The state of Medusozoa genomics: current evidence and future challenges

Mylena D. Santander [1,*], Maximiliano M. Maronna [2,*], Joseph F. Ryan [3,4] and Sónia C. S. Andrade [1]

[1]Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade São Paulo, 277 Rua do Matão, Cidade Universitária, São Paulo 05508-090, Brazil
[2]Departamento de Zoologia, Instituto de Biociências, Universidade de São Paulo, São Paulo, 101 Rua do Matão, Cidade Universitária, São Paulo 05508-090, Brazil
[3]Whitney Laboratory for Marine Bioscience, University of Florida, 9505 Ocean Shore Blvd, St. Augustine, FL 32080, USA
[4]Department of Biology, University of Florida, 220 Bartram Hall, Gainesville, FL 32611, USA
*Correspondence address. Mylena D. Santander, Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade São Paulo, 277 Rua do Matão, Cidade Universitária, São Paulo 05508-090, Brazil. E-mail: mylena.santander@gmail.com; Maximiliano M. Maronna, Departamento de Zoologia, Instituto de Biociências, Universidade de São Paulo, 101 Rua do Matão Cidade Universitária, São Paulo 05508-090, Brazil. E-mail: maxmaronna@gmail.com

Abstract

Medusozoa is a widely distributed ancient lineage that harbors one-third of Cnidaria diversity divided into 4 classes. This clade is characterized by the succession of stages and modes of reproduction during metagenic lifecycles, and includes some of the most plastic body plans and life cycles among animals. The characterization of traditional genomic features, such as chromosome numbers and genome sizes, was rather overlooked in Medusozoa and many evolutionary questions still remain unanswered. Modern genomic DNA sequencing in this group started in 2010 with the publication of the *Hydra vulgaris* genome and has experienced an exponential increase in the past 3 years. Therefore, an update of the state of Medusozoa genomics is warranted. We reviewed different sources of evidence, including cytogenetic records and high-throughput sequencing projects. We focused on 4 main topics that would be relevant for the broad Cnidaria research community: (i) taxonomic coverage of genomic information; (ii) continuity, quality, and completeness of high-throughput sequencing datasets; (iii) overview of the Medusozoa specific research questions approached with genomics; and (iv) the accessibility of data and metadata. We highlight a lack of standardization in genomic projects and their reports, and reinforce a series of recommendations to enhance future collaborative research.

Keywords: annotation, completeness, assembly, genome size, chromosome number, collaborative genomics

Background

Medusozoa subphylum includes nearly 4,055 species of invertebrates distributed in the classes Hydrozoa, Cubozoa, Staurozoa, and Scyphozoa [1], which are found at all latitudes in almost all aquatic environments, from freshwater to marine, and from shallow to deep waters (Fig. 1). Medusozoa species, together with the other cnidarian classes (i.e., Anthozoa and Endocnidozoa), harbor some of the most plastic life cycles and diverse body plans among animals [2] and represent one of its early diverging groups, with all major cnidarian lineages already present 500 million years ago [3].

The Medusozoa clade is characterized by different evolutionary novelties, such as the presence of linear mitochondria and the adult pelagic stage, also known as medusa or jellyfish [4–6]. Most medusozoan life cycles are characterized by the succession of different stages, including a larval, benthic asexually reproducing polyp stage and a sexually reproducing jellyfish stage [6, 7]. This ancestral metagenic life cycle pattern is highly plastic and in some groups has been extensively modified or even lost. For example, several lineages have lost the pelagic medusae or reduced it to a reproductive structure, or acquired colonial lifestyles during the benthic phase [8–10]. Other novel traits have emerged in Medusozoa such as complex body patterns, neuromuscular systems, and sensory organs [11].

The history of Medusozoa genomics started with pioneer cytogenetics reports (e.g., [12, 13]) and was continued later by genome size estimations [14, 15]. Over the past 20 years, technological advances and cost reduction of genome-scale sequencing platforms have led to a steady increase in both the number and diversity of sequenced genomes and transcriptomes [16, 17]. Medusozoa is not an exception, as numerous genomic resources have become available for model and non-model species, especially in the past 3 years. This advance has enabled the study of the genetic basis of many Medusozoa novel traits (e.g., [18–22]. Previous reviews about cnidaria genomics have focused on the small number of species with sequenced genomes available at the time [11, 23, 24], on individual cnidarian lineages (i.e., Myxozoa [25]), or on specific topics such as toxins or evolution of novel traits [11, 26]. Given the increasing amount of genomic information available, an update of the state of Medusozoa genomics is warranted.

Here, we provide a comprehensive review of the major advances in Medusozoa genomics over the past century. To shed light on the understanding of the genomic evolution of the group from High-Throughput Sequencing (HTS) datasets, we report the main trends on the number and quality of available genome projects, taking into account basic information of sequencing datasets, genome assemblies, genome annotations, and accessibility of associated data and metadata.
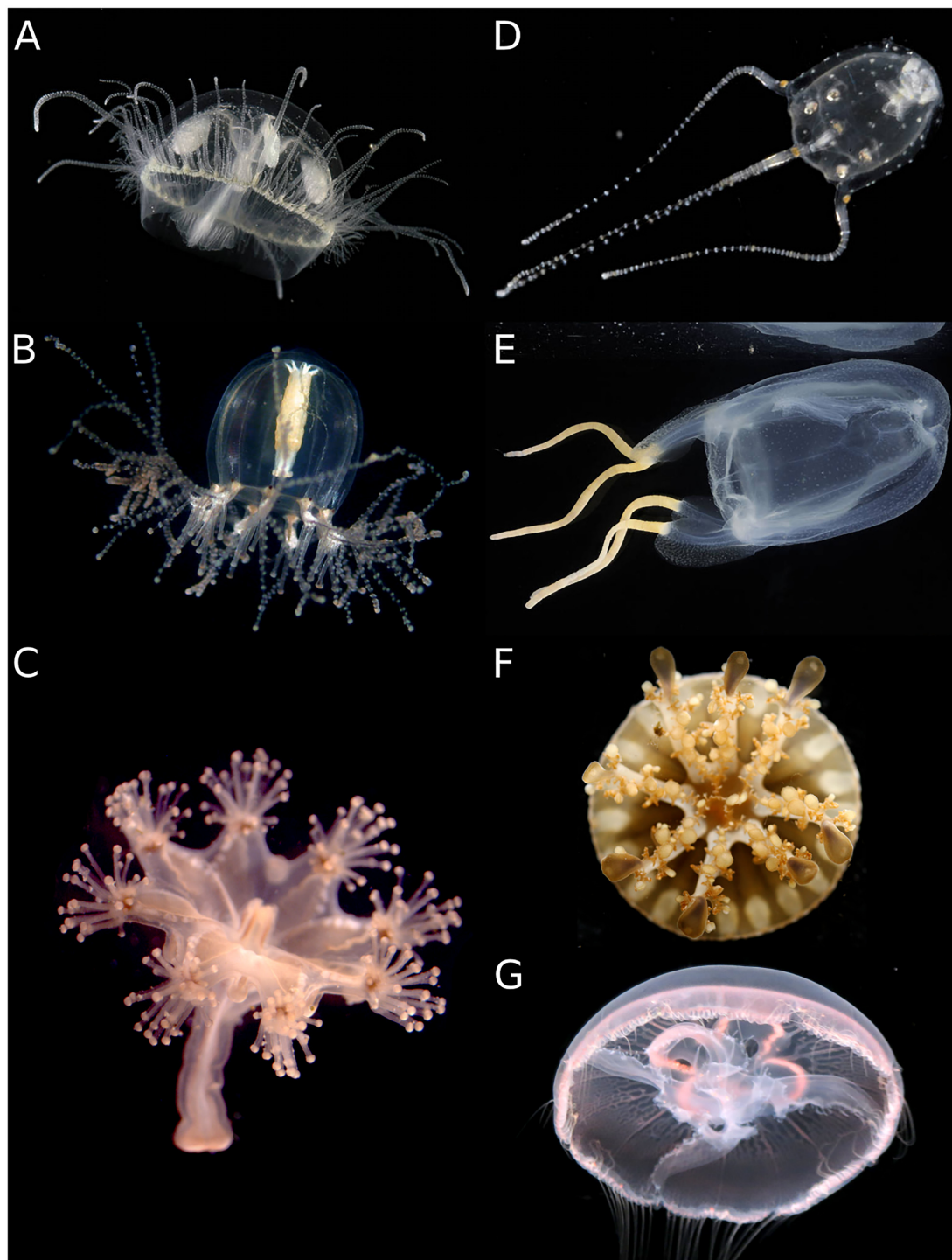
Figure 1: Medusozoa diversity. Examples of different genera covered by this review belong to Hydrozoa (A, B), Staurozoa (C), Cubozoa (D, E), and Scyphozoa (FG). (A) *Craspedacusta sowerbii,* (B) *Cladonema radiatum,* (C) *Haliclystus sanjuanensis,* (D) *Carybdea sivickisi,* (E) *Tamoya haplonema,* (F) *Cassiopea xamachana,* (G) *Aurelia aurita*. Credits to Alvaro E. Migotto (A, B, E), Marta Chiodin (C), Joseph Ryan (F, G), and Cheryl Ames Lewis (D). Photographs A, B, D, E were obtained from Cifonauta [118]. Photographs are not to scale.

Methods

We surveyed literature and databases for cytogenetic reports and genome size estimations. Our main source of genomic information and metadata was NCBI Genome (Assembly, Genomes, Nucleotide, Taxonomy, and SRA [27]). We retrieved data automat-

ically using entrez-direct v.13.9 and NCBI datasets v. 12.12. For information not present in NCBI, we checked published articles for proper information collection, as well as personal repositories mentioned in the associated articles. Owing to recent updates in taxonomic statuses, we modified the attribution of karyotypes,

genome sizes, and assemblies of several species (see main text and Supplementary Materials).

