Genomics 112 (2020) 1686–1693

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

The first complete mitochondrial genome of the sand dollar *Sinaechinocyamus mai* (Echinoidea: Clypeasteroida)

Jih-Pai Lin[a,*], Mong-Hsun Tsai[b], Andreas Kroh[c], Aaron Trautman[d], Denis Jacob Machado[d,e], Lo-Yu Chang[a], Robert Reid[d], Kuan-Ting Lin[b], Omri Bronstein[g,h], Shyh-Jye Lee[f,i], Daniel Janies[d]

[a] Department of Geosciences, National Taiwan University, Taipei, Taiwan
[b] Institute of Biotechnology, National Taiwan University, Taipei, Taiwan
[c] Department of Geology and Palaeontology, Natural History Museum Vienna, Vienna, Austria
[d] Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, USA
[e] Bioinformatics Graduate Program, University of São Paulo, Brazil
[f] Department of Life Science, National Taiwan University, Taipei, Taiwan
[g] School of Zoology, Faculty of Life Sciences, Tel Aviv University, Tel-Aviv, Israel
[h] Steinhardt Museum of Natural History, Tel-Aviv, Israel
[i] Research Center for Developmental Biology and Regenerative Medicine, National Taiwan University, Taipei, Taiwan

ARTICLE INFO

Keywords:
10X genomics
Mitogenome
Echinodermata
Clypeasteroida
Phylogeny

ABSTRACT

Morphologic and molecular data often lead to different hypotheses of phylogenetic relationships. Such incongruence has been found in the echinoderm class Echinoidea. In particular, the phylogenetic status of the order Clypeasteroida is not well resolved. Complete mitochondrial genomes are currently available for 29 echinoid species, but no clypeasteroid had been sequenced to date. DNA extracted from a single live individual of *Sinaechinocyamus mai* was sequenced with 10× Genomics technology. This first complete mitochondrial genome (mitogenome) for the order Clypeasteroida is 15,756 base pairs in length. Phylogenomic analysis based on 34 ingroup taxa belonging to nine orders of the class Echinoidea show congruence between our new genetic inference and published trees based on morphologic characters, but also includes some intriguing differences that imply the need for additional investigation.

1. Introduction

Extant echinoids vary in size and shape, ranging from globose to flat to disk-shaped. Echinoid morphology reflects both their mode of life and feeding strategy [1]. A unique feature for echinoids among extant echinoderms is a pentaradial jaw apparatus internal to the peristome [2,3]. Based on the fossil record, the earliest crown group echinoid is a fossil cidaroid from the Permian of Texas [4]. The divergence of cidaroids and euechinoids is estimated at 268.8 Ma [4]. Hallmarks of the split between Cidaroidea and Euchinoidea are gene network variation in how micromeres and skeletogenic cells are specified (i.e., the presence in euechinoids and absence in cidaroids of the *Pmar1* gene; [5]).

Current studies suggest that evolutionary rates for major echinoid clades are heterogeneous [6]. For example, evolutionary rates for irregular urchins, which are echinoids that lack a pentaradial symmetry due to repositioning of the periproct [7], are higher than those of "regular" urchins [6]. Also, within clypeasteroid families, mean molecular divergence rates differ strongly within individual clades (e.g., in the Mellitidae [8]). The origin of the Clypeasteroida is one of the major issues in echinoid phylogeny discussed by Mooi [9] and Kroh and Smith [10]. In particular, the lantern structure of clypeasteroids is highly modified compared to other echinoids. Mooi [9] stated that "modification of the lantern into a crushing mill, extreme flattening of the test, and proliferation of food-gathering tube feet have allowed clypeasteroids to become epifaunal inhabitants of environments characterized by fine, shifting substrates, a habitat previously inaccessible to most other irregular echinoids".

The radiation of modern clypeasteroids has been rapid and complex [3,11]. For example, workers wielding genetic techniques have uncovered hidden diversification and cryptic speciation events of echinoids [7,12–15]. Bronstein et al. [16] argued that some specific regions in the mitochondrial genome are more suitable than others as phylogenetic markers. Several studies [17–21] have used complete echinoid mitogenomes to derive phylogenies, although their results were based on limited datasets both in terms of genetic and taxonomic sampling. The current study re-evaluates phylogenetic hypotheses for the

* Corresponding author.
E-mail address: alexjplin@ntu.edu.tw (J.-P. Lin).

https://doi.org/10.1016/j.ygeno.2019.10.007
Received 10 April 2019; Received in revised form 18 September 2019
Available online 17 October 2019

Table 1
Comparison of four output methods used in this study.

Assembly variables	*tGalore_raw*	*tGalore_pseudohap*	*kraken-dust_raw*	*kraken-dust_pseudohap*
No. of contigs (> = 0 bp)	5,453,916	245,534	5,506,042	246,212
No. of contigs (> = 1000 bp)	806,447	245,534	806,292	246,212
No. of contigs (> = 5000 bp)	46,632	40,194	45,638	40,484
No. of contigs (> = 10,000 bp)	2770	17,217	2676	17,070
No. of contigs (> = 25,000 bp)	4	5715	0	5638
No. of contigs (> = 50,000 bp)	0	1520	0	1541
Total length (> = 0 bp)	2,884,049,227	1,055,538,852	2,885,237,547	1,054,940,619
Total length (> = 1000 bp)	1,895,158,569	1,055,538,852	1,887,630,099	1,054,940,619
Total length (> = 5000 bp)	313,956,964	589,921,234	306,460,919	588,310,047
Total length (> = 10,000 bp)	33,566,538	436,845,399	32,511,574	432,239,794
Total length (> = 25,000 bp)	135,872	256,922,013	0	254,357,835
Total length (> = 50,000 bp)	0	112,142,064	0	113,023,271
No. of contigs (> = 500 bp)	1,349,506	245,534	1,353,382	246,212
Largest contig	40,267	800,148	24,596	653,181
Total length	2,281,563,514	1,055,538,852	2,276,793,623	1,054,940,619
GC (%)	37	37	37	37
N50	2249	6261	2231	6185
N75	1269	2897	1262	2887
L50	305,440	28,987	307,195	29,468
L75	643,203	93,930	646,159	94,629
No. of N's per 100 kbp	0	5392	0	5338

Echinoidea based on the largest mitogenome dataset to date.

Among 11 extant clypeasteroids recorded from Taiwan [22], *Sinaechinocyamus mai* (Wang, 1984) is one of the most studied sand dollars. Modern investigation of the living populations of *S. mai* began in the early 1990s. Chen and Chen [23] noted that the growth of oral plates is age-dependent but development of plates on the aboral surface is size-dependent. Chen and Chen [23] suggested neoteny (retention of juvenile features into gonadal maturity) rather than progenesis (interruption of growth by the early onset of gonadal maturity) hypothesized by Mooi [24] for the development of *Sinaechinocyamus* with respect to its putative sister taxon *Scaphechinus*.

