CrossMark

Mind the gap! The mitochondrial control region and its power as a phylogenetic marker in echinoids

Omri Bronstein[1,2]* , Andreas Kroh[1] and Elisabeth Haring[2,3]

Abstract

Background: In Metazoa, mitochondrial markers are the most commonly used targets for inferring species-level molecular phylogenies due to their extremely low rate of recombination, maternal inheritance, ease of use and fast substitution rate in comparison to nuclear DNA. The mitochondrial control region (CR) is the main non-coding area of the mitochondrial genome and contains the mitochondrial origin of replication and transcription.
While sequences of the *cytochrome oxidase subunit 1* (*COI*) and *16S rRNA* genes are the prime mitochondrial markers in phylogenetic studies, the highly variable CR is typically ignored and not targeted in such analyses. However, the higher substitution rate of the CR can be harnessed to infer the phylogeny of closely related species, and the use of a non-coding region alleviates biases resulting from both directional and purifying selection. Additionally, complete mitochondrial genome assemblies utilizing next generation sequencing (NGS) data often show exceptionally low coverage at specific regions, including the CR. This can only be resolved by targeted sequencing of this region.

Results: Here we provide novel sequence data for the echinoid mitochondrial control region in over 40 species across the echinoid phylogenetic tree. We demonstrate the advantages of directly targeting the CR and adjacent tRNAs to facilitate complementing low coverage NGS data from complete mitochondrial genome assemblies. Finally, we test the performance of this region as a phylogenetic marker both in the lab and in phylogenetic analyses, and demonstrate its superior performance over the other available mitochondrial markers in echinoids.

Conclusions: Our target region of the mitochondrial CR (1) facilitates the first thorough investigation of this region across a wide range of echinoid taxa, (2) provides a tool for complementing missing data in NGS experiments, and (3) identifies the CR as a powerful, novel marker for phylogenetic inference in echinoids due to its high variability, lack of selection, and high compatibility across the entire class, outperforming conventional mitochondrial markers.

Keywords: Control region, Molecular phylogeny, Mitochondrial markers, NGS, Echinoidea, Sea urchins

Background

Over the past decade, the emerging field of massively parallel sequencing (aka next generation sequencing or NGS) has seen dramatic advances in methods and decrease in costs. In many fields and applications NGS technologies are rapidly replacing the more traditional Sanger sequencing [1]. Nevertheless, despite the

tremendous contribution of NGS technologies to fields such as metagenomics, forensics and clinical diagnostics, these advances are not without limitations. For one, the associated error rate of NGS platforms (~ 0.1-15%) is higher and the read length generally shorter (35-700 bp for short read approaches) [2] than those obtained by traditional Sanger sequencing [3]. Moreover, data accumulating from miscellaneous NGS studies show reduced sequence coverage in non-random regions of both nuclear and mitochondrial genomes [4, 5]. In particular, regulatory regions, such as CpG islands, promoter and 5'-UTR regions have been shown to be particularly prone to reduced coverage and poorer SNP-calling performance on

* Correspondence: omribronstein@gmail.com
[1]Natural History Museum Vienna, Geological-Palaeontological Department, 1010 Vienna, Austria
[2]Natural History Museum Vienna, Central Research Laboratories, 1010 Vienna, Austria
Full list of author information is available at the end of the article

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 2 of 15

NGS platforms [4]. Consequently, Sanger sequencing will most likely remain an essential component in DNA sequence acquisition in the foreseeable future.

Non-coding DNA sequences are segments of an organism's DNA that do not encode protein sequences. While some non-coding DNA is transcribed into functional non-coding RNA molecules (e.g., transfer RNAs, ribosomal RNAs, small nuclear RNAs, micro RNAs), other may function in transcriptional and translational regulation of protein-coding sequences or serve as the origin of DNA replication (to name a few possibilities). The mitochondrial control region (CR) is the longest non-coding region in animal mitochondrial DNA (mtDNA), and is considered the most variable region of the mitochondrial genome [6]. Within the CR, the displacement loop (or D-loop), which is often synonymously used in the literature with CR [7], is in fact a region within the CR comprising a third strand of DNA creating a semi-stable structure [8]. It is this region of the CR that is considered most polymorphic.

Apart from the general advantages of mitochondrial markers in animal phylogenetic studies, namely their maternal inheritance, lack of recombination, and fast rate of evolution [9, 10], several unique qualities make the CR a favoured marker sequence for genetic diversity analyses, in particular, its exceptionally fast evolutionary rate (even in comparison to the rest of the mitochondrial genome [11, 12]), polymorphic nature [13] and presumed selective neutrality as a non-coding region (but see [14, 15]). Consequently, this region has been widely used as a genetic marker in phylogenetic studies of various animals including vertebrate classes such as fish (e.g., [16, 17]), amphibians [18], reptiles [19], birds [20] and mammals [21, 22] as well as numerous invertebrate taxa (e.g., [23–26]. Nevertheless, despite being extremely useful for some species, several factors may hinder the utility of this marker in others. One or several repeat regions within the CR have been found in some species and these may have detrimental effects on PCR amplification, sequencing, or both [27, 28]. Furthermore, some species exhibit segmental duplications involving the CR (e.g., [25, 29, 30]), that, in some cases, leads to the formation of pseudogenes that may be co-amplified by PCR [31, 32]. Further problems might arise from possible homogenization between duplicated copies of the CR (e.g., [32, 33]). Many researchers have, therefore, avoided using this region for phylogenetics, focusing instead on protein or ribosomal RNA coding genes [34–36], and only rarely has the inferential potential of the CR been evaluated in comparison to coding regions (e.g., [37, 38]).

In echinoderms, as in most other taxa, phylogenetic studies have mainly exploited a limited set of markers.

The greater majority of studies utilised fragments of two mitochondrial regions: the *cytochrome c oxidase subunit 1* gene (*COI*) and the *16S ribosomal RNA* gene (*16S*) (e.g., [34–36, 39–45]). Indeed, despite the growing number of echinoderm molecular genetic studies over the past two decades, the limited variety of available markers, and in particular of markers applicable to a broad range of species (often referred to as 'universal primers'), left many gaps in the echinoderm phylogenetic tree. This situation persists even when restricting the discussion to echinoids. Ward et al. [36] for example, utilising a fragment of the *COI* gene, encountered amplification problems for about 10% of their species. Jeffery et al. [34] failed to amplify certain species altogether and had pseudogene complications with others. Smith and Kroh [42] highlighted the incompleteness of available DNA sequence data in camarodonts in their analyses of two nuclear genes (*18S* and *28S rRNA* genes) and the two mitochondrial genes (*COI* and *16S*). Interestingly, the latter authors also stated that "*COI data in isolation found radically different branching orders within individual camarodont families, and placed some bona fide echinometrids as basal members of the Strongylocentridae clade*", emphasizing the need for critical consideration of analyses solely hinging on *COI* data.

Similar to other groups of organisms, NGS data in echinoids suffers from markedly reduced coverage in the CR area (Fig. 1), decreasing the quality and completeness of echinoid mitochondrial genome assemblies. Here we present the development and usability of a new set of primers targeting the echinoid CR and adjacent tRNAs (hereafter termed "CRA", for CR and adjacent areas) to facilitate completing mitogenome assemblies based on NGS data, which often are characterized by low coverage in the control region. We demonstrate the high applicability of this region across the Echinoidea, both in the lab and in resulting phylogenies, and provide a phylogenetic analysis of a wide range of echinoid families based on this marker. Additionally, we utilise data extracted from all publicly available echinoid mitogenomes to evaluate the performance of the two most commonly used phylogenetic markers in echinoid studies (*COI* and *16S*) and compare them to both the CRA and the well-established echinoid morphological consensus phylogeny. CR sequence data hold great prospects for advancing our knowledge of echinoid phylogeny, of both distant and closely related species.

Methods

Taxon sampling

Sequence data from 110 specimens were included in the current analysis. Forty-nine of these sequences were generated as part of the current study, and the remaining were obtained from GenBank (Table 1). Newly generated

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 3 of 15

Fig. 1 Representation of echinoid complete mitochondrial genomes assembled from NGS data, showing gene annotation and coverage. The annotated genomes are represented by four echinoid species: *Hemicentrotus pulcherrimus*, *Strongylocentrotus fragilis*, *Mesocentrotus franciscanus*, and *Strongylocentrotus intermedius*, corresponding to GenBank accession numbers: KC898202, KC898198, KC898199, and KC898200, respectively. Annotations are given at the outer margin of the external circle. Concentric circles represent the corresponding coverage for each of the represented species mitogenomes. Data was obtained from Kober and Bernardi [86, 87]. Enlarged segment illustrates the position of the various primers used in the current study. Coverage was calculated in BRIG [88], after read mapping with Bowtie2 [89] (using the predefined alignment threshold "very-sensitive"). Annotations are based on those for *H. pulcherrimus* (GenBank accession no. NC_023771) and radial plots generated using BRIG

sequences were based on specimens deposited in museum collections as listed in Table 1 and comply with institutional, national, and international guidelines. The sequences obtained from GenBank included all 35 currently available complete echinoid mitochondrial genomes comprising both NGS and Sanger generated sequences. To evaluate the performance of our target region as a phylogenetic marker and the applicability of the primers across the Echinoidea, a broad range of echinoid taxa were sampled. Additionally, to test the applicability of the CRA primers for specimens at varying grades of preservation, we included material of varying quality, from freshly sampled tissue, through ethanol fixed collection material, to dried specimens nearly a century old. In total, 10 orders comprising 17 families, 34 genera and 45 species were represented in the current analysis (Table 1).

