Identification of Mitochondrial DNA (NUMTs) in the Nuclear Genome of *Daphnia Magna*

Krzysztof Kowal

Uniwersytet Przyrodniczy w Lublinie

Angelika Tkaczyk

Uniwersytet Przyrodniczy w Lublinie

Mariusz Pierzchała

Institute of Genetics and Animal Biotechnology Polish Academy of Science

Adam Bownik

Uniwersytet Przyrodniczy w Lublinie

Brygida Ślaska (✉ brygida.slaska@up.lublin.pl)

University of Life Sciences in Lublin <https://orcid.org/0000-0003-3933-2400>

Research article

Keywords: mtDNA, NUMT, water flea, BLAST

Posted Date: May 26th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-29900/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at International Journal of Molecular Sciences on November 18th, 2020. See the published version at <https://doi.org/10.3390/ijms21228725>.

IDENTIFICATION OF MITOCHONDRIAL DNA (NUMTS) IN THE NUCLEAR GENOME OF *DAPHNIA MAGNA*

Krzysztof Kowal¹, Angelika Tkaczyk¹, Mariusz Pierzchała², Adam Bownik³, Brygida Ślaska^{1*}

¹Institute of Biological Bases of Animal Production, Faculty of Animal Sciences and Bioeconomy, University of Life Sciences in Lublin, Akademicka 13, 20-950 Lublin, Poland

²Department of Genomics and Biodiversity, Institute of Genetics and Animal Biotechnology, Polish Academy of Sciences, Postępu 36a Str., 05-552 Jastrzebiec, Poland

³Department of Hydrobiology and Protection of Ecosystems, University of Life Sciences in Lublin, Dobrzańskiego 37, 20-62 Lublin, Poland.

*Corresponding author: brygida.slaska@up.lublin.pl

Abstract

Background : This is the first study in which the *Daphnia magna* (*D. magna*) nuclear genome deposited in the GenBank data-base was analyzed for pseudogene sequences of mitochondrial origin. The first complete information about the genome of *D. magna* was published by Lee et al. in 2019. To date, there is no information about pseudogenes localized in the genome of *D. magna*. The aim of the present study was to identify NUMTs, their length, homology, and location for potential use in evolutionary studies and to check whether their occurrence causes co-amplification during mitochondrial genome analyses.

Results : Bioinformatic analysis showed 1909 fragments of the mitochondrial genome of *D. magna*, of which 1630 fragments were located in ten linkage groups (LG) of the nuclear genome (nDNA). The most frequently occurring fragments of the mtDNA sequence in the nuclear genome included *ND2* (115), *ND3* (113), and *TRNA-CYS* (110)). However, the highest number of NUMTs was observed for the D-loop (147). 253 fragments showed 100% homology (from 16 to 46 bp) with mtDNA gene sequences. The sequence homology for *TRNA-MET* was 100% for all 6 NUMTs (from 16 to 18 bp). The overall length of NUMTs in the nDNA was 44.391 bp (from 16 to 182 bp), which accounted for 0.042% of the entire genome.

Conclusions : The best-matched NUMTs covering more than 90% of the mtDNA gene sequence have been identified for the *TRNA-ARG* (95%), *TRNA-GLU* (97%), and *TRNA-THR* (95%) genes, and they may be included in the functional nuclear tRNA genes. Using the product of total DNA isolation in mtDNA studies, coamplification of nDNA fragments is unlikely in the case of amplification of the whole tRNA genes as well as fragments of other genes and the D-loop with a length exceeding 200 bp. It was observed that *TRNA-MET* fragments had the highest level of sequence homology, which means that they could be evolutionarily the youngest. The lowest degree of homology was found in the pseudogene derived from the mtDNA D-loop sequence. It may probably be the oldest element of mitochondrial DNA incorporated into the nuclear genome; however, further analysis is necessary.

Keywords: mtDNA, NUMT, water flea, BLAST

42   **Background**

43   *Daphnia*, commonly known as the water flea, is a small crustacean usually inhabiting freshwa-
44   ter ponds and lakes on all continents of the globe. It has long been used as a model for elucidation of
45   animal responses and adaptations to environmental changes (1). It has been also used in diverse bio-
46   logical research areas such as ecology, ecotoxicology, evolution, and reproductive biology due to its
47   important position in the aquatic food chain, a high degree of phenotypic plasticity, and cyclical par-
48   thenogenesis responding to environmental stimuli (1-7). Its sensitive behavioral and physiological re-
49   sponses are parameters used as biomarkers of the effect induced by various substances (1-8). Thus, it
50   has been used for reproduction tests, acute toxicity studies, and chronic toxicity tests in the OECD
51   Guidelines (9, 10). The current state of knowledge of the nuclear genome of *Daphnia magna* (*D.*
52   *magna*) is based on reports by Routtu et al. (11, 12), Dukic et al. (13), and Lee et. al. (14). Low- and
53   high-density genetic linkage maps were obtained, in which they assembled the whole genome se-
54   quence of *D. magna*. Specific genetic markers from a high-resolution genetic linkage map of *D. magna*
55   xinb3 were evaluated in toxicological studies by Korea Institute of Toxicology (KIT) (14).

56   Genomic resources are steadily being developed for many species of the genus *Daphnia*. In
57   particular, a database of around 12,000 expressed sequence tags (EST) is currently available
58   (<http://wfleabase.org>) (15), providing a useful resource to isolate polymorphic genetic markers in this
59   species. However, there is no information about pseudogenic sequences of mitochondrial origin
60   (NUMTs) in the *D. magna* genome.

