Chapter 19

Ocean to Tree: Leveraging Single-Molecule RNA-Seq to Repair Genome Gene Models and Improve Phylogenomic Analysis of Gene and Species Evolution

Jan Hsiao, Lola Chenxi Deng, Leonid L. Moroz, Sreekanth H. Chalasani, and Eric Edsinger

Abstract

Understanding gene evolution across genomes and organisms, including ctenophores, can provide unexpected biological insights. It enables powerful integrative approaches that leverage sequence diversity to advance biomedicine. Sequencing and bioinformatic tools can be inexpensive and user-friendly, but numerous options and coding can intimidate new users. Distinct challenges exist in working with data from diverse species but may go unrecognized by researchers accustomed to gold-standard genomes. Here, we provide a high-level workflow and detailed pipeline to enable animal collection, single-molecule sequencing, and phylogenomic analysis of gene and species evolution. As a demonstration, we focus on (1) PacBio RNA-seq of the genome-sequenced ctenophore *Mnemiopsis leidyi*, (2) diversity and evolution of the mechanosensitive ion channel Piezo in genetic models and basal-branching animals, and (3) associated challenges and solutions to working with diverse species and genomes, including gene model updating and repair using single-molecule RNA-seq. We provide a Python Jupyter Notebook version of our pipeline (GitHub Repository: Ctenophore-Ocean-To-Tree-2023 https://github.com/000generic/Ctenophore-Ocean-To-Tree-2023) that can be run for free in the Google Colab cloud to replicate our findings or modified for specific or greater use. Our protocol enables users to design new sequencing projects in ctenophores, marine invertebrates, or other novel organisms. It provides a simple, comprehensive platform that can ease new user entry into running their evolutionary sequence analyses.

Key words Ctenophora, *Mnemiopsis*, De novo transcriptome, Single-molecule sequencing PacBio SMRT, Phylogenomics, Phylogenetic trees, Gene family tree, Mechanosensitive ion channels, Piezo

1 Introduction

Ctenophores are transparent gelatinous, almost alien-like, marine animals of enigmatic, if controversial, origin and biology [1–4]. They exhibit a number of unusual biological features, including rotational symmetry and fourfold structuring of the body plan (unique in animals) [1, 5–7], circumvention of physical constraints in cilia-based swimming that otherwise limit body size across

Leonid L. Moroz (ed.), *Ctenophores: Methods and Protocols*, Methods in Molecular Biology, vol. 2757,
https://doi.org/10.1007/978-1-0716-3642-8_19,

kingdoms and do so by fusing cilia into massive comb plates of 100,000 cilia each [8–11], and a tripartite through-gut with anal pores at one end that make 'thru-ness' of the gut unobvious, a fact that some biologists largely forgot for over a century, and after its rediscovery and modern analysis, the gut's possible homology to tripartite through-guts in other animals remains unclear [12–14]. Moreover, common animal traits such as neurons, synapses, muscles, and mesoderm could potentially be a result of convergent evolution in ctenophores vs. other animals [15–17]. Finally, ctenophore genomes are unusual, even among marine invertebrate oddities, and are recognized as highly divergent in gene sequence and gene families in comparison with other animals [18, 19]. Overall, a deepened understanding of ctenophore diversity and evolution, from sequences to ecosystems, promises new insights into their extraordinary biology.

Initial genome publications for ctenophores were of draft assemblies for *Mnemiopsis leidyi* (*Mnemiopsis*) and *Pleurobrachia bachei* (*Pleurobrachia*) [18, 19]. More recent assembly of *Hormiphora californiensis* (*Hormiphora*) [20] and updated 3D genome assembly in *Pleurobrachia* [21] identified 13 chromosomes in both species and greatly improved available resources. Famously, the initial publications ignited controversy regarding phylogenetic placement of ctenophores in animal evolution [18, 19, 22–27]. The studies contradicted traditional views, indicating ctenophores are the basal-most branch in animals and calling into question conservation vs. convergence of basic animal features, including guts, muscles, and brains [7, 18, 19, 28, 29]. Subsequent studies and commentaries remained deeply divergent, often falling into ctenophore-first vs. sponge-first hypotheses [18, 19, 22–27, 30, 31].

Furthermore, it was shown that technical details regarding evolutionary models and available parameter space in different software packages may help account for dramatically different trees using similar data in different sequence-based studies [4]. Of note, our pipeline here uses maximum-likelihood tree building in IQTree2 and, as might be predicted for the tool [4] and as shown in Fig. 4, places ctenophores at the animal base. Until quite recently, it remained a subject of hot debates which of the two scenarios—ctenophore-first or sponge-first—is correct [2, 4]; although integrative analyses favored the ctenophore-first reconstructions. However, a novel approach using syntenic features across metazoan and unicellular outgroup genomes demonstrated ctenophore-first as correct [32].

Importantly, the divergent nature of ctenophore genes and genomes and their use in characterizing early animal evolution highlight fundamental challenges in sequence analysis when phylogenetic signal is limited [2, 33–35]. Work on ctenophores has led to new tools, approaches, and understanding in the phylogenomics

field, including some of the methods used here [2, 33–35]. Whatever the evolutionary patterns of novelty, conservation, and convergence, phylogenomic study of ctenophores genes and genomes will be critical to understanding early animal evolution and basic principles of animal cell types and systems across phyla and in humans.

Biodiversity offers an incredibly rich potential for discovery of new genes and pathways that can be directly utilized or informatically leveraged in basic research and medicine. Sequencing and phylogenetic characterization of diverse genes in novel organisms have led to the discovery, advancement, and engineering of some of the most powerful genetic tools in science, including Taq for PCR [36], GFP for fluorescent imaging [37], channelrhodopsin for optogenetics [38], and CRISPR-Cas9 for gene editing [39]. Additional discovery and characterization of naturally occurring sequence diversity representing new or novel homologs in related species continue to advance each of these technologies and many others [40–43]. These tools and their advances underscore the importance of genome-scale sequencing of organismal diversity in closely and distantly related species and the importance of phylogenomic characterization of homologs across species; both are areas the protocol here enables for users.

Integrative experimental and phylogenomic approaches are commonly used in protein engineering and can be combined with additional increasingly powerful machine learning, deep learning, and other artificial intelligence (AI) methods. These tools exploit biodiversity and evolution's optimization of diverse sequences and functions through random, potentially near-comprehensive, explorations of sequence space. In this context, presence and distribution of existing biological or estimated ancestral phylogenetically related or unrelated neighboring sequences in sequence space can be highly informative [44], representing successful functional optimization in protein engineering and design by evolution. Leveraging natural sequence diversity at scale, the tools and approaches can massively collapse the sequence space that might otherwise have to be experimentally explored, highlighting a much smaller subset of potential variants. This can save by orders of magnitude the time and resources required, for instance, to rapidly identify or engineer a novel sequence and advance a genetic technology. This is especially true now that it is possible to predict the 3D structure of nearly all proteins based on sequence alignment and better infer potential function of motifs, domains, and regions for engineering [45–51]. Again, these areas in basic and biomedical research highlight the importance of genome-scale sequencing of organismal diversity in closely and distantly related species and the importance of phylogenomic characterization of homologs across species; both are areas the protocol here enables for users.

Phylogenomic analysis of gene family evolution includes three major steps:

1. Candidate homolog identification by searching reference sequences across species genomes and transcriptomes.
2. Multiple sequence alignment of reference and candidate sequences.
3. Gene tree generation based on aligned sequences [52].

Additional generation of a species tree by a similar process may also be needed, particularly when one or more species, or the collective set of species, have never been characterized. Phylogenomic analyses to produce gene and species trees are commonly undertaken in labs for the first time after newly acquired genome or transcriptome data sets arrive. Numerous tools and packages exist and can be run naively, often with reasonable accuracy and without detailed understanding of the command line or coding.

We recommend several such tools below. Still, their often extensive functionality can be overwhelming; some are not free, you are limited to specific subsets of tools developers happened to have packaged for a given task, and underlying steps and code may be inaccessible with little context. In contrast, cutting-edge phylogenomic and phylogenetic tools run at the command line are often freely available and part of a diverse universe of tools offering specific functionalities. These tools enable users to build powerful bioinformatic and phylogenomic pipelines targeted to their particular needs. They may readily scale but require proficiency in working at the command line and coding one or more languages that some researchers may not have yet acquired.

To facilitate sequencing and phylogenomic analysis by new users, we provide here a protocol that includes a high-level workflow of animal collection and single-molecule RNA-seq and a more detailed walk through phylogenomic analysis of species and gene evolution, including production of alignments, trees, and species-gene family heatmaps (Fig. 1).

As a workflow demonstration, we generated single-molecule (long-read) PacBio Sequel II RNA-seq data of an entire adult *Mnemiopsis leidyi* and produced a transcriptome. Our phylogenomic pipeline uses the PacBio transcriptome, which is predominantly accurate full-length transcripts, to improve gene models in the *Mnemiopsis* draft genome, and then performs species and gene evolutionary analysis of Piezo, a mechanosensitive ion channel [53], in basal-branching animals, including ctenophores *Mnemiopsis* and *Hormiphora*.

