Cumhuriyet Sci. J., 44(1) (2023) 36-40
DOI: https://doi.org/10.17776/csj.1221192

Cumhuriyet Science Journal

| csj.cumhuriyet.edu.tr | Founded: 2002 ISSN: 2587-2680 e-ISSN: 2587-246X Publisher: Sivas Cumhuriyet University

Determination of the Optimum Number of Short Reads to Obtain the Mitogenome in some Insect Orders

Mahir Budak [1,a,*]

[1] *Department of Molecular Biology and Genetics, Faculty of Science, Sivas Cumhuriyet University, Sivas, Türkiye.*
**Corresponding author*

Research Article

History
Received: 19/12/2022
Accepted: 16/03/2023

ABSTRACT

Sanger sequencing is frequently used as the final step in time-consuming extraction and enrichment processes for examining the mitochondrial genome (mitogenome). The development of next-generation or massively parallel sequencing has made it possible to consistently gather data at the nucleotide level with comparatively little difficulty. Additionally, reference-based genome assembly is now achievable thanks to the growing amount of mt genome data in databases. Consequently, acquiring the genome with fewer short-read counts reduces the financial load on research projects. The use of mitogenomes, particularly in the studies of systematic and population genetics of insects, have increased, and sequencing mitogenomes in non-model animals have become critical. Twelve species from four insect orders, each having a different-sized genome, were employed in the study. Short reads of these species, used in the study, were acquired from the SRA (The Sequence Read Archive) database. Alignments to the reference genome were carried out in triplicate for five different short read counts. It was observed that 0.092 (*Chrysotoxum bicinctum*) to 14.04 (*Anopheles coluzzii*) sequencing depth was needed to obtain the mitogenome with 100X coverage. This work aims to give researchers a better understanding of how much sequencing depth is necessary for mitogenome investigations.

Keywords: Next-generation sequencing, Mitogenome, Short-reads, Coverage, Insects.

[a] *mbudak@cumhuriyet.edu.tr* *https://orcid.org/0000-0001-5610-486X*

Introduction

Mitochondria are essential organelles for eukaryotic cells due to their crucial functions in the production of bioenergetics intermediates. This organelle also participates in important cellular processes, such as signalling, apoptosis, ageing, metabolic homeostasis, and biosynthesis of lipids [1]. The maternally inherited mitochondrial genome is present in nearly all eukaryotic organisms, possessing double-stranded circular DNA molecules of approximately 16 kb in size in general [2]. They typically appear to be consisted of thousands of copies in each cell with widely varying copy number among different tissue cells. The genome of mitochondria (mitogenome) are also commonly preferred in the studies on the genome architecture, evolutionary relationships and adaptive evolution, providing valuable information from intraspesific level to interspesific or higher level of taxa [3-5]. Performing analyses such as hybridization, population genetics, geographical cline has also become essential to reveal the effects of changing environmental conditions. Mitochondrial genes are frequently used in this type of analyses [6].

Insects have become an important model group for figuring out how organisms respond to global warming from an ecological and evolutionary point of view. They are important to the environment and the economy (e.g., as pollination, pests, and as vectors), but they are also vulnerable to environmental conditions [7]. An insect mitogenome generally consists of 13 protein coding genes (PCGs), 24 RNA genes (22 transfer RNAs (tRNAs) and two ribosomal RNAs (rRNAs)) and A + T-rich region (one large control region) with 14–25 kb in length [3,8]. In the result of a rapid search in the organelle section of the public database of NCBI (November 2022) using the “Insecta, mitochondrion” as keywords and filtering the sequence length >10,000 bp, there were complete insect mitogenomes of more than 5,300 species, corresponding approximately 0.5% of described insect species. Despite the most widely studied marker of insect genomes, some technical difficulties have limited benefit for the full potential of mitogenomes. The purification of mitochondria and isolation its DNA in sufficient quantities to perform direct mitogenome sequencing from the large organism such as vertebrates are relatively possible from fresh material [9,10]. However, this approach is not achievable for insect tissues [11,12].