The systematic position of *S. mai* is also in need of revision. For example, in a recent review of irregular echinoids, Schultz [25] suggested a reevaluation of the placement of *Sinaechinocyamus* among echinoids. Although miniaturized, *S. mai* possesses new characters that do not fit any other clypeasteroid family. Therefore, Wang [26] proposed a new superfamily, Taiwanasteritida Wang, 1984, and a new family Taiwanasteridae Wang, 1984 to place *S. mai* within his new classification of clypeasteroids. This new classification received much attention from echinoderm specialists. Due to the miniaturization of *Sinaechinocyamus*, Wang [26] thought it was related to other microechinoids, such as *Fibulariella acuta* (Yoshiwara, 1898) [27] but their phylogenetic relationships remained unresolved. Mooi [24] pointed out two errors in Wang's [26] study. First, *Taiwanaster* Wang, 1984 is a junior synonym of *Sinaechinocyamus* Liao, 1979 [28], and second, the superfamily Taiwanasteritida Wang, 1984, containing Fibulariidae and Taiwanasteridae, is polyphyletic. Mooi [24] and Mooi and Chen [29] concluded that *Sinaechinocyamus* (=*Taiwanaster*) is a derived scutelline sand dollar. Kroh and Smith [10] found additional phylogenetic evidence for the hypothesis that *Sinaechinocyamus* is a derived scutelline, and placed the family Taiwanasteridae as *incertae sedis* within Scutelliformes. Ziegler et al. [30] interpreted this family as a sister group to all taxa that possess Gregory's diverticulum. The goal of this study is to provide a better understanding of the phylogenetic placement of the enigmatic *Sinaechinocyamus* within Clypeasteroida and provide novel data for echinoid mitogenomes, which shall improve our understanding on echinoid phylogenetics.

2. Materials and methods

2.1. Data acquisition

Live specimens were observed in their native habitat in Miaoli County, Taiwan (120°39′E; 24°29′N) and a dozen live specimens were collected in July of 2017 and May of 2018 by JPL. Presented data is based on genome extraction of a single individual. Genomic DNA was extracted with the RecoverEase DNA isolation kit (Agilent Technologies) and size-selected for molecules of 40 kb or larger using a BluePippin (Sage Science) device. DNA was quantified with Life Technologies Qubit and Agilent Technologies Tapestation 4200. For library preparation 0.625 ng size-selected genomic DNA was used. The barcoded library was prepared with the Chromium gel bead and library kit (10× Genomics), and sequencing was performed with an Illumina Novaseq 6000 sequencer with 151 bp paired-end reads. All read pairs contain a 16-base barcode.

2.2. Identification of sequences

We employed three different strategies to identify and filter the sequences. Filtering sequence reads for both poor quality and adapters were performed via Trim Galore (Babraham Bioinformatics). Trim Galore removes regions of low quality, presumed contaminants, from the pool of reads. In our case, Trim Galore found no contaminated reads. Next, we used Kraken Version 1 [31] to identify foreign reads. Kraken uses user-selected genomes from the NCBI repository to build a database of k-mers that are used to identify foreign reads. We selected a custom-built database from RefSeq bacterial, archaeal, viral, and human genomic libraries as specified in the Kraken manual. Since repetitive DNA can cause false positive matches when comparing two genomes or sequences [32], repeat masking has become a crucial step in many sequence analysis applications like de novo assembly or annotation. Hence, a third strategy included use of dustMasker [33] to remove low complexity regions before assembly.

2.3. Sequence assembly

We used the software Supernova from 10× Genomics for assembly of contigs. Supernova can output multiple types of assembly formats that can be used for different downstream applications. We generated two output formats, raw and pseudohap (Table 1) with Supernova. The

Table 2
Comparison of different filtering strategies for contig assembly.

	A	B	C	D	E	F
# contigs > = 0 bp)	5,535,720	5,453,916	5,319,124	5,399,869	5,506,042	5,533,328
# contigs > = 1000 bp)	808,166	806,447	787,399	793,580	806,292	807,812
# contigs > = 5000 bp)	44,610	46,632	35,758	36,454	45,638	44,522
# contigs > = 10,000 bp)	2586	2770	1740	1798	2676	2572
# contigs > = 25,000 bp)	4	4	2	0	0	4
# contigs > = 50,000 bp)	0	0	0	0	0	0
Total length > = 0 bp)	2,888,834,538	2,884,049,227	2,769,697,116	2,798,260,859	2,885,237,547	2,887,086,409
Total length > = 1000 bp)	1,884,446,997	1,895,158,569	1,761,011,573	1,779,110,006	1,887,630,099	1,882,965,823
Total length > = 5000 bp)	299,342,674	313,956,964	235,523,667	240,764,442	306,460,919	298,743,352
Total length > = 10,000 bp)	31,400,084	33,566,538	20,697,858	21,740,546	32,511,574	31,237,488
Total length > = 25,000 bp)	130,108	135,872	53,664	0	0	130,108
Total length > = 50,000 bp)	0	0	0	0	0	0
No. of contigs > = 500 bp	1,358,943	1,349,506	1,363,143	1,370,186	1,353,382	1,358,460
Largest contig	40,051	40,267	26,832	24,377	24,596	40,051
Total length	2,276,260,880	2,281,563,514	2,170,686,674	2,189,414,574	2,276,793,623	2,274,681,176
GC (%)	36.59	36.6	36.71	36.72	36.6	36.59
N50	2217	2249	2064	2073	2231	2216
N75	1257	1269	1184	1188	1262	1256
L50	309,406	305,440	318,069	319,171	307,195	309,400
L75	650,141	643,203	665,165	667,886	646,159	650,020
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00

A, unfiltered raw output; B, Trim Galore filtering method; C, Kraken with all filtering options; D, Kraken filtering human only; E, Dusted Kraken with all filtering options; F, Dusted Kraken filtering human only

Table 3
Annotation of the mitogenome of *Sinaechinocyamus mai* based on the output from MITOS Server [48] and verification by amino acid translation and comparison to other complete echinoid mitogenomes.

