

Giants among Cnidaria: Large Nuclear Genomes and Rearranged Mitochondrial Genomes in Siphonophores

Namrata Ahuja ^{1,*†}, Xuwen Cao^{2,†}, Darrin T. Schultz^{3,†}, Natasha Picciani¹, Arianna Lord ⁴, Shengyuan Shao², Kejue Jia⁵, David R. Burdick⁶, Steven H.D. Haddock⁷, Yuanning Li², and Casey W. Dunn¹

¹Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

²Institute of Marine Science and Technology, Shandong University, Qingdao 266237, China

³Department of Neuroscience and Developmental Biology, University of Vienna, Vienna 1010, Austria

⁴Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA

⁵Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA

⁶University of Guam Marine Laboratory, Mangilao, GU

⁷Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: namrata.ahuja@yale.edu.

Accepted: March 07, 2024

Abstract

Siphonophores (Cnidaria: Hydrozoa) are abundant predators found throughout the ocean and are important constituents of the global zooplankton community. They range in length from a few centimeters to tens of meters. They are gelatinous, fragile, and difficult to collect, so many aspects of the biology of these roughly 200 species remain poorly understood. To survey siphonophore genome diversity, we performed Illumina sequencing of 32 species sampled broadly across the phylogeny. Sequencing depth was sufficient to estimate nuclear genome size from *k*-mer spectra in six specimens, ranging from 0.7 to 2.3 Gb, with heterozygosity estimates between 0.69% and 2.32%. Incremental *k*-mer counting indicates *k*-mer peaks can be absent with nearly 20× read coverage, suggesting minimum genome sizes range from 1.4 to 5.6 Gb in the 25 samples without peaks in the *k*-mer spectra. This work confirms most siphonophore nuclear genomes are large relative to the genomes of other cnidarians, but also identifies several with reduced size that are tractable targets for future siphonophore nuclear genome assembly projects. We also assembled complete mitochondrial genomes for 33 specimens from these new data, indicating a conserved gene order shared among nonsiphonophore hydrozoans, Cystonectae, and some Physonectae, revealing the ancestral mitochondrial gene order of siphonophores. Our results also suggest extensive rearrangement of mitochondrial genomes within other Physonectae and in Calyophorae. Though siphonophores comprise a small fraction of cnidarian species, this survey greatly expands our understanding of cnidarian genome diversity. This study further illustrates both the importance of deep phylogenetic sampling and the utility of *k*-mer-based genome skimming in understanding the genomic diversity of a clade.

Key words: *k*-mer spectra, genome skimming, genome size, mitochondrial genomes.

© The Author(s) 2024. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Significance

Descriptions of basic genome features, such as nuclear genome size and mitochondrial genome sequences, remain sparse across many clades in the tree of life, leading to over generalizations from very small sample sizes and often limiting selection of optimal species for genome assembly efforts. Here, we use whole-genome shotgun (WGS) skimming to assess a variety of genome features across 35 siphonophores (Cnidaria). This deep dive within a single clade identifies optimal candidates for future genomic work. It further reveals siphonophores have a greater range in nuclear genome size and diversity of mitochondrial genome orders than had been described across all Cnidaria.

Introduction

Siphonophores (Fig. 1) are among the longest (Robison 1995) and most abundant animals in the ocean. They occupy a critical position in the food web of the open ocean (Hetherington et al. 2022) and possess many unique biological features (Mapstone 2014; Munro et al. 2018). Like their coral relatives, siphonophores are colonial animals with many highly integrated, genetically identical bodies that all arise from a single embryo by asexual reproduction. Unlike corals, they are free swimming, and the bodies (zooids) within the colonies are specialized for different tasks, such as feeding, locomotion, and sexual production of new colonies. Studies on siphonophores have been limited, though, because they are fragile and difficult to collect. Improved genomic resources would greatly facilitate work on all aspects of their study and make it possible to learn much more from each of the specimens that are collected.

Molecular data available for siphonophores include a few genes sequenced across a large number of species (Dunn et al. 2005), transcriptomes for developmental and phylogenetic work (Siebert et al. 2011; Munro et al. 2022, 2018), and mitochondrial genomes for three species (Kayal et al. 2015). Chromosome-scale genomes have been assembled for a broad diversity of cnidarians (sea anemones, corals, and jellyfish) in recent years (Fig. 2), but most of these are for hexacorals and relatively few are from Hydrozoa, the clade that includes Siphonophora (Zapata et al. 2015). While no nuclear siphonophore genomes have been sequenced, two genome size estimates based on flow cytometry are available: *Agalma elegans* has been estimated to be 3.482 Gb and *Physalia physalis* at 3.247 Gb (Adachi et al. 2017). These are by far the largest genomes known to exist for cnidarians, similar in size to human genomes. However, with a sample size of 2 out of nearly 200 known species, sampling is insufficient to confidently describe nuclear genome size diversity within Siphonophorae.

In parallel to nuclear genomics, considerable progress has been made in understanding cnidarian mitochondrial genomics, but this sampling is not uniformly distributed. Currently, mitochondrial genome assemblies are available for over 300 anthozoans and 78 medusozoans; however, only 35 medusozoan mitochondrial genomes are complete

(Ling et al. 2023). Almost all sequenced anthozoan mitochondrial genomes are circular. Though medusozoan mitochondrial genomes are thought to be linear (Bridge et al. 1992; Kayal et al. 2015), several medusozoan mitochondrial genomes in public archives are annotated as being circular. However, there has not been enough data on siphonophore mitochondrial genomes to distinguish how they may compare to other Medusozoa, especially given their unresolved position within the clade and the subclade Hydroidolina. Currently, there are three nearly complete mitochondrial siphonophore genomes, including *P. physalis*, *Nanomia bijuga*, and *Rhizophysa eysenhardtii* (Kayal et al. 2015). These have identical gene orders to each other, consistent with few evolutionary changes in mitochondrial genome structure in Siphonophorae. However, the partial nature of these genome sequences and limited taxon sampling is too small to robustly test this.

Genome skimming by shotgun Illumina sequencing has emerged as a powerful tool for gaining insight about multiple genomic features (Heath-Heckman and Nishiguchi 2021; Hogan et al. 2022). Illumina reads are too short to assemble full nuclear genomes, but are sufficient to assemble mitochondrial genomes and to run a variety of analyses on nuclear genome properties based on *k*-mer spectra. With sufficient depth, Illumina sequencing provides estimates of nuclear genome size, heterozygosity, and repeat fraction. Here, we apply genome skimming to a broad diversity of 35 specimens from 32 siphonophore species across the siphonophore phylogeny, representing every major subclade. Our objectives are primarily biological: we seek to expand sampling to understand basic features of genome diversity across siphonophores and understand them relative to other cnidarians. Our objectives are also technical: we hope to identify the species that would be most amenable to chromosome-scale genome sequencing projects and evaluate the resources and approaches that would be needed for such studies at this time.

Results

Specimen Sampling and Sequencing

A total of 2,978 Gb of sequence data were collected across 35 specimens from 32 species (Table 1, Supplementary

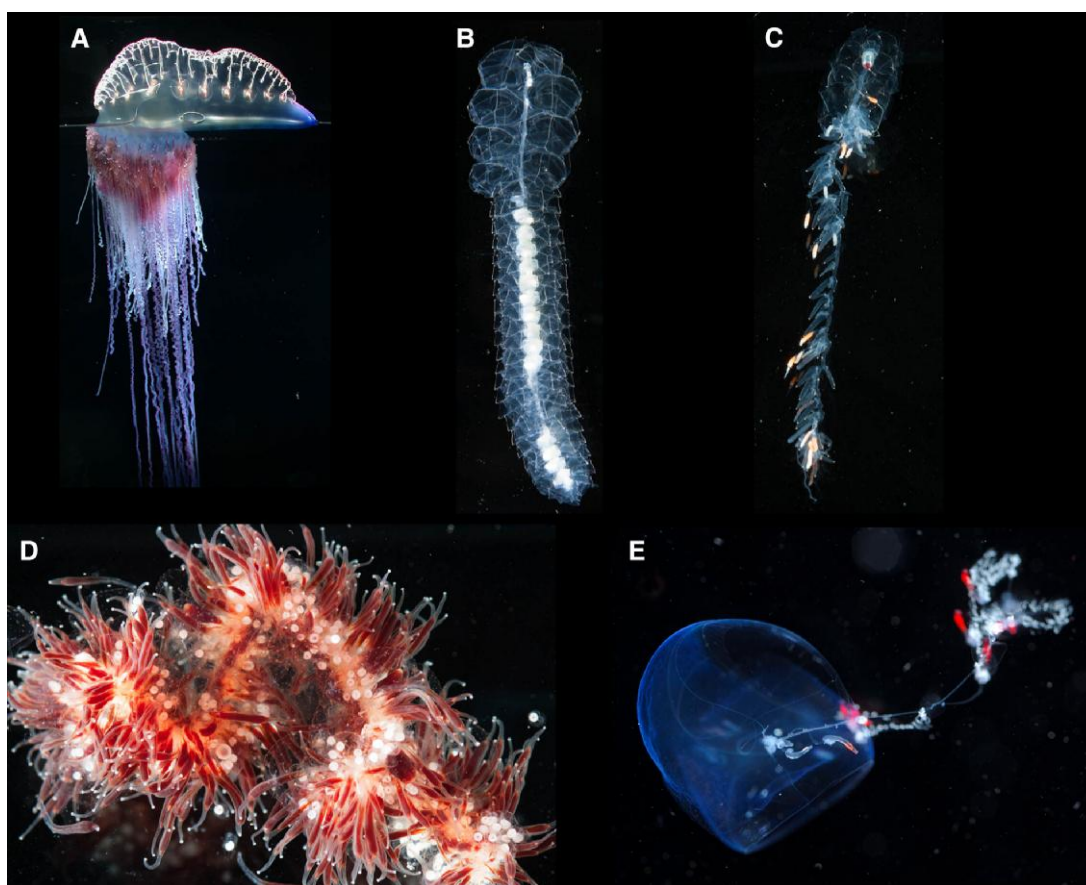


FIG. 1.—Photographs of some siphonophore species included in this study by CWD and SHD. Lengths are approximate. a) *Physalia* (float is about 20 mm across). b) *Frillagalma vityazi* (about 8 cm long). c) *Nanomia bijuga* (about 5 cm long). d) *Apolemia rubriversa* (specimen image about 7 cm across). e) *Gymnopraila lapislazula* (nectophore is about 2 cm across).

