Integrative and Comparative Biology

Integrative and Comparative Biology, volume 61, number 5, pp. 1753–1761
https://doi.org/10.1093/icb/icab152

Society for Integrative and Comparative Biology

SYMPOSIUM

Leveraging Short-Read Sequencing to Explore the Genomics of Sepiolid Squid

Elizabeth Heath-Heckman [*,1] and Michele K. Nishiguchi[†]

[*]Department of Integrative Biology, Michigan State University, East Lansing, MI 48824, USA; [†]Department of Molecular and Cell Biology, University of California Merced, Merced, CA 95343, USA

From the symposium "Genomic Perspectives in Comparative Physiology of Mollusks: Integration Across Disciplines" presented at the annual meeting of the Society for Integrative and Comparative Biology virtual annual meeting, January 3–February 28, 2021.

[1]E-mail: each@msu.edu

Synopsis Due to their large size (~3–5 Gb) and high repetitive content, the study of cephalopod genomes has historically been problematic. However, with the recent sequencing of several cephalopod genomes, including the Hawaiian bobtail squid (*Euprymna scolopes*), whole-genome studies of these molluscs are now possible. Of particular interest are the sepiolid or bobtail squids, many of which develop photophores in which bioluminescent bacterial symbionts reside. The variable presence of the symbiosis throughout the family allows us to determine regions of the genome that are under selection in symbiotic lineages, potentially providing a mechanism for identifying genes instrumental in the evolution of these mutualistic associations. To this end, we have used high-throughput sequencing to generate sequence from five bobtail squid genomes, four of which maintain symbioses with luminescent bacteria (*E. hyllebergi*, *E. albatrossae*, *E. scolopes*, and *Rondeletiola minor*), and one of which does not (*Sepietta neglecta*). When we performed K-mer based heterozygosity and genome size estimations, we found that the *Euprymna* genus has a higher predicted genome size than other bobtail squid (~5 Gb as compared to ~4 Gb) and lower genomic heterozygosity. When we analyzed the repetitive content of the genomes, we found that genomes in the genus *Euprymna* appear to have recently acquired a significant quantity of LINE elements that are not found in its sister genus *Rondeletiola* or the closely related *Sepietta*. Using Abyss-2.0 and then Chromosomer with the published *E. scolopes* genome as a reference, we generated *E. hyllebergi* and *E. albatrossae* genomes of 1.54–1.57 Gb in size, but containing over 78–81% of eukaryotic single-copy othologs. The data that we have generated will enable future whole-genome comparisons between these species to determine gene and regulatory content that differs between symbiotic and non-symbiotic lineages, as well as genes associated with symbiosis that are under selection.

Introduction

While cephalopods have long been model systems for the study of physiology, neurobiology, and a myriad of other disciplines, the first cephalopod genome, that of *Octopus bimaculoides*, was not sequenced until 2015 (Albertin et al. 2015; O'Brien et al. 2018). Since then the list of sequenced cephalopod species has grown significantly to include four octopuses (Albertin et al. 2015; Kim et al. 2018; Zarrella et al. 2019; Li et al. 2020) and three squid, including the Hawaiian bobtail squid *Euprymna scolopes* (Belcaid et al. 2019; da Fonseca et al. 2020; Yoshida et al. 2020). As revealed by these efforts, cephalopod genomes are variable in size but generally large, spanning a range of 2.7–5.1 Gb. However, this large size and range, which was previously hypothesized to be a result of whole-genome duplications (Hallinan and Lindberg 2011; Yoshida et al. 2011) is instead due to high variability in intronic sequences, a high proportion of repetitive sequences, and extensive genomic rearrangements (Ritschard et al. 2019). This variability within the Cephalopoda is thought to underpin the significant evolutionary novelty within the group, such as the drastic neurological innovations in the octopuses and the symbiotic organs found in squid.

Advance Access publication June 30, 2021

The sepiolid, or bobtail squid (Family Sepiolidae) are a family of small, nectobenthic cephalopods that can be found throughout the world. Most bobtail squid form a well-known symbiosis with bioluminescent gram-negative bacteria in the *Vibrionaceae*, and this association is thought to be an ancestral trait of the subfamily Sepiolinae (Pankey et al. 2014). The bacterial partners are housed in a specialized light organ that develops on the ventral surface of the ink sac to allow bacterial luminescence to be directed beneath the animal and therefore enable the host to camouflage itself through counterillumination (Jones and Nishiguchi 2004; McFall-Ngai et al. 2012). However, it is notable that several members of the Sepiolidae have independently lost the ability to form a light-organ symbiosis, and can even be found in the same habitat as their symbiotic counterparts (Fig. 1A, Nishiguchi et al. 2004). Female bobtail squid also elaborate a second symbiotic organ, called the accessory nidamental gland (ANG), which houses a consortium of bacteria that assist in the protection of the female's offspring after the cells are supplemented in the chorion, or "jelly coat" of the egg (Kerwin and Nyholm 2017; Suria et al. 2020).

