scientific reports

Check for updates

OPEN

First draft genome assembly of sponge *Halisarca dujardinii* reveals key components of basement membrane and broad repertoire of aggregation factors

Ilya Borisenko[1✉], Alexander Predeus[2], Andrey Lavrov[3] & Alexander Ereskovsky[4,5]

How features characteristic of multicellular animals emerged in evolution is one of the key topics of modern evolutionary biology. We can get closer to answering that question by studying animals that occupy a sister taxon position to all other animals on the phylogenetic tree, such as sponges (Porifera). We sequenced the genome of the sponge *Halisarca dujardinii* using Oxford Nanopore and Illumina technologies and made an assembly of long reads, followed by polishing with short reads. The resulting assembly had a size of 176 Mb, and an N50 of about 785 Kb. By analyzing transposable elements in the genomes of *H. dujardinii* and a number of other sponges, we found that a significant portion of the genome (more than half for Demospongiae) is occupied by repeats, most of which are evolutionarily young. RNA-seq data were used to predict 13,989 genes in the genome, several times less than in other Demospongiae. By analyzing ortholog groups unique to *H. dujardinii* among sponges and other invertebrates, we found overrepresented genes related to the extracellular matrix. The extracellular matrix of *H. dujardinii* contains, among other components, key basement membrane components such as laminin, nidogen, fibronectin, and collagen IV, which phylogenetic analysis has confirmed that it belongs to this type of nonfibrillar collagen. In addition, we showed in *H. dujardinii* 14 aggregation factor genes responsible for cell recognition and adhesion. Our obtained assembly and annotation will further expand the understanding of genome evolution at the emergence of animal multicellularity.

The phylum Porifera occupies position sister to other metazoans on the phylogenetic tree. Although recent studies challenge their sister relationship with the rest of the Metazoa, they have long been regarded as the first-branching multicellular metazoan organisms (given the controversial position of Placozoa and Ctenophora)[1]. The phyla consist of four classes, each with characteristic features. The class Calcarea includes sponges with a well-developed skeleton composed of calcium carbonate. Sponges in the class Hexactinellida have a syncytial structure in adulthood. The class Demospongiae is the largest, and includes sponges with a skeleton made of silicon oxide or organic substance (spongin)[2]. The class Homoscleromorpha has recently been separated from the class Demospongiae, and includes sponges with the most advanced features of Eumetazoa: basement membrane-contained collagen type IV and adherens-like intercellular junctions[3–5]. Despite extensive research on sponges and significant interest in their biology, many important questions remain unanswered. For instance, how does the non-coding region of the sponge genome compare to that of vertebrates? What molecules, aside from collagen and aggregation factors, constitute the mesohyl of sponges? Why do the epithelia of some sponges possess a basement membrane while others do not? How did aggregation factors—proteins responsible for self-recognition in sponges—arise, and why are they absent in other animals?

Over the last few years, a number of sponges have emerged that can be called model species, facilitated by remarkable features of their biology. For example, sponges often contain a large number of obligate endosymbionts from among bacteria and archaea, therefore, much research has been devoted to the study of symbiosis[6,7].

[1]Biological Faculty, Department of Embryology, Saint Petersburg State University, Universitetskaya Emb. 7/9, Saint Petersburg, Russia199034. [2]Bioinformatics Institute, Kantemirovskaya Str. 2A, Saint Petersburg, Russia197342. [3]Pertsov White Sea Biological Station, Moscow State University, Leninskie Gory 1-12, Moscow, Russia119234. [4]Koltsov Institute for Developmental Biology, Russian Academy of Sciences, 26 Vavilov Str., Moscow, Russia119334. [5]Institut Méditerranéen de Biodiversité et d'Ecologie Marine et Continentale, Aix Marseille Univ, Avignon Univ, CNRS, IRD, 13007 Marseille, France. ✉email: i.borisenko@spbu.ru

Sponges (and their endosymbionts) produce secondary metabolites, which often have high biological activity and are considered potential drugs[8]. Finally, sponges have remarkable regenerative abilities: they can regenerate a small body fragment that has been removed, and the cellular mechanisms of this process differ radically between species[9]. In addition, in some species, when the sponge body is dissociated into individual cells, they are able to come together and form a primmorph, a spherical conglomerate that develops into a new sponge[9].

The first sponge genome, of *Amphimedon queenslandica*, was published in 2010[10]. Then, with a large gap, new genomes began to appear, with assembly levels from shotgun to chromosome level (*Oopsacas minuta* and *Oscarella lobularis*[11]; *Sycon ciliatum*[12]; *Tethya wilhelma*[13]; *Ephydatia muelleri*[14]; *O. minuta*[15]; and two unidentified Cladorhizid and Hexactinellid sponges[1]). In 2023/24, many sponge genomes have been assembled to chromosome level based on PacBio HiFi and Hi-C data by the Aquatic Symbiosis Genomics Project (https://www.aquaticsymbiosisgenomics.org/). Analysis of these genomes again highlighted the substantial diversity of sponges. Despite their similar lifestyles, the number of protein-coding genes and their genome sizes can differ by factors of 2.5 and 6, respectively—ranging from 16,413 genes and 61 Mb in *O. minuta* to 39,245 genes and 326 Mb in *E. muelleri*. Genomic surveys provided further evidence for the presence of true epithelia in sponges, identifying proteins involved in cell–cell adhesion and cell polarity[11]. In the contractile sponge *T. wilhelma*, genes encoding proteins for synaptic transmission, including neurotransmitter metabolic enzymes, transporters, and receptors, were identified[13]. For the calcareous sponge *S. ciliatum*, the embryonic expression of numerous transcription factor families homologous to those in Bilateria has been described[12,16–18]. However, the most prolific model remains *A. queenslandica*[19–22]. Its genome has not only provided a foundation for analyzing coding sequences but was also used to study, for the first time in sponges, the function of non-coding genomic regions that regulate gene expression[23–25].

Even within the class Demospongiae, the evolutionary distance between families is so great that their genome size and gene content can differ by several times (see below[26]). Here we report the draft assembly of the 176-Mbp genome of demosponge *Halisarca dujardinii* Johnston, 1842 (order Chondrillida) as new sponge resource. During our work, we encountered difficulties in DNA purification, sequencing, and assembly. We optimized the isolation protocol and analyze the reasons for the mentioned difficulties. We report significant characteristics of the sponge genome, including the context of genes, transposable elements, analyzed gene families within the Porifera for orthology, and evaluated the repertoire of extracellular matrix proteins with a focus on basement membrane components and aggregation factors.

Materials and methods

Sample collections

Adult H. dujardinii were collected in the area surrounding the Marine biological station «Belomorskaya» of Saint Petersburg University in the Kandalaksha Bay of the White Sea from May to September. Fucus algae with sponges were taken from the sublittoral horizon (1–5 m) using a small grapnel and placed in aquaria with circulating seawater and aeration at 12 ℃. Reproducing sponges were collected in June-July. Individuals with embryos were placed in jars with water and visually inspected several times a day for the emergence of larvae from the maternal organism. After larval emergence began, the cups were placed under a desktop lamp, and larvae were collected near the water surface. Larvae were collected manually by pipette and concentrated using a 40 μm mesh sieve. Body fragments of sponges and larvae were frozen in liquid nitrogen and stored there until nucleic acids were isolated. RNA was isolated from free-swimming larvae, cleavage stages, and regenerating individuals at 3, 12, and 24 h after injury. RNA was isolated using Lumisol (Lumiprobe, Cat. No. 91415) according manufacturer protocol and purified using spin columns (Evrogen, Cat. No. BC033). DNA was isolated using the gravity flow Genomic Tip kit (Qiagen, Cat. No. 10223) and by the phenol–chloroform method[27]. The quality of nucleic acids was assessed by agarose gel electrophoresis (1% for RNA, 0.7% for DNA) and on-chip electrophoresis (Bioanalyzer 2100, Agilent).

Sequencing

Oxford Nanopore libraries were prepared using the SQK-LSK109 (Oxford Nanopore Technologies, ONT) protocol and sequenced on two MinION flowcells (FLO-MIN106) and one PromethION flowcell (FLO-PRO002) for 72 h or until the working pore number fell below 50. The raw signal obtained as fast5 files was converted to fastq files, reflecting nucleotide sequence and signal quality, by the basecaller Guppy v5.0.11 + 2b6dbff in high-fidelity mode, using the configuration file "dna_r9.4.1_450bps_sup.cfg" (for MinION) and "dna_r9.4.1_450bps_sup_prom.cfg" (for PromethION). The data was deposited in GenBank under BioProject PRJNA1055929.

For polishing, we used genomic DNA reads previously obtained by Maja Adamska and deposited in the SRA under the number ERX1222365. DNA in this experiment was also isolated from larvae, libraries were prepared with TruSeq DNA Library Preparation Kit v2, and reads were obtained on an Illumina HiSeq 2000 instrument.

The RNA library was prepared from the polyadenylated fraction, purified with NEBNext® Poly(A) mRNA Magnetic Isolation Module (New England Biolabs), and sequenced on a NovaSeq6000 150 PE with yield of 25–35 million reads per sample. The reads have been deposited in GenBank under the accession numbers SRR27694415 (larvae), SRR27694424 (cleavage), SRR27709540 (3 h after injury), SRR27709534 (12 h after injury) and SRR27709531 (24 h after injury).

Assembly

Illumina reads were processed with the BBDuk v38.07 software to remove adapter sequences. Taxonomic composition of Illumina reads was analyzed using Kraken v2.1.2 software and the databases nt (version dated 01/27/2020) and k2_pluspf_20210517 (version dated 05/17/2021; including *Homo sapiens* and all bacteria, archaea, viruses, and yeasts known to RefSeq).

All Oxford Nanopore reads were also combined, and filtered using the program Filtlong v0.2.0 using the cleaned Illumina sequences as a reference. We chose 5 popular assemblers for assembly: Flye v2.9-b1768, Canu v2.1.1, Shasta v0.7.0, Raven v1.5.1, and Wtdbg2 (Redbean) v2.5. The assembly obtained using Shasta with a minimal read length of 5000 bp was used for further polishing and analysis. For polishing we used the programs Racon v1.4.3 (with polishing by ONT and Illumina reads), Polca v4.0.9 and Pilon v1.24 (with polishing by Illumina reads). To assess the quality of the assembly, we used the program busco v5.2.2 with the "–augustus" option to find de novo genes in the studied genomes. The "metazoan" collection (954 genes) from OrthoDB v10 was used as a database. The most complete Benchmarking Universal Single-Copy Orthologue (BUSCO) values were obtained with a single round of polishing with Illumina reads using Racon. This assembly has been deposited in NCBI WGS under the accession number JBKORC000000000 and was used in further analyses. The parameters to run each application are given in the Github repository (https://github.com/mezoderma/Hdu-genome). The assembly was analysed using the BlobPlot and CovPlot commands from BlobTools v1.1.1. To assess the taxonomic identity of contigs, we performed both a blastn[28] v2.12.0 + search against the nt database (date of download 06/17/2024) and a DIAMOND[29] blastx v2.0.15 search (Buchfink et al. 2015) against the Uniprot database (release 2024_03), assembled in the manner described in the pipeline file. The obtained bam file and BLAST output files were used as input files for BlobTools. To assess whether contigs belong to the sponge we also used mapping of RNA-seq data (see above) to the assembly[30]. First, we mapped a previously assembled de novo transcriptome[31] to the assembly using gmap[32] v2012-12–17, and calculated the fraction of bases covered by RNA-seq reads using samtools coverage. Second, we mapped RNA-seq reads from five independent experiments to assembly using bowtie2 and visualised the mapping as bigwig files using IGV genome viewer[33]. Contigs defined by BlobTools other than sponge's or no-hit were manually checked with regards to these two metrics.

Repeat analysis

Repetitive elements in the genome were analyzed using the Earl Grey pipeline[34], which includes RepeatModeler2, RepeatMasker v4.1.2, and LTR_Finder v1.07. This tool allows de novo identification of transposable elements (TEs), clustering them into families, classifying the families against the open Dfam database, and using the resulting library to mask repeats in the genome. Custom repeat libraries were created for the *H. dujardinii* genome, *Spongilla lacustris*, *Ephydatia muelleri*, *Oscarella lobularis*, *Aphrocallistes beatrix*, *Oopsacas minuta* and *Sycon ciliatum*. References for the genomes used are provided in the Supplementary Table S1.