table S1). We took an iterative approach to sequencing, with a shallow first pass of most samples and then deeper sequencing in a progressively narrower subset of samples. The amount of data per sample therefore ranged widely, from 29 to 311 Gb (Table 1). We prioritized samples according to two criteria. First, we collected more data for a specimen if we observed peaks in the *k*-mer spectrum but needed more data for the GenomeScope2 genome size estimate models to fit well. Second, we prioritized several abundant species that are likely to be the focus of future genomic work, including *Physalia* and *A. elegans*, even though it was clear from preliminary *k*-mer skimming results that their genomes are quite large.

Trimomatic (Bolger et al. 2014) removed from 1.3% to 12.4% of nucleotides per data set resulting in very few nucleotides being trimmed for the majority of specimens sampled (supplementary table S2, fraction_trimmed, and fig. S1, Supplementary Material online). Then, Kraken (Wood and Salzberg 2014) identified from 0.1% to 3.8% of reads as bacterial (supplementary table S2 and fig. S1, Supplementary Material online) and from 1.0% to

36.33% of reads as human (supplementary table S2 and fig. S1, Supplementary Material online). Because Kraken has been shown to have very high false-positive rates (Garrido-Sanz et al. 2022), we further assessed the potential for human contamination by mapping reads to the human genome. When we mapped a million reads from each sample to the human genome, most samples had zero paired mapped reads (supplementary table S2, Supplementary Material online, human_mapped_count). Of the samples that were nonzero, the maximum number of reads that mapped to the human genome was 1,042 (*Stephanomia amphytridis*), well below what could impact results presented here. These results are consistent with the high false classification rate by Kraken. We assessed the effects of putative prokaryotic read contamination on the siphonophore genome size estimates by also estimating genome sizes with the subset of reads that were labeled “unclassified” by Kraken (i.e. not flagged as possible contaminants; supplementary fig. S1, Supplementary Material online). This exclusion of classified, putatively prokaryotic reads had very little impact on genome size

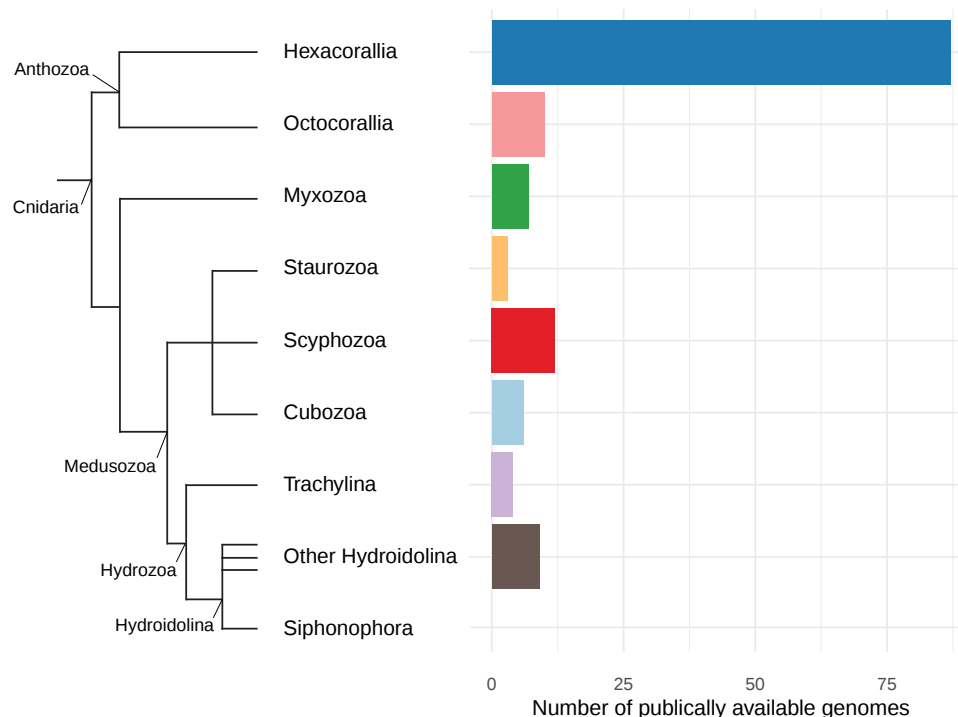


Fig. 2.—Number of nuclear genomes sequenced across Cnidaria, including both scaffold and chromosome-level assemblies. Siphonophores are within Hydroidolina, a subclade within Hydrozoa.

estimates, indicating that our results are robust to the amount of environmental or metagenomic bacterial contamination present in these siphonophore samples. Given the robustness of the genome size estimates, regardless of filtering from Kraken, our remaining analyses focused on the data sets with reads that were only quality trimmed with Trimmomatic.

k-mer Analysis

The *k*-mer results for 35 sequenced samples fall into 4 result categories (Table 1), those with the poorest sampling relative to genome size to the best sampling. First is the “no peak” category. These 25 samples had no peak in the *k*-mer spectra (supplementary table S2, Supplementary Material online, peak_location). Second is the “low peak” category. These two samples, *Agalma okeni* and *Resomia ornicephala* T845-D3, had peaks at very low counts; they were so low that they were out of range in GenomeScope plots, and these samples were not considered further in *k*-mer analyses. Third is the “poor fit” category. These two samples, *Gymnopræia lapislazula* and *R. ornicephala* T898-D2, had peaks that were in the range of GenomeScope plots (Fig. 3), but did not have consistently good model fit across all *k*-mer analyses (Fig. 3; supplementary fig. S2, Supplementary Material online). Fourth is the “good fit” category. These six samples (Fig. 3) resulted in well-defined *k*-mer spectra with good model fit and confident genome size estimates.

GenomeScope2 also provides estimates of heterozygosity and repeat percentage for these specimens (Table 1).

Incremental *k*-mer Counting

We wrote a new *k*-mer counter, sharkmer, that counts *k*-mers incrementally. Rather than providing a single snapshot of the *k*-mer spectrum when all data are considered, it provides a progressive understanding of a *k*-mer spectrum as sequencing depth is increased. We ran sharkmer on the six good fit and two poor fit samples (i.e. all samples with peaks in GenomeScope) to assess *k*-mer spectra with subsets of the data from 1% to 100% in 1% increments. To maintain a tractable memory footprint, we considered only the first billion reads if the data sets were larger than this, capping the largest data sets at 151 Gb (supplementary table S3, Supplementary Material online) since each read was 151 bp.

We first validated sharkmer performance by comparing *k*-mer counts from the final sharkmer sample to two widely used *k*-mer counters, Jellyfish and KMC (Deorowicz et al. 2013, 2015; Kokot et al. 2017). The *k*-mer counts were identical across the three tools (supplementary fig. S2, Supplementary Material online).

We used the incremental *k*-mer counting results to assess the stability of genome size estimation methods. Ideally, we would see a method converge on a genome size as data are added. However, if a method does not converge, results