Name	Anticodon	Start	Stop	Strand	Length	Dist. next feature (nc)	Start/Stop Codons
trnF	GAA	1	74	+	74	−1	
12S		74	966	+	893	0	
trnE	TTC	967	1036	+	70	18	
trnT	TGT	1055	1125	+	71	0	
Control Region		1126	1273	+	148	0	
trnP	TGG	1274	1343	+	70	14	
trnQ	TTG	1358	1428	−	71	8	
trnN	GTT	1437	1508	+	72	0	
trnL1	TAG	1509	1580	+	72	−1	
trnA	TGC	1580	1649	−	70	4	
trnW	TCA	1654	1722	+	69	0	
trnC	GCA	1723	1790	+	68	0	
trnV	TAC	1791	1860	−	70	22	
trnM	CAT	1883	1956	+	74	1	
trnD	GTC	1958	2026	−	69	4	
trnY	GTA	2031	2101	+	71	0	
trnG	TCC	2102	2169	+	68	1	
trnL2	TAA	2171	2243	+	73	4	
ND1		2248	3219	+	972	4	ATG/TAA
trnI	GAT	3224	3295	+	72	0	
ND2		3296	4351	+	1056	3	ATG/TAA
16S		4355	5912	+	1558	−16	
COX1		5897	7450	+	1554	6	ATG/TAG
trnR	TCG	7457	7524	+	68	1	
ND4L		7526	7819	+	294	6	ATT/TAG
COX2		7826	8515	+	690	4	ATG/TAA
trnK	CTT	8520	8589	+	70	0	
ATP8		8590	8757	+	168	−10	GTG/TAA
ATP6		8748	9437	+	690	3	ATG/TAG
COX3		9441	10,223	+	783	−1	ATG/TAA
trnS2	TGA	10,223	10,292	−	70	16	
ND3		10,309	10,659	+	351	19	ATG/TAA
ND4		10,679	12,067	+	1389	−10	ATG/TAG
trnH	GTG	12,058	12,125	+	68	1	
trnS1	GCT	12,127	12,194	+	68	0	
ND5		12,195	14,108	+	1914	1	ATG/TAA
ND6		14,110	14,598	−	489	16	ATG/TAA
CYTB		14,615	1	+	1143	−1	ATG/TAG

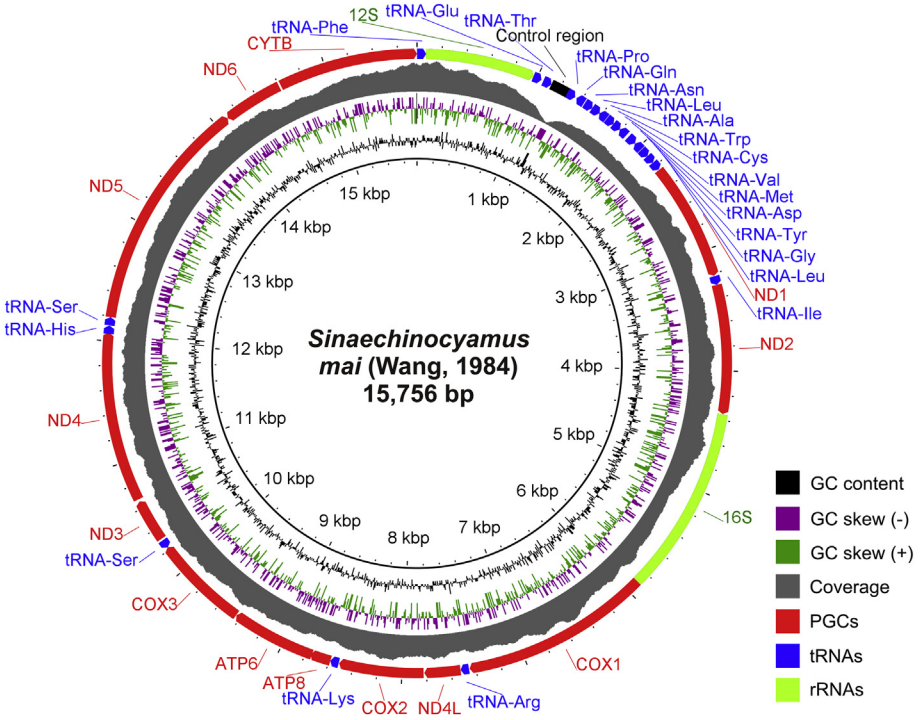

Fig. 1. The complete mitogenome map of the sand dollar *Sinaechinocyamus mai* (Wang, 1984) generated with BRIG [36]. Coverage (gray skyline plot) shows coverage (21× at CR to 819× near 5′ end of 12S); generated by mapping the reads on the mitogenome sequence with Bowtie2 [35].

Table 4
Taxa used in the phylogenetic analysis: 34 ingroup species plus *Holothuria scabra* (indicated with an asterisk) as the outgroup. Key taxa reported with partial mitochondrial genomic information (indicated with two asterisks) were included in the analysis.

Taxon	Order, Family	Base pairs	GenBank accession number	PCGs dataset
Arbacia lixula	Arbacioida, Arbaciidae	15,719	X80396.1	NC001770
*Conolampas sigsbei***	Echinolampadoia, Echinolampadidae	1286	AJ639902; AJ639800	
Diadema setosum	Diadematoida, Diadematidae	15,708	KX385835.1	NC033522
Echinocardium cordatum	Spatangoida, Loveniidae	15,767	FN562581.1	NC013881
*Echinolampas crassa***	Echinolampadoia, Echinolampadidae	642	DQ073744	
Echinometra mathaei	Camarodonta, Echinometridae	15,699	KJ680291.1	NC034767
*Echinoneus cyclostomus***	Echinoneoida, Echinoneidae	1191	AJ639801; AJ639903	
Echinothrix diadema	Diadematoida, Diadematidae	15,712	KX385836.1	NC033523
*Encope micropora borealis***	Clypeasteroida, Mellitidae	2544	MF616991; MF617495; MF617327	
Eucidaris tribuloides	Cidaroida, Cidaridae	15,890	JZLH01S0471648.1	
Glyptocidaris crenularis	Stomopneustoida, Glyptocidaridae	15,760	KX638403.1	NC032365
Heliocidaris crassispina	Camarodonta, Echinometridae	15,702	KC479025.1	NC023774
Hemicentrotus pulcherrimus	Camarodonta, Strongylocentrotidae	15,705	KC490911.1	NC023771
Heterocentrotus mammillatus	Camarodonta, Echinometridae	15,729	KJ680292.1	NC034768
*Holothuria scabra**	Holothuriida, Holothuriidae	15,779	KP257577.1	
Loxechinus albus	Camarodonta, Parechinidae	15,709	KC490910.1	NC023770
Lytechinus variegatus	Camarodonta, Toxopneutidae	15,693	MG676468.1	MG676468
Mesocentrotus franciscanus	Camarodonta, Strongylocentrotidae	15,713	MG676467.1	MG676467
Mesocentrotus nudus	Camarodonta, Strongylocentrotidae	15,709	JX263663.1	NC020771
Mespilia globulus	Camarodonta, Toxopneutidae	15,719	KJ680293.1	NC034769
Nacospatangus altus	Spatangoida, Maretiidae	15,763	KC990834.1	NC023255
Paracentrotus lividus	Camarodonta, Parechinidae	15,696	J04815.1	NC001572
Pseudocentrotus depressus	Camarodonta, Strongylocentrotidae	15,729	KC490913.1	NC023773
Salmacis bicolor raripina	Camarodonta, Temnopleuridae	15,767	KU302104.1	KU302104
Salmacis sphaeroides	Camarodonta, Temnopleuridae	15,762	KU302103.1	NC033528
Sinaechinocyamus mai	Clypeasteroida, Taiwanasteridae	15,756	**This study**	**This study**
Sterechinus neumayeri	Camarodonta, Echinidae	15,707	KJ680295.1	KJ680295
Strongylocentrotus droebachiensis	Camarodonta, Strongylocentrotidae	15,717	AM900391.1	EU054306
Strongylocentrotus intermedius	Camarodonta, Strongylocentrotidae	15,700		KY964300
Strongylocentrotus intermedius	Camarodonta, Strongylocentrotidae	15,700	KC490912.1	KY964299
Strongylocentrotus pallidus	Camarodonta, Strongylocentrotidae	15,712	AM900392.1	NC009941
Strongylocentrotus purpuratus	Camarodonta, Strongylocentrotidae	15,650	X12631.1	NC001453
Temnopleurus hardwickii	Camarodonta, Temnopleuridae	15,696	KP070768.1	NC026200
Temnopleurus reevesii	Camarodonta, Temnopleuridae	15,710	KU302106.1	NC033530
Temnopleurus toreumaticus	Camarodonta, Temnopleuridae	15,722	KU302105.1	NC033529
Tripneustes gratilla	Camarodonta, Toxopneutidae	15,725	KY268294.1	KY268294