Transcriptome assembly and gene prediction

Transcriptome data used for training and as evidence during gene prediction. Raw paired-end sequences of cDNA were quality filtered and trimmed using Trimmomatic v0.32 with default settings. Unpaired reads were discarded. Pooled reads of five samples (larvae, embryos, regenerates at 3, 12 and 24 h) were assembled de novo using Trinity 2.13.2. The assembled transcriptome was cleaned by aligning reads of genomic DNA from larvae. The alignments were done using bowtie2 v2.4.4 with default parameters. Transcriptome contigs not aligned to any of the gDNA read were removed from the assembly. Contigs were clustered and merged with cd-hit-est v4.8.1 with identity threshold 0.95.

Another transcriptome was assembled in genome guided manner. The same five cDNA libraries were mapped to the genome using STAR v2.7.10a. The resulting.bam file was used as input for Trinity v2.15.1 in genome guided mode. The resulting contigs were also clustered with cd-hit-est with the same threshold.

An additional set of transcripts was generated with StringTie v2.2.1. Each cDNA library was independently mapped to the genome with STAR, and the bam file used as input for StringTie. The resulting gtf were assembled with TACO v0.7.3 with default settings. These TACO-produced transcript coordinates provided to PASA pipeline together with de novo and genome guided transcriptomes[35], where only transcripts with more than 90% transcript coverage (parameter –MIN_PERCENT_ALIGNED) and 95% identity (parameter –MIN_AVG_PER_ID) to the genome were merged.

Ab initio gene predictions were performed on the repeat-masked assembly with BRAKER v3.0.6 pipeline based on two different programs—Augustus v3.5.0[36] and Genemark-ES v2.3e[37]. The same five cDNA libraries mapped to the genome were used for training and as evidence in gene prediction.

Gene evidence and predicted gene structure were combined using EVM[35]. The evidence included a) ab initio predictions generated by Augustus and GeneMark, b) PASA generated consensus transcript assemblies based on de novo assembly, genome guided assembly and StringTie transcripts, and c) the miniprot v0.12 generated alignment of UniProtKB/Swiss-Prot (2023-11-08) to the genome. The evidence weights for EVM were set as follows: transcripts assembled by PASA—10, miniprot alignment–2, ab initio genes prediction—1. UTRs were added onto the gene models predicted by EVM by two sequential PASA rounds including annotation loading, annotation comparison and annotation updates to maximize incorporation onto gene models predicted by EVM, as per the suggestion of the authors in the PASA pipeline manual (http://pasapipeline.github.io/). Gene models together with predicted proteins and Trinotate report are uploaded at the Github repository (https://github.com/mezoderma/Hdu-genome).

Gene annotation and orthology analysis

Open reading frames (ORFs) for all genes were predicted using Transdecoder v5.7.1. All best ORF candidates were analysed for protein domains, signal sequences and transmembrane domain using hmmer 3.0, signalp 6, diamond v2.0.15.153 and tmhmm 2.0, and combined to annotate each ORF using Trinotate v4.0.2. Protein sequences for the ortholog search were downloaded from the corresponding genomic projects in GenBank for *A. queenslandica*, *E. muelleri*, *O. lobularis*, *O. minuta, S. ciliatum, Trichoplax adhaerans, Nematostella vectensis* and *Branchiostoma floridae* (Supplementary Table S1). Orthogroups were constructed using OrthoFinder v2.5.5,

and analysis results were visualized using the UpSetR library for R. GO enrichment performed with GOseq v1.54.0; GO terms above threshold 0.0001 were visualized with REVIGO[38].

ECM proteins and aggregation factor analysis

Sequences of fibrillar collagens, type IV collagen, and spongin for *H. dujardinii*, *A. queenslandica*, and *E. muelleri* were obtained using blastp from the proteomes of these species. Vertebrate collagens from Swiss-Prot (Supplementary Table S7) and short-chain collagen from *E. muelleri*[39] (Uniprot accession P18503) were used as a query. Proteins containing conserved domains (NC1 and Wreath) were searched using HMMER 3.4 (hmmscan command) and models available in the Pfam database (PF01410 and PF01413 for NC1) or previously published[40] (for Whreath) under standard settings. Collagen sequences were aligned using MUSCLE v3.8.425 and trimmed manually. Alignments for NC1 domains are given in Supplementary File S1. ML trees were reconstructed using IQ-Tree[41] v1.6.12 with Vt + G4 model selected with ModelFinder[42] implemented in the IQ-Tree.

Wreath domain sequences were aligned with MAFFT v7.526 with L-INS-i option and trimmed using trimAl v1.5.0 with the -gappyout option. Alignments for Wreath domains are given in Supplementary File S2. Model was predicted from alignment with ModelFinder[42] implemented in the IQ-Tree. Phylogenetic tree was built using Mr Bayes v3.2.7a under WAG + I + G4 model with 5,000,000 generations using four chains. Standard deviation of split below 0.01 was reached before end of run. The final tree was drawn using FigTree v1.4.4.

Results and discussion

DNA extraction and sequencing

The adult sponge *Halisarca dujardinii* (Fig. 1A) contains a huge number of cells per unit volume of tissue, so it is a fruitful target for genomic DNA isolation. Thus, we obtained DNA with an average molecular weight of more than 150 kb and A260/A230 > 2.1 with classical phenol–chloroform purification[27,43] and chromatography on gravity-flow columns (Genomic-tip kit, QIAGEN).

In pilot sequencing experiments on MinION, we used the Rapid Sequencing kit (SQK-RAD004), which allows to obtain the maximum read length. However, in two cell experiments alongside a positive control, we found that the sequencing of genomic DNA was inhibited by some substance accompanying the purified DNA. Adapter ligation did not occur during library preparation for Illumina sequencing, and we hypothesized that an unknown substance interferes with ligase activity during library preparation for both ONT and Illumina.

In sponges, a large number of secondary metabolites with different biological activities are known[44]. At least some of these probably serve a protective function for the immobile sponge, whose defense capabilities are severely limited. Assuming that the interfering sequencing agent may be a secondary metabolite of the adult sponge or associated with the extracellular matrix, we isolated genomic DNA from larvae.

Genomic DNA purified from larvae was successfully sequenced on ONT and Illumina. In ONT, we have favored 1D chemistry for library preparation, which, compared to Rapid ligation kit, allows us to obtain shorter read lengths but a larger amount of data. Sequencing on two MinION cells and one PromethION cell resulted in 26.3 Gb of reads. Using only reads that passed the pass/fail filter, we kept 90% of the longest reads, yielding 19 Gb of reads with an N50 of about 17 kb after the filter. The data was deposited in GenBank under BioProject PRJNA1055929.

It is worth noting that the Oxford Nanopore sequencing was performed using the R9.4.1 pore and the SQK-LSK109 ligation kit. While this chemistry has a higher raw error rate compared to the newer R10 versions,

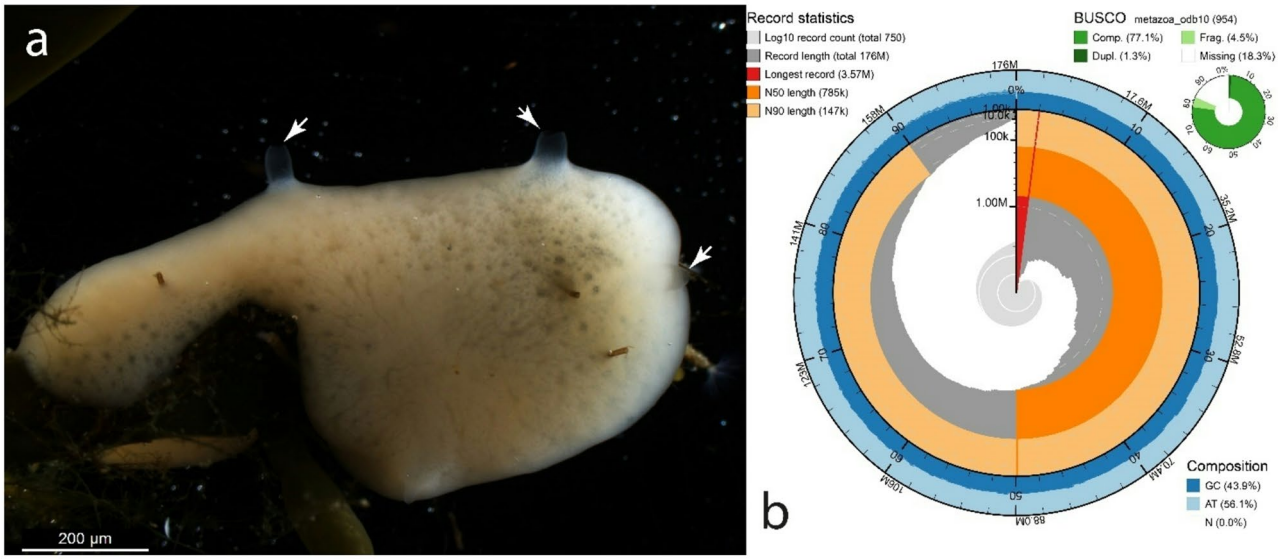

Fig. 1. Demosponge *H. dujardinii*. (**a**) Adult animal in vivo; oscular tubes are shown by arrows. (**b**) The snail plot shows that the final version of the *H. dujardinii* assembly has N50 of 785 Kb, the longest contig is 3.57 Mb long.

especially in homopolymer regions, we mitigated this limitation by using high-accuracy basecalling and, most importantly, by polishing the final assembly with high-quality Illumina short reads, which effectively corrects for such indel-type errors.

Contamination of genomic DNA by reads originating from inevitably present microorganisms and viruses is often a significant source of problems in the assembly, annotation, and analysis of eukaryotic genomes. We analyzed the taxonomic composition of Illumina reads using Kraken2 software. Analysis of Oxford Nanopore reads using Kraken2 is unreliable due to the high error rate in the reads. The result showed little contamination with other species: only 0.01% of the reads were attributed to archaea, 1.10% to bacteria, 0.20% to yeasts, 0.50% to unicellular algae, and 0.13% to viruses. Thus, contamination of the study samples with microorganisms is minimal and we did not expect interference with genome assembly and annotation.

To assess taxonomic purity, we used BlobTools by combining BLAST and read mapping results. At the phylum level, contigs were classified as Porifera (62.4%), no-hit (34.8%), Chordata (2.4% or 18 contigs), Arthropoda (0.26% or two contigs) and Ascomycota (0.13% or one contig). The average GC-content in each of these taxa was 42%. Unexpectedly, the analysis revealed such a taxonomic distribution, given the average size of contigs assigned to Chordata and Arthropoda of 500–600 Kb. The blobplots clearly show that contigs have minimal scatter in GC-content (Fig. 2A, B). We therefore mapped de novo assembled transcripts and RNA-seq reads to the assembly, and additionally examined contigs classified as Chordata, Arthropoda and Ascomycota. From the bam file of mapped transcripts, we extracted the coverage value, the percentage of covered bases of a genomic contig, using the coverage command from samtools. Figure 2C shows that the 'foreign' contigs have transcript coverage ranging from 35 to 85 percent. Most of them (16 out of 21) have coverage greater than 50% of their length. It can also be seen that RNA-seq reads obtained from independent experiments at different times and from different material (e.g., regenerates and larvae) map onto the same contig regions (Fig. 2D). Taken together, these results strongly suggest that the contigs belong to the sponge rather than contaminants.

Genome size estimation

Illumina reads were used to estimate genome size as ONT reads contain too many errors for such an analysis. There are several ways to estimate genome size, both experimental (based on DNA content in the nucleus) and based on genomic reads (pre- and post-assembly). Estimation of genome size and properties before assembly based on *k*-mer distribution is, de facto, a common step in genome assembly.

We used the actively supported software GenomeScope 2.0, which allows us to obtain estimates of genome size, degree of heterozygosity, and number of repeats present. The good (approximately 200-fold) coverage of the genome by Illumina reads allowed us to obtain a stable model that does not change when the size of the *k*-mer used changes from 17 to 27. The estimation results for $k = 21$ is given in Supplementary Fig. S1 and Supplementary Table S2. We determined that the haploid genome size is 188 Mb.

A recent publication described a preliminary and unpublished assembly of the *H. dujardinii* genome, estimating a genome size of 194 Mb based on 44-fold coverage of Illumina reads[45]. This estimate was made using the Kmergenie program. The recently assembled genome of a closely related species, *Halisarca caerulea*, published in GenBank, is estimated to be 195.7 Mb. Results obtained in different laboratories by different methods showed a remarkable similarity in the estimation of haploid genome size. Heterozygosity of the genome was estimated at 0.62%; it should be noted, however, that this estimate is based on *k*-mers and is not capable of estimating heterozygosity of repeat regions.