Recent advances in high-throughput sequencing approaches have overcome many of these difficulties and it is possible to rapidly and economically produce very large numbers of relatively short reads from both the nuclear and mitochondrial DNA contained in an insect sample [12,13]. This facility makes this huge data a valuable resource for extracting and assembling insect mitogenomes. However, we need to determine the minimum short-read data required to identify the mitogenomes from insects with varying genome sizes. We can take advantage of mtDNA being present

approximately 10-100 times more than the nuclear DNA in an animal cell [14]. For this aim, we here selected and analyzed twelve species representing the orders of Hymenoptera, Diptera, Lepidoptera and, Coleoptera with variable genome sizes and sequencing depths. Remarkably, it was found that the genome size of the organism did not correlate with the proportion reads belong to mitogenome in the short-read data. It has been found that approximately 1-2X read depth from mixed DNA sample is required to reliably sequence and assemble a typical insect mitogenome.

Materials and Methods

Data Preparation

Within the scope of the study, 12 species belonging to four distinct insect orders were examined. The short reads and reference mitochondrial genomes representing these species used for the analyses were retrieved from the GenBank database. Table 1 lists the accession numbers for each insect species. Fastq-dump v2.8.0 (https://github.com/ncbi/sra-tools) software was used to convert short reads downloaded in SRA format into forward and reverse *fastq* files. Using a custom python script (https://github.com/budakmah/fastq_sub_reads), sub-datasets with various read counts were produced for each fastq file. Each analysis was conducted in triplicate for each sub-dataset.

Mapping of short-reads to Reference Genomes

For alignment to the reference, the bowtie2 software package [15] was used with default settings. To calculate consensus sequences from aligned reads, sam2consensus program (https://github.com/edgardomortiz/sam2consensus) was used. The sam2consensus program takes as input a SAM file resulting from mapping short reads to a reference, then it calculates the consensus sequence from the aligned reads alone. The findings of the mapping to the reference genome and the consensus computations were recorded in log files that contained statistics of analyses.

Statistical Analysis and Data Visualization

For the results of the reference genome mapping and consensus calculations, log files containing the statistics of the analyzes were generated and exported to R v.4.0.2 [16] for statistical analysis. The R custom script was used to extract information about the read counts of each fastq file (RCF), read counts aligned to the target genome (RCTG), coverage (COV), and overall alignment rate (AR) from the log files. The ggplot2 R package [17] was used to visualize data and generate graphs. Threshold value for COV value was accepted as 100. The following formula was used to determine each fastq file's read depth (RD) based on the genome size (GS):

$$RD = \frac{read\ length\ (RL)\ x\ RCF}{GS}$$

The following equation was used to calculate the ratio of RCF to the mitogenome length (ML):

$$\text{ratio of RCF (RDml)} = \frac{RL\ x\ RCF}{ML}$$

The minimum sequencing depth for 100X (RD_{100}) mitogenome coverage was determined using linear least squares regression analysis. The Kendall correlation method was used to figure out if there was a link between the size of the genome and the number of short reads that were needed to cover 100X of the mitogenome.

Results and Discussion

Twelve species with different genome size from the four largest insect orders—Coleoptera, Hymenoptera, Lepidoptera, and Diptera—were selected for the study (Table 1).

Table 1: GenBank accession numbers for SRA and mitogenome, genome sizes and Orders of the species used in the study

Species	Genome Size (bp)	SRA ID	Order	Mitochondrion ACC	Mitochondiron_l
Anopheles coluzzii	262,617,596	SRR17828126	Diptera	KT382819.1	15.441
Apis mellifera	225,234,541	SRR15173878	Hymenoptera	CM040891.1	16.654
Chironomus riparius	191,837,449	ERR7059573	Diptera	OU815687.1	15.666
Chrysotoxum bicinctum	912,938,338	ERR6054968	Diptera	OU426992.1	19.366
Danaus plexippus	245,173,502	ERR4613992	Lepidoptera	KC836923.1	15.314
Harmonia axyridis	425,524,972	ERR6054991	Coleoptera	OU611935.1	19.884
Heliconius numata	351,227,686	ERR4614539	Lepidoptera	Assembled	14.784
Leptidea sinapis	685,583,065	ERR6054634	Lepidoptera	FR990201.1	15.959
Leptinotarsa decemlineata	641,992,784	SRR1055549	Coleoptera	MZ189364.1	16.741
Pachycephus smyrnensis	225,000,000	SRR18358779	Hymenoptera	KX907846.1	15.203
Syrista parreyssii	160,000,000	SRR15850959	Hymenoptera	OK104785.1	18.666
Tenebrio molitor	287,839,991	ERR5859003	Coleoptera	CM025077.1	15.794

The short read data that was retrieved from the database is used to build sub-data with randomly chosen reads. *de-novo* assembly was carried out for the sake of the study because there was no mitogenome sequence belong to *H. numata* in the database. *C. bicinctum* had the biggest genome (about 912 Mb), while *S. parreyssii* had the shortest genome (about 160 Mb).

