Chromosome-Level Genome Assembly of the Bioluminescent Cardinalfish *Siphamia tubifer*: An Emerging Model for Symbiosis Research

A. L. Gould[1,*], J. B. Henderson[2], and A. W. Lam[2]

[1]Ichthyology Department, Institute for Biodiversity Science and Sustainability, California Academy of Sciences, 55 Music Concourse Dr., San Francisco, California 94118, USA

[2]Center for Comparative Genomics, Institute for Biodiversity Science and Sustainability, California Academy of Sciences, 55 Music Concourse Dr., San Francisco, California 94118, USA

*Corresponding author: E-mail: agould@calacademy.org.

Accepted: 23 March 2022

Abstract

The bioluminescent symbiosis involving the sea urchin cardinalfish *Siphamia tubifer* and the luminous bacterium *Photobacterium mandapamensis* is an emerging vertebrate model for the study of microbial symbiosis. However, little genetic data are available for the host, limiting the scope of research that can be implemented with this association. We present a chromosome-level genome assembly for *S. tubifer* using a combination of PacBio HiFi sequencing and Hi–C technologies. The final assembly was 1.2 Gb distributed on 23 chromosomes and contained 32,365 protein coding genes with a BUSCO score of 99%. A comparison of the *S. tubifer* genome to that of another nonluminous species of cardinalfish revealed a high degree of synteny, whereas a comparison to a more distant relative in the sister order Gobiiformes revealed the fusion of two chromosomes in the cardinalfish genomes. The complete mitogenome of *S. tubifer* was also assembled, and an inversion in the vertebrate WANCY tRNA genes as well as heteroplasmy in the length of the control region were discovered. A phylogenetic analysis based on whole the mitochondrial genome indicated that *S. tubifer* is divergent from the rest of the cardinalfish family, highlighting the potential role of the bioluminescent symbiosis in the initial divergence of *Siphamia*. This high-quality reference genome will provide novel opportunities for the bioluminescent *S. tubifer–P. mandapamensis* association to be used as a model for symbiosis research.

Key words: Apogonidae, symbiosis, HiFi, Hi–C, heteroplasmy.

Significance

This study presents a high-quality chromosome-level assembly of a bioluminescent coral reef fish that is being developed as a vertebrate model for symbiosis research for which there is little genetic information available. This genome will serve as a valuable resource for symbiosis research as well as the study of the evolution of bioluminescence and reef fishes more broadly.

Introduction

The cardinalfish genus *Siphamia* (Kurtiformes: Apogonidae) is comprised of 25 symbiotically bioluminescent species distributed throughout the Indo-Pacific. *Siphamia tubifer* has the broadest distribution, spanning from East Africa to the French Polynesian Islands (Gon and Allen 2012), and is also the most well-studied *Siphamia* species to date (Eibl-Eibesfeldt 1961; Tamura 1982; Gould et al. 2014,

2015, 2016; Gould and Dunlap 2017), including several studies describing its symbiosis with the luminous bacterium, *Photobacterium mandapamensis* (Iwai 1958, 1971; Dunlap and Nakamura 2011; Dunlap et al. 2012; Gould and Dunlap 2019). Unlike most symbiotically luminous fishes, *S. tubifer* is a shallow, reef-dwelling species that can be maintained in aquaria, both with and without its luminous symbiont, rendering it to be experimentally tractable (Dunlap et al. 2012). Thus, the *S. tubifer–P. mandapamensis* association is an emerging model for the study of microbial symbiosis and is especially well-suited for studies of the vertebrate gut microbiome. Despite an accumulation of knowledge of the biology of *S. tubifer* and its symbiosis with *P. mandapamensis*, there is little genetic information available for the fish. A high-quality reference genome of *S. tubifer* will unlock new opportunities to investigate the genetic underpinnings of the symbiosis. We present a chromosome-level assembly of the *S. tubifer* genome produced by combining PacBio HiFi sequencing technology and chromosome conformation capture methods (Hi–C, Lieberman-Aide et *al.* 2009; van Berkum et al. 2010). We then examine synteny between the *S. tubifer* genome and the chromosome-level genomes of two nonluminous relatives. We also assembled the whole mitochondrial genome for *S. tubifer* and use it to infer *S. tubifer*'s phylogenetic position within the Apogonidae.

Results

Genome Size Estimation, Assembly, and Chromosome Mapping

A total of 2,110,443 HiFi circular consensus sequence (CCS) reads consisting of 27,799,385,228 bp were generated from the HiFi library, with a polymerase N50 of 183,061 and subread N50 of 13,439. Over 97% of the reads were between 12,000 and 15,000 bp. From these sequences, the GenomeScope size estimate using kmer lengths 21 and 25 ranged from 947,587,691 to 964,260,239 bp. After contaminant and mitochondrial sequence removal, 2,109,973 sequence reads remained with 6,158,291 bp excluded from the initial reads. These remaining sequences were used as input for the hifiasm assembler to scaffold with the Hi–C reads.

