2023-07-18

MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads

Uliano-Silva, M

10.1186/s12859-023-05385-y
BMC Bioinformatics
Springer Science and Business Media LLC

Uliano-Silva *et al. BMC Bioinformatics* (2023) 24:288
https://doi.org/10.1186/s12859-023-05385-y

BMC Bioinformatics

SOFTWARE Open Access

MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads

Marcela Uliano-Silva[1*†], João Gabriel R. N. Ferreira[2,3†], Ksenia Krasheninnikova[1], Darwin Tree of Life Consortium, Giulio Formenti[4], Linelle Abueg[4], James Torrance[1], Eugene W. Myers[5,6], Richard Durbin[7,1], Mark Blaxter[1] and Shane A. McCarthy[1,7]

†Marcela Uliano-Silva and João Gabriel R. N. Ferreira contributed equally to this work.

*Correspondence: mu2@sanger.ac.uk

[1] Tree of Life, Wellcome Sanger Institute, Cambridge CB10 1SA, UK
[2] Bio Bureau Biotecnologia, Rio de Janeiro, Brazil
[3] Instituto de Biofísica Carlos Chagas Filho, UniversidadeFederal Do Rio de Janeiro, Rio de Janeiro, Brazil
[4] The Rockefeller University, New York, NY, USA
[5] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
[6] Okinawa Institute of Science and Technology, Okinawa, Japan
[7] Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

Abstract

Background: PacBio high fidelity (HiFi) sequencing reads are both long (15–20 kb) and highly accurate (> Q20). Because of these properties, they have revolutionised genome assembly leading to more accurate and contiguous genomes. In eukaryotes the mitochondrial genome is sequenced alongside the nuclear genome often at very high coverage. A dedicated tool for mitochondrial genome assembly using HiFi reads is still missing.

Results: MitoHiFi was developed within the Darwin Tree of Life Project to assemble mitochondrial genomes from the HiFi reads generated for target species. The input for MitoHiFi is either the raw reads or the assembled contigs, and the tool outputs a mitochondrial genome sequence fasta file along with annotation of protein and RNA genes. Variants arising from heteroplasmy are assembled independently, and nuclear insertions of mitochondrial sequences are identified and not used in organellar genome assembly. MitoHiFi has been used to assemble 374 mitochondrial genomes (368 Metazoa and 6 Fungi species) for the Darwin Tree of Life Project, the Vertebrate Genomes Project and the Aquatic Symbiosis Genome Project. Inspection of 60 mitochondrial genomes assembled with MitoHiFi for species that already have reference sequences in public databases showed the widespread presence of previously unreported repeats.

Conclusions: MitoHiFi is able to assemble mitochondrial genomes from a wide phylogenetic range of taxa from Pacbio HiFi data. MitoHiFi is written in python and is freely available on GitHub (https://github.com/marcelauliano/MitoHiFi). MitoHiFi is available with its dependencies as a Docker container on GitHub (ghcr.io/marcelauliano/mitohifi:master).

Keywords: MitoHiFi, Docker, Mitogenome, Heteroplasmy, Long reads, HiFi, Python, Singularity, DToL

Background

Recent advances in genomics have opened the prospect of full genome sequencing of a wide range of species. Both global and taxonomically- or geographically-bounded projects have been initiated to exploit these new technologies to build reference genome libraries for all species on Earth. It is expected that these genome sequences will be a foundational dataset for new understanding in biology, new avenues in conservation and biodiversity monitoring, and new data to promote sustainable bioindustry [1, 2]. The Earth Biogenome Project (EBP) [3] was founded to coordinate and promote biodiversity genomics through affiliated projects such as the Darwin Tree of Life project (DToL) [4], the Vertebrate Genomes Project (VGP) [5] and the Aquatic Symbiosis Genomics project (ASG) [6]. DToL is geographically-focussed, and aims to sequence all eukaryotic species living in and around the islands of Britain and Ireland. The goals of VGP are taxonomically-oriented (to sequence all vertebrate species), while those of ASG are defined by a particular biology (eukaryotes that live in symbiosis with prokaryotic or eukaryotic microbial partners). For each of these and other EBP projects, extensive development is needed in sample collection and preservation, nucleic acid extraction, sequencing, assembly, curation and annotation. A key driver in all these fields is the development of processes that can work at scale, processing hundreds to thousands of species rapidly.

One key advance has been the release of commercial long read sequencing platforms. Here we focus in particular on the PacBio zero mode waveguide, single molecule real time sequencing technology, and the particular deployment of this approach to produce long, high-quality reads using a circular consensus approach (called HiFi for high fidelity). These data have N50 read lengths of 15–20 kb, and accuracies higher than 1 error in 100 (Q20). This radically new data type has changed the landscape of genome assembly [7], and new toolkits (such as HiCanu [8] and HiFiAsm [9]) have been developed to exploit the characteristics of HiFi data.