Table 1

Overview of all samples

Sample	Result category	Haploid genome size estimate (Gb)	Haploid genome size minimum (Gb)	Read pair number (millions)	Total nucleotides (Gb)	Heterozygosity (%)	Repeat (%)
<i>Abylopsis tetragona</i>	No peak	...	1.56	106.342	31.30	...	...
<i>Agalma clausi</i>	No peak	...	1.53	103.939	30.67	...	...
<i>Agalma elegans</i> Atlantic	No peak	...	1.44	96.987	28.77	...	...
<i>Agalma elegans</i> Pacific	Good fit	2.32	...	1034.742	308.00	2.80	57.47
<i>Agalma okeni</i>	Low peak	...	...	255.282	72.97	...	...
<i>Apolemia rubriversa</i>	No peak	...	3.35	249.821	67.03	...	...
<i>Apolemia</i> sp3	No peak	...	2.01	145.572	40.16	...	...
<i>Bargmannia elongata</i>	No peak	...	4.33	299.470	86.61	...	...
<i>Bargmannia lata</i>	No peak	...	2.12	150.626	42.49	...	...
<i>Chelophyes appendiculata</i>	Good fit	1.20	...	353.189	105.00	5.16	56.91
<i>Chuniphyes multidentata</i>	No peak	...	3.11	221.280	62.27	...	...
<i>Cordagalma ordinatum</i>	Good fit	0.69	...	150.794	43.83	5.31	50.76
<i>Craseoa lathetica</i>	No peak	...	2.45	170.722	48.99	...	...
<i>Diphyes dispar</i>	No peak	...	2.90	198.254	57.98	...	...
<i>Erenna sirena</i>	No peak	...	4.60	331.332	92.00	...	...
<i>Forskalia</i>	No peak	...	3.71	267.534	74.23	...	...
<i>Frillagalma vityazi</i>	No peak	...	3.86	284.049	77.16	...	...
<i>Gymnoprora lapislazula</i>	Poor fit	2.33	...	878.095	256.00	6.11	83.98
<i>Halistemma rubrum</i> Atlantic	No peak	...	2.50	170.905	50.04	...	...
<i>Halistemma rubrum</i> Mediterranean	No peak	...	1.54	103.786	30.77	...	...
<i>Hippopodius hippopus</i>	No peak	...	3.65	266.036	73.09	...	...
<i>Lilyopsis fluoracantha</i>	No peak	...	4.02	282.743	80.43	...	...
<i>Nanomia bijuga</i> North Carolina	Good fit	0.71	...	312.832	92.50	4.04	44.55
<i>Nanomia</i> California	Good fit	1.45	...	290.026	86.47	0.67	63.30
<i>Physalia</i> Guam	Good fit	1.65	...	663.882	185.00	3.01	64.85
<i>Physonect</i>	No peak	...	3.30	229.185	65.97	...	...
<i>Praya</i>	No peak	...	3.34	235.932	66.86	...	...
<i>Praya dubia</i>	No peak	...	5.55	383.256	111.00	...	...
<i>Resomia ornicephala</i> T845-D3	Low peak	...	...	167.278	46.81	...	...
<i>Resomia ornicephala</i> T898-D2	Poor fit	1.01	...	566.160	157.00	3.59	60.48
<i>Rhizophysa eysenhardtii</i>	No peak	...	3.44	238.482	68.85	...	...
<i>Rhizophysa filiformis</i>	No peak	...	2.36	159.073	47.29	...	...
<i>Rosacea flaccida</i>	No peak	...	1.49	101.192	29.81	...	...
<i>Stephalia dilata</i>	No peak	...	3.14	219.048	62.75	...	...
<i>Stephanomia amphytridis</i>	No peak	...	2.40	169.178	48.09	...	...

Read pair number is the number of paired reads after trimming. Total length is the sum of the number of nucleotides in paired reads after trimming. Genome size estimates, heterozygosity, and repeat fraction are shown for the poor fit and good fit samples (i.e. samples that had peaks in GenomeScope). Given that the incremental analyses with as much as 16x coverage did not have a peak (supplementary fig. S1, Supplementary Material online), the genome size minimum for those samples without a peak is calculated conservatively as 1/20 the total of read length.

are sensitive to sampling effort. The parametric method implemented in GenomeScope2 does converge well for the six samples with good fit, but not the two with poor fit. This can be seen in plots of genome size and fit versus the percent of reads sampled (Fig. 4). It can also be seen in

animations of GenomeScope plots as sampling is increased (see Github repository). At low sampling, both genome sizes and model fits are low and unstable. This is expected given that much of this range is before there is sufficient coverage to produce a peak. Peaks arise at coverages ranging from

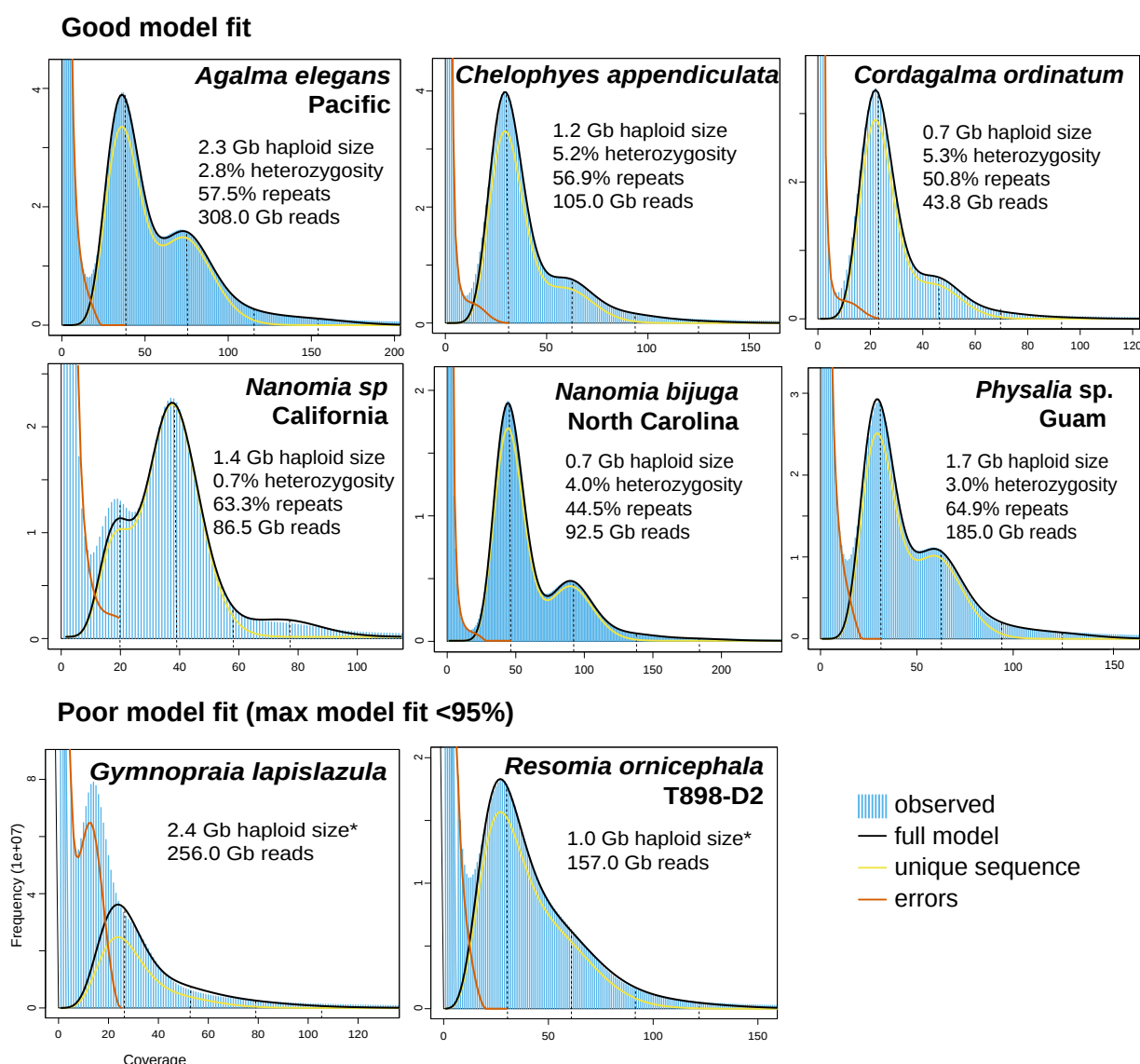


Fig. 3.—Plots of k -mer spectra for samples with peaks. Model fit is assessed by whether the full model curve (outer curved line) has peaks congruent with the underlying k -mer spectrum. The reported genome sizes for the specimens with poor model fit, indicated by asterisks, are from the Jellyfish counting.

8.33× to 15.65× coverage (supplementary table S4, Supplementary Material online). The point at which the maximum model fit reaches 95% corresponds well with the model arriving at the final plateau it converges on, though for most species, estimated genome size still tends to decline slightly as reads are added even on this plateau. This sampling depth that arrives at good model fit is about three times the coverage that was needed for the first peak to arise in the histogram (supplementary table S4, Supplementary Material online). Most of the sequencing effort needed to obtain a credible result is therefore after the first peak has arisen.

We found that a popular manual method of genome size estimation does not converge (supplementary fig. S3, Supplementary Material online; see Materials and

Methods for more information). Genome size estimates decrease as additional sequence data are added. This indicates that the manual method is not reliable, and parametric tools such as GenomeScope should be used.

We calculated minimum genome sizes for all specimens with no peak in the k -mer spectra (Table 1) by dividing the number of base pairs sequenced by 20. This was motivated by the observation that peaks arise in the k -mer spectra by 15.65 coverage in all samples with good model fit (supplementary table S3, Supplementary Material online). Rounding up slightly gives a conservative estimate. Genome sizes could exceed these minimal sizes by a factor of 3, given that we saw peaks with as little as 8.33× coverage (supplementary table S3, Supplementary Material online).