Genome assembly

Given the recommendations for hybrid assembly, the high coverage (about 100 × Oxford Nanopore and about 200 × Illumina reads) conditioned the chosen strategy, namely assembling the genome from long reads, followed by polishing the genome with Illumina reads. We chose 5 popular assemblers for assembly: Flye, Canu, Shasta, Raven and Wtdbg2 (Redbean). The results of the assemblies are presented at the Fig. 3A and Supplementary Table S3.

The results above show that the assembly is problematic: the total length of the haploid assembly differs between some programs by three times. Evaluation with BUSCO software shows that a significant part of the classical single-copy orthologs characteristic of animals is missing from the assembly. We chose the assembly made by the program Shasta with a minimal read length of 5000 bp, which has the one of the highest BUSCO levels.

Despite the low estimate of heterozygosity from *k*-mer analysis, the high taxonomic purity of the reads and the high coverage of both long and short reads lead us to infer that the *H. dujardinii* genome contains a significant number of evolutionarily young repeats that are present in a highly heterozygous state. This is indirectly evidenced by the diagram of contig coverage distribution obtained by Flye assembly (Supplementary Figure S2): we can see a bimodal distribution, as well as a significant number of poorly covered contigs, which are assembly or alignment artifacts.

We applied long-read polishing using the Racon software to improve the sequence quality within resulting assemblies. Nine rounds of iterative polishing were performed. Unfortunately, there was little improvement in the assemblies; after a slight improvement in the first round of polishing, BUSCO values fluctuated around some equilibrium value (Supplementary Figure S3).

Polishing with Illumina short reads improved the assembly to some extent. We tried sequential polishing with Racon, Polca, and Pilon software. An increase in BUSCO values was observed after the first round but decreased in subsequent iterations. Therefore, in the final version, we left only one round of polishing with short reads using Racon (see materials and methods section for details). The BUSCO value for the *H. dujardinii* assembly was 77% for complete gene models (C:77.0% [S:75.7%, D:1.3%], F:4.7%, M:18.3%, n:954), and with the inclusion of partial sequences, the genome accounted for 81.7% of the metazoan reference gene set. The size

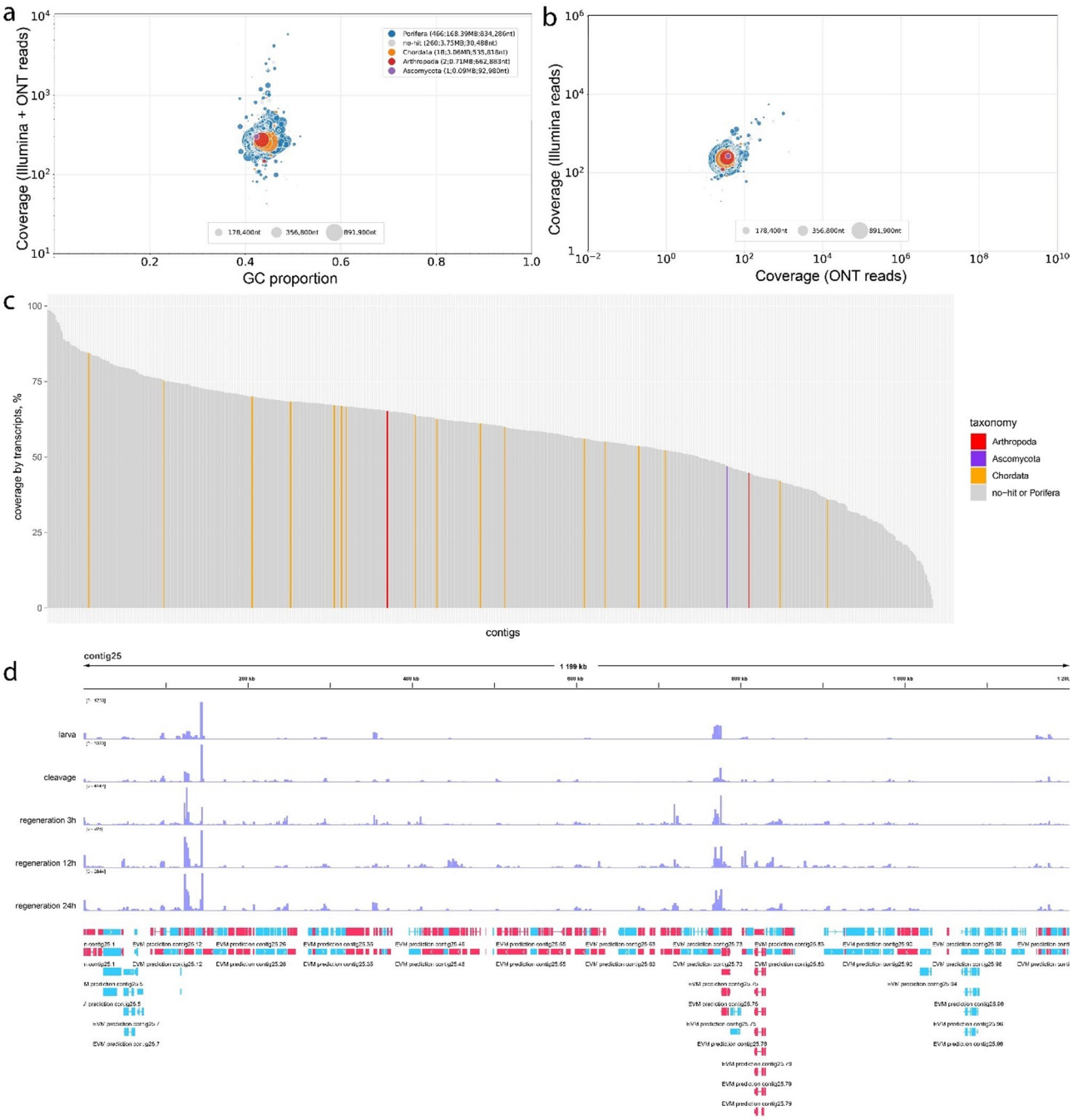

Fig. 2. Analysis of *H. dujardinii* genome assembly for contamination. (**a** and **b**) Blobplots of the genome assembly. Each circle represents a contig, with its size proportional to the length of the contig. Each circle is plotted by GC proportion and coverage with sum of Illumina and ONT reads (**a**) or by coverage of Illumina vs coverage of ONT reads (**b**). The legend detail taxonomic identity according BlobTools; the number of sequences, their combined length and N50 shown in parentheses. (**c**) Coverage of each contig with transcripts from the de novo assembled transcriptome. The contigs are sorted by coverage. Contigs belonging to sponges or whose taxonomic affiliation has not been determined by BlobTools are shown in grey. Suggested contigs of Arthropoda, Ascomyceta and Chordata are shown in different colors. (**d**) Mapping of RNA-seq reads from five different experiments to contig 25, presumably related to Chordata. The RNA-seq reads lie at the same ORF-matched locations in the genome. Gene models are shown below the mapping histograms (blue on the positive strand, red on the negative strand).

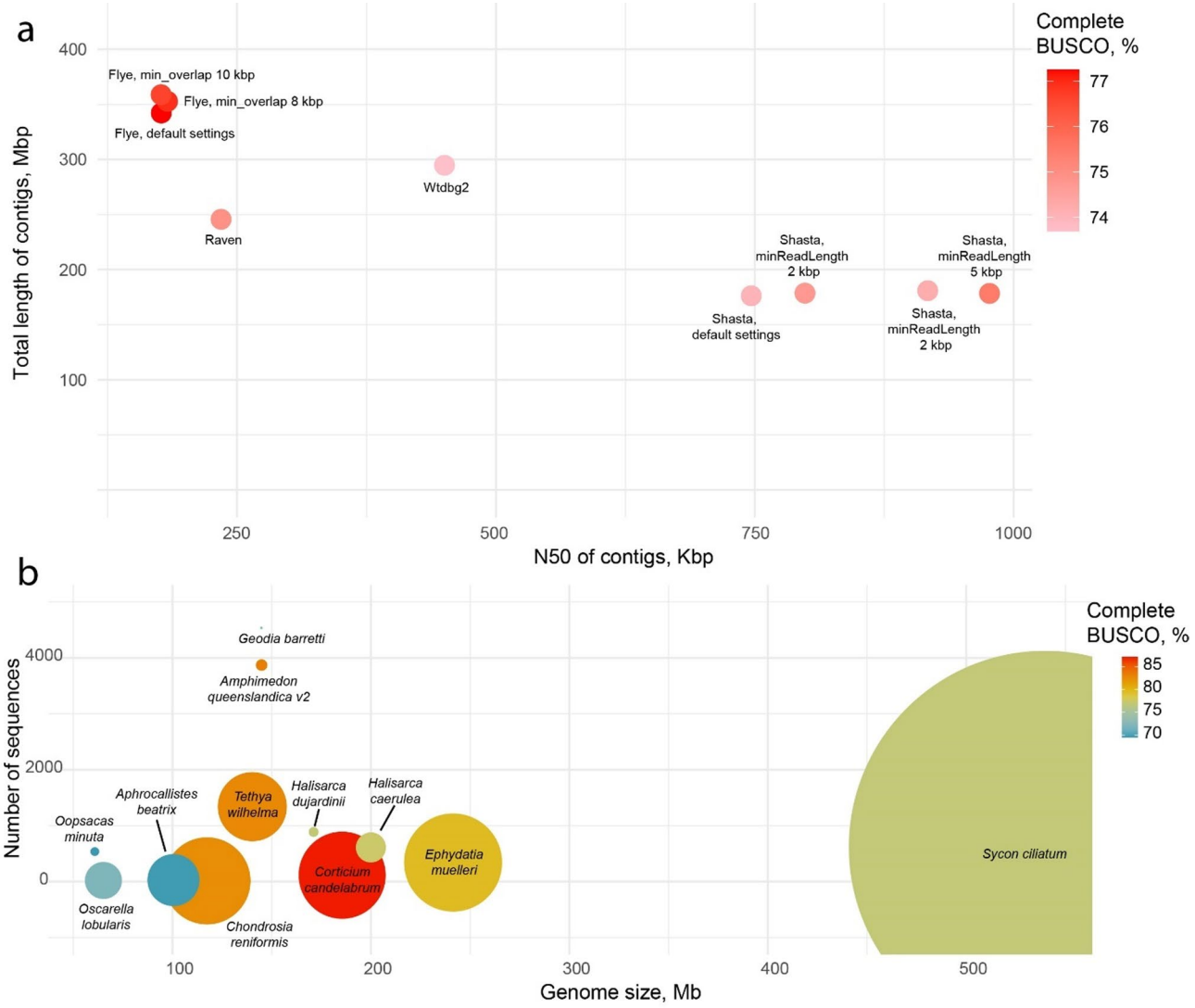

Fig. 3. (**a**) The results of assembling Oxford Nanopore reads with different assemblers. Each circle represents an assembly, the X-axis position represents the N50 value for that assembly, and the Y-axis represents the total length of contigs in the assembly. (**b**) Comparison of sponge genome assemblies from GenBank. Each circle represents an assembly, the X-axis position represents the genome size for that assembly, the Y-axis represents the total number of sequences (contigs or scaffolds) in the assembly, and the size of the circle shows the value of N50. The color shows the proportion of complete genes from the BUSCO set.

of the assembly after polishing was 176 Mb, N50 value was about 785 Kb, and it contains 742 contigs (Fig. 1B). This assembly was deposed in WGS NCBI repository under accession JBKORC000000000, and used for gene models prediction and repetitive elements analysis. We compared parameters of available sponge genome assemblies such as the number of contigs or scaffolds, genome size, N50, and proportion of complete BUSCO (Fig. 3B). While contiguity and completeness approach those of some PacBio-based assemblies, the genome remains fragmented (742 contigs). It should be noted that sponges in general are not characterised by a very high proportion of complete BUSCO (around 80%), as can be seen from other qualitatively assembled genomes. This indicates the completeness of the assembly we obtained.

Repetitive elements analysis

Studies focusing on genome evolution are based mainly on the comparison of protein-coding genes: on the search for families common to different taxa, lineage-specific expansion or loss. At the same time, it is known that in the majority of Metazoa, the transposable elements (TEs) occupy a significant part of the genome and are actively involved in genome evolution. Only a few studies have analyzed TEs of the sponge genome[14,15,46]. The diversity and role of these genetic elements in such ancient multicellular organisms as sponges deserve more attention.