Except for a few groups, all eukaryotes carry an essential cytoplasmic organelle, the mitochondrion. The mitochondrion derives from an ancient symbiosis between the last common ancestor of all eukaryotes and an alphaproteobacterial cell, and while many genes that were on the ancestral bacterial genome have been transferred to the nuclear genome [10], the mitochondrion generally retains a reduced, circular genome [11]. Mitochondria play a fundamental role in aerobic respiration and other key processes and the genes on the mitochondrial genome are essential for these functions [12]. While mitochondrial genome content varies extensively [13], in Metazoa the gene content of the mitochondrial replicon is relatively stable (11 or 12 protein coding genes, two ribosomal RNAs and a set of tRNAs that decode the distinct organellar genetic code). Because the mitochondrial genome is haploid, and generally passes only through the maternal lineage, drift and linked selection tend to purge variants rapidly. Heteroplasmy, the presence of variant mitochondrial genomes within a single individual, has been thought to be generally rare and often associated with disease or normal somatic ageing processes. Although repeats have been reported in mitochondrial genomes [14], the extent of the phenomenon is not fully understood. For example, the assembly and analysis of PacBio CLR and ONT sequencing data identified widespread repeat content and frequent heteroplasmy in vertebrate mitochondrial genomes [14]. These features were not evident in, or accessible through, previous short-read or PCR-and-sequence approaches.

The increase in the rate of generation of reference genomes across biodiversity using accurate long reads offers an opportunity to revisit the evolution of the gene content and structure of mitochondrial genomes. This opportunity requires development of robust pipelines that reliably assemble organellar genomes, correctly report repeat content, resolve heteroplasmy and distinguish true mitochondrial genome sequence from the frequent presence of nuclear genome inserted copies (nuclear mitochondrial transfers or NUMTs). Here we present MitoHiFi, a software designed to use Pacbio HiFi reads to assemble and annotate complete mitochondrial genomes. MitoHiFi was developed within the Darwin Tree of Life Project and has been deployed to assemble mitochondrial genomes from hundreds of Metazoa and Fungi. MitoHiFi is distributed under the MIT licence, and is available on GitHub (https://github.com/marcelauliano) and as a Docker container also on GitHub (ghcr.io/marcelauliano/mitohifi:master).

Results

MitoHiFi is a robust toolkit for mitochondrial genome assembly from HiFi data

MitoHiFi is written in python and orchestrates a series of external tools to select mitochondrial reads from whole genome sequencing datasets, filter out reads that are likely to derive from nuclear-mitochondrial transfers (NUMTs), assembles the genome, circularising it when possible and annotate it for protein and RNA genes. MitoHiFi also identifies possible heteroplasmic variants present in the data (Fig. 1). We describe the processes embodied in MitoHiFi below.

Selection of PacBio HiFi reads based on reference mitochondrial genome(s) from closely-related species

As primary input, MitoHiFi takes Pacbio HiFi reads (-r flag), or contigs assembled from those reads (-c), and a reference mitochondrial genome from a closely-related species along with its annotation in GenBank format. The pipeline includes a script to automatically download closely-related mitochondrial genomes based on the NCBI taxonomy (findMitoReference.py). The defaults are set to download a single reference genome from the same or most-closely related species, but multiple references can be downloaded by setting -n (> 1). When MitoHiFi is starting from reads (-r), it first maps the reads to this reference mitochondrial genome. The aligned read set is filtered to exclude reads that have lengths more than the exact length of the reference genome, as these are likely to derive from NUMT insertions. However, this parameter can be changed when appropriate (*flag –max-read-len*).

Assembly of mitochondrial reads and filtering of contigs to avoid NUMTs

The readset is then assembled with Hifiasm [9] and assembled contigs are parsed through a series of filters to yield final mitochondrial genome candidates. The contigs are compared to the reference mitochondrial genome using BLAST + [15]. Contigs are retained if.

i. Over 50% of their sequence length is present in the BLAST match with the reference (the user can change this threshold with the flag *-p*).

