MitoZ: A toolkit for mitochondrial genome assembly, annotation and visualization

Guanliang Meng\textsuperscript{1,2}, Yiyuan Li\textsuperscript{3}, Chentao Yang\textsuperscript{1,2}, Shanlin Liu\textsuperscript{1,2} *

\textsuperscript{1} BGI-Shenzhen, Shenzhen 518083, China
\textsuperscript{2} China National GeneBank, BGI-Shenzhen, Shenzhen, 518120, China
\textsuperscript{3} Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712

* To whom correspondence should be addressed. Email: liushanlin@genomics.cn

Abstract

Mitochondrial genome (mitogenome) plays important roles in evolutionary and ecological studies. It becomes routine to utilize multiple genes on mitogenome or the entire mitogenomes to investigate phylogeny and biodiversity of focal groups with the onset of High Throughput Sequencing technologies. We developed a mitogenome toolkit MitoZ, consisting of independent modules of \textit{de novo} assembly, findMitoScaf, annotation and visualization, that can generate mitogenome assembly together with annotation and visualization results from HTS raw reads. We evaluated its performance using a total of 50 samples of which mitogenomes are publicly available. The results showed that MitoZ can recover more full-length mitogenomes with higher accuracy compared to the other available mitogenome assemblers. Overall, MitoZ provides a one-click solution to construct the annotated mitogenome from HTS raw data and will facilitate large scale ecological and evolutionary
studies. MitoZ is free open source software distributed under GPLv3 license and available at https://github.com/linzhi2013/MitoZ.

**INTRODUCTION**

With the onset of High Throughput Sequencing (HTS) technologies, we have entered an era in which massive nucleic acid sequencing is becoming routine in phylogenetic and biodiversity monitoring studies (1,2). For example, metabarcoding studies, by taking advantage of complex DNA extracts (e.g. environmental DNA (eDNA)), identify multiple taxa simultaneously from diverse types of samples – stomach contents (3), feces (4,5), sediments (6), soil or water (6-8). In most cases, these studies deal with degraded DNA, thus are in urgent demand for short barcoding fragments for taxonomic identification (9,10). Genes on mitochondrial genomes are preferred due to high copy number per cell, making them more likely be picked up than single-copy nuclear genes. Rapid access to mitochondrial genomes of a myriad of taxa will, firstly, provide critical taxonomic connections between the most abundant and well-constructed DNA barcode COI and those eDNA widely-adopted short markers 12S rRNA, 16S rRNA, CYTB et al. (reviewed by 11); secondly, facilitate the fast-emerging approach – mito-genomics (12-14), which circumvents PCR and requires a taxonomically well covered reference dataset used both for species identification and in gene capture array design (2,15). In addition to its importance in biodiversity monitoring, mitochondrial genome also records maternal inheritance information and is extensively utilized to infer phylogenetic relationship between diverse lineages (1,16).
Apart from the mitogenomes achieved using long-range PCR followed by primer walking strategy and Sanger dideoxy sequencing (17), quite some mitogenomes were obtained using a reference-based method via HTS platform (e.g. 18,19-21). Traditional genome assembly software, for instance, SOAPdenovo2 (22), ALLPATHS-LG (23), Platanus (24), can hardly assemble complete mitogenomes since they are programmed to abandon sequences with extremely-high depth. The two frequently-used mitogenome assembly software, MITObim (25) and NOVOPlasty (26), require closely-related mitochondrial fragments as seeds to anchor short reads and build initial datasets. However, it is often difficult to set an appropriate criterion to define closely-related species – e.g. should an appropriate criterion be congeneric or coordinal in the Linnaean system. The similarities between species also vary a lot between different groups (27). There are also some genera within which none species has a complete mitogenome albeit the plunging cost of sequencing (28). In addition, both software can only generate mitogenome assembly as their final outputs. Thus, separate software, like DOGMA (29), MOSAS (30), MITOS (31) are required for the following genome annotation (32). Besides, all of the three aforementioned annotation software only provide web page version and can hardly deal with assembly with multiple scaffolds.