The TE repertoire of four sponges, *E. muelleri, T. wilhelma, A. queenslandica*, and *S. ciliatum*, have previously been analyzed using de novo repeat modeling[14]. Repeats have been shown to occupy a substantial portion of

the sponge genome (29–49%), and most of them are currently unclassified. It is noticeable that the expansion of different classes of repeats occurs specifically within different taxa. Kenny et al. (2020) provide a logical hypothesis that genome size correlates with the quantity of repeats in the genome, and the proportion of repeats is higher in larger genomes[14]. Interspersed repeats represent 34% of the *O. minuta* genome, and these are most commonly represented by unclassified repeats (18%) and DNA transposons (15%)[15]. The *A. queenslandica* genome contains 43.2% repetitive elements, many of which are DNA transposons (4.16%). Among these non-autonomous DNA transposons, two miniaturized transposable elements, MITEs, named Queen1 and Queen2, have been described[46]. They probably originate from the Tc1/Mariner-like repeats family. Usually MITEs are low-copy numbers, but in *A. queenslandica*, they are abundant in the genome (3800 and 1700 repeats, respectively), and Queen1/Queen2 have not been detected in the genomes of other animals. MITEs can introduce new splice sites, providing domain shuffling and inducing the formation of new multidomain proteins.

We analyzed the diversity of TEs in the genomes of a number of sponges belonging to all four classes: Demospongiae (*Halisarca dujardinii*, *Ephydatia muelleri*, and *Spongilla lacustris*), Calcarea (*Sycon ciliatum*), Homoscleromorpha (*Oscarella lobularis*), and Hexactinellida (*Aphrocallistes beatrix* and *Oopsacas minuta*). TEs were identified de novo and classified by comparison with the Dfam database using a new pipeline based on RepeatModeler, Earl Grey (Supplementary Table S4). Overall, repeats span from 23 to 59% of the genome in sponges (Fig. 4). In the genome of *H. dujardinii*, TEs occupy 53.67% of the current assembly. It should be noted that the genome may actually be larger in size because long repeats cause difficulties during assembly. Therefore, at the current assembly level, we cannot judge the actual proportion of repeats in the genome—only in the

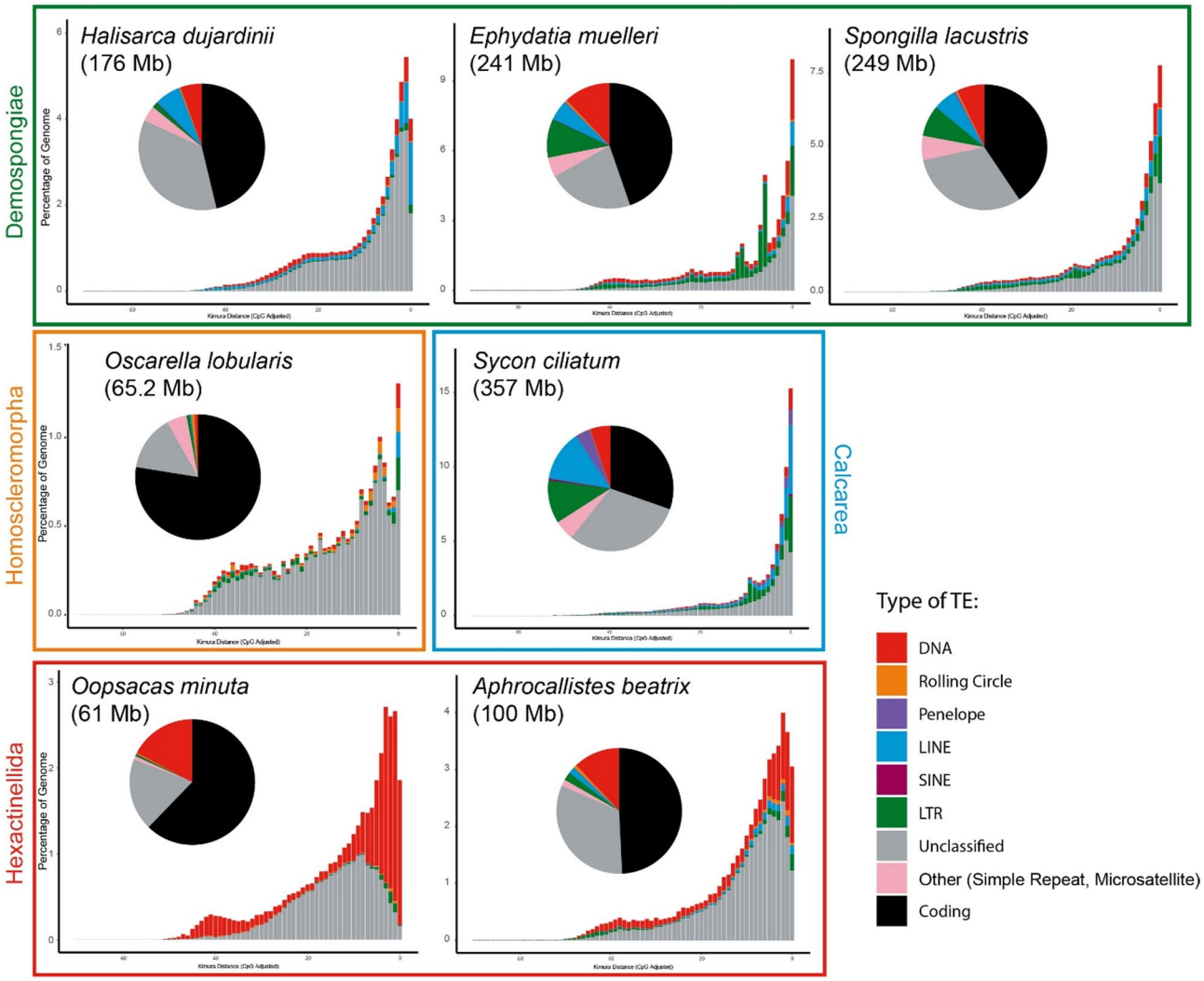

Fig. 4. Repeat landscape plots illustrating TE accumulation history for genome of sponges from different classes, based on Kimura distance-based copy divergence analyses, with sequence divergence (CpG adjusted Kimura substitution level) illustrated on the x-axis, percentage of the genome represented by each TE type on the y-axis, and transposon type indicated by the pie chart on the right-hand side. Major TEs: LINE, long interspersed nuclear element; LTR, long terminal repeat; Penelope, family of retrotransposons; Rolling Circle, family of Helitron transposons; SINE, short interspersed nuclear element.

presented assembly version. The de novo library included 7450 families of repeats. Most of them (approximately 2/3, or 35.7% of the assembly) are represented by repeats that could not be assigned to any described family. The identified TEs of *H. dujardinii* are dominated by retrotransposons of the SLACS CRE and BovB families belonging to LINE, and DNA transposons of the Tourist and Harbinger superfamilies, as well as the Pogo superfamily of the IS630-Tc1-mariner group.

In general, we were unable to describe any taxon-specific expansion of a particular class of repeats in individual sponge species or classes. It can be seen that different species show dominance of different groups and even superfamilies of repeats. For example, class Hexactinellida shows the prevalence of DNA transposons, whereas demosponges and calcareous sponges are dominated by retroelements like LINE and LTR. The genome of *O. lobularis* is generally very poor in repeats (22.28%). The genome of *O. minuta* is almost devoid of retroelements. SINEs were not identified in a number of sponge species (*A. beatrix*, *O. lobularis*, *E. muelleri*, and *S. lacustris*). At the same time, it should be taken into account that a substantial part of TEs remains unclassified, from 14 to 35%, and further studies are needed to elucidate the complete pattern of TE distribution within Porifera.

The interspersed repeat landscape demonstrates very recent (possibly ongoing) expansion of TEs in sponges (Fig. 4). The plot shows how many copies of each repeat are present in the genome as a function of the Kimura distance, which represents the genomic distance between the repeat and the consensus sequence. The distance increases with repeat divergence, *i.e.*, more recent repeats have a small Kimura distance. Peaks show large groups of copies of repeats with the same divergence and indicate a major expansion event of these elements. Broad peaks with a high Kimura distance are characteristic of ancient TEs, and sharp peaks with a small distance are characteristic of young repeats. All sponges studied in this work are characterized by a sharp peak at a small Kimura distance, with substitution level ranging from 1 to 4. Most species have a second small peak between 20 and 42. Interestingly, several peaks are observed in *E. muelleri*, the first two of which are formed due to the expansion of LTR elements.

Evidence-based protein-coding gene annotation

To obtain the most complete set of expressed genes, we sequenced polyadenylated RNA from larvae, regenerating samples, and embryos at different developmental stages. We used RNA from an animal containing embryos or undergoing regeneration to assemble the transcriptome, as this can activate genes that are silent in the intact adult animal. A total of 5 libraries amounting to 63,9 Gb were used to assemble the transcriptome.

For annotation, we modified the strategy described for *Amphimedon*[47] (Fig. 5A). Thus, we combined using PASA the representative de novo assembly, genome-guided Trinity assembly, and independent genome-guided assembly for different stages generated with StringTie. The addition of StringTie-generated transcripts strongly improved the assembly: when these were included in the PASA pipeline, BUSCO scores were elevated from 70 to 88.8% (complete plus fragmented). To predict gene models, we used BRAKER3, an Augustus and GeneMark-ES-based pipeline that automatically performs model training and uses RNA/protein data as evidence. We compared models generated using RNA-seq and protein data (the longest ORFs isolated from the de novo transcriptome) and found that using RNA-seq data yielded a more complete set of BUSCOs (90.1% for RNA-seq, 86.5% for proteins; complete plus fragmented BUSCOs). A nearly identical BUSCO value (90.1%; complete plus fragmented BUSCOs) was determined for the *E. muelleri* chromosome size assembly made with PacBio and Hi-C[14].

All assembled transcripts, gene models, and protein-to-genome alignment were combined using EVM to predict protein-coding gene models[35]. The completed set of *H. dujardinii* genes contains a total of 13,989 transcripts, which includes alternatively spliced gene isoforms that resulted in 18,841 mRNAs. This is a relatively

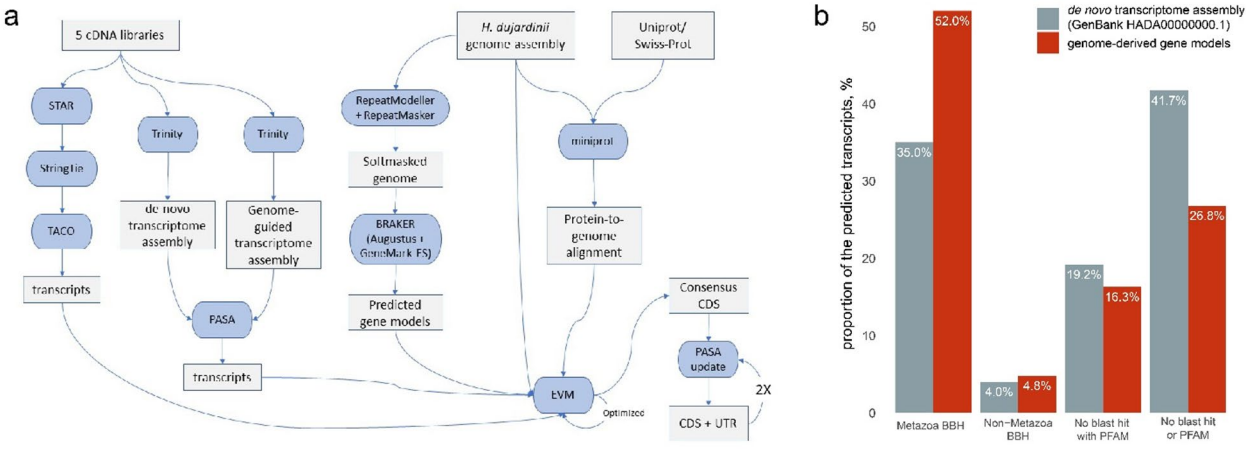

Fig. 5. (**a**) Annotation pipeline. Combined data of RNA-seq generated from embryo, larva and regenerates of *H. dujardinii* were used to annotate the genome. This resulted in the annotation of CDS and UTRs. (**b**) BLASTP BBH annotation comparison. The y-axis shows the percentage of the total number of predicted transcripts. Also shown are proteins that have no significant blast hit yet contain an identifiable Pfam domain, and proteins with no significant blast hit and no identifiable Pfam domain.