For the Hi–C libraries, a total of 742,280,226 and 506,411,380 reads were produced from the muscle and brain tissue, respectively. Of those, 100% of the muscle reads and 99.98% of the brain reads were clean and of high quality with GC contents of 39.3% and 43.9%. The Juicer mapping program found 245,145,667 read pairs with Hi–C contacts (fig. 1*a*). After interactive modification with JuiceBox Assembly Tools (JBAT) (Durand et al. 2016a; Dudchenko et al. 2018), guided by the 3d-dna program contig placement and orientation, the resulting genome assembly was 1.2 Gb distributed on 23 chromosomes (fig. 1*b*), and 1.81% unplaced scaffolds, with a contig N50 of 2.3 Mb and scaffold N50 of 51.1 Mb (table S1, Supplementary Material online) and 37.71% GC content. There are 1,960 contigs constituting chromosomal sequences. An additional two dozen smaller contigs were identified as contaminants by the final nt blastn search and were removed to produce the final assembly with a slightly lower unplaced scaffold percentage (1.74%). The 23 chromosomes in the *S. tubifer* genome assembly are numbered 1 to 22 and 24 based on synteny with the genome of another cardinalfish, *Sphaeramia orbicularis* (GCF_902148855.1), which is based on synteny with the 24 chromosome medaka genome (ASM223467v1). BUSCO completeness assessment from the 3,640 entry Actinopterygii dataset show 99% complete with just 13 of the genes not found (MetaEuk mode: 98% complete, AUGUSTUS mode: 97.2% complete).

Genome Annotation and Statistics

Repeat analysis indicated 626,216,533 bp (52.11% of the genome) classified as repeats, of which, most (23.7% of the genome) are DNA repeat elements. Additionally, 7.03% of the genome contains long interspersed nuclear elements (LINEs), with 16.28% of the genome characterized as unclassified repeats. The extent of repeats may account for the discrepancy between the assembly size and the GenomeScope estimates.

Gene annotation identified 30,117 gene models with a total length of 360,171,123 bp, (29.99% of the genome). Exons comprised of 53,076,342 bp are 4.42% of the genome and averaged 9.64 per gene; fewer than 10% are single exon genes. Additional per chromosome details of genes, exons, and introns are outlined in table S2, Supplementary Material online. The orbiculate cardinalfish, *Sp. orbicularis* was the closest functional annotation reference for 17,079 (56.7%) of the 30,117 *S. tubifer* gene models. This was followed by several other fish species: *Lates calcarifer* (2,317), *Seriola dumerili* (1,357), *Larimichthys crocea* (995), and *Stegastes partitus* (779).

Mitochondrial Genome

There were 5,124,329 total bp in the 392 HiFi reads that matched the 60% query coverage used in the mitochondrial sequence analysis, of which, 176 reads containing 2,302,235 bp had at least 90% of their read length covered. The complete mitochondrial genome averaged 17,905 bp, but varied due to heteroplasmy in the length of the control region (CR; fig. 2*a*) and contained 13 protein coding genes, 22 tRNA genes, and 2 rRNA genes, as expected. However, an inversion was detected within the region that codes for five tRNAs known as the WANCY region, resulting in a WACNY gene order (fig. 2*a*). All of the reads had enough tRNAs to affirm the WACNY order; 174 encompassed all of the five tRNA genes, and the other two reads began with CNY and NY. There were also 135 HiFi reads that