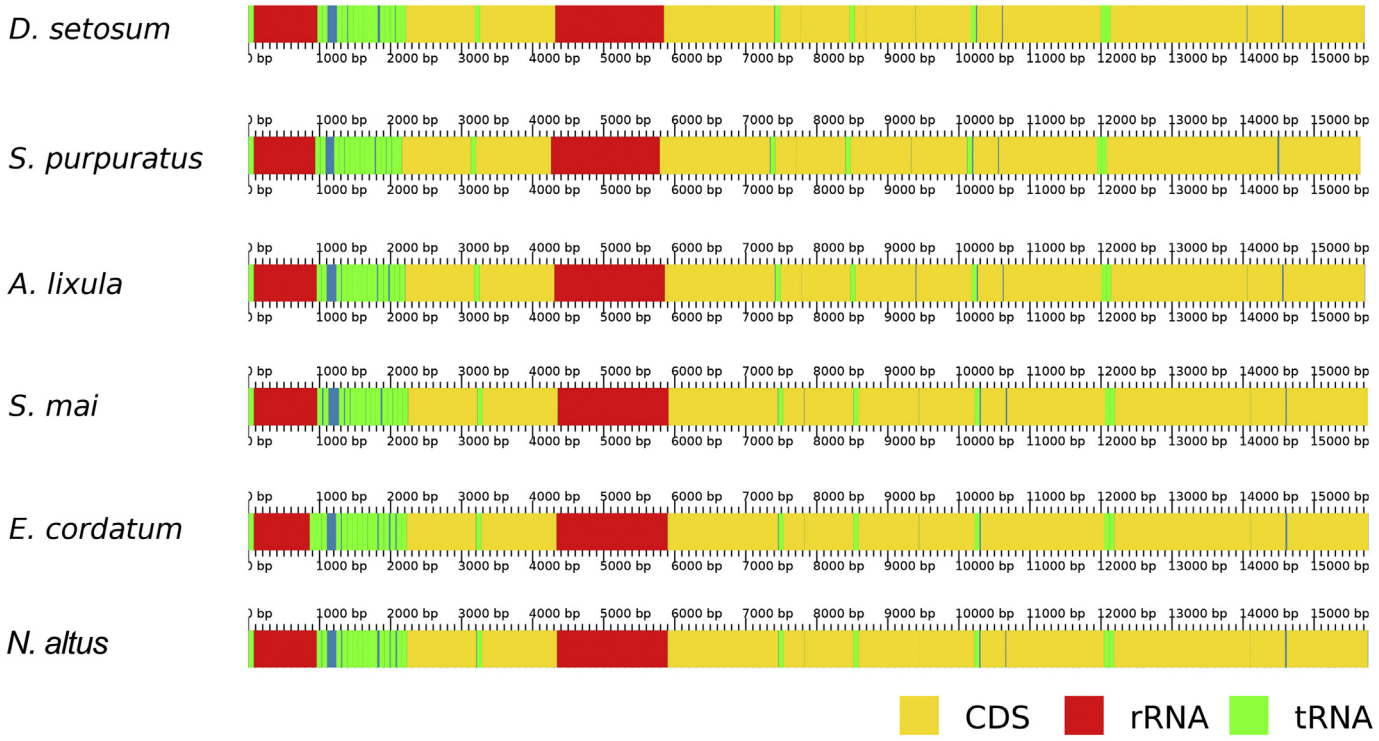

Fig. 2. Comparison of mitogenomes of *S. mai* with *Echinocardium cordatum* (FN562581), *Nacospatangus alta* (KC990834), *Arbacia lixula* (X80396), *Strongylocentrotus purpuratus* (X12631), and *Diadema setosum* (KX385835), based on the software AliTV [49]. The x-axis denotes the site of the feature on the mitochondrial genome. See text for further explanation.

raw output yields contigs without any modifications. The pseudohap output attempts to generate a single record per scaffold, which can provide useful information. We used QUAST to calculate assembly statistics prior to analyzing the generated assemblies and outputs. The statistics of the assemblies generated from the different filtering strategies are presented in Table 2. To compare the raw and pseudohap output types of the 10× Genomics de novo assembly pipeline, we used the two most effective filtering strategies: Trim Galore filtering method and Dusted Kraken filtering human only (B, F in Table 2).

2.4. Identification, annotation and mapping of the mitogenome

We identified contigs representing mitochondrial sequences by blasting the assembly against known echinoid mitogenomes. For the mitogenome assembly and circularization test AWA (https://gitlab.com/MachadoDJ/awa), we used k-mers between 21 and 41 [34]. Most k-mers within this range resulted in identical putative mitogenomes except that a k-mer length of 21 provided the worst assembly among all of the selected k-mer sizes. The other k-mers, when analyzed by AWA, gave longer putative mitogenomes, identical to each other, with slightly better statistics: average coverage = 8.64, average contiguity index = 8.51, average quality = 36.42, average alignment score = −2.03. The complete mitogenome sequence was verified and checked for circularity by mapping the raw reads back to the putative mitogenome sequence with Bowtie2 [35]. We annotated contigs by combining MITOS and GeSeq automatic annotation programs. Annotations were verified by manual searching of individual translated nucleotides with the program blastX. Annotation of *S. mai*'s mitogenome is included in Table 3. The mitogenome map was created with the software BRIG [36] (Fig. 1).