Development of the CRA primers

Primers were designed to flank the control region and D-loop in order to retrieve the full length of this target region. To search for highly conserved regions adjacent to the CR, all publicly available echinoid complete mitochondrial genomes were downloaded, primarily aligned using MAFFT v. 7.2 [46] and subsequently adjusted by eye using Bioedit v. 7.1.3 [47] and AliView v. 1.18.1 [48]. The highly variable nature of the CR, prevented developing a set of 'universal echinoid' primers suitable for a broad range of echinoid species. Consequently, our final CRA target included the genes for *12S rRNA* (partially), $tRNA^{Glu}$, $tRNA^{Thr}$, the CR, $tRNA^{Pro}$, $tRNA^{Gln}$, and a partial sequence of $tRNA^{Asn}$. Two sets of primers were developed: CR15fwd 5'-TACACATCGCCCGTCACTCT-3' (positioned at the 3′ end of the *12S* gene) and CR08rev 5'-TTAACGGCCAAGCGCCTTT-3′ (binding

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 4 of 15

Table 1 Detailed taxonomic placement and GenBank accession numbers of taxa included in the current study

Order, Family	Genus	Species	COI	16S	CRA	Voucher No.
Cidaroida, Cidaridae	*Cidaroid*	sp.	–	–	MG198151[b,d]	MNHN IE-2007-3745
Cidaroida, Cidaridae	*Eucidaris*	*metularia*	–	–	MG198152[b,d]	SMNH Ec 25,624
Cidaroida, Cidaridae	*Prionocidaris*	sp.	–	–	MG198153[b,d]	MNHN IE-2007-3764
Cidaroida, Cidaridae	*Prionocidaris*	sp.	–	–	MG198154[b,d]	MNHN IE-2013-8705
Echinothurioida, Echinothuriidae	*Araeosoma*	*splendens*	–	–	MG198163[b,d]	MNHN IE-2007-1143
Echinothurioida, Echinothuriidae	*Asthenosoma*	*varium*	–	–	MG198164[b,d]	SMNH Ec 25,628
Diadematoida, Diadematidae	*Diadema*	*setosum*	KX385835[c,b]	KX385835[c,b]	KX385835[c,b]	
Diadematoida, Diadematidae	*Diadema*	*setosum*	–	–	MG198159[b,d]	SMNH Ec 25,437
Diadematoida, Diadematidae	*Diadema*	*setosum*	–	–	MG198160[b,d]	SMNH Ec 25,625
Diadematoida, Diadematidae	*Diadema*	*setosum*	–	–	MG198161[b,d]	SMNH Ec 25,626
Diadematoida, Diadematidae	*Diadema*	*setosum*	–	–	MG198162[b,d]	SMNH Ec 25,627
Diadematoida, Diadematidae	*Echinothrix*	*diadema*	KX385836[c]	KX385836[c]	KX385836[c]	
Micropygoida, Micropygidae	*Micropyga*	*tuberculata*	–	–	MG198165[b,d]	MNHN IE-2007-1152
Salenioida, Saleniidae	*Salenia*	*phoinissa*	–	–	MG198166[b,d]	MNHN IE-2007-3765
Stomopneustoida, Glyptocidaridae	*Glyptocidaris*	*crenularis*	KX638403[c,a]	KX638403[c,a]	KX638403[c,a]	
Arbacioida, Arbaciidae	*Arbacia*	*lixula*	X80396[c,b]	X80396[c,b]	X80396[c,b]	
Camarodonta, Echinidae	*Sterechinus*	*neumayeri*	KJ680295[c]	KJ680295[c]	KJ680295[c]	
Camarodonta, Echinidae	*Sterechinus*	*neumayeri*	KF214257[c,a]	KF214257[c,a]	KF214257[c,a]	
Camarodonta, Echinometridae	*Echinometra*	*mathaei*	KJ680291[c]	KJ680291[c]	KJ680291[c]	
Camarodonta, Echinometridae	*Echinometra*	sp.	–	–	MG198122[b,d]	SMNH OB ZNZ16
Camarodonta, Echinometridae	*Echinometra*	sp.	–	–	MG198123[b,d]	SMNH OB ZNZ37
Camarodonta, Echinometridae	*Echinometra*	sp.	–	–	MG198124[b,d]	SMNH OB ZNZ54
Camarodonta, Echinometridae	*Heliocidaris*	*crassispina*	KC479025[c]	KC479025[c]	KC479025[c]	
Camarodonta, Echinometridae	*Heterocentrotus*	*mammillatus*	KJ680292[c]	KJ680292[c]	KJ680292[c]	
Camarodonta, Echinometridae	*Zenocentrotus*	*kellersi*	–	–	MG198150[b,d]	USNM E40502
Camarodonta, Parechinidae	*Loxechinus*	*albus*	JX888466[c,b]	JX888466[c,b]	JX888466[c,b]	
Camarodonta, Parechinidae	*Loxechinus*	*albus*	KC490910[c,b]	KC490910[c,b]	KC490910[c,b]	
Camarodonta, Parechinidae	*Paracentrotus*	*lividus*	J04815[c,b]	J04815[c,b]	J04815[c,b]	
Camarodonta, Parechinidae	*Paracentrotus*	*lividus*	–	–	MG198127[b,d]	SMNH Ec 25,622
Camarodonta, Strongylocentrotidae	*Hemicentrotus*	*pulcherrimus*	KC898202[c,a]	KC898202[c,a]	KC898202[c,a]	
Camarodonta, Strongylocentrotidae	*Hemicentrotus*	*pulcherrimus*	KC490911[c,b]	KC490911[c,b]	KC490911[c,b]	
Camarodonta, Strongylocentrotidae	*Mesocentrotus*	*franciscanus*	KJ526170[c,a]	KJ526170[c,a]	KJ526170[c,a]	
Camarodonta, Strongylocentrotidae	*Mesocentrotus*	*nudus*	JX263663[c,b]	JX263663[c,b]	JX263663[c,b]	
Camarodonta, Strongylocentrotidae	*Mesocentrotus*	*nudus*	KC898201[c,a]	KC898201[c,a]	KC898201[c,a]	
Camarodonta, Strongylocentrotidae	*Pseudocentrotus*	*depressus*	KC490913[c]	KC490913[c]	KC490913[c]	
Camarodonta, Strongylocentrotidae	*Strongylocentrotus*	*droebachiensis*	EU054306[c,a]	EU054306[c,a]	EU054306[c,a]	
Camarodonta, Strongylocentrotidae	*Strongylocentrotus*	*droebachiensis*	AM900391[c,b]	AM900391[c,b]	AM900391[c,b]	
Camarodonta, Strongylocentrotidae	*Strongylocentrotus*	*fragilis*	KC898198[c,a]	KC898198[c,a]	KC898198[c,a]	
Camarodonta, Strongylocentrotidae	*Strongylocentrotus*	*intermedius*	KC490912[c]	KC490912[c]	KC490912[c]	
Camarodonta, Strongylocentrotidae	*Strongylocentrotus*	*intermedius*	KC898200[c,a]	KC898200[c,a]	KC898200[c,a]	
Camarodonta, Strongylocentrotidae	*Strongylocentrotus*	*pallidus*	AM900392[c]	AM900392[c]	AM900392[c]	
Camarodonta, Strongylocentrotidae	*Strongylocentrotus*	*purpuratus*	X12631[c,b]	X12631[c,b]	X12631[c,b]	
Camarodonta, Temnopleuridae	*Mespilia*	*globulus*	KJ680293[c]	KJ680293[c]	KJ680293[c]	
Camarodonta, Temnopleuridae	*Microcyphus*	*rousseaui*	–	–	MG198126[b,d]	SMNH Ec 25,621

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 5 of 15

Table 1 Detailed taxonomic placement and GenBank accession numbers of taxa included in the current study *(Continued)*