Because there have been subtle variations in metrics and statistics between most genome reports, we recalculated some statistics, allowing us to make meaningful comparisons. Briefly, we have generated the following: (i) assembly statistics using the statswrapper.sh script from BBmap v38.73 (BBmap, RRID:SCR_016 965) [28]; (ii) gene statistics from the original annotation files with AGAT v0.6.0 [29] and assessment of completeness of all assemblies using BUSCO v5.0.0+galaxy0 (BUSCO, RRID:SCR_015008) [30] in genome mode and Metaeuk software, using 2 Single Orthologs Databases (eukaryota_odb10, number of genes = 255, number of species = 70; metazoa_odb10, number of genes = 954, number of species = 65), available at the public Galaxy server [31, 32].

Assembly quality was reported following the metric proposed by Earth Biogenome Project [33] (hereafter BGP-metric). This system avoids the use of ambiguous terminology for quality and uses a logarithmic scale where the first 2 numbers are the exponents of the N50 contig and scaffold (1: 0–99 kb; 2: 1–9.9 Mb; 3: 10–99.9 Mb) and the third number corresponds to the level of chromosomal assembly (1: >90% DNA assigned to chromosomes *in silico*; 2: chromosomal rearrangements validated by 2 data sources; 3: >80% DNA assigned to intra-species maps and experimental validation of all breakpoints; see [33]).

All graphs were generated using Python v.3 with ETE Toolkit v.3 [34], Matplotlib v3.3.1 [35], and Seaborn v.0.11 [36] and modified with Inkscape v.0.92 [37] to improve visualization (e.g., font size and spacing). The tree of Figs 2 and 4 represents a simplified phylogenetic hypothesis obtained by combining phylogenies from previous studies (Scyphozoa [38], Medusozoa [5], Hydrozoa [39, 40]), taking into account clades with high congruence and support values. Although the different phylogenetic hypotheses were mostly congruent, no single study nor molecular dataset comprised all the terminals discussed here. We manually compiled all genomic information and HTS metadata referenced in this review using a report model based on previous works and public databases such as NCBI (Supplementary File S1 [29, 41, 42]). The command line used for retrieving genetic information and metadata, for statistics calculation, and the code used for graph generation are available at Supplementary Files S2 and S3. All collected data were updated until 1 May 2021.

Genomic Projects: Whos and Hows of Medusozoa

Chromosome numbers are known for 34 hydrozoan species and 5 scyphozoan, including 3 lineages of the *Aurelia aurita* sp. complex species ([12, 13, 21,43–51]; Supplementary File S4). Older chromosome descriptions for 25 species do not include information about chromosome morphology and often lack photographic records or schematic representations [12, 13, 43–47].

Genome size, a fundamental feature in any genome sequencing project, has been experimentally estimated by flow cytometry or Feulgen densitometry techniques for 24 medusozoan species (Scyphozoa: 7 spp.; Cubozoa: 1 sp.; Hydrozoa: 16 spp.; Supplementary File S4). Genome sizes are highly variable, ranging from 254 to 3,481.68 Mb in *Sanderia malayensis* (Scyphozoa) and *Agalma elegans* (Hydrozoa), respectively [15]. Moreover, an additional 12 genome size estimates are available when considering *k*-mer–based computational assessments, increasing the number of species with genome size information to 30, and including 2 cubozoans (913–2,673 Mb) and 1 staurozoan (230 Mb) (Supplementary Files S1 and

S4). These estimates are considered less accurate, especially for genomes with high heterozygosity, high repetitive content, and large genome size [52]. In fact, *k*-mer–based and experimental estimations from the same species differed by 13–33%.

A total of 34 HTS projects were identified. Of these, 32 had sequencing reads accessible through the NCBI-SRA database, but not all of them were associated with a genome assembly (Table 1, Supplementary File S1). The taxonomic coverage of the assemblies encompassed 7 of the 13 Medusozoa orders and represented ≥1 species per class (Fig. 2): 28 assemblies were accessible for 21 species, representing 0.5% of Medusozoa (Fig. 2, Table 1, Supplementary File S1). Of these 21 species, 12 were Scyphozoa, 4 were Hydrozoa, 4 were Cubozoa, and 1 was Staurozoa. Scyphozoa had the highest number of sequenced families (4 of 22), of which Pelagiidae contained the highest number of sequenced species so far (5 spp.), followed by Ulmaridae, Rhizostomatidae, and Cassiopeiidae with 2 spp. each (Fig. 2), all belonging to subclass Discomedusae (none from Coronamedusae). The remaining assemblies represent 3 of the 8 Cubozoa families and 3 of 135 Hydrozoan families (Fig. 2). In addition to the small fraction of family representation in the hydrozoan genomes, the underrepresentation of Leptothecata is particularly unfavorable because it harbors more than half of Medusozoa species (2,059 spp. [1]).

Much of the assembly effort is biased towards a small number of species. For example, 3 species of Hydrozoa and Scyphozoa presented 2 assemblies each, of which *Hydra viridissima* and *Rhopilema esculentum* were sequenced twice independently; meanwhile *Chrysoaora quinquecirrha* presents 2 versions of the same assembly. Moreover, 3 assemblies were available for 2 different strains of *Hydra vulgaris* (former *Hydra magnipapillata*), 1 of them published as an update of the reference genome called Hydra 2.0. In *Aurelia,* the genomes of 3 different lineages were sequenced and assembled: Baltic sea, Roscoff, and *Aurelia* sp1. strains [19, 20]. Based on a recent taxonomic update of this genus [53], locality and genetic information described in the original articles [19, 20], we decided to refer to these genomic datasets as Baltic sea strain = *Aurelia aurita*; Roscoff strain and *Aurelia* sp1. strains = *Aurelia coerulea*.

Most of the assemblies were deposited in the NCBI Assembly database, 1 was only found in a journal-specific database (i.e., GigaDB [54]), 1 assembly was only in a personal repository (Google Drive), and 1 in the National Human Genome Research Institute site [55]. Some assemblies were additionally deposited in institute-centered repositories such as OIST Marine Genomics Unit [56] and the Marine Invertebrate Models Database (MARIMBA [57]). A significant portion of the publicly available assemblies (total of 8, ~30%) are not yet associated with a formal publication and belong to the IRIDIAN GENOMES project [58]. The most frequent sequencing technology was Illumina (26 assemblies, ~93%), but leaving aside unpublished ones, most works include a combination of different sequencing techniques, library sizes, and platforms (i.e., Sanger, 454, Illumina, long reads, linked reads, and Hi-C sequencing; Supplementary File S1).

Almost all medusozoan genome assemblies were at draft contig or scaffold level, with 1 exception, *R. esculentum*, where chromosome-level scale assembly was reported [59]. The total length, contig and scaffold number, N50, and GC% varied across species and classes (Fig. 3A; references in Supplementary File S5). The assembly continuity and quality was higher in Scyphozoa than in the other classes, as observed by the distribution of contig and scaffold N50 (Fig. 3A) and the BGP-metric for assembly quality (Fig. 3A). In general, they are fragmented (75%) and have contig N50 of <40 kb (Fig. 3A; BGP-metric values of 0.0.0, 0.1.0, and 0.2.0).
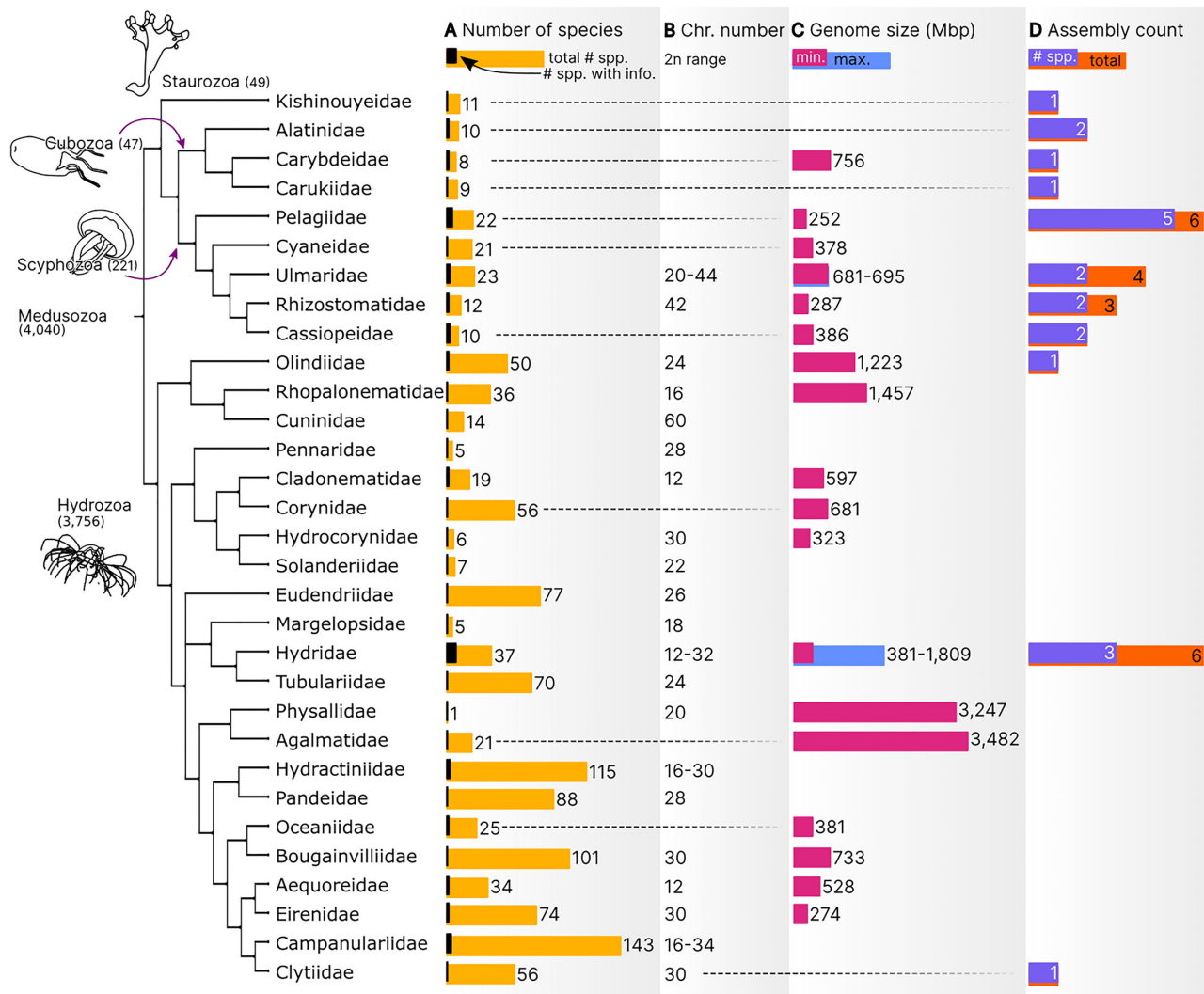
Figure 2: Phylogenetic distribution of genomic information in Medusozoa. (A) Number of described species and number of species with genomic data; (B) chromosome number (2n) range; (C) genome size (Mb) range taking into account flow cytometry and Feulgen densitometry estimations; (D) total number of available assemblies and number of species with assembled genomes. In (B) and (C) single values were also included when only 1 species was characterized. Tree topology is explained in the Methods section. Information used for this graph is available in Supplementary File S5 Table S2.