Here we presented a mitochondrial genome toolkit, MitoZ, providing a one-click solution from HTS raw reads to genome assembly together with annotation and visualization outputs. MitoZ is programmed in Python3 (33) with the assembly module of a modified version of SOAPdenovo-Trans (34), the annotation module of a PERL based script for protein coding genes (PCGs), MiTFi (35) for tRNA and infernal-1.1.1 (36) for rRNA (Fig. 1). We tested the accuracy and efficiency of MitoZ using a batch of mammals and arthropods which have both
mitogenomes obtained by Sanger sequencing in NCBI RefSeq database (37) and shotgun
Paired-End reads in NCBI Sequence Read Archive (SRA) database (38). The results showed
that MitoZ can recover 97.33% of PCGs and rRNA genes of the test samples, of which 94.66%
genomes are in full length and the recovered genes are of high similarity (≥ 97%) to their sanger
sequenced mitogenome.

MATERIAL AND METHODS

1 Samples for test

A total of 30 arthropod and 26 mammal species (Table S1) were selected to evaluate the
performance of MitoZ. Species were picked up by considerations: 1) have mitogenome in
NCBI RefSeq database (37) and were obtained using traditional Sanger sequencing method
(refer as sanger mitogenomes afterward); 2) have HTS data with a volume size ≥ 3 Gb and
Paired-End (PE) read length ≥ 91 bp in NCBI Sequence Read Archive (SRA) database (38);
3) For mammal, data generated from tissue samples were preferred, but 7 blood samples were
included as well for comparison.

We estimated the ratio of mitochondrial derived reads (MDR) for each sample by aligning raw
reads to their corresponding sanger mitogenomes using BWA (version 0.7.12-r1039) (39). It
showed that most samples had a MDR ratio in a range from 0.12% to 0.51% with the mammal
blood samples possessing a significant lower MDR ratio from 0.01% to 0.05% (Table S2). In
addition, we also noticed that six samples (including three mammal non-blood samples, two
mammal blood samples and one arthropod sample) possessed MDR ratio of zero. For those
samples, the MDR could be removed on purpose before data deposit. Thus, we removed them
from the following performance evaluation, leading to a total of 50 samples in the final dataset,
consisting of 29 arthropods, 16 mammal non-blood samples and five mammal blood samples. See Table S3 and Table S4 for details of the procedures of species selection and dataset download.

2 MitoZ

MitoZ consists of multiple modules, including raw data pretreatment, de novo assembly, candidate mitochondrial sequences searching, mitogenome annotation and visualization (Fig. 1). Each module can work independently in case users want a sub part of the entire workflow.

2.1 Raw data pretreatment

A Perl script is designed for filtering the raw data generated by HTS platforms, such as Hiseq 4000. It accepts Pair-End (PE) reads and filters out reads with many ‘N’s, low quality reads, or reads of PCR duplicates (defined as a pair of identical reads). By default, reads will be removed if: 1) of more than 40% low quality (Q<=17) bases; 2) of more than 10 Ns; 3) are PCR duplications

2.2 De novo assembly

Algorithms adopted in SOAPdenovo-Trans (34) fit well for mitogenome assemblies from nuclear DNA extracts, where mitogenome sequences possess considerable higher copy number comparing to nuclear genome sequences - being alike to the differential gene expression patterns it is designed for. We adopted all the algorithms in SOAPdenovo-Trans except for the scaffolding step where we ask for higher linkage support to avoid connections between mitochondrial reads and nuclear mitochondrial DNA segments (NUMTs) (40) and to improve accuracy.
MitoZ has two assembly modes – Quick mode and Multi-Kmer mode. It uses the Quick mode by default, where only one kmer size (K=71) is used for assembly. Users can also use the Multi-Kmer mode to search for the missing PCGs (if any) failed in the Quick mode.

2.3 Mitogenome sequences identification

The assemblies contain both mitogenome sequences and nuclear genome sequences. Thus, we filter out candidate mitogenome sequences using a profile Hidden Markov Model (profile HMM) (41,42) based method which works in an efficient and criteria-relaxed way at first.

HMMER (version 3.1b2) (http://hmmer.org/) (43) is utilized to construct HMM models for both mammal and arthropod clades (2413 and 4007 species, respectively, see Table S5) in the current version. Then, we conduct PCG annotation for candidate mitogenome sequences. The PCG annotation is detailed in section 2.2.4. After that, we use the following three steps to remove potential false positive mitochondrial scaffolds, such as NUMTs and contaminations:

1) Each candidate mitogenome sequence was assigned to a Linnaean taxonomic name using a Python package ETE3 (44) according to their most closely-related PCG homologs in the PCG annotation step. Then, filter out sequences falling outside of a user predefined taxonomic rank, which can be set as order, family or genus.

