Reviewers' Comments

First of all, we would like to thank the reviewers for their thoughtful and very supportive comments. The enclosed manuscript entitled, "*Mining NCBI Sequence Read Archive database: An untapped source of organelle genomes for taxonomic and comparative genomics research*" has been revised according to the insightful suggestions and comments from Reviewers and is hereby submitted for the possible publication in *Diversity*. Below we address the specific points raised by the reviewers and elaborate on the corresponding changes in the manuscript.

Reviewer #1:

We thank the reviewer for the very positive general evaluation and useful suggestions to strengthen our manuscript. Below we have outlined a detailed response to each point in blue.

Comment: "*I have checked OrgaMiner on the website. It seems that SPAdes was the major software for mitochondrial and chloroplast genome assembly from genomic data. I was wondering why the authors did not try other softwares? According to my own experience, different softwares could probably generated different results.*"

Response: We appreciate the reviewer's insightful comment and agree that animal mitogenome and plant plastome assembly methods are sometimes very robust and several softwares are available for this, e.g., MitoZ, SPAdes, and GetOrganelle. Each method has pros and cons in *de novo* organelle genome assembly. Both software tools can be utilized in *de novo* mitochondrial genome assembly studies and integrated into our OrgaMiner workflow. However, for *de novo* organelle genome assembly, we chose the Spades-based GetOrganelle software over the k-mer-based MitoZ software (the extension of SOAP denovo-Trans software). There are two critical reasons for this preference. First, MitoZ employs a k-mer-based approach, offering the advantage of widespread applicability for assembling genomes of novel species. However, k-mer-based approaches like MitoZ are more prone to errors without providing seed sequences from closely related species. The performance of *de novo* assemblers based on k-mers might not be successful in accurately identifying and distinguishing sequences, such as nuclear-embedded mitochondrial DNA sequences (NUMTs). These sequences are mitochondrial DNA sequences integrated into the nuclear genome, posing a significant challenge in *de novo* assembly. For this purpose, we utilize the GetOrganelle software (or Spades). This software uses a short seed sequence from the species or any closely related species sequence information; in other words, it applies a "**seed-and-extend**" approach, along with *De Bruijn* Graphs and the k-mer approach. As using the "seed-and-extend" approach in analyses and the k-mer-based approach allows obtaining a more accurate organelle genome assembly, we adopted the software GetOrganelle as the main *de novo* organelle genome assembler in the OrgaMiner workflow.

Comment: "*How did the authors evaluate the quality of the assembled mitogenomes?*"

Response: The reviewer brings up an excellent point here, which we felt helped strengthen the manuscript significantly through addressing. Indeed, when executing the OrgaMiner workflow,

the accuracy of the mitochondrial genome was assessed and verified following the *de novo* assembly and annotation analysis of the organelle genome. However, we overlooked mentioning the validation method in the current article. Therefore, we also thank the reviewer very much for allowing us to clarify this matter. In the initial phase of *de novo* assembly of mollusk mitochondrial genomes, we downloaded and analyzed NGS data for both mollusk species without a mitochondrial genome sequence and those with known sequences and annotations. Then, the results, obtained through *de novo* assembly and annotation of mollusk species with known mitochondrial genomes, were compared with the corresponding mtDNA sequences downloaded from the NCBI nucleotide database. This comparison encompassed (i) the alignment of already known and *de novo* assembled mitochondrial sequences and (ii) an assessment of alignment identity (percentage identity), alignment coverage, and gene number and composition comparisons. In this context, the following statements have been added to the "*2.3. Validation and application of the OrgaMiner Workflow across diverse clades*" in the material and method section.

2.3. Validation and application of the OrgaMiner Workflow across diverse clades

We performed a series of comparative analyses to validate the accuracy of the OrgaMiner workflow's mitochondrial genome annotations. First, we downloaded the NGS data using the workflow from mollusk species with well-established mitochondrial genome annotations deposited in the NCBI GenBank database. Then, OrgaMiner processed the NGS data and compared the obtained results with those stored in GenBank, demonstrating the reliability of the workflow. The following criteria assessed completeness: sequence alignment identity (percentage identity), alignment coverage, and gene number and composition comparison. Further analyses were conducted to determine the validity of the OrgaMiner workflow in different animal and plant groups. For this reason, the NGS datasets from segmented worms (Phylum Annelida) were selected and analyzed for de novo mitochondrial genome assembly and annotation using the OrgaMiner workflow. Similarly, using the same workflow, the NGS datasets belonging to various species within the dicotyledon Solanum and monocotyledon Zingiberales taxonomic groups were assembled and annotated for their chloroplast genomes. Following the assembly and annotation stages using Chlorobox for chloroplast genomes, a comparative analysis was undertaken between the assembled organelle genomes and those obtained from the NCBI. First, a BlastN database was generated with all gene sequences extracted from the reference assemblies. Subsequently, the BlastN (1e-10) algorithm was employed, using the gene sequences of each assembled genome as queries to identify sequence similarities with the reference NCBI genomes. Additionally, the genes present in both the reference genomes and the assembled genomes of each species were compared to assess the completeness of the organelle genomes.