Order, Family	Genus	Species	*COI*	*16S*	CRA	Voucher No.
Camarodonta, Temnopleuridae	*Salmacis*	*bicolor bicolor*	–	–	MG198130[b,d]	SMNH Ec 25,623
Camarodonta, Temnopleuridae	*Salmacis*	*bicolor rarispina*	KU302104[c]	KU302104[c]	KU302104[c]	
Camarodonta, Temnopleuridae	*Salmacis*	*sphaeroides*	KU302103[c]	KU302103[c]	KU302103[c]	
Camarodonta, Temnopleuridae	*Temnopleurus*	*hardwickii*	KP070768[c,b]	KP070768[c,b]	KP070768[c,b]	
Camarodonta, Temnopleuridae	*Temnopleurus*	*reevesii*	KU302106[c]	KU302106[c]	KU302106[c]	
Camarodonta, Temnopleuridae	*Temnopleurus*	*toreumaticus*	KU302105[c]	KU302105[c]	KU302105[c]	
Camarodonta, Toxopneustidae	*Lytechinus*	*variegatus*	–	–	MG198125[b,d]	SMNH Ec 25,620
Camarodonta, Toxopneustidae	*Pseudoboletia*	*indiana*	–	–	MG198128[b,d]	AIM MA 121531.1
Camarodonta, Toxopneustidae	*Pseudoboletia*	*indiana*	–	–	MG198129[b,d]	AIM MA 121531.2
Camarodonta, Toxopneustidae	*Tripneustes*	*depressus*	–	–	MG198131[b,d]	NHMW-DNAtis_26362
Camarodonta, Toxopneustidae	*Tripneustes*	*depressus*	–	–	MG198132[b,d]	NHMW-DNAtis_26363
Camarodonta, Toxopneustidae	*Tripneustes*	*depressus*	–	–	MG198133[b,d]	NHMW-DNAtis_26364
Camarodonta, Toxopneustidae	*Tripneustes*	*depressus*	–	–	MG198134[b,d]	NHMW-DNAtis_26365
Camarodonta, Toxopneustidae	*Tripneustes*	*depressus*	–	–	MG198135[b,d]	NHMW-DNAtis_26366
Camarodonta, Toxopneustidae	*Tripneustes*	*depressus*	–	–	MG198136[b,d]	NHMW-DNAtis_26367
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	KY268294[c,a]	KY268294[c,a]	KY268294[c,a]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	KJ680294[c]	KJ680294[c]	KJ680294[c]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	KY515261[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	KY515262[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	KY515263[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	KY515264[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198149[b,d]	CAS 187197
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198137[b,d]	NHMW 2017/0125/0001
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198138[b,d]	NHMW 2017/0125/0003
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198139[b,d]	NHMW 2017/0125/0004
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198140[b,d]	NHMW Ev 20,497
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198141[b,d]	NHMW Ev 20,498
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198142[b,d]	NHMW Ev 20,499
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198143[b,d]	NHMW 2016/0329/0001
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198144[b,d]	NHMW 2016/0329/0002
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198145[b,d]	NHMW 2016/0329/0003
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198146[b,d]	NHMW 2016/0329/0004
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198147[b,d]	NHMW Ev 20,500
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla gratilla*	–	–	MG198148[b,d]	NHMW Ev 20,501
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla elatensis*	–	–	KY515254[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla elatensis*	–	–	KY515255[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla elatensis*	–	–	KY515256[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla elatensis*	–	–	KY515257[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla elatensis*	–	–	KY515258[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla elatensis*	–	–	KY515259[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*gratilla elatensis*	–	–	KY515260[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515241[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515242[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515243[b]	

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 6 of 15

Table 1 Detailed taxonomic placement and GenBank accession numbers of taxa included in the current study *(Continued)*

Order, Family	Genus	Species	COI	16S	CRA	Voucher No.
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515244[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515245[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515246[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515247[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515248[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515249[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515250[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515251[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515240[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515252[b]	
Camarodonta, Toxopneustidae	*Tripneustes*	*kermadecensis*	–	–	KY515253[b]	
Clypeasteroida, Clypeasteroidae	*Arachnoides*	sp.	–	–	MG198155[b,d]	QM NO1_F4
Clypeasteroida, Clypeasteroidae	*Arachnoides*	sp.	–	–	MG198156[b,d]	QM NO1_F4B
Clypeasteroida, Clypeasteroidae	*Clypeaster*	*rarispinus*	–	–	MG198157[b,d]	MNHN IE-2013-8702
Clypeasteroida, Clypeasteroidae	*Clypeaster*	*rarispinus*	–	–	MG198158[b,d]	MNHN IE-2013-8704
Spatangoida, Loveniidae	*Echinocardium*	*cordatum*	FN562581[c,b]	FN562581[c,b]	FN562581[c,b]	
Spatangoida, Maretiidae	*Nacospatangus*	*alta*	KC990834[c,b]	KC990834[c,b]	KC990834[c,b]	
Spatangoida, Pericosmidae	*Pericosmus*	*bidens*	–	–	MG198167[b,d]	MNHN IE-2007-1138
Spatangoida, Pericosmidae	*Pericosmus*	*bidens*	–	–	MG198168[b,d]	MNHN IE-2013-8706
Spatangoida, Pericosmidae	*Pericosmus*	*bidens*	–	–	MG198169[b,d]	MNHN IE-2013-8707
Spatangoida, Pericosmidae	*Pericosmus*	*bidens*	–	–	MG198170[b,d]	MNHN IE-2013-8708

COI - cytochrome c oxidase subunit I, *16S rRNA* – ribosomal RNA, *CRA* -control region area (including the control, adjacent *tRNA*s and a part of the *12S rRNA* genes). Sequence type indicates whether the source sequence was generated by Sanger or next generation sequencing
[a]sequence data generated by NGS
[b]sequence data generated by Sanger sequencing
[c]data retrieved from complete mitochondrial genome sequence
[d]sequences generated in the current study
AIM Auckland War Memorial Museum, *CAS* California Academy of Sciences, *MNHN* Muséum national d'Histoire naturelle, *NHMW* Natural History Museum Vienna, *QM* Queensland Museum, *SMNH* Steinhardt Museum of Natural History, *USNM* Smithsonian Institution National Museum of Natural History

within the $tRNA^{Asn}$ gene), complemented by two internal primers: CR_int_fwd 5'-CTTTGGGAGTTGCA AATGTAAGTG-3' and CR_int_rev 5'-TTTAACCCT CTCTCCTGGTTTACA-3' (Fig. 1).

Laboratory procedures
Total genomic DNA was extracted from tube feet and spine muscles using the DNeasy® Blood and Tissue Kit (QIAGEN) following the manufacturer's instructions. PCR amplifications with the TopTaq DNA Polymerase (QIAGEN) were conducted using 1 μl of extracted genomic DNA (approximately 10-15 ng). Reaction conditions using primers CR15fwd and CR08rev were 3 min at 94 °C, followed by 35 cycles of 30 s at 94 °C, 30 s at 57 °C and 60 s at 72 °C, ending with a final extension step of 10 min at 72 °C. PCR products were visualized on a 1.5% agarose gel, purified using ExoSAP-IT (Affymetrix) and sent for sequencing to Microsynth GmbH (Vienna, Austria) using the PCR

primers. In cases of weak amplifications (amplicons of the expected size showing only as faint bands on an agarose gel), target selected re-amplifications were performed following the methods of Bjourson and Cooper [49] using the same primers and reaction conditions as above. For this purpose, the respective PCR fragments were visualised on an agarose gel under UV light and templates for the re-amplification were transferred from the gel to the new PCR tubes using a sterile needle.

Despite yielding a product at the expected length, all amplicons failed to sequence through using the external PCR primers, with the sequencing signal dropping ca. half way through the expected amplicon length (see results for details). Consequently, the internal primers CR_int_fwd and CR_int_rev were used to generate a two-way sequencing of each amplicon (before and after the break-off). The sequences determined in the present study were deposited in GenBank under the accession numbers: MG198122–MG198170.

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 7 of 15

Data assembly and phylogenetic analyses

Four datasets were created to facilitate our analyses. (1) *CRA-all* comprising all publicly available echinoid sequences in addition to the 49 sequences generated in the current study (405 bp long). To facilitate the direct comparisons of the CR performance as a phylogenetic marker with the two most common mitochondrial markers (*COI* and *16S*), three additional datasets were constructed based solely on the 35 publicly available complete echinoid mitogenomes: (2) *Mito-COI* was extracted from the mitogenomes targeting the region defined by the primers EchinoF1 and HCO2198 as in Ward et al. [36] yielding an alignment of 33 unique sequences, 562 bp long. (3) *Mito-16S* was extracted from the mitogenomes targeting the region defined by the primers 16sar-L and 16sbr-H of Palumbi (in [50]) yielding an alignment of 33 unique sequences, 558 bp long. (4) *Mito-CRA* was extracted from the mitogenomes targeting our current CR primers (CR15fwd and CR08rev), yielding an alignment of 33 unique sequences, 405 bp long. In all of the above datasets, ambiguous site removal was performed with trimAl v. 1.2 ([51]; setting: *-automated1*), followed by unique sequences detection using DAMBE6 [52]. The phylogenies based on the above datasets were compared to the current working classification of echinoids as presented by Kroh and Smith [53] and implemented in the World Echinoidea Database [54]. This classification is based on a large-scale numerical cladistic analysis involving representatives of all extant and fossil echinoid families and more than 300 morphological characters.

Phylogenetic analyses were conducted using both Maximum Likelihood (ML) and Bayesian Inference (BI) approaches. A heuristic search under the Bayesian Information Criterion (BIC) [55], as implemented in PartitionFinder2 [56], was employed to determine the optimal partitioning schemes and models of molecular evolution for the phylogenetic analyses (Table 2).

ML analysis was performed using the program RAxML GUI v. 1.5b1 [57]. Settings were 'ML + thorough bootstrap', 100 runs, 1000 replicates, applying the best-fit models as inferred from PartitionFinder2. Bayesian analysis was carried out using the program MrBayes v. 3.2.2 [58]. We ran two independent runs of three 'heated' and one 'cold' chain for 10 million generations and sampled parameters and trees every 100 generations. The runs were also inspected with Tracer v. 1.6 [59] to assess the behaviour of the runs and convergence was assessed using RWTY package [60] implemented in R v. 3.2.1 [61]. In a conservative approach, the first 25% of trees were discarded as burn-in and a 50% majority rule consensus tree was calculated from the remaining trees. Bayesian Posterior Probabilities (PP) were obtained from the 50% majority-rule consensus of the trees sampled during the stationary phase.

