Title:

MitoGeneExtractor: Efficient extraction of mitochondrial genes from next generation sequencing libraries

Running title:

MitoGeneExtractor: mining mitochondrial genes

Authors:

Brasseur*, Marie V., Astrin, Jonas J., Geiger, Matthias F., Mayer, Christoph

Affiliation:

Leibniz Institute for the Analysis of Biodiversity Change, Zoological Research Museum A. Koenig, Adenauerallee 127, 53113 Bonn, Germany

*Correspondence:

Marie V. Brasseur, E-mail: m.brasseur@leibniz-lib.de

Abstract:

Mitochondrial DNA sequences (mtDNA) are often found as byproduct in hybrid enrichment data sets originally created to capture anchored hybrid enrichment (AHE) or ultra-conserved element (UCE) nuclear loci. The mtDNA sequences in these data sets are currently rarely used, even though mitochondrial genes such as COI, ND5, CytB, and 16S are of general interest and often not yet known and deposited in public databases. We developed MitoGeneExtractor to extract mitochondrial genes of interest from genomic libraries. Gene sequences are reconstructed through multiple sequence alignments of sequencing reads to an amino acid reference. We applied MitoGeneExtractor to recently published data created for UCE enrichment and were able to extract complete or nearly complete COI and ND5 sequences for a large proportion of the sequencing libraries. MitoGeneExtractor can be used to extract mitochondrial protein coding genes from a wide range of next generation sequencing data sets.

Key words: Data mining, DNA barcoding, data re-use, mitochondrial genes, COI, ND5

Introduction:

Next generation sequencing (NGS) and high throughput sequencing have become standard tools in biological research and enable the generation of unprecedented amounts of sequencing data (Reuter, Spacek, & Snyder, 2015). Rapidly evolving sequencing technologies and relatively low sequencing costs of ~1,000 USD per genome (30 X coverage on Illumina platforms; genome.gov/sequencingcostsdata; accessed on 03.05.2021) allow researchers to investigate biological processes based not only on one or a few genes. Instead, millions of sequencing reads are generated per run in order to analyze thousands of loci or whole genomes, ranging from individual specimens to entire biological communities. The continuously dropping costs promise the growing exploitation of DNA sequence information in an application-oriented context such as medicine (Lecuit & Eloit, 2015), biomonitoring (Baird & Hajibabaei, 2012) or species conservation (Allendorf, Hohenlohe, & Luikart, 2010). Despite the clear trend towards increased cost-efficiency, generating and analyzing high-throughput sequencing data is still resource demanding with regard to laboratory and computational costs, time and skills.

NGS data potentially harbor much more information than is exploited over the course of the initial experiment. Although it is highly important to incorporate genomic complexity in biological studies, researchers might be particularly interested in specific genes. One example is the mitochondrial cytochrome oxidase I subunit (COI) gene, which is the most commonly used molecular marker in animal species identification (Hebert, Cywinska, Ball, & deWaard, 2003) and related fields, despite some limitations (Eberle, Ahrens, Mayer, Niehuis, & Misof, 2020). Fragments of this gene are further used to assess biotic communities in DNA metabarcoding approaches, using either bulk samples of e.g. trapped invertebrates or free environmental DNA (eDNA) from samples such as water or soil (Cordier et al., 2021; Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012). Organellar DNA sequences are generally present in gDNA sequencing libraries due their high abundances in the cell and therefore in gDNA extracts (Bogenhagen & Clayton, 1974; Samuels et al., 2013) and can be found as byproduct in sequence capture/enrichment data sets (Allio et al., 2020; Amaral et al., 2015; Picardi & Pesole, 2012). Often, these organelle related reads are discarded or ignored during bioinformatic processing, potentially wasting this source of data. Studies that have extracted mitochondrial sequences from ultra-conserved genomic loci enrichment (UCE) data or anchored hybrid enrichment data are rare (e.g. Meiklejohn et al. 2014; Pie et al. 2017; Wang et al. 2017; Caparroz et al. 2018), probably due to the lack of conveniently applicable tools.

Here, we describe an approach to make use of this sequencing byproduct in order to extend the utility of the constantly growing amount of sequencing data beyond the initial study purpose. By aligning DNA sequencing reads to an amino acid reference sequence (e.g. the COI gene), we are able to reconstruct *in silico* the corresponding COI or other mitochondrial sequences, if the mitogenome is sufficiently represented within the genomic read pool. This is especially important for the generation of sequence information in non-model organisms or taxonomic groups in which sample access can be difficult or even impossible, such as rare or extinct taxa. As such, these sources constitute an important but hitherto untapped contribution to the global DNA barcode databases like the Barcode of Life Data System (BOLD) (Ratnasingham & Hebert, 2007).

We selected mitochondrial genes as a case study due to their significance in biological research, because of their usually good representation in sequencing libraries, and since they are typically well conserved and indels are not expected within taxonomic groups.

Several tools such as Phyluce (Faircloth, 2016), MITObim (Hahn, Bachmann, & Chevreux, 2013), Trimitomics (Plese et al., 2019), MitoZ (Meng, Li, Yang, & Liu, 2019) or Mitofinder (Allio et al., 2020) already exist, which aim to reconstruct and extract *in silico* mitochondrial sequences or even whole mitogenomes from genomic read pools. All the mentioned tools are based on assembly results: for example, MITObim aims to reconstruct whole mitogenomes from genomic NGS data sets, relying on the genome assembler MIRA (Chevreux, Wetter, & Suhai, 1999). Based on an iterative selection of reads matching a current intermediate sequence and an assembly of these currently selected reads, MITObim tries to reconstruct mitochondrial genomic regions starting from a seed sequence (Hahn et al., 2013).

The Phyluce pipeline was originally designed to extract UCEs and to subsequently perform phylogenetic analyses with these loci (Faircloth, 2016). Phyluce uses the output of assembly tools such as Trinity (Grabherr et al., 2011) by aligning the produced contigs to a bait (or oligonucleotide probe) reference sequence. Initially designed for standard enrichment baits for UCE loci, Phyluce can in principle be used to extract other loci as well, dependent on the input bait reference. Trimitomics assembles mitochondrial genomes from transcriptomic data (Plese et al., 2019) and MitoFinder is designed to assemble simultaneously both UCE and complementary mtDNA from raw UCE capture libraries (Allio et al., 2020) by using the meta-assembler metaSPAdes (Nurk, Meleshko, Korobeynikov, & Pevzner, 2017) or IDBA (Peng, Leung, Yiu, & Chin, 2010).