	Demospongiae			Homoscleromorpha	Hexactinellida	Calcarea
	Amphimedon queenslandica v1.1 with annotation v2[47]	*Halisarca dujardinii*	*Ephydatia muelleri*[14]	*Oscarella lobularis* (GenBank GCA_947507565.1)	*Oopsacas minuta*[15]	*Sycon ciliatum* (GenBank GCF_964019385.1)
Genome size	164,2 Mb	176 Mb	326 Mb	65,2 Mb	61,4 Mb	539,6 Mb
N50	123,2 Kb	785 Kb	9883 Kb	265,4 Kb	670 Kb	39 Mb
Number of sequences	13 132	742	1 445	21	365	618
Total gene number	40 122	13 989	39 245	16 806	16 340	26 833
Mean gene length	2 521 bp	9162 bp	4507 bp	3 266 bp	1787 bp	14,653 bp
Total mRNA number	47 895	18 841	39 245	20 875	16 413	25 112
Smallest gene	149 bp	150 bp	261 bp	312 bp	130 bp	225 bp
Longest gene	103 992 bp	194 383 bp	116 520 bp	100 729 bp	98 520 bp	579,302 bp
Total exon number	206 984	160 871	265 864	249 730	42 289	237 557
Mean exon length	225 bp	239 bp	293 bp	226 bp	496 bp	314 bp
Mean exons per gene	5,2	8,5	6,8	8,6	2,6	7,9
Total intron in CDS number	166 862	137 686	211 658	204 448	25 876	205 163
Mean intron in CDS length	327 bp	1 073 bp	388 bp	80 bp	341 bp	1 558 bp
CDS						
Total length	42,6 Mb	29,2 Mb	58,7 Mb	43,8 Mb	20,9 Mb	53,4 Mb
Average size	1 062 bp	1 550 bp	1 496 bp	2 099 bp	1 278 bp	2 127 bp
Longest	56 682 bp	39 570 bp	113 766 bp	82 071 bp	41 427 bp	148 583 bp
5' UTRs						
Total number	9873	9092	39,144	19,129		20 989
Total length	1,13 Mb	4,1 Mb	10,0 Mb	6,8 Mb		6,5 Mb
Average size	114 bp	354 bp	257 bp	193 bp		248 bp
Longest	4952 bp	9 062 bp	3 033 bp	8 349 bp		9 623 bp
3' UTRs						
Total number	14,892	9140	39,170	19,641		20 932
Total length	2,8 Mb	5,2 Mb	9,3 Mb	6,0 Mb		14,8 Mb
Average size	188 bp	474 bp	238 bp	216 bp		648 bp
Longest	2 814 bp	6 749 bp	2 846 bp	10 522 bp		10 548 bp

Table 1. Genome assembly statistics for different sponge species.

	H. dujardinii transcriptome assembly, v1 (GenBank HADA00000000.1)	Genome-derived gene models
# of genes	105 862	13 989
# of transcripts	138 992	18 841
# of ORFs	40 754	18 841
BUSCO	Complete 90.9% (single 16.9%, duplicated 74.0%), fragmented 2.5%, missed 6.6%, total number 954	Complete 85.8% (single 84.1%, duplicated 1.7%), fragmented 4.3%, missed 9.9%, total number 954
Metazoa BBH	14 273	9 803
Non-metazoa BBH	1 648	909
No blast hit with PFAM	7 827	3 074
No blast hit or PFAM	17 006	5 055

Table 2. Comparison of completeness and composition of two *H. dujardinii* predicted gene models.

small number of genes for known sponge genomes (Table 1). In our annotation, 44% (8251) of all RNAs have both UTRs, and 53% (9981) of RNAs have at least one UTR. This is higher than in the last published annotation of *Amphimedon* genome, which was the best among sponges (37% of RNAs have at least one UTR)[47]. In addition, de novo assembly of the transcriptome contains a large number of artefacts—fragmented and merged transcripts, errors, and contaminations. We significantly reduced this false diversity by predicting gene models. This is confirmed by an increase in the proportion of transcripts with BBH and a decrease in the number of transcripts without BBH or PFAM domains (Fig. 5B, Table 2).

Orthology analysis

We compared the predicted proteins of *H. dujardinii* with the proteomes of sponges and some other Metazoa using OrthoFinder, a tool for finding orthologs and grouping them into groups (orthogroups)[48]. In addition to *H. dujardinii*, we took *A. queenslandica*, *E. muelleri*, *O. lobularis*, *S. ciliatum*, and *O. minuta* for analysis (references to protein models are given in Supplementary Table S1). First of all, we can see that the number of orthogroups in a species correlates with genome size (Fig. 6A). Thus, the highest number of orthogroups is characteristic of *E. muelleri* and *S. ciliatum* with relatively large genomes (about 10 and 12 thousand orthogroups, 326 and 539 Mb, respectively). The smallest number of orthogroups is in *O. minuta* with a compact genome (about 7000 orthogroups, 61 Mb). It is also noticeable that the largest numbers of orthogroups are either unique for each species or in common for all analyzed species. Thus, the largest number of orthogroups in the analyzed dataset (4554) belongs to *S. ciliatum*; they do not contain sequences of any other examined species. *O. minuta* stands apart, with species-specific proteins united into only 622 orthogroups, while the orthogroups common to other analyzed sponge taxa are slightly larger, 722 orthogroups.

We then added representatives of other phyla, *Trichoplax adhaerens* (Placozoa), *Nematostella vectensis* (Cnidaria), and *Branchiostoma floridae* (Chordata) to the analysis to assess Porifera-specific gene expansion (Supplementary Fig. S4). Again, orthogroups unique to the species or common to all species analyzed were the most represented. The number of species-specific orthogroups is more than twofold reduced in *Trichoplax* compared to *O. minuta*, which has the minimum among sponges (251 vs 581). The addition of taxa did not change the number of orthogroups unique to *H. dujardinii*.

4511 sequences are found in the 1053 orthogroups unique to *H. dujardinii*. 22% of these proteins had a BBH in Uniprot, and 44% had at least one domain annotated in Pfam. 54% of the proteins in these lineage-specific orthogroups did not have any annotation. We performed GO enrichment analysis of these proteins and found that many genes mapping to the terms "Cell communication" (GO:0,007,154 in the 'Biological process' category) and "Collagen-containing extracellular matrix" (GO:0,062,023 in the Cell compartment category) (Fig. 6B, Supplementary Table S5).

The group of proteins associated with the term "Cell communication" contained 80 peptides. Interestingly, 60 sequences transcribed from 25 genes were located at contig 20. Of these 80 proteins, 30 had BBHs with Uniprot, and all of these BBHs were with Metazoa sequences. Most of these sequences from contig 20 belong to aggregation factor family (see below).

Extracellular matrix genes analysis

Manual inspection of sequences in both detected groups ("Cell communication" and "Collagen-containing extracellular matrix") showed that they consisted mainly of proteins related to the extracellular matrix (ECM). Most of the proteins in the "Cell communication" group had the Calx-beta domain, a beta-folding domain consisting of tandem repeats and binding calcium ions. It has been described in deuterostome transmembrane Na/Ca exchangers and in a number of extracellular matrix proteins (e.g., extracellular matrix organizing protein FRAS1, ECM3). A large number of Calx-beta domains are characteristic of adhesion proteins in a wide variety of animals. Thus, an extracellular matrix protein 3 (ECM3), was originally described to enable the migration of mesenchymal cells in blastocoel in the sea urchin[49]. It had five Calx-beta domains, NG2-like (chondroitin sulfate proteoglycan), and transmembrane domains. The primary structure of ECM3 was similar to the sequence of MAFp4, an aggregation factor from demosponge *Clathria prolifera*[50]. Later, a search for mutations responsible for the development of Fraser syndrome, a multisystem malformation usually comprising cryptophthalmos, syndactyly, and renal defects, resulted in the characterization of the FRAS1 protein gene in mammals. The FRAS1 protein sequence showed similarity to sea urchin ECM3, but in addition contained furin-like domains and von Willebrand factor type C domains[51]. Due to these similarities, the gene for the ECM3 protein was named FREM—FRAS1-related extracellular matrix protein[52]. The link between these proteins and AFs in ECM organisation is the Calx-beta domain, whose function in cell-ECM interactions has probably been conserved from sponges to chordates. However, at the peptide sequence level, AFs differ from the proteins of other animals by the presence of a distinct Wreath domain (see below).

To better understand the composition of the extracellular matrix, we extracted the lines containing the term "extracellular matrix" from the Trinotate report and manually examined the resulting list for BBH and domain composition. Thus, we found orthologs with a characteristic set of domains of such structural proteins of the extracellular matrix as Porifera-specific fibrillar collagen[53] and fibrillar collagens of different classes, fibrinogen domain-containing proteins, cartilage matrix protein, mucin, as well as some proteins characterized but not having their own names (Sushi, nidogen, and EGF-like domain-containing protein, or SNED1). More than 100 intermolecular interaction partners have been predicted for this SNED1 protein, including integrin receptors and fibronectin, with which it partially immunolocalizes[54]. SNED1 is widely expressed in mammalian embryos, especially intense in cells undergoing epithelio-mesenchymal transition[55]. Cartilage matrix protein consists of two vWFA and EGF-like domains and has the ability to bind to collagen[56,57]. A membrane-associated protein with a high level of glycosylation, Mucin, forms the backbone of mucus, covering the apical surface of epithelial cells in eumetazoans. Its functions are not limited to barrier function: mucins are able to regulate cell behavior and interactions[58,59]. The enzymes procollagen lysine hydroxylase, glycosyltransferase and peroxidasin have also been found to provide post-translational modifications of collagens (Supplementary Table S6). Many proteins with putative immunoglobulin or EGF domains are detected, but without reliable homology with any sequences from Uniprot.

We also specifically searched for genes of basement membrane components. Type IV collagen, laminin, nidogen (entactin), fibronectin, and the proteoglycan heparan sulfate[60] are characteristic for the basement membrane of metazoan epithelia. Cells (including epithelial cells) express different types of integrins, which are transmembrane receptors that bind a wide range of ligands: fibronectins, collagens, laminins, fibrinogens,

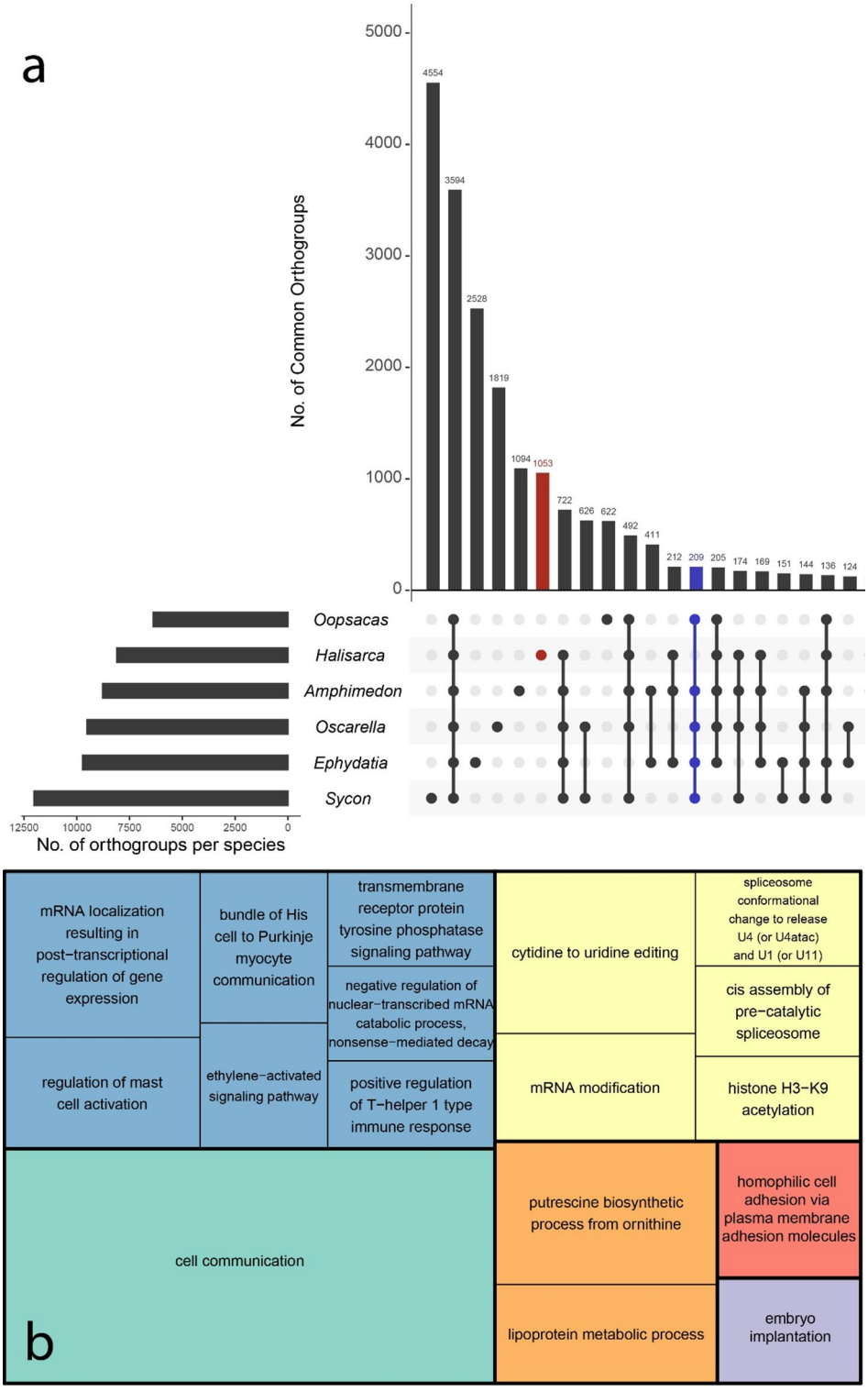

Fig. 6. (**A**) Orthology analysis performed on six species representing all classes of sponges. Each row under the bar chart corresponds to a species. The left side shows how many orthogroups the proteins of a given species represent. On the right is a matrix showing, using dots and lines connecting them, which species are included in the orthogroups. The bar chart shows how many orthogroups have been found involving this set of species. Orthogroups of sequences unique among sponges to *H. dujardinii* are highlighted in red. Orthogroups that include all sponges except *H. dujardinii* are shown in blue. (**B**) GO enrichment of sequences from orthogroups unique among studied sponges to *H. dujardinii*. GO terms in category "Biological process" are indicated inside cells, with cell size proportional to *p*-value.

and others. We found gene orthologs of all basement membrane component proteins and integrin receptors (Supplementary Table S6). All of them contain domains characteristic of Bilateria orthologs (Fig. 7A).