Each sub-data set has been aligned to the reference mitogenomes, and summary statistics is provided in Table 2. The ratio of sequences matched to the reference mitogenome to the number of reads in the used data (AR) was highest in *A. mellifera* (3.21), while it was lowest in *A. coluzzii* (0.02). There are several explanations for why mitogenome reads in total may vary so greatly between species. This is probably due to three main factors: the first is the mitochondrial copy number, the second is potential experimental variances, and third is body part used for sequencing. Insects' energy requirements vary due to their diverse life strategies. This condition also results in variations in mitochondrial copy number present in cells [18–20].

Table 2: Summary results of each sub-dataset

Species	Order	RCF	RCTG	AR	COV	Ns	Gsize	mtGS	MGR	RD	RL
Anopheles coluzzii		6448193.3	2817.7	0.02	26.64	201	262617596	15441	0.0000588	3.68	151
	Diptera	12897770.7	5638.0	0.02	53.29	51	262617596	15441	0.0000588	7.37	151
		19349235.7	8330.3	0.02	78.73	8	262617596	15441	0.0000588	11.05	151
		25797822.7	11122.7	0.02	105.12	2	262617596	15441	0.0000588	14.74	151
		32246096.0	13839.0	0.02	130.81	2	262617596	15441	0.0000588	18.42	151
Apis mellifera		3346972.3	214829.7	3.21	1951.72	182	225234541	16654	0.0000739	2.23	150
	Hymenoptera	13388126.3	858011.7	3.20	7794.95	160	225234541	16654	0.0000739	8.92	150
		23424563.0	1501845.0	3.21	13644.94	43	225234541	16654	0.0000739	15.60	150
		33463792.0	2145393.0	3.21	19492.24	32	225234541	16654	0.0000739	22.29	150
		33463792.0	2145393.0	3.21	19492.24	32	225234541	16654	0.0000739	22.29	150
Chironomus riparius		12275648.7	111738.0	0.46	1069.46	62	191837449	15666	0.0000817	9.60	150
		16369765.7	149404.7	0.46	1429.42	27	191837449	15666	0.0000817	12.80	150
		20463386.7	186594.3	0.46	1786.30	5	191837449	15666	0.0000817	16.00	150
		22097706.7	201586.0	0.46	1929.09	25	191837449	15666	0.0000817	17.28	150
	Diptera	24554698.0	223832.3	0.46	2142.25	0	191837449	15666	0.0000817	19.20	150
Chrysotoxum bicinctum		97351.0	2219.0	1.14	16.70	808	912938338	19366	0.0000212	0.02	151
		194128.3	4493.3	1.16	33.82	40	912938338	19366	0.0000212	0.03	151
		290865.3	6827.0	1.17	51.34	22	912938338	19366	0.0000212	0.05	151
		387865.0	8937.7	1.15	67.27	1	912938338	19366	0.0000212	0.06	151
		485057.0	11227.3	1.16	84.49	0	912938338	19366	0.0000212	0.08	151
Danaus plexippus		1439176.7	7935.0	0.27	78.01	37	245173502	15314	0.0000625	0.88	151
		2158216.0	11747.0	0.27	115.49	7	245173502	15314	0.0000625	1.32	151
	Lepidoptera	2875802.0	15675.3	0.27	154.12	9	245173502	15314	0.0000625	1.76	151
		3596881.7	19699.3	0.27	193.69	1	245173502	15314	0.0000625	2.20	151
		4314262.7	23533.3	0.27	231.41	0	245173502	15314	0.0000625	2.64	151
Harmonia axyridis		412631.3	2289.3	0.28	16.84	262	425524972	19884	0.0000467	0.15	151
		1237167.0	6831.7	0.28	50.24	4	425524972	19884	0.0000467	0.44	151
	Coleoptera	2061552.0	11186.7	0.27	82.25	0	425524972	19884	0.0000467	0.73	151
		2889836.0	15842.3	0.27	116.51	0	425524972	19884	0.0000467	1.02	151
		3715467.3	20627.0	0.28	151.68	0	425524972	19884	0.0000467	1.31	151
Heliconius numata		1580636.0	28642.3	0.91	242.22	53	351227686	14784	0.0000421	0.68	125