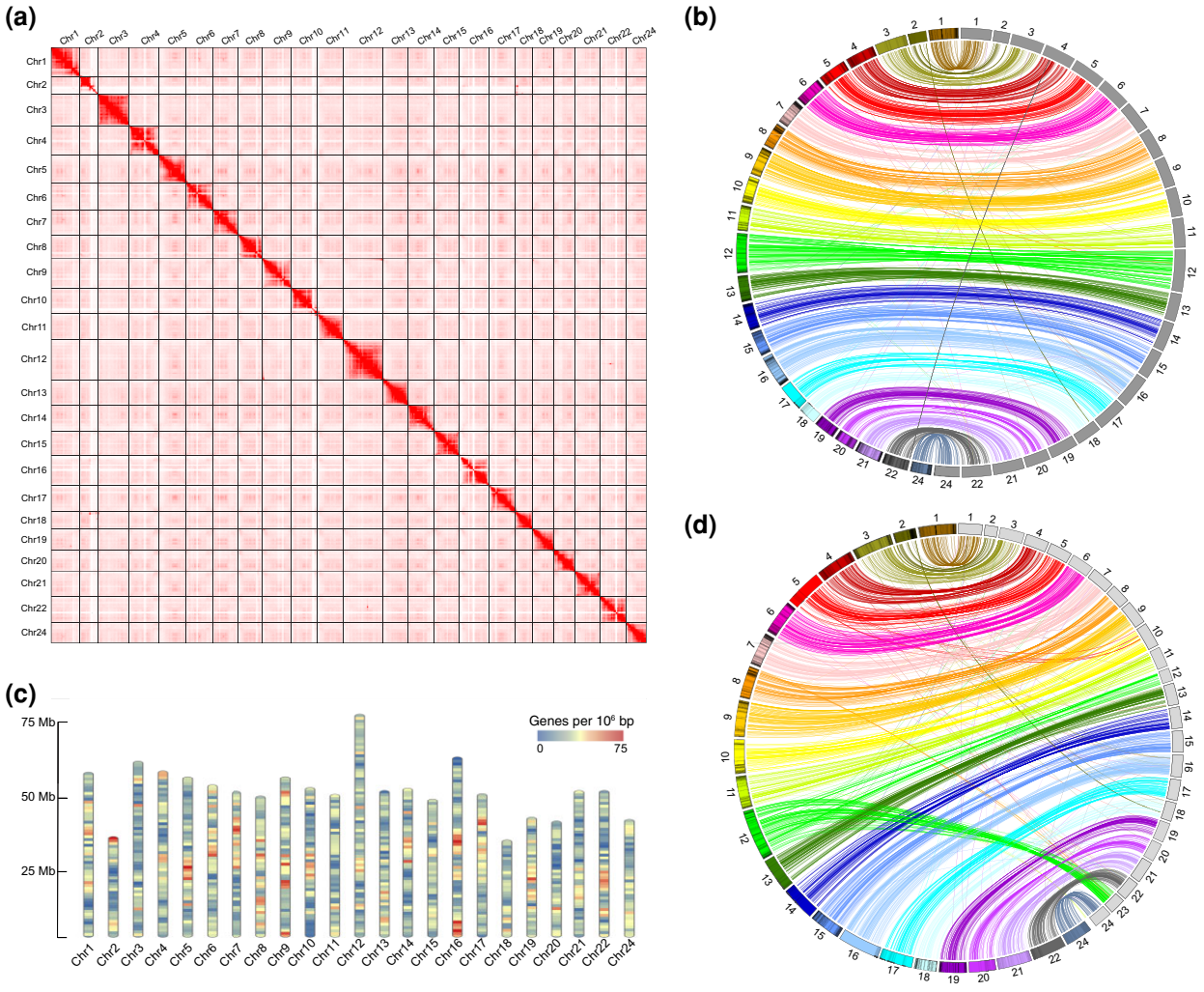

Fig. 1.—(*a*) Hi–C contact heatmap for *Siphamia tubifer*. Black lines indicate chromosome boundaries. (*b*) Gene density distribution across the 23 chromosomes of the *S. tubifer* genome. (*c*) Circos plots depicting synteny between the genomes of *S. tubifer* and the orbiculate cardinalfish, *Sphaeramia orbicularis* (1.3 Gb) and (*d*) the mudskipper *Periophthalmus magnuspinnatus* (702 Mb). Each chromosome in the *S. tubifer* genome is represented by a distinct color, whereas the *Sp. orbicularis* and *P. magnuspinnatus* chromosomes are shown in dark and light gray, respectively. Links between the genomes represent single copy orthologs from the BUSCO Actinopterygii gene set.

encompassed the *Pro* tRNA gene, the entire CR, and the *Phe* tRNA gene from which the CR lengths were determined. The length of the CR ranged from 2,620 to 6,544 bp with a mean of 4,243 bp (fig. 2*b*). Of the 135 sequences, 130 had a 60 bp repeat beginning after *Pro*, and the other five reads had similar repeats. This sequence, or a one to four nucleotide indel or SNP variation of it, was repeated 2–69 times in each read. A goose hairpin sequence (Quinn and Wilson 1993), in this case, C_7TAC_7, was found in 133 of the 135 CR sequences (the two others had C_7TCAC_7 and $C_7TAC_4CAC_8$), which started between 350 and 360 bp from the end of the CR region (fig. 2*a*). The maximum likelihood phylogeny based on the whole mitochondrial genome (excluding the CR) confirms that *S. tubifer* is divergent from the rest of the Apogonidae family but is a member of the Apogonoidei clade, which is sister to the Gobioidei (Ghezelayagh et al. 2021) (fig. 2*c*; fig. S1, Supplementary Material online).

Discussion

Combining PacBio HiFi sequencing with Hi–C technology, we assembled a high-quality, chromosome-level genome for the symbiotically luminous cardinalfish *S. tubifer*.

