pseudo-females/males, and *C. remanei*, *C. brenneri* and *C. japonica* wild-type males/females were downloaded from GSE41367.[16] RNA-Seq data for *C. elegans* and *P. pacificus* wild-type hermaphrodites/males were downloaded from GSE53144.[17] RNA-Seq data for

**Table 1.** Statistics of the mtDNA sequences in *Caenorhabditis* species and *P. pacificus*

| Species | Strain | Accessiono. | Size (bp) | Status[a] | NCR1[b] | NCR2 | NCR3 | NCR4 |
|---|---|---|---|---|---|---|---|---|
| Newly assembled | | | | | | | | |
| *C. afra* | JU1286 | KY552909 | 14245 | Circular | 664 | 216 | 110 | – |
| *C. angaria* | PS1010 | KY552902 | 13596 | Partial[c] | 264 | – | 96 | – |
| *C. brenneri* | PB2801 | KY552900 | 14212 | Circular | 672 | 200 | 112 | – |
| *C. castelli* | JU1956 | KY552910 | 13655 | Circular | 310 | – | 96 | – |
| *C. doughertyi* | JU1771 | KY552904 | 13719 | Circular | 385 | – | 108 | – |
| *C. japonica* | DF5081 | MF370516 | 13988 | Circular | 643 | – | 109 | – |
| *C. macrosperma* | JU2083 | KY552908 | 13633 | Circular | 298 | – | 106 | – |
| *C. nouraguensis* | JU2079 | KY552907 | 13786 | Circular | 443 | – | 104 | – |
| *C. plicata* | SB355 | KY552901 | 14248 | Circular | 929 | – | 101 | – |
| *C. remanei* | PB4641 | KY552899 | 13610 | Partial | 291 | – | 107 | – |
| C. sp. 38 | JU2809 | KY552911 | 13762 | Circular | 428 | – | 100 | – |
| *C. monodelphis* | JU1667 | KY552903 | 13350 | Partial | 62 | – | 106 | – |
| *C. virilis* | JU1968 | KY552905 | 13661 | Partial | 344 | – | 99 | – |
| *C. wallacei* | JU1898 | KY552906 | 13705 | Circular | 372 | – | 109 | – |
| *C. nigoni* | JU1421 | KP259621 | 13856 | Circular | 518 | – | 109 | – |
| **Existing** | | | | | | | | |
| *C. briggsae* | AF16 | NC_009885 | 14420 | Circular | 548 | 222 | 109 | 324 |
| *C. elegans* | N2 | NC_001328 | 13794 | Circular | 466 | – | 109 | – |
| *C. sinica* | JU727 | EU407780 | 13886 | Circular[d] | 383 | 234 | 109 | – |
| *C. tropicalis* | JU1836 | KM403565 | 13874 | Circular | 410 | 130 | 107 | – |
| *P. pacificus* | PS312 | NC_015245 | 15954 | Circular | 2222 | 425[c] | 51 | – |

[a]Circular/complete mtDNA can be achieved during mtDNA assembly.
[b]Size in bp for NCR1-4.
[c]Located in a different tRNA cluster from NCR2 in *Caenorhabditis* species (Fig. 1C).
[d]Previous mtDNA was partial and was rendered circular by Nanopore sequencing.

*C. briggsae* wild-type hermaphrodites/males and *C. nigoni* females/males were generated in our previous study, GSE76306.[18]

## 2.2. Mitochondrial DNA assembly and annotation

Attempt to mtDNA assembly was first made using various existing mtDNA DNA assembly pipelines, including MITObim, NOVOPlasty and ARC.[19,20] However, the NGS data for *Caenorhabditis* species were generated using different sequencing platforms with various read characteristics (Supplementary Table S1). Only relatively small contigs could be assembled. In addition, the annotation steps were not incorporated into these packages. Consequently, a third party webserver, Dogma[21] and MITOS,[22] for example, is required to generate annotation for mtDNA. Therefore, a computational pipeline customized for *Caenorhabditis* mtDNA assembly and annotation was devised to accommodate all data sources described above.

A pipeline for mtDNA assembly and annotation was designed as shown in Supplementary Fig. S1. First, the average read coverage was roughly calculated through dividing the total number of nucleotides by a genome size of 100 Mb. The algorithm works by automatically fetching mitochondrion-specific sequencing reads from an NGS data archive with BWA-MEM[23] using the mtDNA of a closely related species as a reference index. The fetched reads, together with their pair-end partners, were used to assemble contigs with Spades assembler.[24] The contigs were then filtered by read coverage, only those with a 10-times higher coverage than the average read coverage would be chosen for subsequent analysis. And the filtered contigs were then added to the reference index to be used for the next round of sequence fetching. The process was reiterated till the possible mitochondrial reads in a given data source were exhausted or the filtered contigs between two consecutive rounds were identical. The contigs were rendered circular whenever possible followed by automatic annotations of tRNAs and coding genes using annotated *C. elegans* mtDNA as a reference. As a test, *C. briggsae* NGS reads relevant to mtDNA was initially fetched with *C. elegans* mtDNA as a reference. These steps were used as a training set for pipeline design to optimize the parameters in the pipeline. The source code and the raw data used in this study were deposited in the website: https://github.com/runsheng/mitovar). The pipeline was named as Mitovar (Version 0.99).

To benchmark the Mitovar compared with existing softwares for mtDNA assembly, we tested the NGS data of *Caenorhabditis* species from different sequencing platforms with various read characteristics, including Roche 454, Illumina and ION torrent (Supplementary Table S1). NOVOPlasty and ARC crashed when fed with 454 reads, possibly due to a wide range in read length. Therefore, we only tested these pipelines using *C. brenneri* reads generated with Illumina platform only. Using NGS data derived from four different Illumina sequencing libraries (Supplementary Table S1) as the read sources and *C. briggsae* mtDNA sequence as a reference index, we compared the performances between these pipelines by running the programs with default parameters. As a result, Mitovar (V0.99) produced a single contig with a size of 14,212 bp, which was essentially a complete mtDNA; whereas NOVOPlasty (V1.2.3) and ARC (V1.1.3) produced a single contig with a size of 3,870 and 4,959 bp, respectively. Mitobim (V1.9) produced three separated contigs with a size of 1,267, 1,523 and 1,704 bp, respectively. The details of the outs can be found in the link below (https://github.com/Runsheng/mitovar/blob/master/data/benchmark_cbre.fasta), indicating the superiority of the Mitovar over the other pipelines in handling the *Caenorhabditis* mtDNAs.

## 2.3. Phylogenetic inference

For generation of phylogenetic tree, the amino acid sequences of 12 mtDNA-encoded proteins from each species were concatenated with those from *P. pacificus* as an outgroup, protein sequence alignment and maximum likelihood (ML) tree generation with 1000 bootstraps were performed in ETE3 package,[25] using the workflow 'standard_raxml_bootstrap' and 'ptree_raxml_all'.[26] The alignment files in 'fasta' format and the resulting tree file in 'newick' format has been deposited in the website: https://github.com/Runsheng/mitovar/tree/master/data.