2.5. Multiple sequence alignment for phylogenetic analysis

We assembled a nucleotide dataset for the mitogenome of *S. mai*, and 34 additional echinoid mitogenomes representing 16 families and nine orders. The mitogenome of the holothuroid *Holothuria scabra* (KP257577.1) was used as outgroup. We created a multiple sequence alignment with the program Einsi which is part of the MAFFT package [37], and trimmed ragged edges of the alignment and marked missing data with "?". ModelGenerator [38] was used to select the best-fit model of nucleotide substitution for the phylogenetic analysis. Both the Akaike information criterion (AIC) and Bayesian information criterion (BIC) in ModelGenerator indicated GTR + G + I as the best-fitting substitution model for the dataset analyzed (Fig. 3). Under this model, we used the bootstrapping algorithm plus maximum likelihood tree search (commands –f a) with 1000 replicates using RAxML V. 8.2.12 [39]. Runs with standard bootstrap and tree search (commands –f b) were also performed but the results did not change.

In separate analyses, protein coding genes (PCGs) were converted into amino acid sequences and compared with 28 additional published echinoid sequences. For the amino acid translated analyses, all 13 PCGs from each taxon (3695 AA in total) were extracted, concatenated and analyzed using with both Maximum Likelihood and Bayesian Inference approaches following the methodologies in Bronstein and Kroh [17] (Fig. 4 and Table 4). Taxa included in the phylogenetic analysis are listed in Table 4.

2.6. Data deposition

The assembled mitogenome of *S. mai* was deposited to the NCBI GenBank and is available under accession number MN103227.

3. Results and discussion

The Trim Galore strategy (B in Table 2) generated the highest number of long contigs (≥ 10,000 bp) although no potential human contamination had been removed at this stage. Comparing Kraken (C, D in Table 2) with the dusted Kraken (E, F in Table 2) outputs shows that

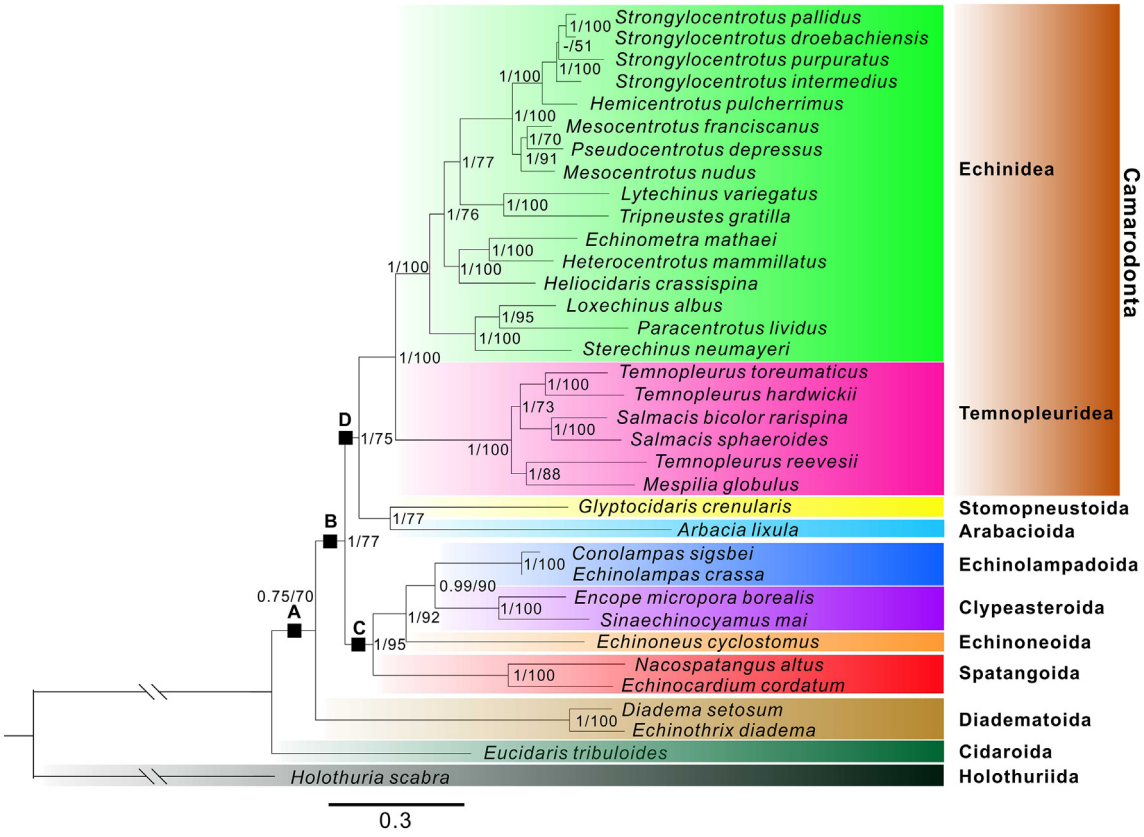

Fig. 3. Phylogenetic reconstruction based on nucleotid sequences of Echinoidea (34 echinoid species; Table 4) with RAxML using 1000 replications, rooted on the outgroup *Holothuria scabra*. The substitution model is GTR + G + I. Bootstrap values are shown at each node. Support values for the Bayesian analyses (Bayesian posterior probabilities with 10 million generations; discarding 25% as burnin) and Maximum Likelihood analyses (bootstrap support with 1000 replications) are shown next to nodes. Major monophyletic clades indicated by solid squares: A, Euechinoidea; B, Acroechinoidea; C, Irregularia; D, Echinacea. Branch lengths were produced with ML analyses.