The topologies of the different phylogenetic trees were visualised and compared using Phylo.io [62], applying a variant of the Jaccard Index (Jaccard similarity coefficient) as the comparison metric as implemented in Phylo.io. This tool performs automated branch-swapping in order to find the best corresponding visualization between two trees and compares branching order. Additionally, trees were also compared using the duplication-aware algorithm treeKO [63] implemented in the ete-compare module of the Python Environment for Tree Exploration (ETE) [64]. In contrast to other algorithms for tree comparison, treeKO does not require complete and exact matching of terminal taxa between compared trees and provides Robinson-Foulds-based distance metrics even in the presence of duplication and loss events.

Substitution saturation decreases the amount of phylogenetic signal to the point that sequence similarities could as likely be the result of chance alone rather than homology. Consequently, when saturation is reached,

Table 2 Sequences summary for the different datasets including best-fitting nucleotide substitution models

	CRA-all	Mito-COI	Mito-16S	Mito-CRA
Number of unique sequences	86	33	33	33
MSA length (bp)	405	562	558	405
%G	20.4	18.9	22.3	20.9
%A	30.9	26.0	31.0	30.7
%C	21.4	24.4	20.5	21.0
%T	27.3	30.8	26.2	27.5
P_{inv}	0.09384	0.34741 (codon position 1&2) 0.04714 (codon position 3)	0.36998	0.17559
Overall mean K2P/p-distance	0.16/0.14	0.21/0.18	0.16/0.143	0.16/0.14
Best-fit model – (ML)	GTR + G	GTR + I + G	GTR + G	GTR + G
Best-fit model – (BI)	HKY + G	SYM + I + G	GTR + G	HKY + G

ML Maximum Likelihood, *BI* Bayesian Inference, *CR* Control region, *COI* cytochrome c oxidase subunit 1, *16S* 16S ribosomal RNA, *MSA* multiple sequence alignment, P_{inv} proportion of invariant sites

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 8 of 15

phylogenetic signal is lost and the sequences are no longer informative about the underlying evolutionary process that produced them [65]. Saturation of substitutions was evaluated by plotting the number of transitions (s) and transversions (v) against the F84 [66] genetic distance, as well as by comparing the information entropy-based index (I_{ss}) with critical values ($I_{ss.c}$) [67, 68] as implemented in DAMBE6 [52]. If I_{ss} is significantly lower than $I_{ss.c}$, sequences have not experienced substitution saturation.

Results

Primer performance and sequence characteristics

Amplifications of the CR and adjacent tRNAs performed well across the diverse groups of echinoids, generally yielding a single distinct product of the expected length (Additional file 1: Figure S1). Notably, arrangement of the mitochondrial genes around the CR (including the *12S* and various tRNA genes) as well as the features within the CR were identical in all of the taxa analysed as well as the taxa obtained from GenBank indicating a conserved organisation of this area. PCR amplifications performed equally well on DNA extracted from tissue of varying quality (considering tissue age and methods of preservation as outlined above). The oldest sample analysed in the current study was a dried specimen of *Zenocentrotus kellersi* A.H. Clark, 1931 (USNM E40502), collected in 1930 together with the holotype. DNA from this specimen was successfully amplified and sequenced, allowing for the determination of the phylogenetic position of *Z. kellersi*.

Regardless of successful amplifications, complete sequencing of amplicons in full length using only the external PCR primers failed. Sequencing using both the forward and reverse primers suffered an abrupt signal loss at a stretch of at least 12 guanine bases, roughly in the middle of the amplified region (see Additional file 1: Figure S2 and text below). Applying the internal primers enabled sequencing the complementary strand (before and after the guanine stretch) thereby ensuring a reliable two-way read of the amplified sections.

The echinoid CR is located within a cluster of 15 tRNA genes, between the genes for $tRNA^{Thr}$ and $tRNA$-^{Pro} confirming the gene order noted previously by Jacobs et al. [69]. This position was confirmed for all taxa examined in the current study. The middle of the CR is composed of a stretch of up to 20 guanine residues (referred to as the poly-G stretch) although the precise length of this string could not be unambiguously determined (due to the sequencing limitations discussed above). Nevertheless, some taxa consistently showed a shorter poly-G stretch than others. The shortest one was recovered in *Asthenosoma varium* and *Araeosoma splendens* (with 12 G residues) followed closely by the

diadematoids *Diadema setosum* and *Echinothrix diadema*, both taxa consistently yielding 13 G residues. The 3′ side of the G stretch is followed by an A + T- rich segment that resembles the TATA box found in the promotor regions of eukaryotes. On the 5′ side of the G stretch is a heterogeneous yet fairly conserved segment of 37 to 40 bp in most cases, preceded by a polypyrimidine tract as observed by Jacobs et al. [69]. As before, the exceptions being the diadematoids which display a shorter, 20 bp segment, and echinothuroids which possess a 30 bp long heterogeneous segment prior to the polypyrimidine tract.

Mitochondrial markers comparison

Phylogenetic reconstructions of the three selected mitochondrial markers (i.e., *COI*, *16S* and CRA), extracted from the 35 complete echinoid mitochondrial genomes are shown in Fig. 2. Taxa from the above datasets were collapsed to genera. The corresponding tree for each marker (left column trees, denoted: A – *COI*, C – *16S* and E – CRA) is shown in comparison to the current [53, 54] echinoid working classification (right column trees, denoted: B, D and F). Topological similarities between the gene trees and the systematic consensus tree are highlighted using the variant Jaccard Index metric as implemented in Phylo.io [62]. The colour coded comparison metric with a score of 1 on the COI gene tree for example (Fig. 2a), denotes an identical subtree structure in the corresponding tree based on morphology (Fig. 2b).

All three markers show high congruence for some of the taxa but marked differences for others. At large, the CRA tree seem to be superior and outperform the other two markers in terms of deep divergence and accuracy in comparison to the systematic consensus tree. The *COI* tree for example misplaces the temnopleurids as the sister group of diadematoids and the Irregularia and places the latter two clades as the sister group of the camarodont *Tripneustes* (Fig. 2a) although these topologies are poorly supported. In the *16S* tree, the Irregularia are misplaced and included within the echinaceans while the temnopleurids are excluded from the latter and resolved as sister group of the diadematoids (Fig. 2c) yet once again this was poorly supported. In general, both *COI* and *16S* reconstructed topologies were inferior to the CRA in terms of resolution and branch support (Fig. 2). While the former topologies retrieved polytomies at the basal nodes of the Camarodonts (also observed in the analysis of Smith et al. [70]) as well as the temnopleurids, both were well resolved in the CRA tree (Fig. 2e). Nevertheless, some discrepancies also occurred within the CRA inferred topology, namely the misplacement of *Arbacia* (a sequence deriving from GenBank, not reconfirmed in the present study) as sister group of the diadematoids (with
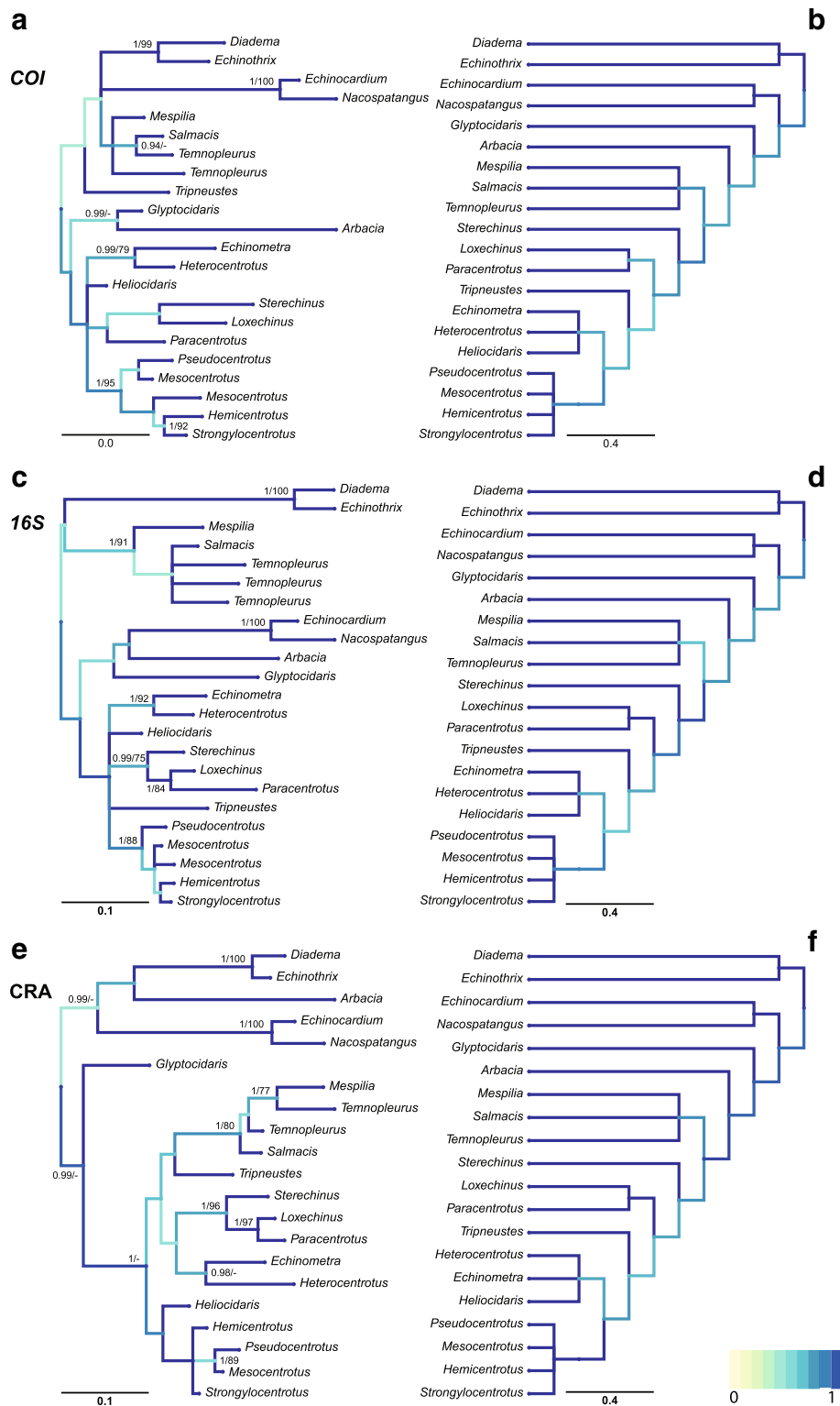
Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 9 of 15

Fig. 2 (See legend on next page.)