## 2.4. Orthology establishment for nucleus-encoded proteins present in mitochondria between human and *Caenorhabditis* species

The *C. elegans* nucleus-encoded mitochondrial genes were defined by the following steps. Human mitochondrial genes were derived from MitoCarta2.0.[27] Their *C. elegans* orthologs were derived from Wormbase (WS254).[28] To identify *C. elegans*'s orthologs in other *Caenorhabditis* species, *C. elegans* mitochondrial genes were used as a query against EggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) database (v4.5), from which ortholog tables for *C. briggsae*, *C. remanei*, *C. brenneri*, *C. japonica* and *P. pacificus* were derived.[29] The NOGs were further annotated by removing the 12 mtDNA encoded terms, and marking all genes directly related to oxidative phosphorylation complexes. The resulting NOGs were listed in Supplementary Dataset S1. Nuclear genes from other *Caenorhabditis* species were assembled using Trinity V2.2.[30] They were used to fetch corresponding mitochondrial terms in the EggNOG database using 'hmmsearch'.[29] The orthologs for the 14 *Caenorhabditis* species were listed in Supplementary Dataset S2.

## 2.5. Quantification of mRNA sequencing data for mtDNA-encoded mitochondrial genes

For quantifying the abundance of gene transcripts, mRNA sequencing reads derived from young adults were mapped to respective reference mtDNA sequences using BWA-MEM with default parameters.[23] Reads with multiple hits or a lower than 20 of MAPping Quality (MAPQ) value were discarded using Samtools (V1.1). Total number of reads mapped to each gene or NCR were counted using HTseq-count V0.6.[31] To generate expression track with RNA-Seq data, count of reads uniquely mapped to mtDNA was normalized against the total number of reads in a given sample. The normalized read counts were used to compute read coverage per nucleotide over the mtDNA. To compare relative expression levels across species using heatmap, Reads Per Kilobase of transcript per Million mapped reads (RPKM) were calculated for all the samples of each species by averaging the read counts from all samples. To make expression levels comparable across species, the relative expression level was calculated for each gene and NCR by normalization against the average expression level of the 12 protein-coding genes with each species. The resulting expression data matrix was used in the downstream analysis. The matrix was used as an input for average linkage clustering analysis with 'hclust' function in R Version 3.1.1[32] with default parameters.

## 2.6. Quantification of mRNA sequencing data for nucleus-encoded mitochondrial genes

Protein sequences encoded by 1,158 human nuclear genes were curated in mitochondria MitoCarta 2.0.[27] Their *C. elegans* orthologs were retrieved from Wormbase (WS254).[28] The orthologs of the *C. elegans* mitochondrial proteins in other nematode species were obtained by querying the individual *C. elegans* mitochondrial genes in EggNOG database 4.5.[29] RPKM for each NOG terms was calculated using 'Salmon' V8.1 with default parameters.[33] For all the nuclear genes in a single NOG term, transcripts from all isoforms were

merged for RPKM calculation. To make the expression levels comparable across species, gene expression levels for each NOG term were normalized in each species by dividing the average RPKM of all genes within each sample to give rise to the relative expression level of each NOG term. The resulting expression values were used in downstream analysis.[34]

### 2.7. Closing the gap in *C. sinica* D-loop region using nanopore sequencing

*C. sinica* D-loop region was first amplified by nested PCR using the genomic DNA extracted from the strain JU727 as a template. The forward and reverse primer sequences for the first round of PCR were TCGGTCTTCCTCACTTTT and TCTTAGCAACCCAAATGC, respectively. The forward and reverse primer sequences for the second round of PCR were TAAGGGTTATACCTTATTTTTAAG and ATTAAACTATTTACTGAAAACCA, respectively. Because the resulting PCR products still contained multiple bands with subtle difference in size even after multiple attempts through tuning PCR parameter, which is consistent with a previous study that also failed in recovering the full-length of the D-loop,[7] it is not possible to resolve the D-loop sequence by Sanger sequencing. Nanopore sequencing technique was used to directly sequence the PCR products. Specifically, the PCR products were size selected by gel purification for multiple bands approximately 490 bp in size, which were used for making sequencing library using 1 D ligation gDNA kit (SQK-LSK108). The library was sequenced on a Nanopore MinION sequencer with R9.4 flow cell (FLO-MIN106) according to manufacturer's descriptions.

Given the complications associated with the D-loop amplification, we reasoned that there could be multiple isoforms of D-loop in the *C. sinica* mtDNA. After self-correction for read errors and haplotyping using Nanocorr (https://github.com/jgurtowski/nanocorr), we chose the contigs with the highest read coverage for building consensus sequence, which may represent only one of the possible D-loop isoforms plus its neighbouring tRNA gene. The Nanopore sequencing data and the resulting D-loop sequence were deposited in SRA under BioProject PRJNA390876 with accession number of SRR5712476.

## 3. Results

### 3.1. *De novo* assembly of mitochondrial genome for 14 *Caenorhabditis* species using Mitovar

We first performed *de novo* genome assembly using Mitovar for four *Caenorhabditis* species whose complete or partial mtDNAs are available except *C. elegans* (Fig. 1A and Table 1). The mtDNA sequences