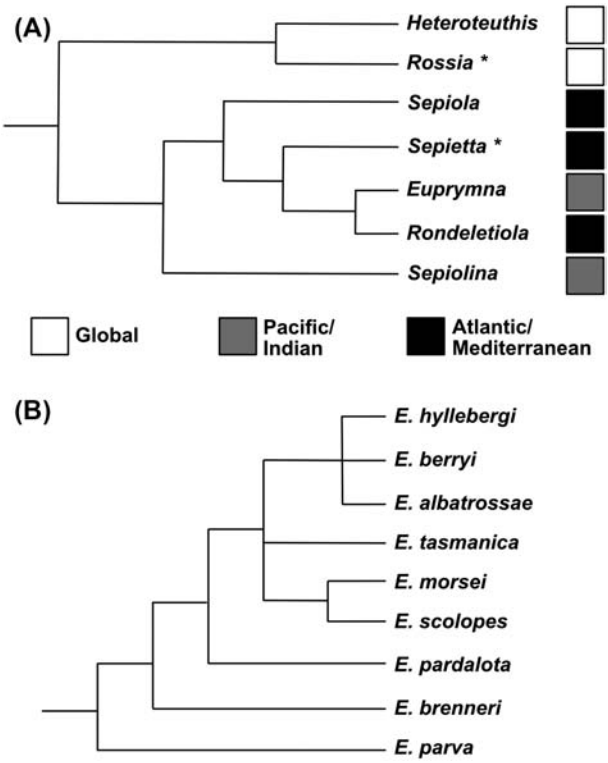

Fig. 1 Phylogenetic relationships within the *Sepiolidae*. (A) Genus-level cladogram of the *Sepiolidae*, adapted from (Sanchez et al. 2018). Geographic range (Global, Atlantic/Mediterranean Ocean, Pacific Ocean) is color-coded. Genera in which species have lost the light-organ symbiosis are shown with an asterisk. (B) Species-level cladogram of species within the genus *Euprymna*, adapted from Fig. 2 in Sanchez et al. (2019).

While ANG symbioses can be found throughout the Decapodiformes (the cephalopod superorder containing squid), bobtail squid offer a rare opportunity to study the evolution of both of these symbiotic organs in the same clade and even the same individuals, from the genomic underpinnings of organ evolution to the functional capacity of specific proteins. In fact, the recent published *E. scolopes* genome provided evidence that evolution of the light organ is associated with the co-option of recently duplicated genes, whereas the ANG is enriched in transcripts that are cephalopod novelties, suggesting separate modes of evolution in these two symbiotic organs (Belcaid et al. 2019).

One clade of particular interest within the family Sepiolidae is *Euprymna*, a genus of bobtail squid distributed across the Indo-West Pacific and Indian Oceans. While the Hawaiian bobtail squid *E. scolopes* is the best-studied species of this genus due to its status as a model system for microbial symbiosis (McFall-Ngai et al. 2012), other species also harbor symbionts and so have characteristics that make them intriguing model systems as well. Two closely related squid, *Euprymna hyllebergi* and *Euprymna albatrossae*, are of great interest due to their dispersal across the south Pacific and well-defined local population structure (Fig. 1B, Jones et al. 2006; Coryell et al. 2018). In fact, the two characterized populations of *E. hyllebergi* and their symbionts are geographically separated due to living on opposite sides of Thailand, resulting in a lack of gene flow between populations as measured by mitochondrial haplotype (Jones et al. 2006). While these data are intriguing, squid genetic studies performed so far have been based on mitochondrial cytochrome oxidase (COI) haplotypes, and so information regarding genes that might be under selection or genomic heterozygosity has yet to be collected.

In this study, we present genome size and heterozygosity estimates for four species of bobtail squids: *Rondeletiola minor*, a symbiotic Mediterranean species with a round light organ, *Sepietta neglecta*, a species found in the Mediterranean that has lost the light organ; *Euprymna hyllebergi* from Thailand; and *Euprymna albatrossae* from the Philippine archipelago (See Fig. 1). These data show that members of the *Euprymna* genus have larger predicted genome sizes and lower genome heterozygosity than that of their Atlantic sister genera, and that lack of a light organ in *S. neglecta* does not appear to change these metrics as compared to *R. minor*. In addition, using the published *E. scolopes* genome as a reference, we were able to assemble 1.54–1.56 Gb of the *E. hyllebergi* and *E. albatrossae* genomes into scaffolds containing 78.88% (*E. hyllebergi*) and 81.19%

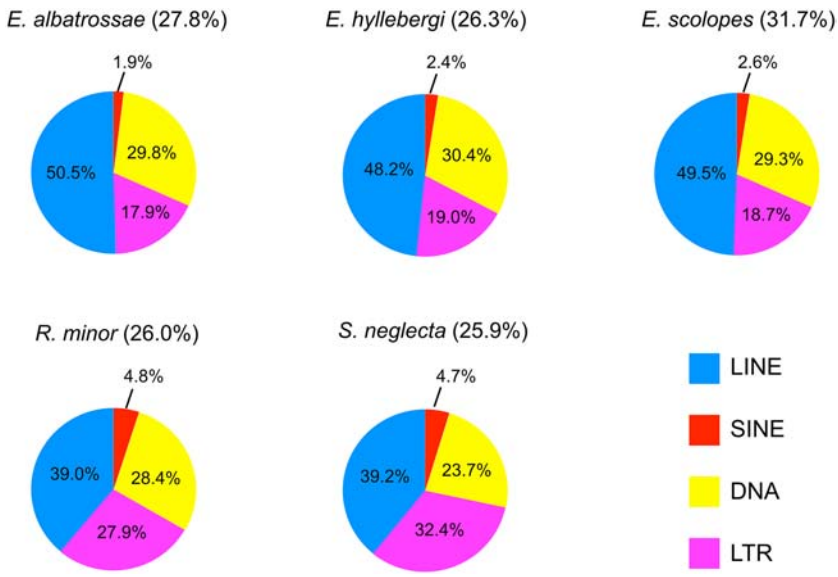

Fig. 2 Composition of repetitive elements in bobtail squid genome sequences. Pie charts show the proportional representation of LINE, SINE, DNA, and LTR elements in the genomes of five bobtail squid. Parenthetical numbers denote the proportion of the genome that is predicted to be composed of these four elements. For complete genomic composition information from which these charts are derived see Fig. S2.