		3160763.7	57245.3	0.91	484.19	14	351227686	14784	0.0000421	1.35	125
		4738753.3	85911.7	0.91	726.66	11	351227686	14784	0.0000421	2.02	125
	Lepidoptera	6321199.0	114477.0	0.91	968.27	1	351227686	14784	0.0000421	2.70	125
		7902389.3	142869.0	0.90	1208.40	1	351227686	14784	0.0000421	3.37	125
Leptidea sinapis		2442035.7	2887.7	0.06	26.29	65	685583065	15959	0.0000233	0.53	151
		4883440.0	6003.0	0.06	54.64	16	685583065	15959	0.0000233	1.07	151
		7326785.3	8897.0	0.06	81.01	2	685583065	15959	0.0000233	1.60	151
		9767001.7	11849.0	0.06	107.86	1	685583065	15959	0.0000233	2.14	151
		12210228.7	14816.7	0.06	134.89	1	685583065	15959	0.0000233	2.67	151
Leptinotarsa decemlineata		17632471.3	9299.0	0.03	56.07	1964	641992784	16741	0.0000261	4.12	101
		35264625.0	18528.0	0.03	111.70	866	641992784	16741	0.0000261	8.24	101
	Coleoptera	35264625.0	18528.0	0.03	111.70	866	641992784	16741	0.0000261	8.24	101
		35264625.0	18528.0	0.03	111.70	866	641992784	16741	0.0000261	8.24	101
		35264625.0	18528.0	0.03	111.70	866	641992784	16741	0.0000261	8.24	101
Pachycephus smyrnensis		1448487.7	9918.7	0.34	97.60	140	225000000	15203	0.0000676	0.97	150
		2896938.7	19712.3	0.34	193.98	116	225000000	15203	0.0000676	1.93	150
		4345744.7	29611.3	0.34	291.36	114	225000000	15203	0.0000676	2.90	150
	Hymenoptera	5793263.0	39278.0	0.34	386.52	109	225000000	15203	0.0000676	3.86	150
		7241341.0	49159.0	0.34	483.73	108	225000000	15203	0.0000676	4.83	150
Syrista parreyssii		3387611.7	79966.7	1.18	640.78	56	160000000	18666	0.0001167	3.18	150
		6772713.3	159137.3	1.17	1275.19	40	160000000	18666	0.0001167	6.35	150
		10159627.7	239514.7	1.18	1919.27	30	160000000	18666	0.0001167	9.52	150
		13548000.0	320148.0	1.18	2565.39	1	160000000	18666	0.0001167	12.70	150
		16935220.7	399503.0	1.18	3201.20	10	160000000	18666	0.0001167	15.88	150
Tenebrio molitor		230950.3	1414.7	0.30	13.33	243	287839991	15794	0.0000549	0.12	151
		461837.7	2920.7	0.32	27.52	21	287839991	15794	0.0000549	0.24	151
	Coleoptera	693059.0	4364.7	0.32	41.25	6	287839991	15794	0.0000549	0.36	151
		1156782.0	7297.7	0.32	68.69	3	287839991	15794	0.0000549	0.60	151
		1617781.0	10097.7	0.31	95.09	0	287839991	15794	0.0000549	0.84	151

RCF: read counts of each fastq file; RCTG: read counts aligned to the target genome; AR: overall alignment rate; COV: coverage; Gsize: genome size; mtGS: mitochondrial genome size; MGR: ratio of mitochondrial genome size to genome size; RD: fastq file's read depth; RL: read length.

The method used to prepare sequencing libraries for next-generation sequencing is another key component. The outcomes of sequencing are significantly impacted by errors and variations in the library preparation procedure.In this situation, the target genome may be represented by fewer reads [21]. Insects have more mitochondrial DNA in their thoracic muscles than their chitinous legs. If the researcher extracts DNA from the insect's legs rather than its thoracic muscles, then the short-read data will contain very few mitogenome sequences. In addition, utilizing the whole insect for genomic DNA isolation implies isolating the microflora that lives in the abdomen [22]. In this case, a portion of the data will be associated with the microorganisms.