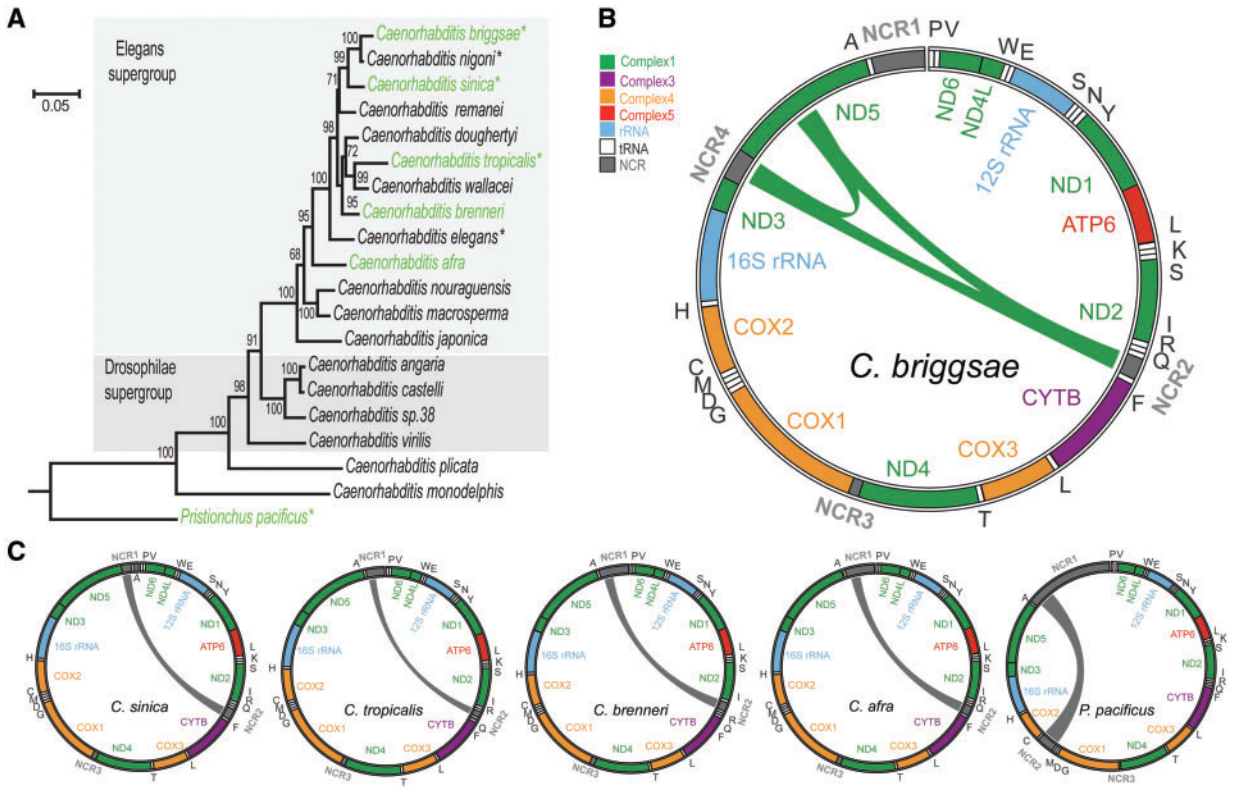


**Figure 1.** Species-specific sampling of *ND5* or non-conserved part of D-loop (also called NCR-1) as a novel NCR2. (A) A phylogenetic tree of *Caenorhabditis* species inferred from the sequences of 12 mitochondrial proteins. Scale of branch length is indicated on the left. Tree was constructed with ML methods implemented in raxML using the concatenated 12 protein sequences from 19 *Caenorhabditis* species with *P. pacifics* as an outgroup. The bootstrap support values in percent from 1000 repeats are indicated on branch node. Five species that contain an NCR2 are highlighted in green. Elegans and Drosophilae supergroup are differentially shaded. The mtDNAs that were previously assembled are indicated with '*'. (B) Diagrams showing the origins of NCR2 or 4 (green ribbons) in the mtDNA of *C. briggsae*, respectively. Sampled sequence size and relative position are shown in scale. (C) Diagrams showing the origins of NCR2 (grey ribbons) in the mtDNA of *C. sinica*, *C. tropicalis*, *C.* brenneri, *C. afra* and *P. pacificus*. Mitochondrial genes are differentially colour coded based on the MRCs they are affiliated with as indicated on the top left. Consensus size in kb is indicated starting from the Phenylalanine tRNA gene (P). NCR, NCR. Note, only *C. briggsae* contains an NCR4 located in the boundary between *ND3* and *ND5*. All six species carry an NCR2 located in a single tRNA gene cluster. The NCR1 and 3 are shared in all *Caenorhabditis* species.

we produced were all identical to the existing ones, demonstrating that Mitovar was effective in retrieving and annotating mtDNA. We therefore performed subsequent sequence analysis using the publicly available mtDNAs for the five species and one outgroup species, *P. pacificus*, except for *C. sinica*, for which we used the complete mtDNA generated here for subsequent analysis (Table 1). We next performed *de novo* genome assembly for the remaining 15 *Caenorhabditis* species whose mtDNAs were not available, although the NGS data of their genomic DNAs were available (Table 1). We successfully generated genome assemblies for 14 of the 15 species using the NGS reads from various sequencing platforms, including Roche 454, Illumina Genome Analyzer II, MiSeq and HiSeq (Supplementary Table S1). Ten out of the 14 mtDNAs were complete as judged by their circular status as well as the presence of all genes expected from *C. elegans* mtDNA, whereas the remaining four mtDNAs seemed to be incomplete because they could not be rendered circular mainly due to the incompleteness in their D-loop regions (Table 1). The D-loop region commonly carries many simple tandem repeats that are rich in AT, which are problematic for retrieving complete sequence of the D-loop, leading to a partial mtDNA in a linear format (Table 1). This is the case of *C. sinica*, whose existing mtDNA is incomplete because of its partial D-loop sequence.[6] We successfully rendered the mtDNA circular using Nanopore sequencing reads (see Materials and methods). We failed in genome assembly for one species, *C. guadeloupensis*, previously known as *C.* sp. 20,[35] presumably due to the insufficient reads (coverage lower than 2×) that could be fetched from the deposited dataset. The size selection steps during their preparation of sequencing library probably eliminated most of the mitochondrial reads.

We annotated the mtDNAs for all 14 species using Mitovar. Unlike human mtDNA-encoded genes that are located on both DNA strands, all the 19 *Caenorhabditis* mtDNAs carry the same number of genes, including 12 genes encoding part of the 4 mitochondrial respiratory chain (MRC) complexes, 2 rRNA (12S and 16S) and 22 tRNA genes that are located on a single DNA strand with identical synteny (Fig. 1B and C). We defined the start position of the circular mtDNA as the start of the phenylalanine tRNA gene (tRNA-P) located immediately downstream of the D-loop region, which is similar to that used in human.[4] To test the use of annotated protein sequences in phylogenetic inference, we built a ML phylogenetic tree of *Caenorhabditis* species using the concatenated amino acid sequences from the 12 proteins in each species (Fig. 1A and Supplementary Fig. S2) as described.[25] The topology of the tree generated here was mostly in agreement with that of a previous phylogenetic tree constructed using the DNA sequences from 11 nuclear genes.[14] Slight discrepancy in node branching within the 'Elegans supergroup' was likely due to the protein sequences used here but the DNA sequences used previously.

### 3.2. Species-specific sampling of D-loop as a novel NCR in a tRNA cluster of mtDNA in *Caenorhabditis* species

The availability of mtDNAs for 19 *Caenorhabditis* species allowed us to examine the evolutionary dynamics and origin of the NCRs across species. D-loop is present in all *Caenorhabditis* species (Figs 1C and 2A). The lengths of the intact D-loop/NCR1 vary substantially among species (Table 1). The most variable regions of the D-loops are located at their 5′ regions. The relatively conserved regions of the D-loops are located at their 3′ regions, which are rich in AT repeats (Fig. 2A). NCR3 located between *ND4* and *COX1* is also present in all *Caenorhabditis* species, which is highly conserved in size (96–112 bp) and in nucleotide sequence (Fig. 2B). In addition to *C. briggsae* and *C. sinica*, we detected three other *Caenorhabditis* species that carried an NCR2 in the same tRNA cluster as that in *C. briggsae*, including *C. afra*, *C. brenneri* and *C. tropicalis* (Fig. 1C and Table 1). The NCR2s from *C. afra*, *C. sinica* and *C. briggsae* are all located between genes tRNA-Q and tRNA-F, whereas the NCR2s from *C. brenneri* and *C. tropicalis* are located between genes tRNA-I and tRNA-R, or between tRNA-Q and tRNA-R, respectively. The sequences of NCR2s are barely alignable between species (Fig. 2C). Apparently, NCR4 is unique to *C. briggsae*.