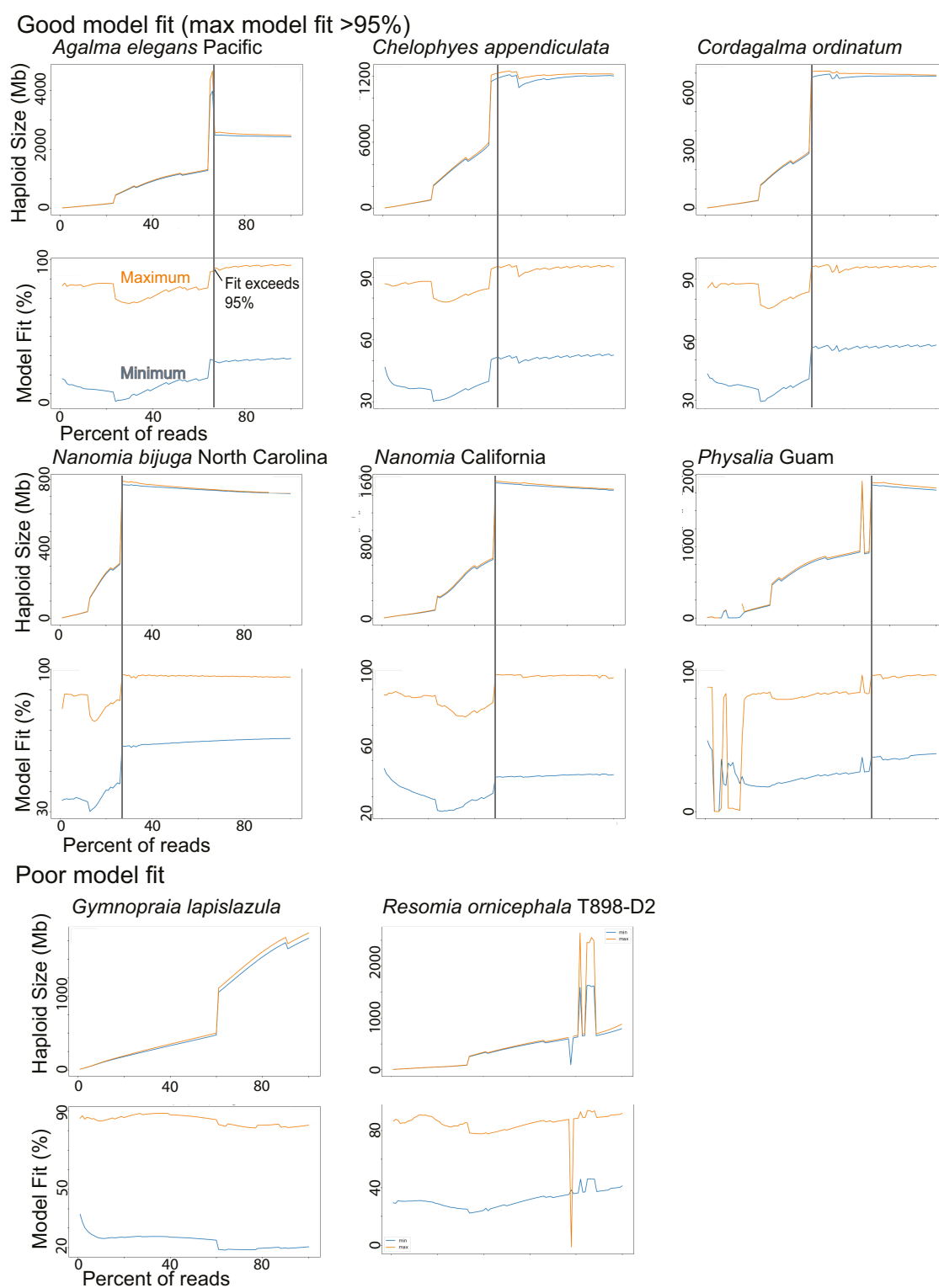


Fig. 4.—GenomeScope convergence for eight siphonophore species shown in Fig. 3. Vertical lines indicate where the models reached 95% model fit, in the six specimens with good model fit.

Mitochondrial Genome Structure and Diversity

Mitochondrial genomes were assembled for all specimens, except for *Hippopodius hippopus* and *Lilyopsis fluoracantha*, which were partially assembled (Fig. 5; [supplementary fig. S4](#), [Supplementary Material](#) online). The assembler indicated that all mitochondrial genomes were linear, except that of *Chelophyes appendiculata*, which was annotated as circular. Closer inspection with read mapping (see Materials and Methods) revealed that the mitochondrial genome of *C. appendiculata* was linear, indicating that the circular annotation by the assembler was incorrect ([supplementary fig. S5](#), [Supplementary Material](#) online). Nine of the siphonophore species' mitochondrial genomes were tested using this read mapping method ([supplementary fig. S5](#), [Supplementary Material](#) online), and they were all linear. These nine species cover the main lineages of siphonophores as resolved in [Munro et al. \(2018\)](#) and include two cystonects, three calyco-phorans, three members of Euphysonectae Clade A physonects, and one member of Euphysonectae Clade B (clade designations can be found in [Munro et al. 2018](#)). These results confirm that siphonophores, like other medusozoans, have linear mitochondrial genomes.

Phylogenetic analysis of the mitochondrial protein-coding genes had generally poor support for relationships within siphonophores and for the relationship of siphonophores to other hydrozoans ([supplementary fig. S6](#), [Supplementary Material](#) online). This is consistent with low support seen for phylogenetic analyses of mitochondrial sequences in other groups ([Kayal et al. 2015](#)). Some nodes of the mitochondrial phylogenetic tree were constrained (Fig. 5) to be consistent with major clades in previous phylogenies ([Munro et al. 2018](#)). These constraints did not violate any nodes with 100% support in the unconstrained tree ([supplementary fig. S6](#), [Supplementary Material](#) online), but given the poor support of the underlying tree and the use of constraints, this should not be considered an independent examination of the siphonophore phylogeny.

Mitochondrial gene order is conserved between multiple nonsiphonophore hydrozoans, Cystonectae, Apolemiidae (Fig. 5), *Bargmannia elongata*, and multiple members of Euphysonectae Clade B. We refer to this conserved gene order as Order 1, which was also found in the two previously published cystonect mitochondrial genomes of *P. physalis* and *Rhizophysa* ([Kayal et al. 2015](#)). *Physonect* sp. (undescribed) and both specimens of *R. ornicephala* showed relatively minor variations on Order 1, with *cox1* shifted from one end to the other and reversed. The gene order of *Nanomia*, *Halistemma*, and *Agalma* was also similar to Order 1, except that the location of *atp8* was changed. However, *Forskalia* sp., *Cordagalma ordinatum*, and all members of Calyco-phorae showed radically rearranged gene orders (Fig. 5). We mapped reads back to all the rearranged

mitochondrial genomes and found strong support for the assemblies. This is evident in both visual assessment of the mappings ([supplementary fig. S7](#), [Supplementary Material](#) online) and through computational confirmation that reads span all adjacent nucleotides in these genomes, apart from the very ends in a few species where there were several unspanned nucleotides. These few locations were all within 10 bp of the termini and had no impact on gene order.

The *atp8* gene has considerable sequence variation in some siphonophores and was initially missing from some mitochondrial genome annotations. We searched for it manually with a variety of methods in species where it was not annotated (see Materials and Methods) and were able to find it in the mitochondrial genomes of all species except *Chuniphyes multidentata*, *Diphyes dispar*, *Abylopsis tetragona*, and *C. appendiculata*. Despite extensive variation at the sequence level, the 3D structure of a single alpha helix is quite conserved.

Most siphonophore mitochondrial genomes we assembled have two tRNAs, one each for methionine (Met) and tryptophan (Trp). This is consistent with the tRNA inventory of many other cnidarians ([Kayal et al. 2015](#)). The exceptions were *D. dispar*, *A. tetragona*, and *C. appendiculata*, for which we identified no tRNAs. These three species form a clade, consistent with a shared loss of tRNA in this group. In addition, tRNA^{Trp} was not found in *Frillagalma vityazi*.

Discussion

Our results confirm that siphonophores have the largest known nuclear genomes within Cnidaria and reveal striking variation in genome size within the group. We were able to confidently estimate genome sizes for six samples that had good model fit ([Table 1](#) and Fig. 3). These specimens include some species with larger genomes that we sequenced more deeply and others that have relatively smaller genomes for siphonophores. All, though, have very large genomes relative to other cnidarians (Fig. 6). For the 25 specimens that had no peaks in the *k*-mer spectra at all, we estimated their minimum genome sizes to be between 1.4 and 5.6 Gb ([Table 1](#)). These minima were obtained by dividing the number of base pairs sequenced by 20, since our incremental *k*-mer counting analyses found a peak in all *k*-mer spectra of samples, which had good model fit by 15.65x coverage ([supplementary table S4](#), [Supplementary Material](#) online). These genome size minima are based on approximate methods and are quite conservative (in part because we rounded up to 20), meaning the genomes could be considerably larger.