61   Nuclear DNA sequences that are homologous to the mitochondrial genome are often referred
62   to as mitochondrial pseudogenes, or NUMTs (16). NUMTs may differ in length and be as large as the
63   full length of the mitochondrial genome (17). It has been reported that the NUMT length is positively
64   correlated with the genome size, suggesting potential roles of non-coding DNA gain and loss in NUMT
65   accumulation (18). NUMTs have been documented in almost all eukaryotic genomes studied (19). The

transfer of mitochondrial DNA (mtDNA) sequences into the nuclear genome is an ongoing evolutionary process (20), which has markedly influenced the evolution and function of eukaryotic genomes (19). Thus, NUMTs are good materials for studying the evolution of nuclear sequence without selective constraints (21). However, because of their homology, NUMTs may confound mtDNA studies, as the NUMT co-amplification product could interfere with sequence analysis (22).

This is the first study in which the *D. magna* nuclear genome deposited in the GenBank database was analyzed for pseudogene sequences of mitochondrial origin. The aim of the present study was to identify NUMTs, their length, homology, and location for potential use in evolutionary studies and to check whether their occurrence causes co-amplification during mitochondrial genome analyses.

75

76

77 **Results**

Bioinformatic analysis showed 1909 fragments of the mitochondrial genome, of which 1630 fragments were located in ten linkage groups (LG) of the nuclear genome of *D. magna*. The other fragments were localized from scaffolds and used during genome sequencing. All the NUMTs found in this research are listed in Supplementary file 1. The total length of the NUMT sequences in the linkage groups corresponded to the number of fragments on individual LG (Table 1). The total length of NUMTs in the *D. magna* genome was 44.391 base pairs (bp), which accounted for 0.042% of the length of the nuclear genome. Their percentage content was 0.037% in the longest linkage group 2, in which 228 NUMTs were identified, and 0.047% in the shortest linkage group 10 (Table 1). The most frequently occurring fragments of the mtDNA sequence in the nuclear genome included *ND2* (115), *ND3* (113), *TRNA-CYS* (110), and 16S rRNA (105). However, the highest number of NUMTs was observed for the non-coding area, i.e. the D-loop (147) (Table 2). The lowest numbers of mtDNA fragments found in nDNA were observed mainly for genes encoding tRNA molecules: *TRNA-PRO* and *TRNA-MET* (6), *TRNA-TYR* and *TRNA-ASN* (4), and *TRNA-SERI* (2). In contrast, fragments of the mitochondrial gene sequence *TRNA-ILE* were the only sequences that were not found in the nuclear

genome of *D. magna*. The highest numbers of mtDNA fragments were recorded on LG2 - 228, and the lowest - on LG8 - 134 (Table 2).

The longest fragments of the mitochondrial genome present in the nuclear genome were observed for the D-loop (182 bp), *ND4* (108 bp), *ND3* (99 bp), and *ND5* and *COX3* (94 bp each) (Table 3). The 182-bp fragment of the D-loop constituted 63% of the entire sequence of this region. In contrast, in the case of the other protein-encoding genes, the fragment size ranged from 4% (*CYTB*) to 32% (*ATP8*). In turn, NUMTs were recorded among genes encoding tRNA, constituting over 90% of the mtDNA gene sequence: *TRNA-ARG* and *TRNA-THR* (95% each) and *TRNA-GLU* (97%). All the analyzed sequence fragments had a minimum length in the range of 16-25 bp (Table 3).

Of the 1630 NUMTs (Table 1), 253 fragments, representing 16% of all NUMTs, showed 100% homology with the mtDNA gene sequences. 100% sequence homology for *TRNA-MET* was found for all 6 NUMTs (Tables 2 and 4). 23 NUMTs whose sequence homology was 100% were observed for the D-loop region and 21 NUMTs for the gene *ND3*. In contrast, in the case of genes *TRNA-SER1* and *TRNA-ALA*, no NUMTs with 100% sequence homology were observed (Table 2). The mean values of the percentage of sequence identity ranged from 88.0% (*ATP8*) to 100% (*TRNA-MET*). The percentage identity for the individual linkage groups was in the range of 90-91%. The largest homology was recorded on LG10 (90.7%) and the lowest - on LG6 (90.2%) (Table 4). At least one sequence with 100% homology was identified on each of the linkage groups.

Table 1 Percentage content of NUMTs in the linkage groups of nuclear genomes

Linkage group (LG)	PHYSICAL LENGTH (BP) FROM LEE ET. AL 2019	SUM OF NUMTS LENGTHS (BP)	PERCENTAGE CONTENT OF NUMTS ON LG
LG1	14,058,888	5,063	0.036%
LG2	16,351,056	6,071	0.037%
LG3	11,081,246	4,389	0.040%
LG4	10,002,879	4,147	0.041%
LG5	10,116,075	4,600	0.045%
LG6	9,588,688	3,997	0.042%
LG7	10,149,764	4,829	0.048%
LG8	9,006,911	3,533	0.039%
LG9	8,299,553	3,966	0.048%
LG10	8,061,327	3,796	0.047%
Total	106,716,387	44,391	0.042%