To trace sequence origin of the NCR2s, we performed homology search against both nuclear and mitochondrial sequences. Strikingly, unlike *C. briggsae* NCR2, which shows the highest sequence homolog to its *ND5*[6], the NCR2s in all the other species show the highest sequence homology to their respective D-loop regions (Fig. 1 and Supplementary Fig. S3), suggesting their origins from the D-loop. A previous study proposed that *C. sinica* NCR2 was originated from *ND5* based on its modest sequence homology with its incomplete D-loop sequence.[7] To confirm this, we performed sequence homology search against NCBI non-redundant DNA sequence database and against the complete D-loop sequence we generated. Although no hit was found in the database, apparent sequence homology between NCR2 and D-loop was observed in *C. sinica* (Supplementary Fig. S3D) with a much higher similarity score than that between *C. sinica* NCR2 and *ND5*. The sequence homology in *C. sinica* is similar to that in *C. afra*, *C. brenneri* and *C. tropicalis*, suggesting its origin from its D-loop rather than from its *ND5*. Although *C. briggsae* is more closely related to *C. nigoni* than to any other *Caenorhabditis* species, the two species do not share NCR number and origin, indicating that the NCR2 and NCR4 are generated after divergence from their common ancestor. Notably, like the NCR2s in all *Caenorhabditis* species which are located in the tRNA cluster between *ND2* and cytochrome b (*CYTB*) genes (Fig. 1B and C), a *P. pacificus* NCR was also located in a tRNA cluster between *COX1* and *COX2*. Sequence alignment shows that the NCR was also originated from the 5′ part of its D-loop (Fig. 1C and Supplementary Fig. S3F). Taken together, our data show that most novel NCRs arise frequently by sampling the D-loop with the exception of *C. briggsae* NCR2 and NCR4. The results also show that tRNA cluster is a hotspot for introduction of novel NCR. It is not clear why *C. briggsae* NCR2 adopts an unusual origin compared with the NCR2 origins in other *Caenorhabditis* species. It is also enigmatic why only a subset of *Caenorhabditis* species produces a NCR2. In summary, we demonstrated that most of the NCR2s were generated by species-specific sampling of the partial D-loop followed by insertion into a unique tRNA cluster. Whether the NCR2s are strain-specific awaits further investigation. Given the species-specific origin of NCRs, a fascinating question is whether the highly variable NCRs are functional *in vivo*.

### 3.3. NCR4 is highly expressed, which is associated with up-regulation of ND3 in *C. briggsae*

To explore the function of these highly variable NCRs, we quantified the expression of all mtDNA-encoded genes and NCRs using the existing RNA-Seq data, mostly derived from mRNA. To make the expression levels comparable between species, we only used the expression data derived from young adult stage from which we produced the relative expression level within each sample (see Materials and methods). It is intriguing that although most NCRs show no or modest expression relative to coding genes, the NCR2 and NCR4 in

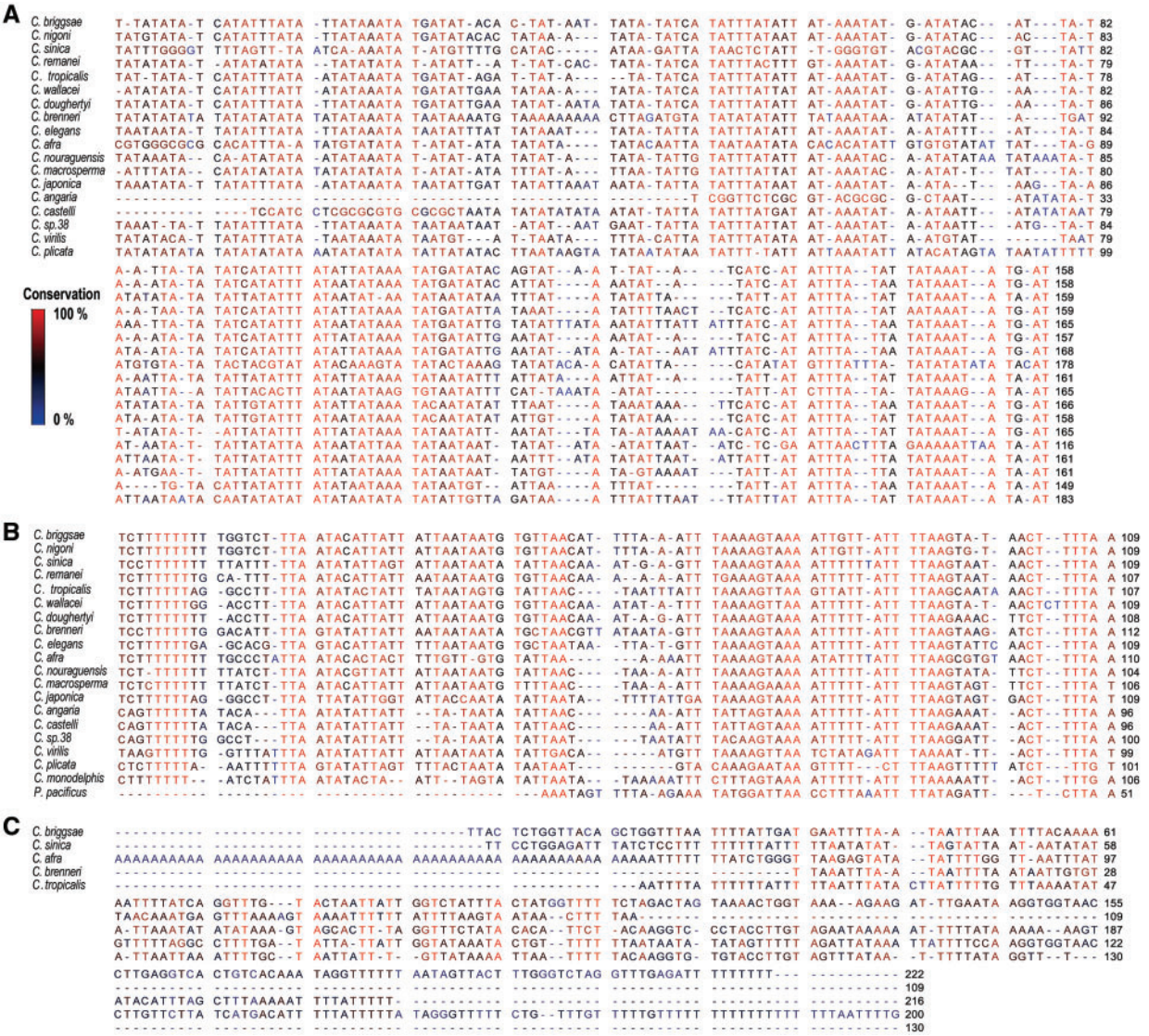


**Figure 2.** Sequence conservation of NCRs in *Caenorhabditis* species. (A) Alignment of NCR1 sequences from the conserved 3′ part of NCR1 in 18 *Caenorhabditis* species whose NCR1 sequence is over 200 bps in length. (B) Alignment of NCR3 full-length sequences from 19 *Caenorhabditis* species and an NCR from outgroup species *P. pacificus*. (C) Sequence alignment of NCR2s between five *Caenorhabditis* species. Note the sequences are barely alignable.

