

# **Population genomics of bluebottles (man o' war, Cnidaria: *Physalia*) and diversity on the high seas**

The bluebottle or Portuguese man o' war (genus *Physalia*), is an open-ocean predator that uses a gas-filled float and muscular crest to sail the surface of the sea. These jellies are distributed across the globe, and like many holoplanktonic organisms, have been hypothesized to be part of a large, panmictic population that extends across ocean basins. We tested this hypothesis by comparing genomic variation of 121 specimens collected by scientists in over a dozen countries. Our results revealed five distinct populations, with multiple lines of evidence indicating that mixing between populations is rare, despite regions of range overlap. We combined these data with an independent dataset of thousands of images of *Physalia* uploaded to the citizen-science website inaturalist.org, which we scored for morphological characters of crest size, tentacle number, and color. From these images, we identified four recognizable morphologies, described their geographical distribution, and linked them to the populations identified with genomic data. Based on these findings, we propose there are at least four species, three of which correspond to species hypothesized by scientists in 18th and 19th centuries: *P. physalis*, *P. utriculus*, and *P. megalista*, along with one as yet unnamed species *Physalia* sp. from the South Pacific. Within each species, we characterize genomic variation against oceanographic distribution, and observe evidence of persistent subpopulations at a regional scale, linked by major ocean currents, along with evidence for individual long-distance dispersal events. Our findings highlight the underappreciated diversity and connected nature of the imperiled ocean surface ecosystem.

## **Introduction**

XXXXXX

I welcome feedback on what would best be included in the introduction, especially suggestions of relevant citations. Ideas include

- introduction to *Physalia* biology

- presentation of current hypotheses
- challenges in predicting population structure in the absence of physical barriers
- relevance to humans, medical and economic impacts
- relevance to open ocean ecosystem

XXXXXX

## Results

### Genomic variation indicates five distinct populations

We generated a new genome assembly for *Physalia* from a specimen collected in Texas, USA in 2017. This assembly, along with its alternate haplotype counterpart, have high contiguity (N50 of 10.4 and 4.6 megabases, see Table S1) and high BUSCO completeness scores (89.7% and 86.9%, Table S2). Like other siphonophores, the genome of *P. physalis* is characterized by a substantial fraction of repeat sequences (~65%, Fig. S1). We estimated the genome size of ten specimens and found that size estimates vary between 1.5 and 2.0 gigabases. These estimates are all smaller than the total assembled length (primary assembly 3.3 Gb), indicating that the assembly process likely inflated the total number of repeat sequences. For all downstream analyses we used only reads mapped to non-repeat regions of the genome. In addition to the genome assembly, we also generated a new transcriptome using full length cDNA generated with the IsoSeq method on an independent specimen of *P. physalis*, collected in Florida in 2023. Downstream analyses were repeated over the genome and transcriptome assemblies to test robustness of results to reference assembly.

Our global collaboration collected >350 specimens, the majority of which are stored at the Yale Peabody Museum (Fig 1B, SI: sample information). Of these, we resequenced the genomes of 124 specimens, along with nine biological replicates. Three samples were removed as biological based on indications of potential contamination (see SI: sample quality), and another 10 were marked as moderate quality and results were tested both including and excluding these samples.

For the remaining 111 samples, we used the software PCANGSD to perform a principal component analysis (PCA) of estimated genotype likelihoods. The results show samples are divided into five clusters along the first principal components of genomic variation (Fig. 1C-D, S2). These five clusters are labeled as A, B, and C, with B and C divided into B1+B2 and C1+C2, given the adjacency of those clusters to one another along the first principal components. We repeated the PCA including an additional ten samples of moderate quality, and as well mapping data to the reference transcriptome, and observed the same five clusters (Fig. S3-S4).