Staurozoa, Cubozoa, and Scyphozoa assemblies have similar percentages of base composition, ~35–43% GC. Consistent with previous reports [60], Hydrozoa genomes have a higher dispersion of GC%, with the GC values of 5 assemblies <35%.

In relation to gene content (Fig. 3B), 17 genomes were annotated using ≥1 source of information (Supplementary File S1) and their total number of genes or total number of protein-coding genes were reported. Further description of coding information was variable among works, and as more detailed information was considered, the number of genomes with reported information decreased. Annotation tracks and gene models were available for only 11 of the 17 datasets. Recalculations of gene features, together with the information recovered from original articles, allowed us to analyze the distribution of 5 different features in 15 genomes of Scyphozoa, Hydrozoa, and Cubozoa (Fig. 3B; Box): number of genes (n = 15), mean exons per CDS (n = 10), mean gene length (n = 11), mean exon length (n = 11), and mean intron length (n = 12). For 3 species, *Cassiopea xamachana* (Scyphozoa; 31,459), *Alatina alata* (Cubozoa; 66,156), and *Calvadosia cruxmeliten-*

sis (Staurozoa; 26,258), the available information was restricted to the number of predicted genes. Some small inconsistencies were detected between original data reported in some articles and our recalculations (Table S5 and S6), and others between data reported in the main text and supplementary materials of some articles.

The determination of repetitive DNA has been an integral step before gene annotation in most genomic projects. Frequently, repeat diversity was not properly reported and the degree of detail also varied between articles: e.g., some published works only referred to the most abundant class of repetitive DNA, meanwhile others described only results at class or family level. Repetitive libraries—consensus sequences representing repeat families—were not properly saved in repositories with the exception of 2 independent articles, and RepeatMasker results were reported in 4 articles (1 reporting only classified repeats). Total repetitive length of 12 species for which coding information was also available is presented in Fig. 3B and discussed in the Box.
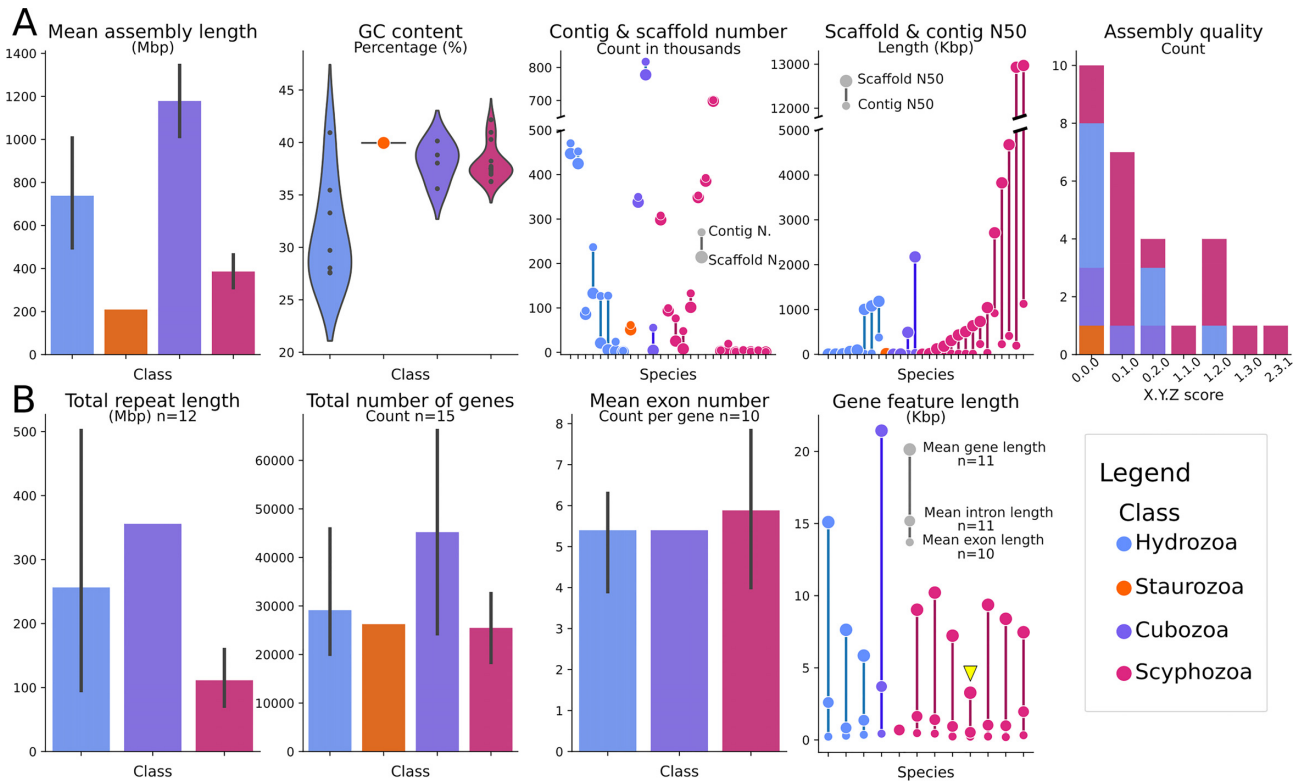
Figure 3: Assembly and genome features. (A) (from left to right): mean assembly length per class, GC content (%) per class, number of contigs and scaffolds per assembly colored by class, contig, and scaffold N50 (in kb) per assembly coloured by class, and count of assemblies of each class corresponding to the different BGP-metric values, where X and Y correspond to contig and scaffold N50, respectively, and Z to chromosome assignment (see Methods section). (B) (from left to right): mean repeat length (Mb) in assembly per class, mean total number of genes per class, mean exon number (count per gene) per class, and mean gene, intron, and exon length (kb) per assembly colored by class. The yellow arrowhead indicates *S. malayensis* gene features (see Box). Within the dotplots, a data point is indicated for each species. When more than one species per class was available, vertical lines were added to barplots to indicate the value dispersion around the mean. All other keys are specified in the figure. Information used for this graph is available in Supplementary File S5 Tables S4–S6.

The degree of completeness of these datasets also varied substantially, as estimated by BUSCO (metazoa_odb10 and eukaryota_odb10; Fig. 4). While all Eukaryota genes were present in ≥1 assembly (Supplementary Files S5 and S6), the level of absence and fragmentation of Metazoa genes was higher (Fig. 4; Supplementary File S5). Seven Metazoa genes were absent in all assemblies and 17 were absent in >20% of them (Fig. 4, indicated in red). Some Metazoa BUSCO genes were absent in lineages with the higher number of assemblies, such as Scyphozoa and Hydrozoa (Fig. 4, indicated in yellow rectangles; Supplementary File S5). This condition was suggested by [20], after detecting the absence of 14 genes in 5 species (version metazoa_o9db), 3 of which coincided with the genes detected as absent here (Orthodb IDs: 460044at33208, 601886at33208, 114954at33208), 1 of which (445034at33208) has a patchy distribution in Medusozoa and 9 of which were removed in later versions of the database (Fig. 4 in boldface).

Moreover, 27 genes were simultaneously recovered as undetectable or fragmented in >80% of the assemblies (Supplementary File S5 Table S7). Based on BUSCO completeness assessment with metazoa_o10db, 13 assemblies present 90–95% of genes (fragmented + complete), while only 1 assembly includes >90% of complete genes; the remaining 15 assemblies present between 57% and 87% of genes (complete + fragmented) or 16–77% com-

plete genes. While the Metazoa database might include genes that are absent, fragmented, or have non-conventional features in all medusozoa species, the utility of the Eukaryota database in the completeness assessment is limited by its low number of genes. Until more specific databases are developed, the combination of both BUSCO databases should be used taking into account their limitations.