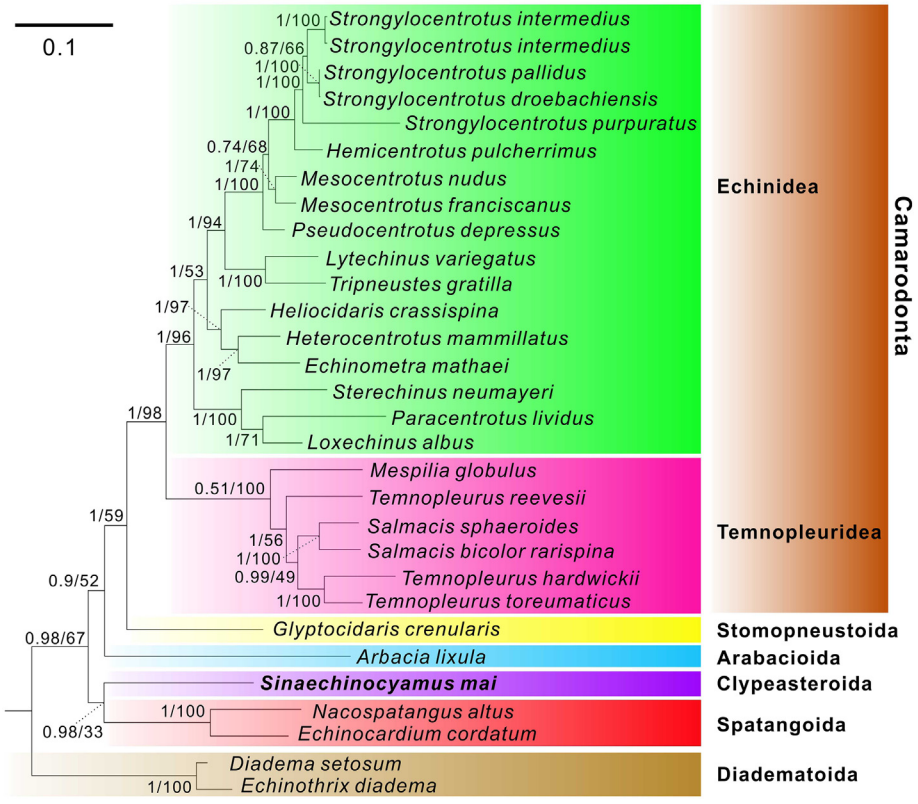

Fig. 4. ML phylogenetic tree based on amino acid sequences (3695 AA) translated from the concatenated PCGs rooted on all other echinoderm classes. Support values for the Bayesian analyses (Bayesian posterior probabilities with 10 million generations) and Maximum Likelihood analyses (bootstrap support with 1000 replications) are shown next to nodes. Branch lengths were produced with ML analyses.

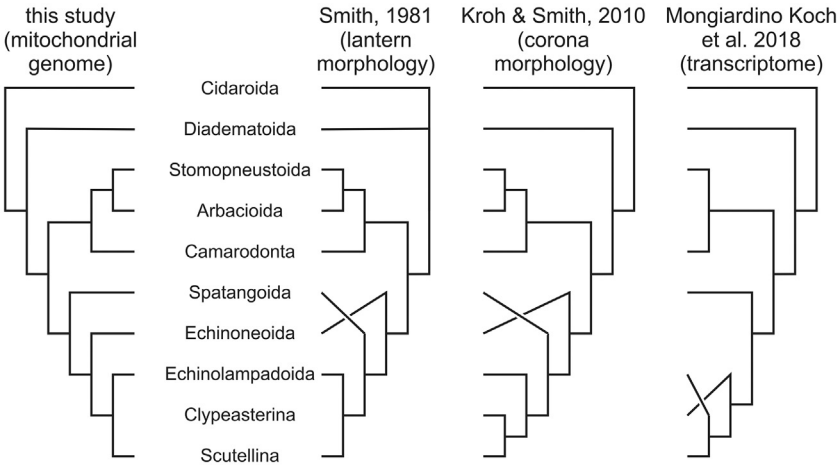

Fig. 5. Comparison of the phylogenetic trees of nine major echinoid clades based on different data. From left to right: mitochondrial genomes (this study), lantern structure (Smith [46]), corona morphology (Kroh and Smith [10]), and transcriptomic data (Mongiardino Koch et al. [43]). Note that some clades are not represented in some of the analyses.

Kraken alone eliminated too many reads and generated unsatisfactory results for all of the evaluation parameters except for the size of the largest contig. Approximately 85% of the raw output contigs produced by Kraken were smaller than 1000 base pairs.

In contrast, for each filtering strategy, none of the pseudohap contigs is smaller than 1000 base pairs. Also, approximately 5% of the pseudohap contigs are > 10,000 base pairs. This demonstrates that the pseudohap output can generate larger and more useful contigs. However, the pseudohap algorithm cannot be used to create an entire mitochondrial genome assembly as it did not capture all of the DNA. The overall assembly size from pseudohap was smaller, about half the length of the raw output and far less than the expected mitogenome size of *S. mai*.

While Kraken is good at identifying sequence fragments that belong to other organisms, an additional metagenomic analysis was performed in order to check for contamination. Thus, Bracken, which is a Bayesian re-estimation of Kraken data, was used to determine per-species abundance. We found that approximately 9% of the total DNA reads were classified by Kraken as belonging to other organisms, either viral, bacterial, archaeal, or human. In addition, Bracken also provides an output of the estimated species abundance that can be used for any number of other metagenomic analyses. In this case, the main concern is human contamination. Bracken recognized that approximately 21% of the identified contamination reads as human in origin. Other types of contamination, including sequences identified as either bacterial, archaeal, or viral, are likely due to the fact that the samples were harvested from their natural environment.

From our multiple sequence alignment of *S. mai* and the other echinoid taxa, we observed a very conserved organization of mitogenomes across Echinoidea. As shown here (Fig. 2), there is no significant rearrangement nor translocation among six Echinoidea taxa belonging to five orders. The gene order among orders is remarkably consistent, and the relative size for each PCG is similar (Fig. 2). Orthologous protein coding genes (CDS) (n = 13), tRNA genes (n = 22), and rRNA (n = 2) genes are found in all echinoid mitogenomes we considered.

The Bootstrap support values are stronger when analyzing the entirety of the mitogenome nucleotide data rather than the amino acid data derived from PCGs (Fig. 4). By increasing the level of genetic information (> 16,779 aligned positions) with improved mitochondrial assembly methods, the tree based on the nucleotide alignment shown here recovers most clades that were originally discovered by morphology and allows interpretations of ingroup relationships in detail (Fig. 3). Deep nodes higher than family level, including Acroechinoidea, Echinacea, and Irregularia, receive good support values (≥80%). In particular, bootstrap support for Irregularia is 98%. Increased resolution within the Echinacea clade is likely due to higher taxon sampling. The Camarodonta are very well-supported as a clade (100% bootstrap support). Within this order, relationships among six families, Temnopleuridae, Echinidae, Parechinidae, Echinometridae, Toxopneustidae, and Strongylocentrotidae, are also well-supported. The tree based on the amino acid alignment (Fig. 4) with Bayesian analyses shows similar phylogenetic relationships with high posterior probabilities, even though fewer ingroup taxa were used.

Incongruence between molecular and morphological data is common for many extant echinoderms [40,41]. Smith [42] reviewed the long-standing issue regarding the lack of molecular support for a monophyletic Clypeasteroida. By comparing the phylogenomic analyses based on nucleotide (Fig. 3) and amino acid data (Fig. 4), deep nodes in the tree are better supported with nucleotide data. Furthermore, even though the key clypeasteroid and echinolampadoid sister taxa of *S. mai* (i.e., *C. sigsbei*, *E. crassa*, *E. cyclostomus*, and *E. micropora borealis*) are represented by partial genetic information in our analysis, they form a clade and provide context, in addition to already published suggestions [9,29], inter alia], for the placement of *S. mai* (Fig. 3). The availability of a mitogenome for *S. mai* sets the stage for exploring the systematic position of this enigmatic taxon once a wider range of other clypeasteroid mitochondrial genomes become available.

