Reconstruction of protein domain evolution using single-cell amplified genomes of uncultured choanoflagellates sheds light on the origin of animals

David López-Escardó1,2, Xavier Grau-Bové1,3,4, Amy Guillaumet-Adkins5,6, Marta Gut5,6, Michael E. Sieracki7 and Iñaki Ruiz-Trillo1,3,8

1Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Maritim de la Barceloneta 37-49, 08003 Barcelona, Catalonia, Spain
2Institut de Ciències del Mar (ICM-CSIC), Passeig Maritim de la Barceloneta 37-49, 08003 Barcelona, Catalonia, Spain
3Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, 08028 Barcelona, Catalonia, Spain
4Department of Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, UK
5CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain
6Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
7National Science Foundation, Arlington, VA 22314, USA
8ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

Understanding the origins of animal multicellularity is a fundamental biological question. Recent genome data have unravelled the role that co-option of pre-existing genes played in the origin of animals. However, there were also some important genetic novelties at the onset of Metazoa. To have a clear understanding of the specific genetic innovations and how they appeared, we need the broadest taxon sampling possible, especially among early-branching animals and their unicellular relatives. Here, we take advantage of single-cell genomics to expand our understanding of the genomic diversity of choanoflagellates, the sister-group to animals. With these genomes, we have performed an updated and taxon-rich reconstruction of protein evolution from the Last Eukaryotic Common Ancestor (LECA) to animals. Our novel data redefines the origin of some genes previously thought to be metazoan-specific, like the POU transcription factor, which we show appeared earlier in evolution. Moreover, our data indicate that the acquisition of new genes at the stem of Metazoa was mainly driven by duplications and protein domain rearrangement processes at the stem of Metazoa. Furthermore, our analysis allowed us to reveal protein domains that are essential to the maintenance of animal multicellularity. Our analyses also demonstrate the utility of single-cell genomics from uncultured taxa to address evolutionary questions.

1. Introduction

Metazoa is the eukaryotic kingdom with most described species so far, around 1.3 million [1], and it is the multicellular group of eukaryotes for which the most differential cell types have been described [2]. Animals’ success might be tightly linked to their multicellular complexity, which has been considered unique in the eukaryotic world [3], compared to other multicellular eukaryotic transitions [3,4]. Thus, the uniqueness of animal multicellularity raises the question of which mechanisms shaped the emergence of such special types of organisms from a unicellular ancestor more than 600 million years ago [4,5].

To address this question, efforts have been made, over the past decade, to reconstruct the Urmetazoan genomic content by comparing the genomic...
sequences of a broad and diverse spectrum of animals [6–8] with the genomes of their closest unicellular relatives: the Choanoflagellata [9,10], Filasterea [11,12] and Teretospora, which include ichthyosporeans and Conilochaecium linacisporum [13,14]. These unicellular lineages and metazoans conform the Holozoa clade. Holozoa, together with Fungi and their unicellular relatives compose the eukaryotic supergroup known as Opisthokonta [15]. Thus, understanding the evolution of opisthokonts is critical to address the origin of animals.

The first genomic comparisons between animals and their unicellular relatives at a genomic level revealed a complex pre-metazoan genetic toolkit, already equipped with a rich repertoire of genes involved in multicellular functions. These included developmental transcription factors (like Brachyury, MYC, Runx, or P53), cell adhesion proteins (ECM elements, integrins, cadherins, and C-type lectins) and cell signalling receptors and transducers [9,10,16–20]. These findings suggested that co-option of ancestral genes into new functions was an important mechanism that occurred in the transition from the unicellular ancestor of animals to the Urmetazoan [20]. However, it was also found that not all the components of many animal signalling pathways, like the Hippo pathway, had a pre-metazoan origin. In some cases, only some ligands or receptors were present before the emergence of multicellularity. Those ancestral ligands and receptors were later on putatively co-opted in functioning within these animal signalling pathways [20,21]. Thus, the acquisition of new genes might have also played an important role in the emergence of animal multicellularity.

