

Supplemental Information for:

## Expanding the search for sperm transmission elements in the mitochondrial genomes of bivalve mollusks

Donald T Stewart <sup>1, a, \*</sup>, Brent M Robicheau <sup>2, a</sup>, Noor Youssef <sup>2</sup>, Manuel A Garrido-Ramos <sup>3</sup>, Emily E Chase <sup>4</sup> and Sophie Breton <sup>5</sup>

**Table S1.** List of all mitochondrial genomes used in this study. Accession codes can be searched in NCBI [23] for additional information on original publication sources, sequence author(s), etc. Taxonomic name listed is as found in the original GenBank report; some taxonomic names are under debate and may have changed since these sequences were first reported.

Taxonomy from Sequence Definition	Accessions	Taxonomic Synonym Given at NCBI
<i>Acanthocardia tuberculata</i>	NC_008452	—
<i>Aculamprotula tientsinensis</i>	NC_029210	—
<i>Acuticosta chinensis</i>	NC_042472	—
<i>Alasmidonta heterodon</i>	NC_037431	—
<i>Alasmidonta varicosa</i>	NC_038155	—
<i>Anodonta anatina</i>	NC_022803	—
<i>Anodonta anatina</i>	KF030962	—
<i>Anodonta anatina</i>	KF030963	—
<i>Anodonta arcaiformis</i>	NC_026674	—
<i>Anodonta cygnea</i>	NC_036488	—
<i>Anodonta euscaphys</i>	NC_026792	<i>Anemina euscaphys</i>
<i>Anodonta lucida</i>	NC_026673	—
<i>Arconaia lanceolata</i>	NC_023955	<i>Lanceolaria lanceolata</i>
<i>Arctica islandica</i>	NC_022709	—
<i>Calypptogena magnifica</i>	NC_028724	—
<i>Cerastoderma edule</i>	NC_035728	—
<i>Coelomactra antiquata</i>	NC_021375	<i>Mactra antiquata</i>
<i>Crassostrea angulata</i>	NC_012648	—
<i>Crassostrea ariakensis</i>	NC_012650	—
<i>Crassostrea belcheri</i>	NC_037851	—
<i>Crassostrea gasar</i>	NC_027653	—
<i>Crassostrea gigas</i>	NC_001276	—
<i>Crassostrea hongkongensis</i>	NC_011518	—
<i>Crassostrea iredalei</i>	NC_013997	—
<i>Crassostrea nippona</i>	NC_015248	—
<i>Crassostrea sikamea</i>	NC_012649	—
<i>Crassostrea sp.</i>	NC_018763	—
<i>Crassostrea virginica</i>	NC_007175	—
<i>Cristaria plicata</i>	NC_012716	—
<i>Cumberlandia monodonta</i>	KU873124	—