Table 2 Distribution of mtDNA gene fragments in the linkage groups and the sum of fragment lengths

mtDNA SEQUENCE	<i>LG</i>										<i>NUMBER OF NUMTs (100% IDENTICAL)*</i>	<i>SUM OF GENE FRAGMENTS NTS (bp)</i>
	1	2	3	4	5	6	7	8	9	10		
<i>TRNA-GLN</i>	3	6	5	7	2	5	3	2	5	4	42 (15)	966
<i>TRNA-MET</i>	2	2	1	1	-	-	-	-	-	-	6 (6)	102
<i>ND2</i>	12	18	15	12	12	9	12	12	6	7	115 (15)	3235
<i>TRNA-TRP</i>	3	2	3	3	2	-	-	3	1	2	19 (6)	447
<i>TRNA-CYS</i>	11	13	9	14	5	14	11	10	11	12	110 (12)	2596
<i>TRNA-TYR</i>	-	1	-	-	-	-	-	1	1	1	4 (1)	100
<i>COX1</i>	1	1	2	2	5	3	7	3	3	4	31 (2)	1175
<i>TRNA-LEU1</i>	5	1	9	3	1	6	5	3	1	2	36 (10)	786
<i>COX2</i>	2	6	3	4	14	7	5	5	2	4	52 (9)	1492
<i>TRNA-LYS</i>	-	2	3	-	2	-	2	2	-	1	12 (3)	290
<i>TRNA-ASP</i>	3	-	1	2	-	2	1	3	-	-	12 (1)	273
<i>ATP8</i>	-	5	1	2	-	1	3	2	1	-	15 (3)	438
<i>ATP6</i>	4	7	4	4	5	2	2	1	2	3	34 (2)	957
<i>COX3</i>	3	2	2	-	4	2	3	-	3	4	23 (3)	717
<i>TRNA-GLY</i>	2	3	4	-	2	1	5	2	2	2	23 (3)	512
<i>ND3</i>	17	17	14	6	9	14	10	9	15	2	113 (21)	2876
<i>TRNA-ALA</i>	1	6	1	5	1	1	1	2	-	1	19 (0)	407
<i>TRNA-ARG</i>	2	1	5	5	2	5	-	1	1	3	25 (4)	695
<i>TRNA-ASN</i>	1	1	-	1	-	-	-	-	1	-	4 (1)	112
<i>TRNA-SER1</i>	-	-	-	-	-	-	-	2	-	-	2 (0)	52
<i>TRNA-GLU</i>	5	7	4	2	4	3	4	4	5	-	38 (8)	924
<i>TRNA-PHE</i>	-	1	-	-	2	1	-	2	2	1	9 (2)	187
<i>ND5</i>	17	15	7	11	7	10	10	3	10	8	98 (10)	3244
<i>TRNA-HIS</i>	-	3	1	2	2	3	7	1	3	2	24 (6)	536
<i>ND4</i>	7	4	3	4	2	1	2	4	3	7	37 (1)	1194
<i>ND4L</i>	4	3	3	1	1	2	1	2	-	6	23 (2)	615
<i>TRNA-THR</i>	5	2	-	-	2	-	3	3	2	-	17 (9)	355
<i>TRNA-PRO</i>	-	-	1	3	-	-	1	-	1	-	6 (1)	151
<i>ND6</i>	9	14	12	7	5	8	14	10	4	8	91 (7)	2445
<i>CYTB</i>	8	14	7	5	6	5	4	3	6	4	62 (8)	1631
<i>TRNA-SER2</i>	9	7	9	3	9	6	5	4	5	11	68 (16)	2644
<i>ND1</i>	15	12	9	13	6	11	13	6	5	6	96 (16)	1486
<i>TRNA-LEU2</i>	-	3	-	-	1	1	-	1	-	-	6 (2)	135
<i>16S rRNA</i>	11	19	3	7	15	8	11	11	10	10	105 (10)	3700
<i>TRNA-VAL</i>	2	2	1	1	3	1	2	2	-	-	14 (4)	320
<i>12S rRNA</i>	12	16	6	7	5	6	12	5	13	10	92 (10)	2617
<i>TRNA-ILE</i>	-	-	-	-	-	-	-	-	-	-	0	0
D-LOOP	13	12	22	13	29	12	12	10	11	13	147 (23)	3979
SUM	189	228	170	150	165	150	171	134	135	138	1630 (253)	
SUM OF FRAGMENTS IN THE LG (bp)	5063	6071	4389	4147	4600	3997	4829	3533	3966	3796		44391

*The counts of NUMTs that were 100% identical with the sequence from mtDNA are indicated in brackets.

118

Table 3 Minimal and maximal length of each mtDNA gene fragments located in the nuclear genome

SEQUENCE	LENGTH (IN BP)	MIN (IN BP)	% MIN	MAX (IN BP)	% MAX
<i>TRNA-GLN</i>	68	16	24%	44	65%
<i>TRNA-MET</i>	65	16	25%	18	28%
<i>ND2</i>	987	18	2%	58	6%
<i>TRNA-TRP</i>	64	16	25%	38	59%
<i>TRNA-CYS</i>	64	16	25%	40	63%
<i>TRNA-TYR</i>	64	18	28%	35	55%
<i>COX1</i>	1537	19	1%	72	5%
<i>TRNA-LEU1</i>	68	16	24%	39	57%
<i>COX2</i>	679	18	3%	52	8%
<i>TRNA-LYS</i>	70	17	24%	35	50%
<i>TRNA-ASP</i>	63	17	27%	32	51%
<i>ATP8</i>	168	17	10%	54	32%
<i>ATP6</i>	675	18	3%	47	7%
<i>COX3</i>	789	19	2%	94	12%
<i>TRNA-GLY</i>	63	16	25%	32	51%
<i>ND3</i>	354	17	5%	99	28%
<i>TRNA-ALA</i>	62	19	31%	28	45%
<i>TRNA-ARG</i>	64	16	25%	61	95%
<i>TRNA-ASN</i>	67	16	24%	45	67%
<i>TRNA-SER1</i>	65	25	38%	27	42%
<i>TRNA-GLU</i>	65	16	25%	63	97%
<i>TRNA-PHE</i>	68	16	24%	26	38%
<i>ND5</i>	1708	19	1%	94	6%
<i>TRNA-HIS</i>	63	16	25%	42	67%
<i>ND4</i>	1315	19	1%	108	8%
<i>ND4L</i>	306	17	6%	38	12%
<i>TRNA-THR</i>	63	16	25%	60	95%
<i>TRNA-PRO</i>	64	16	25%	34	53%
<i>ND6</i>	504	18	4%	64	13%
<i>CYTB</i>	1133	18	2%	46	4%
<i>ND1</i>	927	18	2%	88	9%
<i>TRNA-SER2</i>	69	16	23%	35	51%
<i>TRNA-LEU2</i>	67	16	24%	27	40%
<i>16S rRNA</i>	1373	19	1%	78	6%
<i>TRNA-VAL</i>	72	16	22%	38	53%
<i>12S rRNA</i>	752	18	2%	72	10%
<i>TRNA-ILE</i>	64	-	-	-	-
D-LOOP	289	17	6%	182	63%