(*E. albatrossae*) of the conserved complete or fragmented eukaryotic single-copy orthologs, suggesting that these scaffolds are enriched in protein-coding sequences as compared to the genome as a whole. It is our hope that this study can serve as a jumping-off point for future population genetic and genomic studies in bobtail squid.

Results and discussion

Genome characteristics of different genera within the sepiolidae

To gather genomic information for the four species of interest, we first generated short-read sequencing data from those species using Illumina NovaSeq or HiSeq Sequencers. As the genomes of marine species tend to be more heterozygotic than those of their freshwater and terrestrial counterparts (DeWoody and Avise 2000), we used a single animal for each library. After sequencing, we filtered the resultant reads to remove those with low quality scores and were shorter than 15 bp (see Materials and Methods). The number of reads and total bases sequenced in each library can be found in Table 1.

We processed the resultant filtered reads with Jellyfish (Marçais and Kingsford 2011) to generate a histogram of 21 nucleotide k-mers that we then analyzed with GenomeScope 2.0 (Ranallo-Benavidez et al. 2020) to determine the estimated size and heterozygosity of the four genomes (Table 2). Graphs showing the kmer frequency and coverage estimated by GenomeScope can be found in Fig. S1. We found that *E. hyllebergi* had a predicted genome heterozygosity of 1.362–1.512% and a predicted genome size of between 4.247 and 4.294 Gb, which is similar to the predicted heterozygosity of 0.910–1.196% and genome size of 4.356–4.421 Gb of *E. albatrossae*. These values are in contrast to those of the sister genera that we sequenced, with *R. minor* exhibiting a predicted heterozygosity of 1.883–1.996% and predicted size of 3.335–3.381 Gb, and *S. neglecta* having a predicted heterozygosity of 2.149–2.247% and predicted genome size between 3.389 and 3.411 Gb. Since the sequenced *E. scolopes* genome is approximately 5 Gb in size and that k-mer based estimation techniques such as GenomeScope can underestimate the size of repeat-rich genomes (Pflug et al. 2020), it is likely that the true genome sizes of *E. albatrossae* and *E. hyllebergi* are closer to 5.0 Gb. However, the large discrepancy between those species along with *R. minor* and *S. neglecta* suggests that even if these genome sizes are underestimated, the genome size of both species is about 1 Gb smaller than those in the genus *Euprymna*.

Transposable elements make up a large proportion of eukaryotic genomes, and can often contribute to evolution of particular groups (Wessler 2006). To determine whether the transposable element composition differed among these bobtail squid genomes, we performed a preliminary analysis of repetitive element composition using DNAPipeTE, a tool which allows for the estimation of repetitive content proportion and composition from low-coverage genomic data (Goubert et al. 2015). In addition to the four species mentioned above, we generated 46.18 Gb of 150 bp paired-end

Table 1 Species gathered in this study with sequencing library information

	Data (Gb)	Number of reads (pre-filtering)	Number of reads (post-filtering)
Euprymna hyllebergi	130.6	1,034,728,000	1,012,887,000
Euprymna albatrossae	122.3	953,767,994	935,975,416
Euprymna scolopes	46.18	322,739,228	300,217,312
Rondeletiola minor	104.7	873,945,708	849,829,818
Sepietta neglecta	171.97	1,151,520,128	1,128,800,612

Table 2 Estimates of genome size and heterozygosity for representative sepiolid squids

	Heterozygosity Min	Heterozygosity Max	Genome Size Min (Gb)	Genome Size Max (Gb)	% Repeats
E. hyllebergi	1.362%	1.512%	4.247	4.294	46.4%
E. albatrossae	0.910%	1.196%	4.356	4.421	46.5%
R. minor	1.883%	1.996%	3.335	3.381	49.0%
S. neglecta	2.149%	2.247%	3.389	3.411	44.8%

genomic sequence for *E.* scolopes, which, although not enough coverage to enable genome size estmation with GenomeScope, allowed us to determine the repeat landscape using DNAPipeTE. Our analysis showed that while the proportion of repetitive content remained relatively stable (Fig. S2), the content of these regions differed between genera. In *Euprymna* LINEs (Long Interspersed Nuclear Elements) comprise between 48.2 and 50.5% of the identifiable repetitive elements (Fig. 2), whereas they represent 39.0% in *R. minor* and 39.2% in *S. neglecta*. This is offset by a relative increase in the proportion of SINEs (Short Interspersed Nuclear Elements, 1.9%–2.6% in *Euprymna*; 4.8% in *R. minor*; 4.7% in *S. neglecta*) and LTRs (Long Terminal Repeats, 17.9–19.0% in *Euprymna*; 27.9% in *R. minor*; 32.4% in *S. neglecta*) in *S. neglecta* and *R. minor*. The sequence divergence between reads and contigs in the DNAPipeTE dataset can be used to estimate the age distribution of repetitive elements in the genome (Goubert et al. 2015), which can suggest when these elements were acquired in the lineages examined. Our data showed that, in accordance with the published analysis of the *E. scolopes* genome (Belcaid et al. 2019), all three *Euprymna* species exhibited a recent large-scale acquisition of LINEs that is not present in *R. minor* or *S. neglecta* (Fig. 3). Belcaid *et al.* suggested that this recent acquisition was responsible for the drastic difference in genome size between *O. bimaculoides* and *E. scolopes*, but our data suggest that the acqusition of the LINEs was recent enough that it may explain the large difference between the estimated genome sizes of *Euprymna* species and *R. minor* and *S. neglecta*.