<i>Cumberlandia monodonta</i>	NC_034846	—
<i>Cuneopsis capitata</i>	NC_042469	<i>Cuneopsis capitatus</i>
<i>Cuneopsis heudei</i>	NC_042471	—
<i>Cuneopsis pisciculus</i>	NC_026306	—
<i>Cyclina sinensis</i>	NC_029478	—
<i>Dahurinaia dahurica</i>	NC_023942	<i>Margaritifera dahurica</i>
<i>Donax semiestriatus</i>	NC_035984	—
<i>Donax trunculus</i>	NC_035985	—
<i>Donax variegatus</i>	NC_035986	—
<i>Donax vittatus</i>	NC_035987	—
<i>Dosinia altior</i>	NC_037916	—
<i>Dosinia japonica</i>	NC_038063	—
<i>Dosinia troscheli</i>	NC_037917	—
<i>Echyriddella menziesii</i>	KU873122	—
<i>Echyriddella menziesii</i>	NC_034845	—
<i>Fulvia mutica</i>	NC_022194	—
<i>Gari elongata</i>	NC_042422	—
<i>Gibbosula crassa</i>	NC_037942	—
<i>Hiatula acuta</i>	NC_042421	—
<i>Hippopus hippopus</i>	NC_039944	—
<i>Hyriopsis cumingii</i>	NC_011763	—
<i>Hyriopsis schlegelii</i>	NC_015110	<i>Sinohyriopsis schlegelii</i>
<i>Hyriopsis schlegelii</i>	HQ641407	<i>Sinohyriopsis schlegelii</i>
<i>Katharina tunicata</i> (Outgroup for tree)	NC_001636	—
<i>Lamprotula caveata</i>	KX091842	—
<i>Lamprotula caveata</i>	NC_030336	—
<i>Lamprotula coreana</i>	NC_026035	—
<i>Lamprotula gottschei</i>	KJ627225	—
<i>Lamprotula gottschei</i>	NC_023806	—
<i>Lamprotula leai</i>	NC_023346	<i>Lamprotula leaii</i>
<i>Lamprotula leai</i>	KC847114	<i>Lamprotula leaii</i>
<i>Lamprotula scripta</i>	NC_030258	—
<i>Lamprotula tortuosa</i>	NC_021404	—
<i>Lampsilis ornata</i>	NC_005335	—
<i>Lampsilis powellii</i>	NC_037720	—
<i>Lampsilis powellii</i>	MF326972	—
<i>Lampsilis siliquoidea</i>	MF326974	—
<i>Lampsilis siliquoidea</i>	NC_037721	—
<i>Lanceolaria grayana</i>	NC_026686	<i>Lanceolaria grayii</i>
<i>Lasmigona compressa</i>	NC_015481	—
<i>Lepidodesma languilati</i>	NC_029491	—
<i>Leptodea leptodon</i>	NC_028522	<i>Potamilus leptodon</i>
<i>Lutraria maxima</i>	NC_036766	—

<i>Lutraria rhynchaena</i>	NC_023384	—
<i>Mactra chinensis</i>	NC_025510	—
<i>Margaritifera falcata</i>	NC_015476	—
<i>Margaritifera margaritifera</i>	NC_043836	—
<i>Margaritifera marocana</i>	NC_034911	<i>Pseudunio marocanus</i>
<i>Margaritifera marocana</i>	KY131954	<i>Pseudunio marocanus</i>
<i>Meretrix lamarckii</i>	NC_016174	—
<i>Meretrix lamarckii</i>	KP244452	—
<i>Meretrix lusoria</i>	NC_014809	—
<i>Meretrix lyrata</i>	NC_022924	—
<i>Meretrix meretrix</i>	NC_013188	—
<i>Meretrix petechialis</i>	NC_012767	—
<i>Modiolus modiolus</i>	NC_033537	—
<i>Moerella iridescent</i>	NC_018371	<i>Iridona iridescent</i>
<i>Monodontina vondembuschiana</i>	NC_044112	—
<i>Mutela dubia</i>	NC_034844	—
<i>Mytilus californianus</i>	NC_015993	—
<i>Mytilus edulis</i>	AY823624	—
<i>Mytilus edulis</i>	NC_006161	—
<i>Mytilus galloprovincialis</i>	AY363687	—
<i>Mytilus galloprovincialis</i>	NC_006886	—
<i>Mytilus galloprovincialis</i>	DQ399833	—
<i>Mytilus trossulus</i>	GU936626	—
<i>Mytilus trossulus</i>	KM192127	—
<i>Mytilus trossulus</i>	NC_007687	—
<i>Nodularia douglasiae</i>	NC_026111	—
<i>Nuttallia olivacea</i>	NC_018373	—
<i>Ostrea denselamellosa</i>	NC_015231	—
<i>Ostrea edulis</i>	NC_016180	—
<i>Ostrea lurida</i>	NC_022688	—
<i>Paphia amabilis</i>	NC_016889	—
<i>Paphia euglypta</i>	NC_014579	—
<i>Paphia textile</i>	NC_016890	<i>Paratapes textilis</i>
<i>Paphia undulata</i>	NC_016891	<i>Paratapes undulatus</i>
<i>Pilsbryconcha exilis</i>	NC_044124	—
<i>Potamilus alatus</i>	NC_033858	—
<i>Potomida littoralis</i>	NC_030073	—
<i>Potomida littoralis</i>	KT247375	—
<i>Pyganodon grandis</i>	FJ809755	—
<i>Pyganodon grandis</i>	NC_013661	—
<i>Quadrula quadrula</i>	FJ809751	—
<i>Quadrula quadrula</i>	NC_013658	—
<i>Ruditapes decussatus</i>	NC_035757	—