The smallest observed genomes are scattered across the siphonophore phylogeny and nested within groups with much larger genomes. Thus, our findings suggest that the most recent common ancestor of siphonophores already had a very large genome relative to other cnidarians, and

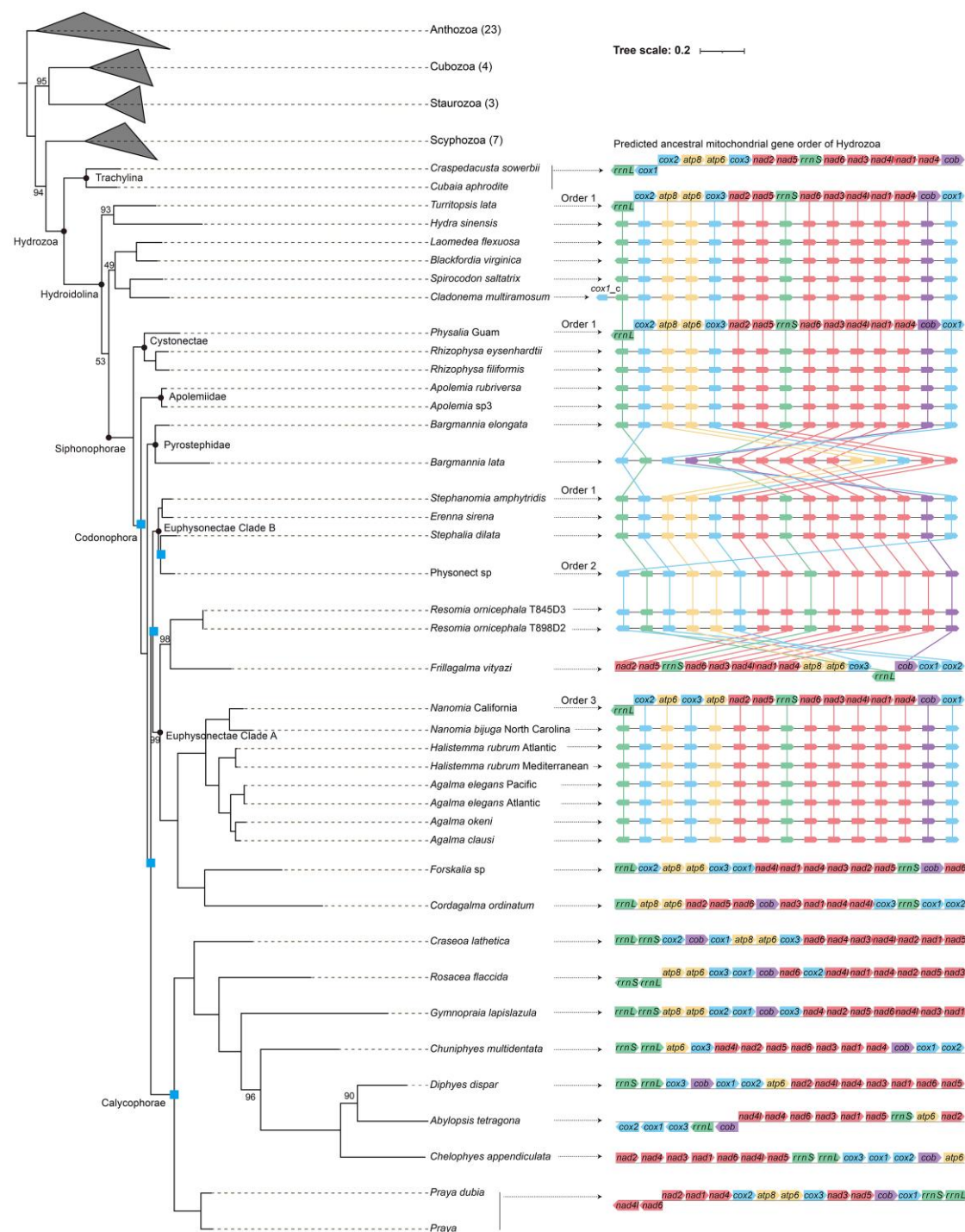


FIG. 5.—Diversity of siphonophore mitochondrial gene order. Squares indicate nodes constrained to be congruent with Munro et al. (2018). None of the constrained nodes had strong conflict between the mitochondrial and nuclear data; they were well supported in the nuclear data and had poor support in the mitochondrial data (see [supplementary fig. S6, Supplementary Material](#) online, for the unconstrained mitochondrial phylogeny). Bootstrap values less than 100 are indicated at the nodes; where no numerical value is indicated at unconstrained nodes, the bootstrap value is equal to 100. Mitochondrial rRNA and protein-coding gene arrangements of siphonophores are depicted on the right. *cox1_c* is an incomplete copy of *cox1* and lacks the 5' end of the gene.

that there have been multiple independent reductions in genome size within isolated lineages. The most striking reduction of genome size we identified was within *Nanomia*.

We included two specimens, one from California and one from North Carolina. The California specimen has a genome size of 1.5 Gb, within the range of many other

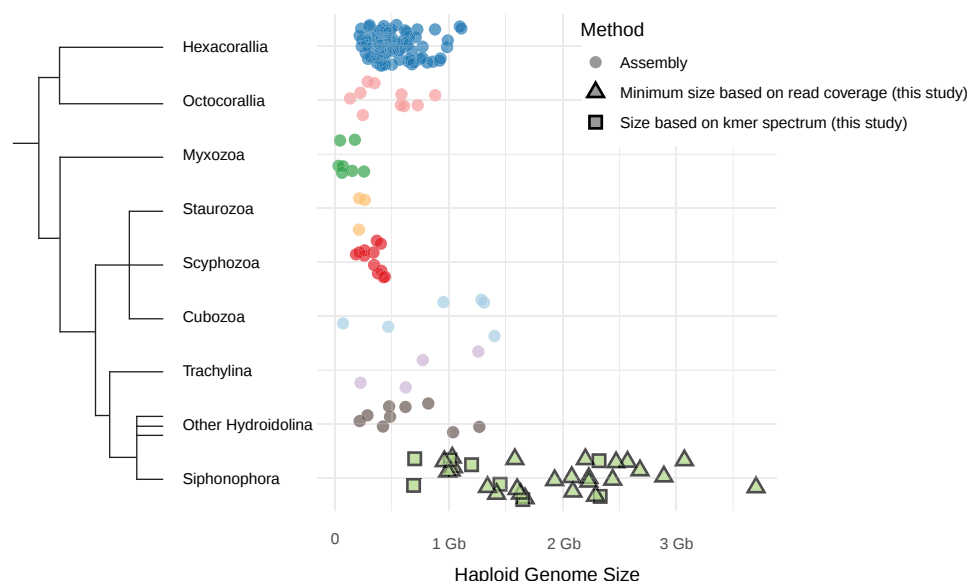


Fig. 6.—Cnidarian genome sizes and genome size estimates arranged by group. All Siphonophora represented here are newly sequenced for this project and are bolded.

siphonophores, while the North Carolina specimen has a genome of only 700 Mb, one of the two smallest siphonophore genomes we found (Fig. 3 and Table 1). Remarkably, both of these specimens come from populations that are commonly referred to as a single species, *N. bijuga*, though the branch length separating them is longer than for other conspecific specimens (Dunn et al. 2005). Haeckel (1888) described eight distinct species, which Bargmann and Totton (1965) later synonymized as all *N. bijuga*, indicating that there are morphological differences among *Nanomia* sp. populations. Taken together, this evidence strongly suggests these are different species, though a taxonomic revision is beyond the scope of this paper. In the meantime, we refer to the North Carolina specimen as *N. bijuga* and the California specimen as *Nanomia* California. These results show that genome skimming, even without genome assembly, can provide insight into questions about systematic biology. It is notable that the genomes differ in size by approximately 2-fold, and that *N. bijuga* North Carolina has a lower fraction of repeat sequences, 44.7%, than does *Nanomia* California, 63.4% (Table 1). This suggests that a possible reduction in the size of the *N. bijuga* North Carolina could be due to a reduction in repetitive regions. Full genome sequences for both of these taxa will be a fascinating opportunity to further understand radical changes in cnidarian genome size over relatively short evolutionary time scales.

Genome sizes were previously estimated with cell flow cytometry for two siphonophores collected in Japan, *A. elegans* at 3.482 Gb and *Physalia* at 3.247 Gb (Adachi et al. 2017). These sizes differ considerably from the estimates we obtained of 2.3 Gb for *A. elegans* (Pacific) and

1.7 Gb for *Physalia* Guam. The specimens were collected at different locations for the two studies and, similar to *Nanomia*, there may be more diversity than is commonly appreciated within these species, including variation in genome sizes across populations, which may even reflect different species. Alternatively, there could be technical variations between cell flow cytometry and *k*-mer analyses. Additional samples will be required to resolve this.

Siphonophore mitochondrial genome assemblies revealed remarkable conservation of gene order in early siphonophore evolution, followed by bursts of changes in gene order within Eucladophora (Damian-Serrano et al. 2021). Previous studies have identified 4 main mitochondrial gene orders in Medusozoa (Kayal et al. 2015), but we found more than 10 gene orders in siphonophores. Mitochondrial gene Order 1 (Fig. 5) spans the most recent common ancestor of siphonophores and is also present in outgroup taxa, suggesting that this is the ancestral mitochondrial gene order of siphonophores. The mitochondrial gene order within medusozoans is well conserved, and the predicted Medusozoa ancestral mitochondrial gene order is shared by most of the Scyphozoa, Staurozoa, Cubozoa (ignoring genome fragmentation), and Trachylina (a subclade of Hydrozoa). However, the order underwent some changes in the ancestor of Hydroidolina (the clade that includes siphonophores), such as loss of two nonstandard protein-coding genes *polB* and *ORF314* and inversion and translocation of *cox1* (Kayal et al. 2015). The ancestral mitochondrial gene order of siphonophores (Order 1) is consistent with the ancestral mitochondrial gene order of Hydroidolina.