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 10 of 15

low support 0.8/−) with both being the sister group of the Irregularia. Interestingly, *Heliocidaris* was consistently misplaced outside the Echinometridae in all three data sets, either placed in a basal polytomy amongst members of the Camarodonta (*COI* and *16S*) or as sister group of the Strongylocentrotidae (CRA).

Summary statistics for the comparison of the different mitochondrial markers are given in Table 3. As the trees used in the current comparison were not strictly symmetric and contained duplicate attributes (i.e., tip names), only the duplication aware distance of the TreeKO method (treekoD) was appropriate for the comparison. Duplicated attributes were formed during the phylogenetic analysis as genera were recovered as non-monophyletic and split into several clades. The CRA tree had a significantly larger effective size in comparison to *COI* and *16S* (i.e., more items from this tree were used for the comparison with the systematic consensus tree). The treekoD metric showed similar values for all three markers, with the values for *COI* and CRA being virtually identical (0.60 and 0.61, respectively).

Substitution saturation was evaluated for all markers and datasets. No saturation was detected in the CRA as reflected from the linear correlation of the number of transitions and transversions plotted against genetic distance (Fig. 3), as well as from a significantly lower value of I_{ss} in comparison to $I_{ss.c}$ (for both symmetrical and asymmetrical weighed topologies) (Table 4). Concerning the *16S* marker, no saturation was detected assuming a symmetrical topology, while incipient saturation was detected assuming an asymmetrical topology for 32 OTUs and above (Additional file 1: Figure S3 and Table S1). In *COI* saturation were more prominent at

the third codon position (Additional file 1: Figure S3 and Table S1).
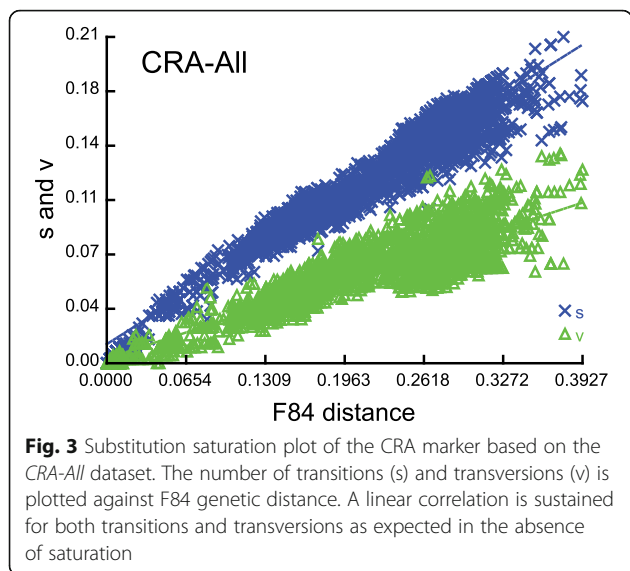
Echinoid phylogeny based on the CR

Both BI and ML analyses for the complete set of CRA sequences (data set *CRA-all*) rooted on the Cidaridae produced highly congruent topologies for all major clades and subclades. Consequently, BI inferred topologies are presented with both posterior probabilities and bootstrap support (from ML analyses) for each clade (Fig. 4). The strict consensus tree shows good resolution and branch support in most parts of the tree. Most clades received high nodal support except for some members of the Strongylocentroidae and members of the genus *Temnopleurus*. The phylogeny based on the new CRA marker successfully retrieved most putative species as highly supported monophyletic clades. Nevertheless, in several instances the retrieved topology contradicted conventional systematics. The genus *Temnopleurus* for instance, was not monophyletic, as *Temnopleurus reevesii* was retrieved as the sister group of *Mespilia globulus*. Although receiving very poor support for this split, the latter clade formed the sister clade of the other temnopleurids in the current data set, *T. hardwickii* and *T. toreumaticus*. Similarly, sequences of *Diadema setosum* mostly collapsed into a single, well supported clade, although a single *D. setosum* sequence formed the sister of *Echinothrix diadema*. As in the case of *Temnopleurus*, this latter split received very poor nodal support. *Zenocentrotus kellersi* A.H. Clark, 1931 (USNM E40502), is presented here for the first time in a molecular phylogenetic context (Fig. 4). It was recovered as the sister group of *Heterocentrotus mammillatus*,

Table 3 Phylogenetic tree comparisons using the duplication-aware algorithm TreeKO as implemented in the Python Environment for Tree Exploration (ETE)

Source tree	Reference tree	Effective tree size	nRF	RF	maxRF	%src_br	%ref_br	Subtrees number	treekoD
COI	Syst	10.0	0.60	9	15	0.38	0.43	1	0.60
16S	Syst	11.0	0.53	9	17	0.44	0.50	4	0.53
CRA	Syst	18.5	0.61	18.5	31	0.36	0.43	2	0.61

Genetically inferred trees based on: *COI, 16S*, CRA are compared to current echinoid classification (Syst) (see text for details)
RF Robinson-Foulds symmetric distance, *nRF* normalized RF distance (RF/maxRF); %src_br – frequency of edges in target tree found in the reference; %ref_br – frequency of edges in the reference tree found in the target; Subtrees – number of subtrees used for the comparison; treekoD – average p distance among all possible subtrees in the original target trees to the reference tree (lower treekoD values are indicative of higher similarity between trees)

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 11 of 15

Fig. 3 Substitution saturation plot of the CRA marker based on the *CRA-All* dataset. The number of transitions (s) and transversions (v) is plotted against F84 genetic distance. A linear correlation is sustained for both transitions and transversions as expected in the absence of saturation

Discussion

Apart from their use in echinoid phylogeny, the primers used herein can be used as a complementary tool to NGS assemblies of echinoid mitogenomes, many of which show very low read coverage in the CR (Fig. 1). Coverage in NGS studies was shown to be significantly correlated to GC content of the analysed sequences (e.g., [72] for mitochondrial DNA). The extreme drop in coverage we observed at the echinoid CR, however, is not easily explained as being a direct result of low GC content in this region. Figure 5 illustrates variation in GC contents throughout the mitogenome of *Hemicentrotus pulcherrimus* in relation to coverage. Although reduced coverage generally does seem to coincide with lower GC content, the extreme drop in coverage at the CR could not be explicitly accounted for by the former, as other parts of this echinoid mitogenome show even lower GC content but do not suffer from coverage reduction as substantial as the CR. Similarly, low coverage values have been reported (or can be inferred by read mapping of published data) for the CR of other organisms (insects: [73] [based on mapping of reads from GenBank short read archive SRR835993 on sequence KP216766]; [74]: Fig. 1; crustaceans: [75]: Fig. 1; humans: [76, 77]: 130, Fig. 1). The reasons for this phenomenon are not clear and may be in part related to technical issues (assembly, read mapping) or biochemial qualities of this region (see [3]). Regardless of the driving mechanism, supplementing NGS data with Sanger generated sequences of this target region will improve the quality of mitogenomes assemblies.

By far the most thoroughly studied control region of all organisms is that of humans (see [78]). Although nearly nine times shorter than the human CR and bearing little sequence similarity, several analogies between human and echinoid CR have been proposed [69, 79]. In particular, sequence motifs in the echinoid CR such as the polypurine and polypyrimidine tracts (see above), were regarded as homologous to the mammalian conserved sequence blocks [80, 81]. In line with the interpretations of Jacobs et al. [69], Cantatore et al. [80], and De Giorgi et al. [82], based on merely three echinoid

forming with the latter the sister clade of *Echinometra* and *Microcyphus*, in congruence to the current morphological classification.

On two occasions, putative species and subspecies had identical sequences. *Tripneustes gratilla gratilla*, *Tripneustes gratilla elatensis* and *Tripneustes depressus* all shared a single haplotype that was retrieved as sister of *Tripneustes kermadecensis* (Fig. 4). This result was expected, due to the phenomenon of mitochondrial capture between some members of *Tripneustes* [71]. *Strongylocentrotus droebachiensis* and *Strongylocentrotus pallidus* similarly shared a single haplotype, closely related to another sequence of *S. pallidus*. One taxon, *Heliocidaris crassispina*, an echinometrid, was placed as sister group to the family Strongylocentrotidae (in agreement with the *COI* and *16S* trees, but in contrast to current classification), albeit this grouping received very poor nodal support. The only taxon to display markedly contradictory placement (in comparison to current classification) with strong nodal support is a sequence of *Salenia phoinissa* (1/99) that was placed as sister group of *Araeosoma splendens*, in the midst of the expected echinothuriid clade.