<i>Ruditapes philippinarum</i>	NC_031332	—
<i>Saccostrea cucullata</i>	NC_027724	—
<i>Saccostrea echinata</i>	NC_036478	—
<i>Saccostrea glomerata</i>	NC_036483	—
<i>Saccostrea kegaki</i>	NC_030533	—
<i>Saccostrea mordax</i>	NC_013998	—
<i>Saccostrea mytiloides</i>	NC_036479	—
<i>Sanguinolaria ovalis</i>	NC_042423	—
<i>Saxidomus purpuratus</i>	NC_026728	<i>Saxidomus purpurata</i>
<i>Schistodesmus lampreyanus</i>	NC_042470	<i>Schistodesmus lampreyanus</i>
<i>Semele scabra</i>	NC_018374	—
<i>Sinanodonta woodiana</i>	KM434235	—
<i>Sinanodonta woodiana</i>	NC_024943	—
<i>Sinonovacula constricta</i>	NC_011075	—
<i>Solecurtus divaricatus</i>	NC_018376	—
<i>Solen grandis</i>	NC_016665	—
<i>Solen strictus</i>	NC_017616	—
<i>Solenia carinata</i>	NC_023250	—
<i>Solenia carinata</i>	KC848655	—
<i>Solenia oleivora</i>	NC_022701	—
<i>Solenia oleivora</i>	KY007143	—
<i>Solenia rivularis</i>	NC_039839	<i>Parvasolenia rivularis</i>
<i>Solenia rivularis</i>	KY007142	<i>Parvasolenia rivularis</i>
<i>Soletellina chinensis</i>	NC_042420	<i>Hiatula chinensis</i>
<i>Soletellina diphos</i>	NC_018372	<i>Hiatula diphos</i>
<i>Toxolasma parvus</i>	NC_015483	<i>Toxolasma parvum</i>
<i>Tridacna derasa</i>	NC_039945	—
<i>Tridacna squamosa</i>	NC_026558	—
<i>Unio crassus</i>	KY290450	—
<i>Unio crassus</i>	NC_033976	—
<i>Unio delphinus</i>	NC_033854	—
<i>Unio delphinus</i>	KT326918	—
<i>Unio douglasiae</i>	KP970613	<i>Nodularia douglasiae</i>
<i>Unio japonensis</i>	AB055624	<i>Pronodularia japonensis</i>
<i>Unio pictorum</i>	NC_015310	—
<i>Unio tumidus</i>	NC_033977	—
<i>Hyriopsis cumingii</i>	KC150028	—
<i>Limecola balthica</i>	MN528029	—
<i>Utterbackia imbecillis</i>	NC_015479	—
<i>Utterbackia peninsularis</i>	NC_015477	—
<i>Venerupis philippinarum</i>	AB065374	<i>Ruditapes philippinarum</i>
<i>Venustaconcha ellipsiformis</i>	NC_013659	—



and A; threshold = amount of minimum similarity (bp) required for a match.

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Tue Aug 6 11:48:36 2019
Modified on Fri May 21 2021
@author: Noor Youssef

Identifying STE motifs in random sequences
"""
import numpy as np

## Parameters to change ##
nuc_motif = "CCATAAATGTTGAAAATAAGG"

#specify sequence length
length = {"MgaRM" : 3590, "MgaSM": 1529, "MgaF": 1225, "MtrRM": 1070, "MedM": 993, "MedF": 1226,
"Sco": 1602, "Pte": 1986,
         "Can": 1285, "Fmu": 4368, "Mly": 4620, "Lba": 3968, "Hhi": 3027, "Ais": 1497, "Lle":
830, "Lsi": 575, "Tpa": 647, "Pgr": 575, "Ucr": 557}