The BUSCO score of 99% completeness indicates that this is a near-complete genome and will serve as a valuable resource for future research. This is only the second cardinalfish genome assembly to date, and our comparison of the two indicates there is significant synteny between them, despite the divergence of *S. tubifer* from the rest of the cardinalfish family. An additional comparison to a

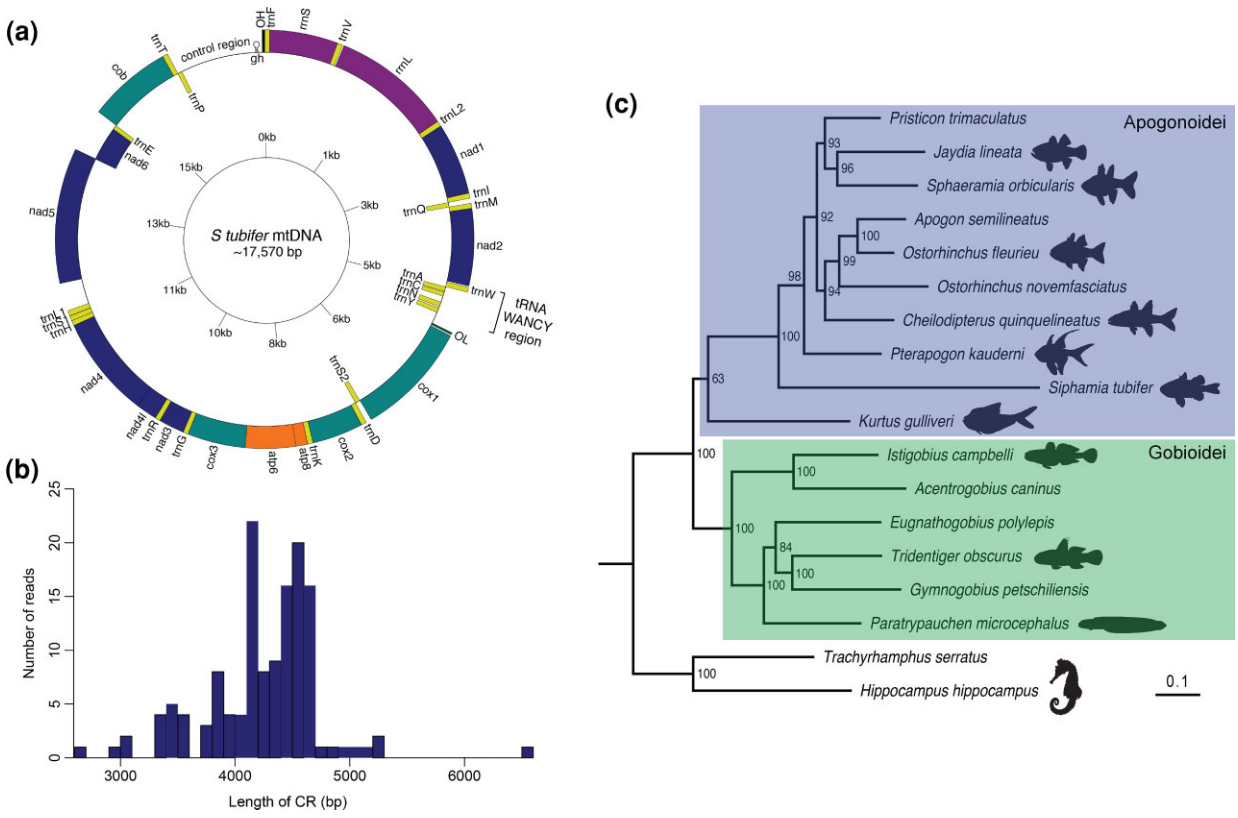

Fig. 2.—(*a*) Gene map of the complete mitogenome of *Siphamia tubifer*. All genes are labeled including the tRNA WANCY region as well as the control region and the approximate location of the goose hairpin (gh) within the control region. (*b*) Histogram depicting heteroplasmy in the length of the control region observed for the HiFi sequence reads spanning the entire region. (*c*) Maximum likelihood tree depicting the phylogenetic relationships of several cardinalfish species for which there is whole mitochondrial genome data available, including *S. tubifer* from this study, in relation to another member of the Apogonoidei clade, *Kurtus gulliveri*, and several species of gobies in the sister clade Gobioidei. Two Syngnathiformes species are included as an outgroup. The relationships are based on whole mitochondrial DNA sequences excluding the control region using the GTR + F+I + G4 model of substitution. Bootstrap support values (500 replicates) are listed at the nodes. The scale bar indicates nucleotide substitutions per site. Associated GenBank accession numbers for each species are listed in table S3, Supplementary Material online.

more distant genome in the Gobiiformes, revealed a merging of two chromosomes resulting in 23 chromosomes in both cardinalfish genomes. Determining whether this is a common feature of all cardinalfish and when this merge occurred would require additional chromosome-level assemblies of other species' genomes in the two orders.