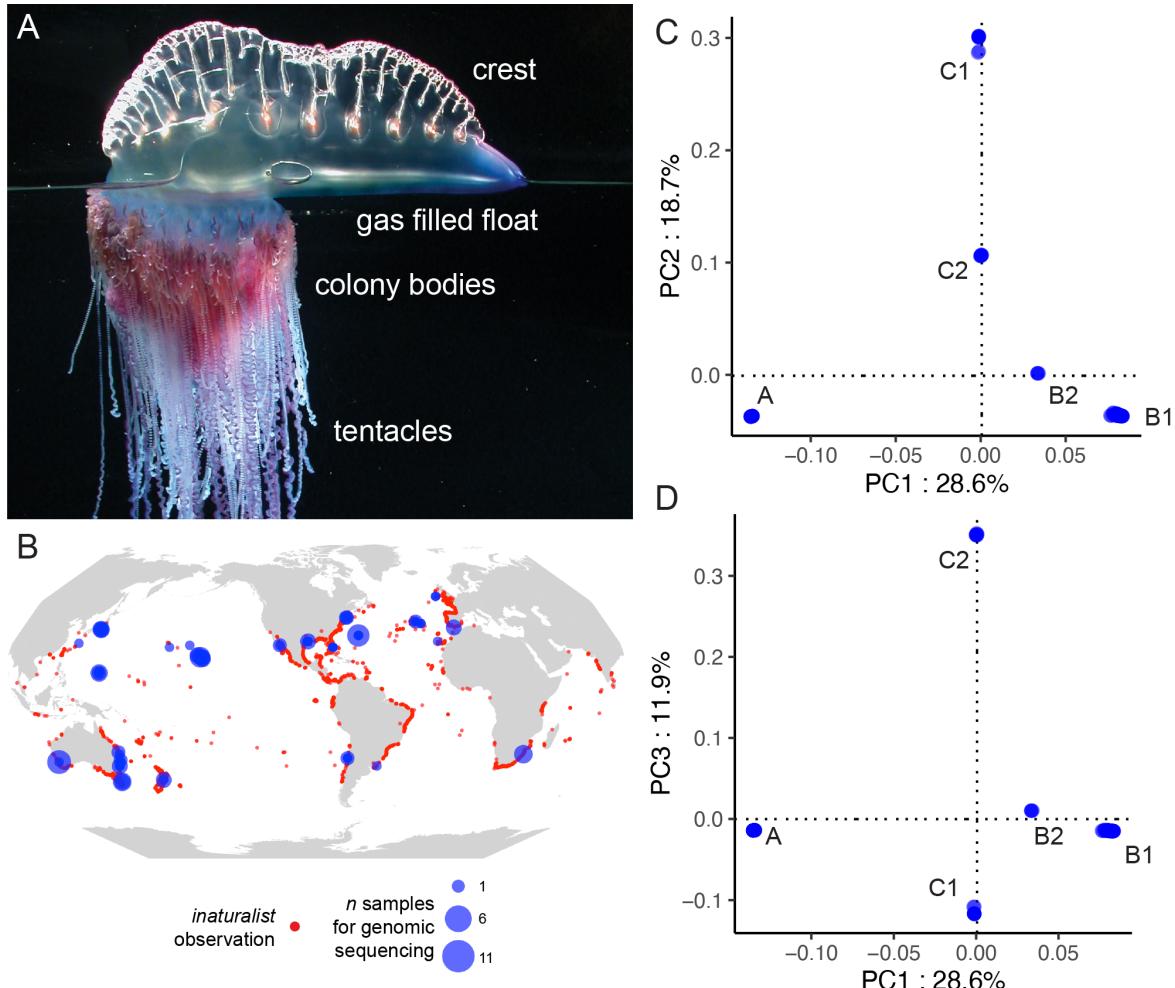


Figure 1: Biology, distribution, and genomic variation of *Physalia*. A, *Physalia* colonies comprise a muscular crest attached to a gas-filled float which maintains the mature animal at the surface of the water. Colony bodies (zooids), including those specialized for feeding (gastrozooids), prey-capture (palpons with tentacles) and reproduction (gonozoooids) are added to the float via asexual reproduction. Tentacles drape below to trap, sting, and retrieve fish. B, *Physalia* are observed throughout the world, as shown by observations posted to [inaturalist.org](https://inaturalist.org) (red). Samples for genomic analysis (blue) were collected across ocean basins by our global collaboration of scientists. C-D, The first three principal components of variation reveal five populations, labeled as clusters A, B1, B2, C1, and C2.

## **Evidence of strong genomic differentiation between populations**

The distribution of these five populations shows that at least two clusters were recovered from multiple ocean basins: cluster B1 was found in the S. Atlantic, S. Indian, and S. and N. Pacific; cluster C1 was found on both sides of the S. Indian and S. Pacific oceans. In contrast, cluster A was observed only in the Atlantic, B2 on both sides of the N. Pacific, and C2 only in New Zealand and Tasmania. We evaluated population differentiation by calculating the reciprocal fixation index ( $F_{ST}$ ), averaged across non-repeat genomic windows. Average  $F_{ST}$  values range from 0.29, between B1 and B2, to 0.65, between A and C1, suggesting little genetic exchange between populations (Fig. 2B, see Fig. S5 for range across genomic windows). Estimates of nucleotide diversity,  $\pi$ , indicate that cluster A has the lowest overall diversity and clusters B1 and C2 have the highest (Fig. S6), consistent with estimates of individual heterozygosity (Fig. S1C).