*C. briggsae* are highly expressed (Fig. 3A and B). Given the unique origin of the two genes from *ND5* as opposed to the NCR2s in other four species which were originated from a NCR, i.e. D-loop, we asked whether the elevated expression of the two *C. briggsae* NCRs served to provide an extra dosage of the duplicated *ND5* or played a regulatory role for their neighbouring genes. To this end, we performed hierarchical clustering of the relative expression levels of all protein-coding genes encoded by mtDNA using existing RNA-Seq data in the nine nematode strains, including seven nematode species (Supplementary Datasets S1 and S2). Surprisingly, the results showed that *C. briggsae ND3* expression level was roughly two orders of magnitude higher than that in the remaining six species (Fig. 3B). The NCR4 is duplicated from the mid-part of *ND5* in *C. briggsae* (Fig. 1B) with at least two independent single-nucleotide deletions, leading to no open reading frame (ORF) that was continuous with its neighbouring genes, i.e. *ND3* and *ND5* (Supplementary Fig. S4). In addition, the duplicated fragment is flanked by the sequences showing poor alignment with *ND5*, indicating that it is unlikely that the expression of NCR4 serves to increase the *ND5* dosage, consistent with previous observations.[6,7,36] Given a significantly elevated expression of *ND3* in *C. briggsae* versus its counterpart in other species and that the NCR4 is located immediately downstream of the gene, we reasoned that the NCR4 might serve as a 3′ UTR for *ND3* to up-regulate its expression. To examine this possibility, we counted the number of RNA-Seq reads with poly-A tail immediately after the end of *ND3* or NCR4. To our surprise, in contrast to other protein-coding neighbouring genes, for example, *ND6* and *ND4L*, whose transcripts are individually polyadenylated, most of the RNA-Seq reads were continuous from *ND3* to NCR4, hereafter referred to as continuous read, indicating that the two genes were processed into a single transcript post-transcriptionally. Notably, only 12 of 2,287 (0.52%) continuous reads were polyadenylated immediately at the end of *ND3*, whereas 120 of the 129 (93.0%) reads mapped to the 3′ end of NCR4 were polyadenylated immediately at the end of