Interestingly, these data suggest that the evolution of the *Euprymna* genus may be associated with an increase in genome size and a decrease in genetic variation. One potential explanation for this phenomenon is that the origin of the *Euprymna* genus was associated with an increase in genome size due to LINE proliferation in this lineage before its diversification and expansion across the Pacific (Soto et al. 2012; Soto and Nishiguchi 2014). LINEs are autonomous, Type I retrotransposons, and thus are able to replicate themselves if not silenced through epigenetic targeting or other means (Wessler 2006). They have been implicated in the rapid diversification of other genera, such as in the Antarctic fish genus *Trematomus*, where it is postulated that a breakdown of epigenetic control of transposable elements, such as LINEs, possibly due to environmental stressors, led to TE reactivation and proliferation (Auvinet et al. 2018). In *Hydra*, accumulation of sequences from a single LINE subgroup, CR1, led to the origination and diversification of a subset of the genus, the brown hydras (Wong et al. 2019). Similarly, the LINE accumulation in the genus *Euprymna* is also dominated by the CR1 subfamily, suggesting that it may promote diversification in several invertebrate lineages. After the establishment of the genus, our data suggest that it is possible that particular *Euprymna* species were generated by the restriction of small founder populations to sites in the Pacific that led to reduced genetic variation from the original population (Jones et al. 2006). More recently, it has been shown that divergence times among *Euprymna* species are large (5.1–18.6%), but less than those between *Euprymna* and *Sepiola* species (33.8–33.4%; Sanchez et al. 2019). It has also been shown that *E. albatrossae* mitochondrial haplotype distribution is strongly affected by geography and therefore can easily become genetically isolated (Coryell et al. 2018). This