Assembly guided sequence reconstruction approaches have several drawbacks: i.) assemblies are highly parameter dependent, ii.) the quality of assemblies quickly drops if read coverage values are low (see results of this study). iii.) An assembly process is always computationally intensive, especially for large data sets. This can prevent or at least hamper the fast and efficient sequence reconstruction for hundreds or thousands of individuals/taxa. iv.) Existing approaches rely on reference sequences from a closely related species or at least seeding sequences such as the barcode region. Finally, (v) in the presence of NUMTs (nuclear mitochondrial DNA), a sequence variation is introduced which can prevent a successful assembly of the reads. In preliminary analyses, we have found that MITObim suffers from this problem. Potentially, implementing another assembler than MIRA within MITObim could produce better results for multi allelic data and uneven read coverage. Altogether, assemblers require a substantial amount of sequence reads for being able to reconstruct the target region, particularly in the presence of only partially similar sequences such as NUMTs.

The here presented workflow does not require the assembly of reads but instead is based on an alignment of the DNA sequencing reads to an amino acid reference. For this purpose, we developed the tool MitoGeneExtractor, which utilizes the program Exonerate (ebi.ac.uk/about/vertebrate-genomics/software/exonerate) to align DNA reads to a provided amino acid reference (Figure 1). MitoGeneExtractor uses the Exonerate output (i.e. vulgar file format, containing information about the start/end position of the read alignment in the reference, whether the forward or the reverse complement orientation aligned and an alignment score) to generate a multiple sequence alignment (MSA) of the reads. Due to the degeneracy of the genetic code, this allows a considerable DNA sequence variation of reads that can successfully be aligned to the reference. This makes it possible to use the same amino acid reference for a broad spectrum of taxa in particular when mining genes from the conserved mitochondrial genome. The subsequently resulting MSAs can be used to reconstruct a consensus gene sequence for the individual sample. When implemented in a data analysis management system such as Snakemake (Köster & Rahmann, 2012), it is possible to analyze and extract sequence information from hundreds or even thousands of genomic DNA data sets automatically and simultaneously.

We tested our approach with a large avian data set from Harvey et al. (2020), which upon publication had been used for a comprehensive phylogenomic analysis of songbirds (*Passeriformes*) in a tropical biodiversity hotspot. With the presented approach, we were able to reconstruct sequence information (≥ 90 % of the sequence) for two mitochondrial genes, the cytochrome *c* oxidase 1 (COI) and NADH dehydrogenase subunit 5 (ND5) gene for 85 % and 80 % of the samples, respectively. We compared MitoGeneExtractor with MitoFinder (Allio et al., 2020) regarding the sequence reconstruction success and computational time. Further, we evaluated the taxonomic assignment based on our reconstructed sequences obtained with MitoGeneExtractor. As no full-length COI sequence information was present for any of the bird species in NCBI, we evaluated our approach via the comparison of our reconstructed sequences with COI barcodes from BOLD (Ratnasingham & Hebert, 2007) and compared our taxonomic assignment inferred via the reconstructed barcodes with the taxonomic assignment of the initial study from Harvey et al. (2020).

Material and Methods:

During the initial study of Harvey et al. (2020), the authors generated target enrichment data of UCEs and exons for 1,993 individuals. Their final data set comprised 1,287 neotropical bird species, represented by 1-38 individuals per species. We used this data set to attempt the *in silico* reconstruction of complete COI and ND5 sequences for all of the 1,993 individuals. The DNA extracts were obtained from genomic resource collections at natural history museums and from field excursions. gDNA extracts were enriched for UCEs and conserved exons and sequenced on Illumina HiSeq platforms (Harvey et al., 2020).

Obtaining and pre-processing of data:

Raw sequence data was downloaded from the NCBI Sequence Read Archive PRJNA655842 using prefetch from the SRA-toolkit v 2.11.2 (http://ncbi.github.io/sra-tools/). The sra files were transformed to the fastq format with fastq-dump (SRA-toolkit). We specified the options --split-e in order to extract the data in separate files, if paired-end read data was generated, and --readids to retain unique read sequence IDs. Paired-end read information cannot be exploited with Exonerate because each read is individually aligned to the reference, either in forward or reverse complement orientation. Therefore, we concatenated paired-end libraries and treated them as single-end libraries. This artificially doubled read numbers in paired-end libraries (Table S1, S2) but allowed to retain one read of a read pair, when the other read was discarded during quality trimming. Raw sequencing reads were quality trimmed using the cutadapt v 1.18 (Martin, 2011) wrapper script TrimGalore! v 0.0.6 (https://github.com/FelixKrueger/TrimGalore) with auto-detection of Illumina adapters and a quality cut-off at Phred < 20. Fastq files were transformed to the fasta format using bash shell commands. Data transformation and quality processing was conducted within a Snakemake workflow in order to improve reproducibility of data analysis.

Generation of reference protein sequences:

The NCBI protein database (https://www.ncbi.nlm.nih.gov/protein/) was searched for full length sequence information of the COI and ND5 genes for all passerine birds (*Passeriformes*). All sequences were downloaded and one sequence per genus was retained. The sequences were visually inspected with AliView v 1.26 (Larsson, 2014) and irregular sequences (which corrupted the alignment) were removed. Then, the sequences were aligned (385 for COI and 331 for ND5) using the MUSCLE algorithm (Edgar, 2004) and the resulting consensus amino acid sequences were used as reference for the MSAs.

Alignment of reads – MitoGeneExtractor

We developed MitoGeneExtractor which creates consensus gene sequences in the following three steps. In step one, MitoGeneExtractor calls Exonerate, which needs to be installed independently, to align the amino acid reference sequence to the input (i.e. quality filtered) DNA reads (both input files are expected to be in fasta format). Two important Exonerate command line parameters, which alter the alignment settings, can be specified when calling MitoGeneExtractor and are passed to Exonerate: the genetic code used for translating the reads prior to the alignment and the frameshift penalty. Further, the user can specifiy the minimum alignment score threshold used by Exonerate, if desired. In step two, MitoGeneExtractor uses the Exonerate output in vulgar format (see Exonerate manual for details) to create an alignment of all input reads. Parameters can be specified to control e.g.

the minimum coverage and the minimum alignment score relative to the read length to control the alignment quality. Finally in step three, MitoGeneExtractor determines consensus sequences for the gene of interest and provides the user with an alignment fasta file and the desired consensus sequence as final output.

When calling MitoGeneExtractor, the most time-consuming step is the generation of the Exonerate vulgar files (although this only takes on the order of 30 seconds for 1 million reads using a single core on a modern laptop). For existing vulgar files, the MSAs are generated by MitoGeneExtractor in a few seconds, allowing a fast re-analysis with adjusted parameters once the vulgar files are already produced. Exonerate writes alignment information to the vulgar file only for those reads that could successfully be aligned to the target gene. From this information MitoGeneExtractor determines not only the MSA of successfully aligned reads, but also the corresponding consensus sequence. The MSA of reads can be used for subsequent data exploration and analyses.