We tested the monophyly and phylogenetic relationships of these lineages using two approaches: First, we assembled mitochondrial genomes for each sample and inferred a mitochondrial tree. For this analysis, we combined the mitogenomes generated in this study with all publicly available *Physalia* genomes, as well as mitogenomes from the sister genus *Rhizophysa* as an outgroup. The most likely tree shows clusters are monophyletic, with relatively little sequence variation within clusters (Fig. 2C). The clusters B1 and B2 were found to be sister to one another with high bootstrap support, and the clade of B1+B2 sister to cluster A. Support values were lower for the relationships at the base of the *Physalia* phylogeny.

Second, we estimated the phylogeny from a dataset of 800k high-quality SNPs, using the coalescent-based software SVDQuartets. A phylogeny of all specimens confirmed the monophyly of the five populations (Fig. S7). We examined the relationships between populations by estimating a tree with individuals assigned to their respective cluster. Our results again indicated a split between clusters C1, C2, and the clade of A and B1+B2 (XXX SI). Bootstrap support values for both partitions in this unrooted tree were universally high (100, Fig. S7C).

We used an admixture analysis to test for evidence of mixing across populations. The results favored five populations, corresponding to the five clusters above, and showed little evidence of mixture between populations (Fig. 2D). Repeating this analysis including 10 samples of moderate quality returned the same general results (Fig. S3), with the exception of three moderate-quality specimens of C2 that showed a moderate proportion of admixture with C1. Repeating analyses using the reference transcriptome returned the same results (Fig. S4).

Several studies have examined *Physalia* diversity using single genetic markers as a point of comparison. In order to place those data in the context of our findings, we inferred individual genetrees for four genes: CO1, ITS, 16S, and 18S (Fig. S8-S11). We combined publicly available sequences from NCBI with assembled marker sequences from our specimens, inferred using *in silico* PCR as implemented in our custom software **sharkmer**. These results furthered our understanding of *Physalia* diversity in the following ways: [1] a specimen reported from the

Sargasso Sea (N. Atlantic) extended the predicted range of B1; [2] a specimen reported from Pakistan (N. Indian) extended the predicted range of B2; [3] using the internally transcribed spacer gene ITS, we were able to assign three clans, described in New Zealand, to clusters we describe here: clan 1 = cluster C2, clan 2 = cluster B1, clan 3 = cluster C1; however, using COI we found an incongruent result for the identity of clan 3. Without further information we cannot distinguish whether this result is due to potential exchange of mitochondrial sequences between clusters C1 and C2 in New Zealand, or an artifact of the marker amplification and assembly techniques from the original study.

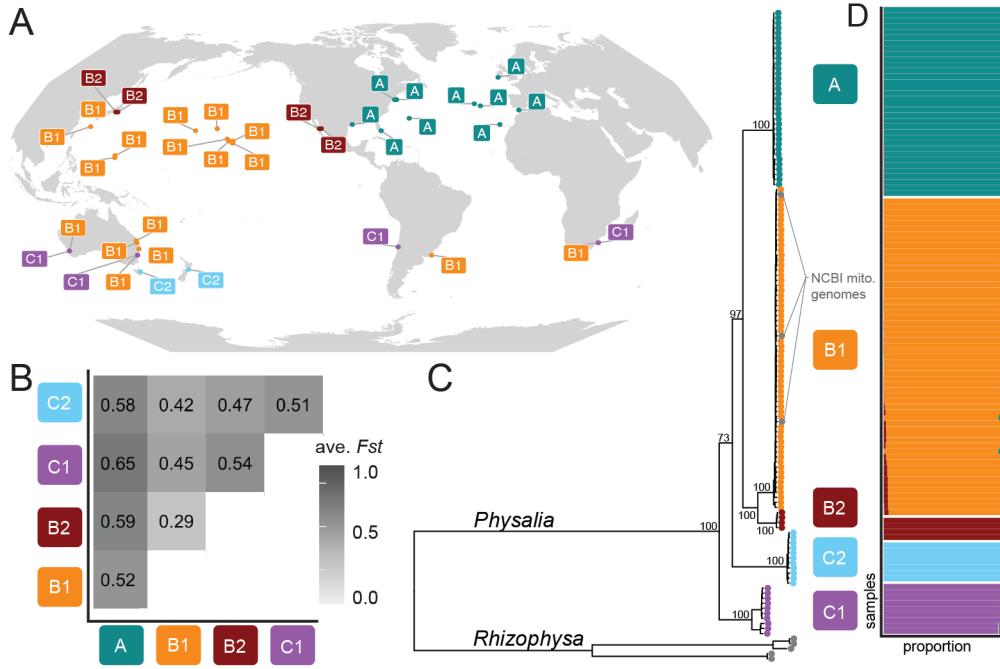


Figure 2: Multiple lines of evidence indicate reproductive isolation between populations. A, The distribution of the five clusters from Fig. 1 shows populations span ocean basins (e.g. B1, C1) and populations are restricted to smaller areas (e.g. C2 observed in New Zealand and Tasmania). B, Reciprocal fixation index ( $F_{ST}$ ) averaged across non-repeat genomic windows indicate high levels of reproductive isolation between all populations, with the weakest between B1 and B2. C, Phylogenetic analysis of assembled mitochondrial sequences shows reciprocal monophyly of populations. Bootstrap values are shown at internal *Physalia* nodes. D, Admixture analysis recovers five populations with little evidence of mixture.