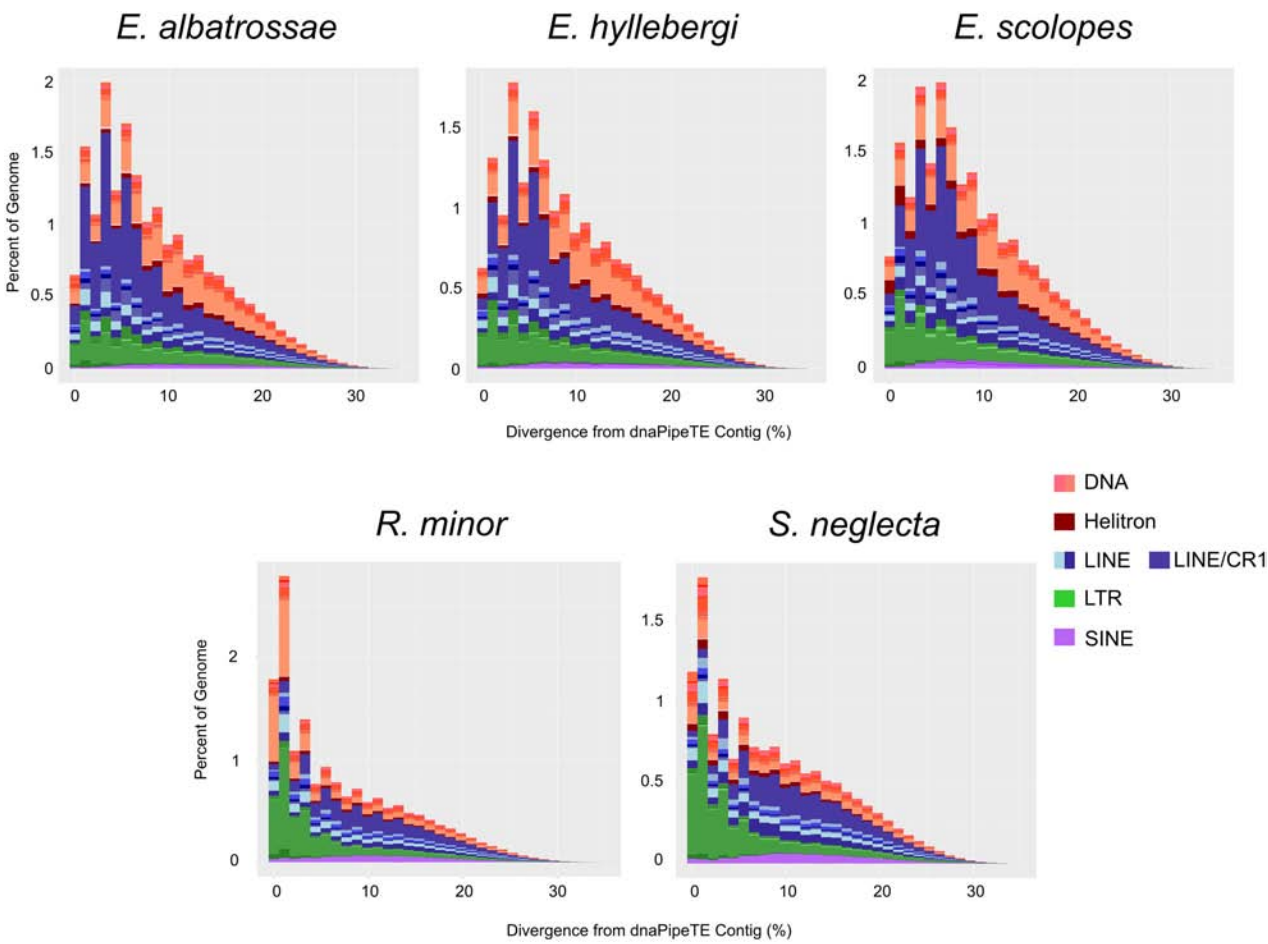

Fig. 3 Timing of TE acquisition in bobtail squid genomes as estimated by sequence divergence. Stacked bar charts denote the proportion of the genome composed of TEs relative to the divergence of the TEs from the assembled DNAPipeTE contigs, a reliable metric of the timing of TE acquisition. Colors denote different TE families, with DNA Transposons (DNA) shown in red/orange hues, Helitron TEs in maroon, LINE (Long Interspersed Nuclear Element) families in blue/blue-violet (including CR1), Long Terminal Repeats (LTR) in green, and Short Interspersed Nuclear Elements (SINEs) shown in purple.

trend is also shown in *R.* minor, in which mitochondrial haplotypes are completely segregated between populations in the Mediterranean Sea and the Bay of Biscay (Zamborsky and Nishiguchi 2011). Interestingly, this 2011 study showed little to no within-population variation at the mitochondrial sequence level between these two allopatric populations, suggesting that the species is contiguous between the Mediterranean Sea and Atlantic Ocean, with high levels of introgression. Future sequencing efforts including more individuals of each species and more species within the family will allow us to determine relative population sizes and gene flow between populations to tease apart these hypotheses. In addition, as the non-symbiotic *S. neglecta* does not appear to have major genome losses or gains relative to all other taxa sampled as measured by genome size and repeat divergences, it is unlikely that the loss of the light organ in this species is due to a large-scale event that drastically altered genome size.

Assembly of *E. albatrossae* and *E. hyllebergi* genomes

To leverage our short-read data, we decided to assemble the genomes of *E. hyllebergi* and *E. albatrossae* using a reference-based approach. While the genomes of *S. neglecta* and *R. minor* are of great interest, we decided not to pursue them further at this time, as the only reference genome available within the Sepiolidae is for *E. scolopes*, and our genome size estimation suggested that the genomes of *R. minor* and *S. neglecta* may be different enough to preclude scaffolding on a species from a different genus. *E. hyllebergi* and *E. albatrossae*, however, have estimated genome sizes comparable to the 5.1 Gb *E. scolopes* genome (Belcaid et al. 2019). To assemble the genomes of interest, we first used Abyss 2.0 to perform *de novo* genome assemblies on our *E. hyllebergi* dataset using k-mers of 44 (k44), 66 (k66), and 88 (k88) nucleotides. We then used Quast (Gurevich et al. 2013) to determine which assembly appeared to be the most

Table 3 *E. hyllebergi* and *E. albatrossae* genome assembly metrics

	# Contigs	Total length (Gb)	N50/L50	Ns/100 kbp	Complete BUSCOs (%)	Fragmented BUSCOs (%)
E. hyllebergi						
k44	70,574,706	7.56	3798/12,434	166.35	4.62	1.98
k66	31,978,991	6.78	4028/21,276	67.03	9.57	3.3
k88	24,714,025	6.95	3958/15,063	58.01	8.58	2.64
Redundans k66	4,712,130	2.87	4035/21,101	67.4	9.57	3.3
Rk66 + Chromosomer	36,568	1.54	1,208,991/342	11,626	58.09	20.79
E. albatrossae						
k66	28,343,354	6.27	3768/18,922	66.68	11.88	2.31
Redundans k66	4,345,250	2.88	3772/18,802	66.87	11.88	2.31
Rk66 + Chromosomer	35,168	1.57	1,218,260/346	10,749	66.67	14.52
E. scolopes (Belcaid *et al.*)	50,192	5.1	3,171,000/na	na	96.9	2.1

promising for future analysis. Quast compares genome assemblies by generating statistics such as the N50 value (Thrash et al. 2020) of the genome build and calculating the number of conserved single-copy orthologs (BUSCOs; Simão et al. 2015) that the *de* novo genome contains. As the k66 assembly had the highest N50 of 4028 and the highest proportion of complete and fragmented BUSCO groups (9.57 and 3.3%, respectively), we decided to use it for further assembly.

To reduce the amount of alternative contigs generated because of genome heterozygosity, we used the Redundans software package (Pryszcz and Gabaldón 2016) to remove redundant contigs from our *E. hyllebergi* and *E. albatrossae* datasets. The reduction changed the total length of both assemblies by about 2/3 (6.78 to 2.87 Gb for *E. hyllebergi* and 6.27 to 2.88 Gb for *E. albatrossae*), but did not change the total number of BUSCOs present in the data, suggesting that little to no coding information was lost in the process (Table 3, Redundans k66).

Finally, to scaffold the representative sequences onto the published *E. scolopes* genome, we used Chromosomer, a software package that, in conjunction with the NCBI BLAST package, allows for the generation of larger genomic scaffolds using the genome of a closely related species as a reference (Altschul et al. 1990; Tamazian et al. 2016). This process increased the size and completeness of both genome assemblies. For *E. hyllebergi*, the final N50 was 1.21 Mb and the assembly contained all or some portion of 78.88% of the eukaryotic single-copy conserved orthologs, whereas for *E. albatrossae*, the final N50 was 1.22 Mb and 81.19% of eukaryotic BUSCOs were detected (Table 3, Rk66 + Chromosomer). The increase seen in the represented BUSCOs is due to the decrease in fragmentation due to scaffolding by Chromosomer, and can be seen in other assemblies that utilize this pipeline (Zarrella et al. 2019).