The clades of Echinolampadoida and Clypeasteroida both receive 100% bootstrap support. Furthermore, the clade of Echinolampadoida + Clypeasteroida is supported at 90% bootstrap. Firmly supported ingroup taxa also reflect the good support values obtained in the deeper corresponding nodes (e.g., 100% for Camarodonta, 95% for Irregularia, 75% for Echinacea, 77% for Acroechinoidea, and 70% for Euechinoidea). The latter clades were also recovered by other molecular studies [42,43]. Recent molecular studies [43,44] highlighted the disagreements between molecular trees and morphological trees in certain areas of the tree (e.g. monophyly of the clypeasteroids and acroechinoids). The cause of the mismatch between phylogenetic interpretations based on morphological and molecular data currently remains unresolved [45]. Recent phylogenomic work shed new light on this question. The analyses of Mongiardino Koch and colleagues [43] based on a transcriptomic dataset confirmed the results of Smith and coworkers [44] in placing *Conolampas sigsbei* (a member of the Echinolampadoida) as sister-group to Scutellina. Mongiardino Koch and colleagues [43] argued that many of the features used to support the monophyly of clypeasteroids in fact show substantial differences in expression in clypeasterines vs. scutellines. While our results indicate a good corroboration of the echinoid phylogeny based on lantern evolution as outlined in Smith [46] (Fig. 5), they cannot shed light on the question of clypeasteroid and acroechinoid monophyly since no mitogenomes for taxa crucial to investigate these questions are yet available. The main difference between our tree and that of Smith [46] (Fig. 5) is that our phylogenomic analyses agree with the hypothesis proposed in Mooi [9] that the presence of lanterns in adults was suppressed first in

the Irregularia, then the lantern re-appeared in adult clypeasteroids.

4. Conclusions

Due to advances in genetic research, the wealth of molecular information is thought by many to be superior to morphological data, and molecular analyses became the most commonly used method for phylogenetic studies among living organisms [47]. Our molecular analysis, based on > 15,000 bp for most ingroup taxa, is only in part congruent with Smith's (1981) classification based on morphological data (e.g., Fig. 5). A major difference demonstrated by the present data in comparison to previous phylogenetic trees is the position of echinoneoids (Fig. 5). Whereas morphological data (lantern evolution: Smith [46]; corona, lantern and spine structure: Kroh and Smith [10]) identify Echinoneoida as sister to a clade containing Neognathostomata (cassidulids and clypeasteroids) and Atelostomata (spatangoids and holasteroids), the genetic data presented here support a more derived position, namely a sister-group relationship between Neognathostomata and Echinoneoida. Additional data will be needed to verify if this is a function of taxon sampling or truly a new topology.

Acknowledgements

Kwen-Shen Lee and Chang-Po Chen are acknowledged for providing background information about *S. mai*. R. Swisher commented on an earlier draft of this work. This work is supported by Taiwan-ROC Ministry of Science and Technology grants (MOST 105-2116-M-002-012; MOST 106-2116-M-002-018; and MOST 107-2116-M-002-007). AK's and OB's research was funded by the Austrian Science Fund (FWF): project P29508-B25. We acknowledge the support of Department of Bioinformatics and Genomics, College of Computing and Informatics, and Graduate School of the University of North Carolina at Charlotte. The authors declare that there are no conflicts of interest.