## Trait-based analysis of citizen-science images reveals four morphologies

We tested for the evidence of distinct morphologies of *Physalia* by analyzing a dataset of thousands of research-grade images of *Physalia* uploaded to the citizen-science website [inaturalist.org](https://inaturalist.org). While most of these images are of beached and deceased specimens, many aspects of the gross morphology are often identifiable. We scored these aspects on a subset of XXX of images, focusing on the following characters: the height and length of the crest, relative to the float; the color of the crest apex and of the colony bodies (primarily gastrozooids); the number of major tentacles (defined as having dense tentilla); and the visible presence of a gap between zooid growth zones. For each trait, we grouped them into broad categories (e.g. one, two, or many major tentacles), and we had three independent observers score the same set of 100 images to confirm reproducibility.

From these images, we identified four distinct morphologies (Fig. 3A-B). These were defined by describing sets of rules for positive identification based on suites of characters, excluding images of poor quality or of juvenile specimens (Fig. S12). These rules constitute a strict definition for a high confidence observation of each type; for example, images were positively identified as the *P. physalis* morphology if they had reddish feeding bodies, multiple tentacles, and a crest that as tall as the float extends nearly to the anterior end. While specimens of *P. physalis* may deviate from these characters (e.g. if the crest is not raised), the rules were designed to minimize overlaps between morphologies and allow for high-confidence identifications.

Three of the morphologies we identified correspond to species hypothesized by scientists centuries ago. *P. physalis*, was named by Linneaus based off of specimens from the Atlantic that had large sails and multiple tentacles (Fig. 3A). *P. utriculus* was named by Gmelin in 1788 based of illustrations of a Pacific specimen that had a single tentacle, yellow-tipped gastrozooids, and a flared posterior growth zone. *P. megalista* was described based off illustrations by Lesueuer and Petit made in 1807 of specimens from the Southern Ocean that had an incomplete crest and a sinuously postured float. Each of these species names was synonymized with *P. physalis* in later centuries; our results indicate these synonymizations to have been unjustified.

Based on the positive identification of these images, we examined the distribution of morphologies across ocean basins (Fig. 3C). We found that *P. physalis* was observed in the N. and Southwest Atlantic; *P. utriculus* was found throughout the Pacific and Indian and extending into the SW Atlantic and Gulf of Mexico; *P. megalista* was found in the Southern edges of the Atlantic, Indian, and Pacific; and the unnamed morph *P. sp* was found in New Zealand, Tasmania, and E. Australia. We compared the morphology of the specimens collected and sequenced for genomic analysis to the morphotypes identified from inaturalist data, using images taken upon collection, when available, and the morphology of fixed specimens (see SI: sample information for images). Our results confirm that cluster A corresponds to *P. physalis*, B1 to *P. utriculus*, C1 to *P. megalista*, and C2 to *P. sp*. No images of collected specimens were available for cluster B2, and analysis of the morphology of the single available fixed specimen showed gross similarity to specimens of B1 / *P. utriculus*.