A byproduct of producing HiFi reads is the large percentage of the mitogenome captured in an individual read (Formenti et al. 2021). These genomes typically range from 16,000 to 22,000 bp, which makes them amenable to discover reordering, duplicated regions leading to pseudogenes, duplicated CRs, CR repeats, and heteroplasmy within an individual. With 176 mitochondrial HiFi reads we were able to determine the unique WACNY ordering of the WANCY region of tRNAs of this individual not reported in 3,034 MitoFish website annotations (June 3, 2021), although mitochondrial gene-order rearrangements have been documented for other teleost fishes (e.g., Inoue et al 2003; Poulsen et al. 2013), including rearrangements in the WANCY region (Poulsen et al. 2019). Future HiFi sequencing of the mitogenomes of additional *S. tubifer* specimens and other *Siphamia* species would indicate whether the WACNY gene order observed here is unique to this individual or is a common feature of the species or genus. We also observed heteroplasmy in the length of the CR likely caused by repeat expansion and/or contraction, which has been documented for other fish species, including the three-spined stickleback (Lindén et al. 2004), several sardine species (Samonte et al. 2000), and the flatfish *Platichthys flesus* (Hoarau et al. 2002). Thus, such variability in the copy number of tandem repeats in the CR could be a more common occurrence that has been undetected with other sequencing approaches. As HiFi sequencing becomes more widely implemented, heteroplasmy in the mitogenome might be documented more frequently for other organisms (e.g., Formenti et al. 2021).

The phylogenetic analysis based on whole mitogenomes indicated that *S. tubifer* is divergent from the rest of the cardinalfish family (Apogonidae), a placement previously

supported and estimated to have occurred approximately 50 Ma (Thacker 2014). The evolutionary relationship of *S. tubifer* as sister to the rest of the cardinalfishes raises the possibility that the bioluminescent symbiosis played a role in the host's initial divergence and speciation from a common ancestor. The acquisition of bacterial endosymbionts as a primary mechanism by which new species can arise was proposed nearly a century ago (Wallin 1927), and speciation by symbiosis has since been documented (Brucker and Bordenstein 2012). Future studies identifying host genes involved in the *S. tubifer–P. mandapamensis* symbiosis are now possible with the reference genome of *S. tubifer* and will help determine whether the symbiosis played a role in host speciation for *Siphamia*.

Materials and Methods

Tissue Collection, DNA Extraction, and Sequencing

All tissue was obtained from a single female *S. tubifer* specimen collected in Okinawa, Japan (26.66°N, 127.88°E). The fish was collected and euthanized following approved protocols and permits for the capture, care, and handling of fish by the California Academy of Science's Institutional Animal Care and Use Committee. Immediately following euthanasia, fresh muscle tissue was sampled from the flanking region of the fish for high molecular weight (HMW) DNA extraction using a phenol–chloroform protocol provided by Pacific Biosciences of California, Inc. Fresh muscle and brain tissue were also sampled from the fish for Hi–C methods. The HMW DNA was prepared for PacBio HiFi sequencing at UC Berkeley's QB3 Genomics Sequencing Lab (Berkeley, CA) and sequenced on one Sequel II 8 M SMRT Cell.

Hi–C Library Preparation and Sequencing

In situ Hi–C libraries were prepared from the freshly homogenized muscle and brain tissues following a previously described protocol (Rao et al. 2014) with slight modifications. After the Streptavidin pull-down step, the biotinylated Hi–C products underwent end-repair, ligation, and enrichment using the NEBNext® Ultra™II DNA Library Preparation kit (New England Biolabs Inc.). Titration of the number of PCR cycles was performed as previously described (Belton et al. 2012). The final libraries were then sequenced as paired-end 150 bp reads on the Illumina NovaSeq 6000 platform by Novogene Corporation, Inc.

Genome Size Estimation, Assembly, and Chromosome Mapping

CCSs were generated using ccs v5.0.0 (https://github.com/PacificBiosciences/pbbioconda), from 35.95 M subreads, representing 442.25G bases, and filtered to produce HiFi reads, defined as having at least two circular passes and minimum of 99.9% accuracy. A custom script created a.fastq file containing the HiFi reads extracted from the.bam output file of the ccs step. Jellyfish (Marcais and Kingsford 2012) was then used to count and create histograms of kmers size 21 and 25 from the HiFi reads, and GenomeScope v2.0 (Ranallo-Benavidez et al. 2020) was run on each set to estimate the genome size.

