Supplementary Materials: Materials and Methods

Materials and Methods

Sample collection and DNA extraction

Physalia specimens were collected by a global collaboration of scientists. Full sampling details and required permit information can be found in the supplementary text: specimen collection information. The majority of specimens were collected after they washed ashore, using appropriate safety protocols to avoid stings. A few specimens were collected directly from the water, either sampling from a boat or while diving (for juvenile specimens that hadn't yet surfaced). Specimens were preserved in >70% alcohol (ethanol, when available), and stored at room temperature, with the exception of two samples collected from Hawai'i that were stored in DNAshield, and several samples from the Eastern United States that were flash-frozen.

When possible, whole specimens were collected and shipped to the Yale Peabody Museum. All sequenced specimens were photographed, and images are shown in the supplementary text. Additional specimens were loaned from the Western Australia Museum, the Tasmanian Museum and Art Gallery, and the Field Museum in Chicago.

High molecular weight DNA for genome sequencing and assembly was extracted from a flash-frozen specimen, Yale Peabody Museum (YPM) No. 110876, collected in Texas, USA in 2017. Extractions were performed following the protocol described by Chen and Dellaporta, 1994 (*1*), with modifications. In this protocol, tissue was homogenized under liquid nitrogen and extracted with 5 mL of a urea-based extraction buffer for 15 minutes at 65 degrees C. Three 25:24:1 phenol:chloroform:isoamyl-alcohol (P:C:I) extractions were performed, each allowed to rock for 5 minutes before centrifugation. The P:C:I extractions were followed up with extraction with one volume of chloroform prior to precipitation with isopropanol. Extracted DNA and RNA was resuspended in 100 l Tris-EDTA buffer, analyzed by gel electrophoresis, then brought to a volume of 400 L and subjected to RNase treatment with 3 L RNAse L and 2 L of RNAse cocktail for 60 minutes at 37 degrees C. The RNA-free DNA was then brought to 500 l with 5M NaCl and extracted with 400 l P:C:I. The aqueous phase was removed and

400 L of 5M NaCl, 500 mM EDTA, 10mM Tris was added to the P:C:I for back extraction. The back-extracted aqueous phase was combined with the first aqueous phase and 0.3 volumes of 100% ethanol was added to precipitate polysaccharides, pelleted at 17,000 x G. DNA was precipitated with 1.7 volumes of 100% ethanol, pelleted, and washed with 70% ethanol and resuspended in Tris-EDTA. The purified DNA was examined by pulse-field gel electrophoresis and showed a strong band at >98kb.

RNA for transcriptome sequencing was extracted from a flash-frozen specimen, YPM. No. 110436, collected in Florida, USA in 2023. Tissue was homogenized using a mortar and pestle chilled to -80 degrees C, and RNA was extracted using the RNAqueous Total RNA Isolation Kit following manufacturers instructions and including a lithium-chloride precipitation step. RNA was processed for library preparation and PacBio Iso-Seq sequencing by the Keck Microarray Shared Resource at Yale University. Delivered reads were clustered with the PacBio `isoseq cluster2` command, version 1.0.1. These were then further deduplicated with treeinform (*2*) as implemented in the code available at https://github.com/dunnlab/isoseq, commit `f0b69a8`. This tool implements a phylogenetically informed refinement of the transcriptome to remove species-specific variants, by building gene trees from the target transcriptome (here *Physalia*) and gene predictions for related species (here 23 species, including 10 cnidarians). Clades with short total branch length and that contain only sequences from the target species are collapsed to the longest sequence.

For samples intended for population genomic analysis, tentacle pieces were dissected from whole specimens and stored in 95% ethanol prior to DNA extraction for genome sequencing. DNA extractions were performed using the EZNA Mollusk kit following manufacturers instructions and an overnight digestion, with the exception of several samples from Japan, Guam, and Texas that were extracted using the urea-based phenol-chloroform protocol described above, as well as one sample from the Gulf of California, extracted at Monterey Bay Aquarium Research Institute with the DNeasy DNA Blood and Tissue kit (Qiagen), following manufacturer's instructions. Whole genome DNA was processed for library preparation and sequencing by the Yale Center for Genome Analysis.