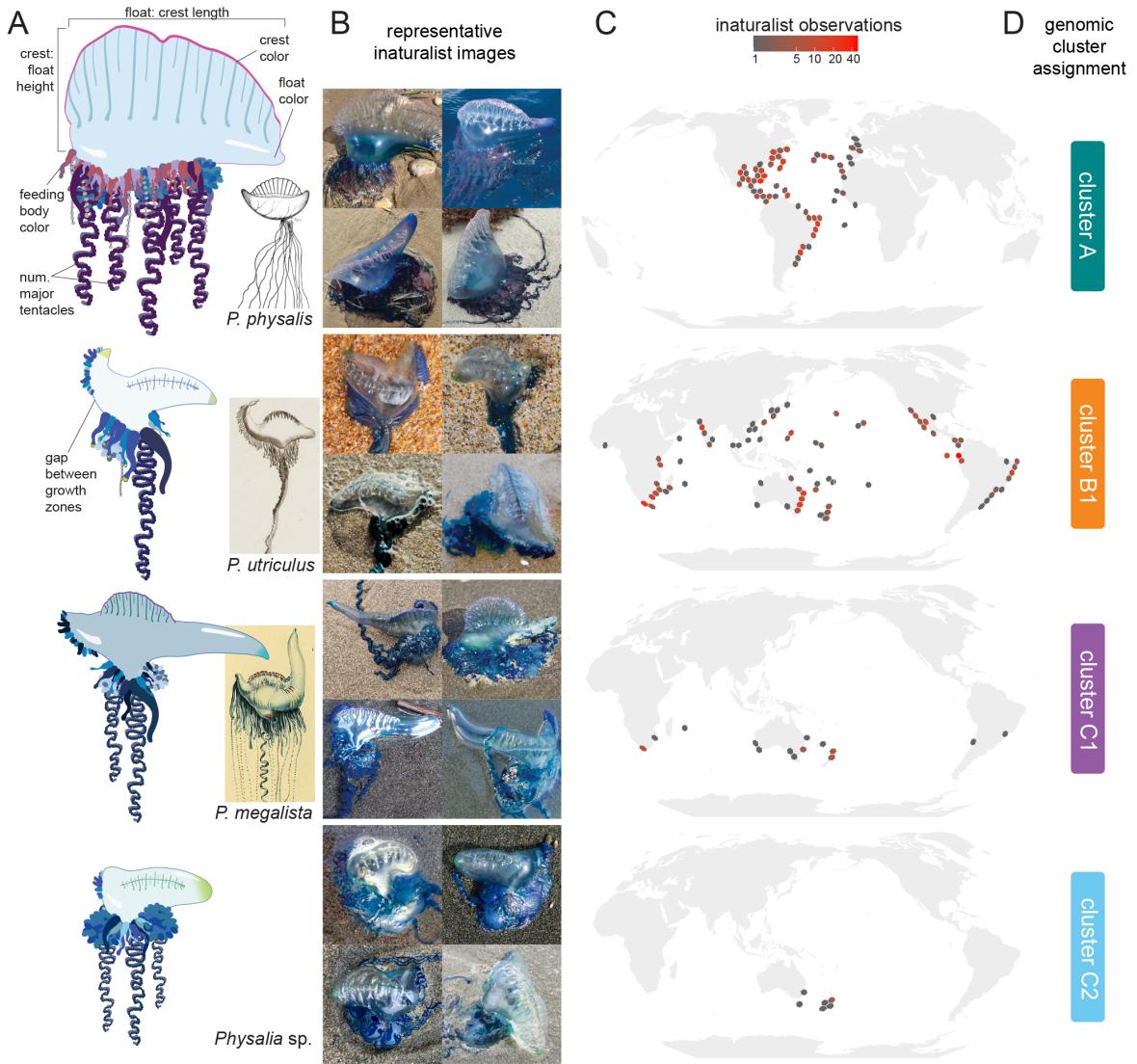


Figure 3: Distinct morphologies are detectable in citizen science images. A, Morphological traits such as aspects of size, color, and tentacle number were scored for thousands of images on inaturalist.org. From these, four morphologies were identified, and are illustrated here. Three of these correspond to historically hypothesized species. B, Representative photos of each morphology from inaturalist. C, Ranges of positively identified inaturalist records for each morphology, using a rule-based analysis of morphological traits. D, Morphologies were assigned to a genomic cluster by scoring the same traits of genomic specimens.

## Oceanographic structure and long-distance migration within species

Given that these animals are capable of wind-powered movement, combined with the evidence of distributions extended across ocean basins, we tested for evidence of long-distance dispersal and subpopulations structure within species. We repeated PCA on samples in each of the four clusters *P. physalis*, *P. utriculus*, *P. megalista*, and *P. sp C2*. For *P. utriculus* we repeated this analysis as both inclusive and exclusive of cluster B2.

Within each of the species compared, samples largely grouped by oceanographic region where collected (Fig. 4). In addition, PC1 was largely aligned with longitudinal distance between samples. Samples did not group by date of collection in the first principal components for any species (Fig. S13-S15). The observation of a strong oceanographic signature, persistent even when samples were collected at sites over the span of multiple years, indicates that most *Physalia* subpopulations largely stay in place over time.

Several exceptions to the pattern of persistent regional subpopulations are found in the dataset. For *P. physalis*, samples from Florida, Bermuda, and New England are highly similar, without substructure corresponding to collection sites, indicating these samples are part of one large population aligned with the Gulf Stream current and wind pattern (Fig. 4A-B). In addition, we detected three samples within *P. physalis* and one sample within *P. utriculus* that showed “ectopic” genomic signatures, indicating individual histories of long-distance dispersal. In *P. physalis* the three samples are found in Bermuda and the Azores, suggesting samples occasionally exchange across the central N. Atlantic. In *P. utriculus*, one sample collected in Hawai’i showed similarity to samples collected in Guam and Japan, suggesting transport across the N. Pacific.