119

120

121

122

123

124

125

126

127

128

129

130

131 **Table 4** Mean % identity of mtDNA gene fragments located in the linkage groups
132

SEQUENCE	LG1	LG2	LG3	LG4	LG5	LG6	LG7	LG8	LG9	LG10	MEAN % IDENTITY FOR EACH GENE
<i>TRNA-GLN</i>	93.8	88.8	94.3	92.9	92.3	85.1	97.1	89.4	90.5	95.0	91.6
<i>TRNA-MET</i>	100.0	100.0	100.0	100.0	-	-	-	-	-	-	100.0
<i>ND2</i>	88.7	91.4	90.2	88.0	89.3	88.4	91.2	90.0	87.2	87.2	89.5
<i>TRNA-TRP</i>	90.9	95.2	88.3	90.2	93.8	-	-	89.7	100.0	94.0	91.8
<i>TRNA-CYS</i>	92.6	89.7	89.1	89.9	88.4	93.3	91.4	87.6	88.7	92.4	90.5
<i>TRNA-TYR</i>	-	85.7	-	-	-	-	-	100.0	94.7	97.1	94.4
<i>COX1</i>	79.5	92.3	82.6	94.6	87.8	82.2	88.3	90.7	92.8	95.4	89.1
<i>TRNA-LEU1</i>	91.0	100.0	92.5	91.9	100.0	93.6	91.5	95.2	86.2	83.3	92.2
<i>COX2</i>	81.6	95.4	86.5	87.5	92.1	90.2	91.1	89.0	85.0	89.0	90.2
<i>TRNA-LYS</i>	-	95.2	86.5	-	98.6	-	95.1	91.9	-	90.5	92.6
<i>TRNA-ASP</i>	96.8	-	95.0	84.4	-	85.6	90.9	91.5	-	-	90.9
<i>ATP8</i>	-	88.4	88.5	84.8	-	100.0	84.5	87.7	90.9	-	88.0
<i>ATP6</i>	86.9	85.1	90.4	88.4	93.1	90.3	86.9	95.2	85.5	92.7	88.9
<i>COX3</i>	84.9	90.5	87.6	-	90.5	91.3	97.1	-	88.3	94.6	90.9
<i>TRNA-GLY</i>	90.4	92.1	92.0	-	95.2	90.9	94.3	85.0	95.2	88.3	92.0
<i>ND3</i>	90.2	91.5	89.2	88.5	96.1	91.1	90.3	89.9	91.5	95.7	91.0
<i>TRNA-ALA</i>	90.5	93.7	94.7	89.8	94.7	90.9	88.0	90.7	-	94.7	91.9
<i>TRNA-ARG</i>	89.0	90.5	92.0	94.3	89.7	87.0	-	90.9	94.7	97.7	91.7
<i>TRNA-ASN</i>	84.6	100.0	-	88.0	-	-	-	-	95.6	-	92.0
<i>TRNA-SERI</i>	-	-	-	-	-	-	-	90.3	-	-	90.3
<i>TRNA-GLU</i>	89.8	90.8	89.3	90.2	93.0	95.1	87.6	91.6	96.2	-	91.5
<i>TRNA-PHE</i>	-	100.0	-	-	90.1	90.5	-	95.5	92.6	91.3	93.1
<i>ND5</i>	89.5	88.0	86.1	92.5	89.4	92.4	86.8	86.0	89.5	88.6	89.2
<i>TRNA-HIS</i>	-	95.1	100.0	100.0	80.7	87.6	91.2	88.9	89.0	100.0	91.8
<i>ND4</i>	89.5	86.1	90.9	87.3	93.0	89.7	87.3	88.5	88.7	94.1	89.8
<i>ND4L</i>	86.1	87.4	84.6	82.9	100.0	88.9	95.0	87.3	-	90.8	88.5
<i>TRNA-THR</i>	98.3	89.5	-	-	95.0	-	98.3	100.0	94.0	-	96.6
<i>TRNA-PRO</i>	-	-	91.3	90.7	-	-	100.0	-	92.6	-	92.7
<i>ND6</i>	92.9	89.0	89.7	92.6	87.8	90.9	89.1	89.2	88.7	92.6	90.2
<i>CYTB</i>	91.1	89.4	89.8	93.0	91.1	87.1	88.9	93.4	92.2	90.9	90.5
<i>ND1</i>	90.7	92.7	93.1	88.3	92.3	88.3	91.5	92.5	90.6	86.1	90.6
<i>TRNA-SER2</i>	95.3	92.3	94.4	95.2	89.4	93.6	96.6	90.2	90.3	88.8	92.3
<i>TRNA-LEU2</i>	-	92.0	-	-	85.2	85.2	-	100.0	-	-	91.1
<i>16S rRNA</i>	86.3	87.6	81.7	91.9	87.1	88.4	86.4	87.9	86.1	92.7	87.8
<i>TRNA-VAL</i>	100.0	83.5	90.5	92.0	90.9	90.5	83.3	100.0	-	-	91.4
<i>12S rRNA</i>	89.1	89.9	88.2	87.8	90.2	88.9	89.2	93.0	93.1	85.5	89.5
D-LOOP	91.9	90.0	91.7	91.8	88.8	91.0	90.3	92.6	91.5	88.0	90.6
MEAN % IDENTITY FOR EACH LG	90.6	90.3	90.3	90.5	90.4	90.2	90.4	90.6	90.6	90.7	90.4

133

134 Discussion

135 This is the first study in which the *D. magna* nuclear genome deposited in the GenBank data-
136 base was analyzed for pseudogene sequences of mitochondrial origin. The first complete information

137 about the genome of *D. magna* was published by Lee et al. in 2019 (14). To date, there is no infor-
138 mation about pseudogenes localized in the genome of the water flea. Our research provides complete
139 bioinformatic information about the location of NUMTs found in the reference genome, which may
140 be useful for future phylogenetics, evolution, and/or population analyses.