Differences in sequencing strategy and platforms are expected to be linked with assembly quality, in terms of both continuity and completeness. For example, hybrid sequencing plus proximity ligation maps and combined evidence-based annotation should generate better results than a short-read sequencing and single-evidence annotation [61, 62]. Although this general trend was observed in this review, with most Illumina-only datasets showing lower BGP-metric (Fig. 3) and lower completeness (Fig. 4), it is not a granted condition. Certain specific cases can exemplify biological and methodological issues that impose limitations to genome sequencing and assembly: e.g., the difficulty in obtaining chromosome-scale assemblies despite small genome sizes and combined sequencing strategies (Hi-C + short reads + long reads) [63, 64] or the difficulty in extracting high molecular weight DNA [20]. Because of the heterogeneity of Medusozoa genomic projects in terms of time periods, objectives, methods, and resources, a proper quantitative analysis of the relationship be-

Table 1: Genomic projects related to Medusozoa HTS

Project	Release year (NCBI-SRA)	Class (No. genomes)	Species	Main research topics
Chapman et al. [76]	2008	Hydrozoa (1)	*Hydra vulgaris*	Gene evolution; micro-synteny
IISER Pune	2014–2015	Hydrozoa (1)	*Hydra vulgaris*	not_informed
NHGRI [55]	No SRA	Hydrozoa (1)	*Hydra vulgaris*	not_informed
NHGRI [116]	2016	Hydrozoa (1)	*Hydractinia echinata**	not_informed
Gold et al. [19]	2018	Scyphozoa (1)	*Aurelia coerulea*	Life cycle; gene evolution; intraspecies variability; HOX
IRIDIAN GENOMES [58]	2018	Hydrozoa (1)	*Craspedacusta sowerbii*	not_informed
Kim et al. [73]	2018	Scyphozoa (1)	*Nemopilema nomurai*	Life cycle; jellyfish body patterning; gene evolution; toxins
IRIDIAN GENOMES [58]	2019	Hydrozoa (1)	*Scolionema suvaense*	not_informed
Khalturin et al. [20]	2019	Scyphozoa (1)	*Aurelia aurita***, *Aurelia coerulea***	Life cycle; jellyfish body plan; gene evolution; synteny
		Cubozoa (1)	*Morbakka virulenta*	
Leclère et al. [21]	2019	Hydrozoa (1)	*Clytia hemisphaerica*	Life cycle; gene evolution; micro-synteny; TF
Odhera et al. [22]	2019	Scyphozoa (1)	*Cassiopea xamachana*	Gene evolution; micro-synteny; Homeobox; toxins
		Cubozoa (1)	*Alatina alata*	
		Staurozoa (1)	*Calvadosia cruxmelitensis*	
Vogg et al. [117]	2019	Hydrozoa (1)	*Hydra oligactis*; *Hydra viridissima*	Gene evolution; RTKs; developmental genes
Hamada et al. [72]	2020	Hydrozoa (1)	*Hydra viridissima*	Symbiosis; immune response; repetitive DNA; Homeobox
IRIDIAN GENOMES [58]	2020	Cubozoa (1)	*Alatinidae sp.*	not_informed
			Carybdea marsupialis	
			Tamoya ohboya	
		Hydrozoa (2)	*Cladonema radiatum*	
			Eutima sp. BMK-2020	
		Scyphozoa (4)	*Aurelia coerulea*	
			Chrysaora achlyos	
			Chrysaora chesapeakei	
			Chrysaora fuscescens	
		Staurozoa (1)	*Calvadosia cruxmelitensis*	
Li et al. [59]	2020	Scyphozoa (1)	*Rhopilema esculentum*	Gene evolution; toxins
Nong et al. [63]	2020	Scyphozoa (2)	*Sanderia malayensis*, *Rhopilema esculentum*	Gene evolution; small RNAs; micro-synteny; Homeobox
Xia et al. [74]	2020	Scyphozoa (1)	*Chrysaora quinquecirrha*	Gene and gene feature evolution; repetitive DNA
Xia et al. [64]	2020	Scyphozoa (1)	*Chrysaora quinquecirrha*	Assembly improvement report
UMCG	2021	Scyphozoa (1)	*Cassiopea andromeda*	not_informed

Sequencing projects with no current related publication are remarked with capital letters. Column "Main research topics" describes keywords according to references, restricted to a maximum of 4; "gene evolution" refers to the study of gene gains/losses and also of specific gene families. Species with reported assemblies were re-analyzed in this review (boldface; Supplementary File S5 Table S3). IISER PRune: Indian Institute of Science Education and Research, Pune; NHGRI: National Human Genome Research Institute; RTK: receptor tyrosine kinase; TF: transcription factors; UMCG: University Medical Center Groningen. *"preliminary" assembly available at the institutional site; **species with taxonomic updates. For further details see Supplementary File S1.

tween methods and outcome quality would not be feasible, and we prefer to refer to articles specialized in assessing methods (e.g., [61, 62]).

The State of Medusozoa Genomics: Inner and Derived Knowledge

The first glimpse of the Medusozoa genomic organization was obtained by cytogenetic studies [12, 13, 21, 43–51], but in contrast to other animals, the available information is still sparse. Many cytogenetic questions essential to the understanding of genome evolution are unanswered in Medusozoa, either at species or population scale, including the distribution of the chromosome number (2n), fundamental number of chromosome arms (FN), genome size, ploidy level, and heterochromatin content. These are questions that have gained renewed interest since the arrival of the genomic era.

Regarding the phylogenetic distribution of the chromosome number, no inferences can yet be made on the sparse available information, apart from the presence of some chromosome vari-
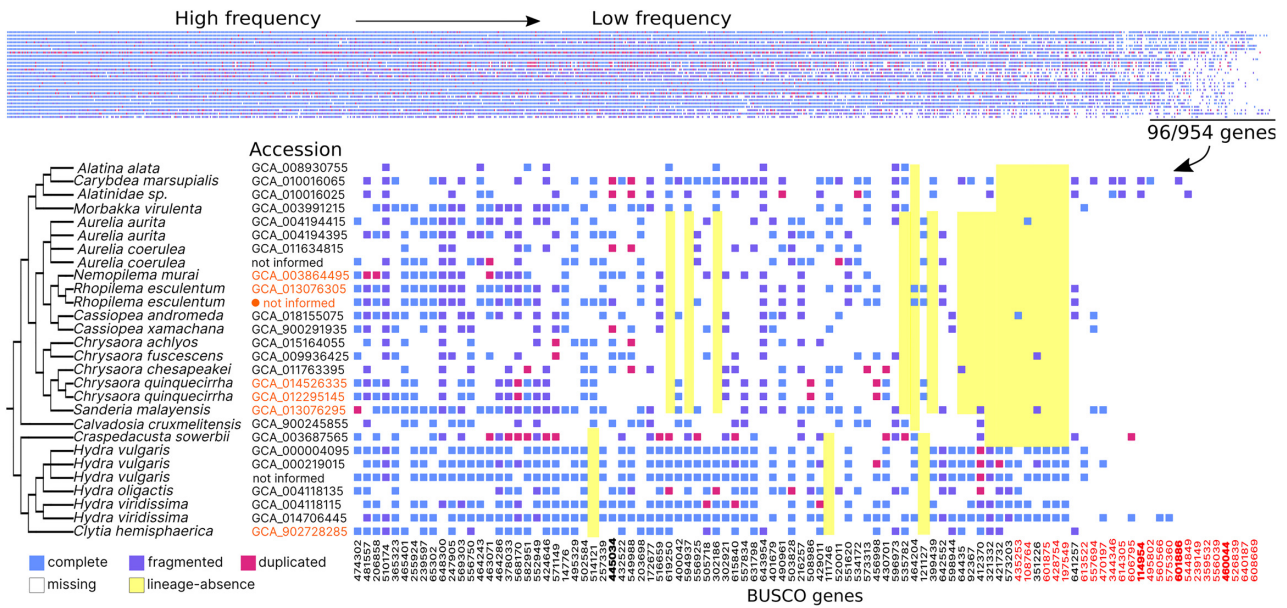
Figure 4: BUSCO Metazoa gene distribution in Medusozoa assemblies. Each column corresponds to a gene and each row an assembly. Columns were ordered based on presence from left to right and the least present genes (n = 96) are shown in detail. Genes absent in all or almost all assemblies (>80% of absence) are indicated in red; genes also reported absent [20] are indicated in boldface; genes absent in specific lineages are indicated with yellow rectangles. Higher quality assemblies are indicated in orange (BGP-metric > 1.0.0). The assembly with the highest quality score for BGP-metric is indicated by an orange circle and corresponds to *Rhopilema esculentum* [59]. Information used for this graph and full BUSCO gene names are available in Supplementary File S5 Table S7.

ation throughout Medusozoa. A special case was reported in *Hydra,* where, according to recent descriptions, many species shared a 2n = 30 karyotype with metacentric or submetacentric chromosomes ([51]; Supplementary File S4). This suggests that the 2n = 30 karyotype could be widely distributed in the genus and even in other Hydrozoa groups because it was also described for 1 species of Hydrocorynidae, Hydractiniidae, Campanulariidae, Bougainvilliidae, and Clytiidae, and 3 Eirenidae (Supplementary File S4; references therein). Interestingly, in Anthozoa, a few sea anemones and several scleractinian corals have karyotypes between 2n = 28 and 2n = 30 [65–67]. Nevertheless, a higher sampling effort should be conducted to test the extent of this apparent karyotype stability.

Scyphozoa genomes tend to be smaller (∼250 to ∼700 Mb) than those of Hydrozoa, which encompass a larger range (∼380 to ∼3,500 Mb) (Fig. 2; Supplementary File S4, references therein), but owing to the scarcity of estimations that represent ∼1% of the subphylum, these ranges should be considered preliminary. The evolution of eukaryotic genome size is a long-standing question that has been called the "C-value enigma" [41]. This name stems from the difficulty elucidating the evolutionary forces (e.g., drift and natural selection) that have given rise and serve to maintain variations in genome size, the mechanisms of genome size change, and the consequences of these variations at an organismal level [41]. Several conflicting hypotheses have been postulated to explain this puzzle, with most having experimental support in some but not all lineages (reviewed in [68]). The molecular basis of these variations in Medusozoa has only been studied in detail for *Hydra* [69] and for *S. malayensis* [63]; their trends have been related to repetitive DNA and gene length, respectively (Box). Meanwhile, the ecological and historical factors underlying genome size diversity and its extent in Medusozoa are topics that remain to be elucidated.

- **Gene content and length:** it is straightforward to imagine that the evolution of these 2 characteristics has potential impacts in macroevolution of organisms. The distribution of gene number in Medusozoa (Fig. 3B) ranged from 17,219 in the Scyphozoan *Rhopilema esculentum* [59] to 66,156 in the Cubozoan *Alatina alata* [22], which is higher than the range (18,943 ± 451.82) described for animals [41]; however, most species of all classes have gene counts near the median (26,258). The upper limit described in the highly fragmented *A. alata* genome deviates from that observed in *Morbakka virulenta* (24,278 genes), the only other sequenced Cubomedusae [20, 22]. Species with varying genome sizes of Hydrozoa, Scyphozoa, and *M. virulenta* (Cubozoa) had similar mean CDS lengths (1,414, 1,214, 1,387 bp), mean numbers of exons per gene (5, 6, 5.4), and mean exon lengths (306, 293, 432 bp) but had different gene lengths (9,530, 7,855, and 21,444 bp, respectively) owing to the presence of longer introns in Hydrozoa and Cubozoa when compared to Scyphozoa (Hydrozoa: 1,600; Cubozoa: 3,705 vs 1,146 bp in Scyphozoa). This is best exemplified in the genome of the scyphozoan *S. malayensis*, which has the smallest cnidarian genome reported to date [63] and has also the smallest introns of any sequenced medusozoan genome (Fig. 3B, yellow arrowhead). Nevertheless, these ranges are rough estimates and sometimes heterogeneous, e.g., resulting from different filtering parameters, and their implications should be tested as new assemblies and annotations become available.