For this study, we installed Exonerate version 2.2.4 and called MitoGeneExtractor with the options -t 0.5 (consensus threshold; i.e. an unambiguous nucleotide in the consensus sequence is inferred only if it is supported by 50% of the nucleotides at this site), -r 1 (minimum relative alignment score; alignment score from Exonerate divided by the length of the alignment) and -n 0. Setting the -n parameter to a value greater than 0 would instruct MitoGeneExtractor to include bases of the read beyond the alignment region Exonerate has found. Less conservative parameter combinations were tested as well, and the resulting statistics can be found in supplementary tables S1-S4. Depending on the analyzed taxon, the genetic code (parameter -C) used by Exonerate needs to be adjusted. The genetic code is supplied by the corresponding integer, according to the synopsis from Osawa, Jukes, Watanabe, & Muto (1992) and Jukes & Osawa (1993), also adapted by NCBI (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi).

Evaluation:

We first evaluated the general sequence reconstruction success in terms of recovered nucleotides for each of the 1,993 samples. Based on the data set of Harvey et al., (2020), we compared the COI and ND5 consensus sequences mined with MitoGeneExtractor with the sequences mined with MitoFinder v. 1.4 (Allio et al., 2020) regarding number of reconstructed sequences, their completeness and computation time. MitoFinder assemblies were generated by the assembly tool IDBA (Peng et al., 2010) in paired-end mode (except for 41 single-end libraries), using the mitogenome of *Eremophila alpestris* (NCBI PRJNA636471, downloaded 09.09.2021) as reference and the vertebrate mitochondrial code (-o 2). To assess the run time of both tools, a data subset (n = 100) was re-analyzed, including only samples which were known to perform well (i.e. a complete coding region of COI was reconstructed with both tools). Analyses were run on a Linux based HPC server and 10 cores were provided for each program. MitoFinder samples were assembled using 10 GB of RAM (-m 10) per sample.

The nucleotide recovery in each reconstructed gene sequence obtained with both tools was visualized with the python3 (Van Rossum & Drake, 2009) package seaborn (Waskom, 2021). The violin plots show the estimated kernel density curve of the data distribution (bandwidth scale factor=0.04).

Nucleotide recoveries of at least 90 % of the full length of the corresponding gene were treated as a successfully reconstructed gene sequence. To test whether the reconstructed COI sequences can be used to correctly identify the corresponding species, we queried the sequences against the NCBI nucleotide database. Since no full-length sequence information was available for COI and ND5 for any of the corresponding species, (NCBI nucleotide database accessed on 03.05.2021), a direct comparison of the full length sequences was not possible. Therefore, we extracted the 658 bp barcode region from the reconstructed COI sequences (nucleotide positions 45 - 702, flanked by the primer pair of Folmer, Black, Hoeh, Lutz, & Vrijenhoek, 1994) and compared the barcode sequences to entries in BOLD (Ratnasingham & Hebert, 2007). The *in silico* generated barcodes were taxonomically assigned using BOLDigger v.1.2.2 (Buchner & Leese, 2020) and our inferred taxonomic assignment was compared with the taxonomic assignment from Harvey et al. (2020). We included only COI barcodes without any gaps in the barcode region (= 1,611) although for species identification purposes also shorter barcodes or slightly incomplete sequences can be sufficient. For those samples in which the best BOLD hit did not match the species assignment of Harvey et al. (2020), we used the R package ‘bold’ (Chamberlain, 2021) to check whether COI sequence information for the species was present at all in BOLD.

Results:

Data extraction, quality filtering:

When used as single-end libraries (i.e. paired-end libraries were concatenated), the 1,993 libraries downloaded from NCBI yielded in total 14,215,651,594 reads, with 6,431,834 (median) per library. Read numbers ranged from 2,984 to 32,059,194, presenting a very heterogeneous test data set. After quality trimming, 6,298,529 (median) per sample were retained (Table S1; for individual sample statistics, see Table S2).

Cytochrome c *oxidase subunit one:*

Per sample, 2,927 (median) reads were successfully aligned to the COI reference with MitoGeneExtractor. Mean base coverage of the alignments (normalized by gene length) ranged from 0 to 16,644 nucleotides per position. The amino acid sequences of avian COI typically comprise 517 amino acids, resulting in 1,551 nucleotides, including the stop codon. If a sequence segment is not covered by reads, MitoGeneExtractor inserts gaps in the consensus sequence. In case nucleotides cannot be inferred unambiguously according to the consensus threshold (here: 50 %), Ns are inserted. We first evaluated the COI sequences based on the number of recovered nucleotides. From 1,993 analyzed samples, we reconstructed complete full-length COI sequences for 621 specimens (= 31.2 %). In total, we were able to generate 1,682 COI sequences with a base recovery of at least 90 % of the full gene length (Figure 2). First evidence for correctly reconstructed sequences with exon character is the absence of stop codons within the open reading frame (ORF). We detected 26 stop codons within the 1,993 sequences that were not found at the 3' end of the reconstructed sequences. Only 7 samples failed completely (= 0.35 %) and 158 COI sequences showed a poor base recovery of lower than 60 %. We extracted 1,611 full length COI barcodes from the COI gene sequences (Figure 2). All reconstructed COI consensus sequences can be found in supplementary file 2.

NADH dehydrogenase subunit five:

Per sample, 3,277 (median) reads aligned to the ND5 reference sequence, with a mean base coverage ranging between 0 and 12,694 bases per position. The avian coding region of the ND5 gene can have 605-608 amino acids, depending on the taxon of interest (e.g. Gao et al., 2021; Gao, Yin, & Zhu, 2021). Based on visual inspection of our results, we found that the reconstructed ND5 genes in our data set typically comprised 605 amino acids (including stop codon). Based on that, we recovered ND5 sequences with a base coverage of ≥ 90 % from 1,595 specimens (80 %). Despite this overall high sequence recovery success, only a small proportion, i.e. 174 sequences were recovered in full length, which is low compared to the COI gene (Figure 2). In total, 21 stop codons were detected in this data set, which were not located at the 3' end of the sequence. Only 13 samples (0.65 %) failed completely, i.e. no reads were mapped to consensus sequence and 200 reconstructed sequences showed a nucleotide recovery of less than 60 % of the complete gene sequence.

All reconstructed ND5 consensus sequences can be found in supplementary file 3.

Comparison to MitoFinder:

We compared the performance of MitoGeneExtractor with the existing tool MitoFinder (Allio et al., 2020), which was designed to assemble mitogenomes from NGS sequence data. For 981 samples, COI sequence information was assembled, from which 719 full length genes were reconstructed (36.08 %).

ND5 sequence information was assembled for 674 samples (33.8 %), but no gene was completely assembled (highest nucleotide recovery = 1,782 positions).

Although MitoFinder was able to reconstruct a slightly higher number of full-length COI genes, the overall assembly success was inferior to the sequence reconstruction with MitoGeneExtractor. For the majority of samples, no sequence information was recovered, contrasting the generally high reconstruction rate obtained with MitoGeneExtractor, which is consistent for both mitochondrial genes (Figure 3).