Finally, we provide a Python Jupyter Notebook that runs a fully fledged demonstration of the pipeline analysis. The notebook offers a simple means and user-friendly environment for nonexperts to run and adapt the pipeline. One can see the underlying code in

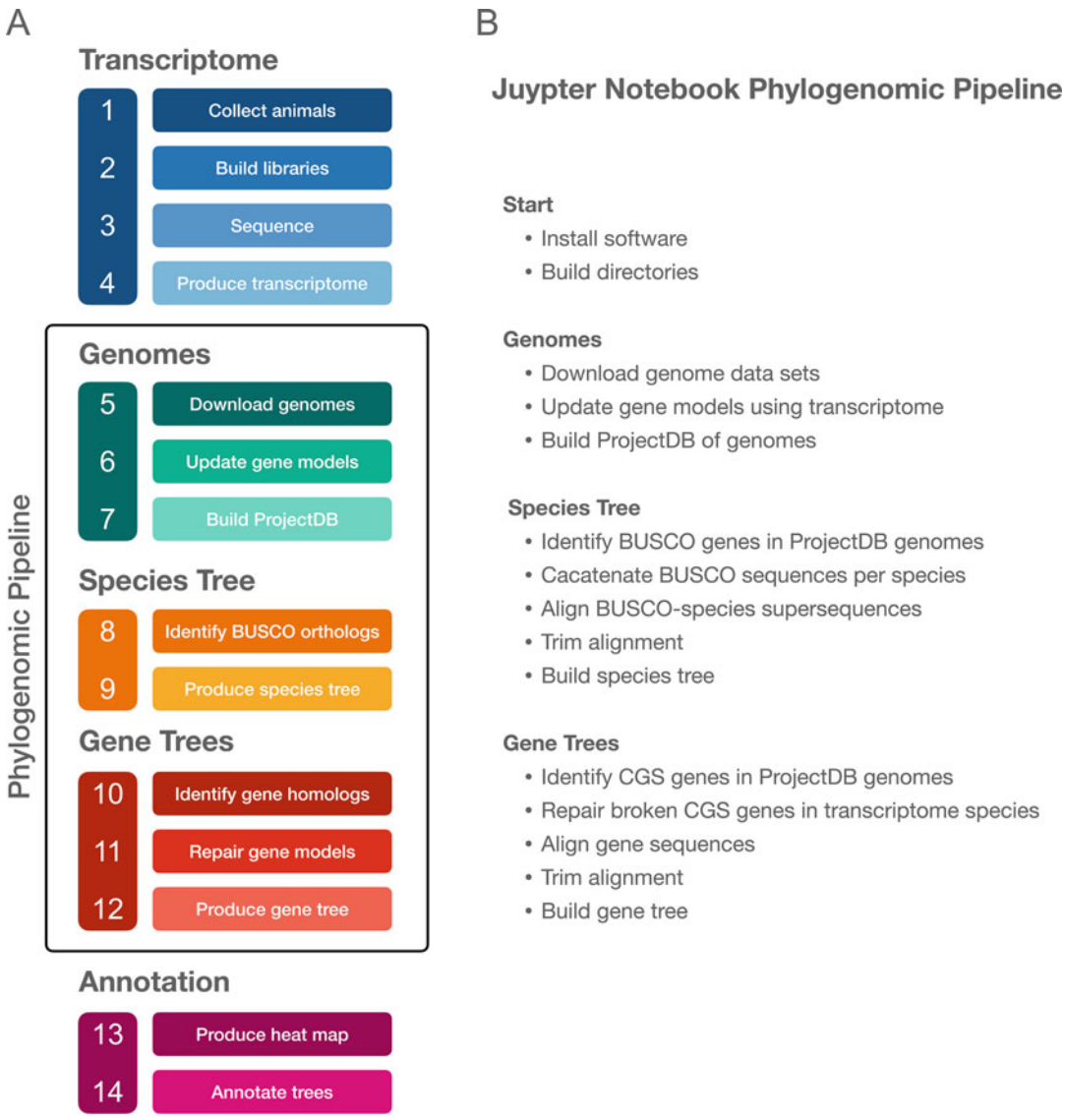

Fig. 1 Overview of the ocean-to-tree workflow and phylogenomics pipeline. (**a**) Ocean-to-tree workflow. (**b**) Outline of the phylogenomic pipeline that is run in the associated Jupyter Notebook that is provided at GitHub (https:/github.com/000generic/Ctenophore-Ocean-To-Tree-2023)

action and gain familiarity with the methods in the process of running the notebook. At the same time, the notebook provides full access to all scripts, allowing expert users to build off of our pipeline and advance their own specific projects. The notebook can be run for free in the Google Colab cloud, locally, or elsewhere.

2 Materials

2.1 Collection

1. Dock or seawall access to ocean or sea.
2. Large or aquarium dip net or dip cup.
3. 5–20 glass or plastic jars of varying sizes.
4. 5–20 L bucket or cooler.
5. Watertight insulated water bottle.
6. Cooler or Styrofoam box.
7. Optional: Temperature logger.
8. Optional: Blue ice or heat packs.
9. Small aquarium dip net.
10. Air pump.
11. Air line.
12. Air stone.

2.2 RNA Extraction

1. Small aquarium dip net.
2. Air pump.
3. Air line.
4. Air stone.
5. If collecting seawater: Bag filter.
6. If preparing seawater: Sea salt (Instant Ocean or others).
7. If preparing seawater: Hydrometer or refractometer.
8. Sterile filter that fits screw-top bottles.
9. Vacuum source for sterile filtration.
10. Five 1-L sterilized glass bottles (other sizes can work—need to hold 2.5–5 L total).
11. Five 1-L sterilized glass beakers.
12. Large Kimwipes.
13. Electric homogenizer (Polytron 1200 or others).
14. *Alternate: Glass mortar and pestle.*
15. RNA extraction kit (Qiagen RNeasy or others).
16. Surface decontamination solution (RNase Away or others).
17. RNase-free plasticware.
18. Low-binding DNA microcentrifuge tubes (DNA LoBind or others).
19. Fume hood.
20. Spectrophotometer (NanoDrop or others).
21. Fluorometer (Qubit or others).
22. Automated small-volume electrophoresis system (TapeStation, Bioanalyzer, or others).

2.3 Sequencing

1. Single-molecule sequencing system (PacBio Sequel II or others).
2. Single-molecule sequencing kit (PacBio Sequel II or others).
3. Spectrophotometer (NanoDrop or others).
4. Fluorometer (Qubit or others).
5. Automated small-volume electrophoresis system (TapeStation, Bioanalyzer, or others).

2.4 Phylogenomics

Python 3 programming language	https://tinyurl.com/yx2vt8mu
Python package installer: Anaconda	https://tinyurl.com/yrxzaase
Unix command line editor: Vim	https://tinyurl.com/4sxe3amu
Alternate: Emacs	https://tinyurl.com/yc8pjtdv
Lightweight editor: Atom	https://tinyurl.com/2p8ec7z7
Alternate: Sublime ($)	https://tinyurl.com/5yxekktu
Full IDE editor: Spyder	https://tinyurl.com/yeywt9pw
Alternate: PyCharm (free or $)	https://tinyurl.com/58kw5pv7
Alternate: Visual Studio Code	https://tinyurl.com/2p8f2xte
Pipeline run documentation: Jupyter Notebook	https://tinyurl.com/2w4trfku
Pipeline push-button automation: Snakemake	https://tinyurl.com/2cra9mxx
Alternate: YAML	https://tinyurl.com/ymr4c7ty
Pipeline standalone functionality: Singularity	https://tinyurl.com/jyy3sv7c
Computing: Cloud Google Colab (free or $)	https://tinyurl.com/489ttan7
Alternate: Google Cloud Life Sciences ($)	https://tinyurl.com/5e2wfvwn
Alternate: Cloud Amazon Web Services ($)	https://tinyurl.com/bdu4p6vj

(continued)

Alternate: Local research-grade machine or cluster ($)	NA
Alternate: Local laptop or desktop ($)	NA
Transcriptome QC: EvidentialGene	https://tinyurl.com/4mdkkrm9
Sequence searcher: Blast Suite	https://tinyurl.com/ycktxjsd
Single-gene gene family identifier: BUSCO	https://tinyurl.com/2p8mvjau
Sequence aligner: MAFFT	https://tinyurl.com/ve3cdzd2
Alignment trimmer: ClipKit	https://tinyurl.com/rvkyp4a7
Tree builder maximum likelihood-like: FastTree2	https://tinyurl.com/366ajd6t
Tree builder maximum likelihood: IQTree2	https://tinyurl.com/3mna5kte
Alternate: PhyloBayes (Bayesian Inference)	https://tinyurl.com/4kpbph6a
Alignment viewer: AlignmentViewer	https://tinyurl.com/2f7xscr7
Tree viewing: FigTree	https://tinyurl.com/5n8zay9z
Alternate: iTOL—Interactive Tree of Life (free or $)	https://tinyurl.com/4avjbunw
Spreadsheet data analysis: Google Sheets	https://tinyurl.com/mtdzdkf3
Graphics-friendly software: Google Slides	https://tinyurl.com/2p993bcn
Phylogenomic pipeline alternate: Geneious ($)	https://tinyurl.com/2bbr5y7u
Alternate: Galaxy (free or $)	https://tinyurl.com/4etvu7sv
Phylogenetic pipeline alternate: CIPRES (free or $)	https://tinyurl.com/2p8pc48v
Alternate: NGPhylogeny	https://tinyurl.com/2p88uss6
Alternate: PhyloToL	https://tinyurl.com/44etynyk