- **Repetitive content:** repetitive DNA represents a significant part of eukaryotic genomes and is highly diverse, composed by different kinds of transposable elements

(TEs), tandem repeats, and multigene families (e.g., rRNA and tRNA). Many of these sequences, especially TEs and satellite DNA, were initially considered as an expendable sector of the genome, although their impact on genomic evolution has since been recognized (reviewed in [70]). For example, fusions between TEs and host genes have occurred multiple times in vertebrates and have contributed to the evolution of novel features [71]. Likewise, TEs and other repetitive DNA have been associated with genomic rearrangements and changes in DNA content (e.g., [69, 70]). The *Hydra* genus, which has been more extensively studied from this point of view, has experienced a rapid genomic evolutionary rate and presents a 3-fold genome size increase resulting from the amplification of a single long interspersed nuclear element family [69]. Moreover, *Hydra* genomes include an overrepresentation of transposase-related domains [72]. It is interesting to note that many of the Medusozoa species studied so far have relatively small genomes but unusually high proportions of repetitive DNA [20, 63, 73, 74]. Nevertheless, the lack of standardization in the description of its diversity, and the discrepancy in the degree of detail in which these have been described, limits the potential to make inferences. Repetitive DNA is a complex study subject, limited by assembly continuity and annotation effort, but restricting genomic studies to the "functional" part of the genome (sensu [75]) may lead us to a narrowed view of the Medusozoa genome evolution.

Modern Medusozoa genomics formally started with the sequencing and publication of the *H. vulgaris* genome, which in Cnidaria was only preceded by *Nematostella vectensis* [65, 76]. *H. vulgaris* is one of the earliest models in biology, mainly used for the study of development, regeneration, and more recently, of aging (reviewed in [77, 78]). The study of these 2 early genomes was fundamental for the reconstruction of a more complex ancient eumetazoan genome than first suggested by the comparison of vertebrates and insects [16, 23, 65, 76].

Unlike most other medusozoan species, *Hydra* lives in fresh water, lacks a medusa, and has a genome that has experienced a very rapid rate of evolution [21]. It therefore is not the ideal species for reconstructing historical nodes on the Medusozoa tree of life. As such, more recent medusozoa genomes have led to important updates in our understanding of Medusozoa-relevant research topics, including phylogenetic reconstructions, the genetic basis of the medusae, the evolution of symbiosis, toxin characterization, and Homeobox gene evolution, to name a few examples (Table 1). Nevertheless, Medusozoa genomes include thousands of single-copy genes and repetitive elements; however, only a limited number of them have been analyzed in detail.

The complex nature of Medusozoa venom has been investigated by a number of transcriptomic, proteomic, and genomic studies (reviewed in [26]). Several putative toxin genes and domains have been identified, covering a significant part of the wide range of known toxins [20, 22, 59, 73]. In Scyphozoa, toxin-like genes were often recovered as multicopy sets [20, 59]. Moreover, in *R. esculentum* toxin-like genes were also tandemly arranged and several of them were located nearby in chromosome 7, suggesting that the observed organization might influence toxin co-expression [59]. Minicollagens, which are major components of nematocysts, also had a clustered organization and a pattern of

co-expression in *Aurelia* [20]. These examples add to various clustered genes described in Cubozoa, Hydrozoa, and Anthozoa and would indicate that gene clustering and operon-like expression of toxin genes is widespread in Cnidaria ([20] and references therein).

The determination of lineage-specific genes and increases and decreases of gene content is one of the recurrent questions found in Medusozoa genomic studies (e.g., [20, 21]), and it has been conducted using different methodologies and sets of species. Recent evidence proved that the detection of lineage-specific genes, and other analyses relying on accurate annotation and orthology prediction, can be significantly biased by methodological artifacts [79–83]; several problems have been identified, such as low taxon sampling, heterogeneous gene predictions, and failure of detecting distant homology and fast-evolving orthologues. These considerations are highly relevant in Medusozoa because comparisons are often made, by necessity, with distantly related species (e.g., Anthozoa has been estimated to have diverged from Medusozoa ∼800 million years ago [84]). In Cnidaria, the most elevated rates of loss have been estimated in the hydrozoan branch leading to *Clytia hemisphaerica* and *Hydra* [21, 76], followed by slightly lower rates of gene loss in Scyphozoa and substantially lower rates in Anthozoa [19]. Gene families that have experienced expansion and contraction have been studied in relation to complex life cycle patterns [19, 21], simplification of the body plan [72, 76], and the evolution of symbiosis [72], among others (Table 1). Expression patterns of identified taxonomically restricted medusozoan genes have been mainly studied in the context of life cycle stages (e.g., [20, 21]).

The complex life cycle of Medusozoa has resulted from the combination of both ancestral and novel features. *Aurelia, Morbakka virulenta,* and *C. hemisphaerica* have significantly different patterns of gene expression across stages and during transitions [19–21]. Differentially expressed genes include many conserved ancestral families of transcription factors [19–21]; there is also a considerable amount of the putative lineage-restricted genes that show differential expression in the adult stages [20, 21]. A few of these "novel" medusozoan genes have been described, such as novel myosin-tail proteins that are absent from Anthozoa and represent markers of the medusae striated muscles [20]. It was suggested that the evolution of the Medusozoa complex life cycle would therefore have involved the rewiring of regulatory pathways of ancestral genes and the contribution of new ones [19–21]. As such, the body plan and life cycle simplifications observed in Clytia and Hydra, respectively, would be the result of loss of transcription factors involved in their development [21]. Finally, the significance of many of the putative Medusozoa and species-specific genes remains to be elucidated.

On the other hand, synteny was also analyzed several times, including species of Hydrozoa, Cubozoa, and Scyphozoa, and analyses were carried on at different scales depending on assembly continuity (i.e., microsynteny and macrosynteny), and often comparing the focus species to species from sister clade Anthozoa [19–21, 67, 76]. High synteny conservation was found within Anthozoa (*N. vectensis* vs *Scolanthus callimorphus* [65–67]) and within Hydrozoa (*H. vulgaris* vs *C. hemisphaerica* [21]). Meanwhile, conservation of synteny at a lesser degree was also observed between Anthozoa and Scyphozoa (*N. vectensis* vs *R. esculentum*; *N. vectensis* vs *Aurelia* strains [19, 20, 67]) and only a few shared syntenic blocks between Hydozoa and Anthozoa (*H. vulgaris* vs *N. vectensis* [21, 67, 76]), Hydrozoa and Scyphozoa (*H. vulgaris* vs *A. aurita* [19]), and Scyphozoa and Cubozoa (*A. aurita* vs *M. virulenta* [20]). It is particularly interesting to note that *H. vulgaris, N. vectensis*, and *S. callimorphus* present 2n = 30 but shared fewer syntenic blocks than either of

the 2 anthozoans with *R. esculentum*, which has a different karyotype (2n = 22) ([67] [non peer-reviewed]). These results suggest that there is evidence for the conservation of an ancient genome architecture in Anthozoa and Scyphozoa, but less conservation in Hydrozoa and Cubozoa, coincident with a more rapid rate of genome reorganization in the last 2 classes [21, 67].

Prospects on Genomic Data and General Resources

The increasing amount of genomic information available for diverse organisms has enabled statistical inferences of trends in eukaryotic genomic evolution. Examples of such studies are available at small and large phylogenetic scales and have enabled evolutionary analyses of the distribution of gene numbers, gene features (e.g., intron size), and repetitive content (e.g., [41]). Nevertheless, the power of eukaryotic genomic comparative analyses is hindered by a lack of data and metadata standardization [41, 85], which is especially evident in Medusozoa.

There is much to learn from decades-old references of cytogenetic studies, but some studies, especially older ones, lack complete material and methods (e.g., pretreatment, references, designs and photographs; general metadata as locality, taxonomic identification) and therefore should be considered carefully in a comparative framework (e.g., [86]).

Similar problems can be expected in relation to genomic data because metadata are often not specified in great detail. We analyzed hundreds of fields including genetic information and metadata (methods, metrics and registry codes; table in Supplementary File S1), of which no dataset presents most of them, whatever the area or section (e.g., processing area, section trimming). This could be a future problem because reusing previously published datasets is becoming routine, and tracking of information (e.g., BioProjects, Biosamples, methodologies, filtering parameters) would be misleading [85, 87].

Descriptions of bioinformatic methods in genome studies are often even less comprehensive than database metadata. For example, we identified ≥3 independent projects, each of which applied different criteria for gene model filtering, and another 3 articles applied slightly different criteria for repeat library filtering (Supplementary File S1). Although differences at this stage can seem small on the surface, they can result in hard-to-detect biases downstream that can lead to flawed biological conclusions. For example, resistance genes have been underestimated in some flowering plant genomes owing to inconsistencies of genome annotation stemming from differences in repeat masking [88]. Likewise, in the present review, we identify discrepancies in BUSCO genome completeness comparisons that are caused by differences in database versions, which are frequently unspecified in the associated articles.

An alternative solution for comprehensive comparative analyses is to (re)annotate all genomes with the same pipeline, a task that is laborious and time consuming. Some programs were designed for achieving this task simultaneously in many related species (e.g., [89, 90]). Another alternative is to use specific software developed to improve genome annotations by leveraging data from multiple species (e.g., [91, 92]) or targeting specific gene families [93, 94]. Finally, differences in annotation due to methodological artifacts can be accommodated in comparative analysis if considered as a variable in the statistical tests (e.g., comparing tRNA genes in high- and low-quality avian genomes [95]).