Next, filtering was performed to remove contaminant sequences. Since using blastn (Altschul et al. 1990) and other similar tools are inefficient with long reads, we first used minimap2 (Li 2018) with the genome of the orbiculate cardinalfish, *Sp. orbicularis*, to exclude matching reads from further contaminant analysis. For the remaining sequences, blastn was used against a database of the fish's luminous symbiont, *P. mandapamensis* (Urbanczyk et al. 2011), and the first 500 bases of the remaining long reads were used as blastn queries against the nt database with option -taxidlist restricting the search to bacteria, excluding those with e-value greater than -1e10. Mitochondrial DNA sequences were also identified and removed for separate analysis by using blastn against a database of three Apogonidae mitochondrial genomes: *Sp. orbicularis*, *Ostorhinchus fleurieu*, and *Pristicon trimaculatus*. Subsequent nuclear genome analysis used the remaining long read HiFi sequences with contaminant and mitochondrial sequences removed.

The remaining HiFi sequences were assembled with hifiasm v0.13-r308 (Cheng et al. 2021), with purge_dups (Guan et al. 2020) to separate out duplicate haplotigs, producing a primary assembly of the higher quality contigs and an alternate assembly of contigs with duplicates. The combined brain and muscle tissue Hi–C reads were then mapped using juicer v1.6 (Durand et al. 2016b) against the hifiasm assembled contig-level genome. We ran 3d-dna v180922 (Dudchenko et al. 2017) with its early-exit flag to create an input file for JBAT (Durand et al. 2016a; Dudchenko et al. 2018) that represents the assembly with contigs ordered and oriented in a candidate chromosome-level depiction. Using JBAT, we interactively updated the location and orientation of contigs and their delineation within chromosomes (fig. 1*a*). This assembly was also queried against the nt database to identify any additional contaminants for removal.

To assess the level of genome completeness, we ran BUSCO v5.12 (Simão et al. 2015) with the 3,640 entry Actinopterygii dataset in both MetaEuk (Karin et al. 2020) and AUGUSTUS (Keller et al. 2011) modes. We then used a custom script to update BUSCOs found by AUGUSTUS that were missing in the MetaEuk results and another to report the combined scores.

Gene Annotation and Synteny

Prior to gene annotation, *de novo* repeats were identified using RepeatModeler v2.0.1(Flynn et al. 2020). First,

the.fasta file representing these species-specific repeat models and the vertebrate repeat models from Repbase RepeatMasker libraries v20181026 were combined and used in Repeatmasker v4.0.9 (Smit et al. 2013–2015) with the options -small -xsmall and -nolow to create the soft-masked repeat version of the assembly file used for gene model annotation. BRAKER2 (Brůna et al. 2021), using GeneMark-EP+ (Brůna et al. 2020) and AUGUSTUS, combined with the vertebrate protein database from OrthoDB v10 (Kriventseva et al. 2019), was used for gene annotation. The output of potential gene models represented in.gff3, amino acid, and DNA files was subject to additional filtering and functional annotation.

To check for protein domains, we ran InterProScan v5.51–85.0 (Jones et al. 2014) on the amino acid sequences found in the BRAKER2 results. These sequences were also used as queries for a blastp run on three databases: SwissProt, TrEMBL, and the vertebrate proteins from OrthoDB v10. The DNA versions of the sequences were queried with blastn against the nt database (February 13, 2021). Gene models were kept for those sequences with an InterProScan determined protein domain and one of the four database searches yielding a match with an e-value 0.1e−6 or less. Protein domain IDs and Gene Ontology terms, as determined by the InterProScan output, were added to the.gff3 file for each gene model as was the functional annotation description. tRNAscan-SE v2.0.8 (Chan et al. 2021) was implemented to identify tRNAs.

Synteny between the *S. tubifer* genome and the chromosome-level genome assemblies of *Sp. orbicularis* and *Periophthalmus magnuspinnatus* (GCA_009829125.1) was characterized using the set of single copy orthologs identified from the BUSCO (Simão et al. 2015) Actinopterygii gene set, and the output was visualized in Circos (Krzywinski et al. 2009).

Mitochondrial Genome Assembly and Analysis

Mitochondrial genome analysis was based on sequences matching at least 60% query coverage in a blastn match (qcovus format specifier) to one of the three Apogonidae mitochondrial genomes previously mentioned. When matched to the reverse strand, sequences were reverse complemented ("_RC" appended to the name) so that all sequences have the same orientation. Megahit (Li et al. 2015) was then run on these sequences to assemble a draft mitogenome, and MITOS2 (Bernt et al. 2013) was used to annotate the mitogenome. Mitfi (Jühling et al. 2012) was used to identify tRNAs from 176 reads that matched at least 90% query coverage to one of the three closely related species' mitogenomes. Tandem Repeat Finder (Benson 1999) was run to find repeats in the CR. The phylogenetic placement of *S. tubifer* within the cardinalfishes and Kurtiformes order was then inferred using the mitochondrial genome sequence. Whole mitogenomes (excluding the CR) were aligned using MAFFT (Katoh et al. 2002), and maximum likelihood trees were constructed with raxml-ng (Kozlov et al. 2019) using the substitution model with the lowest BIC score as predicted by IQtree (Nguyen et al 2015) and 500 bootstrap replicates.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgements

This work was supported by the National Institutes for Health (grant number NIH-DP5-OD026405-01).