141 A computer-based search for NUMTs in the nuclear genome of 85 species of animals, plants,
142 fungi, and protists showed that the total length of detected NUMTs varied from 0 to 823.9 kb per
143 nuclear genome (24). For instance, the NUMT content was 0 in *Anopheles gambiae*, 263.478 bp in
144 *Homo sapiens*, and more than 800 kbp in *Oryza sativa* (24). In the case of *D. magna* (Table 1), the
145 overall length of NUMT in the nuclear genome was 44.391 bp. The total length of 24 NUMTs was
146 9.989 bp in the *Pteromalus puparum* genome, and 42.972 bp in *Nasoni vitripennis* (25), and more than
147 230 kbp in *Apis mellifera* (26). There seems to be a positive correlation between the haploid genome
148 sizes (C-values) and NUMT amount/prevalence in eukaryotes (27).

149 Another explanation for the differences in the total length of NUMTs is the fact that they accu-
150 mulate in the genomes in a continuous evolutionary process (18). Like in other species, e.g. *M. lucifu-*
151 *gus* (28), the percentage of NUMTs in the genome was less than 0.1%. In contrast, the number and
152 length of NUMTs may vary depending on the computer-based query for NUMTs in the BLAST tool,
153 as in the case of the genome of *Canis lupus familiaris* (22, 29). In our study, the sequence search in
154 *BLASTN 2.6.0* implemented in CLC Genomics Workbench 12.0 yielded 1909 results, although 279
155 results were found in various scaffolds used during sequencing of the *D. magna* genome. In this paper,
156 however, the NUMT results from the scaffold were excluded due to the potential occurrence of arti-
157 facts created during sequencing, as observed by Shi et al. (28), where surprisingly, an entire mitochon-
158 drial genome was found in the scaffold AAPE02072785 in the *M. lucigufus* genome. It is also worth
159 considering that, in this work, the BLAST search result took into account all search results, even those
160 fragments whose length was only 16 bp (Table 3). The results of the NUMTs found in scaffolds are
161 listed in Supplementary file 1.

NUMTs were used to define characteristics and to clarify phylogenetic inconsistencies suggested by paralog sequences (18, 30). The analysis performed by Mishmar et al. (31) revealed that mtDNA fragments, which were integrated into the nucleus before the radiation of modern human mtDNAs, confirming that mtDNAs similar to today's African macro-haplogroup L were the first human mtDNAs.

The analysis of NUMTs in *D. magna* revealed that the latest evolutionary sequences are pseudogenes derived from the sequence *TRNA-MET*, since the homology of the entire gene sequence was 100%. The lowest homology (70.33%) was characteristic for the pseudogene derived from the mtDNA D-loop sequence located on LG5 (no tabulated data). It may probably be the oldest element of mitochondrial DNA incorporated into the nuclear genome. However, due to the high degree of mutation in the D-loop, the thesis requires further verification. Similarly, NUMTs derived from *TRNA-ALA* and *TRNA-SER1* may have been one of the first sequences derived from mtDNA. However, by assessing only sequence homology, the sequence of incorporation of mitochondrial pseudogenes into the nuclear genome cannot be determined. Nevertheless, as observed by Mishmar et al. (31), the nuclear genome accumulates mutational changes at a much slower rate than mtDNA. Hence, the sequences of "recent" NUMTs can provide valuable information about the mtDNA sequences of the earliest humans.

NUMTs exhibit different degrees of homology to their mitochondrial counterparts. They are variable in size, evenly distributed within and among chromosomes, and, in some cases, they are highly rearranged and/or fragmented (32). However, the size of the mitochondrial chromosome does not correlate with the NUMT frequency or size distribution (19). The transfer to the nucleus can be influenced by the vulnerability of mitochondria to stress and other factors which may cause the escape of mtDNA to the cytoplasm (32). Mutations in mtDNA may occur in the entire mitochondrial genome; however, they are most frequently detected in the hypervariable regions of D-loops (33-35).

186   The number of mutations as well as their incidence in the D-loop area may be related to the
187 number of NUMTs occurring in the nuclear genome.

188   The higher the mutation rate, the greater the likelihood of transfer of the D-loop fragment into
189 the cytoplasm followed by its incorporation into the nuclear genome as a pseudogene. In the *D. magna*
190 genome, the greatest numbers of pseudogenes from the D-loop (147) were observed, and only 23 of
191 them had 100% sequence homology (Table 1). However, no pseudogenes derived from the *TRNA-ILE*
192 gene sequence were observed, although this does not mean that they cannot appear in the future during
193 the evolution of this species (Table 1).

194   The number of somatic cell divisions from the zygote to meiosis (and the loss of the nuclear
195 envelope during each division) should influence the frequency of mitochondrion-to-nucleus DNA
196 transfer (36). It is therefore possible for fragments from all mtDNA genes, including *TRNA-ILE*, to
197 occur in the nuclear genome during embryogenesis independently in different individuals.