We tested the strength of population differentiation between subpopulations, defined using k-means clustering within species (Fig. S16). This analysis recovered two subpopulations within *P. physalis*, three in *P. utriculus*, two in *P. megalista*, and no significant substructure within *P. sp C2*. Population B2 was treated as a single cluster, given limited sampling. Average *Fst* values between subpopulations were largely small (<0.05), with the exception of the division within *P. megalista* (average *Fst* of 0.12, Fig. S17). This division is reflected in the mitochondrial and nuclear phylogenetic results (Fig. 2C, S7A), and suggests a potential barrier to reproduction within *megalista*. We also calculated *Fst* values between subpopulations defined by the combination of species and sampling region (Fig. S18). This comparison confirmed that *Fst* values are not lower between subpopulations of different species collected in the same region (e.g. *utriculus* and *megalista* that co-occur in the SW Pacific, *Fst* value of 0.43).

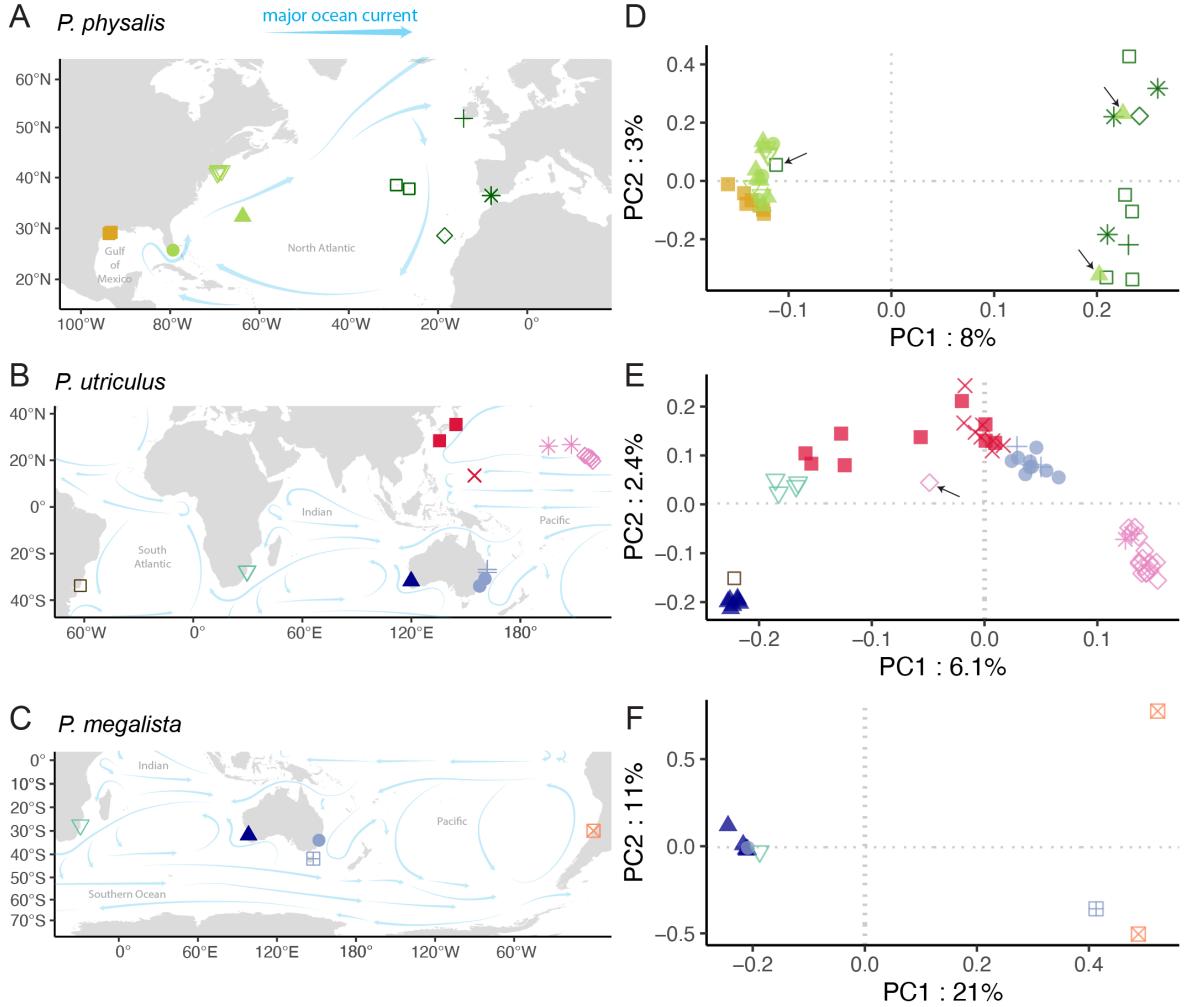


Figure 4: Within species, genomic variation indicates persistent populations within ocean regions, as well as individual long-distance dispersal events. For each of *P. physalis* (A), *P. utriculus* (B), and *P. megalista* (C), we visualized genomic variation and oceanographic distribution. Colors indicate regions of the ocean (e.g. Northwest Atlantic), and shapes indicate sampling location (e.g. Florida). Major ocean currents, as compiled by NOAA, are shown with blue dashed arrows. D-F, Principal component analyses within species show that subpopulations are largely defined by region. Exceptions to this pattern are marked with black arrows; these individuals suggest long-distance dispersal events across regions. Similarity between regions is largely explained by longitude and ocean currents.