Table 4 Substitution saturation analysis of the CRA region based on the index of substitution saturation as implemented in DAMBE6

Marker dataset	Number of OTU[a]	Iss[b]	I$_{ss.cSym}$[c]	df[d]	*p* value[e]	I$_{ss.cAsym}$[f]	df[d]	*p* value[e]
CRA-All	4	0.240	0.789	366	< 0.0001	0.757	366	< 0.0001
	8	0.259	0.743	366	< 0.0001	0.631	366	< 0.0001
	16	0.242	0.703	366	< 0.0001	0.494	366	< 0.0001
	32	0.260	0.692	366	< 0.0001	0.363	366	0.0001

[a]Number of sequences used in the random resampling
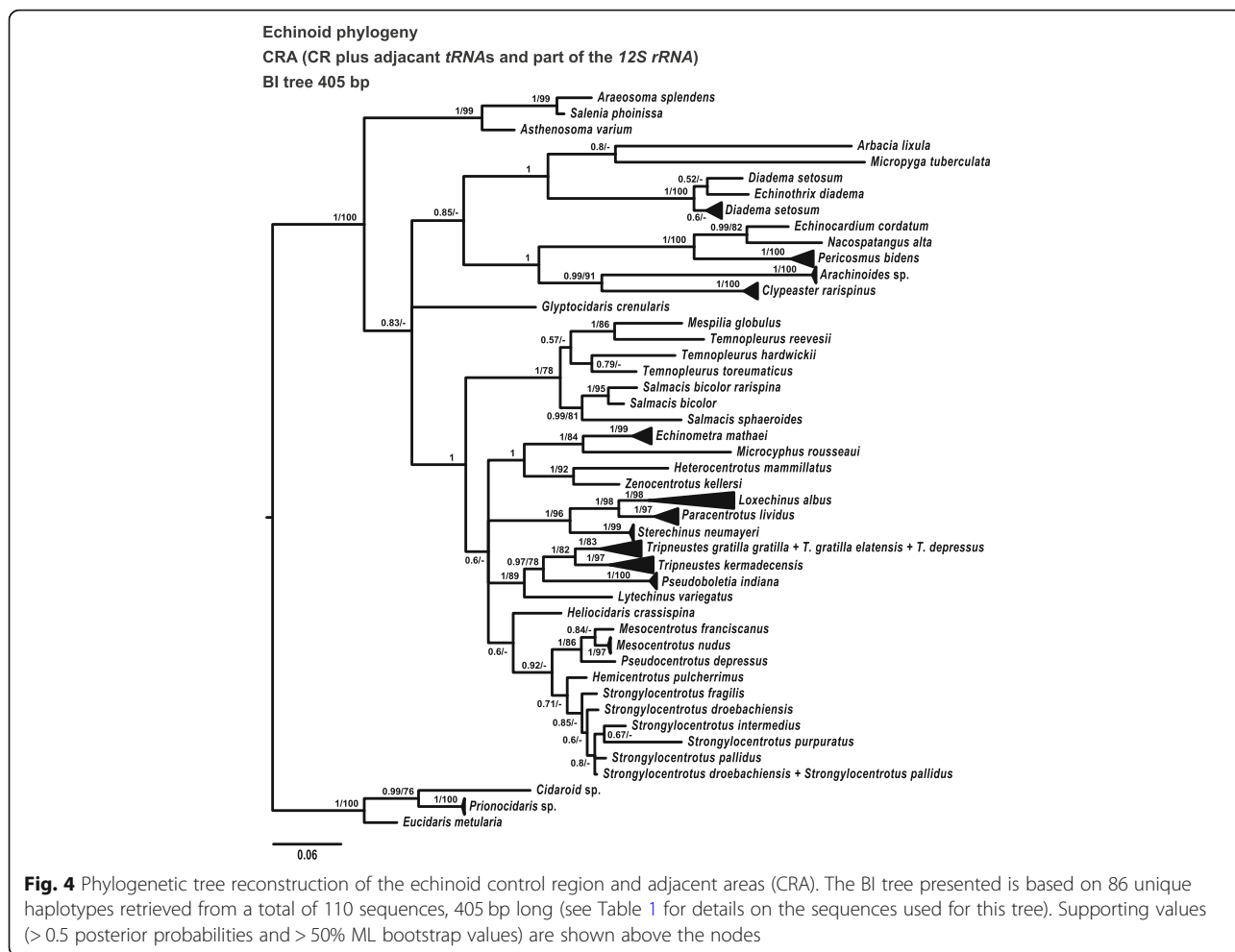[b]index of substitution saturation
[c]critical value for a symmetrical tree topology
[d]degrees of freedom
[e]probability that I$_{ss}$ is significantly different from the critical value (I$_{ss.cSym}$/I$_{ss.cAsym}$)
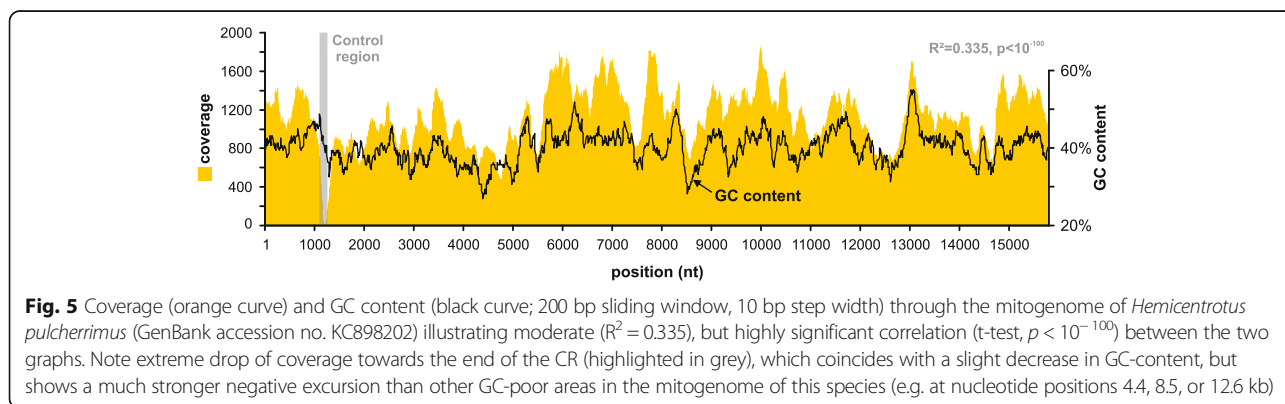[f]critical value for an asymmetrical tree topology
Note: two-tailed tests are used

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 12 of 15

Fig. 4 Phylogenetic tree reconstruction of the echinoid control region and adjacent areas (CRA). The BI tree presented is based on 86 unique haplotypes retrieved from a total of 110 sequences, 405 bp long (see Table 1 for details on the sequences used for this tree). Supporting values (> 0.5 posterior probabilities and > 50% ML bootstrap values) are shown above the nodes

species (*Strongylocentrotus purpuratus*, *Paracentrotus lividus*, and *Arbacia lixula*, respectively), our data from 42 additional echinoid species, tightly conforms to previous observations.

The two most widely used mitochondrial markers, *COI* and *16S*, showed evidence of substitution saturation in our dataset, which may compromise their potential for phylogenetic inference. While the high rate of substitution makes the underlying genes powerful genetic markers, the fast rate of evolution in third codon positions of protein-coding genes, often makes them vulnerable to substantial substitution saturation between highly diverged taxonomic groups [83]. Although we observed substantial saturation at the third codon position in *COI* (Additional file 1: Figure S3 and Table S1), the underlying topologies independently constructed based

Fig. 5 Coverage (orange curve) and GC content (black curve; 200 bp sliding window, 10 bp step width) through the mitogenome of *Hemicentrotus pulcherrimus* (GenBank accession no. KC898202) illustrating moderate ($R^2 = 0.335$), but highly significant correlation (t-test, $p < 10^{-100}$) between the two graphs. Note extreme drop of coverage towards the end of the CR (highlighted in grey), which coincides with a slight decrease in GC-content, but shows a much stronger negative excursion than other GC-poor areas in the mitogenome of this species (e.g. at nucleotide positions 4.4, 8.5, or 12.6 kb)

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 13 of 15

on codon positions 1 and 2 and codon position 3, respectively, were largely congruent with each other and with tree topology based on morphology. The combination of high topological resolution and lack of substitution saturation of CRA, in contrast, highlights its advantages as a powerful and reliable phylogenetic marker.

Placement of Irregularia in the CRA tree as a sister-group to diadematoids is at odds with the results of recent phylogenetic studies based both on morphology [53, 70] and molecular data [53]. It conforms, however, with earlier phylogenetic hypotheses [84, 85] that resolved irregular echinoids as sister-group to pedinoids and diadematoids. It is worth noting, however, that taxon sampling for irregular echinoids in the present dataset is rather low and more data is needed to verify this result, particularly for basal Irregularia (holectypoids) and holasteroids.

Conclusions

This study represents the first thorough investigation of the mitochondrial control region across a wide range of echinoid taxa. It provides a tool for complementing incomplete mitochondrial genomes based on NGS experiments. In comparison to the conventional mitochondrial markers such as *16S* and *COI*, the mitochondrial control region of echinoids was found to outperform the traditional markers. As such, it is a powerful, novel marker for phylogenetic inference in echinoids showing high variability, lack of selection, and high compatibility across the entire class.