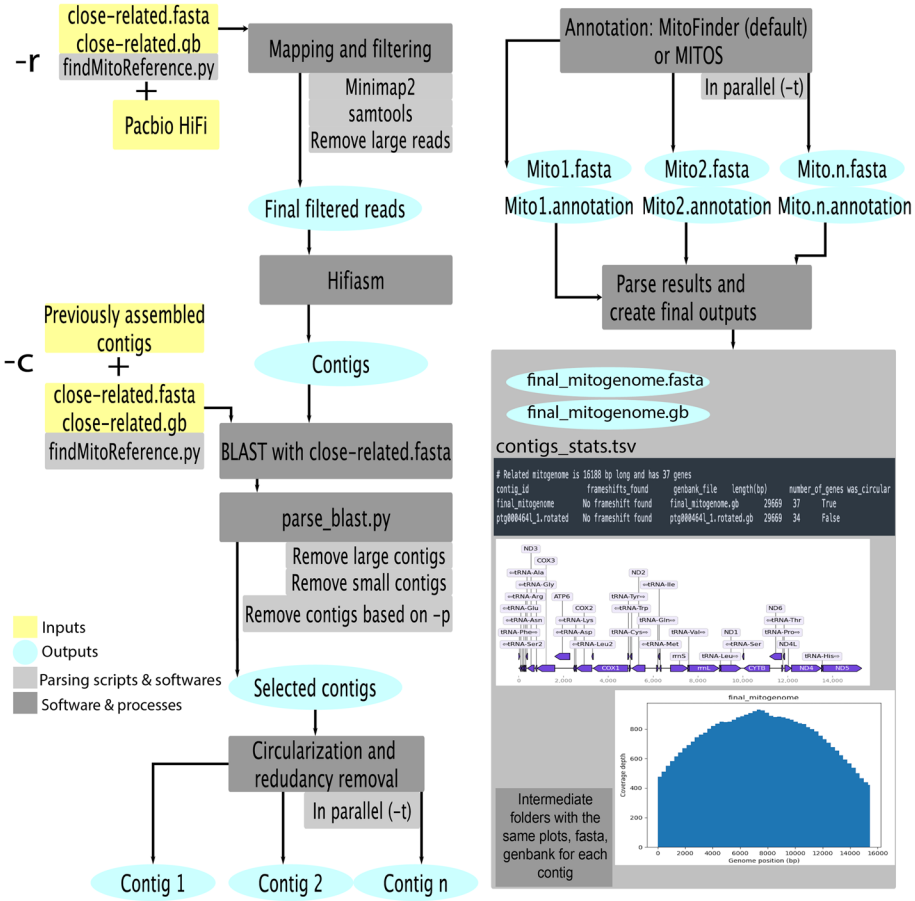

Fig. 1 Outline of the MitoHiFi workflow. Flow diagram of MitoHiFi processes. The inputs to MitoHiFi are a closely-related reference mitochondrial genome with either (i) raw Pacbio HiFi reads (−r parameter), or (ii) previously assembled contigs (−c parameter). Example of outputs generated by MitoHiFi. A summary of the outputs generated by the pipeline, including the contigs_stats.tsv table summarising the metrics for the assembly, output sequence and annotation files, and graphical representations of coverage and annotation

ii. The contig is less than five times the length of the reference (long contigs are likely to be NUMTs).

iii. The contig is > 80% of the length of the reference (smaller contigs are likely to be incomplete).

All BLAST parsing information is saved in intermediate files *parsed_blast.txt* and *parsed_blast_all.txt* (see GitHub page for intermediate output column details).

Circularization

All candidate contigs are processed through circularizationCheck [16]. This module uses self-BLAST of each contig to identify sequence redundancy between its ends. As default > 220 bp of overlap is required but the user can change this by setting *–circular-size* to a required length. MitoHiFi outputs circularisation information to *all_contigs.circularisationCheck.txt*.

Annotation and rotation

Candidate contigs are annotated in parallel (-t flag for number of threads) using MitoFinder (the default) [16] or MITOS [17] (using flag –mitos). The mitochondrial genetic code should be set with the flag -o. MitoFinder finds protein coding genes through BLAST similarity searches using the reference nucleotide and protein sequences. MITOS performs de novo annotation based on its database of orthologous genes. Both annotation pipelines use mitfi [18] to identify and classify tRNAs. MitoHiFi outputs the annotation for all assembled contigs in multiple formats (GenBank [gb], general feature format [gff], FASTA) and produces plots of the annotated features and of reads mapping coverage, if initiated from reads (-r). After annotation, MitoHiFi identifies the tRNA-Phe locus and rotates the mitochondrial genome to have its start at the first base of this locus, following established convention. If no contig has a tRNA-Phe locus, then the most frequent tRNA locus among all contigs is chosen as reference for rotation.

Choice of representative mitochondrial genome

We find that Hifiasm [9] within MitoHiFi frequently assembles more than one candidate mitochondrial genome from raw read data. This may be due to read error and high coverage, or true heteroplasmy (see Discussion). MitoHiFi outputs a sequence file (FASTA format) and annotation (GenBank format) for each contig, and selects a final representative mitochondrial genome using the following criteria:

(i) All potential contigs are sorted by number of genes annotated in relation to the reference, then:

a. It searches for contigs that were classified as circular (and has had sequence redundancy removed)
b. Have a similar size to the reference mitochondrial size
c. And its annotation includes no genes that contain a frameshift.

(ii) If no contig passes the criteria in I, MitoHiFi will select the circular contig that follows at least two of the criteria above in order A>B>C.