2.5 Source Data

Name	Type	Database ID	URL
Ichthyosporea *Sphaeroforma arctica*	Genome	NCBI RefSeq GCF_001186125.1	https://tinyurl.com/yrk8aeca
Filasterea *Capsaspora owczarzaki*	Genome	NCBI RefSeq GCF_000151315.2	https://tinyurl.com/yeu3dpsm
Choanoflagellata *Monosiga brevicollis*	Genome	NCBI RefSeq GCF_000002865.3	https://tinyurl.com/bdzh3huk
Porifera *Ephydatia muelleri*	Genome	EphyBase v1	https://tinyurl.com/2c3de66a
Porifera *Amphimedon queenslandica*	Genome	NCBI RefSeq GCF_000090795.1	https://tinyurl.com/yc69a5ae
Ctenophora *Mnemiopsis leidyi*	Genome	Ensembl Metazoa 51 MneLei_Aug2011	https://tinyurl.com/fmakzcz6
Ctenophora *Hormiphora californiensis*	Genome	GitHub Hormiphora Hcv1.av93	https://tinyurl.com/2p83cvpd
Cnidaria *Nematostella vectensis*	Genome	Stowers Institute NVEC200	https://tinyurl.com/2p893t24
Cnidaria *Morbakka virulenta*	Genome	OIST MG MOR05_r06	https://tinyurl.com/bd8jkvn6
Cnidaria *Rhopilema esculentum*	Genome	GigaDB 100720	https://tinyurl.com/2k8pxmcv
Cnidaria *Hydra vulgaris*	Genome	NIH NHGRI Hydra2.0	https://tinyurl.com/yckzbjew
Placozoa *Trichoplax adhaerens*	Genome	NCBI Genome GCF_000150275.1	https://tinyurl.com/mnw3z5cj
Chordata *Homo sapiens*	Genome	NCBI RefSeq GCF_000001405.39	https://tinyurl.com/4fd2c75v
Arthropoda *Drosophila melanogaster*	Genome	NCBI RefSeq GCF_000001215.4	https://tinyurl.com/mvbt9hw4

(continued)

Name	Type	Database ID	URL
Nematoda *Caenorhabditis elegans*	Genome	NCBI RefSeq GCA_000002985.3	https://tinyurl.com/2p8c6wnv
Ctenophora *Mnemiopsis leidyi*	Transcriptome	NCBI SRA SRR18002386	https://tinyurl.com/mxyy7jca
Chordata *Homo sapiens* Piezo 1	Gene	UniProt Q92508	https://tinyurl.com/ycknp27k
Chordata *Homo sapiens* Piezo 2	Gene	UniProt Q9H5I5	https://tinyurl.com/tpebj8fm
Arthropoda *Drosophila melanogaster* Piezo	Gene	UniProt M9MSG8	https://tinyurl.com/2p8evtrd
Arthropoda *Drosophila melanogaster* Piezo-like	Gene	UniProt A0A126GUQ2	https://tinyurl.com/253fvf54
Nematoda *Caenorhabditis elegans* Piezo	Gene	UniProt A0A061ACU2	https://tinyurl.com/4a7fsuz2

3 Methods

3.1 Collection

1. Ctenophore collections vary depending upon locations and conditions. Readily available genera include *Mnemiopsis*, *Pleurobrachia*, and *Beroe*. Capture one or more individuals of *Mnemiopsis leidyi* (Atlantic Ocean: University of Chicago Marine Biological Laboratory, Woods Hole, MA), *Pleurobrachia bachei* (Pacific Ocean: University of Washington Friday Harbor Laboratories, Friday Harbor, WA), other small-sized ctenophores, or other species found at a dock, seawall, or ocean using a dip net, dip cup, or small plankton net on a line (Fig. 2a–d) (*see* **Note 1**).
2. Maintain animals individually in small jars, ideally aerated, or in a 5–20 L bucket or cooler with fresh seawater during collection (Fig. 2b). When possible, replenish water periodically, particularly if air and water temperatures are very different. Avoid overloading a bucket with animals or plankton as water conditions can deteriorate quickly.

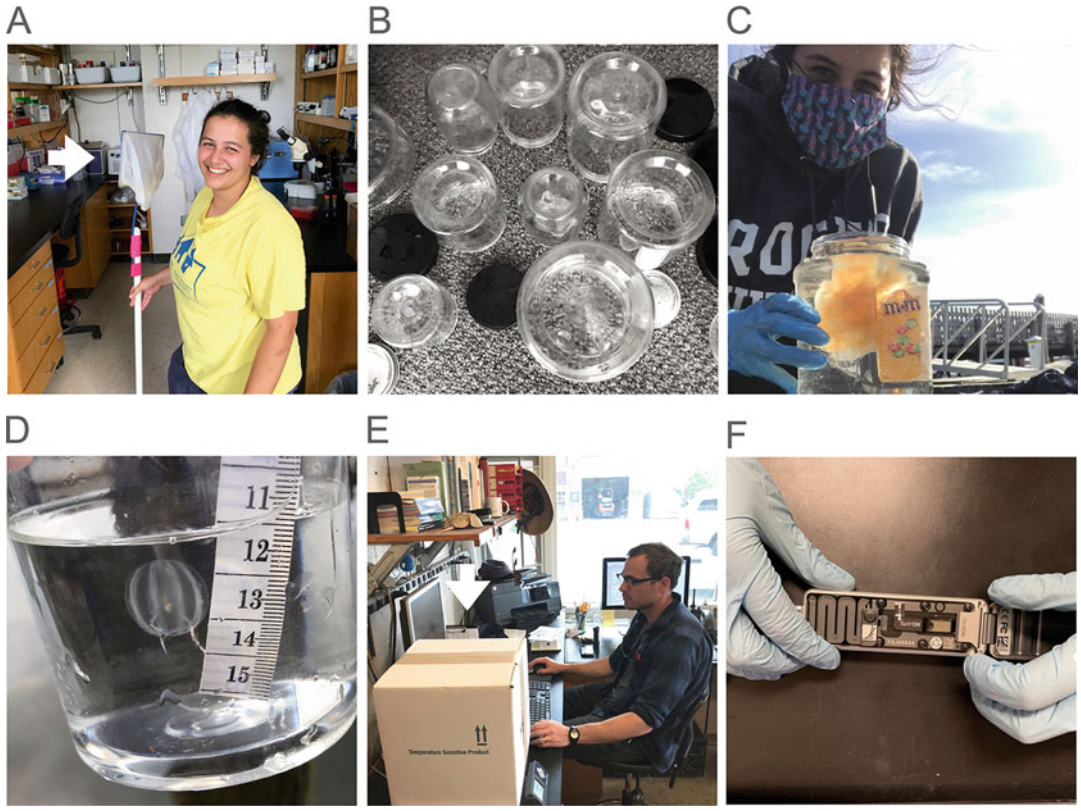

Fig. 2 Field collection, shipping, and sequencing of ctenophores, jellyfish, or other surface-dwelling marine organisms. (**a**) A medium-sized aquarium net duct-taped onto a piece of PVC piping (white arrow) can provide a simply effective tool for collecting surface-dwelling animals when walking along docks or seawalls at the ocean. (**b**) Individual ctenophores or jellyfish can be held in small jars after collection to help avoid damage. (**c**) Collecting ctenophores and jellyfish off docks in Bristol, RI, dock walker M. Cordeiro is pictured with the lion's mane jellyfish Cnidaria *Cyanea capillata*. (**d**) Ctenophora *Pleurobrachia pileus* collected by M. Cordeiro off docks in Bristol, RI. (**e**) Shipment of Ctenophora *Mnemiopsis leidyi* (boxed up and indicated by white arrow) by E. Edsinger and S. Bennet (pictured) from the University of Chicago Marine Biological Laboratory Marine Resources Center to the University of Florida Whitney Laboratory for Marine Bioscience Moroz Laboratory for single-cell sequencing. (**f**) Oxford Nanopore Technologies MinION and related sequencers provide inexpensive, user-friendly means of doing single-molecule genome and RNA-seq in a laboratory using a laptop and molecular biology tools and reagents

3. Transfer ctenophores to a watertight insulated bottle or cooler filled with fresh seawater using a small aquarium dip net (*see* **Note 2**). Aim for a density of five to ten animals golf ball or smaller in size in 1–2 L of seawater. Fill to brim to minimize sloshing that could damage animals.
4. Transport or ship animals to laboratory shortly after collection or else transport, maintain in aquaria with appropriate temperature and aeration, and ship later. Ship the insulated watertight container containing animals in a Styrofoam box (Fig. 2e). Blue ice for polar-to-cooler temperate species or heat packs for

warmer temperate-to-tropical species can be added to the box to help maintain appropriate temperatures for the animals (*see* **Note 3**).

5. Set up ctenophores with aeration upon arrival after transport or shipping. Polar-to-cooler temperate species can be maintained in a 4 °C cold room for 1–3 days without food. Similarly, warmer temperate-to-tropical species can be maintained at room or elevated temperatures for several days without food.

3.2 RNA Extraction

1. Process animals for RNA extraction shortly after arrival in the laboratory, or maintain with aeration at a species-appropriate temperature for up to several days, and then process (*see* **Note 4**). Ensure animals have been starved at least 1 day to clear their guts and avoid sequence contamination by prey.
2. Collect or prepare 5–10 L of seawater (*see* **Note 5**). If collecting seawater, bag or sterile filter after collection to minimize or remove organic and inorganic material. If preparing seawater from sea salt, match its salinity to that of the seawater animals were shipped in. The goal is to minimize osmotic shock to the animals and associated stress-related gene expression prior to RNA extraction.
3. Sterile filter 3–5 L of seawater using vacuum filtration or other methods.
4. Optional: Store seawater in a watertight container to prevent evaporation, or sterile-filtered seawater in sterilized 1 L or other sized glass or plastic bottles, for several days or weeks. Ideally, keep cold and in the dark to minimize growth by microorganisms for non-sterile water.
5. Equilibrate seawater and sterile seawater to the temperature of seawater animals were shipped in or to short-term culturing temperature, if animals will be maintained for a short time after arrival in the lab. Again, the goal is to minimize temperature shock to the animals and associated stress-related gene expression prior to RNA extraction.
6. Decontaminate lab bench and fume hood areas for work with RNA (*see* **Note 6**).
7. Prepare five beakers of temperature-equilibrated sterile filtered seawater at 500 mL to 1 L volumes.
8. Wash ctenophores one at a time in sterile seawater by gently collecting and transferring between beakers of sterile seawater using a clean aquarium dip net. Allow 30 s or longer in each beaker. It is possible to accumulate sterile seawater-washed animals in the final beaker.