While not complete, our genome assemblies of *E. albatrossae* and *E. hyllebergi* contain all or portions of 78.88–81.19% of the predicted conserved orthologs of eukaryotic protein-coding genes and are about 1.5 Gb in size; 29.4% of the *E. scolopes* genome size of about 5.1 Gb. This discrepancy suggests that our assembled sequence is enriched in protein-coding areas of the genome, which is expected as those regions are less likely to be repetitive and are easier to assemble with short-read data. Therefore, our analysis suggests that genome assembly using affordable short-read data, which, as opposed to long-read sequencing, can often be obtained from archival samples, can generate useful information about protein sequence and content, even in large repetitive genomes such as those found in the Cephalopoda.

Future directions

Our analysis serves as a jumping-off point for further analysis of genomes within the bobtail squid. One line of inquiry that is of great interest to us is the multiple instances of symbiosis loss within this clade. The sequencing of complete genomes of taxa within the Sepiolidae, particularly those that have lost the ability to form light-organ symbioses in the genera *Sepietta* and *Rossia* (see Fig. 1), will allow us to compare gene content and genome architecture within the group to determine the genomic signatures of symbiosis loss. Population-level analyses using whole genome sequencing or reduced representation techniques will also allow us to determine what, if any, genes are under selection in different lineages, and therefore gain a picture of the evolutionary trajectory of host genes associated with symbiosis. Population genetics using whole genomes within the cephalopods is in its infancy, and the geographic range, differential isolation, and contrasts in environ-

ment of sepiolid squid mean that the clade presents a unique opportunity to examine how these factors affect the genetics and evolution of cephalopods. In addition, the variable presence of symbiosis within the group means that we can leverage host and symbiont genomes to address co-evolution between partners at the level of the population to determine whether host or symbiont drive changes in protein-coding genes associated with the mutualism.

Materials and methods

Field collection

Adult *Euprymna hyllebergi* (Rayong, Thailand) and *E. albatrossae* (San Juan Barotac Viejo, Iloilo, Panay, Philippines) squid (∼2–4 cm in mantle length) were acquired either by dip or seine net during the evening when dark. Captured squids were brought back to the laboratory and placed on ice to anesthetize them prior to dissection or immediately placed in RNA later for total RNA or DNA extraction. For *E. hyllebergi* and *E. albatrossae*, the species designations were confirmed by sequencing of the mitochondrial gene COI as previously described (Nishiguchi et al. 2004; Guerrero-Ferreira and Nishiguchi 2007; Guerrero-Ferreira et al. 2013; Coryell et al. 2018). Adult *Sepietta neglecta* were collected by SCUBA in 7–10 m water in the Bai des Elmes (Banyuls Sur-Mer, France) and subsequently anesthetized and frozen at −80°C for subsequent nucleotide extraction. Adult *Rondeletiola minor* were collected by bottom trawling in 60–120 m depth in Banyuls Sur-Mer, France. Adult *Euprymna scolopes* were collected using dip nets in Paiko Lagoon in Oahu, Hawaii. After collection, the animals were anesthetized and then preserved in ethanol.

DNA extraction

For all genomic surveys, a single individual was sequenced due to the high rate of heterozygosity found in marine organisms (DeWoody and Avise 2000). *E. albatrossae and E. hyllebergi* DNA was extracted using approximately 25 mg of preserved tissue that was dissected from the gill or mantle of each squid. Dissected tissues were washed with 100 μL of nuclease-free water to remove any residual preservative. DNA was extracted using the DNeasy© blood and tissue kit protocol for animal tissues (Qiagen, Valencia, CA). All genomic DNA extractions were visualized on a 1% agarose gel and quantified using a Nanodrop 9600 (ThermoFisher Scientific, Waltham, MA). *R. minor* and *S. neglecta* DNA samples were extracted by the TCGB Genomics core at UCLA.

Library preparation, sequencing, and read processing

For each animal, a single genomic DNA library with an insert size of approximately 200 bp was generated at the UCLA TCGB Genomics core using the Kapa Hyperprep Kit (Roche Diagnostics, Indianapolis, IN). The quality and quantity of input DNA and the resultant sequencing libraries were verified using a TapeStation (Agilent Technologies, Santa Clara, CA). Sequences were generated by sequencing of the resultant libraries on the Illumina NovaSeq 4000 using 150 bp, paired-end sequencing (*E. albatrossae, E. hyllebergi, R. minor*) or with an Illumina HiSeq 3000 using 150 bp, paired-end sequencing (*S. neglecta*). Reads were filtered and trimmed using FastP software (Chen et al. 2018). Reads that contain at least 40% bases with a quality score lower than 15 and those that have a length of 15 bp or smaller were removed from the analysis. Specific parameters used for our data were that the 5" and 3" ends of the reads were trimmed if the mean quality of the sliding window was below the default quality cutoff (–5 and –3 options), polyG read tails were trimmed (–g option). Results of the sequencing, filtering, and trimming process can be found in Table 1.