The selected main mitochondrial genome assembly files are renamed *final_mitogenome.fasta* and *final_mitogenome.gb*, and a graphical representation of the annotation is produced (*final_mitogenome.annotation.png)* along with a plot of the reads (gbk.HiFiMapped.bam.filtered.fasta*)* mapped to it (*final_mitogenome.coverage.png*). Mapping quality filtering for the final coverage plots can be set with the flag *-covMap*. MitoHiFi also outputs annotation and coverage plots for all other potential contigs (*contigs_annotations.png*, *coverage_plot.png*). To produce all the potential contigs coverage plot, they are concatenated in a file to use all as reference for mapping. Coverage plots are only produced when MitoHiFi is started with the flag -r.

MitoHiFi also produces a summary file *contigs_stats.tsv* that reports which closely-related reference was used and gives details of each assembled mitochondrial genome (including assembly size, number of genes annotated, presence of frameshifts in genes and circularisation data) (Fig. 1). Finally, MitoHiFi produces a *shared_genes.tsv* file that

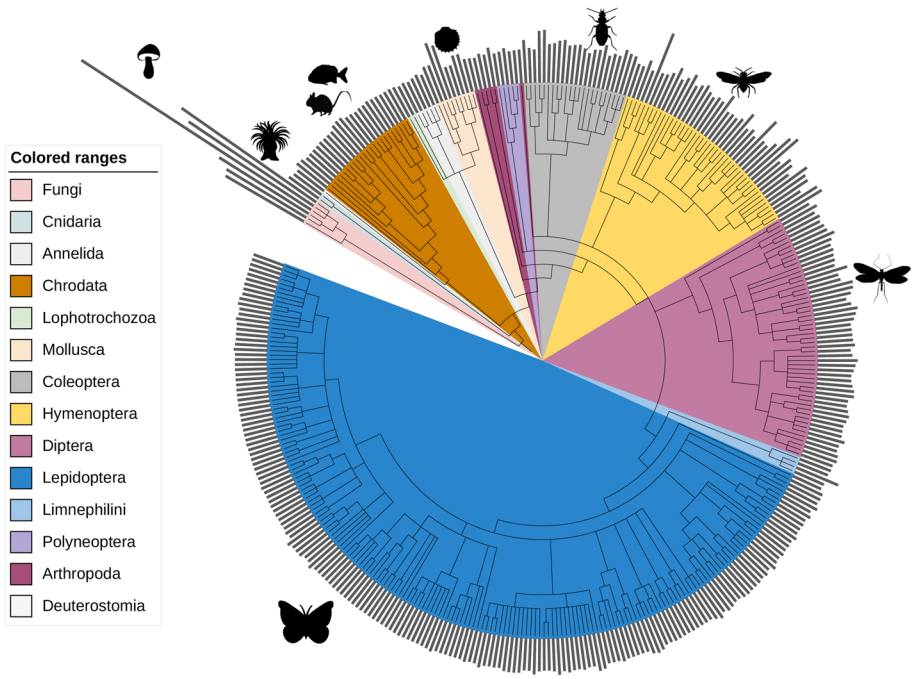

Fig. 2 Mitochondrial genomes assembled with MitoHiFi for DToL, ASG and VGP species. The 374 mitochondrial genomes assembled using MitoHiFi are shown. The three is derived from NCBI TaxonomyDB [20, 21], and was visualised in iToL [22]. The grey histograms represent the mitochondrial genome length for each species. Lepidoptera median length assembled was 15732 bp. For a high-resolution with species names and sizes visit https://github.com/marcelauliano/MitoHiFi/blob/master/docs/Figure_2_HR.pdf. Some species icons are from PhyloPics2.0

presents a comparison of genes annotated in the reference in relation to genes annotated in each potential contigs (including final) for a quick inspection of the annotation.

Using MitoHiFi for mitochondrial genomes from non-metazoan species and for plastid genomes

Metazoan mitochondrial genomes are generally relatively small and have limited gene content. An exception is found in Cnidaria, where gene content and size can be larger. For other lineages, where gene content and size can vary greatly even between closely related taxa, MitoHiFi offers alternative approaches. For the assembly of mitochondrial genomes from Viridiplantae (plants) the parameter *-a plant* should be called. This will prevent parse_blast.py from filtering out large contigs as plant mitochondrial genomes can greatly vary in size even between closely-related taxa [19]. Plastid genomes are similarly variable in size and content. For plastid genome assembly findMitoReference.py should be run with the flag -type chloroplast to find a closely-related plastid genome. While MitoFinder is used as default for annotation, if the user wishes to use MITOS for fungal genomes, the parameter *-a fungi* should be used. It's worth noting that MitoHiFi is not optimised to assemble plants or fungal organelles. Nevertheless, we successfully assembled fungal mitochondrial genomes from 43 to 133 kb (Fig. 2). Plant organellar genomes present two challenges: high levels

of heteroplasmy and presence of long repeat sections. Currently we recommend that plant organellar genome estimates generated by MitoHiFi should always be checked and manually finished.