We selected 100 samples which showed full-length COI gene reconstruction with both tools, MitoGeneExtractor and MitoFinder and compared their computation times. MitoGeneExtractor reconstructed the 100 COI consensus sequences in 00:26:49 minutes on 10 cores including the time consuming Exonerate alignment step, whereas MitoFinder required 24:26:38 hours for the assembly and gene extraction on the same computer and using the same number of cores.

Evaluation based on taxonomic assignment:

From the 1,611 full length COI barcode sequences, we obtained a similar taxonomic assignment as first hit, i.e. the same bird species as in Harvey et al. (2020), for 1,031 individuals (64 %). The sequence identity to database entries of these barcodes ranged from 92 – 100 % similarity. 56 samples showed a similar taxonomic assignment to that in Harvey et al. (2020), which was not the first hit in BOLD but was present among the 20 best hits (Figure 4). From these 1,087 'correctly' assigned samples, 998 showed barcode identities of ≥ 97 %, which is a commonly applied threshold for species delimitation based on COI (Hebert et al., 2003). Eight samples were morphologically assigned by Harvey et al. (2020) only to genus level, preventing a taxonomic comparison on species level.

The 524 samples with a diverging taxonomic assignment and a sequence similarity of the first BOLD hit between 86.88 – 100 %, were mainly assigned to the same genus (427 individuals) as in Harvey et al. (2020). These 524 samples represented 443 taxa (439 species and four morphotaxa on genus level), for which in most cases (402 species), no COI sequence information was available in BOLD (Figure 4). Altogether, 97 reconstructed COI barcode sequences were not assigned to the same genus as in the original study. In most of these cases (77), the sequence similarity to database entries was below 95 %.

Notably, 15 individual samples (13 morphotaxa) were molecularly assigned to different genera with a high sequence similarity ranging from 97-100 %, such as the sample referred to as *Sclerurus caudacutus* in Harvey et al. (2020), which has a barcode identity of 97.36 % to *Poospiza lateralis* or the sample *Aphrastura spinicauda*, which has a barcode identity of 98.84 with *Poospiza thoracica* in BOLD. Interestingly, we found COI sequences in BOLD for four of these taxa (*Aphrastura spinicauda,* with 17 % divergence from the sequences in the study under the same name, *Lepidocolaptes falcinellus* (13 % divergence), *Phyllomyias virescens* (17 % divergence), or *Sclerurus caudacutus* (18 % divergence).

Discussion:

MitoGeneExtractor shows a high sensitivity and specificity when mining reads from NGS sequencing libraries. The success of sequence reconstruction mainly depends on the number of reads of the specific gene that are found in the NGS library. The decreased gene sequence reconstruction success of the ND5 gene with both tools, MitoGeneExtractor and MitoFinder, might be due to a lower number of reads for this locus in the sequencing library compared to the COI gene, which could be explained if the COI gene was enriched in the study of Harvey et al. (2020), even though this was not mentioned in the publication.

Comparing MitoGeneExtractor and MitoFinder, both reconstruct roughly the same number of full-length COI sequences. Including sequences with a nucleotide recovery of ≥ 90 %, MitoGeneExtractor reconstructed about twice as many COI sequences compared to MitoFinder. This pattern is consistent with the reconstruction success of the ND5 gene and highlights the potential drawback of assembly-guided sequence reconstruction: if the read coverage at a given position is too low, the extension of the reconstructed sequence is aborted, preventing the potential usage of reads, which cover subsequent positions of the gene. For specific genes of interest, MitoGeneExtractor is more efficient and faster than assembly guided tools such as MitoFinder, which aim to assemble complete mitogenomes. For the reconstruction of the same 100 COI sequences, MitoGeneExtractor was 54 x faster than MitoFinder.

Due to the high sequence identity between COI barcode reference database entries and our generated COI barcodes for most taxa, we conclude that our approach of sequence reconstruction works and that NGS read data can be exploited beyond the initial study purpose. The majority of detected stop codons occur at the end of the extracted gene sequences. If not, they should either result from sequencing errors or from incorporating reads from nuclear mitochondrial pseudogenes (NUMTs) (Gaziev & Shaikhaev, 2010). Again, read depth is crucial for a reliable reconstruction in assembly or MSA based approaches. In high coverage regions, these ‘wrong’ reads will be overruled by reads originating from the true loci and parameter settings might play a subordinate role (see Tables S3, S4). In gene regions which are covered only by a low number of reads, incorrect nucleotides have a higher likelihood of contributing to the resulting consensus sequences. MitoGeneExtractor has different options to handle these issues: using the coverage filter parameter --minSeqCoverageInAlignment demands a minimum number of reads for the computation of consensus sequences. More parameters exist which allow to find a trade-off between sensitivity and specificity, e.g. the – r and – n parameters (see the MitoGeneExtractor manual for details). Decreasing the specificity will improve base recovery but potentially introduces erroneous bases (Table S3, S4). Therefore, the increase of this parameter should be done only based on previous observations, followed by subsequent inspection of the alignments, and is generally not recommended. A certain trade off might be necessary since despite the high conservation of most mitochondrial genes, the first and last 30 bp of the full COI gene are often more variable in larger taxonomic groups.

Since Exonerate produces an alignment score based on the number of aligned bases of the read to the reference, reads which only partially overlap with the reference at the beginning or the end might be omitted because they have a position-dependent low alignment score. This can result in missing sequence information at the beginning/end of the reconstructed consensus sequence. In MitoGeneExtractor, the minimum alignment score (corrected for read length) can be adjusted with the parameter -r. If this value is decreased, reads with lower alignment score will be incorporated, which can result in more complete sequences. In our analyses of the ND5 gene, we were able to reconstruct more complete ND5 sequences when the minimum relative score -r was lowered from 1 to 0.8 (Table S4). Finally, the alignment files produced by MitoGeneExtractor should be visually inspected in uncertain cases in order to optimize the alignment quality thresholds. The default values provide a good but conservative setting for a heterogeneous range of data sets, but must be adjusted for specific cases, particularly when read coverage is expected to be low.

With the *in silico* reconstructed COI barcode sequences, 1,095 specimens were assigned to the corresponding morphotaxa, although the sequence similarity was in some cases clearly below 97 %, which is a commonly used as a species cutoff value (Hebert et al., 2003). Although the genetic divergence in the COI gene was shown to be generally low within avian species, higher intraspecific variability might be expected for tropical faunas which might contribute to the high genetic distances observed within our barcodes (Hebert, Stoeckle, Zemlak, & Francis, 2004). Diverging taxonomic assignments can further be the result of cryptic diversity or intraspecific divergence, which was reported for some of the taxa in Harvey et al. (2020). Furthermore, genetic differences in low coverage gene regions between the generated COI sequences and database entries might be the result of artefacts such as the incorporation of NUMTs reads (Gaziev & Shaikhaev, 2010), sequencing errors, or contaminations. In principle, difference with respect to a database can also be due to erroneous database entries. Interestingly, some specimens with different taxonomic assignments between Harvey et al. (2020) and our study, e.g. *Aphrastura spinicauda*, which was identified as *Poospiza thoracica*, have a very distinct morphological appearance, so that misidentification seems unlikely. Additionally, the overall divergence level might be inflated due to geographically biased sampling of taxa and their underrepresentation in databases (Kerr et al., 2009; Phillips, Gillis, & Hanner, 2019). Although birds are among the most intensively studied taxonomic groups, many of the here analyzed species are rare in the wild (most specimens were sampled at natural history collections), which explains the limited or even completely absent COI sequence information on NCBI/BOLD for some of the taxa.