9. Work on 1–3 large Kimwipes layered on one another on the decontaminated bench. Kimwipes can be replaced periodically as needed.
10. Transfer the animal onto scrunched up Kimwipes to dry briefly.
11. Transfer lightly dried animal to 50 mL tube containing an appropriate volume of RNA extraction buffer based on kit instructions (*see* **Note 7**).
12. Homogenize immediately in fume hood using electric homogenizer for 10–30 s (*see* **Note 8**).
13. Continue extraction according to kit protocol, and make the final elution into RNase-free water. Process the sample within a few days when possible, as samples will degrade over time in pure water.
14. Determine RNA purity using a spectrophotometer according to equipment instructions. Pure RNA has a 260/280 ratio of 2.0, and pure nucleic acid has a 260/230 ratio of 2.0–2.2. The lower values suggest contamination. Samples that are contaminated might be re-extracted per RNA extraction kit, RNA cleanup and concentration kit, or other protocols (*see* **Note 9**). RNA concentrations determined by spectrophotometry are typically less accurate than other methods.
15. Determine RNA concentration using a fluorometer according to equipment directions. General 2–10 μg is desirable, as PacBio RNA-seq libraries require 1 μg for preparation, though it is possible to use less.
16. Determine RNA quality using a small volume electrophoresis system according to equipment directions. RIN values greater than 8.0 and closer to 9 or 10 are needed to ensure full-length RNA molecules for full-length single-molecule sequencing (*see* **Note 10**).

3.3 Single-Molecule Sequencing

1. An overview of our sequencing strategy is provided (Fig. 2).
2. Prepare sequencing libraries for high-accuracy single-molecule RNA-seq according to sequencing technology kit directions (*see* **Note 11**). In our demonstration workflow, we used PacBio SMRT Sequel II single-molecule sequencing technology and sequencing library kit (https://www.pacb.com/). At this time (Feb 2022), it offers the highest available accuracy in single-molecule sequencing.
3. Perform sequencing of sequencing libraries according to sequencing technology equipment directions. In our demonstration workflow, we used PacBio Sequel II sequencing based on the technology's accuracy. Pictured is the incredibly small, inexpensive, and user-friendly Oxford Nanopore Technologies MinION sequencer (Fig. 2f).

4. Perform post-sequencing quality control and production of a final transcriptome using sequencing technology software. In our demonstration workflow, we used the PacBio SMRT Link software associated with the sequencing system.
5. Deposit reads and transcriptome in NCBI SRA, SRR, and TSA databases, respectively. Demonstration data sets for *Mnemiopsis leidyi* are available at NCBI (BioProject PRJNA806463; BioSample SAMN25884795; Reads SRR18002386; Transcriptome TSA) (*see* **Note 12**).

3.4 Genome Gene Model Updates

1. A demonstration of the following phylogenomic pipeline can be run using our provided Python Jupyter Notebook (GitHub Repository: Ctenophore-Ocean-To-Tree-2023 https://github.com/000generic/Ctenophore-Ocean-To-Tree-2023) in the Google Colab Free cloud or elsewhere. It requires no external input and produces Metazoa15 species (see below) and Piezo gene family alignments and trees.
2. Select computing and software options, and install all software in preparation of running the phylogenomic pipeline. *See* Subheading 2.4 for suggested options. Specific scripts and outside software used in the Jupyter Notebook version of our pipeline were selected or optimized in part to allow the pipeline to run on Google Colab Free (you will need a Google account; https://colab.research.google.com/). Thus, notebook specifics may not be the ideal option if adapting the pipeline to a specific project but can be a good place to start.
3. Download genome gene model gene sets (mRNA, CDS, and AA) and GFF or GTF files for each single-molecule sequenced species (*see* **Note 13**).
4. If isoforms are present in genome gene model gene sets (mRNA, CDS, and AA), collapse isoforms to the longest or best representative transcript. Processing can vary species to species and genome source to genome source. It can typically be done using GFF/GTF and/or fasta header information. Biopython (https://biopython.org/) tools or in-house Python scripts can be used to do this. In the case of the *Mnemiopsis leidyi* genome at Ensembl Metazoa (*see* Subheading 2.5 for link), no isoforms are present in the genome gene model gene set, so the downloaded data is ready to use for sequence repair.
5. Download or copy the single-molecule transcriptome data set (mRNA or CDS) to make it locally available, if it is not already. *See* Subheading 2.5 for link to our *Mnemiopsis* transcriptome.
6. Translate the transcriptome, and select the best transcript per "locus" independent of the genome using EvidentialGene [54] or other tools like CD-Hit [55] to produce a "T1"

transcriptome. EvidentialGene will produce many files, including mRNA (cdna), CDS (cds), and AA (aa) "okay" transcriptomes that include only EvidentialGene's designated best transcript per "locus" (*see* **Note 14**).

7. Build a Blastn database of T1 transcriptome mRNA using Blast + suite makeblastdb.
8. Blast genome CDS gene models against the T1 mRNA transcriptome database using Blastn (*see* **Note 15**).
9. Parse the blast report to identify T1 transcripts that have only a single gene model hit of e-value 0.0 (or can use other or additional Blast statistics and thresholds) (*see* **Notes 16** and **17**).
10. Update genome gene model mRNA, CDS, and AA to T1 transcriptome sequences. In our demonstration pipeline, this is referred to as the UPDATED genome for *Mnemiopsis*.

3.5 Project Database

1. An overview of our phylogenomics pipeline is provided (Fig. 1).
2. Download genome gene model gene sets (mRNA, CDS, and AA) and GFF or GTF files for species of interest (*see* **Notes 18** and **19**). For our demonstration pipeline, we focus on basal-branching animals (ctenophores, sponges, placozoans, and cnidarians) and unicellular outgroups, collectively referred to here as Metazoa15.
3. If isoforms are present in genome gene model gene sets (mRNA, CDS, and AA), collapse isoforms to the longest or the best representative transcript. Processing can vary species to species and genome source to genome source. It can typically be done using GFF/GTF and/or fasta header information. Biopython (https://biopython.org/) tools or in-house Python scripts can be used to do this.
4. Standardize file names and header information. To make visual interpretation of trees easier and provide phylogenetic context on an alignment or tree, our pipeline replaces header information with just the Phylum Genus species details and uses simple sequence identifiers (pdb0000000000) specific to the pipeline run. A map of source and pipeline header details is produced, so things can be mapped back and forth, if needed. There updated files are the pipeline's ProjectDB fastas.
5. Produce a Blastp database for each genome.

3.6 Species Tree Homologs

1. Run BUSCO and its latest Metazoa HMMs [56] on each genome to identify single-copy gene family orthologs to later use in generating a species tree.

2. Process the BUSCO gene fasta files produced by BUSCO Metazoa to reflect species names and ProjectDB identifiers and to provide sets of all BUSCO single-copy sequences per genome in a single file.

3.7 Species Alignment and Tree

1. Concatenate all BUSCO sequences per species genome. The order of sequences must be identical across species. For genes that are absent in a given genome, place holder sequence, such as 10 X's in a row, or simply no sequence, can be used.
2. Align the concatenated BUSCO sequences using MAFFT [57].
3. Trim the aligned sequences using ClipKit with the smartgap setting (*see* **Note 20**) [58].
4. Build a species tree using FastTree2 or IQTree2 for maximum-likelihood methods and/or PhyloBayes for Bayesian inference methods (*see* **Note 21**) [59–61].

3.8 Gene Family Homologs

1. Identify representative reference genes for the gene family of interest in the literature or elsewhere. These genes will be used to scan and identify homologs in Metazoa15 genomes. Ideally, sequences will be from species having high-quality well-annotated genomes. For our demonstration pipeline, we focus on three genetic models in neuroscience, human (*Homo sapiens*), fly (*Drosophila melanogaster*), and worm (*Caenorhabditis elegans*) (*see* **Note 22**). Critically, our pipeline uses a Reciprocal Top Family (RTF) method (detailed below). It requires all homologs of a gene in a genome be included as reference genes for the method to work correctly.
2. Collect CDS or AA fasta sequences for the selected reference genes from UniProt (protein sequences only; https://www.uniprot.org/), GenBank (https://www.ncbi.nlm.nih.gov/nuccore/), genome database websites, or other sources (*see* **Note 23**). This collection of reference sequences is referred to here as the reference gene set (RGS).
3. Blast RGS sequences against each Metazoa15 ProjectDB genome database (*see* **Note 24**).
4. Parse the blast reports and genome fastas to produce a single nonredundant fasta of all RGS hits in all Metazoa15 genomes, the All Hits Fasta.
5. Blast RGS sequences against just the RGS genomes (*see* **Note 25**).
6. Parse the blast report and genome fastas, and then update headers of identified RGS genes in each genome with the header used in the RGS fasta file.

7. Produce a blast database of the RGS-header updated RGS genomes combined.
8. Blast the All Hits Fasta sequences against the RGS genome database.
9. Parse the blast report, and produce a fasta file of all hits that any one of the RGS sequences as a top hit in the combined RGS genomes. This collection of sequences represents all identified potential homologs searched by RBF in the genomes. The sequences are referred to here as the candidate gene set (CGS).
10. Combine the RGS and CGS fasta sequences to produce a final gene set (FGS) fasta file. The FGS fasta will be used for subsequent phylogenetic characterization of the gene family.