Type IV collagen attracted our attention in particular. The most part of the molecule of type IV collagen is occupied by a helical domain consisting of triplets of Gly-Xaa-Yaa repeats, where X and Y are often proline and hydroxyproline. At the C-terminal end of the peptide is the globular domain NC1 (non-collagenous), so named because it differs from the typical fibrillar collagen helix built of repeats. In Metazoa, collagens are represented by a group of proteins (humans have 22 of its orthologs), some of which form fibrils (types I–III, V, and XI) and others form nonfibrillar structures (types IV, VI, VII, VIII, IX, etc.)[61]. When the mesohyl of sponges was studied by biochemical methods based on selective extraction of fibrillar proteins, the presence of two collagen proteins named spongins was shown[62]. A study of the ultrastructure of demosponges showed that their mesohyl contains cross-striated fibers 20–25 nm in diameter, resembling collagen fibers, and in a number of structures (cement between the spicules, inorganic skeleton of bath sponges, envelope of gemmules) spongin fibers 10 nm thick are present[63]. The study of spongins has shown that they are short-chain collagens containing a small helical domain and an NC1 domain[64]. Their orthologs are common in invertebrate animals from sponges to ascidians. Phylogenetic analysis suggests that spongins are precursors of type IV collagens[64]. Recent work divides sponge collagens into 4 types: spongins, collagen IV-like proteins, fibrillar collagens of Demospongiae, and collagens of Hexactinellida[2].

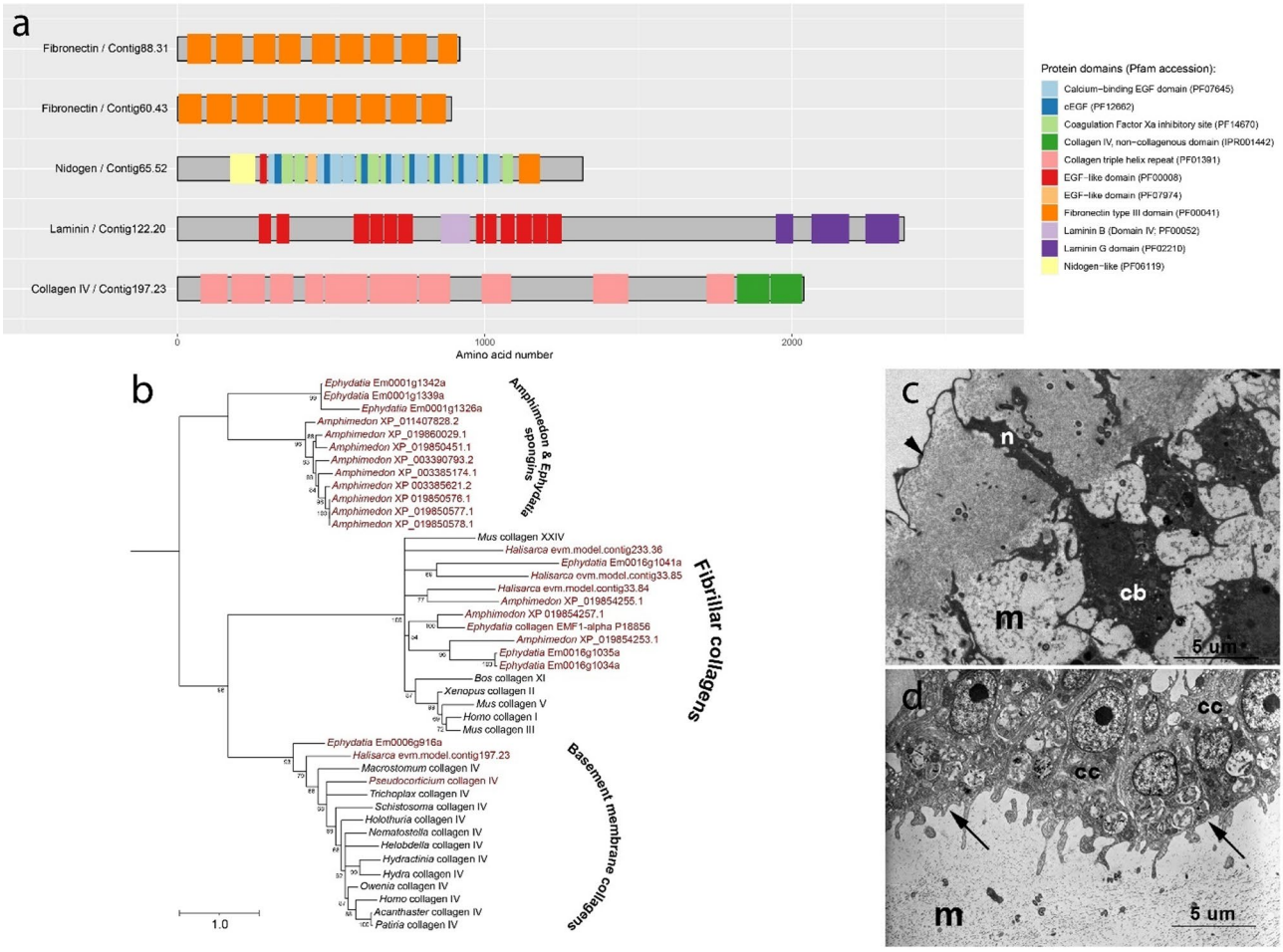

Fig. 7. Basement membrane components in *H. dujardinii* genome. (**A**) Domain architecture components of key proteins evaluated by HMMER against Pfam database. (**B**) Maximum likelihood tree constructed from the alignment of the C-terminal NC1 domains of type IV collagens and different fibrillar collagens from phylogenetically distant animals as well as spongins. Numbers in nodes are bootstrap. Sponge's sequences are in red. (**C**) Transmission electron microscopy of exopinacocytes, the outer layer of *H. dujardinii* cells. They are T-shaped: the cell consists of an outer part (cap, shown by the *arrowhead*), a neck (*n*) and a cell body with a nucleus (*cb*). (**D**) The choanocytes (*cc*) form another epithelium-like tissue in *H. dujardinii*. Their basal surfaces (indicated by *arrows*) face the mesohyl (*m*). Fibers are visible in the mesohyl, but no structures resembling the basement membrane are observed. Scale bar is the same for both TEM images. Sponge tissues were fixed in 2.5% glutaraldehyde on 0.1 M cacodylate buffer, then post-fixed in 1% OsO_4 and embedded in EMbed-812 resin (Electron Microscopy Science) after dehydration. Sections with a thickness of 70 nm were stained with lead citrate and uranyl acetate.

Type IV collagen was previously discovered among sponges by BLAST and phylogenetic analysis in *Corticium candelabrum* (Homoscleromorpha) and *Sycon coactum* (Calcarea)[65]. Experimental confirmation of the localization of this protein in the basal membrane was obtained in homoscleromorphs *Pseudocorticium jarrei*[3] and *Oscarella lobularis*[5,66].

We searched the proteomes of sponges for proteins containing NC1 domains—homologues of collagen IV, fibrillar collagens, and spongins—using HMMER and models available in the Pfam database. As Aouacheria et al. noted, "from the collagen nomenclature, noncollagenous domains have been named purely on the basis of their position from the C-terminus of the collagen chain, that is, the most C-terminal noncollagenous regions have been defined as NC1 domains although their sequences are often unrelated"[64]. Indeed, a number of fibrillar collagens as well as collagen IV have an NC1 domain at the C-terminal end. At the same time, the Pfam database contains two different hidden Markov models for NC1 domains, PF01410 (NC1 domain of fibrillar collagens) and PF01413 (NC1 domain of collagen IV). Using HMMER, we searched for proteins containing these domains in the proteomes of *H. dujardinii*, *A. queenslandica*, and *E. muelleri*. We were unable to identify any NC1 domain in the previously described spongins[64], so we used them for BLAST against the same proteomes.

A search with HMMER showed the presence of one protein with the NC1 domain of collagen IV in *H. dujardinii*. The presence of a type IV collagen gene in *E. muelleri* has been shown previously[14]. No such proteins were detected in *A. queenslandica*. Collagens with NC1 domain PF01410, presumably fibrillar, were identified in all three species: nine in *A. queenslandica* and *E. muelleri*, and sixteen in *H. dujardinii*. Three paralogs of each were taken into phylogenetic analysis. Using BLAST, several sequences of spongin short-chain collagens were obtained from *E. muelleri* and *A. queenslandica*; these had a helical domain and a C-terminal domain that did not correspond to either PF01410 or PF01413. In *H. dujardinii*, among the top-10 BBHs to *E. muelleri* spongins[64], there were either collagens with a PF01410 domain or no NC1 domain at all. Sequences of collagens from different Metazoa were also taken for analysis—a complete list of collagens is given in Supplementary Table S7. To find out to which group certain sponge proteins belong, we performed phylogenetic analysis (Fig. 7B). The *H. dujardinii* and *E. muelleri* proteins with domain PF01413 belong to the clade with type IV collagens. Fibrillar metazoan collagens form a group with spongin proteins containing the PF01410 domain. Spongin short-chain collagens of *E. muelleri* and *A. queenslandica* are clustered separately from these two groups.

Leys and Riesgo [65] considered various cases of detection in sponges of proteins similar in domain organization to one or another component of the basement membrane and the absence of their complete set in the then available genome of *A. queenslandica*[65]. It is likely that collagen IV and laminin were in the genome of the last common ancestor of multicellular animals, and the absence of one or the other genes in sponges and comb jellies is the result of secondary loss[67]. Here we described for the first time a complete set of basement membrane proteins in the genome of *H. dujardinii*. Despite the presence of the corresponding proteins, we were unable to find morphological features of any structures resembling the basement membrane of epithelia in the layers of pinacocytes or choanocytes (Fig. 7C, D). The presence of basement membrane proteins and enzymes ensuring the assembly of the collagen network in *H. dujardinii* suggests the following possible scenarios for sponges: (1) the protein components of the basement membrane arose independently before its formation as components of the ECM and then, as a result of convergent evolution, formed this structure, (2) they form a homologous structure not detectable at the electron microscopic level, or (3) a morphologically formed basement membrane may be present at a particular stage of the life cycle. For example, in *O. lobularis* from the class Homoscleromorpha, the basement membrane has been characterised by electron microscopy and by antibodies to type VI collagen in larvae and in adult animals[66,68,69]. No such evidence is yet available for Demospongiae. This question remains to be resolved in the future using immunocytochemistry and gene expression data.

Analysis of aggregation factors

Another large group of extracellular matrix proteins that are unique to demosponges are called aggregation factors (AF). These large extracellular glycoproteins, responsible for species-specific and allorecognition, got their name for their participation in the recognition that occurs during cell aggregation[50,70]. At the beginning of the twentieth century, the phenomenon of sponge development from a suspension of cells obtained by mechanical or chemical (with the help of chelators of divalent cations) dissociation was discovered[71]. In experiments on mixing sponge cells of two different species, the cells aggregated only with cells of their own species but not with those of another. Cells interact in this process using receptors on the surface and AFs connecting receptors of neighboring cells. Allorecognition is enabled by high levels of polymorphism and RNA editing[40].