Genome assembly

Eight Single-Molecule Real-Time (SMRT) sequencing cells of PacBio HiFi data were assembled with `canu`, v. 2.2 (`-pacbio-hifi` option) (*3–5*), with the estimated genome size parameter set to three gigabases. HiFi reads were mapped to this assembly with `minimap2`, v. 2.22-r1101 (*6*), to determine the appropriate cutoffs for purging duplicated contigs. These were removed using `purge_haplotigs`, v. 1.1.2 (low, medium, and high cutoffs set at 5x, 40x, and 200x respectively) (*7*), and overlapping contig ends were clipped with the same program. The parameters for `purge_haplotigs` were modified to avoid memory limitation (`-I` was set to 1G, `-p` was dropped, and `-N` was set to 1000).

A foreign contamination screen (FCS, via the National Center for Biotechnology Information, NCBI) was performed on both the purged and haplotype assemblies, using the tool provided for GenBank submissions which detected and removed one adapter sequence. We used the tool `LongStitch`, v. v1.0.4 (*8*), to scaffold both the purged (primary) and haplotype (alternate) assemblies. Scaffolding was performed first using the eight HiFi cells used for assembly and the `ntLink-arks` functionality, and then using a dataset of 225 gigabases of linked-read data sequenced with 10XGenomics Chromium sequencing, interleaved with `LongRanger` (provided by 10X Genomics). The FCS was repeated on this assembly and detected no further foreign contaminants.

Repeat regions were detected and masked with `RepeatModeler` and `RepeatMasker`, v. 4.1.5 (*9*), to build a general feature format (gff) file, used to exclude repeats from downstream analyses. `BUSCO`, v 5.4.4 (*10*), and `BBMap stats.sh` were used to evaluate final assemblies. Genome assembly is made publicly available at NCBI, BioProject number PRJNA1040906.

Genome mapping

Paired-end genome sequencing targeting a read length of 150 base pairs was performed for 145 libraries using an Illumina NovaSeq at the Yale Center for Genome Analysis. Full details on quality control, mapping statistics, and final library parameters are available in the GitHub document https://github.com/shchurch/Physalia_population_genomics/manuscript_files/quality_control.html. Briefly, sequencing depth range varied across samples from a target of 10-60x genome size. These 145 samples included two replicate libraries, generated from repeated DNA extractions from the same specimens. In addition, from sequenced libraries we generated two technical replicates by randomly splitting read files. These replicates were used to evaluate reproducibility and were excluded from the main analyses presented in this work.

Overall sequence quality (e.g. GC content, adapter content) was evaluated using `FastQC`, v. 0.11.9. Reads were trimmed for Illumina adapters using `Trimmomatic`, v. 0.39 (*11*). Potential human, bacterial, and viral DNA contamination was evaluated using `Kraken2`, v. 2.1.2 (*12*), standard database. Additional cross-species contamination was evaluated using *in silico* PCR of the ribosomal 18S gene from genomic reads, and comparing results to publicly available datasets with a basic local alignment search tool, `BLAST`. Potential kinship or cross-contamination between *Physalia* samples was evaluated by calculating the kinship-based inference for genomes (KING-robust) relatedness score on reads mapped to the assembled genome using `PLINK2`, v. 2.00a5LM (*13*), calculated only using SNPs within Hardy-Weinberg equilibrium (p-value <1e-7), and excluding those with missing alleles >0.1 or a minor allele frequency >0.01).

Based on the results of the quality control analyses, six samples were identified as contaminated and an additional four samples were identified as replicated sampling events from a single specimen (e.g. multiple tentacle tips taken from the same animal in the field). These samples

were excluded from downstream analyses, such that the final dataset, excluding technical and biological replicates, consisted of 133 samples. Of those, 123 were marked as high quality based on overall sequencing depth, read quality, and proportion of missing sites. Analyses were performed on a strict dataset of only high-quality samples, and repeated on the full dataset of high- and moderate-quality samples.

Reads were mapped to the reference genome using `BWA`, v. 0.7.17-r1188 (*14*). Mapped reads were sorted, deduplicated, and indexed using `picard`, v. 2.25.6. Alleles were called using `BCFtools`, v. 1.16, `mpileup` (*15*). To test the robustness of downstream analyses to reference assembly, reads were mapped to the independent transcriptome assembly, using only R1 reads as single-end data.