Incomplete reference databases or wrongly assigned COI barcodes represent the major limitations of molecular species identification (Moritz & Cicero, 2004; Pentinsaari, Ratnasingham, Miller, & Hebert, 2020). One example for an ambiguous taxonomy found in data bases represent the *Phylloscartes* specimens, which were identified as *Pogonotriccus* individuals by us. According to The Global Biodiversity Information Facility (GBIF) (www.gbif.org, accessed 03.03.2022) the genus name *Pogonotriccus* is often synonymized with the genus *Phylloscartes* but both names are still in use although genetic differences were shown to be low (Tello, Moyle, Marchese, & Cracraft, 2009).

Finally, the morphological species delimitation is not always consistent with genetic divergence and evolutionary history of a single gene (Bilton, Turner, & Foster, 2017; Weigand et al., 2017). Especially the disproportionally high biodiversity from tropical regions is (taxonomically) underexplored (Balakrishnan, 2005; Dirzo & Raven, 2003) and needs ongoing research effort to be resolved.

This highlights the value of the opportunity to further exploit NGS data if researchers work with non-model organisms or taxa from which sample accession is difficult due to various reasons (e.g. ancient DNA, protected species, remote occurrence). Thus, nucleotide database managers may consider automatically running MitoGeneExtractor as a wrapper to routinely harvest genetic information, e.g. to add new barcode data to BOLD from the growing number of available NGS datasets, thus adding species entirely new to the database (as in the present example) or adding data that help in better monitoring genetic diversity at the population level. An important use case of MitoGeneExtractor should be the extraction of COI sequences from sequencing libraries in order to identify misidentifications of specimen and contaminations in the sequencing library. Sequencing projects should use MitoGeneExtractor routinely to exclude these potential problems.

One can imagine that the sequence information is even scarcer for genes other than COI, which are not commonly used as marker gene for population genetics or as molecular barcode for metazoans. In the case of the ND5 gene, only 50 full length DNA sequences for all passerine birds are deposited in the NCBI nucleotide database (accessed on 29.04.2021) from which 15 belong to *Phylloscopus occisinensis*, 15 to *Phylloscopus griseolus* and 15 to *Phylloscopus affinis*, all from the same study.

Besides the possibility of additional data mining from database resources, the approach can be used to extract reads originating from specific loci, although many more loci were sequenced in actual experiments (e.g. in hybrid enrichment experiments). Since the read origin is ‘identified’ via MSAs to an amino acid reference, only DNA sequences can be extracted that directly translate into amino acid sequences. Perfect candidates for such loci are eukaryotic organellar genes such as COI. Due to the degeneracy of the genetic code, many different individuals within a broad taxonomic spectrum can be analyzed with the same reference. The number of available amino acid sequences used to produce the consensus reference as well as the taxonomic level (e.g. order, class, phylum) can potentially influence the MSAs and the sequence reconstruction process: a very general consensus sequence (e.g. a vertebrate reference) can be more useful when analyzing a broader taxonomic spectrum of individuals, although less conserved sequence parts of the gene might be inaccurately reconstructed. The higher the taxonomic specificity of the reference sequence, the more accurate the reconstructed DNA sequence. The taxonomic level of the reference as well as the parameters for the MSA have to be adjusted to individual needs.

Currently, we advertise MitoGeneExtractor only for mitochondrial genes, since in the current implementation, indel information coming from Exonerate is not used and reads that align with indels are discarded. The assumption that no indels are present is well met for the majority of mitochondrial genes and taxonomic groups.

Typical distances of sequencing reads to amino acid references, the potential presence of splicing variants and the fact that indels are not considered in the current implementation, limit the application of MitoGeneExtractor for eukaryotic nuclear genes. In contrast, its utility for extracting mitochondrial sequences has been demonstrated and opens the door to extract mitochondrial genes routinely from genomic sequencing resources such as hybrid enrichment data. We also tested MitoGeneExtractor on RNA-seq data (results not shown) and were able to reconstruct COI sequences.

In conclusion, we demonstrated that extraction of sequencing reads from specific loci through alignment to an amino acid reference allows an accurate reconstruction of the corresponding DNA sequence for mitochondrial genes. When incorporated in workflow management tools such as Snakemake, sequence information can be generated for hundreds or even thousands of individuals within a broad taxonomic spectrum without the need for reference sequences of the same or closely related species. When researchers are interested in specific mitochondrial genes, MitoGeneExtractor is faster and more efficient than assembly guided software such as MitoFinder. In principle, the approach can be used to reconstruct any protein coding gene (organelle or prokaryotic genes, RNA-seq data, exon sequencing data) and if gene/locus of interest contributed to the sequence read population within a given NGS library. Genomic resources from which good results are expected are sequencing libraries from hybrid enrichment experiments, transcriptomes and low coverage genomes, although the latter was not tested here.

Acknowledgements:

The work of MB was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG, MA 3684/3-1). MG, JA and CM were supported by a grant from the German Federal Ministry of Education and Research (BMBF) for the GBOL project (FKZ 01LI1101 and 01LI1501).

References:

Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, *11*(10), 697–709. doi: https://doi.org/10.1038/nrg2844

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources*, *20*(4), 892–905. doi: https://doi.org/10.1111/1755-0998.13160

Amaral, F. R. do, Neves, L. G., Jr, M. F. R. R., Mobili, F., Miyaki, C. Y., Pellegrino, K. C. M., & Biondo, C. (2015). Ultraconserved Elements Sequencing as a Low-Cost Source of Complete Mitochondrial Genomes and Microsatellite Markers in Non-Model Amniotes. *PLOS ONE*, *10*(9), e0138446. doi: https://doi.org/10.1371/journal.pone.0138446

Baird, D. J., & Hajibabaei, M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, *21*(8), 2039–2044. doi: https://doi.org/10.1111/j.1365-294X.2012.05519.x

Balakrishnan, R. (2005). Species Concepts, Species Boundaries and Species Identification: A View from the Tropics. *Systematic Biology*, *54*(4), 689–693. doi: https://doi.org/10.1080/10635150590950308

Bilton, D. T., Turner, L., & Foster, G. N. (2017). Frequent discordance between morphology and mitochondrial DNA in a species group of European water beetles (Coleoptera: Dytiscidae). *PeerJ*, *5*. doi: https://doi.org/10.7717/peerj.3076