Repetitive content estimation

The repetitive content of the genomes was estimated by the use of DNAPipeTE, a software package that allows for the estimation of repetitive element proportion and composition from low-coverage short-read sequencing (Goubert et al. 2015). Using the trimmed and filtered read dataset generated in section 3.3, a subset of 10 million reads was randomly selected by DNAPipeTE after normalization and then standard software parameters were used.

Genome statistics and assembly

Kmer-based genome size and heterozygosity estimations were performed by generating a kmer frequency histogram using Jellyfish with a kmer size of 21 (Marçais and Kingsford 2011). The diagram was then used as input to GenomeScope2.0 (Ranallo-Benavidez et al. 2020) which estimates the genome size and heterozygosity.

We used Abyss 2.0 to generate the initial *de novo* genome contigs for *E. hyllebergi* and *E. albatrossae* (Jackman et al. 2017). To determine which k-mer was most appropriate, we generated *de novo* assemblies from our *E. hyllebergi* data using k-mers of 44, 66, and 88 and then compared the resultant genomic scaffolds using Quast (Gurevich et al. 2013) to determine which k-mer generated the most complete assembly. Because of the high level of heterozygosity in marine samples, we

also tested the redundance reduction utility Redundans (Pryszcz and Gabaldón 2016) on our assembly, however, the Quast analysis suggested that it significantly reduced the assembled genome size after Chromosomer scaffolding. After identifying the best input assembly, we used Chromosomer (Tamazian et al. 2016) to scaffold the *de novo* contigs using the published *E. scolopes* genome (Belcaid et al. 2019). Finally, the protein-coding content of the genomes were predicted by Benchmarking Universal Single-Copy Orthologs, or BUSCO, analyses using the built-in utility in Quast with the eukaryotic ortholog (eukaryota_odb9) dataset which contains 303 total BUSCO groups (Simão et al. 2015).

Acknowledgments

The authors would like to thank D. Cheam and B. Pipes for assistance with sample preparation, M. Whitehorn for enabling manuscript preparation, and the Technology Center for Genomics and Bioinformatics and David Jacobs at UCLA for their help and expertise.

Funding

This work was supported by NASA [EXO 80NSSC21K0256 to MKN]; and the School of Natural Sciences at UC Merced.

Supplementary data

Supplementary data available at *ICB* online.

Data availability

The data underlying this article, including raw sequencing reads and assembled genomic sequences, are available in the NCBI BioProject Database at https://www.ncbi.nlm.nih.gov/bioproject/, and can be accessed under the IDs PRJNA718258 (E. albatrossae), PRJNA718261(*E. hyllebergi*), PRJNA736156 (*E. scolopes*), PRJNA718269 (*R. minor*), and PRJNA718263 (*S. neglecta*).

References

Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, Brenner S, Ragsdale CW, Rokhsar DS. 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature 524:220–4.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–10.

Auvinet J, Graça P, Belkadi L, Petit L, Bonnivard E, Dettaï A, Detrich WH, Ozouf-Costaz C, Higuet D. 2018. Mobilization of retrotransposons as a cause of chromosomal diversification and rapid speciation: the case for the Antarctic teleost genus *Trematomus*. BMC Genomics 19:339.

Belcaid M, Casaburi G, McAnulty SJ, Schmidbaur H, Suria AM, Moriano-Gutierrez S, Pankey MS, Oakley TH, Kremer N, Koch EJ, et al. 2019. Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. Proc Natl Acad Sci 116:3030–5.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–90.

Coryell RL, Turnham KE, Ayson EG de J, Lavilla-Pltogo C, Alcala AC, Sotto F, Gonzales B, Nishiguchi MK. 2018. Phylogeographic patterns in the Philippine archipelago influence symbiont diversity in the bobtail squid–Vibrio mutualism. Ecol Evol. 8:7421–35.

da Fonseca RR, Couto A, Machado AM, Brejova B, Albertin CB, Silva F, Gardner P, Baril T, Hayward A, Campos A, et al. 2020. A draft genome sequence of the elusive giant squid, *Architeuthis dux*. GigaScience 9:1–12.

DeWoody JA, Avise JC. 2000. Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. J Fish Biol 56:461–73.

Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. 2015. De Novo Assembly and Annotation of the Asian Tiger Mosquito (*Aedes albopictus*) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (*Aedes aegypti*). Genome Biol Evol 7:1192–205.

Guerrero-Ferreira R, Gorman C, Chavez AA, Willie S, Nishiguchi MK. 2013. Characterization of the Bacterial Diversity in Indo-West Pacific Loliginid and Sepiolid Squid Light Organs. Microb Ecol 65:214–26.

Guerrero-Ferreira RC, Nishiguchi MK. 2007. Biodiversity among luminescent symbionts from squid of the genera *Uroteuthis*, *Loliolus* and *Euprymna* (Mollusca: Cephalopoda). Cladistics 23:497–506.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–5.

Hallinan NM, Lindberg DR. 2011. Comparative analysis of chromosome counts infers three paleopolyploidies in the mollusca. Genome Biol Evol 3:1150–63.

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, et al. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res 27:768–77.

Jones BW, Lopez JE, Huttenburg J, Nishiguchi MK. 2006. Population structure between environmentally transmitted vibrios and bobtail squids using nested clade analysis. Mol Ecol 15:4317–29.

Jones BW, Nishiguchi MK. 2004. Counterillumination in the Hawaiian bobtail squid, *Euprymna scolopes* Berry (Mollusca: Cephalopoda). Mar Biol 144:1151–5.