198   In the human genome, NUMTs are commonly associated with repetitive elements, suggesting
199 a possible role for transposable elements in mtDNA integration in the nuclear genome (31). Certain
200 NUMTs are repeated multiple times within the human genome (32, 37). In the case of the *D. magna*
201 genome, some NUMTs were also observed, which were repeated many times in different linkage
202 groups (no tabulated data). However, their association with repetitive elements in the nuclear genome
203 requires additional research. The average levels of NUMT sequence homology for the individual link-
204 age groups do not differ significantly from each other, which may indicate a random and even inclusion
205 of sequence fragments into each of them (Table 4).

206   The cytochrome C oxidase subunit I (*COI*) has possibly been the most commonly studied
207 marker. However, its popularity is mainly associated with its use as a maker for DNA barcoding of
208 animal diversity (38). There are several factors causing inadequacy of mtDNA in general and *COI*
209 individually, such as male-biased gene flow, selection on any mtDNA nucleotide(s) (as the whole

genome is one linkage group), retention of ancestral polymorphism, and introgression following hybridization (39). Presently, there are huge numbers of *COI* sequences in public databases, and most of them have a limited length, generally close to the length of the barcoding region. It is known that the possibility of the presence of NUMTs in the existing data should not be ignored (40). Since the success of taxonomic differentiation is positively correlated with the barcode length, the minibarcode length is usually kept above 100 bp. For example, an approximately 250-bp region of 16S rRNA can be successfully amplified from various medicinal preparations and food products. It provides correct identification of animal species (41, 42). Gene fragments that are often used for species identification in *D. magna* are in the following ranges: *COXI* (19-72 bp), *CYTB* (18-46 bp), 12s rRNA (18-72 bp), and 16s rRNA (19-78 bp); each of them constitutes less than 10% of the length of the entire gene (Table 3). Hence, the NUMT sequences of frequently analyzed genes are generally shorter than the respective mitochondrial sequence; thus, the possibility of NUMT coamplification should decrease with an increased length of the targeted mitochondrial marker (24, 43, 44). However, it is worth paying attention to the coverage of NUMTs derived from the *TRNA-ARG* (95%), *TRNA-GLU* (97%), and *TRNA-THR* (95%) genes (Table 3). Perhaps, in these cases, they are not NUMTs but functional genes coding for nuclear tRNA molecules, and the differences in homology and sequence length are evolutionary modifications resulting from the function performed in the nucleus, such as changes in the anti-codon region.

NUMTs are highly polymorphic in terms of the sequence, homo/heterozygosity status, and presence/absence at a specific locus (45). These features facilitate the use of NUMTs as specific population markers, as proposed for the human population by Lang et al. (46). The biological importance of NUMTs may correlate with their location on the chromosome. Depending on the location of the insertion, NUMTs may perturb the function of the genes (44). Additionally, *de novo* integration of NUMT pseudogenes into the nuclear genome has an adverse effect in some cases: promoting various disorders

234 and aging, as observed in humans (47). Chatre and Ricchetti (48) report that migratory mitochondrial
235 DNA can also have an impact on the replication of the nuclear region in *Saccharomyces cerevisiae*.

236

237 **Conclusions**

238 This article described the first occurrence of mitochondrial pseudogenic sequences (NUMTs)
239 in the nuclear genome of *D. magna*. There was no full sequence homology for two genes: *TRNA-SER1*
240 as well as *TRNA-ALA* and NUMTs. The total length of NUMTs in the nuclear genome was 44.391 bp
241 (from 16 to 182 bp), which accounted for 0.042% of the entire genome. The best-matched NUMTs
242 covering more than 90% of the mtDNA gene sequence were identified for the *TRNA-ARG* (95%),
243 *TRNA-GLU* (97%), and *TRNA-THR* (95%) genes, and they may be included in the functional nuclear
244 tRNA genes. The NUMT length varied from 16 to 63 bp for *tRNA* genes, from 17 to 108 bp for coding
245 genes, from 18 to 78 bp for *rRNA* genes, and from 17 to 182 bp for the D-loop region. Therefore, using
246 the product of total DNA isolation in mtDNA studies, coamplification of nDNA fragments is unlikely
247 in the case of amplification of the whole *tRNA* genes as well as fragments of other genes and the D-
248 loop with a length exceeding 200 bp. It was observed that fragments *TRNA-MET* (from 16 to 18 bp
249 length) had the highest level of sequence homology, which means that they could be evolutionarily the
250 youngest. The lowest degree of homology was found in the pseudogene derived from the mtDNA D-
251 loop sequence. It may probably be the oldest element of mitochondrial DNA incorporated into the
252 nuclear genome; however, due to the high degree of mutation in the D-loop, the thesis requires further
253 analysis and elucidation.

254

255 **Methods**

256 The whole sequence and annotation of the nuclear and mitochondrial genome of *D. magna*
257 were obtained from GenBank (the accession numbers for the nuclear and mitochondrial genome are
258 GCA_003990815.1 and NC_026914.1, respectively). The presence of NUMTs in the *D. magna*

nuclear genome GCA_003990815.1 was evaluated using the BLAST (BLASTN 2.6.0) (23) program implemented within the CLC Genomics Workbench 12.0 software package (<https://www.qiagenbioinformatics.com/>). The following parameters of the BLAST approach were implemented: number of threads 64; low complexity filter (to avoid hits to sequences that are not indeed related); Match/Mismatch and Gap Costs = Match 2 Mismatch 3 Existence 5 Extension 2; Max number of hit sequences 100.