Assembly of mitochondrial genomes at scale

MitoHiFi has been deployed to assemble mitochondrial genomes from species analysed by DToL (350 species to December 2022), the VGP (22 species), and ASG (2 species) from 39 orders of Metazoa and 6 Fungi species (Fig. 2). The assemblies have been submitted to the International Nucleotide Sequence (INSDC) databases (ENA,

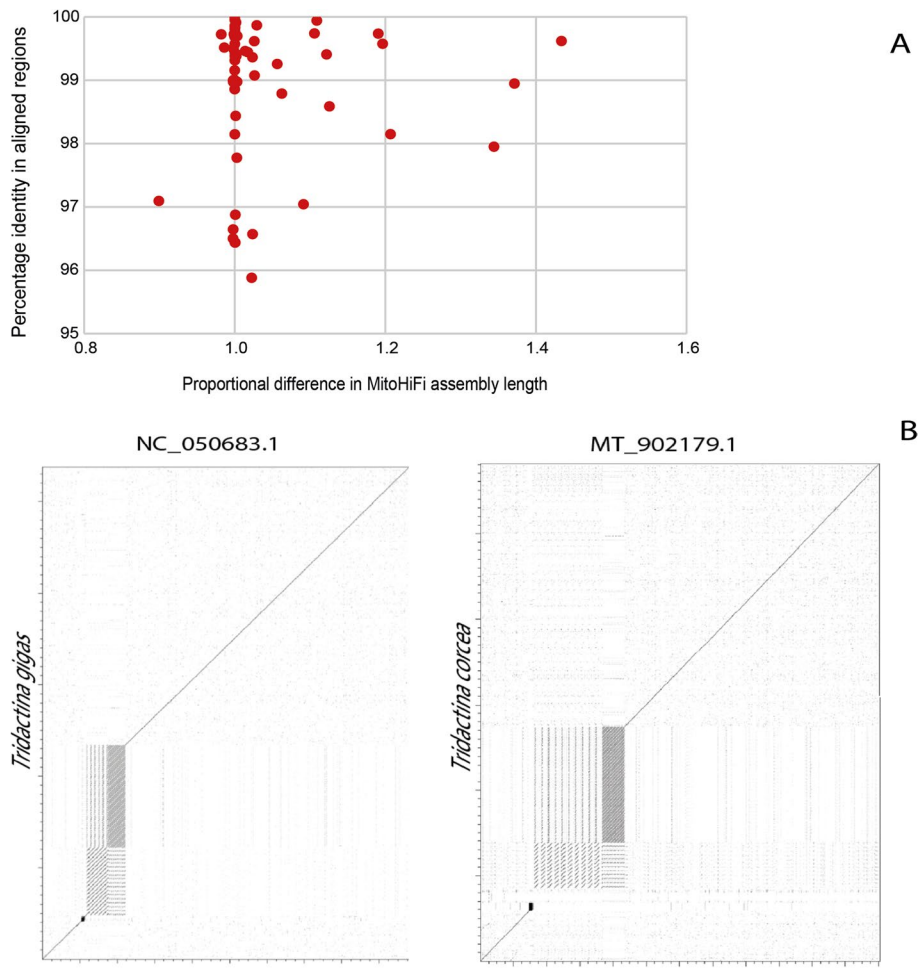

Fig. 3 MitoHiFi assemblies often include sequence absent from previous mitochondrial genome assemblies. **A** For 60 species (Additional file 2: Table S2), the new MitoHiFi assemblies were compared to sequences available for the same species in INSDC databases. For each pair the percent nucleotide identity of the aligned sequences and the proportional size of the MitoHiFi assembly were measured. **B** Comparison of MitoHiFi and previously published mitochondrial genomes of *Tridacna gigas* and *Tridacna crocea* was performed using dotter, and the results were visualised in dotplots. The dotplots show sequence conservation along the diagonal line, while deviations from the diagonal indicate variations such as repeat copies and missing sequences. The MitoHiFi assemblies include multiple copies of repeat sequences, supported by individual long reads, that are likely collapsed in the previously published sequences

NCBI, DDBJ) or to the VGP GenomeArk and a list of accession numbers can be found in Additional file 1: Table S1.

MitoHiFi identifies additional repeat content in previously published mitogenomes

Sixty of the species assembled here with MitoHiFi had mitochondrial genomes previously sequenced and submitted to the INSDC databases (Additional file 2: Table S2, Fig. 3A). To assess the performance of MitoHiFi we compared the new MitoHiFi assemblies to these published assemblies (Fig. 3). The majority of species had assemblies of a similar size and with a nucleotide identity above 96%. Only three MitoHiFi assemblies were smaller than the previously assembled reference. The *Flammulina velutipes* (Fungi, Basidiomycota, Agaricomycetes) MitoHiFi assembly was 8927 bp smaller than the database reference JN190940.1. Apart from that 8 Kb portion, both assemblies have the same t-RNAS, rRNAs and protein coding genes annotated and a nucleotide similarity of 97%. We investigated the 8 kb sequence unique to JN190940.1 through mapping our Pacbio HiFi reads to it, and found no evidence of reads spanning that sequence. We also compared this additional sequence using BLAST [15] against the NCBI nucleotide database and found no evidence of it being present in other fungal mitochondrial genomes. The additional segment in JN190940.1 contains no essential conserved genes. The difference between the MitoHiFi assembly and the published one could be technical due to accidental inclusion of a segment due to the technology used in cloning and PCR or true biological variation between isolates.