## Discussion

XXXXXX

As with the introduction, I welcome suggestions on what are the most salient points to include in the discussion. A few ideas are spelled out below.

XXXXXX

The evolutionary history of *Physalia* suggests that diversification may have originated from the southwest Pacific, where three species are currently known to co-occur. The most recent divisions are between *P. physalis* and *P. utriculus*, perhaps as the ancestral population spread to occupy the Atlantic ocean, and between population B2 and *P. utriculus*, here described from the North Pacific. In addition, strong genomic differentiation between subpopulations of *P. megalista* suggests a potential nascent split, with samples from the SW. Pacific falling on both sides of this division.

The southwest regions of ocean basins consistently represent hotspots of *Physalia* diversity: three species are found in the SW Pacific (*utriculus*, *megalista*, and C2), two species in the SW. Indian ocean (*utriculus* and *megalista*), and three species in the SW. Atlantic (*physalis*, *utriculus*, and *megalista*). In contrast, the eastern boundaries of ocean basins are often lacking in *Physalia* species and observations: *Physalia* are not reported north of Baja California, and there are few observations in W. Africa or between Peru and Chile. These patterns are consistent with dynamics XXX.

For centuries, naturalists have speculated on the number of *Physalia* species (a complete treatment of *Physalia* taxonomy over time is available in XXXPughXXX). While many dozens of species have been hypothesized, three names have persisted through time: *P. physalis*, *P. utriculus*, and *P. megalista*. In the case of *P. utriculus*, most authors from the 17th-19th centuries predicted there were at least two species, one centered in the Atlantic and one in the Indopacific. Twentieth century naturalists synonymized *utriculus* based on the observation that juvenile Atlantic specimens are morphologically similar to the typical Indopacific morphs (quote from Phil..?). In the case of *P. megalista*, the concept of a Southern Ocean species was invoked through several iterations, often associated with diagrams of a *Physalia* with an short crest and long anterior projection. Other authors found illustrations of *megalista* improbable (quotes about the posture). As recent as 2010, the taxonomy was again revised, and in an attempt to reconstruct a Pacific species, new material was designated as a neotype for both *utriculus* and *megalista* as proposed synonyms of one another. Our results build upon this history, vindicating several early predictions and adding new complexity to others. It's clear that the original descriptions of *physalis*, *utriculus*, and *megalista* accurately captured three distinct species. While *physalis* is restricted to the Atlantic, we show that *utriculus* is also found in the S. Atlantic, and at least occasionally as far north as the Gulf of Mexico, and therefore is not correctly designated as an Indopacific species. Our morphological analysis also robustly confirms that the original drawing of *megalista* represents a realistic posture for specimens from this species. Our work rejects the recent synonym between *utriculus* and

*megalista*, but pending further analysis of the neotype material, we can't yet determine to which name it should be assigned.

## Methods

### Sample collection and DNA extraction

*Physalia* specimens were collected by a global collaboration of scientists, full sampling details can be found in the Supplemental Information: specimen info, SIXXX. The majority of specimens were collected from the beach following being 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). Specimens were preserved in >70% alcohol (ethanol, when available), and stored at room temperature. When possible, whole specimens were collected and shipped to the Yale Peabody Museum. All sequenced specimens were photographed, images and accession numbers can be found in SIXXX. Several specimens were also loaned from the Western Australia Museum and the Tasmanian Museum and Art Gallery, and one specimen from the Field Museum in Chicago.

Tentacle tips were dissected from whole specimens and stored in alcohol prior to DNA extraction for genome resequencing. Two juvenile samples collected from Hawai'i were stored in DNAshield. Several samples from Texas, Florida, and New England were directly flash-frozen. DNA extractions were performed using the EZNA Mollusk kit with an overnight digestion, with the exception of one sample from the Gulf of California, extracted at Monterey Bay Aquarium Research Institute with the XXXprotocolXXX, and for XXXKristin samplesXXX. Whole genome DNA was processed for library preparation by the Yale Center for Genome Analysis.