265

266 **Abbreviations**

267 *D. magna* - *Daphnia magna*, NUMTs - nuclear copies of mitochondrial DNA, bp - base pairs, LG -
268 linkage groups, KIT - Korea Institute of Toxicology, EST - expressed sequence tags, nDNA - nuclear
269 DNA, mtDNA - mitochondrial DNA, BLAST - Basic Local Alignment Search Tool, *ND1-6 and ND4L*
270 - NADH dehydrogenase subunits 1–6 and subunit 4L, *COI-3 or COX1-3* - cytochrome oxidase subu-
271 nits I-III, *ATP6 and 8* - ATPase subunit 6 and 8, *CYTB* - cytochrome b, 12S rRNA - gene for small
272 subunit ribosomal RNA, 16S rRNA - Gene for large subunit ribosomal RNA, TRNA - gene coding
273 transfer RNAs

274

275

276 **Declarations**

277 **Acknowledgements**

278 Not applicable.

279

280 **Authors' contributions**

281 KK and BŚ designed this research, KK performed bioinformatics analyses, KK, AT, and BŚ analyzed
282 the data and wrote the manuscript, BŚ, AB, and MP reviewed and provided editorial advice.
283 BŚ supervised this research. All authors have read and approved the final version of the manuscript.

284

285 **Funding**

286 Not applicable.

287

288 **Ethics approval and consent to participate**

289 Not applicable.

290

291 **Availability of data and materials**

292 All data generated or analysed during this study are included in this published article in Supplementary
293 file 1.

294

295 **Consent for publication**

296 Not applicable.

297

298 **Competing interests**

The authors declare that they have no competing interests.

References

1. Ebert D. A Genome for the Environment. *Science*. 2011;331(6017):539-40.
2. Lampert W, Sommer U. *Limnoecology: the ecology of lakes and streams*. New York: Oxford university press; 2007.
3. Lampert W, *Daphnia: Development of a Model Organism in Ecology and Evolution*. 2011; Oldendorf/Luhe, Germany: International Ecology Institute.
4. Miner BE, De Meester L, Pfrender ME, Lampert W, Hairston NG, Jr. Linking genes to communities and ecosystems: *Daphnia* as an ecogenomic model. *Proc Biol Sci*. 2012;279(1735):1873-82.
5. Bownik A, Slaska B, Bochra J, Gumieniak K, Galek K. Procaine penicillin alters swimming behaviour and physiological parameters of *Daphnia magna*. *Environ Sci Pollut Res Int*. 2019;26(18):18662-73.
6. Bownik A, Jasieczek M, Kosztowny E. Ketoprofen affects swimming behavior and impairs physiological endpoints of *Daphnia magna*. *Science of The Total Environment*. 2020;725.
7. Bownik A, Szabelak A, Kulinska M, Waleka M. Effects of L-proline on swimming parameters of *Daphnia magna* subjected to heat stress. *J Therm Biol*. 2019;84:154-63.
8. Bownik A, Ślaska B, Dudka J. Cisplatin affects locomotor activity and physiological endpoints of *Daphnia magna*. *Journal of Hazardous Materials*. 2019;121259.
9. Persoone G, Baudo R, Cotman M, Blaise C, Thompson KC, Moreira-Santos M, et al. Review on the acute *Daphnia* magnatoxicity test – Evaluation of the sensitivity and the precision of assays performed with organisms from laboratory cultures or hatched from dormant eggs. *Knowledge and Management of Aquatic Ecosystems*. 2009(393).
10. OECD. Test No. 211: *Daphnia magna* Reproduction Test, OECD Guidelines for the Testing of Chemicals, Section 2. OECD Publishing, Paris. 2012.
11. Routtu J, Hall MD, Albere B, Beisel C, Bergeron RD, Chaturvedi A, et al. An SNP-based second-generation genetic map of *Daphnia magna* and its application to QTL analysis of phenotypic traits. *BMC genomics*. 2014;15:1033.
12. Routtu J, Jansen B, Colson I, De Meester L, Ebert D. The first-generation *Daphnia magna* linkage map. *BMC genomics*. 2010;11(1):508.
13. Dukić M, Berner D, Roesti M, Haag CR, Ebert D. A high-density genetic map reveals variation in recombination rate across the genome of *Daphnia magna*. *BMC Genetics*. 2016;17(1).
14. Lee BY, Choi BS, Kim MS, Park JC, Jeong CB, Han J, et al. The genome of the freshwater water flea *Daphnia magna*: A potential use for freshwater molecular ecotoxicology. *Aquatic Toxicology*. 2019;210:69-84.
15. Colson I, Du Pasquier L, Ebert D. Intragenic tandem repeats in *Daphnia magna*: structure, function and distribution. *BMC Res Notes*. 2009;2:206.
16. Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. Numt, a Recent Transfer and Tandem Amplification of Mitochondrial DNA to the Nuclear Genome of the Domestic Cat. *Journal of molecular evolution*. 1994;39(2):174-90.
17. Sun X, Yang A. Exceptionally large mitochondrial fragments to the nucleus in sequenced mollusk genomes. *Mitochondrial DNA Part A*. 2016;27(2):1409-10.
18. Song H, Moulton MJ, Hiatt KD, Whiting MF. Uncovering historical signature of mitochondrial DNA hidden in the nuclear genome: the biogeography of *Schistocerca* revisited. *Cladistics*. 2013;29(6): 643-62.
19. Leister D. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet*. 2005;21(12):655-63.
20. Williams ST, Knowlton N. Mitochondrial pseudogenes are pervasive and often insidious in the

349 snapping shrimp genus *Alpheus*. *Molecular Biology and Evolution*. 2001;18(8):1484-93.

350 21. Bensasson D, Zhang DX, Hartl DL, Hewitt GM. Mitochondrial pseudogenes: evolution's
351 misplaced witnesses. *Trends in ecology & evolution*. 2001;16(6): 314-21.