Sixteen mitochondrial genomes assembled by MitoHiFi were larger than their previous version. Alignment and gene content investigations show that repeats are the cause of longer MitoHiFi assemblies (Fig. 3B, Additional file 3: Figure S1). A dotplot and a nucleotide inspection of *Tridacna gigas* (Metazoa, Mollusca, Bivalvia) assemblies showed three types of repeats that are present in both assemblies, but those repeats have more copies present in the assembly built by MitoHiFi and are supported by HiFi data. The same is true for *Tridacna crocea*. Other dotplots are shown in Additional file 3: Figure S1. For these 16 assemblies where MitoHiFi genomes are larger than the previous reference, the percentage of nucleotide identity in the aligned portions is always above 95% making it unlikely that MitoHiFi is including NUMTS in the final assembled sequence.

Discussion

In the emerging era of reference genomics, where chromosomally-complete assemblies are sought, PacBio HiFi reads are a key data type [7]. Because of their length and quality, HiFi data support rapid and robust assembly of any genome, including those of organelles. Existing organelle genome assembly toolkits ([14, 16, 23]) were designed for short reads, and deal efficiently with their limitations. However short reads cannot resolve tandem and other repeats and may not be able to distinguish mitochondrial insertions in the nuclear genome from true organellar sequence. We have built MitoHiFi to best make use of long, accurate HiFi reads. MitoHiFi analysis can be initiated from raw HiFi data or from reads pre-assembled into primary contigs, for example using Hifiasm [9] or HiCanu [8]. These assemblers are not designed to assemble organellar genomes, which

will be present at high relative coverage compared to the target nuclear genome, and usually emit assembly estimates that include a multiplicity of subgenomes and misassemblies or reject the mitochondria as being a multicopy repeat.

MitoHiFi performs best when using raw read data, as this allows the tool to effectively avoid reads derived from NUMTs (and equivalent plastid transfers, NUPTs) and to identify likely assembled NUMT/NUPT loci using length cutoffs. MitoHiFi uses a relatively generous length cutoff because historical use of short read or PCR-based methods to retrieve organellar genomes may have resulted in collapse of repeat regions and thus underestimation of true genome length. In our analyses we identified many examples where the previously published assemblies differed from MitoHiFi assemblies because of a relative lack of repeat sequences in the published assemblies (Fig. 3B, Additional file 3: Figure S1). Inspection of MitoHiFi outputs shows that these filters are effective in avoiding NUMTs/NUPTs (see Additional file 4: Figure S2). We note that the toolkit also emits these "other" assemblies, and would advise that if the assembly generated for a species does not fit within expected parameters (presenting a small number of genes, very large repeats present or exhibiting many gene frameshifts), further manual validation should be performed.

The choice of the final reference assembly is based on comparison to previously assembled genomes available in public databases. The default parameter for coverage is set to 50% (*i.e.*, at least 50% of the sequence of the contig has to be present in the BLAST match with the closely-related species mitochondrial genome), but this may be raised as high as 90% where the species being analysed is part of a richly-sampled group with highly constrained genome content and structure, such as Vertebrata. However, for other taxa such as Hymenoptera and Mollusca, the 50% match length cutoff is required (and indeed may be too strict) because these clades have mitochondrial genomes with highly variable gene orders [24] and a diversity of repeat lengths and copy numbers. For taxa with known variability, it is recommended that the pipeline is run using multiple different references (using the findMitoReference.py -n flag) and exploring different match length proportions.

MitoHiFi incorporates two tools for protein coding gene finding and annotation, MitoFinder (the default) [16] and MITOS (flag *–mitos*) [17]. Both annotators use mitfi [18] particularly for organellar genomes that contain genes not in the core set, and where intron splicing of RNA and protein coding genes is common. Because of this, annotations produced by MitoHiFi should be checked manually before submission to databases.

Although MitoHiFi was optimised to assemble metazoan mitochondrial genomes, it has performed well with Fungi where the mitochondrial genomes assembled ranged from 43 to 133 kb (Fig. 2). While MitoHiFi can also assemble plant mitochondria and plastid genomes, the outcomes may vary. This is due to the significant variations in size, gene content, and repeat composition observed in plant mitochondrial and plastid genomes. The final output completeness of these genomes will depend on the complexity of the genome assembly graph.