3.9 Gene Model Repair

1. Align FGS sequences using MAFFT [57].
2. Trim the aligned sequences using ClipKit with the smartgap setting (*see* **Note 20**) [58].
3. Build a gene family tree using FastTree2 [59–61].
4. Examine the alignment and tree in viewing software, such as AlignmentViewer, FigTree or iTOL, and/or Geneious [62].
5. Focus in particular on branches having multiple homologs of cluster together for a single species and on the length of sequences in alignment relative to RGS sequences. Use this comparison to identify possible partial, expanded, or broken gene model artifacts. For the demonstration pipeline, note that there are four copies of Piezo from the *Mnemiopsis* draft genome clustered together on the tree but only a single copy from the *Mnemiopsis* PacBio transcriptome (Fig. 3a, b). There is also only a single homolog of Piezo in the *Hormiphora* genome (Fig. 3a, b). Based only on the tree, it would appear that there was an expansion of the Piezo gene family along the *Mnemiopsis* lineage within ctenophores and that only one of the four Piezo homologs in *Mnemiopsis* was detected in the PacBio transcriptome. However, in the alignment it is clear that all four copies from the draft genome are partial sequences and roughly line up in a 5′ to 3′ series relative to full-length RGS, *Hormiphora*, and *Mnemiopsis* PacBio sequences (Fig. 3a). Additional examination of genomic coordinates (GFF file) for the four gene models indicates they reside next to each other in the genome as neighbors. Based on this evidence, it appears there is a single copy of Piezo in *Mnemiopsis*, but the gene was broken into four gene models during the process of genome annotation for the draft genome. Thus, it seems reasonable to remove the four broken gene models from the FGS fasta and keep only the full-length PacBio sequence to represent the Piezo gene family in the *Mnemiopsis* genome.

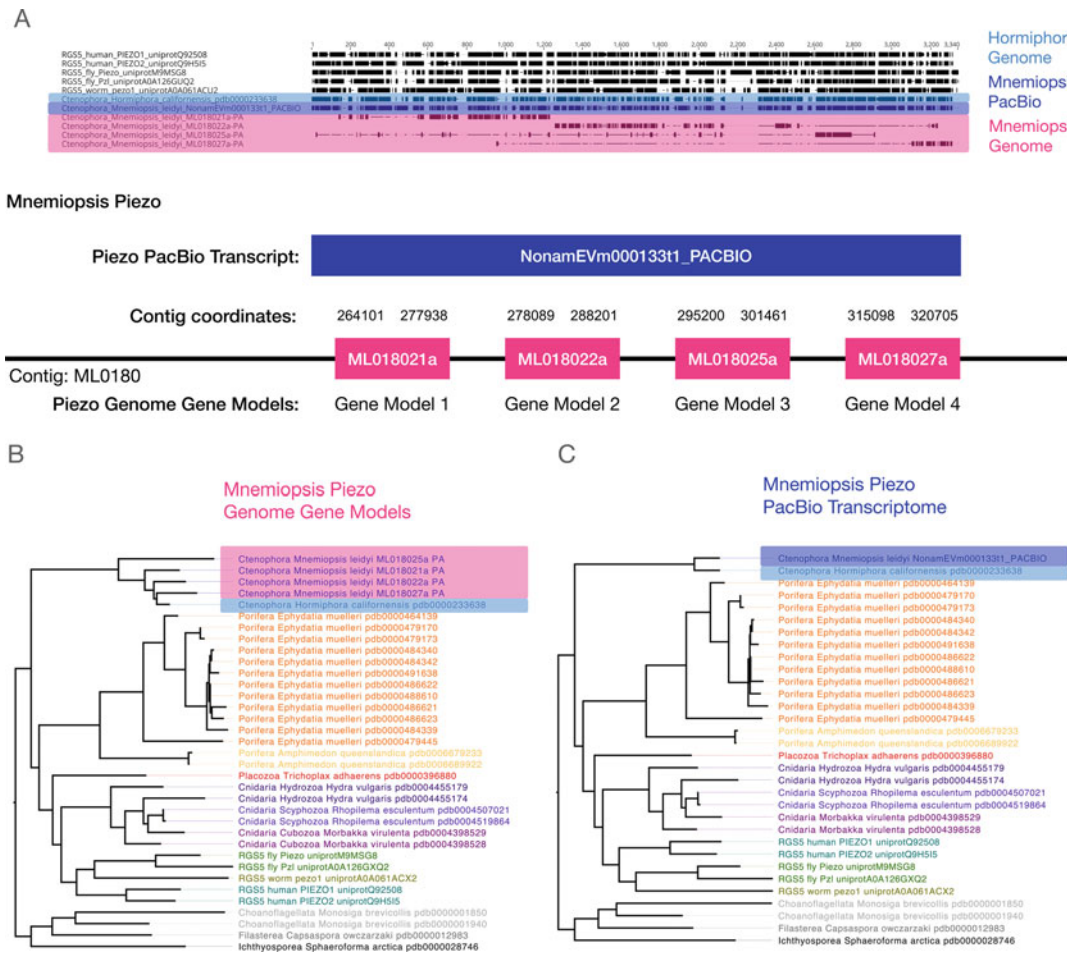

Fig. 3 Gene model repair in *Mnemiopsis* and evolution of the Piezo gene family tree in Metazoa15 genomes. (**a**) MAFFT alignment of Piezo reference sequences in human, fly, and worm and homologs identified in genomes of the ctenophores *Hormiphora* and *Mnemiopsis* and in the PacBio transcriptome of *Mnemiopsis*. *Note*: Multiple gene models appear to be partial sequences of a full-length sequence present as a single copy in the *Mnemiopsis* transcriptome and *Hormiphora* genome. (**b**) IQTree2 maximum-likelihood Piezo gene family tree for Metazoa15 genomes only. The *Mnemiopsis* lineage appears to have expanded the number of copies of Piezo in its genome. However, the alignment in 3A suggests Piezo in the *Mnemiopsis* genome is one gene broken up into four gene models. (**c**) IQTree2 maximum-likelihood Piezo gene family tree for Metazoa15 genomes but with *Mnemiopsis* Piezo gene models repaired by replacement with the *Mnemiopsis* transcriptome Piezo sequence. A single copy of Piezo appears present in the ancestor and is conserved in the *Mnemiopsis*–*Hormiphora* lineage in ctenophores

6. Remove partial, expanded, or broken gene models that can be represented instead by more accurate transcriptome sequences.
7. Remove any transcriptome sequences that are not being used as replacements for gene models exhibiting artifacts. In the case of our demonstration pipeline, FGS sequences have now been repaired for *Mnemiopsis* using the PacBio transcriptome, and the gene set is referred to here as REPAIRED.

3.10 Gene Family Alignment and Tree

1. Align the FGS REPAIRED sequences using MAFFT [57].
2. Trim the aligned sequences using ClipKit with the smartgap setting (*see* **Note 20**) [58].
3. Build a gene tree using FastTree2 or IQTree2 for maximum-likelihood methods and/or PhyloBayes for Bayesian Inference methods (*see* **Note 21**) [59–61].

3.11 Annotations and Heatmap

1. Visualize the species tree in tree viewing software such as FigTree, iTOL, or Geneious [61]. For the demonstration pipeline, this is the dot "fasttree." file for FastTree2 and the dot "treefile." for IQTree2.
2. Color annotate branches or taxon labels using the software, as preferred. Trees can also be rooted and branches rotated at this point.
3. Export the color annotated tree as a pdf or other scalable vector file (Fig. 4a).

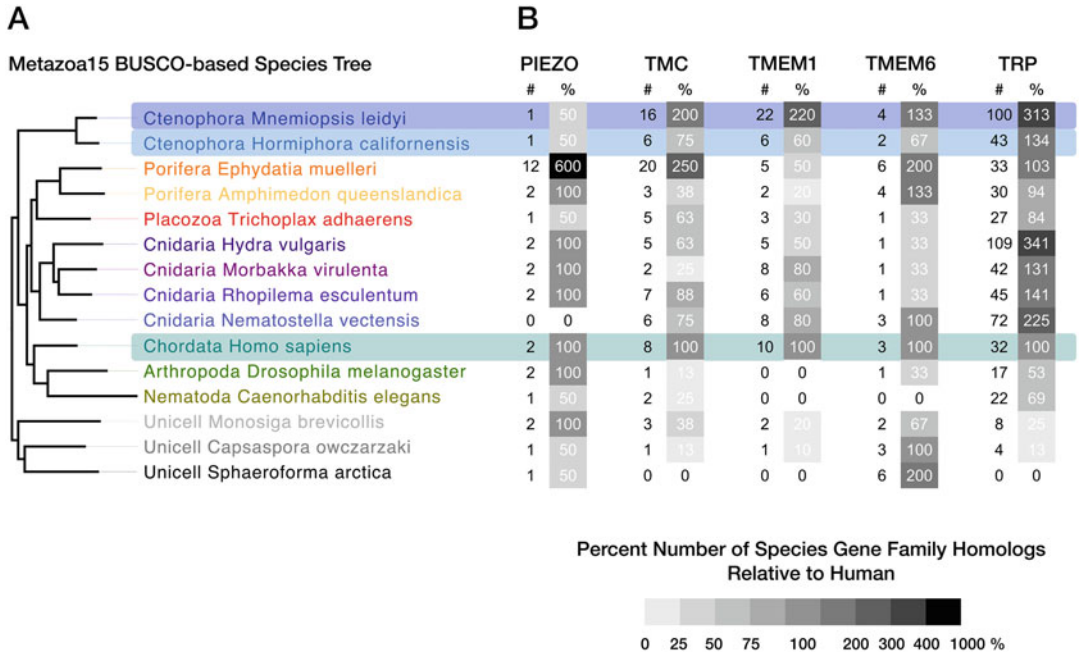

Fig. 4 BUSCO-based phylogenetic tree of Metazoa15 species evolution and associated heatmap of number and percentage relative to humans of homologs per gene family per species. (**a**) BUSCO Metazoa-based phylogenetic species tree using up to 982 genes per species in superalignment and IQTree2 maximum likelihood and C+60 model of evolution. *Note*: Ctenophores as the basal-most branch within animals is similar to recent studies that used similar methods, while the more traditional placement of sponge as the basal-most branch is commonly found using Bayesian methods. It remains unresolved which tree is correct in regard to ctenophore vs. sponge placement, but regardless, the situation highlights the importance of exploring and considering alternative methods and tools. (**b**) Heatmap of homologs per gene family per species