High molecular weight DNA was extracted from a flash frozen specimen collected in Texas, USA in 2017. XXXHMW protocolXXX.

### Genome assembly

Eight Single-Molecule Real-Time (SMRT) sequencing cells of PacBio HiFi data were assembled with `canu`, version 2.2 (`-pacbio-hifi` option), with the estimated genome size parameter set to three gigabases. HiFi reads were mapped to this assembly with `minimap2`, v. 2.22-r1101, 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), 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) 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, 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**. 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) to build a general feature format (gff) file, used to exclude repeats from downstream analyses. **BUSCO**, v 5.4.4, and **BBMap stats.sh** were used to evaluate final assemblies. Assemblies are made publicly available at NCBI, project number PRJNA1040906, raw reads are available at XXX.

### Genome mapping

Paired-end genome sequencing targeting a length of 150 basepairs was performed for 133 libraries using an Illumina NovaSeq at the Yale Center for Genome Analysis. Sequencing depth ranged is reported in the Supplemental Information: sequence quality information SIXXX. The 133 samples included several biological replicates (multiple 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 downstream analyses.

Overall sequence quality (GC content, adapter content) was evaluated using **FastQC** (v. 0.11.9). Reads were trimmed for Illumina adapters using **Trimmomatic** (v. 0.39). Potential human, bacterial, and viral DNA contamination was evaluated using **Kraken2** (v. 2.1.2, 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 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, 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). Full quality control reports are available in SIXXX.

Based on the results of the quality control analyses, 3 samples were identified as contaminated and excluded from downstream analyses. The final dataset, excluding technical and biological replicates, consisted of 121 samples. Of those, 111 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). Mapped reads were sorted, deduplicated, and indexed using **picard** (v. 2.25.6). Alleles were called using **BCFtools** (v. 1.16, **mpileup**). To test robustness of downstream analyses to reference assembly, reads were mapped to the independent transcriptome assembly (mapping only R1 reads as single-end data).

## Phylogenetics

Mitochondrial genomes were assembled from trimmed reads using the software **GetOrganelle** (v. 1.7.7.0, using the **animal\_mt** database and default parameters). **GetOrganelle** failed to circularize the assemblies, in line with recent observations of linear mitochondrial genomes in siphonophores XXX; 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 sister genus *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). A mitochondrial phylogeny was inferred using **IQtree2** (v. 2.2.6, model autoselected and 1,000 ultrafast bootstraps XXXmultiple citationsXXX), with *Rhizophysa* used as the outgroup.

Individual marker sequences were assembled from raw reads using *in silico* PCR as implemented in **sharkmer** (v. XXX). Four markers were selected to infer individual gene trees: mitochondrial cytochrome oxidase I (CO1), internal transcribed spacer (ITS), and large ribosomal subunit 16S, as well as 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). 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 basepairs, 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) on reads mapped to a random sample of 100,000 non-repeat

genomic regions, each larger than 1,000 basepairs. Sites were included based on the following filters: maximum p-value of variability of 1e-6, minimum Phred quality score of 40, minimum and maximum depth of 2x and 99x respectively, and present in a minimum of 91 individuals (~75% of 121 total samples). PCA and admixture analyses were performed using PCANGSD (v. 1.21, -admix-alpha set to 50), allowing the software to choose the optimal number of components.

PCA was repeated on subsets of samples within each population. Subpopulations were classified using k-means clustering of the covariance matrix, 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) 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 basepairs, and summary statistics were averaged over regions. The final dataset contained Statistics were calculated between populations as assigned using PCA and admixture analyses; between subpopulations, as defined using PCA and admixture within species; and between population + sampling location combinations.

### **Morphological scoring of images**

A dataset of XXX ID numbers for research-grade photos of *Physalia* were downloaded on XXX. Of these, a subset of XXX 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 (New Zealand, South Africa, 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 anterior projection of float, binned as  $<1/4$  crest length,  $>1/4$  and  $<3/4$  crest length, and  $>3/4$  crest length
- presence of pink or purple coloration on crest
- presence of yellow or reddish coloration on gastrozooids
- clear, glassy float coloration
- number of mature fishing tentacles (defined as having tentilla tightly packed), binned as one tentacle, two tentacles, or more than two tentacles
- presence of a gap between the oral and posterior 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 (Samuel Church, River Abedon, and Namrata Ahuja). To ensure consistency, researchers first scored the same 100 randomly sampled photos, and compared results to bring qualitative assignments into alignment. Images classified as being poor quality, from a ventral perspective, or of a juvenile specimen were excluded from downstream analyses.

Four distinct morphologies were identified after examining thousands of 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 SIXXX. Given plasticity of the traits of the characters 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 specimens processed for genomic analyses.

When image assignments extended the known range of a population, as defined by the genomic results, 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.