2) We calculated a confident score $S_j$ for each scaffold to determine the final mitogenome sequences. $S_j$ is calculated using a formula as follows:

$$S_j = A * \sum_{i=1}^{n} C_i$$

where $A$ represents the assemble reliability of each assembly. It is a weight factor representing the reliability of de bruijin route selection – the higher the better. Its
value is mainly consisted of two factors - Kmer depth of contigs and read supportive number when connect contigs to scaffolds. For sequences generated using other assembly software, $A$ will be surrogated by average depth information calculated according to the number of reads that can be mapped to the targeted sequences. $n$ indicates the number of PCGs in sequence $j$ and $C_i$ indicates the completeness (in percentage) of $i$ th PCG, which is calculated by dividing the length for each gene by the length of its shortest reference counterpart in the annotation database.

3) Sequences are ranked by their confident scores. Then, MitoZ tries to find all 13 PCGs from the sequence with the highest confident score. In case mitogenome is not assembled as an integrate single sequence, MitoZ finds the rest PCGs from sequences by rank and skips sequences containing conflict genes, e.g. a complete $COX1$ gene is located in a former sequence, then the latter sequences with lower $S_j$ score containing $COX1$ gene (complete or not) will be skipped. However, if former $COX1$ is incomplete, the latter sequences containing also an incomplete $COX1$ genes will be kept for PCG searching. The searching stops when 13 PCGs are all located, or sequences are run out. In addition, Sequences, regardless of gene conflicts, containing $\geq 5$ PCGs (complete or not) will be retained to confirm identities, e.g. parasites.

2.4 Genome annotation

protein coding genes

An in-house Perl script is designed for PCG annotation. Basically, the script finds candidate PCG sequences by aligning nucleotide sequences to a local protein sequences database using tBlastn in BLAST (version 2.2.19) (45), then uses Genewise (version 2.2.0) (46) to determine
the boundaries of each PCG. MitoZ further tries to determine the precise position of start codons and stop codons by translating the nucleotide sequence with proper mitochondrial genetic code. MitoZ tries to find ‘TA’ or ‘T’ bases, assuming TAA stop codon is completed by adding 3’ A residues to the mRNA in case of absence of the standard stop codons. The current version includes protein database of both Mammals and Arthropods (Table S6).

tRNA genes

MiTFi (35), a covariance models (CMs) (42) based method, is adopted to annotate mitochondrial tRNA genes. It contains 22 manually curated tRNAs database for metazoan and employs a step-wise procedure to search for tRNA candidates in query sequences. MitoZ, by default, outputs tRNA annotation results of e-value $\leq 0.001$ by setting MiTFi parameters as “-cores 1 -evalue 0.001 -onlycutoff -code 2/5(representing Mammal/Arthropod)"

rRNA genes

The 12S rRNA and 16S rRNA genes are annotated using infernal-1.1.1 (36) with the published rRNA CM models (31). MitoZ searches for rRNA with the local searching mode implemented in infernal-1.1.1 in case no candidates are detected using the global searching mode.

2.5 Visualization

Mitochondrion genome features can be illustrated with an independent module, which employed Circos (47) to depict gene elements features, such as PCGs, rRNA genes, tRNA genes, GC content, and sequencing depth distribution. The color of each element can be set as personal preference.

3 Performance Evaluation

MitoZ assembly
We firstly applied MitoZ quick mode to all the 50 test samples, and tried to filter out contamination sequences by mitochondrial PCG annotation (set --requiring_taxa to be taxonomic rank of order for each species). For those didn’t get 13 PCGs, we applied another run using multi-Kmer mode.

**Assembly quality**

We examined the similarities between those mitochondrial genes (PCGs + rRNA) MitoZ recovered and that of Sanger mitogenomes using megablast in BLAST+ (2.6.0) (48). The genes with similarities < 97% to their sanger counterparts were regarded as false positive, but those false positives who can find matches (similarity≥ 97%) to their corresponding genes of the same species on NCBI (detailed in Table S7 and Table S8 and Table S9) were regarded as correct assemblies in the following statistics. We used MAFFT (version 7.309) (49,50) and Unipro UGENE (version 1.26.1) (51) to conduct global alignment between each pair and check mismatches, respectively.