MitoHiFi successfully generated a complete mitochondrial genome for *Climacium dendroides*, which is approximately 100 kb in size and contains 64 genes (data not shown). However, MitoHiFi encountered challenges in producing a complete version of *C. dendroides* plastid genome. The plastid genome is characterised by a large inverted repeat region that includes small and large subunit ribosomal RNA genes, and occasionally other loci [25]. This repeat region can exceed the length of a single HiFi read, making it difficult to fully resolve the orientation of the non-repeat segments. For the *C. dendroides* plastid run, MitoHiFi assembled 85 contigs and selected a final representative which was 75 kb-long and contained 78 genes, representing only a partial portion of the plastid (data not shown).

To address these limitations and achieve a comprehensive structure of mitochondrial and chloroplast genomes, we are developing algorithms that employ a new assembly logic and large k-mer sizes that can span the repeats within these organelles. It is important to note that plastid genomes typically exist as two isomers generated by inversion across the repeats [26, 27]. Linear representations of plastid assemblies must select one of the two possible paths through the repeats. Graph-based representations of these genomes would better depict the actual structures present. Plant mitochondrial genomes also exhibit multiple isomers and can contain various repeat segments of different lengths. Once again, HiFi reads may not span these repeats, necessitating a graph-based approach for a complete representation of the genome. We are actively working on methods to employ graph representations in plastid genome assembly to accurately resolve the structures of different organellar genome populations found in plants.

Conclusions

MitoHiFi efficiently assembles and annotates mitochondrial genomes using PacBio HiFi reads. We have used it to assemble 374 mitochondrial genomes from major biodiversity genomics projects. MitoHiFi is openly available on GitHub as code and as a Docker container under the MIT licence.

Methods

Species analysed

The data analysed here were generated as part of the DToL, VGP and ASG projects. Each of the target species was sequenced to ~ 25 fold coverage in PacBio HiFi reads of the nuclear genome following standard protocols. The nuclear genomes have been assembled using HiFi and Illumina Hi-C data, and will be reported elsewhere. Additional file 1: Table S1 presents the INSDC genome ascension numbers, BioSample IDs, SRA reads IDs, and/or VGP links for the genome sequences and reads of each species analysed.

MitoHiFi pipeline

MitoHiFi was written in python3. It incorporates other tools as shown in Fig. 1. MitoHiFi was run from reads (-r) with default parameters for all species presented here apart from some fungi species. For *Mucor piriformis*, *Flammulina velutipes*, *Pleurotus ostreatus* and *Agaricus bisporus*, reads were first assembled with MBG ([28] parameters: -k 1001 -w 250 -a 5 -u 150) to obtain contigs that were then input to MitoHiFi with the -c flag. The GitHub page can be accessed for detailed documentation on all final and

intermediate outputs by MitoHiFi. To reproduce all the assemblies generated by this study follow instructions on our GitHub page (MitoHiFi/scripts_paper) https://github.com/marcelauliano/MitoHiFi/tree/master/scripts_paper.

Comparative analyses

Sixty mitochondrial genomes assembled by MitoHiFi (Additional file 2: Table S2) were compared with their previous reference found on INSDC with FastANI [29] for nucleotide identity and dotter [30] (Fig. 3B and Additional file 3: Figure S1).

Graphical representations

The tree in Fig. 2 was derived from the TaxonomyDB tree using ETE based on the NCBI TaxIDs for each species. The figure was generated in ITOL [22]. Further modification, including addition of species silhouettes from PhyloPics2.0 and others was performed in Adobe Illustrator.

Data and software accessibility

The Additional file 1: Table S1 presents the INSDC accession numbers of the genome sequences of all the species analysed. The mitochondrial genomes have been submitted along with the nuclear genome sequences. MitoHiFi is available on GitHub (https://github.com/marcelauliano/MitoHiFi) and as a Docker container (ghcr.io/marcelauliano/mitohifi:master) under the MIT licence.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05385-y.

Additional file 1. INSDC or GenomeArk accession numbers of the genome sequences of all the species analysed and presented in Fig. 2.

Additional file 2. List of species that were assembled by MitoHiFi and that also had previous mitogenomes available on online databases.

Additional file 3. **Figure 1**. Dotplots of MitoHiFi mitogenomes (x axis) with their pre-existing mitogenome assemblies (y axis) for each species.

Additional file 4. **Figure 2**. Screenshots of IGV plots of reads mapped back to *Andrea bucephala* mitogenome before and after NUMTs reads were removed.