The submission of raw sequencing data and fundamental metadata to the NCBI-SRA or EMBL-ENA remains a vital step in ensuring the usability and transparency of genome data [96, 97]. Also, project-centric repositories serve to store assemblies and associated datasets, and enable comparative studies by basic tools. Taxon-restricted databases including cnidarian data have been used in the past, but these are often not maintained owing to lack of upkeep funding and other factors (e.g., [98, 99]). In addition, submission to the large databases like SRA and GenBank can lead to the automatic detection of specific issues such as contamination or annotation errors that might otherwise not be detected. For these reasons, the large general databases should remain the primary repositories for sequence and metadata [100]. Nevertheless, this is not always the case. For example, the assembly with the highest continuity as estimated by the BGP-metric, corresponding to *R. esculentum* [59], is only found in a journal-specific database and lacks a stable identifier (e.g., NCBI accession). A similar situation is observed for 1 of the *H. vulgaris* assemblies (Hydra 2.0), which is only found in a project-specific database [55].

There is a growing number of community-driven guidelines, standards, databases, and resources based on the Findable, Accessible, Interoperable, and Reusable principles (FAIR principles) for digital research outputs [100]. Furthermore, global initiatives of large-scale genome sequencing included in Earth Biogenome Project have adopted a set of standardized protocols for the different stages of the genome projects, such as specimen collection, DNA extraction, sequencing, assembly and annotation methods, and reporting, in order to generate datasets that could "be useful to the broadest possible scientific community" [33]. Standards should also be implemented by independent research groups publishing genomes. The main goal of standardization is to promote evaluation, discovery, and reuse of genomic information, providing long-term benefits for science.

The following are suggestions to enhance genome projects and outcomes, and to promote open and collaborative research. These suggestions can be broadly applied to any genome project and are in line with those proposed by many initiatives and consortia (e.g., [33, 100, 101]). Nevertheless, it is worth reinforcing and discussing them in the context of this review because genome projects are more and more often being initiated in research laboratories that have historically been more focused on other aspects of medusozoan biology and may not be as familiar with these general practices:

1. Deposit all data and metadata in public specialized databases (e.g., NCBI), at least once associated articles are accepted for publication. Provide comprehensive metadata, including those not considered as priority for the aforementioned project. Frequently, data and metadata that are described in the original articles or deposited in repositories are not submitted to public databases. Tracking information from multiple sources is time consuming and prone to error. Databases and repositories enable the improvement of metadata after the initial releases, by the addition of new or corrected information (e.g., publication information) from the authors. We believe that this kind of data curation would improve the state of Medusozoa genomics not only by enabling downstream analysis after the publication but also by enabling the detection of methodological options (e.g., tissue selection; sequencing technology) that would improve the quality of the results.

2. Consider providing standardized genome statistics in an easily accessible format (e.g., Supplementary File S1 presented here). Alternatively, use specialized tools that standardize reports for multiple samples and datasets (e.g., [42,

102, 103]). This will facilitate meta-analyses, prompt new genome studies to make accurate comparisons to previously published studies, and prevent the propagation of erroneous information.

3. Deposit output results that were fundamental in any of the steps of the analysis (e.g., gene models, repetitive libraries, and annotation tracks). A Medusozoa-centric database with long-term maintenance is still lacking for the community (e.g., Mollusca clade [104]), but many open repositories can serve this purpose with low or no costs considering the size of the aforementioned outputs. There are open topic-centric repositories (e.g., Dfam [105] for repetitive DNA), general repositories (e.g., FigShare, Zenodo; or even NCBI for annotation tracks), as well as personal or institutional ones. Many of the reviewed genomic projects already made use of these repositories but failed to deposit some of the outputs. A solution for this inconvenience is to update submissions or create novel ones (e.g., submit annotations to NCBI or ENA) to deposit the missing outputs.

4. Inform as much as possible if a dataset was edited (e.g., removal of exogenous DNA; gene and repetitive sequence filtering criteria).

5. Use and clearly identify software, database versions, and references in all instances (e.g., RRID, BUSCO version, and repetitive database version).

6. Deposit command lines and scripts used to handle data (from reads to full annotation).

The latter suggestions (3–6) are mainly related to providing detailed methodologies of bioinformatic analyses. First, proper method and results descriptions can help to recover metadata and criteria usually not available in large sequence repositories. Second, comparative analyses depend upon standardization at different levels and significant sample sizes. The inclusion of species in downstream analyses is limited by data availability and proper description of previous analyses, custom software, and results.

7. Engage in community-driven conversations about standards, guidelines, and species priorities. There are a number of taxon-specific meetings that would be appropriate venues to engage in these conversations including the International Conference on Coelenterate Biology (∼decennial [106]), the International Jellyfish Blooms Symposium (∼triennial), Cnidofest (∼biennial [107]), Tutzing workshop (∼biennial [108]), and Cnidofest zoom seminar series. In addition, satellite meetings at larger annual meetings (e.g., the Society for Integrative and Comparative Biology [SICB] or the Global Invertebrate Genomics Alliance [GIGA] [101]) could provide appropriate venues to facilitate discussions on how the community can best move forward as more and more genomic data come online.

The adoption of best practices in the Medusozoa genomics community will pave the way for major breakthroughs regarding understanding the genomic basis for several evolutionary innovations that arose within and in the stem lineage of Medusozoa. Similar advances were achieved with extensive taxon sampling at broader scales, where 25 novel core gene groups enriched in regulatory functions might be underlying the emergence of animals [109, 110]. Medusozoa innovations have puzzled the community for decades [5, 7, 11, 111] and include the origin of the medusa, the loss of polyp structures, the establishment of symbiosis, the blooming potential, and the evolution of an extremely potent venom. A deeper understanding of the genomic events driving these innovations will require accurate identifications of a number of key genomic features including (but not limited to) single-copy orthologs, gene losses, lineage-specific genes, gene family expansions, and non-coding regulatory sequences.

Conclusions

The pace of genomic development in Medusozoa is far more rapid than more traditional disciplines such as cytogenetics, where gaps still remain. As the effect of chromosome structural variants in evolution is increasingly tested and recognized, it is expected that these disciplines will gain a revived interest as has been seen in other animal groups [112]. In spite of the great advances in Medusozoa genomics, we found a general lack of standardization in methodologies and genome reports across independent sequencing projects. Efforts to incorporate standards would benefit future studies and could promote the identification of hitherto undiscovered evolutionary patterns.

It is safe to anticipate that standardization will become increasingly easier as chromosome-level assemblies become more commonplace and as new integrated workflows of data reporting and submission are developed (e.g., [113]). It will be possible to perform standardized annotation and analyses to identify patterns in medusozoa genome evolution.

The distribution of genetic and genomic information presented significant taxonomic gaps in Medusozoa. It is a reasonable scenario because genomic sequencing data are accumulating in many medusozoan lineages. Even so, some of the most species-rich clades with a diverse array of phenotypic and ecological traits have not yet had their genomes sequenced (e.g., Scyphozoa:Coronamedusae, Hydrozoa:Macrocolonia). These, and other, heretofore genomically underexplored lineages provide golden opportunities from which to make major contributions to understanding the evolution of Medusozoa genomes and would be a wonderful contribution to the rest of the Medusozoa research community. Defining candidate species for sequencing can avoid unnecessary doubled efforts. Different international projects recognized this situation and proposed a set of criteria for prioritizing species at other scales, such as the GIGA ([101]).

Conversations about how best to promote such efforts and best practices for medusozoan genomics will help move the field forward. Such conversations could lead to new standards and potentially a powerful cnidarian genomics database. This latter goal would be most effective if accompanied by a strong alliance that spans the growing cnidarian genomics community.

Data Availability

All collected information, outputs, and scripts supporting new results are available in the Supplementary Files S1–S9 in Figshare [114]. All genomic resources from previous articles and projects are publicly available and their sources are referenced in Supplementary File S4 Table S3. The most up-to-date copy of Table S1 is available in Figshare [114] and can be updated upon the original author's request.

Additional Files

Supplementary File S1. Dataset 1. Genome report sheet.
Supplementary File S2. Dataset 2. Command line to retrieve data from NCBI and to generate new results.
Supplementary File S3. Dataset 3. Scripts used for graph construction.

Supplementary File S4. Table S1. Species information considering chromosome number, genome size, and genomic datasets.

Supplementary File S5. Tables S2–S8. All information used for constructing graphs presented in this work. Includes summary information of Fig. 2 (Table S2), genome resources used in this study (Table S3), assembly statistics for Fig. 3A (Table S4), genome features of Fig. 3B (Tables S5 and S6), and BUSCO results for Fig. 4 and Supplementary Fig. S1 (Tables S7 and S8).

Supplementary File S6. Figure S1. BUSCO Eukaryota gene distribution in Medusozoa assemblies. Each column corresponds to a gene and each row an assembly. Information used for this graph is available in Supplementary File S5 Table S8.

Supplementary File S7. Dataset 4. Original metadata from NCBI.

Supplementary File S8. Dataset 5. Original results from AGAT and Galaxy server (BUSCO).

Supplementary File S9. Dataset 6. Figures in vectorial format.

Abbreviations

BUSCO: Benchmarking Universal Single-Copy Orthologs; CDS: coding sequence; HTS: high-throughput sequencing; kb: kilobase pairs; Mb: megabase pairs; NCBI: National Center for Biotechnology Information; rRNA: ribosomal RNA; SRA: Sequence Read Archive; tRNA: transfer RNA.

Competing Interests

The authors declare that they have no competing interests.

Funding

Authors' Contributions

M.D.S. collected the information, ran the analysis, conceived the study, and drafted the manuscript; M.M.M. collected the information, conceived the study, and drafted and reviewed the manuscript; J.F.R. drafted and reviewed the manuscript; S.C.S.A. conceived the study and drafted and reviewed the manuscript. All authors gave final approval for publication.

Acknowledgments

References

1. World Register of Marine Species. Cnidaria. http://www.marinespecies.org/aphia.php?p=taxdetails&id=1267. Accessed 24 November 2021.

2. Bosch, TCG, Adamska, M, Augustin, R, *et al*. How do environmental factors influence life cycles and development? An experimental framework for early-diverging metazoans. *Bioessays* 2014;**36**(12):1185–94.