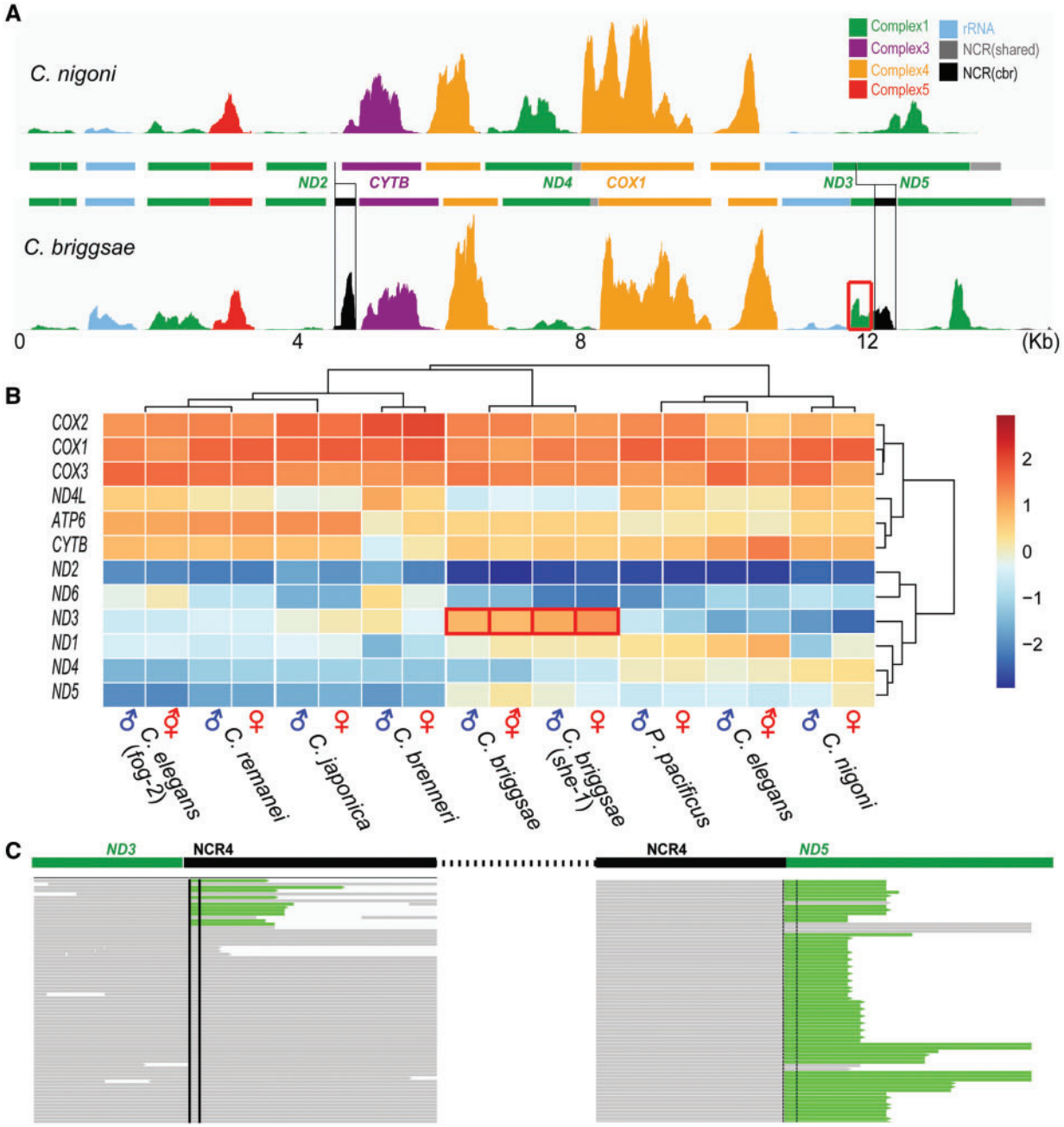


**Figure 3.** Association of the elevated expression of a recent *ND5* duplication, i.e. NCR4 with up-regulation of *ND3* expression in *C. briggsae*. (A) RNA-Seq tracks showing the read coverage (vertical axis) aggregated from both sexes of *C. nigoni* (top) and *C. briggsae* (bottom) along their respective reference genomes (two horizontal bars in the middle). Peaks and genes are colour coded based on their origin of respective respiratory chain complex and two *C. briggsae* unique NCRs are coloured in black (highlighted in box). Note the up-regulated expression of *ND3* in *C. briggsae* versus that in its closest relative *C. nigoni* is highlighted with a red box. *ND3* expression is significantly lower in *C. nigoni*. (B) Heatmap showing the normalized expression of mtDNA-encoded proteins with names indicated on the left. *C. briggsae*-specific higher expression of *ND3* is highlighted in red boxes. Note that relatively low expression of various ND proteins versus that of other mitochondrial proteins in *Caenorhabditis* species. (C) Sequence alignment for RNA-Seq reads against a mitochondrial genomic region spanning *ND3,* NCR4 and *ND5* (shown as thick bars on the top). Alignments of RNA-Seq reads are shown at the bottom with poly-A sequences coloured in green. Note that the most reads are continuous between the end of *ND3* and NCR4 followed by poly-A tailing (shown in green on the right); whereas only a small portion of *ND3* reads are poly-A tailed at its own end (green on the left). The NCR4 contain multiple stop codons which was annotated a pseudogene of *ND5* (not shown).

NCR4 (Fig. 3C), indicating that the elevated expression of *ND3* was mainly contributed by the continuous reads spanning both *ND3* and NCR4. In addition, the transcript of the NCR4 was predicted to form a stem-loop structure (Supplementary Fig. S5), typical of a bacterial 3′ UTR.[37] Taken together, we demonstrate that the *C. briggsae* NCR4 possibly functions as a 3′ UTR for *ND3* to up-regulate its expression. *C. briggsae* NCR2 was also highly expressed with a poly-A tail, but only as an independent transcript (Fig. 3A). It remains unclear why the NCR2 shows a relatively higher expression level in *C. briggsae* than NCR2s in other *Caenorhabditis* species (Fig. 4).

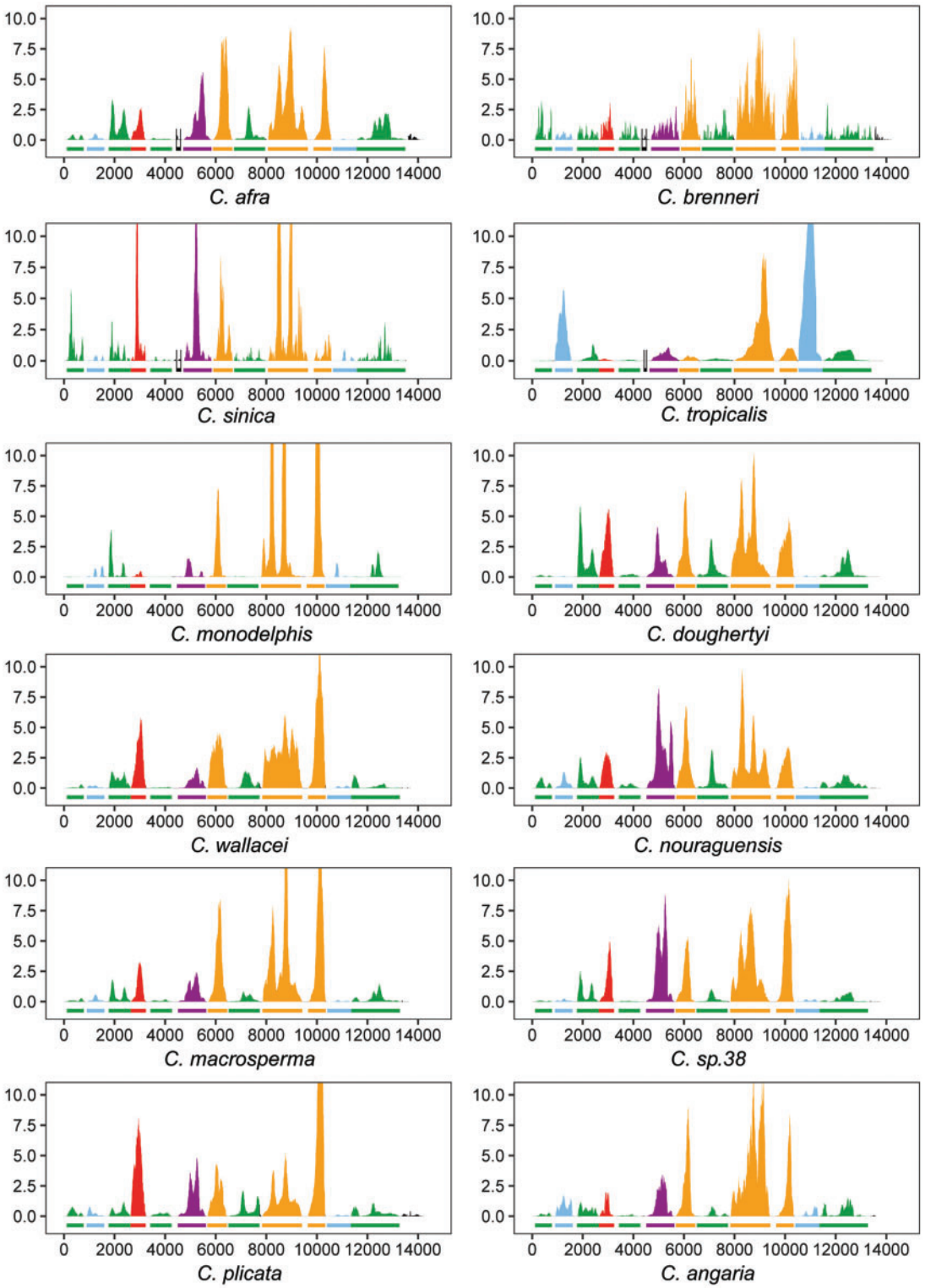


**Figure 4.** RNA-Seq tracks showing the coverage of reads (vertical axis) derived from 12 *Caenorhabditis* species along their respective reference genomes (horizontal bars). Peaks are colour-coded as indicated in Fig. 3. Note the apparent expression of NCR2 (highlighted in black line) in *Caenorhabditis* species carrying an NCR2 except *C. tropicalis*.

For nucleus-encoded mitochondrial genes, we observed an overall dichotomy of gene expression between male and female in *Caenorhabditis* species (Supplementary Fig. S6), suggesting a differential metabolic loading between the males and females in these species.