NGS technologies are revolutionizing our understanding of evolutionary biology, allowing us to generate phylogenetic datasets on the scale of genomes. Nevertheless, Sanger sequencing still offers a rapid, low cost, and accessible sequencing technology that secures its current status as the primary 'work horse' of molecular genetics. The present study provides a cheap and efficient tool for species identification in echinoids and for tracking the evolutionary history of the entire class Echinoidea when NGS experiments are beyond the scope of a project.

Additional file

Additional file 1: Figure S1. Gel image showing PCR products. **Figure S2.** Sequencing results shown as a four-color chromatogram. **Figure S3.** Substitution saturation plots. **Table S1.** Substitution saturation analysis. (PDF 1735 kb)

Abbreviations

5′–UTR: Five prime untranslated region; AIM: Auckland War Memorial Museum; BI: Bayesian Inference; BIC: Bayesian Information Criterion; bp: Base pairs; CAS: California Academy of Sciences; COI: Cytochrome c oxidase subunit I; CpG: 5′–C–phosphate–G–3′; CR: Control region; CRA: Control region and adjacent tRNAs; ML: Maximum Likelihood; MNHN: Muséum national d'Histoire naturelle; mtDNA: Mitochondrial DNA; NGS: Next generation sequencing; NHMW: Natural History Museum Vienna; PP: Posterior

probabilities; QM: Queensland Museum; rRNA: Ribosomal RNA; SMNH: Steinhardt Museum of Natural History; SNP: Single nucleotide polymorphism; tRNA: Transfer RNA; USNM: Smithsonian Institution National Museum of Natural History

Availability of data and materials

The datasets generated and analysed during the current study are included in this published article and its supplementary information files. DNA sequences data have been deposited in GenBank under accession numbers MG198122–MG198170.

Authors' contributions

OB, AK, and EH conceived and designed the study. OB and AK performed the experiments. OB, AK, and EH analysed the data and wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Natural History Museum Vienna, Geological-Palaeontological Department, 1010 Vienna, Austria. [2]Natural History Museum Vienna, Central Research Laboratories, 1010 Vienna, Austria. [3]Department of Integrative Zoology, University of Vienna, Vienna, Austria.

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7.
2. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:1–11.

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 14 of 15

3. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.

4. Wang W, Wei Z, Lam T-W, Wang J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. Sci Rep. 2011;1:55.

5. Ekblom R, Smeds L, Ellegren H. Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. BMC Genomics. 2014;15(1):467.

6. Stoneking M, Hedgecock D, Higuchi RG, Vigilant L, Erlich HA. Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. Am J Hum Genet. 1991;48(2):370–82.

7. Aquadro CF, Greenberg BD. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics. 1983;103(2):287–312.

8. Kasamatsu H, Robberson DL, Vinograd J. A novel closed-circular mitochondrial DNA with properties of a replicating intermediate. Proc Natl Acad Sci U S A. 1971;68(9):2252–7.

9. Avise JC. Molecular markers, natural history and evolution. New York: Chapman and Hall; 1994.

10. Brown WM, George M, Wilson AC. Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci. 1979;76(4):1967–71.

11. McMillan OW, Palumbi SR. Rapid rate of control-region evolution in Pacific butterflyfishes (Chaetodontidae). J Mol Evol. 1997;45(5):473–84.

12. Meyer A. Evolution of mitochondrial DNA in fishes. In: Hochachka PW, Mommsen TP, editors. Biochemistry and molecular biology of fishes, vol. 2. Amsterdam: Elsevier Science; 1993. p. 1–38.

13. Ghatak S, Lallawmzuali D, Mukherjee S, Mawia L, Pautu JL, Kumar NS. Polymorphism in mtDNA control region of Mizo-Mongloid breast Cancer samples as revealed by PCR-RFLP analysis. Mitochondrial DNA Part A. 2014; 27(3):2205–8.

14. Pereira F, Soares P, Carneiro J, Pereira L, Richards MB, Samuels DC, Amorim A. Evidence for variable selective pressures at a large secondary structure of the human mitochondrial DNA control region. Mol Biol Evol. 2008;25(12): 2759–70.

15. Rech GE, Sanz-Martín JM, Anisimova M, Sukno SA, Thon MR. Natural selection on coding and noncoding DNA sequences is associated with virulence genes in a plant pathogenic fungus. Genome Biol Evol. 2014;6(9): 2368–79.

16. Jamandre BW, Durand J-D, Tzeng W-N. High sequence variations in mitochondrial DNA control region among worldwide populations of flathead mullet Mugil cephalus. Int J Zool. 2014;2014:9.

17. Beltrán-López RG, Domínguez-Domínguez O, Guerrero JA, Corona-Santiago DK, Mejía-Mojica H, Doadrio I. Phylogeny and taxonomy of the genus *Ilyodon* Eigenmann, 1907 (Teleostei: Goodeidae), based on mitochondrial and nuclear DNA sequences. J Zool Syst Evol Res. 2017;55(4):340–55.

18. Huang ZH, Tu FY. Characterization and evolution of the mitochondrial DNA control region in Ranidae and their phylogenetic relationship. Genet Mol Res. 2016;15(3):gmr8491.

19. Jiang Y, Nie LW, Huang ZF, Jing WX, Wang L, Liu L, Dai XT. Comparison of complete mitochondrial DNA control regions among five Asian freshwater turtle species and their phylogenetic relationships. Genet Mol Res. 2011; 10(3):1545–57.

20. Kryukov AP, Spiridonova LN, Mori S, Arkhipov VY, Red'kin YA, Goroshko OA, Lobkov EG, Haring E. Deep Phylogeographic breaks in magpie *Pica pica* across the Holarctic: concordance with bioacoustics and phenotypes. Zool Sci. 2017;34(3):185–200.

21. Boyko AR, Boyko RH, Boyko CM, Parker HG, Castelhano M, Corey L, Degenhardt JD, Auton A, Hedimbi M, Kityo R, et al. Complex population structure in African village dogs and its implications for inferring dog domestication history. Proc Natl Acad Sci. 2009;106(33): 13903–8.

22. Firestone KB. Phylogenetic relationships among quolls revisited: the mtDNA control region as a useful tool. J Mamm Evol. 2000;7(1):1–22.

23. Bronstein O, Kroh A, Tautscher B, Liggins L, Haring E. Cryptic speciation in pan-tropical sea urchins: a case study of an edge-of-range population of *Tripneustes* from the Kermadec Islands. Sci Rep. 2017;7(1):5948.

24. Diniz FM, Maclean N, Ogawa M, Cintra IHA, Bentzen P. The hypervariable domain of the mitochondrial control region in Atlantic spiny lobsters and its potential as a marker for investigating phylogeographic structuring. Mar Biotechnol. 2005;7(5):462–73.

25. Shao R, Barker SC, Mitani H, Aoki Y, Fukunaga M. Evolution of duplicate control regions in the mitochondrial genomes of Metazoa: a case study with Australasian *Ixodes* ticks. Mol Biol Evol. 2005;22(3):620–9.

26. Zhang JJ, Duan JR, Zhou YF, Peng JY, Fang DA. Genetic diversity of mitochondrial control region (D-loop) polymorphisms in *Coilia ectenes taihuensis* inhabiting Taihu Lake, China. Genet Mol Res. 2017;16(1): gmr16019457.

27. Irwin JA, Saunier JL, Niederstätter H, Strouss KM, Sturk KA, Diegoli TM, Brandstätter A, Parson W, Parsons TJ. Investigation of Heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. J Mol Evol. 2009;68(5):516–27.

28. Ludwig A, May B, Debus L, Jenneckens I. Heteroplasmy in the mtDNA control region of sturgeon (*Acipenser*, *Huso* and *Scaphirhynchus*). Genetics. 2000;156(4):1933–47.

29. Bensch S, Härlid A. Mitochondrial genomic rearrangements in songbirds. Mol Biol Evol. 2000;17(1):107–13.

30. Nittinger F, Haring E, Pinsker W, Wink M, Gamauf A. Out of Africa? Phylogenetic relationships between Falco biarmicus and the other hierofalcons (Aves: Falconidae). J Zool Syst Evol Res. 2005;43(4):321–31.

31. Singh TR, Shneor O, Huchon D. Bird mitochondrial gene order: insight from 3 warbler mitochondrial genomes. Mol Biol Evol. 2008;25(3):475–7.

32. Cadahía L, Pinsker W, Negro JJ, Pavlicev M, Urios V, Haring E. Repeated sequence homogenization between the control and pseudo-control regions in the mitochondrial genomes of the subfamily Aquilinae. J Exp Zool B Mol Dev Evol. 2009;312B(3):171–85.

33. Eberhard JR, Wright TF, Bermingham E. Duplication and concerted evolution of the mitochondrial control region in the parrot genus Amazona. Mol Biol Evol. 2001;18(7):1330–42.

34. Jeffery CH, Emlet RB, Littlewood DTJ. Phylogeny and evolution of developmental mode in temnopleurid echinoids. Mol Phylogenet Evol. 2003;28:99–118.

35. Littlewood DTJ, Smith AB. A combined morphological and molecular phylogeny for sea urchins (Echinoidea: Echinodermata). Phil Trans R Soc Lond B. 1995;347:213–34.