3. Cartwright, P, Collins, AG. Fossils and phylogenies: integrating multiple lines of evidence to investigate the origin of early major metazoan lineages. *Integr Comp Biol* 2007;**47**(5):744–51.

4. Bridge, D, Cunningham, CW, Schierwater, B, *et al*. Class-level relationships in the phylum Cnidaria: evidence from mitochondrial genome structure. *Proc Natl Acad Sci U S A* 1992;**89**(18):8750–3.

5. Kayal, E, Bentlage, B, Pankey, MS, *et al*. Phylogenomics provides a robust topology of the major cnidarian lineages and insights on the origins of key organismal traits. *BMC Evol Biol* 2018;**18**:68.

6. Marques, AC, Collins, AG. Cladistic analysis of Medusozoa and cnidarian evolution. *Invertebr Biol* 2004;**123**:23–42.

7. Collins, AG. Phylogeny of Medusozoa and the evolution of cnidarian life cycles. *J Evol Biol* 2002;**15**(3):418–32.

8. Boero, F, Bouillon, J. Zoogeography and life cycle patterns of Mediterranean hydromedusae (Cnidaria). *Biol J Linn Soc* 1993;**48**(3):239–66.

9. Da Silveira, FL, Morandini, AC. *Nausithoe aurea* n. sp. (Scyphozoa: Coronatae: Nausithoidae), a species with two pathways of reproduction after strobilation: sexual and asexual. *Contrib Zool* 1997;**66**:235–46.

10. Straehler-Pohl, I, Jarms, G. Morphology and life cycle of *Carybdea morandinii*, sp. nov. (Cnidaria), a cubozoan with zooxanthellae and peculiar polyp anatomy. *Zootaxa* 2011;**2755**(1):36–56.

11. Forêt, S, Knack, B, Houliston, E, *et al*. New tricks with old genes: the genetic bases of novel cnidarian traits. *Trends Genet* 2010;**26**(4):154–8.

12. Harvey, EB. A review of the chromosome numbers in the Metazoa. Part I. *J Morphol* 1916;**28**(1):1–63.

13. Makino, S. *An Atlas of the Chromosome Numbers in Animals*. 2nd ed. Ames: Iowa State College Press; 1951.

14. Goldberg, RB, Crain, WR, Ruderman, JV, *et al*. DNA sequence organization in the genomes of five marine invertebrates. *Chromosoma* 1975;**51**(3):225–51.

15. Adachi, K, Miyake, H, Kuramochi, T, *et al*. Genome size distribution in phylum Cnidaria. *Fish Sci* 2017;**83**(1):107–12.

16. Dunn, CW, Ryan, JF. The evolution of animal genomes. *Curr Opin Genet Dev* 2015;**35**:25–32.

17. Goodwin, S, McPherson, JD, McCombie, WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**(6):333–51.

18. Ames, CL, Ryan, JF, Bely, AE, *et al*. A new transcriptome and transcriptome profiling of adult and larval tissue in the box jellyfish *Alatina alata*: an emerging model for studying venom, vision and sex. *BMC Genomics* 2016;**17**:650.

19. Gold, DA, Katsuki, T, Li, Y, *et al*. The genome of the jellyfish *Aurelia* and the evolution of animal complexity. *Nat Ecol Evol* 2019;**3**(1):96–104.

20. Khalturin, K, Shinzato, C, Khalturina, M, *et al*. Medusozoan genomes inform the evolution of the jellyfish body plan. *Nat Ecol Evol* 2019;**3**(5):811–22.

21. Leclère, L, Horin, C, Chevalier, S, *et al*. The genome of the jellyfish *Clytia hemisphaerica* and the evolution of the cnidarian life-cycle. *Nat Ecol Evol* 2019;**3**(5):801–10.

22. Ohdera, A, Ames, CL, Dikow, RB, *et al*. Box, stalked, and upside-down? Draft genomes from diverse jellyfish (Cnidaria, Acraspeda) lineages: *Alatina alata* (Cubozoa), *Ca lvadosia cruxmelitensis* (Staurozoa), and *Cassiopea xamachana* (Scyphozoa). *Gigascience* 2019;**8**(7):giz069.

23. Steele, RE, David, CN, Technau, U. A genomic view of 500 million years of cnidarian evolution. *Trends Genet* 2011;**27**(1):7–13.

24. Technau, U, Schwaiger, M. Recent advances in genomics and transcriptomics of cnidarians. *Mar Genomics* 2015;**24**:131–8.

25. Alama-Bermejo, G, Holzer, AS. Advances and discoveries in myxozoan genomics. *Trends Parasitol* 2021;**37**(6):552–68.

26. D'Ambra, I, Lauritano, C. A review of toxins from Cnidaria. *Mar Drugs* 2020;**18**(10):507.

27. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2015;**43**(D1):D6–17.

28. Bushnell, B. BBMap v38.73. https://sourceforge.net/projects/bbmap/. Accessed 25 May 2021.

29. Dainat, J, Hereñú, D, Pucholt, P. NBISweden/AGAT: AGAT-v0.6.0. *Zenodo* 2021. https://doi.org/10.5281/zenodo.5336786.

30. Simão, FA, Waterhouse, RM, Ioannidis, P, *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**(19):3210–2.

31. Afgan, E, Baker, D, Batut, B, *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 2018;**46**(W1):W537–44.

32. Galaxy. https://usegalaxy.org/. Accessed 10 August 2021.

33. Lewin, HA, Robinson, GE, Kress, WJ, *et al.* Earth BioGenome Project: sequencing life for the future of life. *Proc Natl Acad Sci U S A* 2018;**115**(17):4325–33.

34. Huerta-Cepas, J, Serra, F, Bork, P. ETE 3: Reconstruction, analysis and visualization of phylogenomic data. *Mol Biol Evol* 2016;**33**(6):1635–8.

35. Caswell, TA, Droettboom, M, Lee, A, *et al.* Matplotlib release v3.3.1. *Zenodo* 2020. https://doi.org/10.5281/zenodo.3984190.

36. Waskom, ML. Seaborn: statistical data visualization. *J Open Source Softw* 2021;**6**(60):3021.

37. Inkscape Project IW. Inkscape. https://inkscape.org/. Accessed 21 January 2022.

38. Bayha, KM, Dawson, MN, Collins, AG, *et al.* Evolutionary relationships among scyphozoan jellyfish families based on complete taxon sampling and phylogenetic analyses of 18S and 28S ribosomal DNA. *Integr Comp Biol* 2010;**50**(3): 436–55.

39. Maronna, MM, Miranda, TP, Peña Cantero, ÁL, *et al.*. Towards a phylogenetic classification of Leptothecata (Cnidaria, Hydrozoa). *Sci Rep* 2016;**6**:18075.

40. Mendoza-Becerril, MA, Jaimes-Becerra, AJ, Collins, AG, *et al.* Phylogeny and morphological evolution of the so-called bougainvilliids (Hydrozoa, Hydroidolina). *Zool Scr* 2018;**47**(5):608–22.

41. Elliott, TA, Gregory, TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos Trans R Soc Lond B Biol Sci* 2015;**370**(1678):20140331.

42. Wilbrandt, J, Misof, B, Niehuis, O. COGNATE: comparative gene annotation characterizer. *BMC Genomics* 2017;**18**(1):535.

43. Tardent, P. *Coelenterata, Cnidaria*. 1st ed. Jena/Stuttgart: Gustav Fischer; 1978.

44. Kubota, S. Systematic study on a bivalve-inhabiting hydroid *Eucheilota intermedia* Kubota from central Japan. *J Fac Sci Hokkaido Univ Ser VI Zool* 1985;**24**:pl. I.

45. Kubota, S. Taxonomic study on *Hydrocoryne miurensis* (Hydrozoa). *Publ Seto Mar Biol Lab* 1988;**33**(1-3):1–18.

46. Kubota, S. Second finding of *Stylactaria piscicola* (Komai, 1932) comb. nov. (Hydrozoa: Hydractiniidae) from off Atsumi Peninsula, Japan. *Publ Seto Mar Biol Lab* 1991;**35**(1-3):11–5.

47. Kubota, S. Chromosome number of a bivalve-inhabiting hydroid, *Eugymnanthea japonica* (Leptomedusae: Eirenidae) from Japan. *Publ Seto Mar Biol Lab* 1992;**35**(6):383–6.

48. Guo, P. The karyotype of *Rhopilema esculenta*. *J Fish China* 1994;**18**:253–5.

49. Anokhin, B, Kuznetsova, V. Chromosome morphology and banding patterns in *Hydra oligactis* Pallas and *H. circumcincta* Schultze (Hydroidea, Hydrida). *Fol Biol Krakow* 1999;**47**:91–6.

50. Anokhin, B, Nokkala, S. Characterization of C-heterochromatin in four species of Hydrozoa (Cnidaria) by sequence specific fluorochromes Chromomycin A $_3$ and DAPI. *Caryologia* 2004;**57**(2):163–6.

51. Anokhin, BA, Kuznetsova, VG. FISH-based karyotyping of *Pelmatohydra oligactis* (Pallas, 1766), *Hydra oxycnida* Schulze, 1914, and *H. magnipapillata* Itô, 1947 (Cnidaria, Hydrozoa). *Comp Cytogenet* 2018;**12**(4):539–48.

52. Pflug, JM, Holmes, VR, Burrus, C, *et al.*. Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). *G3 (Bethesda)* 2020;**10**(9):3047–60.

53. Lawley, JW, Gamero-Mora, E, Maronna, MM, *et al.* The importance of molecular characters when morphological variability hinders diagnosability: systematics of the moon jellyfish genus *Aurelia* (Cnidaria: Scyphozoa). *PeerJ* 2021;**9**:e11954.

54. GigaDB. http://gigadb.org. Accessed 1 April 2021.

55. Hydra 2.0 Web Portal. https://research.nhgri.nih.gov/hydra/. Accessed 1 April 2021.

56. OIST Marine Genomics Unit Genome Browser. https://marinegenomics.oist.jp/gallery. Accessed 1 April 2021.

57. MARIMBA. http://marimba.obs-vlfr.fr/. Accessed 1 April 2021.

58. IRIDIAN GENOMES. https://www.iridiangenomes.com/. Accessed 1 April 2021.

59. Li, Y, Gao, L, Pan, Y, *et al.* Chromosome-level reference genome of the jellyfish *Rhopilema esculentum*. *Gigascience* 2020;**9**(4):giaa036.