Phylogenetics

Mitochondrial genomes were assembled from a subset of ten million trimmed reads for each sample, using the software `GetOrganelle`, v. 1.7.7.0 (*16*), using the `animal_mt` database and default parameters. `GetOrganelle` failed to circularize the assemblies, in line with the expected linear mitochondrial genomes in siphonophores (*17*); the resulting top path assembly was used as the final linear genome. Assembled sequences were combined with publicly available mitochondrial assemblies for *Physalia* and their outgroup *Rhizophysa* from NCBI, accession numbers: OQ957220, KT809328, LN901209, KT809335, NC_080942, NC_080941, OQ957206, OQ957199. Mitochondrial genomes were aligned using `MAFFT`, v. 7.505, `--adjustdirectionaccurately` option (*18*). A mitochondrial phylogeny was inferred using `IQtree2`, v. 2.2.6 (*19*), model autoselected (*20*) and 1,000 ultrafast bootstraps (*21*), with *Rhizophysa* selected as the outgroup.

Individual marker sequences were assembled from raw reads using *in silico* PCR as implemented in `sharkmer` (available at https://github.com/caseywdunn/sharkmer, commit `c43cfc2`). Four markers were selected to infer individual gene trees: mitochondrial cytochrome oxidase I (CO1), mitochondrial large ribosomal subunit 16S, nuclear ribosomal internal transcribed spacer (ITS), as well as small nuclear ribosomal subunit 18S. These markers were combined with all publicly available *Physalia* and *Rhizophysa* sequences for the same genes, from NCBI. Sequences were aligned with `MAFFT`, and gene trees inferred with `IQtree2`, as described above.

A phylogeny of single nucleotide polymorphisms (SNPs) was assembled using `SVDquartets`, as implemented in PAUP*, v. 4.0a (*22*). SNPs were selected based on the following filters: minimum Phred quality of 40, minimum and maximum depth of 2x and 99x respectively, maximum proportion of missing data of 25%, minimum distance between SNPs set to 100 base pairs, excluding sites with only alternative alleles called, and only selecting bi-allelic SNPs. The final dataset contained 839,510 SNPs. `SVDquartets` was used to infer a phylogeny of all specimens without population-level information, and a phylogeny with specimens assigned to

populations based on results of the principal component and admixture analyses. For the latter, support was evaluated using 100 bootstraps.

Principal components analysis

Principal component analysis (PCA) was performed on estimated genotype likelihoods, calculated using `ANGSD`, v. 0.935 (*23*), on reads mapped to a random sample of 100,000 non-repeat genomic regions, each larger than 1,000 base pairs. Sites were included based on the following filters: p-value of variability below 1e-6, minimum Phred quality score of 40, minimum and maximum depth of 2x and 99x respectively, and present in a minimum of 92 individuals (75% of 123 samples). PCA and admixture analyses were performed using `PCANGSD`, v. 1.21 (*24*), `-admix-alpha` set to 50 and allowing the software to choose the optimal number of components.

PCA and admixture analyses were repeated on the full subset of samples, using a minimum of 100 samples (75% of 133 samples), as well as with reads mapped to the reference transcriptome. PCA was also performed within each lineage detected in this study. Subpopulations were classified using k-means clustering of the resultant covariance matrices, with the optimal number of clusters chosen using an elbow plot of eigenvalues.

Population statistics

Populations genomic statistics (*pi*, *Dxy*, and *Fst*) were calculated using `pixy`, v. 1.2.7 (*25*), on a dataset of alleles filtered with the following metrics: minimum Phred quality score of 40, minimum and maximum depth of 2x and 99x respectively, maximum missingness of 25%. Statistics were calculated on a random sample of 100,000 non-repeat genomic regions, each larger than 1,000 base pairs, and summary statistics were averaged over these regions. Statistics were calculated between lineages as assigned using PCA and admixture analyses; between subpopulations, as defined using PCA and admixture within species; and between lineage + sampling location combinations.

Morphological scoring of images

ID numbers for ~11,000 research-grade photos of *Physalia* were downloaded in October, 2023. Of these, a subset of 4,047 images were scored, selected to include multiple images from all represented countries and time zones, as well as to maximize representation in areas hypothesized to have increased diversity (specifically New Zealand, South Africa, and Brazil). Images were categorized based on quality and perspective on the animal (e.g., ventral, dorsal, or lateral), and were scored for the following traits:

- sail height, binned into four categories: as tall as float, >1/3 the height of float, <1/3 the height of float, or flush with float / no visible height
- length of float anterior to the end of the sail, binned as <1/4 sail length, >1/4 and <3/4 sail length, and >3/4 sail length
- presence of pink or purple coloration on the sail
- presence of yellow or reddish coloration on gastrozooids
- clear, glassy float coloration
- arrangement of principal fishing tentacles (defined as having tentilla tightly packed), categorized as having one central tentacle, two central tentacles, or many
- presence of a gap between the central (main) and posterior colony zone of zooids
- juvenile morphology, defined as having a globular float with one or no major tentacles, no sail height, and few zooids.