**Factors influencing mitogenome assembly**

A+T content plays an important role in HTS experiments since the known bias in the library preparation step leading to genome regions that possess extreme A+T content tend to have low sequencing depth and are difficult to be assembled (52). Plus, the heterozygosity rate works against the genome assembly quality using HTS platforms (53). The “heteroplasmy” of our samples may come from pooled individuals aiming to produce enough DNA extracts for HTS library construction. We investigated the influences of sequence characteristics on the assembly qualities, including MDR ratio, depth, A+T content and mitogenome heteroplasmy. We aligned HTS reads on their corresponding mitogenomes using BWA (39) and calculated
We calculated the heterozygosity value for each sample based on Site Frequency Spectrum (SFS) obtained using ANGSD.

**Comparison between MitoZ and NOVOPlasty**

We also ran NOVOPlasty (version 2.7.2) to assemble mitogenomes of the 50 test samples with default parameters and the corresponding sanger mitogenome of each species was used as the reference seed. We then annotated the NOVOPlasty results with the annotate module in MitoZ and examined the similarities between those mitochondrial genes (PCGs + rRNA) obtained by NOVOPlasty and that of Sanger mitogenomes (detailed in Table S10 and Table S11).

**RESULTS**

**Mitogenome completeness**

Of the 750 genes ((13 PCGs + 2 rRNAs) * 50 species), 691 (92.13%) genes were full-length recovered, 39 genes (5.20%) were partially recovered and 20 (2.67%) genes were not assembled by MitoZ (Fig. 2 (iii)). The Multi-Kmer mode contributed a total of 46 genes that were either failed (33) or partially (13) assembled in the Quick mode.

We categorized our assembly results into four types: 1) Type A, all the genes (13 PCGs and 2 rRNA) were recovered and represented by one single sequence; 2) Type B, all the genes were assembled, however represented by ≥ 2 sequences; 3) Type C, not all but more than half (≥ 8) of the genes were recovered; 4) Type D, the number of recovered genes was less than eight. In total, we got 37 (74.00%) mitogenomes of type A, four (8.00%) mitogenomes of type B, nine (18.00%) mitogenomes of type C, and none mitogenome of type D. See Fig. 2 (i) and (ii).

**Gene similarities**
For the 735 PCGs and rRNA genes recovered, 724 (98.50%) genes matched their Sanger counterparts well (similarity ≥ 97%, Fig. 3). We further checked these single nucleotide variances (SNV) between genes assembled by MitoZ and their Sanger counterparts. Although the SNVs can arise from individual variances or mitochondrial heteroplasmy, it is also worth to note that those non-perfect-match genes always possess high sequencing depth in HTS assemblies except for the ones located around “Ns” regions (Fig. S1 (i)) and those SNVs are in most cases located in homopolymers or A+T-rich regions (Fig. S1 (ii)), where are regions that typical Sanger sequencing errors happen.

The 5 false positive genes (the red dots in Fig. 3) with low similarities to their Sanger counterparts could be contributed to insufficient sequencing depth in HTS sequencing and sequencing errors in Sanger mitogenomes, see Table S9 for details. Fig. 4 shows an example of mitogenome visualization by MitoZ.

**Comparison between MitoZ and NOVOPlasty**

NOVOPlasty successfully recovered 570 (76.00%) PCGs and rRNA genes of full length, partially assembled 30 (4.00%) genes and failed to assemble 150 (20.00%) genes (Fig. 5 (i)). A total of 133 (17.73%) NOVOPlasty-failed genes were successfully assembled by MitoZ. The gene similarities between NOVOPlasty and Sanger mitogenomes were in concert with that of MitoZ, indicating these genes of low similarities were probably attributed to intra-species genetic variation (Fig. 5 (ii)).

**Factors influencing mitogenome assembly**

The portion of mitochondrial DNA varies in different samples. Our results showed that the assemblies of type A and B mitogenomes tended to possess higher MDR ratios than the assemblies of type C and D (Fig. S2 (i) and (ii)). However, no significant correlation was
detected between mitogenome assembly qualities and factors of both A+T content and heterozygosity ratio (Fig. S2 (iii) and (iv)).