References

[1] P.M. Kier, Rapid evolution in echinoids, Palaeontology 25 (1982) 1–9.
[2] J.W. Durham, H.B. Fell, A.G. Fischer, P.M. Kier, R.V. Melville, D.L. Pawson, C.D. Wagner, Echinoids, in: R.C. Moore (Ed.), Treatise on Invertebrate Paleontology Part U Echinodermata 3, The Geological Society of America Inc. Lin et al. and The University of Kansas Press, Boulder and Lawrence, 1966, pp. U211–U640.
[3] A. Ziegler, L. Schroder, M. Ogurreck, C. Faber, T. Stach, Evolution of a novel muscle design in sea urchins (Echinodermata: Echinoidea), PLoS One 7 (2012) e37520.
[4] J.R. Thompson, E. Petsios, E.H. Davidson, E.M. Erkenbrack, F. Gao, D.J. Bottjer, Reorganization of sea urchin gene regulatory networks at least 268 million years ago as revealed by oldest fossil cidaroid echinoid, Sci. Rep. 5 (2015) 15541.
[5] J.R. Thompson, E.M. Erkenbrack, V.F. Hinman, B.S. McCauley, E. Petsios, D.J. Bottjer, Paleogenomics of echinoids reveals an ancient origin for the double-negative specification of micromeres in sea urchins, PNAS 114 (2017) 5870–5877.
[6] M.J. Hopkins, A.B. Smith, Dynamic evolutionary change in post-Paleozoic echinoids and the importance of scale when interpreting changes in rates of evolution, PNAS 112 (2015) 3758–3763.
[7] T. Saucède, R. Mooi, B. David, Phylogeny and origin of Jurassic irregular echinoids (Echinodermata: Echinoidea), Geol. Mag. 144 (2007) 333–359.
[8] S.E. Coppard, K.S. Zigler, H.A. Lessios, Phylogeography of the sand dollar genus *Mellita*: cryptic speciation along the coasts of the Americas, Mol. Phylogenet. Evol. 69 (2013) 1033–1042.
[9] R. Mooi, Paedomorphosis, Aristotle's lantern, and the origin of the sand dollars (Echinodermata: Clypeasteroida), Paleobiology 16 (1990) 25–48.
[10] A. Kroh, A.B. Smith, The phylogeny and classification of post-Palaeozoic echinoids, J. Syst. Palaeontol. 8 (2010) 147–212.
[11] J. Ghiold, H. A, Biogeography and biogeographic history of clypeasteroid echinoids, J. Biogeogr. 13 (1986) 183–206.
[12] O. Bronstein, A. Kroh, E. Haring, Do genes lie? Mitochondrial capture masks the Red Sea collector urchin's true identity (Echinodermata: Echinoidea: *Tripneustes*), Mol. Phylogenet. Evol. 104 (2016) 1–13.
[13] O. Bronstein, A. Kroh, B. Tautscher, L. Liggins, E. Haring, Cryptic speciation in pantropical sea urchins: a case study of an edge-of-range population of *Tripneustes* from the Kermadec Islands, Sci. Rep. 7 (2017) 5948.
[14] O. Bronstein, Y. Loya, The taxonomy and phylogeny of *Echinometra* (Camarodonta: Echinometridae) from the Red Sea and Western Indian Ocean, PLoS One 8 (2013) e77374.
[15] S.E. Coppard, A new genus of mellitid sand dollar (Echinoidea: Mellitidae) from the eastern Pacific coast of the Americas, Zootaxa 4111 (2016) 158–166.
[16] O. Bronstein, A. Kroh, E. Haring, Mind the gap! The mitochondrial control region and its power as a phylogenetic marker in echinoids, BMC Evol. Biol. 18 (2018) 80.
[17] E.S. Balakirev, V.A. Pavlyuchkov, F.J. Ayala, Complete mitochondrial genome of the phenotypically-diverse sea urchin *Strongylocentrotus intermedius* (Strongylocentrotidae, Echinoidea), Mitochond. DNA Part B 2 (2017) 613–614.
[18] C. De Giorgi, A. Martiradonna, C. Lanave, C. Saccone, Complete sequence of the mitochondrial DNA in the sea urchin *Arbacia lixula*: conserved features of the echinoid mitochondrial genome, Mol. Phylogenet. Evol. 5 (1996) 323–332.
[19] H.T. Jacobs, D.J. Elliott, V.B. Math, A. Farquharson, Nucleotide sequence and gene organization of sea urchin mitochondrial DNA, J. Mol. Biol. 202 (1988) 185–217.
[20] C. Li, G. Wu, W. Fu, X. Zen, The complete mitochondrial genome of *Diadema setosum* (Aulodonta: diadematidae), Mitochond. DNA Part B 1 (2016) 873–874.
[21] M. Perseke, D. Bernhard, G. Fritzsch, F. Brümmer, P.F. Stadler, M. Schlegel, Mitochondrial genome evolution in Ophiuroidea, Echinoidea, and Holothuroidea: insights in phylogenetic relationships of Echinodermata, Mol. Phylogenet. Evol. 56 (2010) 201–211.
[22] H. Lee, J.-P. Lin, H.-C. Li, L.-Y. Chang, K.-S. Lee, S.-J. Lee, W.-J. Chen, A. Sankar, S.-C. Kang, Young colonization history of a widespread sand dollar (Echinodermata; Clypeasteroida) in western Taiwan, Quat. Int., (In press).
[23] C.-P. Chen, C.-M. Chao, Reduction of growth rate as the major process in the miniaturization of the sand dollar *Sinaechinocyamus mai*, Biol. Bull. 193 (1997) 90–96.
[24] R. Mooi, Progenetic miniaturization in the sand dollar *Sinaechinocyamus*: Implications for clypeasteroid phylogeny, in: C. De Ridder, P. Dubois, M.-C. Lahaye, M. Jangoux (Eds.), Echinoderm Research: Proceedings of the Second European Conference on Echinoderms, A. A. Balkema, Brussels, Belgium, 18–21 September, 1989, 1990, pp. 137–143.
[25] H. Schultz, Echinoidea, Volume 2: Echinoidea with Bilateral Symmetry, Irregularia, Walter de Gruyter GmbH, Berlin, 2017.
[26] C.-C. Wang, New classification of clypeasteroid echinoids, Proc. Geol. Soc. China 27 (1984) 119–152.
[27] T. Mortensen, A Monograph of the Echinoidea IV. 2. Clypeasteroida. Clypeasteridæ, Arachnoidæ, Fibulariidæ, Laganidæ and Scutellidæ, Atlas, C. A. Reitzel, Copenhagen, 1948.
[28] Y. Liao, A.M. Clark, The Echinoderms of Southern China, Science Press, Beijing, China, 1995.
[29] R. Mooi, C.-P. Chen, Weight belts, diverticula, and the phylogeny of the sand dollars, Bull. Mar. Sci. 58 (1996) 186–195.
[30] A. Ziegler, J. Lenihan, L.G. Zachos, C. Faber, R. Mooi, Comparative morphology and phylogenetic significance of Gregory's diverticulum in sand dollars (Echinoidea: Clypeasteroida), Organ. Divers. Evol. 16 (2016) 141–166.
[31] D.E. Wood, S.L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments, Genome Biol. 15 (2014) R46.
[32] M.C. Frith, A new repeat-masking method enables specific detection of homologous sequences, Nucleic Acids Res. 39 (2010) e23.
[33] A. Morgulis, E.M. Gertz, A.A. Schäffer, R. Agarwala, A fast and symmetric DUST implementation to mask low-complexity DNA sequences, J. Comput. Biol. 13 (2006) 1028–1040.
[34] J.D. Machado, D. Janies, C. Brouwer, T. Grant, A new strategy to infer circularity applied to four new complete frog mitogenomes, Ecol. Evol. 8 (2018) 4011–4018.
[35] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359.
[36] N.F. Alikhan, N.K. Petty, N.L. Ben Zakour, S.A. Beatson, BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons, BMC Genomics 12 (2011).
[37] K. Katoh, H. Toh, Recent developments in the MAFFT multiple sequence alignment program, Brief. Bioinform. 9 (2008) 286–298.
[38] T.M. Keane, C.J. Creevey, M.M. Pentony, T.J. Naughton, J.O. McLnerney, Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified, BMC Evol. Biol. 6 (2006) 29.
[39] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313.
[40] D. Janies, Phylogenetic relationships of extant echinoderm classes, Can. J. Zool. 79 (2001) 1232–1250.
[41] A.K. Miller, A.M. Kerr, G. Paulay, M. Reich, N.G. Wilson, J.I. Carvajal, G.W. Rouse, Molecular phylogeny of extant Holothuroidea (Echinodermata), Mol. Phylogenet. Evol. 111 (2017) 110–131.
[42] A.B. Smith, Sea Urchins (Echinoidea), Oxford University Press, Oxford, 2009.
[43] N. Mongiardino Koch, S.E. Coppard, H.A. Lessios, D.E.G. Briggs, R. Mooi, G.W. Rouse, A phylogenomic resolution of the sea urchin tree of life, BMC Evol. Biol. 18 (2018) 189.
[44] A.B. Smith, D. Pisani, J.A. Mackenzie-Dodds, B. Stockley, B.L. Webster, D.T.J. Littlewood, Testing the molecular clock: molecular and paleontological estimates of divergence times in the Echinoidea (Echinodermata), Mol. Biol. Evol. 23 (2006) 1832–1851.
[45] A.B. Smith, A. Kroh, Phylogeny of Sea Urchins, in: J.M. Lawrence (Ed.), Sea Urchins: Biology and Ecology, Elsevier, 2013, pp. 1–14.
[46] A.B. Smith, Implications of lantern morphology for the phylogeny of post-Palaeozoic echinoids, Palaeontology 24 (1981) 779–801.
[47] W.C. Wheeler, G. Giribet, Molecular data in systematics: A promise fulfilled, a future beckoning, in: D. Williams, M. Schmitt, Q. Wheeler (Eds.), The Future of Phylogenetic Systematics: The Legacy of Willi Hennig, Cambridge University Press, Cambridge, 2016, pp. 329–343.
[48] M. Bernt, A. Donath, F. Jühling, F. Externbrink, C. Florentz, G. Fritzsch, J. Pütz, M. Middendorf, P.F. Stadler, MITOS: Improved de novo metazoan mitochondrial genome annotation, Mol. Phylogenet. Evol. 69 (2013) 313–319.
[49] M.J. Ankenbrand, S. Hohlfeld, T. Hackl, F. Förster, AliTV—interactive visualization of whole genome comparisons, PeerJ Comp. Sci. 3 (2017).