Bogenhagen, D., & Clayton, D. A. (1974). The Number of Mitochondrial Deoxyribonucleic Acid Genomes in Mouse L and Human HeLa Cells: QUANTITATIVE ISOLATION OF MITOCHONDRIAL DEOXYRIBONUCLEIC ACID. *Journal of Biological Chemistry*, *249*(24), 7991–7995. doi: https://doi.org/10.1016/S0021-9258(19)42063-2

Buchner, D., & Leese, F. (2020). BOLDigger – a Python package to identify and organise sequences with the Barcode of Life Data systems. *Metabarcoding and Metagenomics*, *4*, e53535. doi: https://doi.org/10.3897/mbmg.4.53535

Caparroz, R., Rocha, A. V., Cabanne, G. S., Tubaro, P., Aleixo, A., Lemmon, E. M., & Lemmon, A. R. (2018). Mitogenomes of two neotropical bird species and the multiple independent origin of mitochondrial gene orders in Passeriformes. *Molecular Biology Reports*, *45*(3), 279–285. doi: https://doi.org/10.1007/s11033-018-4160-5

Chamberlain, S. (2021). *bold: Interface to Bold Systems API*. https://CRAN.R-project.org/package=bold

Chevreux, B., Wetter, T., & Suhai, S. (1999). Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *German conference on bioinformatics*.

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., … Lanzén, A. (2021). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, *30*(13), 2937–2958. doi: https://doi.org/10.1111/mec.15472

Dirzo, R., & Raven, P. H. (2003). Global State of Biodiversity and Loss. *Annual Review of Environment and Resources*, *28*(1), 137–167. doi: https://doi.org/10.1146/annurev.energy.28.050302.105532

Eberle, J., Ahrens, D., Mayer, C., Niehuis, O., & Misof, B. (2020). A Plea for Standardized Nuclear Markers in Metazoan DNA Taxonomy. *Trends in Ecology & Evolution*, *35*(4), 336–345. doi: https://doi.org/10.1016/j.tree.2019.12.003

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. doi: https://doi.org/10.1093/nar/gkh340

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786–788. doi: https://doi.org/10.1093/bioinformatics/btv646

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, 7.

Gao, J., Wang, G., Zhou, C., Price, M., Ma, J., Sun, X., … Yue, B. (2021). Complete Mitochondrial Genome of Fulvetta cinereiceps (Sylviidae: Passeriformes) and Consideration of its Phylogeny within Babblers. *Pakistan Journal of Zoology*, *53*(6). doi: https://doi.org/10.17582/journal.pjz/20180524050555

Gao, Y., Yin, S., & Zhu, L. (2021). The complete mitochondrial genome of the Thick-billed Flowerpecker (Dicaeum agile).

Gaziev, A. I., & Shaikhaev, G. O. (2010). Nuclear mitochondrial pseudogenes. *Molecular Biology*, *44*(3), 358–368. doi: https://doi.org/10.1134/S0026893310030027

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

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—A baiting and iterative mapping approach. *Nucleic Acids Research*, *41*(13), e129–e129. doi: https://doi.org/10.1093/nar/gkt371

Harvey, M. G., Bravo, G. A., Claramunt, S., Cuervo, A. M., Derryberry, G. E., Battilana, J., … Derryberry, E. P. (2020). The evolution of a tropical biodiversity hotspot. *Science*, *370*(6522), 1343–1348. doi: 10.1126/science.aaz6970

Hebert, P., Cywinska Alina, Ball Shelley L., & deWaard Jeremy R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1512), 313–321. doi: 10.1098/rspb.2002.2218

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of Birds through DNA Barcodes. *PLoS Biology*, *2*(10), e312. doi: https://doi.org/10.1371/journal.pbio.0020312

Jukes, T. H., & Osawa, S. (1993). Evolutionary changes in the genetic code. *Comparative Biochemistry and Physiology. B, Comparative Biochemistry*, *106*(3), 489–494. doi: 10.1016/0305-0491(93)90122-l

Kerr, K. C., Birks, S. M., Kalyakin, M. V., Red'kin, Y. A., Koblik, E. A., & Hebert, P. D. (2009). Filling the gap—COI barcode resolution in eastern Palearctic birds. *Frontiers in Zoology*, *6*(1), 29. doi: https://doi.org/10.1186/1742-9994-6-29

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. doi: https://doi.org/10.1093/bioinformatics/bts480

Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276–3278. doi: 10.1093/bioinformatics/btu531

Lecuit, M., & Eloit, M. (2015). The potential of whole genome NGS for infectious disease diagnosis. *Expert Review of Molecular Diagnostics*, *15*(12), 1517–1519. doi: https://doi.org/10.1586/14737159.2015.1111140

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. doi: https://doi.org/10.14806/ej.17.1.200

Meiklejohn, K. A., Danielson, M. J., Faircloth, B. C., Glenn, T. C., Braun, E. L., & Kimball, R. T. (2014). Incongruence among different mitochondrial regions: A case study using complete mitogenomes. *Molecular Phylogenetics and Evolution*, *78*, 314–323. doi: https://doi.org/10.1016/j.ympev.2014.06.003

Meng, G., Li, Y., Yang, C., & Liu, S. (2019). MitoZ: A toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Research*, *47*(11), e63. doi: 10.1093/nar/gkz173

Moritz, C., & Cicero, C. (2004). DNA Barcoding: Promise and Pitfalls. *PLOS Biology*, *2*(10), e354. doi: https://doi.org/10.1371/journal.pbio.0020354

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. doi: https://doi.org/10.1101/gr.213959.116

Osawa, S., Jukes, T. H., Watanabe, K., & Muto, A. (1992). Recent evidence for evolution of the genetic code. *Microbiological Reviews*, *56*(1), 229–264. doi: 10.1128/mr.56.1.229-264.1992

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2010). IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. In B. Berger (Hrsg.), *Research in Computational Molecular Biology* (S. 426–440). Berlin, Heidelberg: Springer. doi: https://doi.org/10.1007/978-3-642-12683-3_28

Pentinsaari, M., Ratnasingham, S., Miller, S. E., & Hebert, P. D. N. (2020). BOLD and GenBank revisited – Do identification errors arise in the lab or in the sequence libraries? *PLOS ONE*, *15*(4), e0231814. doi: https://doi.org/10.1371/journal.pone.0231814

Phillips, J. D., Gillis, D. J., & Hanner, R. H. (2019). Incomplete estimates of genetic diversity within species: Implications for DNA barcoding. *Ecology and Evolution*, *9*(5), 2996–3010. doi: https://doi.org/10.1002/ece3.4757

Picardi, E., & Pesole, G. (2012). Mitochondrial genomes gleaned from human whole-exome sequencing. *Nature Methods*, *9*(6), 523–524. doi: https://doi.org/10.1038/nmeth.2029