#specify frequencies pi_T, pi_C, pi_G, pi_A
frequencies = { "MgaRM" : [28.2,15.8,20.9,35.0],
                "MgaSM" : [30.1,15.2,18.9,35.8],
                "MgaF" : [28.0,14.6,25.5,31.9],
                "MtrRM" : [27.9,15.3,19.5,37.3],
                "MedM" : [28.9,15.7,19.4,35.9],
                "MedF" : [28.1,14.3,25.1,32.5],
                "Sco" : [33.1,8.40,24.7,33.8],
                "Pte" : [26.8,11.9,25.4,35.9],
                "Can" : [31.4,27.6,13.3,27.6],
                "Fmu" : [30.7,12.3,22.1,34.9],
                "Mly" : [41.5,5.80,20.8,31.9],
                "Lba" : [35.8,10.6,20.9,32.7],
                "Hhi" : [33.4,12.1,25.7,28.8],
                "Ais" : [37.4,10.2,17.4,35.0],
                "Lle" : [37.5,8.40,26.5,27.6],
                "Lsi" : [34.8,11.1,23.8,30.3],
                "Tpa" : [32.0,10.8,25.8,31.4],
                "Pgr" : [33.7,11.7,19.8,34.8],
                "Ucr" : [31.2,13.5,20.3,35.0]}

# specify threshold
threshold = {"MgaRM" :16,
            "MgaSM" :17,
            "MgaF" :16,
            "MtrRM" :22,
            "MedM" :21,
            "MedF" :16,
            "Sco" :16,
            "Pte" :16,
            "Can" :16,
            "Fmu" :16,
            "Mly" :16,
            "Lba" :18,
            "Hhi" :17,
            "Ais" :16,
            "Lle" :16,
            "Lsi" :16,
            "Tpa" :16,
            "Pgr" :16,
            "Ucr" :16}

#####
nucleotide_converter = {"T":0, "C":1, "G":2, "A":3}
motif = [nucleotide_converter[i] for i in nuc_motif]

def count_num_of_hits(seq, motif, threshold):
    """
```

Counts the number of times (i.e., hits) a given motif is observed in a sequence given a threshold (e.g, 16 nucleotide matches)

```
seq: full sequence
motif: segment we are looking for
threshold: the number of positions that must match
'''
number_hits = 0

# sliding window over sequence
for i in range(len(seq) - len(motif) + 1):
    window = seq[i:i+len(motif)]

    #count the number of similar nucleotides between motif and window
    similarity = sum([1 for x in range(len(motif)) if window[x] == motif[x]])

    #if similarity greater than threshold, count as hit
    if similarity >= threshold:
        number_hits += 1
return(number_hits)

# for each organism extract sequence length, nucleotide frequencies, and threshold
for org in length.keys():
    seq_length = length[org]
    freq       = frequencies[org]
    thresh     = threshold[org]

num_seqs_with_motif = []; num_seqs_unifrom_with_motif = []
for trial in range(1000):

    #generate a random sequence generated from the empirical nucleotide frequency
    seq = np.random.choice(4, seq_length, p= [x/sum(freq) for x in freq])
    seq_unifrom = np.random.choice(4, seq_length, p= [0.25, 0.25, 0.25, 0.25])

    #count the number of times motif was present in the random sequence
    number_hits_per_seq = count_num_of_hits(seq, motif, thresh)
    number_hits_per_seq_unifrom = count_num_of_hits(seq_unifrom, motif, thresh)

    if number_hits_per_seq > 0:
        num_seqs_with_motif.append(number_hits_per_seq)

    if number_hits_per_seq_unifrom > 0:
        num_seqs_unifrom_with_motif.append(number_hits_per_seq_unifrom)

print("Sequences generated from empirical frequencies")
print(org, "Total number of hits", sum(num_seqs_with_motif), "Number of sequences with at-
least one hit", len(num_seqs_with_motif))

print("Sequences generated from uniform frequencies")
print(org, "Total number of hits", sum(num_seqs_unifrom_with_motif), "Number of sequences
with atleast one hit", len(num_seqs_unifrom_with_motif))
```