**DISCUSSION**

The development of HTS technique and low sequencing cost greatly facilitates mitogenome sequencing which used to require a tedious process of long range PCR followed by primer walking (56). Nowadays, we are able to obtain mitogenomes from shotgun reads with higher efficiency and lower cost. At the same time, however, the large volume of data challenges our ability to efficiently analyze the exponentially growing dataset. MitoZ, a versatile mitogenome toolkit, aims to achieve mitogenome assemblies together with annotation results from whole genome shotgun reads. It is by far the easiest method to deliver human-readable outcomes for mitogenome studies – require no special pretreatment in either DNA extraction or nucleotide sequencing, and it combines the key bioinformatic steps – clean data filtering, *de novo* assembly, annotation and visualization, thus provides users an ‘one-step’ solution from raw data to publishable outcomes and will accelerate the accumulation of mitogenomes. In addition, MitoZ conducts assembly without the aid of reference sequences from closely-related species, which can be a crucial feature when the mitochondrial genes from closely-related species are unavailable.

The boundaries of PCGs, especially the stop codons, of many mitochondrial PCGs are not precisely determined in Genbank. Aside from tBlastn and Genewise, MitoZ developed an in-house Python script to further determine the start and stop codons around the boundary of each gene and in most cases was able to precisely locate the start and stop codons (Table S7 and Table S8). In addition, the current annotation module, especially PCGs annotation, works well
mainly for arthropods and mammals and needs further improvement to support more domains of life.

In summary, MitoZ shows the ability to assemble and annotate mitogenomes efficiently and accurately. With the rapid accumulation of mitogenomes and robust reference databases of specific environments or groups, it will facilitate the developments of several important fields in the foreseeable future, such as phylogenetic inference, quarantine inspection, aquatic and agriculture ecosystems scrutiny.

**AUTHOR’S CONTRIBUTION**

S.L. and G.M. designed this study. Most of the coding work was conducted by G.M., with minor contributions from S.L., Y.L. and C.Y. S.L. and G.M. drafted the manuscript. All the authors contributed to the manuscript revise.

**ACKNOWLEDGEMENT**

We sincerely thank to Chengran Zhou, Min Tang and Xin Zhou for their trying out the MitoZ’s beta version and their positive feedbacks. S.L., G.M. and C.Y. is supported by Free-oriented Project from Shenzhen Government (JCYJ20170817150755701).
**Fig. 1. MitoZ toolkit components.** Ellipses indicate input data files, rectangles (solid line) represent functional modules in MitoZ which can be run independently when users provide corresponding input files.
**Fig. 2. MitoZ test result.** (i) the proportion of assembled mitogenomes in different assembly categories; (ii) diagram of different assembly types where the boxes represent PCGs or rRNA genes and the solid lines stand for the other parts of mitogenomes and the upper grey boxes represent sanger mitogenome while the lower colorful ones are assembled by MitoZ; (iii): the gene completeness distribution of PCGs and rRNA genes of the 50 test species. “full length” means gene completeness ≥ 95%, “partial” means gene completeness < 95%, and “missing” means the genes that were not recovered by MitoZ.
Fig. 3. Gene similarities (MitoZ vs Sanger). The blue dots present three species whose genes possessed similarities < 97% to their Sanger mitogenomes but can find high similarity genes of the same species in NCBI NT database. Such incongruences could derive from intraspecies variances. The red points (5 in total) present genes possessed similarities < 97% to their Sanger mitogenomes and could not find better hits in NCBI. The rest genes and samples were presented by green dots.
Fig. 4. Demonstration of mitogenome visualization using MitoZ
Fig. 5. Performance comparisons between MitoZ and NOVOPlasty. (i) Left: mitogenome types, see Fig. 2(i) and Fig. 2(ii) for the categories of mitogenome types, while type D indicates the total number of PCG and rRNA genes recovered was less than eight. Right: gene (PCGs and rRNA genes) completeness distribution, see Fig. 2(iii) for the meanings of “full length”, “partial” and “missing”. (ii) Diagram of gene similarities to the Sanger genes. Genes that recovered by both MitoZ and NOVOPlasty (n=592) were included in the analysis, and paired samples t-test was used.
Supplementary files

Fig. S1. Sequencing depth and A+T content around the indel regions.

Fig. S2. Factors influencing mitogenome assembly.