AFs were well characterized biochemically in three species of Demospongiae[50,72,73]. Five clustered AF genes were then characterized in *A. queenslandica* using BLAST[40]. A search in the transcriptomes of 24 sponge species revealed that AFs are present only within the class Demospongiae, suggesting a monophyletic origin of these molecules. A conserved region is present at the C-terminal end of the peptide, identified by Grice et al. using multiple sequence alignment and named Wreath domain[40]. We used the authors' hidden Markov model of the Wreath domain and the AF identification algorithm to search for AFs in the *H. dujardinii* genome based on the presence of the Wreath and Calx-beta domains. We detected 48 transcripts encoding peptides with Wreath domain transcribed from 17 genes. Of these, 15 genes are assembled in a cluster located on contig 20, and another two genes are scattered in other contigs (Fig. 8A). Typical (group 1 according to [40]) AFs of *H. dujardinii* are assembled in a cluster. They are characterized by the presence of a signal peptide, the Wreath domain at the C-terminus, and several Calx-beta domains (Fig. 8B). Calx-beta domains bind calcium cations and are responsible for adhesion[74,75]; their number varies from 1 to 66 in typical AFs. The number of Wreath domains varies in typical AFs from 1 to 3, and due to alternative splicing, their number may vary in isoforms of transcripts of the same gene. Some peptides also contain EGF-like, von Willebrand factor type D (vWFD), and other domains. Proteins encoded by two genes not belonging to the cluster are classified as AF-like proteins. One

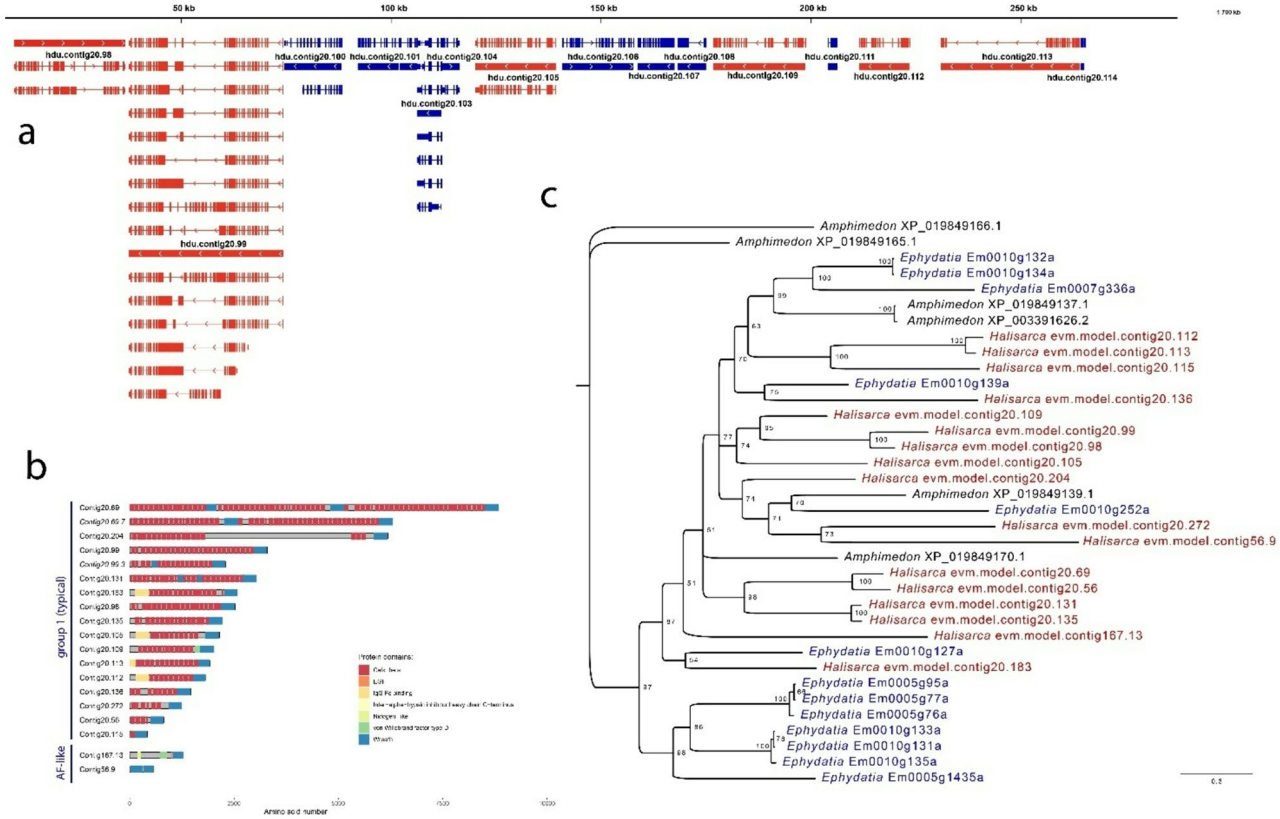

Fig. 8. Aggregation factors (AF) of *H. dujardinii*. (**A**) Genome browser view of a 250-kb fragment of contig 20 containing six AF genes in *H. dujardinii*. AF genes are in red, other genes are blue. Thick blocks represent coding exons, thin blocks non-coding exons and lines introns. Small arrows on introns denote the direction of transcription. Isoforms are shown for transcripts with alternative splicing. Scale bar in kb shown on top. (**B**) Domain composition of the AF and AF-like proteins in *H. dujardinii*. Isoforms with different numbers of Wreath domains are shown for two proteins (names in italics). (**C**) Unrooted Bayesian inference protein tree inferred from the alignment of Wreath domain of sponges. The sequences of *E. muelleri* are shown in blue, *A. queenslandica* in black, and *H. dujardinii* in red. The values at the tree nodes are posterior probabilities for each split in percent.

of them contains Wreath domain but does not contain Calx-beta, and the another one consists almost entirely of Wreath domain.

We retrieved 6 proteins with the Wreath domain from the *A. queenslandica* genome and 13 from the *E. muelleri* genome. The tree was constructed by aligning these proteins with 14 AFs of *H. dujardinii*. Only Wreath domain sequences were used to construct the tree, since the number of Calx-beta domains varies in AFs from one to several tens. It is not possible to use outgroup in the tree building: a search for proteins containing the Wreath domain in Uniprot did not identify any sequences.

At the level of amino acid sequences, the Wreath domain appeared to be low conservative. We calculated the proportion of identical amino acids and those with similar physicochemical properties in pairwise comparisons of sequences from the alignment. Expectedly, the highest proportion of identical and similar amino acids was for sequences from the same species (76% and 87%, respectively, for contig20.6 and contig20.5 from *H. dujardinii*), and the lowest proportion was between species (8% and 12%, respectively, for the pair Em0005g76a from *E. muelleri* and contig20.1 from *H. dujardinii*). However, the median percentage of identical amino acids is only 36.4%, which is quite small for a protein domain within a single animal phylum.

The phylogenetic tree constructed using Bayesian inference contained two nodes in the middle with low posterior probability (51%). We were unable to overcome this by iterative alignment, fitting different replacement models, and using other reconstruction methods. Adding Wreath domain sequences from the demosponge proteins of *Halichondria panicea*, *Dysidea avara* and *Geodia barretti*, whose proteomes are available at NCBI, to the alignment worsens the topology of the tree, turning it into a comb. Although the overall phylogeny of AFs cannot be unambiguously resolved using these three species, several features are notable (Fig. 8C). (1) Basically, it can be noted that clades with acceptable support are formed by sequences of the same species. (2) Some of *H. dujardinii* genes are clearly the result of duplications, as indicated by their position in the tree and close proximity on the chromosome. (3) Some clades include proteins from all three species, but support for them is relatively low (about 70%). (4) A group of seven *E. muelleri* paralogues forms an isolated, well-supported clade. Two *A. queenslandica* proteins are more closely related to this clade than to the other sequences. These data

demonstrate the complexity of the evolutionary history of aggregation factors within the class Demospongiae, and their expansion in *H. dujardinii*.

AFs were previously thought to be a monophyletic group unique to the class Demospongiae. However, a recent analysis using a search based on structural similarity rather than amino acid sequence revealed that the central beta-sandwich of the wreath domain is highly similar to the vWFD domain and to the C-terminal domain of Repulsive Guidance Molecules[76]. In addition, vWFD proteins are able to multimerise to form ring-like structures, just like AF. This indicates a possible homology between the domain present in Bilateria (vWFD) and the Demospongiae-specific Wreath domain at both structural and functional levels: both proteins are highly glycosylated and provide cell–cell interactions with ECM.

Conclusion

The study of non-bilaterian phylogenetic lineages is of particular importance in order to understand the origin and early evolution of structures, processes, and functions in Metazoa. Phylum Porifera, located as sister to at least most other Metazoa, is a convenient taxon to search for answers to such questions. Carefully assembled and annotated genomes of non-model species are still extremely scarce, although next-generation sequencing has become routine. In the present work, we presented the results of analyzing the assembly of the demosponge *Halisarca dujardinii* genome made with Oxford Nanopore long reads. In the genome, we analyzed genes encoding components of basement membrane and aggregation factors, proteins responsible for calcium-dependent adhesion and recognition. Our data on the AF gene set and the presence of basement membrane components indicate substantial differences among demosponges (such as *A. queenslandica, E. muelleri*, and *H. dujardinii*). To date, the number of predicted genes in the genome of *H. dujardinii* is the smallest among the available Porifera genomes, although the nearest species in terms of the number of genes (*Oopsacas minuta* and *Oscarella lobularis*) have genomes almost 3 times smaller. The other two species of Demospongiae species considered have at least 2.5 times more genes. These findings highlight the considerable heterogeneity within Porifera and the complexity of its evolutionary history. In addition, we have obtained a resource for further study of gene expression mechanisms by modern methods (scRNA-seq, ChIP-seq) in dynamic processes such as development and regeneration.

Data availability

The datasets generated and analyzed during the current study are available in the SRA NCBI repository PRJNA1055929, PRJNA1068026, PRJNA1067540, in the WGS NCBI repository JBKORC000000000, in Github repository [https://github.com/mezoderma/Hdu-genome](https:/github.com/mezoderma/Hdu-genome), and as supplementary files. Additional information is available upon request to the corresponding author (I.B.).

Received: 16 July 2025; Accepted: 13 November 2025
Published online: 29 December 2025

References

1. Schultz, D. T. et al. Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).
2. Ehrlich, H., Wysokowski, M., Żółtowska-Aksamitowska, S., Petrenko, I. & Jesionowski, T. Collagens of Poriferan origin. *Mar. Drugs* **16**, 79 (2018).
3. Boute, N. et al. Type IV collagen in sponges, the missing link in basement membrane ubiquity. *Biol. Cell* **88**, 37–44 (1996).
4. Gazave, E. *et al.* No longer Demospongiae: Homoscleromorpha formal nomination as a fourth class of Porifera. In *Ancient Animals, New Challenges* (eds Maldonado, M., Turon, X., Becerro, M. & Jesús Uriz, M.) 3–10 (Springer, Dordrecht, 2011). https://doi.org/10.1007/978-94-007-4688-6_2
5. Vernale, A. et al. Evolution of mechanisms controlling epithelial morphogenesis across animals: New insights from dissociation-reaggregation experiments in the sponge Oscarella lobularis. *BMC Ecol. Evol.* **21**, 160 (2021).
6. Busch, K. et al. Biodiversity, environmental drivers, and sustainability of the global deep-sea sponge microbiome. *Nat. Commun.* **13**, 5160 (2022).
7. Carrier, T. J. et al. Symbiont transmission in marine sponges: Reproduction, development, and metamorphosis. *BMC Biol.* **20**, 100 (2022).
8. Sipkema, D. et al. Large-scale production of pharmaceuticals by marine sponges: Sea, cell, or synthesis?. *Biotechnol. Bioeng.* **90**, 201–222 (2005).
9. Ereskovsky, A., Borisenko, I. E., Bolshakov, F. V. & Lavrov, A. I. Whole-body regeneration in sponges: Diversity, fine mechanisms, and future prospects. *Genes* **12**, 506 (2021).
10. Srivastava, M. et al. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* **466**, 720–726 (2010).
11. Belahbib, H. *et al.* New genomic data and analyses challenge the traditional vision of animal epithelium evolution. *BMC Genomics* **19**, 393 (2018).
12. Fortunato, S. et al. Genome-wide analysis of the sox family in the calcareous sponge Sycon ciliatum: Multiple genes with unique expression patterns. *EvoDevo* **3**, 14 (2012).
13. Francis, W. R. et al. The genome of the contractile demosponge tethya wilhelma and the evolution of metazoan neural signalling pathways. *BioRxiv* https://doi.org/10.1101/120998 (2017).
14. Kenny, N. J. et al. Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge Ephydatia muelleri. *Nat Commun* **11**, 3676 (2020).
15. Santini, S. et al. The compact genome of the sponge Oopsacas minuta (Hexactinellida) is lacking key metazoan core genes. *BMC Biol.* **21**, 139 (2023).
16. Fortunato, S. A. V. et al. Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. *Nature* **514**, 620–623 (2014).
17. Fortunato, S. A. V., Vervoort, M., Adamski, M. & Adamska, M. Conservation and divergence of bHLH genes in the calcisponge Sycon ciliatum. *EvoDevo* **7**, 23 (2016).
18. Leininger, S. *et al.* Developmental gene expression provides clues to relationships between sponge and eumetazoan body plans. *Nat. Commun.* **5**, 3905 (2014).
19. Adamska, M. et al. Structure and expression of conserved Wnt pathway components in the demosponge Amphimedon queenslandica. *Evol. Dev.* **12**, 494–518 (2010).