The minimum sequencing depth for 100X (RD_{100}) mitogenome coverage was determined using linear least squares regression analysis. Table 3 displays RD_{100} and p-values for each species. The RD_{100} value is highest for *A. coluzzii* (14.04), and it is lowest for *C.bicinctum* (0.09). According to these findings, the RD_{100} values for the majority of species are quite close to 1, and there is no discernible trend according to the different orders of insects.

Table 3: The minimum sequencing depth for 100X mitogenome coverage

Species	RD_{100}	p value
Anopheles coluzzii	14.04	<0.01
Apis mellifera	0.12	<0.01
Chironomus riparius	0.92	<0.01
Chrysotoxum bicinctum	0.09	<0.01
Danaus plexippus	1.14	<0.01
Harmonia axyridis	0.87	<0.01
Heliconius numata	0.28	<0.01
Leptidea sinapis	1.98	<0.01
Leptinotarsa decemlineata	7.37	<0.01
Pachycephus smyrnensis	0.99	<0.01
Syrista parreyssii	0.51	<0.01
Tenebrio molitor	0.88	<0.01

RD_{100}: The minimum sequencing depth for 100X mitogenome coverage

Insect species with different-sized genomes were chosen to see how genome size affects mitogenome assembly. The Kendall rank correlation coefficient or *Kendall's tau* statistic is used to estimate a rank-based measure of association between RD_{100} and genome size. It is likely that as the size of the genome increases, the proportion of short reads that come from the genome will also grow. Concurrently, it is expected that the ratio of reads from mitogenomes will decrease. However, the values for R(tau) and p- value were determined to be 0.03 and 0.95, respectively (Figure 1).

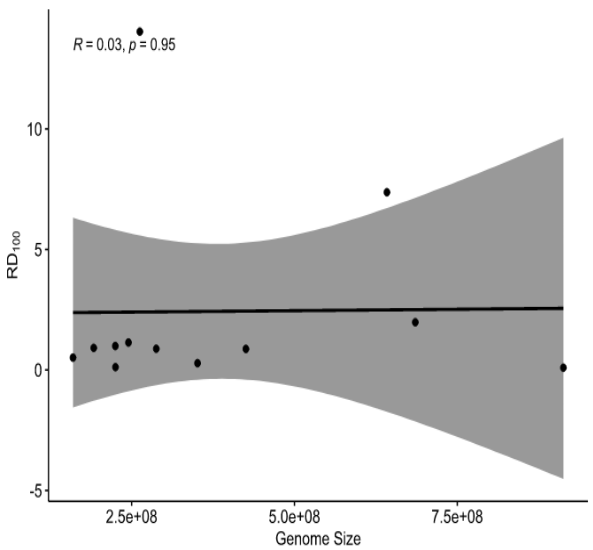

Figure 3. The Kendall rank correlation coefficient statistic graph estimate a rank-based measure of association between RD_{100} and genome size.

It is safe to say that there was no link between the size of the genome and the number of mitogenome reads in the short reads. Perhaps even more surprising is the finding that the insect with the largest genome size (*C. bicinctum*) also had the lowest RD_{100} score (0.09) of all the examined insects (Table 3). As a result, 1-2X reading depth seems sufficient to obtain a full mitogenome with high throughput sequencing approach.

For instance, for a next-generation sequencing analysis employing short-reads of 150 bp in length, roughly 4 million short-reads would be adequate to retrieve the mitogenome of an insect with a genomic size of approximately 500 Mb.

Conclusion

This study determined the minimum short-read data required to identify the mitochondrial genomes of twelve species belonging to the orders Hymenoptera, Diptera, Lepidoptera and, Coleoptera with varied genome sizes. Remarkably, it was found that the genome size of the organism did not correlate with the proportion reads belong to mitogenome in the short-read data. It has been found that approximately 1-2X read depth is required to reliably sequence a typical insect mitogenome.

Conflicts of Interest

The author declares no conflicts of interest. No competing financial interests exist.

References

[1] Ladoukakis, E. D., Zouros, E., Evolution and inheritance of animal mitochondrial DNA: Rules and exceptions, *Journal of Biological Research-Thessaloniki*, 24 (2) (2017) 1–7.