The findings from the mollusk mitochondrial genome assembly and annotation validation are further described in the "*3.2. Molluscan mtDNA annotation findings and implication*" subtitle to clarify this point.

3.2. Molluscan mtDNA annotation findings and implication

To evaluate the accuracy of mollusk mitochondrial genome annotations generated by the OrgaMiner workflow, NGS data from species with known mitochondrial genome sequences and annotations were

analyzed using the workflow. The resulting mitochondrial genome and annotation files were then compared to the corresponding data available in the NCBI Nucleotide database. These analyses demonstrated that the OrgaMiner workflow produces mitochondrial genomes highly similar to those in the NCBI Nucleotide database, as evidenced by metrics such as alignment identity (percentage identity), alignment coverage, and gene number/composition comparisons. We processed a total of 10 WGS and 10 RNA-Seq mollusk data using the OrgaMiner workflow and compared the results to their mitochondrial genomes in the NCBI nucleotide database (Table S3). Compared to reference mtDNA, regardless of coding or non-coding genes, the alignment identity were found to be >99.18% for WGS data and >98.60% for RNA-Seq data, respectively. However, WGS data appears to be more successful in extracting mitochondrial genes than RNA-Seq data ("genome comparison" in Table S3). The validation results largely indicate the reliability of the OrgaMiner workflow in analyzing NGS data from the NCBI SRA. In some cases, partial mitochondrial and chloroplast genomes may be obtained from the OrgaMiner workflow. This occurrence could be attributed to the nature of the RNA-Seq or WGS NGS datasets in the NCBI SRA database rather than the OrgaMiner workflow itself. While the read length, depth, and Phred quality score are satisfactory, several factors inherent to NGS data in the ".fastq" format can negatively impact the analysis. These factors include: (i) unlike WGS data, organellar coding genes, and particularly tRNA genes, are represented at low levels in RNA-Seq data; (ii) the high duplication rates of certain NGS reads, such as high rRNA duplication common in RNA-Seq data; and (iii) the potential presence of contaminant sequences (non-target organisms).

The results regarding the application of the OrgaMiner workflow to various animal datasets (Annelida) and plant(dicotyledon Solanum and monocotyledon Zingiberales) NGS datasets is investigated in "*3.3. The applicability of OrgaMiner workflow across different clades*".

3.3. The applicability of OrgaMiner workflow across different clades

To test the validity of the OrgaMiner workflow for other organism groups, we performed mitochondrial genome analyses on animals with the taxonomic keyword Annelida. As for plants, we analyzed the chloroplast genomes of both dicotyledons from the Solanum genus and monocotyledons from the Zingiberales order. A total of 37 annelid mitochondrial genomes and annotations were obtained, 27 from whole-genome sequencing (WGS) data and 10 from RNA-Seq data. When examining the mitochondrial genome assembly and annotation of these species in the NCBI nucleotide database, it is observed that accurate results are obtained using the OrgaMiner workflow (Table S4). Compared to reference mitogenomes, for annelid species, the alignment identity was >99.65% for WGS data and >96.67% for RNA-Seq data, respectively. Using the OrgaMiner workflow, nearly all mitochondrial genes of annelids were comprehensively covered. As observed in the mollusk species, WGS data appeared to outperform RNA-Seq data in the de novo retrieval and annotation of annelid organelle genomes regarding the number of genes retrieved ("mitochondrial comparison" in Table S4). As for plant species, we analyzed 15 species from the Solanum genus, including 6 with WGS and 9 with RNA-Seq data, as well as 26 species from the Zingiberales order, all with WGS data. The results from both groups were compared to their reference chloroplast genome sequence and annotations. Regardless of plant species, all species showed over 96% alignment identity. In WGS data analysis of species from the genus Solanum and the order Zingiberales,

the chloroplast genome and genes are largely represented and compatible with the reference genome/annotation ("plastid comparison" in Table S4). Although de novo assembled chloroplast genomes from both the WGS and RNA-seq data have shown promising results with many unique genes aligned and commonly found in the reference genomes, there were some inconsistencies observed between the reference genomes and OrgaMiner-assembled genomes, as gene numbers and compositions differed more than expected. This difficulty may have arisen due to the challenging nature of annotating plastid genomes, a task necessitating manual curation that users may contemplate following the assembly of plastid genomes [58].

Reference

Qu, X. J.; Zou, D.; Zhang, R. Y.; Stull, G. W.; Yi, T. S. Progress, challenge and prospect of plant plastome annotation. Front Plant Sci 2023, 14.

Comment: "*Line 178: One "sequence" should be deleted.*"

Response: We thank the reviewer for pointing this out. Modified as suggested.

Comment: "*Line 293: Here "two species" might be changed to "three species" since three names were listed in the brackets*."

Response: We thank the reviewer for pointing this out. Modified as suggested.