4. Import into spreadsheet software, such as Google Sheets, a file of traits per species you would like to map onto the species tree. The file should be structured with species names in one column and one or more columns of traits after it. For the demonstration pipeline, we include a file with counts of homologs per species genome.
5. Sort species names to match their vertical order on the species tree.
6. If useful, analyze the trait data to have additional features mapped onto the species tree.
7. If useful, generate thresholds, or use other methods to produce a heatmap that quantifies aspects of the traits based on a color scale that recolors each trait cell in the spreadsheet appropriately. This can also be done using more technical methods using tools such as Python, R, MatLab, or other programming language.
8. Export the heatmap as a pdf or other scalable vector file (Fig. 4b).
9. Combine the species tree and heatmap using any number of graphics-related software, including Google Slides, Keynote, PowerPoint, Gimp, Illustrator, or others. Using files generated by the demonstration pipeline for Piezo, we have combined our species tree and heatmap and include in the heatmap additional mechanosensitive gene families (TMC, TMEM16, TMEM63, and TRP; Fig. 4).

3.12 Summary

As highlighted above, there are many tools and alternatives to consider and work with in performing a phylogenomic analysis. Our Jupyter Notebook pipeline offers a simple push-button platform that can be used to replicate our work and generate species and gene alignments and trees for Piezo and the Metazoa15 species. We have also made it easy to use the notebook but run it for other gene families of interest. The underlying scripts provide an opportunity to build off of the platform and tailor things to a specific project and requirements. However, this increasingly requires expertise in coding and bioinformatics. Alternatively, there are a number of user-friendly desktop and online software platforms that offer extensive tools and options. In particular, ourselves and others commonly use or recommend Geneious or Galaxy for molecular biology, genomic, and phylogenomic tasks and NGPhylogeny or PhyloTol.

4 Notes

1. Ctenophores can also be collected in the surface or midwaters offshore from a boat using a plankton net or other collection device, particularly devices designed for delicate gelatinous animals.
2. A watertight bottle and Styrofoam container can also be used. The main thing is to maintain ambient seawater temperature for animals during shipping by immediately insulating the freshly collected seawater at ambient temperature.
3. For longer periods of transport or shipping, and particularly when there is a large difference in air and ideal seawater temperature, shipping boxes can include blue ice for polar to colder temperature species or heat packs for hotter temperate to tropical species. Test any shipping box for thermostability during shipping time by running a mock shipment in the lab with a temperature logger in the box. Adjust the amount of blue ice or heat packs, as needed. Also be sure to provide one or more pinhole-sized holes through any Styrofoam as otherwise oxygen will be used up by the heat packs. Travel times without aeration and with temperature control other than insulation and the addition of blue ice or heat packs can be 1–2 days without any issue.
4. Deep water and various other species can be delicate and difficult to maintain and might be processed immediately upon collection, when possible.
5. Preparing seawater from sea salt is relatively quick and easy but does require accurate measures of salinity. It is recommended to add around 80% of the sea salt, mix, and then gradually add the remaining salt while mixing and checking salinity. If time permits, the salts and seawater can be left to sit overnight, mixed, and salinity checked again the next day. Salinity can be measured with a hydrometer (inexpensive and less accurate) or refractometer (more expensive but highly accurate). For details on differences and use of each, see https://youtu.be/dQUSbruh7s4. Plunging a 1 L graduated cylinder to mix and dissolve sea salts provides highly efficient mixing.
6. Decontaminate lab bench, pipettes, tube holders, and other equipment using RNase Away or similar product that destroys RNAses. Periodically decontaminate or change gloves. Avoid breathing into tubes or solution bottles. This can be done by tilting them away from you when open. Do not leave tubes or bottles open. Open tube lids from sides without contacting the interior surface. Lids can be placed loosely on top, if helpful. Lids should be placed interior surface down on fresh Kimwipes when taken off.

7. Ctenophores are largely water, and kit recommendations for volume of homogenization buffer based on weight will often result in low final concentrations of RNA. Therefore, use 1–2× volumes for up to silver dollar-sized animals. It can be useful to test several volumes for a given species.
8. Homogenization can be done by any number of methods. For ctenophore preparations we have used variations of fresh tissue and electric or glass mortar and pestle homogenization and/or liquid nitrogen flash frozen material ground up in a ceramic mortar and pestle under liquid nitrogen and with the extraction buffer added and ground up before thawing. Avoid extended homogenization using an electric homogenizer, as buildup of heat and shearing can degrade samples.
9. There are two strategies to consider when faced with contaminated RNA at the end of extraction. If animals are not limited, you might simply redo the extraction with a new animal to see if it performs better. However some animals and samples seem to be highly resistant to producing clean samples after initial extraction. It is unclear why this is so, but we have even worked with companies making the latest kits, and they have had issues both in getting clean initial extractions. Ctenophores have been less problematic but it's worth noting. Alternatively, initially extracted contaminated samples can be cleaned up by any number of kits or older molecular biology methods. If the amount of RNA is limited, we find older methods can perform better in terms of minimizing loss of RNA or DNA after cleanup. If RNA is abundant, newer kits are quick and easy for cleanup. However, it is often the case that the amount of RNA after cleanup is greatly reduced after kit cleanup. It is unclear why this is so, but for both extraction and cleanup it may have something to do with interactions between the marine invertebrate samples and kits optimized for mammalian tissues. However, RIN values identification of ribosomal bands and assumes vertebrate sizes. However, many invertebrates have ribosomal bands that differ in size or even ribosomal molecules that separate under electrophoresis conditions and run at smaller sizes. A final evaluation of the molecular weight and distribution of the RNA smear, ribosomal bands, and RIN values might be used to make a final determination if the RNA quality should pass QC or not.
10. Ideally, three to five biological replicates will be sequenced. Here we use PacBio SMRT with Sequel II Chemistry, which offered at the time (Spring 2020) the highest single-molecule accuracy of any technology but with relatively high costs. Thus, our sequencing strategy was to use a single animal sequenced on two flow cells.

11. You will need to register with NCBI to submit data. Registering or updating your profile to link to an ORCID ID is a good idea. Data can be set to release at a later date during the submission process. It is useful to move data to NCBI once the initial processing is completed, as NCBI provides free archival storage; deposition in a public database like NCBI is generally required or encouraged in a publication, so taking care of the upload and processing now will streamline analyses and manuscript preparations later.
12. Ideally, genome gene models would be annotated after single-molecule sequencing; however, it is not always possible. Here, we provide a semi-automated method to repair candidate gene sequences after initial identification. Similar approaches could be more fully automated at genome-scale but can produce artifacts, and it might be worth considering re-annotating the genome.
13. EvidentialGene will collapse some paralogs and will fail to collapse some isoforms. These artifacts should be minimal but are an important caveat.
14. A PacBio Sequel II single-molecule transcript is likely to be a correct representation of an isoform of a gene. In contrast, genome gene models often integrate diverse data sets and methods and can be prone to artifacts, particularly in draft genome like that of *Mnemiopsis leidyi*, and can include partial, expanded, or broken gene models. For these reasons, our pipeline seeks to replace gene model sequences with transcriptome sequences when possible.
15. We have found that using genome gene model mRNA leads to many spurious Blastn hits that have surprisingly good statistics, including e-values of 0.0. Using CDS seems to greatly reduce these false positives. Potentially, the issue is related to untranslated regions in poorly called gene models, but we have not formally characterized things and do not fully understand what features of the mRNA and gene models are causing the false positive hits to arise when using mRNA for the gene models.
16. When there is only a single hit of e-value 0.0 (the best possible e-value score in Blast), the match of a gene between genome and transcriptome is clear, and the gene model can be updated to the transcriptome sequence with confidence. In cases where there are unassembled or unannotated paralogs, a transcriptome paralog could end up replacing a gene model ortholog sequence. This can be difficult to detect but should be rare and is a caveat to be aware of. Gene families with many sequence-similar paralogs will fail to be improved in these steps, as local alignments by Blastn will have highly similar or identical statistics. Importantly, BUSCO genes are generally single-gene

families, and we use them here for building a species tree. Because BUSCO genes generally have only a single gene per family, our gene model updating process will likely improve many BUSCO genes, particularly in draft genomes where gene model quality can vary a lot. Updating BUSCO gene models by the single-molecule transcriptome will improve the species tree, as gene models updated to their full sequence will perform better in alignment and tree building.