Author Contributions

A.L.G. conceived of the project and secured funding for the work. A.L.G. carried out tissue dissections and A.W.L. performed the high molecular weight DNA extractions and HI–C library preparations. J.B.H. carried out the genome assembly and associated bioinformatics. A.L.G. performed the phylogenetic analyses and data analyses. A.L.G. and J.B.H. contributed to the discussion and interpretation of the results and writing of the manuscript. All authors approve of the submitted version of this manuscript.

Data Availability

All raw sequencing and genome assembly data of *Siphamia tubifer* have been deposited at NCBI under accession number PRJNA736963. Genome assembly and associated sequencing data are available under NCBI Bioproject PRJNA736963.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Belcaid M, et al. 2019. Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. Proc Natl Acad Sci USA 116: 3030–3035.

Belton JM, et al. 2012. Hi–C: a comprehensive technique to capture the conformation of genomes. Methods 58:268–276.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.

Bernt M, et al. 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 69:313–319.

Brucker RM, Bordenstein SR. 2012. Speciation by symbiosis. Trends Ecol Evol. 27(8):443–451.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinformatics 3:lqaa108.

Brůna T, Lomsadze A, Borodovsky M. 2020. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genomics Bioinformatics 2:lqaa026.
Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 44:e147.
Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. BioRxiv:614032.
Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat Methods 18:170–175.
Dudchenko O, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356:92–95.
Dudchenko O, et al. 2018. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. BioRxiv:254797
Dunlap PV, Gould AL, Wittenrich ML, Nakamura M. 2012. Symbiosis initiation in the bacterially luminous sea urchin cardinalfish *Siphamia versicolor*. J Fish Biol. 81:1340–1356.
Dunlap PV, Nakamura M. 2011. Functional morphology of the luminescence system of *Siphamia versicolor* (Perciformes: Apogonidae), a bacterially luminous coral reef fish. J Morph. 272:897–909.
Durand NC, et al. 2016a. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Systems. 3:99–101.
Durand NC, et al. 2016b. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Systems. 3:95–98.
Eibl-Eibesfeldt I. 1961. Eine symbiose zwischen fischen (*Siphamia versicolor*) und seeigeln. Z Tierpsychol. 18:56–59.
Faber JE, Stepien CA. 1998. Tandemly repeated sequences in the mitochondrial DNA control region and phylogeography of the pike-perchesstizostedion. Mol Phylogenet Evol. 10:310–322.
Farrer RA. 2017. Synima: a Synteny imaging tool for annotated genome assemblies. BMC Bioinform. 18:1–4.
Flynn JM, et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA 117: 9451–9457.
Formenti G, et al. 2021. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. Genome Biol. 22:1–22.
Ghezelayagh A, et al. 2021. Prolonged morphological expansion of spiny-rayed fishes following the end-Cretaceous. bioRxiv.
Gon O, Allen GR. 2012. Revision of the Indo-Pacific cardinalfish genus *Siphamia* (Perciformes: Apogonidae). Zootaxa 3294:1–84.
Gould AL, Dougan KE, Koenigbauer ST, Dunlap PV. 2016. Life history of the symbiotically luminous cardinalfish *Siphamia tubifer* (Perciformes: Apogonidae). J Fish Biol. 89:1359–1377.
Gould AL, Dunlap PV. 2017. Genomic analysis of a cardinalfish with larval homing potential reveals genetic admixture in the Okinawa Islands. Mol Ecol. 26:3870–3882.
Gould AL, Dunlap PV. 2019. Shedding light on specificity: population genomic structure of a symbiosis between a coral reef fish and luminous bacterium. Front Microbiol. 10:2670.
Gould A, Fritts-Penniman A, Gaisiner A. 2021. Museum genomics illuminate the high specificity of a bioluminescent symbiosis across a genus of reef fish. Front Ecol Evol. 9:18.
Gould AL, Harii S, Dunlap PV. 2014. Host preference, site fidelity, and homing behavior of the symbiotically luminous cardinalfish, *Siphamia tubifer* (Perciformes: Apogonidae). Mar Biol. 161: 2897–2907.
Gould AL, Harii S, Dunlap PV. 2015. Cues from the reef: olfactory preferences of a symbiotically luminous cardinalfish. Coral Reefs. 34: 673–677.
Guan D, et al. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36:2896–2898.
Haas BJ, Delcher AL, Wortman JR, Salzberg SL. 2004. DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics 20:3643–3646.
Hoarau G, Holla S, Lescasse R, Stam WT, Olsen JL. 2002. Heteroplasmy and evidence for recombination in the mitochondrial control region of the flatfish *Platichthys flesus*. Mol Biol Evol. 19:2261–2264.
Inoue JG, Miya M, Tsukamoto K, Nishida M. 2003. Evolution of the deep-sea gulper eel mitochondrial genomes: large-scale gene rearrangements originated within the eels. Mol Biol Evol. 20: 1917–1924.
PacificBiosciences. 2020. Ipa hifi genome assembler.
Iwai T. 1958. A study of the luminous organ of the apogonid fish *Siphamia versicolor* (Smith and Radcliffe). J Wash Acad Sci. 48: 267–270.
Iwai T. 1971. Structure of luminescent organ of apogonid fish, *Siphamia versicolor*. Jap J Ichthy. 18:125–127.
Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240.
Jühling F, et al. 2012. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. Nucleic Acids Res. 40:2833–2845.
Kaeding AJ, et al. 2007. Phylogenetic diversity and cosymbiosis in the bioluminescent symbioses of "*Photobacterium mandapamensis*". Appl Environ Microbiol. 73:3173–3182.
Karin EL, Mirdita M, Söding J. 2020. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome 8:1–15.
Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics 27:757–763.
Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35: 4453–4455.
Kriventseva EV, et al. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 47:D807–D811.
Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639–1645.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100.
Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31: 1674–1676.
Lieberman-Aiden E, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–293.
Lindén M, Påhlsson C, Stärner H. 2004. Tandem repeat polymorphism and heteroplasmy in the mitochondrial DNA control region of threespine stickleback (Gasterosteus aculeatus). Behaviour 141(11–12):1357–1369.
Ludwig A, May B, Debus L, Jenneckens I. 2000. Heteroplasmy in the mtDNA control region of sturgeon (*Acipenser*, *Huso* and *Scaphirhynchus*). Genetics 156:1933–1947.
Marcais G, Kingsford C. 2012. Jellyfish: a fast k-mer counter. Tutorialis e Manuais 1:1–8.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32:268–274.