## 4. Discussion

Function and evolution of mtDNA have long been a hot topic in phylogenetic studies[34]. Despite the highly conserved coding elements

within mtDNAs, its NCRs undergo fast evolution. The function of the highly variable NCRs remains mysterious. Comparative genomics is the method of choice for studying mtDNA evolution and function. With the sharply decreased sequencing costs, sequencing nuclear genome becomes affordable for nearly all species even by an individual research laboratory. Despite the quality of nuclear genome assemblies varies substantially from one another, especially for those that are generated exclusively by NGS data, these sequencing data provide an opportunity to generate a mtDNA with relatively high quality. This is because the mtDNAs usually co-exist with nuclear genome in NGS data. Here we presented mtDNAs from 19 *Caenorhabditis* species mostly by mining the existing NGS data, 14 of which were reported for the first time. Those mtDNAs form an invaluable resource for further functional and evolutionary studies.[38,39,40]

Discovery of independent duplication of D-loop as a novel NCR in multiple *Caenorhabditis* species has possible implications in mtDNA replication and transcription. DNA replication in human adopts a model of strand displacement-loop (D-loop) or RNA incorporation throughout the lagging strand (RITOLS)[41] starting from the D-loop region, although the exact replication mechanism remains heavily debated.[42] Interestingly, mtDNA replication in *C. elegans* was proposed to adopt a rolling cycle mode, in which the nascent mtDNA formed branched circular lariat structures with concatenated tails that are ultimately resolved into monomeric circles.[43] How these polymeric fragments are resolved into a circular one remains an open question. The D-loop serves as both a DNA replication origin and a transcription start site in human.[42,44] In addition, mtDNA replication starts with a short RNA transcript as a primer that is also derived from D-loop in human, indicating a coupling between DNA replication and transcription.[42] Notably, the D-loop sequences in *Caenorhabditis* species contain numerous tandem and inverse repeats (Supplementary Fig. S7). Whether the D-loop-derived NCR2s function as an alternative DNA replication origin or a transcription start site awaits further investigation.

It is well-established that 3′ UTR plays a critical role in stabilizing nuclear mRNA transcript. These UTRs are usually at least tens of bps in length that are bound by RNA-binding proteins. Few mtDNA-encoded genes carry such a long 3′ UTR. There are only a short stretch of RNA sequence with handful of bps in length that are present between stop codon and poly-A signal in most mtDNA encoded genes, for example, in human[44,45] and rodents.[46] It is unclear whether these short sequences work as an effective 3′ UTR as that in nuclear genes or not. Our observation that *C. briggsae* ND3 mRNAs can be alternatively poly-A tailed either immediately after its stop codon or at the end of NCR4 is the first of its kind to our knowledge. This is based on the observation that the majority of ND3 transcripts contain the sequences derived from NCR4, which is associated with its elevated expression in *C. briggsae* compared with its counterpart in other species (Fig. 3). This raises the possibility the newly duplicated NCR4 could possibly serve as an effective 3′ UTR to boost the expression of its host gene. Functional consequences of the increased expression of ND3 remains unclear.

Incorporation of NCR2 specifically into a specific tRNA cluster but not into other position suggests that the cluster is a hot spot for novel NCR insertion. We speculate that this cluster may serve as a punctuation site for resolving the branched-circular lariat structure into a circular format during mtDNA replication, but further evidence is needed to test this hypothesis. Availability of mtDNAs in the 19 *Caenorhabditis* species provides an opportunity for further addressing the evolution, DNA replication and transcription regulation of mtDNA. It will also facilitate investigation into the mechanism of uniparental inheritance of mitochondria as evidenced in both *C. elegans*[47] and *C. briggsae*.[48]

## Acknowledgements

This work was supported by General Research Fund (HKBU263512, HKBU12103314, HKBU12123716) and Collaborative Research Fund (HKBU5/CRF/11G) from Hong Kong Research Grant Council to Z.Z. and by National Natural Science Foundation of China (31701089) to R.L. R.L. designed the Mitovar pipeline and wrote its code, performed *de novo* assembly of all mitochondrial genomes. X.R. performed Nanopore sequencing. Y.B., Q.D. and V.H. contributed to data extraction and polishing from public databases. Z.Z. conceived and coordinated the study and wrote the manuscript. The authors declare that they have no conflicts of interest with the contents of this article. We thank Mr. Chung Wai Shing for logistic support and the members of Zhao's lab for helpful discussion and comments.

## Accession numbers

All the mtDNA sequences generated in this study were deposited in Genbank with accession numbers listed in Table 1. Nanopore sequencing data were deposited in BioProject with accession no.: PRJNA390876. Source code for Mitovar and all the raw data used by Mitovar for mtDNA assembly were deposited in website: https://github.com/runsheng/mitovar. Details information on the source data used were also listed in Supplementary Table S1.

## Conflict of interest

None declared.

## Supplementary data

Supplementary data are available at *DNARES* online.

## References

1. Anderson, S., Bankier, A. T., Barrell, B. G., et al. 1981, Sequence and organization of the human mitochondrial genome, *Nature*, **290**, 457.
2. Johnson, S. C., Yanos, M. E., Kayser, E. B., et al. 2013, mTOR inhibition alleviates mitochondrial disease in a mouse model of Leigh syndrome, *Science*, **342**, 1524–8.
3. Ojala, D., Montoya, J. and Attardi, G. 1981, tRNA punctuation model of RNA processing in human mitochondria, *Nature*, **290**, 470–4.
4. Lemire, B. 2005, Mitochondrial genetics. *WormBook: the online review of C. elegans biology*, The C. elegans Research Community, WormBook, doi/10.1895/wormbook.1.7.1, http://www.wormbook.org.
5. Nicholls, T. J. and Minczuk, M. 2014, In D-loop: 40 years of mitochondrial 7S DNA, *Exp. Gerontol.*, **56**, 175–81.
6. Howe, D. K. and Denver, D. R. 2008, Muller's Ratchet and compensatory mutation in Caenorhabditis briggsae mitochondrial genome evolution, *BMC Evol. Biol.*, **8**, 62.
7. Raboin, M. J., Timko, A. F., Howe, D. K., Felix, M. A. and Denver, D. R. 2010, Evolution of Caenorhabditis mitochondrial genome pseudogenes and Caenorhabditis briggsae natural isolates, *Mol. Biol. Evol.*, **27**, 96.
8. Stein, L. D., Bao, Z., Blasiar, D., et al. 2003, The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics, *PLoS Biol.*, **1**, E45.
9. Molnar, R. I., Bartelmes, G., Dinkelacker, I., Witte, H. and Sommer, R. J. 2011, Mutation rates and intraspecific divergence of the mitochondrial genome of Pristionchus pacificus, *Mol. Biol. Evol.*, **28**, 2317–26.
10. Bratic, I., Hench, J. and Trifunovic, A. 2010, Caenorhabditis elegans as a model system for mtDNA replication defects, *Methods*, **51**, 437–43.