17. It is a good idea to include multiple single-cell outgroup species that are distantly related to one another but relatively close to animals for phylogenetic gene family trees of deep animal evolution. Losses of gene families do occur, and having distance can increase the odds of detecting ancient origins outside animals.
18. Metazoa15 genome datasets are from high-quality, published, and publicly available genomes: Ichthyosporea *Sphaeroforma arctica* [63], Filasterea *Capsaspora owczarzaki* [64], Choanoflagellata *Monosiga brevicollis* [65], Porifera *Ephydatia muelleri* [66], Porifera *Amphimedon queenslandica* [67], Ctenophora *Mnemiopsis leidyi* [18, 19], Ctenophora *Hormiphora californiensis* [20], Cnidaria *Nematostella vectensis* [68, 69], Cnidaria *Morbakka virulenta* [70], Cnidaria *Rhopilema esculentum* [71], Cnidaria *Hydra vulgaris* [72], Placozoa *Trichoplax adhaerens* [73], Chordata *Homo sapiens*, Arthropoda *Drosophila melanogaster*, and Nematoda *Caenorhabditis elegans*.
19. Other tools commonly used for trimming such as GBlocks [74] or TrimAl [75] can also be used; however, ClipKit provided superior trimming in our informal testing.
20. FastTree offers rapid fairly rigorous trees using maximum-likelihood-like methods. However, IQTree provides higher-quality trees, including better branch support and closer in matching to expectations. IQTree also includes ultrafast bootstraps which are statistically more rigorous and offer a declared threshold for evaluating branch support. FastTree jobs typically run seconds to hours on our machines, while the same data takes hours to days or even weeks for IQTree. IQTree is limited to the number of parameters it can explore in its model of evolution. PhyloBayes uses Bayesian inference to build trees and is not limited in parameters for its model of evolution. PhyloBayes typically takes longer than IQTree on a given data set, to the point we have killed jobs after several weeks in realizing they would take months to complete. All of the longer times are for large species trees. For most gene trees, run times will be very reasonable, in the seconds to days range in most cases.

21. There are many possible strategies and methods for identifying gene homologs in a species. Our demonstration pipeline uses what we refer to here as a blast-based reciprocal best family (RBF) strategy. It is lightweight (meaning it works well on smaller machines) and provides rapid accurate discovery of homologs. We like it because it performs well in regard to avoiding false-positive identification of homologs in genomes. In comparison with more sensitive and/or iterative detection methods, like HMMs and HMMer3 [76], Blast does miss some homologs in genomes that are highly sequence divergent from RGS sequences. More generally, most homolog detection methods have substantial overlap in true positive identification and varying levels of remote homolog detection and avoidance of false positives. In challenging cases, we find that one single method is rarely the best, so using multiple strategies can be a good idea, but it can also confound things without substantial gain. These are all things to consider or test for a given project. In regard to RBF, it is a variant of commonly used reciprocal best hit methods. Reference sequences are blasted against genomes of interest. Identified hit sequences are collected and blasted back against the genomes of the reference species. For RBF, we then keep all initial hits that have as a top hit at least one of the RGS sequences in at least one of the reference genomes. These filtered hits then form the candidate gene set. CGS and RGS sequences are subsequently combined and used for downstream phylogenetic analysis of the gene family.
22. Our demonstration pipeline is focused on deep evolution of Piezo in basal-branching animals and therefore requires the use of protein sequence, as phylogenetic signal is retained longer in protein sequences due to the redundant nature of codons. For more recent evolutionary comparisons, typically staring around 100 million years or less, DNA sequences can be considered and will be required for the most recent comparisons, as the amount of phylogenetic signal is increasingly limited (i.e., there are fewer and fewer sequence differences between genes of different species or individuals).
23. UniProt, GenBank, and other public databases are excellent sources for protein sequences that can be used to build reference gene sets. In addition, if human genes are of interest in building the reference gene set, the HUGO Genome Nomenclature Committee (HGNC) website (https://www.genenames.org/) provides curated gene groups that often represent entire gene families or superfamilies and can be useful to rapidly build reference gene sets. However, some gene groups are defined by function and include diverse gene families. In addition, a given gene can be in multiple gene groups, so care is required in selecting gene groups, but once

selected it can greatly facilitate the creation of a reference gene set for phylogenetic analysis.

24. Omissions of genes and gene families in genome gene model gene sets can occur, particularly in draft genomes, and we find that phylogenetic redundancy can be useful to compensate for these random losses. When possible, we include three distantly related species per major group, which in this case is three distantly related species within each phylum and for the unicellular outgroup.
25. RGS genes can potentially come from many sources, and their identifiers might not match those used for the same gene in their reference genome. Blasting RGS genes against the reference genomes enables identification of RGS genes in the reference genome and their associated reference genome identifiers. This is then used in later steps in sorting through initial CGS blast hits against the reference genomes. Importantly, for paralogs that are similar in sequence, or when there are highly conserved domains across homologs, it is possible to exceed the e-value sensitivity of BLAST, and RGS gene might technically be assigned the incorrect genome gene identifier internally. This is unlikely to impact the filtering of false positives based on initial CGS blast hits in the reference genomes; however, it is important to retain only genes outside RGS and then add the initial RGS sequences to the gene set prior to sequence alignment and tree building. In addition, RGS sequences on later gene family trees should be checked for sequence-identical or near-sequence-identical siblings that may represent overlooked RGS sequences that snuck through due to the limits of BLAST sensitivity.

Acknowledgments

We wish to thank the 2019 and 2020 field collectors at the University of Chicago Marine Biological Laboratory Marine Resources Center for providing animals, including S. Bennet, and similarly M. Cordeiro, as an undergraduate at Roger Williams University.
This work was supported in part by a National Institute of Mental Health of the National Institutes of Health award (MH119646) to S.C. and E.E., a National Institute of Neurological Disorders and Stroke of the National Institutes of Health award (R01NS114491) to L.L.M, a National Science Foundation award (IOS-1557923) to L.L.M, a Human Frontiers Science Program award (RGP0060/2017) to E.E. and L.L.M, Vetlesen Foundation funding, and a Connecticut Research Fund Grant (2018) award to E.E. The content is solely the authors' responsibility and does not necessarily represent official views of the funding agencies.

References

1. Hernandez-Nicaise M-L (1991) Ctenophora. In: Westfall FW, Harrison JA (eds) Microscopic anatomy of invertebrates: Placozoa, Porifera, Cnidaria, and Ctenophora. Wiley, pp 359–418
2. Redmond AK, McLysaght A (2021) Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. Nat Commun 12:1783. https://doi.org/10.1038/s41467-021-22074-7
3. Moroz LL, Romanova DY, Kohn AB (2021) Neural versus alternative integrative systems: molecular insights into origins of neurotransmitters. Philos Trans R Soc Lond Ser B Biol Sci 376:20190762. https://doi.org/10.1098/rstb.2019.0762
4. Li Y, Shen X-X, Evans B, Dunn CW, Rokas A (2021) Rooting the animal tree of life. Mol Biol Evol 38:4322–4333. https://doi.org/10.1093/molbev/msab170
5. Nielsen C (2012) Animal evolution: interrelationships of the living phyla. OUP, Oxford. Available: https://play.google.com/store/books/details?id=kr7HeXXq0o4C
6. Tamm SL (1982) Ctenophora. In: Shelton GAB (ed) Electrical conduction and behaviour in "simple" invertebrates. Clarendon Press, Oxford; New York. Available: https://www.worldcat.org/title/electrical-conduction-and-behaviour-in-simple-invertebrates/oclc/8894059
7. Nielsen C (2019) Early animal evolution: a morphologist's view. R Soc Open Sci 6:190638. https://doi.org/10.1098/rsos.190638
8. Heimbichner Goebel WL, Colin SP, Costello JH, Gemmell BJ, Sutherland KR (2020) Scaling of ctenes and consequences for swimming performance in the ctenophore *Pleurobrachia bachei*. Invertebr Biol 139:e12297. https://doi.org/10.1111/ivb.12297
9. Omori T, Ito H, Ishikawa T (2020) Swimming microorganisms acquire optimal efficiency with multiple cilia. Proc Natl Acad Sci U S A 117:30201–30207. https://doi.org/10.1073/pnas.2011146117
10. McDonald KA, Grünbaum D (2010) Swimming performance in early development and the "other" consequences of egg size for ciliated planktonic larvae. Integr Comp Biol 50:589–605. https://doi.org/10.1093/icb/icq090
11. Tamm SL (2014) Cilia and the life of ctenophores. Invertebr Biol 133:1–46. https://doi.org/10.1111/ivb.12042
12. Dunn CW, Leys SP, Haddock SHD (2015) The hidden biology of sponges and ctenophores. Trends Ecol Evol 30:282–291. https://doi.org/10.1016/j.tree.2015.03.003
13. Presnell JS, Vandepas LE, Warren KJ, Swalla BJ, Amemiya CT, Browne WE (2016) The presence of a functionally tripartite throughgut in Ctenophora has implications for metazoan character trait evolution. Curr Biol 26:2814–2820. https://doi.org/10.1016/j.cub.2016.08.019
14. Agassiz L (1850) Contributions to the natural history of the acalephæ of North America. Part I: on the naked-eyed medusæ of the shores of Massachusetts, in their perfect state of development. Mem Am Acad Arts Sci 4:221–316. https://doi.org/10.2307/25058163
15. Moroz LL (2014) The genealogy of genealogy of neurons. Commun Integr Biol 7:e993269. https://doi.org/10.4161/19420889.2014.993269
16. Moroz Leonid L, Kohn Andrea B (2016) Independent origins of neurons and synapses: insights from ctenophores. Philos Trans R Soc Lond Ser B Biol Sci 371:20150041. https://doi.org/10.1098/rstb.2015.0041
17. Moroz LL (2015) Convergent evolution of neural systems in ctenophores. J Exp Biol 218:598–611. https://doi.org/10.1242/jeb.110692
18. Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS et al (2014) The ctenophore genome and the evolutionary origins of neural systems. Nature 510:109–114. https://doi.org/10.1038/nature13400
19. Ryan JF, Pang K, Schnitzler CE, Nguyen A-D, Moreland RT, Simmons DK et al (2013) The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science 342:1242592. https://doi.org/10.1126/science.1242592
20. Schultz DT, Francis WR, McBroome JD, Christianson LM, Haddock SHD, Green RE (2021) A chromosome-scale genome assembly and karyotype of the ctenophore *Hormiphora californensis*. G3 11:jkab302. https://doi.org/10.1093/g3journal/jkab302
21. Hoencamp C, Dudchenko O, Elbatsh AMO, Brahmachari S, Raaijmakers JA, van Schaik T et al (2021) 3D genomics across the tree of life reveals condensin II as a determinant of