There are several striking features of the diversity of mitochondrial gene order within Siphonophora. We found that

tRNA^{Met}(CAU) and tRNA^{Trp}(UCA) are lost in three species, *D. dispar*, *A. tetragona*, and *C. appendiculata*, that form a clade in Calyophorae. In addition, tRNA^{Trp}(UCA) could not be found in *F. vityazi*. There are documented cases of tRNA loss in other nonbilaterian species (Lavrov and Pett 2016), including in ctenophores (Pett et al. 2011; Pett and Lavrov 2015), some sponge lineages (Wang and Lavrov 2008), and some cnidarians (Kayal et al. 2012). Octocorals lacking tRNA^{Trp}(UCA) rarely use TGA codons in their mitochondrial coding sequences (Pont-Kingdon et al. 1998; Brugler and France 2008; Pett and Lavrov 2015). However, this phenomenon was not found in tRNA^{Trp}(UCA)-deficient *D. dispar*, *A. tetragona*, *C. appendiculata*, and *F. vityazi*, which, like *N. bijuga*, *C. multidentata*, *G. lapolisazula*, and *R. ornicephala* with tRNA^{Trp}(UCA), preferred to use TGA rather than TGG to encode Trp in mitochondrial coding sequences (supplementary table S5, Supplementary Material online). This indicates that there may be atypical tRNA^{Trp}(UCA) in the mitochondrial genomes of *D. dispar*, *A. tetragona*, *C. appendiculata*, and *F. vityazi* that we have not been able to identify. Alternatively, there may be unknown mechanisms for decoding mitochondrial TGA codon, such as in *Leishmania tarentolae*, and the tRNA^{Trp}(CCA) in nuclear genome undergoes a specific C-to-U nucleotide modification in the first position of the anticodon to allow the decoding of the mitochondrial TGA codons as Trp (Alfonzo et al. 1999), which may also be present in ctenophore *Mnemiopsis leidyi* (Pett et al. 2011).

The remarkable conservatism of mitochondrial gene order across Hydrozoa and many siphonophores contrasts sharply with the radical rearrangements seen within *Forskalia* sp., *C. ordinatum*, and all members of Calyophorae (Fig. 5). They are so rearranged that it is difficult to find any pattern, including any correlation with nuclear genome size because of the few species with good model fits. With the exception of *Praya* and *Craseoa lathetica*, specimens with highly reorganized mitochondrial genomes also have exceptionally long branches in the mitochondrial protein sequence phylogeny (supplementary fig. S6, Supplementary Material online). Rearrangements of the mitochondrial gene order in siphonophores therefore tend to be associated with an elevated rate of evolution of the gene sequences themselves (Fig. 5). Previous studies have shown that many members of Hydroidolina have duplicated *cox1* at each end of their mitochondrial chromosome(s) (Kayal et al. 2012, 2015). We found that the mitochondrial genomes of Cystonectae (including *Physalia* Guam, *R. eysenhardtii*, and *Rhizophysa filiformis*), *B. elongata*, *S. amphytridis*, *Erenna sirena*, and *Stephalia dilata* have a complete *cox1* at one end (downstream of *cob*) and a very short duplicated 3' end of *cox1* (20 to 30 amino acids) at the other end (downstream of *rrnL*). However, similar gene duplication events were not detected in other siphonophores, suggesting that duplicated genes at the ends of mitochondrial chromosomes are not universal in Hydroidolina. Another notable feature is that *Physonect* sp. and the two *R. ornicephala*

specimens share a unique gene order, but are not monophyletic, suggesting convergence.

The assembler (NOVOPlasty) indicated that our assembled mitochondrial genome of *C. appendiculata* was circular. This would be very surprising given the consistent linear mt genomes of Medusozoa (Bridge et al. 1992). Further inspection with read mapping indicated that this genome is linear (supplementary fig. S5, Supplementary Material online), and that the circular annotation was spurious. We note that some other medusozoan genomes deposited in GenBank are annotated as circular, such as *Nemopilema nomurai* (KY454767), *Rhopilema esculentum* (KY454768), *Catostylus townsendi* (OK299144), *Chrysaora pacifica* (MN448506), and *Haliclystus antarcticus* (KU947038), among others. We suspected that these were also spurious annotations. SRA data is available for *N. nomurai*, *R. esculentum*, and *C. townsendi* so we tested whether the reads for these species passed through the head–tail junction. They did not pass through the junctions (supplementary fig. S8, Supplementary Material online), corroborating our hypothesis that these mitochondrial genomes are actually linear. These results show that assembler software seems to be biased toward circular annotations, and these results should always be critically evaluated.

In addition to the biological insight gained with these genome skimming analyses, they have technical value for prioritizing future siphonophore genome projects. Though largely relative to other cnidarian genomes, siphonophore genomes for which we obtained confident genome size estimates can feasibly be sequenced and assembled to chromosome scale using standard modern approaches. Given their tractable genome sizes, phylogenetic distribution, and interesting biology, all six species with good model fit (Fig. 3) are promising priorities for future de novo siphonophore genome assembly projects.

Though the roughly 190 described siphonophores make up a tiny fraction of the more than 10,000 cnidarian species, this deep dive within this single clade greatly enriches our understanding of nuclear and mitochondrial genome diversity in Cnidaria. Other poorly known clades may have similar surprises to offer and may benefit from genome skimming to better understand their genome diversity in the context of evolution.

Materials and Methods

Summary of Genomes across Cnidaria

Cnidarian genome sizes were retrieved in November 2023 from <https://www.ncbi.nlm.nih.gov/data-hub/genome/>.

Sampling

Out of the 35 specimens examined in this study, 23 were collected in a previous study of siphonophore phylogenetics

(Dunn et al. 2005). The remaining 12 specimens are new. Data for each specimen can be found in [Supplementary table S1](#). Most were collected by deep-sea submersible off the coast of California by Monterey Bay Aquarium Research Institute from the research vessel *R/V Western Flyer*. We also collected a *Physalia* in Guam that was preserved in ethanol. Specimens collected by the *R/V Western Flyer* were frozen with liquid nitrogen, homogenized by mortar and pestle, and gDNA extracted with E.Z.N.A. Mollusc DNA Kit (Cat. #D3373-01). The *Physalia* specimen was ground up via mortar and pestle and then gDNA extracted using the QIAGEN DNeasy Blood & Tissue Kit (Cat. #69504). Samples were quantified via NanoDrop and Qubit.

Sequencing

Whole-genome shotgun (WGS) library preparation and sequencing was completed by the Yale Center for Genome Analysis (YCGA). The WGS libraries were prepared with the IDT xGen DNA Library Prep EZ Kit (Cat. #10009821). They were then sequenced on an Illumina NovaSeq6000 instrument using a S4 flow cell in a 2 × 150 bp paired-end run. The numbers of reads and bases sequenced for each specimen are indicated in [Table 1](#).

Genome *k*-mer Analysis

We developed a Snakemake pipeline to complete the following analysis (Köster and Rahmann 2012). Low-quality regions and adapters were trimmed with Trimmomatic v0.39 (Bolger et al. 2014) prior to all downstream *k*-mer analyses. Only reads that remained paired after trimming were retained. Potential contamination was evaluated with Kraken v2.1.3 (Wood and Salzberg 2014) and its standard database of 100 GB, which includes bacteria, archaea, virus, and human genomes along with a collection of known vectors. This partitioned the full set of reads into two read sets, classified reads with similarity to organisms in the Kraken database and therefore potential contamination and unclassified reads for which there is no identified potential contamination. Unless otherwise noted, analyses were conducted on the full set of reads. Given the high false-positive rate of Kraken2 for human sequences (Garrido-Sanz et al. 2022), we also mapped 1 million reads from each sample to the human reference genome GRCh38 with bwa v0.7.17 (Li and Durbin 2009) to more rigorously evaluate the possibility of human contamination.

We used Jellyfish v2.3.0 (Marçais and Kingsford 2011) to count *k*-mers (*k* = 21) and GenomeScope 2.0 (Vurture et al. 2017; Ranallo-Benavidez et al. 2020) to analyze *k*-mer spectra. We conducted Jellyfish + GenomeScope analyses on the full read set and on the unclassified read set (i.e. reads that were not classified as contaminants by

Kraken). In preliminary analyses, we also tested *k* = 17 on the full read set, but found little difference from *k* = 21 analyses and did not pursue these analyses further.

Manual Genome Size Estimation

In addition to the parametric models of GenomeScope2, we also applied a standard manual method for genome size estimation from *k*-mer spectra with one or more peaks. This generally follows the approach outlined at <https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/>. The histogram produced by Jellyfish has two columns, which we refer to as *X* and *Y*. *X* is the count of *k*-mer occurrences in the input sequences, and *Y* is the number of *k*-mers with that count. We identified two *X* values by manual inspection of each histogram. *X*_{min} is the *X* position of the first minimum. All counts to the left of *X*_{min} are considered to be sequencing errors and are excluded from further consideration. *X*_{peak} is the *X* position of the first peak. We then estimated the haploid genome length *L* as follows:

$$L = \frac{\sum_{n=X_{\min}}^{10,000} X_n Y_n}{2 X_{\text{peak}}}.$$

The 2 is in the denominator since the first peak is the heterozygous peak. The sum is through 10,000 since this is the maximum included in the Jellyfish histogram output.