11. Okimoto, R., Macfarlane, J., Clary, D. and Wolstenholme, D. 1992, The mitochondrial genomes of two nematodes, Caenorhabditis elegans and Ascaris suum, *Genetics*, **130**, 471–98.
12. Li, R., Ren, X., Bi, Y. and Zhao, Z. 2015, Mitochondrial genome of Caenorhabditis nigoni (Rhabditida: habditidae), *Mitochond. DNA*, **27**, 3107–3108.
13. Yang, S. S., Li, S. and Wang, G. X. 2014, The complete mitochondrial genome of Caenorhabditis tropicalis n. sp. (Rhabditida: Rhabditidae), *Mitochond. DNA*, **27**, 1763–1764.
14. Kiontke, K. C., Felix, M. A., Ailion, M., et al. 2011, A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits, *BMC Evol. Biol.*, **11**, 339.
15. Kumar, S., Schiffer, P. H. and Blaxter, M. 2012, 959 Nematode genomes: a semantic wiki for coordinating sequencing projects, *Nucleic Acids Res.*, **40**, D1295–300.
16. Thomas, C. G., Li, R., Smith, H. E., Woodruff, G. C., Oliver, B. and Haag, E. S. 2012, Simplification and desexualization of gene expression in self-fertile nematodes, *Curr. Biol.*, **22**, 2167–72.
17. Albritton, S. E., Kranz, A. L., Rao, P., Kramer, M., Dieterich, C. and Ercan, S. 2014, Sex-biased gene expression and evolution of the x chromosome in nematodes, *Genetics*, **197**, 865–83.
18. Li, R., Ren, X., Bi, Y., et al. 2016, Specific downregulation of spermatogenesis genes targeted by 22G RNAs in hybrid sterile males associated with an X-Chromosome introgression, *Genome Res.*, gr. 204479.204116.
19. Dierckxsens, N., Mardulyn, P. and Smits, G. 2017, NOVOPlasty: de novo assembly of organelle genomes from whole genome data, *Nucleic Acids Res.*, **45**, e18.
20. Hahn, C., Bachmann, L. and Chevreux, B. 2013, Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach, *Nucleic Acids Res.*, **41**, e129.
21. Wyman, S. K., Jansen, R. K. and Boore, J. L. 2004, Automatic annotation of organellar genomes with DOGMA, *Bioinformatics*, **20**, 3252–3255.
22. Bernt, M., Donath, A., Jühling, F., et al. 2013, MITOS: improved de novo metazoan mitochondrial genome annotation, *Mol. Phylogenet Evol.*, **69**, 313–319.
23. Li, H. 2013, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv Preprint arXiv*, **1303**, 3997.
24. Bankevich, A., Nurk, S., Antipov, D., et al. 2012, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J Comput. Biol.*, **19**, 455–477.
25. Huerta-Cepas, J., Serra, F. and Bork, P. 2016, ETE 3: reconstruction, analysis, and visualization of phylogenomic data, *Mol. Biol. Evol.*, **33**, 1635–1638.
26. Stamatakis, A. 2014, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics*, **30**, 1312–1313.
27. Calvo, S. E., Clauser, K. R. and Mootha, V. K. 2015, MitoCarta2. 0: an updated inventory of mammalian mitochondrial proteins, *Nucleic Acids Res.*, gkv1003.
28. Howe, K. L., Bolt, B. J., Cain, S., et al. 2015, WormBase 2016: expanding to enable helminth genomic research, *Nucleic Acids Res*, **44**, D774–780.
29. Huerta-Cepas, J., Szklarczyk, D., Forslund, K., et al. 2015, eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences, *Nucleic Acids Res*, **44**, D286–293.
30. Haas, B. J., Papanicolaou, A., Yassour, M., et al. 2013, De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity, *Nat. Protoc.*, **8**, 1494.
31. Anders, S., Pyl, P. T. and Huber, W. 2015, HTSeq—a Python framework to work with high-throughput sequencing data, *Bioinformatics*, **31**, 166–169.
32. R. Core Team, 2016, R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.
33. Patro, R., Duggal, G. and Kingsford, C. 2015, Salmon: accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment. *bioRxiv*, 021592.
34. Yang, Z. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol.Biol. Evol.*, **24**, 1586–1591.
35. Felix, M.-A., Braendle, C. and Cutter, A. D. 2014, A streamlined system for species diagnosis in Caenorhabditis (Nematoda: Rhabditidae) with name designations for 15 distinct biological species, *PLoS One*, **9**, e94723.
36. Phillips, W. S., Coleman-Hulbert, A. L., Weiss, E. S., et al. 2015, Selfish mitochondrial DNA proliferates and diversifies in small, but not large, experimental populations of caenorhabditis briggsae. *Genome Biol. Evol.*, **7**, 2023–2037.
37. Hill, K. E., Lloyd, R. S. and Burk, R. F. 1993, Conserved nucleotide sequences in the open reading frame and 3'untranslated region of selenoprotein P mRNA, *Proc. Natl Acad. Sci. U.S.A.*, **90**, 537–541.
38. Havird, J. C. and Sloan, D. B. 2016, The roles of mutation, selection, and expression in determining relative rates of evolution in mitochondrial versus nuclear genomes, *Mol. Biol. Evol.*, **33**, 3042–3053.
39. Houtkooper, R. H., Mouchiroud, L., Ryu, D., et al. 2013, Mitonuclear protein imbalance as a conserved longevity mechanism, *Nature*, **497**, 451–457.
40. Nabholz, B., Ellegren, H. and Wolf, J. B. 2013, High levels of gene expression explain the strong evolutionary constraint of mitochondrial protein-coding genes, *Mol. Biol. Evol.*, **30**, 272–284.
41. Yasukawa, T., Reyes, A., Cluett, T. J., et al. 2006, Replication of vertebrate mitochondrial DNA entails transient ribonucleotide incorporation throughout the lagging strand, *Embo J.*, **25**, 5358–5371.
42. Agaronyan, K., Morozov, Y. I., Anikin, M. and Temiakov, D. 2015, Replication-transcription switch in human mitochondria, *Science*, **347**, 548–551.
43. Lewis, S. C., Joers, P., Willcox, S., Griffith, J. D., Jacobs, H. T. and Hyman, B. C. 2015, A rolling circle replication mechanism produces multimeric lariats of mitochondrial DNA in Caenorhabditis elegans, *PLoS Genet.*, **11**, e1004985.
44. Mercer, T. R., Neph, S., Dinger, M. E., et al. 2011, The human mitochondrial transcriptome, *Cell*, **146**, 645–658.
45. Temperley, R. J., Wydro, M., Lightowlers, R. N. and Chrzanowska-Lightowlers, Z. M. 2010, Human mitochondrial mRNAs—like members of all families, similar but different, *Biochim. Biophys. Acta*, **1797**, 1081–1085.
46. Markova, S., Filipi, K., Searle, J. B. and Kotlik, P. 2015, Mapping 3' transcript ends in the bank vole (Clethrionomys glareolus) mitochondrial genome with RNA-Seq, *BMC Genomics*, **16**, 870.
47. Zhou, Q., Li, H., Li, H., et al. 2016, Mitochondrial endonuclease G mediates breakdown of paternal mitochondria upon fertilization. *Science*, **353**, 394–399.
48. Ross, J. A., Howe, D. K., Coleman-Hulbert, A., Denver, D. R. and Estes, S. 2016, Paternal mitochondrial transmission in intra-species Caenorhabditis Briggsae hybrids. *Mol. Biol. Evol.*, **33**, 3158–3160.