Kerwin AH, Nyholm SV. 2017. Symbiotic bacteria associated with a bobtail squid reproductive system are detectable in the environment, and stable in the host and developing eggs. Environ Microbiol 19:1463–75.

Kim B-M, Kang S, Ahn D-H, Jung S-H, Rhee H, Yoo JS, Lee J-E, Lee S, Han Y-H, Ryu K-B, et al. 2018. The genome of common long-arm octopus *Octopus minor*. GigaScience 7:1–7.

Li F, Bian L, Ge J, Han F, Liu Z, Li X, Liu Y, Lin Z, Shi H, Liu C, et al. 2020. Chromosome-level genome assembly of the East Asian common octopus (*Octopus sinensis*) using PacBio sequencing and Hi-C technology. Mol Ecol Resour 20:1572–82.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–70.

McFall-Ngai M, Heath-Heckman EAC, Gillette AA, Peyer SM, Harvie EA. 2012. The secret languages of coevolved symbioses: insights from the *Euprymna scolopes-Vibrio fischeri* symbiosis. Semin Immunol 24:3–8.

Nishiguchi MK, Lopez JE, Boletzky Sv. 2004. Enlightenment of old ideas from new investigations: more questions regarding the evolution of bacteriogenic light organs in squids. Evol Develop 6:41–9.

O'Brien CE, Roumbedakis K, Winkelmann IE. 2018. The Current State of Cephalopod Science and Perspectives on the Most Critical Challenges Ahead From Three Early-Career Researchers. Front Physiol 9:700; 1–21.

Pankey MS, Minin VN, Imholte GC, Suchard MA, Oakley TH. 2014. Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid. Proc Natl Acad Sci 111:E4736–42.

Pflug JM, Holmes VR, Burrus C, Johnston JS, Maddison DR. 2020. Measuring Genome Sizes Using Read-Depth, k-mers, and Flow Cytometry: Methodological Comparisons in Beetles (Coleoptera). G3: Genes|Genomes|Genetics 10:3047–60.

Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res 44:e113–1-10.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun 11:1432; 1–10.

Ritschard EA, Whitelaw B, Albertin CB, Cooke IR, Strugnell JM, Simakov O. 2019. Coupled Genomic Evolutionary Histories as Signatures of Organismal Innovations in Cephalopods. Bioessays 41:1900073; 1–11.

Sanchez G, Jolly J, Reid A, Sugimoto C, Azama C, Marlétaz F, Simakov O, Rokhsar DS. 2019. New bobtail squid (Sepiolidae: Sepiolinae) from the Ryukyu islands revealed by molecular and morphological analysis. Commun Biol 2:1–15.

Sanchez G, Setiamarga DHE, Tuanapaya S, Tongtherm K, Winkelmann IE, Schmidbaur H, Umino T, Albertin C, Allcock L, Perales-Raya C, et al. 2018. Genus-level phylogeny of cephalopods using molecular markers: current status and problematic areas. PeerJ 6:e4331; 1–19.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–2.

Soto W, Nishiguchi MK. 2014. Microbial experimental evolution as a novel research approach in the Vibrionaceae and squid-Vibrio symbiosis. Front Microbiol 5:593; 1–14.

Soto W, Punke EB, Nishiguchi MK. 2012. Evolutionary perspectives in a mutualism of sepiolid squid and bioluminescent bacteria: combined usage of microbial experimental evolution and temporal population genetics. Evolution 66:1308–21.

Suria AM, Tan KC, Kerwin AH, Gitzel L, Abini-Agbomson L, Bertenshaw JM, Sewell J, Nyholm SV, Balunas MJ. 2020. Hawaiian Bobtail Squid Symbionts Inhibit Marine Bacteria via Production of Specialized Metabolites, Including New Bromoalterochromides BAC-D/D′. mSphere 5:e00166–20.

Tamazian G, Dobrynin P, Krasheninnikova K, Komissarov A, Koepfli K-P, O'Brien SJ. 2016. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. GigaScience 5:38; 1–11.

Thrash A, Hoffmann F, Perkins A. 2020. Toward a more holistic method of genome assembly assessment. BMC Bioinform 21:249; 1–8.

Wessler SR. 2006. Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci 103:17600–1.

Wong WY, Simakov O, Bridge DM, Cartwright P, Bellantuono AJ, Kuhn A, Holstein TW, David CN, Steele RE, Martínez DE. 2019. Expansion of a single transposable element family is associated with genome-size increase and radiation in the genus *Hydra*. Proc Natl Acad Sci 116:22915–7.

Yoshida M, Imoto J, Kawai Y, Funahashi S, Minei R, Akizuki Y, Ogura A, Nakabayashi K, Yura K, Ikeo K. 2020. Genomic and Transcriptomic Analyses of Bioluminescence Genes in the Enope Squid *Watasenia scintillans*. Mar Biotechnol 22: 760–71.

Yoshida M, Ishikura Y, Moritaki T, Shoguchi E, Shimizu KK, Sese J, Ogura A. 2011. Genome structure analysis of molluscs revealed whole genome duplication and lineage specific repeat variation. Gene 483:63–71.

Zamborsky DJ, Nishiguchi MK. 2011. Phylogeographical Patterns among Mediterranean Sepiolid Squids and Their Vibrio Symbionts: Environment Drives Specificity among Sympatric Species. Appl Environ Microbiol 77:642–9.

Zarrella I, Herten K, Maes GE, Tai S, Yang M, Seuntjens E, Ritschard EA, Zach M, Styfhals R, Sanges R, et al. 2019. Scientific Data 6:13; 1–8.