Incremental *k*-mer Counting and Minimum Genome Sizes

To identify the read coverage (i.e. the total length of sequencing reads divided by the haploid genome size) required to observe a peak in the *k*-mer histogram and to assess the stability of the genome size estimates relative to depth of sequencing, we performed incremental *k*-mer analyses that assessed progressively larger subsets of the data for the eight samples that had poor fit or good fit. These data subsets ranged in size from 1% to 100% of the data in 1% increments.

We initially attempted to construct these incremental subsamples by subsetting reads and rerunning Jellyfish on each subset, but that was not tractable given the large computational and disk space needs. We therefore wrote *sharkmer* (<https://github.com/caseywdunn/sharkmer>), a new incremental *k*-mer counter implemented in rust. Commit aed4b426 was used for the analyses presented here. It counts the *k*-mers in *n* hashes and then combines these chunks to provide incremental *k*-mer counts. We set *n* = 100 so that we could obtain *k*-mer counts at 1% intervals. This strategy is faster than creating incrementally larger data sets and analyzing each with Jellyfish, but requires a large amount of RAM. We therefore capped the number of reads considered at 1 billion. This truncated the very large data sets of *A. elegans* Pacific, *Physalia* Guam,

and *R. ornicephala* T898-D2 (Supplementary table S3). We compare the results of sharkmer to those of KMC and Jellyfish, using the truncated data sets where the read number exceeded 1 billion, to validate this new tool.

The output was used to generate videos that show how the *k*-mer spectra change as data are added, as well as summary statistics that reveal how many reads are needed to estimate genome size and how stable those estimates are.

Mitochondrial Genome Assembly and Annotation

Most mitochondrial genomes were de novo assembled with NOVOPlasty v4.3.1 (Dierckxsens et al. 2017). The mitogenomes of *Bargmannia lata* and *F. vityazi* were assembled by GetOrganelle v1.7.6 (Jin et al. 2020). Partial mitogenomes of *H. hippopus* and *L. fluoracantha* were assembled by SPAdes v3.13.0 (Prjibelski et al. 2020). To determine whether the siphonophore mitochondrial genome is linear or circular, we split the genome sequences at *nad5* (*nad5* was chosen because it is the longest gene in the mitochondrial genome) and manually reconnected the head and tail of the original sequence, with the head of the new sequence being the second half of *nad5* and the tail being the front half of *nad5*. Sequencing reads were then mapped onto these spliced sequences. If reads spanned the head–tail junction, it indicated that the sequence is circular; otherwise, it is linear (supplementary fig. S5, Supplementary Material online). The software used for read mapping were Bowtie 2 v2.4.1 (parameters and other information are in supplementary tables S6 to S8, Supplementary Material online; Langmead and Salzberg 2012) and SAMtools v1.17 (Danecek et al. 2021) with Tablet v1.21.02.08 (Milne et al. 2013) visualization.

Annotation of the mitochondrial genomes was conducted with MITOS2 web server (Donath et al. 2019) with the Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code (NCBI translation table 4), using default settings and canceling the circular setting. Then, genome annotation of start and stop positions of each gene was manually adjusted using Artemis (Rutherford et al. 2000). We verified the presence and absence of tRNA in siphonophores using the tRNA annotation tools MITOS2 and tRNAscan-SE.

We also used read mapping to verify the assembly of mitochondrial genomes that showed evidence of rearrangement; 10M to 50M reads (supplementary table S7, Supplementary Material online) were mapped to each of these genomes using Bowtie 2 v2.4.1 in end-to-end mode. These were assessed in two ways. First, by visual inspection (supplementary fig. S7, Supplementary Material online). Second, we wrote a python script (analyses_mitochondria/count_spanning_reads.py in the git repository) that parses cigar strings from each read mapping and counts the number of mappings that span from one

location to the adjacent location to its right. A value of zero indicates that there are no mapped reads that connect a position in the assembly to the position to its immediate right.

ATP8 had low sequence similarity to known sequences in some species, which required a manual search process. This involved using a combination of methods to find *atp8*, including BLAST version 2.2.26 and BLAST+ 2.13.0 (Altschul et al. 1990, 1997; Camacho et al. 2009) and a deep learning–based homolog detection tool, PROST (Rives et al. 2021; Kilinc et al. 2023), and looking for *atp8* start and stop sites within open reading frames. Each potential *atp8* sequence was folded using AlphaFold2 and ColabFold v1.5.3 (Jumper et al. 2021; Mirdita et al. 2022; Varadi et al. 2022) to see if they matched the single alpha helix shape of other annotated *atp8* genes of siphonophores. Additionally, we noted a pattern of *atp8* genes being neighbors with *atp6*, *cox3*, and *nad2*, so we looked for open reading frames in the gaps between these genes. Mitochondrial genomes were updated to reflect newly found *atp8* genes as seen in Fig. 5.

Phylogenetic Analysis

The amino acid sequences of 13 mitochondrial protein encoding genes for siphonophores (33 specimens of the complete mitochondrial genome were assembled), 22 other medusozoans, and 23 anthozoans were used for phylogenetic analysis (see supplementary table S5, Supplementary Material online, for accession numbers of previously published genomes included in these analyses). Each gene was individually aligned in MAFFT v7.310 (Katoh and Standley 2013) and trimmed using Trimal v1.4.1 (Capella-Gutiérrez et al. 2009) with default parameters to discard ambiguously aligned sites. Gene alignments were then concatenated into a final supermatrix. Unconstrained maximum likelihood (ML) phylogenetic analysis was then conducted using IQ-Tree v2.0.3 (Minh et al. 2020) with the option -m MFP and evaluated with 1,000 fast-bootstrapping replicates (supplementary fig. S6, Supplementary Material online). Given the broader sequence sampling of the transcriptome phylogeny, we ran constrained inferences after clamping the five nodes (squares in Fig. 5) to be consistent with the topology of the tree in Munro et al. (2018). Given these constraints, the present paper should not be considered an independent analysis of phylogenetic relationships within Siphonophora.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

Thanks to Lourdes Rojas and Eric Lazo-Wasem at the Yale Peabody Museum of Natural History. Thank you to the

NIH T32 Training Grant in Genetics at Yale. All library preparation and sequencing were conducted at the Yale Center for Genomic Analysis. All analyses were performed on computer clusters operated by the Yale Center for Research Computing.

Author Contributions

N.A.: new sample collection, new DNA extractions, coordination with sequencing center, implemented *k*-mer analysis workflow, and manuscript writing. D.T.S.: project design, developed *k*-mer analysis workflow, and interpretation of *k*-mer spectra. C.W.D.: project design, data analysis, implementation of sharkmer, data submission, and manuscript writing. N.P.: implemented *k*-mer analysis workflow and manuscript edits. A.L.: summary of existing cnidarian genomes. X.C.: assembly, annotation, and analysis of mitochondrial genomes and manuscript writing. S.H.D.H.: new sample collection and project design. Y.L.: assembly, annotation, and analysis of mitochondrial genomes and manuscript writing. K.J.: *atp8* annotation in siphonophore mitochondrial genomes. D.R.B.: collected *Physalia* specimens. SS: assistance with mitochondrial genome assembly.

Funding

Support for this project came from multiple sources including the National Science Foundation grants NSF-OCE to C.W.D. and NSF-OCE 1829805 to S.H.D.H., the Tal Waterman Fund at Yale, the Shandong Provincial Natural Science Foundation (grant number: ZR2023QC268), and the Qingdao Postdoctoral Applied Research Project (grant number: QDBSH20220202163). D.T.S. was supported by European Research Council (ERC) under the European Union's Horizon 2020 Programme (grant number 945026). S.H.D.H. was supported by the David and Lucile Packard Foundation. D.R.B. was supported by NSF Established Program to Stimulate Competitive Research (EPSCoR) grant (grant number OIA-1946352).

Data Availability

Code and other analysis components are available at https://github.com/dunnlab/siph_skimming. Sequence reads have been deposited in NCBI SRA as BioProject PRJNA925656. Annotated mitochondrial genomes have been deposited at NCBI with accession numbers OQ957189 to OQ957220.