36. Ward RD, Holmes BH, O'Hara TD. DNA barcoding discriminates echinoderm species. Mol Ecol Resour. 2008;8(6):1202–11.

37. Barker FK, Benesh MK, Vandergon AJ, Lanyon SM. Contrasting evolutionary dynamics and information content of the avian mitochondrial control region and ND2 gene. PLoS One. 2012;7(10):e46403.

38. Kruckenhauser L, Haring E, Pinsker W, Riesing MJ, Winkler H, Wink M, Gamauf A. Genetic vs. morphological differentiation of old world buzzards (genus *Buteo*, Accipitridae). Zool Scr. 2004;33(3):197–211.

39. Bribiesca-Contreras G, Solís-Marín FA, Laguarda-Figueras A, Zaldívar-Riverón A. Identification of echinoderms (Echinodermata) from an anchialine cave in Cozumel Island, Mexico, using DNA barcodes. Mol Ecol Resour. 2013;13(6): 1137–45.

40. Bronstein O, Loya Y. The taxonomy and phylogeny of *Echinometra* (Camarodonta: Echinometridae) from the Red Sea and western Indian Ocean. PLoS One. 2013;8(10):e77374.

41. Hoareau TB, Boissin E. Design of phylum-specific hybrid primers for DNA barcoding: addressing the need for efficient COI amplification in the Echinodermata. Mol Ecol Resour. 2010;10(6):960–7.

42. Smith AB, Kroh A. Phylogeny of sea urchins. In: Lawrence JM, editor. Sea urchins: biology and ecology. 3rd ed. Amsterdam: Elsevier; 2013. p. 1–14.

43. Soliman T, Omar H, Abdel Razek FA, El-Sayed A-FM, Elmasry E, Davis Reimer J. Phylogenetic characterization of two echinoid species of the southeastern Mediterranean, off Egypt. Egypt J Aquat Res. 2015;41(4):359–65.

44. Zigler KS, Lessios HA. Speciation on the coasts of the new world: phylogeography and the evolution of bindin in the sea urchin genus *Lytechinus*. Evolution. 2004;58(6):1225–41.

45. Palumbi SR. What can molecular genetics contribute to marine biogeography? An urchin's tale. J Exp Mar Biol Ecol. 1996;203(1):75–92.

46. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80.

47. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. Nucleic Acids Symp Ser. 1999;41:95–8.

48. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014;30(22):3276–8.

49. Bjourson AJ, Cooper JE. Band-stab PCR: a simple technique for the purification of individual PCR products. Nucleic Acids Res. 1992;20(17):4675.

Bronstein *et al. BMC Evolutionary Biology* (2018) 18:80

Page 15 of 15

50. Palumbi S, Martin A, Romano S, McMillan WO, Stice L, Grabowski G: Simple Fool's guide to PCR. Honolulu: Department of Zoology and Kewalo Marine Laboratory; 2007.

51. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.

52. Xia X. DAMBE6: new tools for microbial genomics, Phylogenetics, and molecular evolution. J Hered. 2017;108(4):431–7.

53. Kroh A, Smith AB. The phylogeny and classification of post-Palaeozoic echinoids. J Syst Palaeontol. 2010;8(2):147–212.

54. Kroh A, Mooi R. The World Echinoidea Database - Version 3.0. 2017. http://www.marinespecies.org/echinoidea/index.php. Accessed 20 Aug 2017.

55. Schwarz S. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4.

56. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol Biol Evol. 2017;34(3):772–3.

57. Silvestro D, Michalak I. raxmlGUI: a graphical front-end for RAxML. Organ Divers Evol. 2012;12(4):335–7.

58. Ronquist F, Teslenko M, van der Mark P, Ayres D, Darling A, Hohna S, Larget B, Liu L, Suchard M, Huelsenbeck J. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61(3):539–42.

59. Rambaut A, Suchard MA, Xie D, Drummond AJ. Tracer v1.6. 2013. Available from http://beast.community/tracer.

60. Warren DL, Geneva AJ, Lanfear R. RWTY (R we there yet): an R package for examining convergence of Bayesian phylogenetic analyses. Mol Biol Evol. 2017;34(4):1016–20.

61. Team RC. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2013. Available from https://www.r-project.org/.

62. Robinson O, Dylus D, Dessimoz C. Phylo.io : interactive viewing and comparison of large phylogenetic trees on the web. Mol Biol Evol. 2016; 33(8):2163–6.

63. Marcet-Houben M, Gabaldón T. TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. Nucleic Acids Res. 2011;39(10):e66.

64. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of Phylogenomic data. Mol Biol Evol. 2016;33(6):1635–8.

65. Salemi M. Nucleotide substitution models. PRACTICE: the PHYLIP and TREE-PUZZLE software packages. In: Salemi M, Vandamme A-M, editors. The phylogenetic handbook a practical approach to phylogenetic analysis and hypothesis testing. Cambridge: Cambridge University Press; 2003. p. 88–97.

66. Felsenstein J. Distance methods for inferring phylogenies: a justification. Evolution. 1984;38(1):16–24.

67. Xia X, Lemey P. Assessing substitution saturation with DAMBE. In: Lemey P, Salemi M, Vandamme A-M, editors. The phylogenetic handbook: a practical approach to DNA and protein phylogeny. 2nd ed. Cambridge: Cambridge University Press; 2009. p. 615–30.

68. Xia X, Xie Z, Salemi M, Chen L, Wang Y. An index of substitution saturation and its application. Mol Phylogenet Evol. 2003;26(1):1–7.

69. Jacobs HT, Elliott DJ, Math VB, Farquharson A. Nucleotide sequence and gene organization of sea urchin mitochondrial DNA. J Mol Biol. 1988;202(2): 185–217.

70. Smith AB, Pisani D, Mackenzie-Dodds JA, Stockley B, Webster BL, Littlewood DTJ. Testing the molecular clock: molecular and paleontological estimates of divergence times in the Echinoidea (Echinodermata). Mol Biol Evol. 2006; 23:1832–51.

71. Bronstein O, Kroh A, Haring E. Do genes lie? Mitochondrial capture masks the Red Sea collector urchin's true identity (Echinodermata: Echinoidea: *Tripneustes*). Mol Phylogenet Evol. 2016;104:1–13.

72. McElhoe JA, Holland MM, Makova KD, MS-W S, Paul IM, Baker CH, Faith SA, Young B. Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. Forensic Sci Int. 2014;13:20–9.

73. Peng X-Y, Zhou P, Qiang Y, Qian Z-Q. Characterization of the complete mitochondrial genome of *Bombyx huttoni* (Lepidoptera: Bombycidae). Mitochondrial DNA Part A. 2015;27(6):4112–3.

74. Wu F, Kumagai L, Cen Y, Chen J, Wallis CM, Polek M, Jiang H, Zheng Z, Liang G, Deng X. Analyses of Mitogenome sequences revealed that Asian Citrus psyllids (*Diaphorina citri*) from California were related to those from Florida. Sci Rep. 2017;7(1):10154.

75. Gan HM, Schultz MB, Austin CM. Integrated shotgun sequencing and bioinformatics pipeline allows ultra-fast mitogenome recovery and confirms substantial gene rearrangements in Australian freshwater crayfishes. BMC Evol Biol. 2014;14:19.

76. Bouhlal Y, Martinez S, Gong H, Dumas K, Shieh JTC. Twin mitochondrial sequence analysis. Mol Genet Genomic Med. 2013;1(3):174–86.

77. King JL, LaRue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warshauer DH, Davis CP, Parson W, Sajantila A, et al. High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int. 2014;12:128–35.

78. Brandon MC, Lott MT, Nguyen KC, Spolim S, Navathe SB, Baldi P, Wallace DC. MITOMAP: a human mitochondrial genome database—2004 update. Nucleic Acids Res. 2005;33(Database Issue):D611–3.

79. Walberg MW, Clayton DA. Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. Nucleic Acids Res. 1981;9(20):5411–21.

80. Cantatore P, Roberti M, Rainaldi G, Gadaleta MN, Saccone C. The complete nucleotide sequence, gene organization, and genetic code of the mitochondrial genome of *Paracentrotus lividus*. J Biol Chem. 1989;264(19): 10965–75.

81. Jacobs HT, Herbert ER, Rankine J. Sea urchin egg mitochondrial DNA contains a short displacement loop (D-loop) in the replication origin region. Nucleic Acids Res. 1989;17(22):8949–65.

82. De Giorgi C, Martiradonna A, Lanave C, Saccone C. Complete sequence of the mitochondrial DNA in the sea urchin *Arbacia lixula*: conserved features of the echinoid mitochondrial genome. Mol Phylogenet Evol. 1996;5(2):323–32.

83. Xia X. The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. Mol Biol Evol. 1998;15(3):336–44.

84. Smith AB. Implications of lantern morphology for the phylogeny of post-Palaeozoic echinoids. Palaeontology. 1981;24(4):779–801.

85. Smith AB. Echinoid Palaeobiology. London: Allen & Unwin; 1984.

86. Kober KM, Bernardi G. Erratum to: Phylogenomics of strongylocentrotid sea urchins. BMC Evol Biol. 2017;17(1):50.

87. Kober KM, Bernardi G. Phylogenomics of strongylocentrotid sea urchins. BMC Evol Biol. 2013;13(1):1–14.

88. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 2011;12:402.

89. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9:357–9.