Acknowledgements

We acknowledge the huge effort of many colleagues at each stage in the generation of the genomes by the Darwin Tree of Life Project, including species sampling and processing, DNA extraction and sequencing, genome assembly and curation, and database construction. We thank Sujai Kumar and Martin Jones for their valuable help in the improvement of our Docker images and GitHub page. Mark Blaxter[1], Nova Mieszkowska[8,9], Neil Hall[10], Peter Holland[11], Richard Durbin[7,1], Thomas Richards[11], Paul Kersey[12], Peter Hollingsworth[13], Willie Wilson[8,14], Alex Twyford[13,15], Ester Gaya[12], Mara Lawniczak[1], Owen Lewis[11], Gavin Broad[16], Fergal Martin[17], Michelle Hart[13], Ian Barnes[16] . [8]Marine Biological Association of the United Kingdom, Citadel Hill, Plymouth PL1 2PB, UK, [9]University of Liverpool, Liverpool L69 3BX, UK, [10]Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK, [11]Department of Biology, University of Oxford, Mansfield Road, Oxford OX1 3SZ, UK, [12]Royal Botanic Gardens, Kew, Richmond, London TW9 3AE, UK, [13]Royal Botanic Garden Edinburgh, Edinburgh EH3 5LR, [14]University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK, [15]Institute of Ecology and Evolution, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, [16]Natural History Museum, Cromwell Road, London SW7 5BD, UK, [17]EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

Author contributions

MitoHiFi was idealised and written by M. U. S. and J. G. R. N. F. Authors G. F., E. W. M., R. D., M. B., and S. A. M. contributed ideas for software development. M. U. S., J. G. R. N. F., K. K., G. F., L. A., J. T., S. A. M. ran and/or submitted assemblies and gave software feedback. M. U. S. drafted the manuscript. All authors contributed to the final manuscript and approved it.

Funding

This research was funded by the Wellcome Trust Darwin Tree of Life Discretionary Award (218328) and the Wellcome Sanger Core Award (220540/Z/20/A).

Availability of data and materials
The datasets generated and analysed during the current study are available in the INSDC repositories. Additional file 1: Table S1 presents all the accession numbers of the genome sequences of all the species analysed. MitoHiFi is available on GitHub (https://github.com/marcelauliano/MitoHiFi) and as a Docker container (ghcr.io/marcelauliano/mitohifi:master) under the MIT licence.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 3 March 2023 Accepted: 13 June 2023
Published online: 18 July 2023

References

1. Richards S. It's more than stamp collecting: how genome sequencing can unify biological research. Trends Genet TIG. 2015;31:411–21.
2. Blaxter M, Archibald JM, Childers AK, Coddington JA, Crandall KA, Di Palma F, et al. Why sequence all eukaryotes? Proc Natl Acad Sci. 2022;119:2115636118.
3. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth biogenome project: sequencing life for the future of life. Proc Natl Acad Sci USA. 2018;115:4325–33.
4. Blaxter M, Mieszkowska N, Di Palma F, Holland P, Durbin R, et al. Sequence locally, think globally: the darwin tree of life project. Proc Natl Acad Sci. 2022;119:e2115642118.
5. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:737–46.
6. Aquatic symbiosis genomics project - Wellcome Sanger Institute [Internet]. [cited 2022 Sep 6]. Available from: https://www.sanger.ac.uk/collaboration/aquatic-symbiosis-genomics-project/
7. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62.
8. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 2020;30(9):1291–305.
9. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5.
10. Lane N, Martin W. The energetics of genome complexity. Nature. 2010;467:929–34.
11. Gray MW, Burger G, Lang BF. Mitochondrial evolution. Science. 1999;283:1476–81.
12. Wallace DC. Mitochondrial DNA in evolution and disease. Nature. 2016;535:498–500.
13. Formaggioni A, Luchetti A, Plazzi F. Mitochondrial genomic landscape: a portrait of the mitochondrial genome 40 years after the first complete sequence. Life. 2021;11:663.
14. Formenti G, Rhie A, Balacco J, Haase B, Mountcastle J, Fedrigo O, et al. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. Genome Biol. 2021;22:120.
15. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinform. 2009;10:421.
16. Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Mol Ecol Resour. 2020;20:892–905.
17. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 2013;69:313–9.
18. Jühling F, Pütz J, Bernt M, Donath A, Middendorf M, Florentz C, et al. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. Nucleic Acids Res. 2012;40:2833–45.
19. Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, Christensen AC. The alternative reality of plant mitochondrial DNA: one ring does not rule them all. PLoS Genet. 2019;15(8):e1008373.
20. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI taxonomy: a comprehensive update on curation, resources and tools. Database. 2020;2020:baaa062.
21. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2019;47:D94–9.
22. Letunic I, Bork P. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47:W256–9.
23. Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21:241.
24. Malkócs T, Viricel A, Becquet V, Evin L, Dubillot E, Pante E. Complex mitogenomic rearrangements within the Pectinidae (*Mollusca: Bivalvia*). BMC Ecol Evol. 2022;22:29.

25. Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011;76:273–97.
26. Palmer JD. Chloroplast DNA exists in two orientations. Nature. 1983;301:92–3.
27. Stein DB, Palmer JD, Thompson WF. Structural evolution and flip-flop recombination of chloroplast DNA in the fern genus *Osmunda*. Curr Genet. 1986;10:835–41.
28. Rautiainen M, Marschall T. MBG: minimizer-based sparse de Bruijn graph construction. Bioinformatics. 2021;37(16):2476–8.
29. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114.
30. Sonnhammer ELL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene. 1995;167:GC10.

Publisher's Note