Literature Cited

- Adachi K, Miyake H, Kuramochi T, Mizusawa K, Okumura S-I. Genome size distribution in phylum Cnidaria. *Fish Sci.* 2017;83(1):107–112. <https://doi.org/10.1007/s12562-016-1050-4>.
- Alfonzo JD, Blanc V, Estevez AM, Rubio MA, Simpson L. C to U editing of the anticodon of imported mitochondrial tRNA^{Trp} allows decoding of the UGA stop codon in *Leishmania tarentolae*[J]. *EMBO J.* 1999;18(24):7056–7062. <https://doi.org/10.1093/emboj/18.24.7056>.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389–3402. <https://doi.org/10.1093/nar/25.17.3389>.
- Bargmann HE, Totton AK. A synopsis of the Siphonophora. By A.K. Totton assisted by H.E. Bargmann, etc. London: Trustees of the British Museum (Natural History); 1965.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30(15):2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>.
- Bridge D, Cunningham CW, Schierwater B, DeSalle R, Buss LW. Class-level relationships in the phylum Cnidaria: evidence from mitochondrial genome structure. *Proc Natl Acad Sci U S A.* 1992;89(18):8750–8753. <https://doi.org/10.1073/pnas.89.18.8750>.
- Brugler MR, France SC. The mitochondrial genome of a deep-sea bamboo coral (Cnidaria, Anthozoa, Octocorallia, Isididae): genome structure and putative origins of replication are not conserved among octocorals. *J Mol Evol.* 2008;67(2):125–136. <https://doi.org/10.1007/s00239-008-9116-2>.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10(1):421. <https://doi.org/10.1186/1471-2105-10-421>.
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25(15):1972–1973. <https://doi.org/10.1093/bioinformatics/btp348>.
- Damian-Serrano A, Haddock SHD, Dunn CW. The evolution of siphonophore tentilla for specialized prey capture in the open ocean. *Proc Natl Acad Sci U S A.* 2021;118(8):e2005063118. <https://doi.org/10.1073/pnas.2005063118>.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. *GigaScience.* 2021;10(2). <https://doi.org/10.1093/gigascience/giab008>.
- Deorowicz S, Debudaj-Grabysz A, Grabowski S. Disk-based k-mer counting on a PC. *BMC Bioinformatics* 2013;14(1):160. <https://doi.org/10.1186/1471-2105-14-160>.
- Deorowicz S, Kokot M, Grabowski S, Debudaj-Grabysz A. KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics* 2015;31(10):1569–1576. <https://doi.org/10.1093/bioinformatics/btv022>.
- Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 2017;45(4):e18. <https://doi.org/10.1093/nar/gkw955>.
- Donath A, Jühling F, Al-Arab M, Bernhart SH, Reinhardt F, Stadler PF, Middendorf M, Bernt M. Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic Acids Res.* 2019;47(20):10543–10552. <https://doi.org/10.1093/nar/gkz833>.
- Dunn CW, Pugh PR, Haddock SHD. Molecular phylogenetics of the siphonophora (Cnidaria), with implications for the evolution of functional specialization. *Syst Biol.* 2005;54(6):916–935. <https://doi.org/10.1080/10635150500354837>.
- Garrido-Sanz L, Ángel Senar M, Piñol J. Drastic reduction of false positive species in samples of insects by intersecting the default output of two popular metagenomic classifiers. *PLoS One* 2022;17(10):e0275790. <https://doi.org/10.1371/journal.pone.0275790>.
- Haeckel E. Report on the Siphonophorae collected by H.M.S. Challenger during the years 1873-1876. UK: H.M. Stationery Office; 1888.

- Heath-Heckman E, Nishiguchi MK. Leveraging short-read sequencing to explore the genomics of sepiolid squid. *Integr Comp Biol*. 2021;61(5):1753–1761. <https://doi.org/10.1093/icb/ibab152>.
- Hetherington ED, Damian-Serrano A, Haddock SHD, Dunn CW, Choy CA. Integrating siphonophores into marine food-web ecology. *Limnol Oceanogr Lett*. 2022;7(2):81–95. <https://doi.org/10.1002/lol2.10235>.
- Hogan RI, Hopkins K, Wheeler AJ, Yesson C, Allcock AL. Evolution of mitochondrial and nuclear genomes in Pennatulacea. *Mol Phylogenet Evol*. 2022;178:107630. <https://doi.org/10.1016/j.ympev.2022.107630>.
- Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, Li D-Z. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol*. 2020;21(1):241. <https://doi.org/10.1186/s13059-020-02154-5>.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021; 596(7873):583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772–780. <https://doi.org/10.1093/molbev/mst010>.
- Kayal E, Bentlage B, Cartwright P, Yanagihara AA, Lindsay DJ, Hopcroft RR, Collins AG. Phylogenetic analysis of higher-level relationships within Hydrozoa (Cnidaria: Hydrozoa) using mitochondrial genome data and insight into their mitochondrial transcription. *PeerJ* 2015;3:e1403. <https://doi.org/10.7717/peerj.1403>.
- Kayal E, Bentlage B, Collins AG, Kayal M, Pirro S, Lavrov DV. Evolution of linear mitochondrial genomes in medusozoan cnidarians. *Genome Biology and Evolution*. 2012;4(1):1–12. <https://doi.org/10.1093/gbe/evr123>.
- Kilinc M, Jia K, Jernigan RL. Improved global protein homolog detection with major gains in function identification. *Proc Natl Acad Sci U S A*. 2023;120(9):e2211823120. <https://doi.org/10.1073/pnas.2211823120>.
- Kokot M, Długosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 2017;33(17):2759–2761. <https://doi.org/10.1093/bioinformatics/btx304>.
- Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012;28(19):2520–2522. <https://doi.org/10.1093/bioinformatics/bts480>.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012;9(4):357–359. <https://doi.org/10.1038/nmeth.1923>.
- Lavrov DV, Pett W. Animal mitochondrial DNA as we do not know it: mt-genome organization and evolution in nonbilaterian lineages. *Genome Biol Evol*. 2016;8(9):2896–2913. <https://doi.org/10.1093/gbe/evw195>.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>.
- Ling MK, Yap NWL, Iesa IB, Yip ZT, Huang D, Quek ZBR. Revisiting mitogenome evolution in Medusozoa with eight new mitochondrial genomes. *iScience* 2023;26(11):108252. <https://doi.org/10.1016/j.isci.2023.108252>.
- Mapstone GM. Global diversity and review of Siphonophorae (Cnidaria: Hydrozoa). *PLoS One* 2014;9(2):e87737. <https://doi.org/10.1371/journal.pone.0087737>.
- Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;27(6):764–770. <https://doi.org/10.1093/bioinformatics/btr011>.
- Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*. 2013;14(2):193–202. <https://doi.org/10.1093/bib/bbs012>.
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;37(5):1530–1534. <https://doi.org/10.1093/molbev/msaa015>.
- Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods*. 2022;19(6):679–682. <https://doi.org/10.1038/s41592-022-01488-1>.
- Munro C, Siebert S, Zapata F, Howison M, Damian-Serrano A, Church SH, Goetz FE, Pugh PR, Haddock SHD, Dunn CW. Improved phylogenetic resolution within Siphonophora (Cnidaria) with implications for trait evolution. *Mol Phylogenet Evol*. 2018;127:823–833. <https://doi.org/10.1016/j.ympev.2018.06.030>.
- Munro C, Zapata F, Howison M, Siebert S, Dunn CW. Evolution of gene expression across species and specialized zooids in Siphonophora. *Mol Biol Evol*. 2022;39(2):msac027. <https://doi.org/10.1093/molbev/msac027>.
- Pett W, Lavrov DV. Cytonuclear interactions in the evolution of animal mitochondrial tRNA metabolism. *Genome Biol Evol*. 2015;7(8):2089–2101. <https://doi.org/10.1093/gbe/evw124>.
- Pett W, Ryan JF, Pang K, Mullikin JC, Martindale MQ, Baxevanis AD, Lavrov DV. Extreme mitochondrial evolution in the ctenophore *Mnemiopsis leidyi*: insight from mtDNA and the nuclear genome. *Mitochondrial DNA* 2011;22(4):130–142. <https://doi.org/10.3109/19401736.2011.624611>.
- Pont-Kingdon G, Okada NA, Macfarlane JL, Beagley CT, Watkins-Sims CD, Cavalier-Smith T, Clark-Walker GD, Wolstenholme DR. Mitochondrial DNA of the coral *Sarcophyton glaucum* contains a gene for a homologue of bacterial MutS: a possible case of gene transfer from the nucleus to the mitochondrion. *J Mol Evol*. 1998;46(4):419–431. <https://doi.org/10.1007/PL00006321>.
- Prijbelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de novo assembler. *Current Protocols in Bioinformatics*. 2020;70(1). <https://doi.org/10.1002/cpbi.102>.
- Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020;11(1):1432. <https://doi.org/10.1038/s41467-020-14998-3>.
- Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A*. 2021;118(15):e2016239118. <https://doi.org/10.1073/pnas.2016239118>.
- Robison BH. Light in the ocean's midwaters. *Scientific American*. 1995;273:60–64. <http://www.jstor.org/stable/24981452>.
- Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, Barrell B. Artemis: sequence visualization and annotation. *Bioinformatics* 2000;16(10):944–945. <https://doi.org/10.1093/bioinformatics/16.10.944>.
- Siebert S, Robinson MD, Tintori SC, Goetz F, Helm RR, Smith SA, Shaner N, Haddock SHD, Dunn CW. Differential gene expression in the siphonophore *Nanomia bijuga* (Cnidaria) assessed with multiple next-generation sequencing workflows. *PLoS One* 2011;6(7):e22953. <https://doi.org/10.1371/journal.pone.0022953>.
- Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*. 2022;50(D1):D439–D444. <https://doi.org/10.1093/nar/gkab1061>.
- Vurtture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 2017;33(14):2202–2204. <https://doi.org/10.1093/bioinformatics/btx153>.

Wang X, Lavrov DV. Seventeen new complete mtDNA sequences reveal extensive mitochondrial genome evolution within the Demospongiae. *PLoS One* 2008;3(7):e2723. <https://doi.org/10.1371/journal.pone.0002723>.

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 2014;15(3):R46. <https://doi.org/10.1186/gb-2014-15-3-r46>.

Zapata F, Goetz FE, Smith SA, Howison M, Siebert S, Church SH, Sanders SM, Ames CL, McFadden CS, France SC, et al. Phylogenomic analyses support traditional relationships within Cnidaria. *PLoS One* 2015;10(10):e0139068. <https://doi.org/10.1371/journal.pone.0139068>.

Associate editor: Dennis Lavrov