60. Galliot, B, Schummer, M. 'Guessmer' screening strategy applied to species with AT-rich coding sequences. *Trends Genet* 1993;**9**(1):3–4.

61. Hoff, K, Stanke, M. Current methods for automated annotation of protein-coding genes. *Curr Opin Insect Sci* 2015;**7**:8–14.

62. Peona, V, Blom, MPK, Xu, L, *et al.* Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol Ecol Resour* 2021;**21**(1):263–86.

63. Nong, W, Cao, J, Li, Y, *et al.* Jellyfish genomes reveal distinct homeobox gene clusters and conservation of small RNA processing. *Nat Commun* 2020;**11**(1):3051.

64. Xia, W-X, Li, H-R, Ge, J-H, *et al.* High-continuity genome assembly of the jellyfish *Chrysaora quinquecirrha*. *Zool Res* 2021;**42**(1):130–4.

65. Putnam, NH, Srivastava, M, Hellsten, U, *et al.* Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 2007;**317**(5834):86–94.

66. Taguchi, T, Tagami, E, Mezaki, T, *et al.* Recent progress of molecular cytogenetic study on scleractinian (stony) corals. *Kuroshio Sci* 2017;**11**.73–81

67. Zimmermann, B, Robb, S, Genikhovich, G, *et al.* Sea anemone genomes reveal ancestral metazoan chromosomal macrosynteny. *bioRxiv* 2021;doi:10.1101/2020.10.30.359448.

68. Blommaert, J. Genome size evolution: towards new model systems for old questions. *Proc Biol Sci* 2020;**287**(1933):20201441.

69. Wong, WY, Simakov, O, Bridge, DM, *et al.* Expansion of a single transposable element family is associated with genome-size increase and radiation in the genus Hydra. *Proc Natl Acad Sci U S A* 2019;**116**(46):22915–7.

70. Schrader, L, Schmitz, J. The impact of transposable elements in adaptive evolution. *Mol Ecol* 2019;**28**(6):1537–49.

71. Cosby, RL, Judd, J, Zhang, R, *et al*. Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* 2021;**371**(6531):eabc6405.

72. Hamada, M, Satoh, N, Khalturin, K. A reference genome from the symbiotic hydrozoan, *Hydra viridissima*. *G3 (Bethesda)* 2020;**10**(11):3883–95.

73. Kim, H-M, Weber, JA, Lee, N, *et al*. The genome of the giant Nomura's jellyfish sheds light on the early evolution of active predation. *BMC Biol* 2019;**17**(1):28.

74. Xia, W, Li, H, Cheng, W, *et al*. High-quality genome assembly of *Chrysaora quinquecirrha* provides insights into the adaptive evolution of jellyfish. *Front Genet* 2020;**11**:535.

75. Graur, D, Zheng, Y, Azevedo, RBR. An evolutionary classification of genomic function. *Genome Biol Evol* 2015;**7**(3):642–5.

76. Chapman, J, Kirkness, EF, Simakov, O, *et al*. The dynamic genome of *Hydra*. *Nature* 2010;**464**(7288):592–6.

77. Galliot, B. *Hydra*, a fruitful model system for 270 years. *Int J Dev Biol* 2012;**56**(6-7-8):411–23.

78. Tomczyk, S, Fischer, K, Austad, S, *et al*. *Hydra*, a powerful model for aging studies. *Invertebr Reprod Dev* 2015;**59**(sup1):11–6.

79. Weisman, CM, Murray, AW, Eddy, SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol* 2020;**18**(11):e3000862.

80. Weisman, CM, Murray, AW, Eddy, SR. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. bioRxiv 2022;doi:10.1101/2022.01.13.476251.

81. Chen, Y, González-Pech, RA, Stephens, TG, *et al*. Evidence that inconsistent gene prediction can mislead analysis of dinoflagellate genomes. *J Phycol* 2020;**56**(1):6–10.

82. Martín-Durán, JM, Ryan, JF, Vellutini, BC, *et al*. Increased taxon sampling reveals thousands of hidden orthologs in flatworms. *Genome Res* 2017;**27**(7):1263–72.

83. Natsidis, P, Kapli, P, Schiffer, PH, *et al*. Systematic errors in orthology inference and their effects on evolutionary analyses. *iScience* 2021;**24**(2):102110.

84. Quattrini, AM, Rodríguez, E, Faircloth, BC, *et al*. Palaeoclimate ocean conditions shaped the evolution of corals and their skeletons through deep time. *Nat Ecol Evol* 2020;**4**(11):1531–8.

85. Schriml, LM, Chuvochina, M, Davies, N, *et al*. COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci Data* 2020;**7**(1):188.

86. Martinez, PA, Jacobina, UP, Fernandes, RV, *et al*. A comparative study on karyotypic diversification rate in mammals. *Heredity* 2017;**118**(4):366–73.

87. Toczydlowski, RH, Liggins, L, Gaither, MR, *et al*. Poor data stewardship will hinder global genetic diversity surveillance. *Proc Natl Acad Sci U S A* 2021;**118**(34):e2107934118.

88. Bayer, PE, Edwards, D, Batley, J. Bias in resistance gene prediction due to repeat masking. *Nat Plants* 2018;**4**(10):762–5.

89. Fiddes, IT, Armstrong, J, Diekhans, M, *et al*. Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation. *Genome Res* 2018;**28**(7):1029–38.

90. König, S, Romoth, LW, Gerischer, L, *et al*. Simultaneous gene finding in multiple genomes. *Bioinformatics* 2016;**32**(22):3388–95.

91. Dunne, MP, Kelly, S. OMGene: mutual improvement of gene models through optimisation of evolutionary conservation. *BMC Genomics* 2018;**19**(1):307.

92. Dunne, MP, Kelly, S. OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations. *BMC Genomics* 2017;**18**(1):390.

93. Hua, Z, Early, MJ. Closing target trimming and CTTdocker programs for discovering hidden superfamily loci in genomes. *PLoS One* 2019;**14**(7):e0209468.

94. Kim, S, Cheong, K, Park, J, *et al*. TGFam-Finder: a novel solution for target-gene family annotation in plants. *New Phytol* 2020;**227**(5):1568–81.

95. Ottenburghs, J, Geng, K, Suh, A, *et al*. Genome size reduction and transposon activity impact tRNA gene diversity while ensuring translational stability in birds. *Genome Biol Evol* 2021;**13**(4):evab016.

96. Arita, M, Karsch-Mizrachi, I, Cochrane, G. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res* 2021;**49**(D1):D121–4.

97. Kodama, Y, Shumway, M, Leinonen, R. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* 2012;**40**(D1):D54–6.

98. Hemmrich, G, Bosch, TC. Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation. *Bioessays* 2008;**30**(10):1010–8.

99. Ryan, JF, Finnerty, JR. CnidBase : The Cnidarian Evolutionary Genomics Database. *Nucleic Acids Res* 2003;**31**(1):159–63.

100. Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.<number>101.</number>GIGA Community of Scientists. The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes. *J. Hered* 2014;**105**(1):1–18.

102. Ewels, P, Magnusson, M, Lundin, S, *et al*. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**(19):3047–8.

103. Challis, R, Richards, E, Rajan, J, *et al*. BlobToolKit – interactive quality assessment of genome assemblies. *G3 (Bethesda)* 2020;**10**(4):1361–74.

104. Liu, F, Li, Y, Yu, H, *et al*. MolluscDB: an integrated functional and evolutionary genomics database for the hyper-diverse animal phylum Mollusca. *Nucleic Acids Res* 2021;**49**(D1):D988–97.

105. Storer, J, Hubley, R, Rosen, J, *et al*. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* 2021;**12**(1):2.

106. Fautin, DG, Westfall, JA, Cartwright, P, *et al*. Coelenterate Biology 2003: trends in research on Cnidaria and Ctenophora. *Hydrobiologia* 2005;**530**:11–3.

107. He, S, Grasis, JA, Nicotra, ML, *et al*. Cnidofest 2018: the future is bright for cnidarian research. *Evodevo* 2019;**10**:20.

108. Funayama, N, Frank, U. Meeting report on "At the roots of bilaterian complexity: insights from early emerging metazoans," Tutzing (Germany) September 16-19, 2019. *Bioessays* 2020;**42**(2):1900236.

109. Paps, J, Holland, PWH. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nat Commun* 2018;**9**(1):1730.

110. Guijarro-Clarke, C, Holland, PWH, Paps, J. Widespread patterns of gene loss in the evolution of the animal kingdom. *Nat Ecol Evol* 2020;**4**(4):519–23.

111. Dawson, MN, Hamner, WM. A character-based analysis of the evolution of jellyfish blooms: adaptation and exaptation. *Hydrobiologia* 2009;**616**(1):193–215.

112. Deakin, JE, Potter, S, O'Neill, R, *et al*. Chromosomics: bridging the gap between genomes and chromosomes. *Genes* 2019;**10**(8):627.

113. Dimitrova, M, Meyer, R, Buttigieg, PL, *et al*. A streamlined workflow for conversion, peer review, and publication of ge-

nomics metadata as omics data papers. *Gigascience* 2021;**10**(5): giab034.

114. Santander, MD, Maronna, MM, Ryan, JF, *et al*. The state of Medusozoa genomics: supplementary material. figshare. 2022; http://dx.doi.org/10.6084/m9.figshare.17155676.

116. Hydractinia Genome Project Portal. https://research.nhgri.nih.gov/hydractinia/. Accessed 1 April 2021.

117. Vogg, MC, Beccari, L, Iglesias Ollé, L, *et al*. An evolutionarily-conserved Wnt3/β-catenin/Sp5 feedback loop restricts head organizer activity in Hydra. *Nat Commun* 2019;**10**(1): 312.

118. Migotto, AE, Vellutini, BC. Cifonauta: Banco de Imagens de Biologia Marinha. http://cifonauta.cebimar.usp.br/. Accessed 2 February 2022.