352 22. Verscheure S, Backeljau T, Desmyter S. In silico discovery of a nearly complete mitochondrial
353 genome Numt in the dog (*Canis lupus familiaris*) nuclear genome. *Genetica*. 2015;143(4):453-8.

354 23. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and
355 PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*.
356 1997;25(17):3389-402.

357 24. Hazkani-Covo E, Zeller RM, Martin W. Molecular Poltergeists: Mitochondrial DNA Copies
358 (numts) in Sequenced Nuclear Genomes. *PLoS Genet*. 2010;6(2):e1000834.

359 25. Viljakainen L, Oliveira DC, Werren JH, Behura SK. Transfers of mitochondrial DNA to the
360 nuclear genome in the wasp *Nasonia vitripennis*. *Insect Mol Biol*. 2010;19 Suppl 1:27-35.

361 26. Behura SK. Analysis of nuclear copies of mitochondrial sequences in honeybee (*Apis*
362 *mellifera*) genome. *Mol Biol Evol*. 2007;24(7):1492-505.

363 27. Song H, Moulton MJ, Whiting MF. Rampant nuclear insertion of mtDNA across diverse
364 lineages within Orthoptera (Insecta). *PLoS One*. 2014;9(10):e110508.

365 28. Shi H, Xing Y, Mao X. The little brown bat nuclear genome contains an entire mitochondrial
366 genome: Real or artifact? *Gene*. 2017;629:64-7.

367 29. Ishiguro N, Nakajima A, Horiuchi M, Shinagawa M. Multiple nuclear pseudogenes of
368 mitochondrial DNA exist in the canine genome. *Mamm Genome*. 2002;13(7):365-72.

369 30. Berthier K, Chapuis M-P, Moosavi SM, Tohidi-Esfahani D, Sword GA. Nuclear insertions and
370 heteroplasmy of mitochondrial DNA as two sources of intra-individual genomic variation in
371 grasshoppers. *Systematic Entomology*. 2011;36(2):285-99.

372 31. Mishmar D, Ruiz-Pesini E, Brandon M, Wallace DC. Mitochondrial DNA-like sequences in
373 the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration.
374 *Hum Mutat*. 2004;23(2):125-33.

375 32. Woischnik M, Moraes CT. Pattern of organization of human mitochondrial pseudogenes in the
376 nuclear genome. *Genome Res*. 2002;12(6):885-93.

377 33. Bertagnolli AC, Soares P, van Asch B, Amorim A, Cirnes L, Maximo V, et al. An assessment
378 of the clonality of the components of canine mixed mammary tumours by mitochondrial DNA analysis.
379 *Vet J*. 2009;182(2):269-74.

380 34. Murgia C, Pritchard JK, Kim SY, Fassati A, Weiss RA. Clonal origin and evolution of a
381 transmissible cancer. *Cell*. 2006;126(3):477-87.

382 35. Slaska B, Grzybowska-Szatowska L, Bugno-Poniewierska M, Surdyka M, Śmiech A. Nuclear
383 and mitochondrial DNA mutation in human and canine tumors. *Med Weter*. 2013;69(4):195-202.

384 36. Walbot V, Evans MM. Unique features of the plant life cycle and their consequences. *Nat Rev*
385 *Genet*. 2003;4(5):369-79.

386 37. Tourmen Y, Baris O, Dessen P, Jacques C, Malthiery Y, Reynier P. Structure and chromosomal
387 distribution of human mitochondrial pseudogenes. *Genomics*. 2002;80(1):71-7.

388 38. Hebert PD, Dewaard JR, Landry JF. DNA barcodes for 1/1000 of the animal kingdom. *Biol*
389 *Lett*. 2010;6(3):359-62.

390 39. Moritz C, Cicero C. DNA barcoding: promise and pitfalls. *PLoS Biol*. 2004;2(10):e354.

391 40. Bertheau C, Schuler H, Krumbock S, Arthofer W, Stauffer C. Hit or miss in phylogeographic
392 analyses: the case of the cryptic NUMTs. *Mol Ecol Resour*. 2011;11(6):1056-9.

393 41. Arulandhu AJ, Staats M, Hagelaar R, Voorhuijzen MM, Prins TW, Scholtens I, et al.
394 Development and validation of a multi-locus DNA metabarcoding method to identify endangered
395 species in complex samples. *Gigascience*. 2017;6(10):1-18.

396 42. Coghlan ML, Haile J, Houston J, Murray DC, White NE, Moolhuijzen P, et al. Deep sequencing
397 of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and
398 health safety concerns. *PLoS Genet*. 2012;8(4):e1002657.

- 399 43. Richly E, Leister D. NUMTs in sequenced eukaryotic genomes. *Mol Biol Evol.*
400 2004;21(6):1081-4.
- 401 44. Gaziev AI, Shaikhaev GO. Nuclear mitochondrial pseudogenes. *Molecular Biology.*
402 2010;44(3):358-68.
- 403 45. Hazkani-Covo E, Graur D. A comparative analysis of numt evolution in human and
404 chimpanzee. *Mol Biol Evol.* 2007;24(1):13-8.
- 405 46. Lang M, Sazzini M, Calabrese FM, Simone D, Boattini A, Romeo G, et al. Polymorphic NumtS
406 trace human population relationships. *Hum Genet.* 2012;131(5):757-71.
- 407 47. Dayama G, Emery SB, Kidd JM, Mills RE. The genomic landscape of polymorphic human
408 nuclear mitochondrial insertions. *Nucleic Acids Res.* 2014;42(20):12640-9.
- 409 48. Chatre L, Ricchetti M. Nuclear mitochondrial DNA activates replication in *Saccharomyces*
410 *cerevisiae*. *PLoS One.* 2011;6(3):e17235.

411

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryfile1.xlsx](#)