20. Conaco, C. et al. Transcriptome profiling of the demosponge Amphimedon queenslandica reveals genome-wide events that accompany major life cycle transitions. *BMC Genom.* **13**, 209 (2012).
21. Kerner, P., Degnan, S. M., Marchand, L., Degnan, B. M. & Vervoort, M. Evolution of RNA-binding proteins in animals: insights from genome-wide analysis in the sponge Amphimedon queenslandica. *Mol. Biol. Evol.* **28**, 2289–2303 (2011).
22. Yuan, H., Hatleberg, W. L., Degnan, B. M. & Degnan, S. M. Gene activation of metazoan Fox transcription factors at the onset of metamorphosis in the marine demosponge *Amphimedon queenslandica*. *Dev Growth Differ* **64**, 455–468 (2022).
23. Cornejo-Páramo, P., Roper, K., Degnan, S. M., Degnan, B. M. & Wong, E. S. Distal regulation, silencers, and a shared combinatorial syntax are hallmarks of animal embryogenesis. *Genome Res* **32**, 474–487 (2022).
24. Gaiti, F., Calcino, A. D., Tanurdžić, M. & Degnan, B. M. Origin and evolution of the metazoan non-coding regulatory genome. *Dev. Biol.* **427**, 193–202 (2017).
25. Fernandez-Valverde, S. L. & Degnan, B. M. Bilaterian-like promoters in the highly compact Amphimedon queenslandica genome. *Sci Rep* **6**, 22496 (2016).
26. Wörheide, G. *et al.* Chapter one—deep phylogeny and evolution of sponges (Phylum Porifera). In *Advances in Marine Biology* (eds Becerro, M. A., Uriz, M. J., Maldonado, M. & Turon, X.) vol. 61 1–78 (Academic Press, 2012).
27. Sambrook, J., Russell, D. W., Irwin, C. A. & Janssen, K. A. *Molecular cloning: A laboratory manual* (Cold Spring Harbor Laboratory Press, 2001).
28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
29. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
30. Steffen, K. et al. Whole genome sequence of the deep-sea sponge Geodia barretti (Metazoa, Porifera, Demospongiae). *G3 Genes/Genomes/Genetics* **13**, 192 (2023).
31. Borisenko, I., Adamski, M., Ereskovsky, A. & Adamska, M. Surprisingly rich repertoire of Wnt genes in the demosponge Halisarca dujardini. *BMC Evol. Biol.* **16**, 123 (2016).
32. Wu, T. D. & Watanabe, C. K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
33. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
34. Baril, T., Galbraith, J. & Hayward, A. Earl Grey: A fully automated user-friendly transposable element annotation and analysis pipeline. https://doi.org/10.1101/2022.06.30.498289 (2023).
35. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
36. Stanke, M. et al. AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
37. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **42**, e119 (2014).
38. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
39. Exposito, J.-Y., Ouazana, R. & Garrone, R. Cloning and sequencing of a Porifera partial cDNA coding for a short-chain collagen. *Eur. J. Biochem.* **190**, 401–406 (1990).
40. Grice, L. F. et al. Origin and evolution of the sponge aggregation factor gene family. *Mol. Biol. Evol.* **34**, 1083–1099 (2017).
41. Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A. & Minh, B. Q. W-IQ-TREE: A fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **44**, W232-235 (2016).
42. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
43. Blin, N. & Stafford, D. W. A general method for isolation of high molecular weight DNA from eukaryotes. *Nucleic Acids Res.* **3**, 2303–2308 (1976).
44. Mayer, A. M. S. et al. Marine pharmacology in 2018: Marine compounds with antibacterial, antidiabetic, antifungal, anti-inflammatory, antiprotozoal, antituberculosis and antiviral activities; affecting the immune and nervous systems, and other miscellaneous mechanisms of action. *Pharmacol. Res.* **183**, 106391 (2022).
45. Adameyko, K. I. et al. Conservative and atypical ferritins of sponges. *Int. J. Mol. Sci.* **22**, 8635 (2021).
46. Erpenbeck, D. et al. First evidence of miniature transposable elements in sponges (Porifera). *Hydrobiologia* **687**, 43–47 (2012).
47. Fernandez-Valverde, S. L., Calcino, A. D. & Degnan, B. M. Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge Amphimedon queenslandica. *BMC Genomics* **16**, 387 (2015).
48. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
49. Hodor, P. G., Illies, M. R., Broadley, S. & Ettensohn, C. A. Cell-substrate interactions during sea urchin gastrulation: Migrating primary mesenchyme cells interact with and align extracellular matrix fibers that contain ECM3, a molecule with NG2-like and multiple calcium-binding domains. *Dev. Biol.* **222**, 181–194 (2000).
50. Fernàndez-Busquets, X. & Burger, M. M. Cell adhesion and histocompatibility in sponges. *Microsc. Res. Tech.* **44**, 204–218 (1999).
51. McGregor, L. et al. Fraser syndrome and mouse blebbed phenotype caused by mutations in FRAS1/Fras1 encoding a putative extracellular matrix protein. *Nat. Genet.* **34**, 203–208 (2003).
52. Shafeghati, Y., Kniepert, A., Vakili, G. & Zenker, M. Fraser syndrome due to homozygosity for a splice site mutation of FREM2. *Am. J. Med. Genet. A* **146A**, 529–531 (2008).
53. Exposito, J. Y. & Garrone, R. Characterization of a fibrillar collagen gene in sponges reveals the early evolutionary appearance of two collagen gene families. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 6669–6673 (1990).
54. Vallet, S. D. et al. Computational and experimental characterization of the novel ECM glycoprotein SNED1 and prediction of its interactome. *Biochem. J.* **478**, 1413–1434 (2021).
55. Barqué, A. et al. Knockout of the gene encoding the extracellular matrix protein SNED1 results in early neonatal lethality and craniofacial malformations. *Dev. Dyn.* **250**, 274–294 (2021).
56. Argraves, W. S., Deák, F., Sparks, K. J., Kiss, I. & Goetinck, P. F. Structural features of cartilage matrix protein deduced from cDNA. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 464–468 (1987).
57. Kiss, I. et al. Structure of the gene for cartilage matrix protein, a modular protein of the extracellular matrix. Exon/intron organization, unusual splice sites, and relation to alpha chains of beta 2 integrins, von Willebrand factor, complement factors B and C2, and epidermal growth factor. *J. Biol. Chem.* **264**, 8126–8134 (1989).
58. McShane, A. et al. Mucus. *Curr. Biol.* **31**, R938–R945 (2021).
59. Pajic, P. et al. A mechanism of gene evolution generating mucin function. *Sci. Adv.* **8**, eabm8757 (2022).
60. Pozzi, A., Yurchenco, P. D. & Iozzo, R. V. The nature and biology of basement membranes. *Matrix Biol* **57–58**, 1–11 (2017).
61. Exposito, J., Cluzel, C., Garrone, R. & Lethias, C. Evolution of collagens. *Anat. Rec.* **268**, 302–316 (2002).
62. Gross, J., Sokal, Z. & Rougvie, M. Structural and chemical studies on the connective tissue of marine sponges. *J. Histochem. Cytochem.* **4**, 227–246 (1956).
63. Garrone, R. *Phylogenesis of connective. Tissue morphological aspects and biosynthesis of sponge intercellular matrix* (Wiley, 1978).
64. Aouacheria, A. et al. Insights into early extracellular matrix evolution: Spongin short chain collagen-related proteins are homologous to basement membrane type IV collagens and form a novel family widely distributed in invertebrates. *Mol. Biol. Evol.* **23**, 2288–2302 (2006).
65. Leys, S. P. & Riesgo, A. Epithelia, an evolutionary novelty of metazoans. *J. Exp. Zool. (Mol. Dev. Evol.)* **318**, 438–447 (2012).

66. Rocher, C. et al. The buds of Oscarella lobularis (Porifera, Homoscleromorpha): A new convenient model for sponge cell and evolutionary developmental biology. *J. Exp. Zool. B Mol. Dev. Evol.* **342**, 503–528 (2024).
67. Renard, E., Le Bivic, A. & Borchiellini, C. Origin and evolution of epithelial cell types. In *Origin and Evolution of Metazoan Cell Types* (CRC Press, 2021).
68. Boury-Esnault, N., Ereskovsky, A., Bézac, C. & Tokina, D. Larval development in the Homoscleromorpha (Porifera, Demospongiae). *Invertebr. Biol.* **122**, 187–202 (2005).
69. Gazave, E. et al. Systematics and molecular phylogeny of the family Oscarellidae (Homoscleromorpha) with description of two new oscarella species. *PLoS ONE* **8**, e63976 (2013).
70. Moscona, A. A. Cell aggregation: Properties of specific cell-ligands and their role in the formation of multicellular systems. *Dev. Biol.* **18**, 250–277 (1968).
71. Wilson, H. V. On some phenomena of coalescence and regeneration in sponges. *J. Exp. Zool.* **5**, 245–258 (1907).
72. Müller, W. E., Koziol, C., Müller, I. M. & Wiens, M. Towards an understanding of the molecular basis of immune responses in sponges: the marine demosponge Geodia cydonium as a model. *Microsc. Res. Tech.* **44**, 219–236 (1999).
73. Wiens, M. et al. Innate immune defense of the sponge Suberites domuncula against bacteria involves a MyD88-dependent signaling pathway. Induction of a perforin-like molecule. *J. Biol. Chem.* **280**, 27949–27959 (2005).
74. Fernàndez-Busquets, X., Körnig, A., Bucior, I., Burger, M. M. & Anselmetti, D. Self-recognition and Ca^{2+}-dependent carbohydrate-carbohydrate cell adhesion provide clues to the cambrian explosion. *Mol. Biol. Evol.* **26**, 2551–2561 (2009).
75. Hilge, M., Aelen, J. & Vuister, G. W. Ca^{2+} regulation in the Na+/Ca^{2+} exchanger involves two markedly different Ca^{2+} sensors. *Mol. Cell* **22**, 15–25 (2006).
76. Ruperti, F. et al. Proteomic analysis of the sponge Aggregation Factor implicates an ancient toolkit for allorecognition and adhesion in animals. *Proc. Natl. Acad. Sci.* **121**, e2409125121 (2024).

Acknowledgements

We thank the resource centers 'Computer Centre SPbU' and 'Molecular and Cell Technologies' of St Petersburg State University for technical support, and the Educational and Research Station of SPbU "Belomorskaia". The research was supported by the Russian Science Foundation Grant No. 22-74-00042, https://rscf.ru/project/22-74-00042/ (I.B.). The study was conducted in frame of Government basic research program for the Koltzov Institute of Developmental Biology no. 0088-2024-0012 (A.E.).

Author contributions

I.B. and A.L. was responsible for animal connection and nucleic acid extraction. A.P. contributed to the genome assembly. A.P. and I.B. are responsible for the initial manuscript preparation. I.B. contributed to study design, sequencing, bioinformatic analysis. A.E. contributed to study design. All authors contributed to final editing of the manuscript.

Funding

The research was supported by the Russian Science Foundation, Grant No. 22–74-00042, https://rscf.ru/project/22-74-00042/ (I.B.). The study was conducted in frame of Government basic research program for the Koltzov Institute of Developmental Biology No. 0088-2024-0012 (A.E.).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-28962-y.

Correspondence and requests for materials should be addressed to I.B.