[2] Carlucci, A., Lignitto, L., Feliciello, A., Control of mitochondria dynamics and oxidative metabolism by cAMP, AKAPs and the proteasome, *Trends in Cell Biology*, 18 (12) (2008) 604–613.

[3] Cameron, S. L., Insect mitochondrial genomics: Implications for evolution and phylogeny, *Annual Review of Entomology*, 59 (2014) 95–117.

[4] Aydemir, M. N., Korkmaz, E. M., Comparative mitogenomics of hymenoptera reveals evolutionary differences in structure and composition*, International Journal of Biological Macromolecules*, 144 (2020) 460–472.

[5] Ballard, J. W. O., Pichaud, N., Mitochondrial DNA: More than an evolutionary bystander, *Functional Ecology*, 28 (1) (2014) 218–231.

[6] Okamura, Y., Sato, A., Kawaguchi, L., Nagano, A. J., Murakami, M., Vogel, H., Kroymann, J., Microevolution of pieris butterfly genes involved in host plant adaptation along a host plant community cline, *Molecular Ecology*, 31 (11) (2022) 3083–3097.

[7] González-Tokman, D., Córdoba-Aguilar, A., Dáttilo, W., Lira-Noriega, A., Sánchez-Guillén, R. A., Villalobos, F., Insect responses to heat: Physiological mechanisms, evolution and ecological implications in a warming world, *Biological Reviews*, 95 (3) (2020) 802–821.

[8] Boore, J. L., Animal mitochondrial genomes, *Nucleic Acids Research*, 27 (8) (1999) 1767–1780.

[9] Güler, M., Güler, F. T., Korkmaz, E. M., Budak, M., Böcek dokularından DNA İzolasyonu yöntemlerinin kalite, verim ve maliyet açısından karşılaştırılması, *Selçuk Üniversitesi Fen Fakültesi Fen Dergisi*, 44 (2018) 135–148.

[10] Budak, M., Korkmaz, E. M., Basibuyuk, H. H., A molecular phylogeny of the cephinae (hymenoptera, cephidae) based on mtDNA COI gene: A test of traditional classification, *ZooKeys*, 130 (2011) 363–378.

[11] Hu, M., Jex, A. R., Campbell, B. E., Gasser, R. B., Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing, *Nature Protocols,* 2 (10) (2007) 2339–2344.

[12] Jex, A. R., Hall, R. S., Littlewood, D. T. J., Gasser, R. B., An integrated pipeline for next-generation sequencing and annotation of mitochondrial genomes, *Nucleic Acids Research,* 38 (2) (2010) 522–533.

[13] Ye, F., Samuels, D. C., Clark, T., Guo, Y., High-throughput sequencing in mitochondrial DNA research, *Mitochondrion*, 17 (2014) 157–163.

[14] Al-Nakeeb, K., Petersen, T. N., Sicheritz-Pontén, T., Norgal: Extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data, *BMC Bioinformatics*, 18 (1) (2017) 1–7.

[15] Langmead, B., Salzberg, S. L., Fast gapped-read alignment with bowtie 2, *Nature Methods*, 9 (2012) 357–359.

[16] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (2022).

[17] Wickham, H., ggplot2: Elegant graphics for data analysis. 2nd ed., New York: Springer-Verlag (2016) 18 – 30.

[18] Reinhold, K., Energetically costly behaviour and the evolution of resting metabolic rate in insects, *Functional Ecology,* 13 (1999) 217–224.

[19] Li, F., Zhao, X., Li, M., He, K., Huang, C., Zhou, Y., Li, Z., Walters, J. R., Insect genomes: Progress and challenges, *Insect Molecular Biology*, 28 (6) (2019) 739–758.

[20] Cameron, S. L., How to sequence and annotate insect mitochondrial genomes for systematic and comparative genomics research, *Systematic Entomology*, 39 (3) (2014) 400–411.

[21] Gómez-Rodríguez, C., Intraspecific genetic variation in complex assemblages from mitochondrial metagenomics: Comparison with DNA barcodes, *Methods in Ecology and Evolution*, 8 (2) (2017) 248–256.

[22] Patzold, F., Zilli, A., Hundsdoerfer, A. K., Advantages of an easy-to-use DNA extraction method for minimal-destructive analysis of collection specimens*, PloS one*, 15 (7) (2020) e0235222.