Table S1. Species list used for MitoZ performance testing

Table S2. A+T content of sanger mitogenome and sanger mitochondrial genes, mitochondrial reads ratio and heterozygosity of SRA data

Table S3. Procedures and species list for selecting test species for arthropods

Table S4. Procedures and species list for selecting test species for mammals

Table S5. Species list for building PCGs HMM models

Table S6. Species list for building mitochondrial PCGs annotation database

Table S7. MitoZ genes vs. sanger genes of 29 arthropod samples

Table S8. MitoZ genes vs. sanger genes of 21 mammal samples

Table S9. Information of the five false positive genes and samples with low similarity (<97%) genes in MitoZ assembly

Table S10. NOVOPlasty genes vs. sanger genes of 29 arthropod samples

Table S11. NOVOPlasty genes vs. sanger genes of 21 mammal samples
1. Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K.E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G. et al. (2013) Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet*, 45, 1176-1182.

2. Tang, M., Hardman, C.J., Ji, Y., Meng, G., Liu, S., Tan, M., Yang, S., Moss, E.D., Wang, J., Yang, C. et al. (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods Ecol Evol*, 6, 1034-1043.

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

4. Bohmann, K., Monadjem, A., Lehmkuhl Noer, C., Rasmussen, M., Zeale, M.R., Clare, E., Jones, G., Willerslev, E. and Gilbert, M.T. (2011) Molecular diet analysis of two African free-tailed bats (molossidae) using high throughput sequencing. *PLoS One*, 6, e21441.

5. Kartzinel, T.R., Chen, P.A., Coverdale, T.C., Erickson, D.L., Kress, W.J., Kuzmina, M.L., Rubenstein, D.I., Wang, W. and Pringle, R.M. (2015) DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proc Natl Acad Sci USA*, 112, 8019-8024.

6. Shaw, J.L.A., Clarke, L.J., Wedderburn, S.D., Barnes, T.C., Weyrich, L.S. and Cooper, A. (2016) Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. *Biological Conservation*, 197, 131-138 %@ 0006-3207.

7. Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., Minamoto, T. and Miya, M. (2017) Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. *Sci Rep*, 7, 40368.

8. Taberlet, P., Prud'Homme, S.M., Campione, E., Roy, J., Miquel, C., Shehzad, W., Gielly, L., Rioux, D., Choler, P., Clement, J.C. et al. (2012) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Mol Ecol*, 21, 1816-1820.

9. Rees, H.C., Maddison, B.C., Middleditch, D.J., Patmore, J.R.M., Gough, K.C. and Crispo, E. (2014) REVIEW: The detection of aquatic animal species using environmental DNA - a review of eDNA as a survey tool in ecology. *Journal of Applied Ecology*, 51, 1450-1459.

10. Goldberg, C.S., Turner, C.R., Deiner, K., Klymus, K.E., Thomsen, P.F., Murphy, M.A., Spear, S.F., McKee, A., Oyler-McCance, S.J. and Cornman, R.S. (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution* %@ 2041-210X.

11. Pompanon, F., Deagle, B.E., Symondson, W.O., Brown, D.S., Jarman, S.N. and Taberlet, P. (2012) Who is eating what: diet assessment using next generation sequencing. *Mol Ecol*, 21, 1931-1950.
12. Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., Tang, M., Fu, R., Li, J. and Huang, Q. (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience*, 2, 4.

13. Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S., Song, W., Li, Y., Wu, Q., Zhang, A. *et al.* (2014) Multiplex sequencing of pooled mitochondrial genomes-a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Res.*, 42, e166.

14. Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M.J.T.N., Baselga, A., Vogler, A.P. and Gilbert, M. (2015) Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, 6, 883-894.

15. Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., Zhang, H., Misof, B., Kjer, K.M., Tang, M. *et al.* (2016) Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Mol Ecol Resour.*, 16, 470-479.

16. Cameron, S.L. (2014) Insect mitochondrial genomics: implications for evolution and phylogeny. *Annu Rev Entomol.*, 59, 95-117.

17. Hu, M., Jex, A.R., Campbell, B.E. and Gasser, R.B. (2007) Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing. *Nature Protocols*, 2, 2339-2344.

18. Hunt, V.L., Tsai, I.J., Coghlan, A., Reid, A.J., Holroyd, N., Foth, B.J., Tracey, A., Cotton, J.A., Stanley, E.J., Beasley, H. *et al.* (2016) The genomic basis of parasitism in the Strongyloides clade of nematodes. *Nature Genetics*, 48, 299-307.