Pie, M. R., Ströher, P. R., Bornschein, M. R., Ribeiro, L. F., Faircloth, B. C., & McCormack, J. E. (2017). The mitochondrial genome of Brachycephalus brunneus (Anura: Brachycephalidae), with comments on the phylogenetic position of Brachycephalidae. *Biochemical Systematics and Ecology*, *71*, 26–31. doi: https://doi.org/10.1016/j.bse.2016.12.009

Plese, B., Rossi, M. E., Kenny, N. J., Taboada, S., Koutsouveli, V., & Riesgo, A. (2019). Trimitomics: An efficient pipeline for mitochondrial assembly from transcriptomic reads in nonmodel species. *Molecular Ecology Resources*, *19*(5), 1230–1239. doi: https://doi.org/10.1111/1755-0998.13033

Ratnasingham, S., & Hebert, P. D. N. (2007). bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364. doi: https://doi.org/10.1111/j.1471-8286.2007.01678.x

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, *58*(4), 586–597. doi: https://doi.org/10.1016/j.molcel.2015.05.004

Samuels, D. C., Han, L., Li, J., Quanghu, S., Clark, T. A., Shyr, Y., & Guo, Y. (2013). Finding the lost treasures in exome sequencing data. *Trends in Genetics*, *29*(10), 593–599. doi: https://doi.org/10.1016/j.tig.2013.07.006

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050. doi: https://doi.org/10.1111/j.1365-294X.2012.05470.x

Tello, J. G., Moyle, R. G., Marchese, D. J., & Cracraft, J. (2009). Phylogeny and phylogenetic classification of the tyrant flycatchers, cotingas, manakins, and their allies (Aves: Tyrannides). *Cladistics*, *25*(5), 429–467. doi: https://doi.org/10.1111/j.1096-0031.2009.00254.x

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Wang, N., Hosner, P. A., Liang, B., Braun, E. L., & Kimball, R. T. (2017). Historical relationships of three enigmatic phasianid genera (Aves: Galliformes) inferred using phylogenomic and mitogenomic data. *Molecular Phylogenetics and Evolution*, *109*, 217–225. doi: https://doi.org/10.1016/j.ympev.2017.01.006

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. doi: 10.21105/joss.03021

Weigand, H., Weiss, M., Cai, H., Li, Y., Yu, L., Zhang, C., & Leese, F. (2017). Deciphering the origin of mito-nuclear discordance in two sibling caddisfly species. *Molecular Ecology*, *26*(20), 5705–5715. doi: https://doi.org/10.1111/mec.14292

Data Accessibility and Benefit-Sharing

Data Accessibility

Raw data were downloaded from NCBI under Bioproject accession number PRJNA655842. A snapshot of MitoGeneExtractor source code is publicly available under https://doi.org/10.5281/zenodo.6373959. The most recent version is available at GitHub, where also Snakemake workflows and example analyses can be found: github.com/cmayer/MitoGeneExtractor.

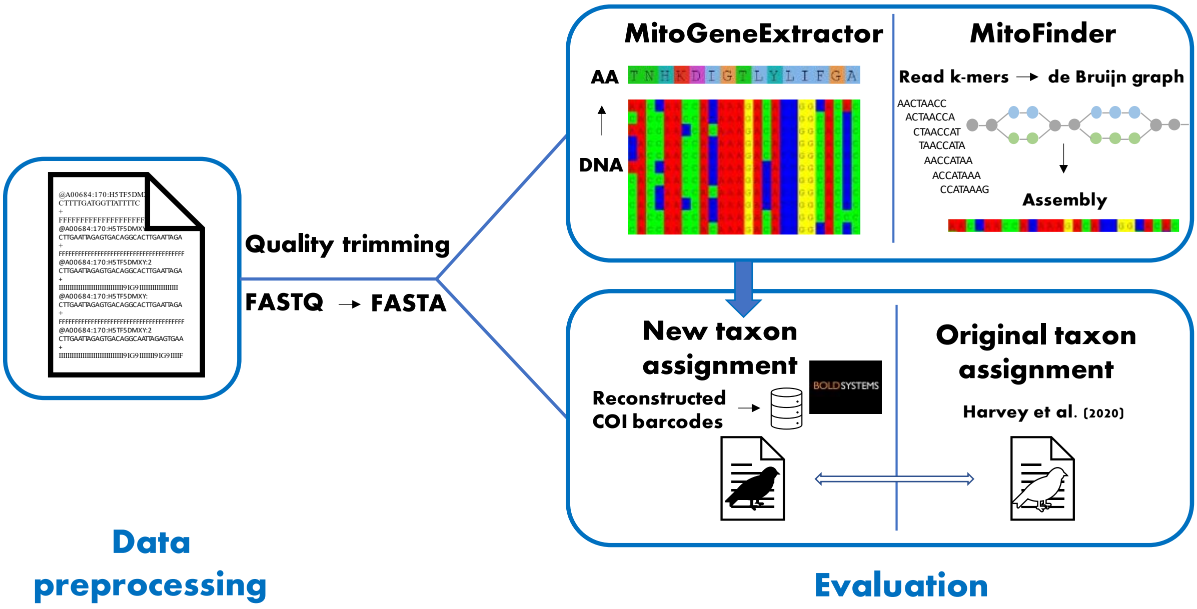
Author Contributions

MB and CM designed the study, CM developed the MitoGeneExtractor program, MB performed the analyses and wrote the manuscript with the help of CM, MG and JA. All authors approved the final version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Figures:



Graphical abstract

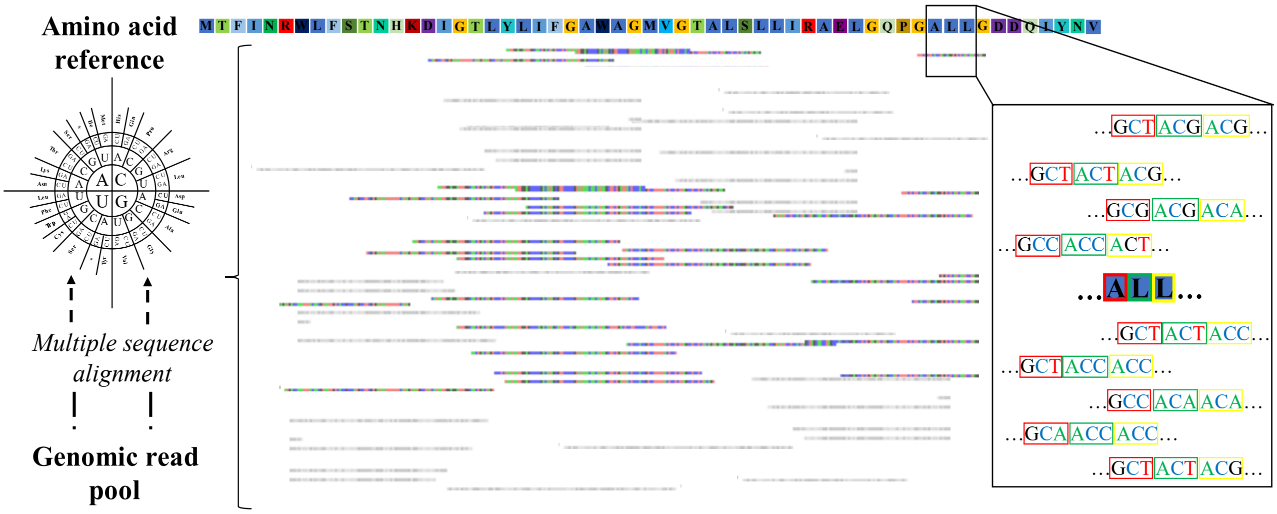


Fig. 1: Illustration of MitoGeneExtractor algorithm. DNA sequence reads are aligned to an amino acid reference taking into account the specified genetic code. With the alignment information coming from Exonerate, a multiple sequence alignment is produced from which the consensus sequence is inferred.