Poulsen JY, et al. 2013. Mitogenomic sequences and evidence from unique gene rearrangements corroborate evolutionary relationships of Myctophiformes (Neoteleostei). BMC Evol Biol. 13:111.

Poulsen JY, Sado T, Miya M. 2019. Unique mitochondrial gene order in *Xenodermichthys copei* (Alepocephalidae: Otocephala)—a first observation of a large-scale rearranged 16S–WANCY region in vertebrates. Mitochondrial DNA Part B Resour. 4: 511–514.

Quinn TW, Wilson AC. 1993. Sequence evolution in and around the mitochondrial control region in birds. J Mol Evol. 37: 417–425.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 20 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 11:1432.

Rao SS, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159: 1665–1680.

Samonte IE, Pagulayan RC, Mayer WE. 2000. Molecular phylogeny of Philippine freshwater sardines based on mitochondrial DNA analysis. J Heredity. 91:247–253.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Smit A, Hubley R, Green P. 2013–2015. *RepeatMasker Open-4.0*. Available from: http://www.repeatmasker.org

Tamura R. 1982. Experimental observations on the association between the cardinalfish (*Siphamia versicolor*) and the sea urchin (*Diadema setosum*). Galaxea 1:1–10.

Thacker CE. 2014. Species and shape diversification are inversely correlated among gobies and cardinalfishes (Teleostei: Gobiiformes). Org Divers Evol. 14:419–436.

Turanov SV, Lee YH, Kartavtsev YP. 2019. Structure, evolution and phylogenetic informativeness of eelpouts (Cottoidei: Zoarcales) mitochondrial control region sequences. Mitochondrial DNA Part A 30:264–272.

Urbanczyk H, et al. 2011. Genome sequence of *Photobacterium mandapamensis* strain svers. 1.1, the bioluminescent symbiont of the cardinal fish *Siphamia versicolor*. J Bacteriol. 193:3144–3145.

Van Berkum NL, et al. 2010. Hi-C: a method to study the three-dimensional architecture of genomes. J. Vis. Exp. 39:e1869.

Wada M, et al. 2006. Lux A gene of light organ symbionts of the bioluminescent fish *Acropoma japonicum* (Acropomatidae) and *Siphamia versicolor* (Apogonidae) forms a lineage closely related to that of *Photobacterium leiognathi* ssp. Mandapamensis. FEMS Microbiol Lett. 260:186–192.

Wallin JE. 1927. Symbionticism and the origin of species. Baltimore: Wilhams & Wilkins.

Xu L, et al. 2019. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 47:W52–W58.

Yoshiba S, Haneda Y. 1967. Bacteriological study on the symbiotic luminous bacteria cultivated from the luminous organ of the apogonid fish, *Siphamia versicolor*, and the Australian pine cone fish, *Cleidopus gloriamaris*. Sci Rep. 13:82–84.

Associate editor: Bonnie Fraser