architecture type. Science 372:984–989. https://doi.org/10.1126/science.abe2218

22. Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G et al (2017) Ctenophore relationships and their placement as the sister group to all other animals. Nat Ecol Evol 1:1737–1746. https://doi.org/10.1038/s41559-017-0331-3
23. Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H et al (2015) Genomic data do not support comb jellies as the sister group to all other animals. Proc Natl Acad Sci U S A 112:15402–15407. https://doi.org/10.1073/pnas.1518127112
24. Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H et al (2016) Reply to Halanych et al.: Ctenophore misplacement is corroborated by independent datasets. Proc Natl Acad Sci U S A 113:E948–E949. https://doi.org/10.1073/pnas.1525718113
25. Halanych KM, Whelan NV, Kocot KM, Kohn AB, Moroz LL (2016) Miscues misplace sponges. Proc Natl Acad Sci U S A 113:E946–E947. https://doi.org/10.1073/pnas.1525332113
26. Whelan NV, Kocot KM, Moroz LL, Halanych KM (2015) Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci U S A 112:5773–5778. https://doi.org/10.1073/pnas.1503453112
27. Feuda R, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N et al (2017) Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. Curr Biol 27:3864–3870.e4. https://doi.org/10.1016/j.cub.2017.11.008
28. Jékely G, Paps J, Nielsen C (2015) The phylogenetic position of ctenophores and the origin (s) of nervous systems. EvoDevo 6:1. https://doi.org/10.1186/2041-9139-6-1
29. Rokas A (2013) Genetics. My oldest sister is a sea walnut? Science 342:1327–1329. https://doi.org/10.1126/science.1248424
30. Kapli P, Telford MJ (2020) Topology-dependent asymmetry in systematic errors affects phylogenetic placement of Ctenophora and Xenacoelomorpha. Sci Adv 6:eabc5162. https://doi.org/10.1126/sciadv.abc5162
31. Telford MJ, Moroz LL, Halanych KM (2016) Evolution: a sisterly dispute. Nature 529:286–287. https://doi.org/10.1038/529286a
32. Schultz DT, Haddock SHD, Bredeson JV, Green RE, Simakov O, Rokhsar DS (2023) Ancient gene linkages support ctenophores as sister to other animals. Nature 618:1–8. https://doi.org/10.1038/s41586-023-05936-6
33. Pandey A, Braun EL (2020) Phylogenetic analyses of sites in different protein structural environments result in distinct placements of the metazoan root. Biology 9:64. https://doi.org/10.3390/biology9040064
34. Hernandez AM, Ryan JF (2021) Six-state amino acid recoding is not an effective strategy to offset compositional heterogeneity and saturation in phylogenetic analyses. Syst Biol 70:1200. https://doi.org/10.1093/sysbio/syab027
35. Natsidis P, Kapli P, Schiffer PH, Telford MJ (2020) Systematic errors in orthology inference: a bug or a feature for evolutionary analyses? Cold Spring Harbor Laboratory, p 2020.11.03.366625. https://doi.org/10.1101/2020.11.03.366625
36. Ishino S, Ishino Y (2014) DNA polymerases as useful reagents for biotechnology—the history of developmental research in the field. Front Microbiol 5:465. https://doi.org/10.3389/fmicb.2014.00465
37. Swaminathan S (2009) GFP: the green revolution. Nat Cell Biol 11:S20. https://doi.org/10.1038/ncb1953
38. Hegemann P, Nagel G (2013) From channelrhodopsins to optogenetics. EMBO Mol Med 5:173–176. https://doi.org/10.1002/emmm.201202387
39. Ishino Y, Krupovic M, Forterre P (2018) History of CRISPR-Cas from encounter with a mysterious repeated sequence to genome editing technology. J Bacteriol 200:e00580-17. https://doi.org/10.1128/JB.00580-17
40. Raghunathan G, Marx A (2019) Identification of Thermus aquaticus DNA polymerase variants with increased mismatch discrimination and reverse transcriptase activity from a smart enzyme mutant library. Sci Rep 9:590. https://doi.org/10.1038/s41598-018-37233-y
41. Nidhi S, Anand U, Oleksak P, Tripathi P, Lal JA, Thomas G et al (2021) Novel CRISPR-Cas systems: an updated review of the current achievements, applications, and future research perspectives. Int J Mol Sci 22:3327. https://doi.org/10.3390/ijms22073327
42. Lambert GG, Depernet H, Gotthard G, Schultz DT, Navizet I, Lambert T et al (2020) *Aequorea*'s secrets revealed: new fluorescent proteins with unique properties for bioimaging and biosensing. PLoS Biol 18:e3000936. https://doi.org/10.1371/journal.pbio.3000936
43. Zabelskii D, Alekseev A, Kovalev K, Rankovic V, Balandin T, Soloviov D et al (2020) Viral rhodopsins 1 are an unique family of light-gated cation channels. Nat Commun

11:5707. https://doi.org/10.1038/s41467-020-19457-7

44. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM (2019) Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 16:1315–1322. https://doi.org/10.1038/s41592-019-0598-1
45. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G et al (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50:D439–D444. https://doi.org/10.1093/nar/gkab1061
46. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2
47. Baek M, Baker D (2022) Deep learning and protein structure modeling. Nat Methods 19:13–14. https://doi.org/10.1038/s41592-021-01360-8
48. Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S et al (2021) Computed structures of core eukaryotic protein complexes. Science 374:eabm4805. https://doi.org/10.1126/science.abm4805
49. Woodall NB, Weinberg Z, Park J, Busch F, Johnson RS, Feldbauer MJ et al (2021) De novo design of tyrosine and serine kinase-driven protein switches. Nat Struct Mol Biol 28:762–770. https://doi.org/10.1038/s41594-021-00649-8
50. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR et al (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373:871–876. https://doi.org/10.1126/science.abj8754
51. Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J et al (2021) De novo protein design by deep network hallucination. Nature 600:547–552. https://doi.org/10.1038/s41586-021-04184-w
52. Kapli P, Yang Z, Telford MJ (2020) Phylogenetic tree building in the genomic age. Nat Rev Genet 21:428. https://doi.org/10.1038/s41576-020-0233-0
53. Lewis AH, Grandl J (2021) Piezo1 ion channels inherently function as independent mechanotransducers. Elife 10:e70988. https://doi.org/10.7554/eLife.70988
54. Gilbert DG (2019). Longest protein, longest transcript or most expression, for accurate gene reconstruction of transcriptomes? BioRxiv. https://doi.org/10.1101/829184
55. Fu L, Niu B, Zhu Z et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152. https://doi.org/10.1093/bioinformatics/bts565. Epub 2012 Oct 11. PMID: 23060610; PMCID: PMC3516142
56. Seppey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. Methods Mol Biol 1962:227–245. https://doi.org/10.1007/978-1-4939-9173-0_14
57. Katoh K, Misawa K, Kuma K-I, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066. Available: http://www.ncbi.nlm.nih.gov/pubmed/12136088
58. Steenwyk JL, Buida TJ 3rd, Li Y, Shen X-X, Rokas A (2020) ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol 18:e3001007. https://doi.org/10.1371/journal.pbio.3001007
59. Price MN, Dehal PS, Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490
60. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A et al (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015
61. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288. https://doi.org/10.1093/bioinformatics/btp368
62. Letunic I, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293. https://doi.org/10.1093/nar/gkab301
63. Dudin O, Ondracka A, Grau-Bové X, Haraldsen AA, Toyoda A, Suga H et al (2019) A unicellular relative of animals generates a layer of polarized cells by actomyosin-dependent cellularization. Elife 8:e49801. https://doi.org/10.7554/eLife.49801
64. Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E et al

(2013) The *Capsaspora* genome reveals a complex unicellular prehistory of animals. Nat Commun 4:2325. https://doi.org/10.1038/ncomms3325

65. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J et al (2008) The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature 451:783–788. https://doi.org/10.1038/nature06617
66. Kenny NJ, Francis WR, Rivera-Vicéns RE, Juravel K, de Mendoza A, Díez-Vives C et al (2020) Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge *Ephydatia muelleri*. Nat Commun 11:1–11. https://doi.org/10.1038/s41467-020-17397-w
67. Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T et al (2010) The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature 466:720–726. https://doi.org/10.1038/nature09201
68. Zimmermann B, Robb SMC, Genikhovich G, Fropf WJ, Weilguny L, He S et al (2020) Sea anemone genomes reveal ancestral metazoan chromosomal macrosynteny. bioRxiv:2020.10.30.359448. https://doi.org/10.1101/2020.10.30.359448
69. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A et al (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94. https://doi.org/10.1126/science.1139158
70. Khalturin K, Shinzato C, Khalturina M, Hamada M, Fujie M, Koyanagi R et al (2019) Medusozoan genomes inform the evolution of the jellyfish body plan. Nat Ecol Evol 3:811–822. https://doi.org/10.1038/s41559-019-0853-y
71. Li Y, Gao L, Pan Y, Tian M, Li Y, He C et al (2020) Chromosome-level reference genome of the jellyfish *Rhopilema esculentum*. Gigascience 9:giaa036. https://doi.org/10.1093/gigascience/giaa036
72. Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T et al (2010) The dynamic genome of *Hydra*. Nature 464:592–596. https://doi.org/10.1038/nature08830
73. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T et al (2008) The *Trichoplax* genome and the nature of placozoans. Nature 454:955–960. https://doi.org/10.1038/nature07191
74. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577. https://doi.org/10.1080/10635150701472164
75. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348
76. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform 23:205–211. Available: https://www.ncbi.nlm.nih.gov/pubmed/20180275