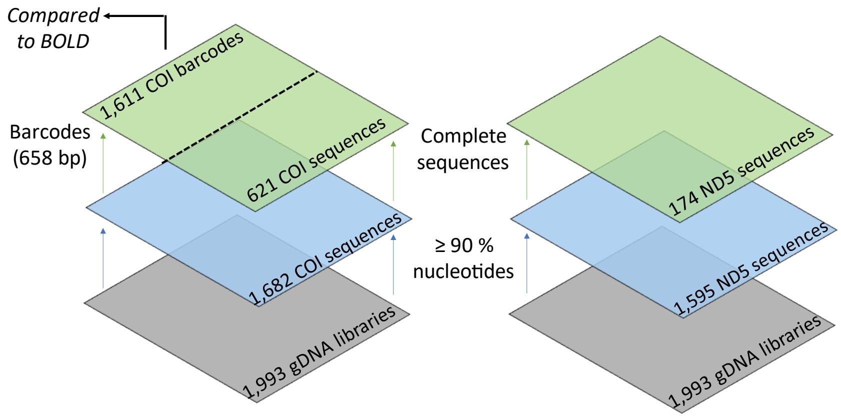


Fig. 2: Success of gene sequence reconstruction. Full length sequences were reconstructed for a large number of specimen (green plane), close to full length sequences, i.e. ≥ 90 % of the complete coding DNA sequence, are available for most specimen (blue plane). For the taxonomic evaluation, COI barcodes were extracted and compared to the Barcode of Life Database. Left: COI gene sequences, right: ND5 gene sequences.

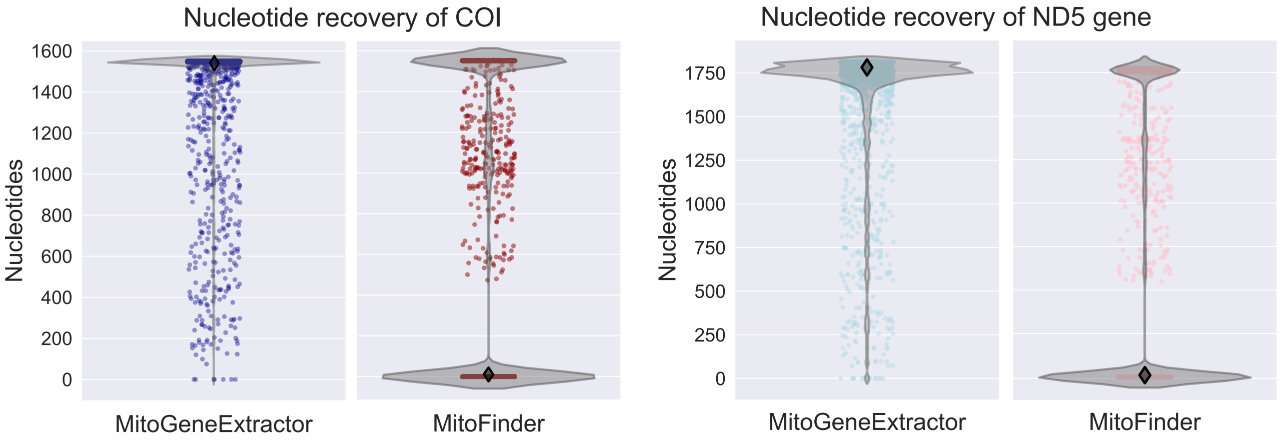


Fig. 3: COI (left) and ND5 (right) reconstruction success with MitoGeneExtractor (blue) and MitoFinder (red). Density plots indicate the probability density curve of the data. Colored dots show the number of nucleotides in individual consensus sequences obtained with MitoGeneExtractor and MitoFinder. Diamonds indicate the median of reconstructed sequences with MitoGeneExtractor (COI = 1,545, ND5 = 1,755) and MitoFinder (COI = 0, ND5 = 0).

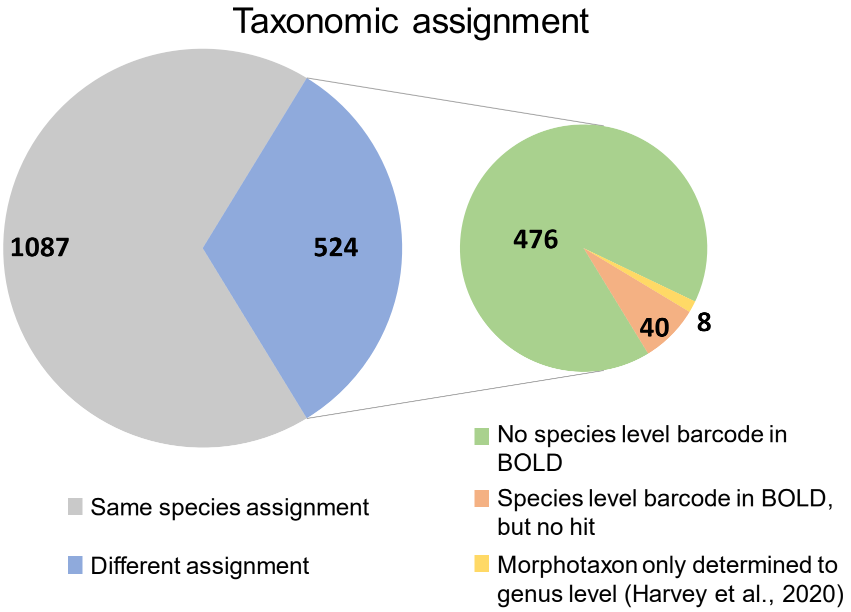


Fig. 4: Taxonomic assignment based on reconstructed COI barcode sequences. Numbers refer to individuals and their reconstructed barcode sequences. For species with barcode sequence information available in BOLD, the taxonomic assignment was consistent to the original study for a large proportion of the specimens. When a specimen was morphologically not determined on species level (yellow), a comparison was not possible.