19. Smeds, L., Warmuth, V., Bolivar, P., Uebbing, S., Burri, R., Suh, A., Nater, A., Bureš, S., Garamszegi, L.Z., Hogner, S. *et al.* (2015) Evolutionary analysis of the female-specific avian W chromosome. *Nature Communications*, 6.

20. Babbucci, M., Basso, A., Scupola, A., Patarnello, T. and Negrisol, E. (2014) Is It an Ant or a Butterfly? Convergent Evolution in the Mitochondrial Gene Order of Hymenoptera and Lepidoptera. *Genome Biology and Evolution*, 6, 3326-3343.

21. Wang, Z., Wang, Z., Shi, X., Wu, Q., Tao, Y., Guo, H., Ji, C. and Bai, Y. (2018) Complete mitochondrial genome of Parasesarma affine (Brachyura: Sesarmidae): Gene rearrangements in Sesarmidae and phylogenetic analysis of the Brachyura. *Int J Biol Macromol.*, 118, 31-40.

22. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1, 18.

23. Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA*, 108, 1513-1518.

24. Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H. *et al.* (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.*, 24, 1384-1395.

25. 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 Research*, 41, e129-e129.
26. Dierckxsens, N., Mardulyn, P. and Smits, G. (2017) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.*, **45**, e18.

27. Rebijith, K.B., Asokan, R., Krishna, V., Ranjitha, H.H., Krishna Kumar, N.K. and Ramamurthy, V.V. (2014) DNA Barcoding and Elucidation of Cryptic Diversity in Thrips (Thysanoptera). *Florida Entomologist*, **97**, 1328-1347.

28. Wolfsberg, T.G., Schafer, S., Tatusov, R.L. and Tatusova, T.A. (2001) Organelle genome resources at NCBI. *Trends in Biochemical Sciences*, **26**, 199-203.

29. Wyman, S.K., Jansen, R.K. and Boore, J.L. (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252-3255.

30. Sheffield, N.C., Hiatt, K.D., Valentine, M.C., Song, H. and Whiting, M.F. (2010) Mitochondrial genomics in Orthoptera using MOSAS. *Mitochondrial DNA*, **21**, 87-104.

31. Bernt, M., Donath, A., Juhling, F., Externbrink, F., Florentz, C., Fritzsch, G., Putz, J., Middendorf, M. and Stadler, P.F. (2013) MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol*, **69**, 313-319.

32. Xu, P., Velozo Timbó, R., Coiti Togawa, R., M. C. Costa, M., A. Andow, D. and Paula, D.P. (2017) Mitogenome sequence accuracy using different elucidation methods. *Plos One*, **12**, e0179971.

33. Van Rossum, G. (2007), Vol. 41, pp. 36.

34. Leinonen, R., Sugawara, H., Shumway, M. and International Nucleotide Sequence Database, C. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19-21.

35. Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G.X. (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

36. Wheeler, T.J. and Eddy, S.R. (2013) nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, **29**, 2487-2489.
44. Huerta-Cepas, J., Serra, F. and Bork, P. (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol*, **33**, 1635-1638.
45. Gertz, E.M., Yu, Y.K., Agarwala, R., Schaffer, A.A. and Altschul, S.F. (2006) Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol*, **4**, 41.
46. Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and Genomewise. *Genome Res*, **14**, 988-995.
47. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res*, **19**, 1639-1645.
48. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
49. Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*, **30**, 3059-3066.
50. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, **30**, 772-780.
51. Okonechnikov, K., Golosova, O., Fursov, M. and team, U. (2012) Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, **28**, 1166-1167.
52. Sims, D., Sudbery, I., Iott, N.E., Heger, A. and Ponting, C.P. (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**, 121-132.
53. Pryszcz, L.P. and Gabaldon, T. (2016) Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res*, **44**, e113.
54. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**.
55. Korneliussen, T.S., Albrechtsen, A. and Nielsen, R. (2014) ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, **15**, 356.
56. Crampton-Platt, A., Yu, D.W., Zhou, X. and Vogler, A.P. (2016) Mitochondrial metagenomics: letting the genes out of the bottle. *Gigascience*, **5**, 15.