Each trait was only scored when visible, therefore absence of a score is not evidence of trait absence. Images were scored in batches by three different researchers (SHC, RBA, and NA). To ensure consistency, researchers independently scored the same set of 100 randomly sampled photos, and compared results to bring qualitative assignments into alignment. Images classified as being of poor quality, taken from a ventral perspective, or of a juvenile specimen as defined above, were excluded from downstream analyses.

Four morphological types were identified from scored images in combination with descriptions and diagrams of historically hypothesized species. Rules for assigning images to one of these four morphologies were established based on combinations of characters, see Fig. S12. Given the potential plasticity of the traits in question (e.g., color, size), no single trait was considered diagnostic. Genomic clusters were associated with these morphologies by scoring the same traits on the specimens processed for genomic analyses.

When image assignments extended the known range of a genomically defined lineage, these images were independently rescored by two researchers. If there was any discrepancy in the resulting scores for a trait relative to the morphological assignment, the image was excluded from the rule-based analysis.

1. J. Chen, S. Dellaporta, "Urea-based plant DNA miniprep" in *The Maize Handbook* (Springer, 1994), pp. 526–527.

2. A. Guang, M. Howison, F. Zapata, C. Lawrence, C. W. Dunn, Revising transcriptome assemblies with phylogenetic information. *PLoS One* **16**, e0244202 (2021).

3. S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, A. M. Phillippy, Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**, 722–736 (2017).

4. S. Koren, A. Rhie, B. P. Walenz, A. T. Dilthey, D. M. Bickhart, S. B. Kingan, S. Hiendleder, J. L. Williams, T. P. Smith, A. M. Phillippy, De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology* **36**, 1174–1182 (2018).

5. S. Nurk, B. P. Walenz, A. Rhie, M. R. Vollger, G. A. Logsdon, R. Grothe, K. H. Miga, E. E. Eichler, A. M. Phillippy, S. Koren, HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research* **30**, 1291–1305 (2020).

6. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

7. M. J. Roach, S. A. Schmidt, A. R. Borneman, Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 1–10 (2018).

8. L. Coombe, J. X. Li, T. Lo, J. Wong, V. Nikolic, R. L. Warren, I. Birol, LongStitch: High-quality genome assembly correction and scaffolding using long reads. *BMC Bioinformatics* **22**, 1–13 (2021).

9. J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, A. F. Smit, RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).

10. M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, E. M. Zdobnov, BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular biology and evolution* **38**, 4647–4654 (2021).

11. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

12. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 1–13 (2019).

13. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742–015 (2015).

14. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013).

15. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

16. J.-J. Jin, W.-B. Yu, J.-B. Yang, Y. Song, C. W. DePamphilis, T.-S. Yi, D.-Z. Li, GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology* **21**, 1–31 (2020).

17. N. Ahuja, X. Cao, D. T. Schultz, N. Picciani, A. Lord, S. Shao, K. Jia, D. R. Burdick, S. H. Haddock, Y. Li, others, Giants among cnidaria: Large nuclear genomes and rearranged mitochondrial genomes in siphonophores. *Genome Biology and Evolution* **16**, evae048 (2024).

18. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular biology and evolution* **30**, 772–780 (2013).

19. B. Minh, H. Schmidt, O. Chernomor, D. Schrempf, M. Woodhams, A. Von Haeseler, R. L. IQ-TREE, IQ-TREE2: New models and efficient methods for phylogenetic inference in the genomic era. *DOI: https://doi. org/10.1093/molbev/msaa015* **37**, 1530–1534 (2020).

20. S. Kalyaanamoorthy, B. Q. Minh, T. K. Wong, A. Von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587–589 (2017).

21. D. T. Hoang, O. Chernomor, A. Von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35**, 518–522 (2018).

22. J. Chifman, L. Kubatko, Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**, 3317–3324 (2014).

23. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).

24. J. Meisner, A. Albrechtsen, Inferring population structure and admixture proportions in low-depth NGS data. *Genetics* **210**, 719–731 (2018).

25. K. L. Korunes, K. Samuk, Pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources* **21**, 1359–1368 (2021).