In two recent studies aiming at better reconstructing the Urmetazoan genome, bursts of new genic innovation were shown at the stem of Metazoa. One of those studies focused on the search of animal genetic innovations using a new method to infer homology [22], while the other expanded the genomic information of choanoflagellates by sequencing the transcriptomes of 19 choanoflagellate species [23]. Both studies claimed that approximately 1500–1700 genes were acquired during the transition towards animal multicellularity (three times more gene acquisition than has been reported in their unicellular ancestors). In particular, the most conserved animal-specific genes in extant metazoans were genes related to major signalling pathways such as components of TGF-beta or Wnt signalling pathways, and transcription factors like ETS or POU functions and redefines the origin of some genes thought to be metazoan-specific, which appeared earlier in the unicellular ancestor of animals like the POU transcription factor.

2. Methods

(a) Cell collection and whole genome amplification

Cells for single-cell genomics were collected from the Mediterranean sea and different places in the Indian Ocean during the Tara Oceans expedition [35] and cryopreserved as described before [36]. Flow cytometry cell sorting, single cell lysis and whole genome amplification by Multiple Displacement Amplification (MDA) [37] were performed at the Bigelow Single-cell genomics facility (Boothbay, Maine, USA), as previously described [38–40]. The SAGs obtained were screened by PCR using universal eukaryotic 18S, as in previous studies [34,40]. Four SAGs were placed in distant phylogenetic positions compared to choanoflagellates for which there are available transcriptomic or genomic data (figure 1). Thus, they were deemed worthy of further analysis. Associated environmental data are summarized in electronic supplementary material, table S1, and further details can be found in PANGAEA [31,43].

(b) Library preparation and genome sequencing

Four SAGs (UC1, UC2, UC3 and UC4) were sent for sequencing at CNAG (Barcelona, Spain). The libraries were constructed with the TruSeq Nano DNA Library Preparation Kit according to the manufacturer’s protocol. Briefly, aiming for an insert size of 550 bp, 200 ng of gDNA were sheared by sonication using Covaris E210 (Covaris). Fragmented DNA was purified with Agencourt AMPure XP beads. Afterwards, end repair and size selection were performed, following 3’ adenylation reaction and ligation.
Figure 1. Phylogenetic position of the new choanoflagellate SAGs. Phylogenetic tree based on 117 sequences of the 18S rDNA gene, representing all that is known of the molecular diversity of choanoflagellates and unicellular holozoans, including environmental lineages. The phylogenetic analysis was inferred by maximum likelihood under the GTR+Γ with IQ-TREE. Clades marked by a bullet (•) present high statistical split support, with values greater than 80% of SH-aLRT (bootstraps of single branch test) and greater than 95% of ultrafast bootstrap. Both indexes were computed with IQ-TREE. The remaining split supports obtained can be found at Figshare (https://doi.org/10.6084/m9.figshare.7819571.v1) in the tree file. The order and class names given are based on [30,41,42]. Choanoflagellates with transcriptomic data available are depicted with a red asterisk, and those with genomic data available are depicted with a blue hash. Choanoflagellates’ craspedidan clades were named according to our phylogenomic analysis (figure 3). Clade 3 nomenclature and nomenclature within Acanthoecida are the same as in [27]. The Acanthoeicida picture was taken from [28] and Craspedida pictures were taken in Nicole’s King laboratory.
of the Illumina adapter indexes. DNA fragments were enriched by eight cycles of PCR and then purified with Agencourt AMPure XP beads. The Agilent Technologies 2100 Bioanalyzer DNA 1000 assay was used for library quality control and quantification.

Each library was sequenced using one lane of MiSeq reagent kit v2 (Illumina). The sequencing run was performed according to standard Illumina operation procedures in paired-end mode, with a read length of 2 × 251 bp and a yield of greater than 11 Gb. Primary data analysis, image analysis, base calling and quality scoring of the run were processed using the manufacturer’s software Real Time Analysis (RTA 1.18.54) and followed by generation of FASTQ sequence files by CASAVA v. 1.8. The reads obtained were used to perform a downsampling analysis as described in [34]; UC1 and UC4 presented longer assemblies and the curve was still not saturated. Thus, we decided to apply more sequencing depth for them. Finally, UC1 and UC4 were sequenced in 3 MiSeq lanes with a yield of greater than 34 Gb. Raw reads can be found at ENA (European Nucleotide Archive: ebi.ac.uk/ena), project accession number PRJEB31614.

(c) Genome assembly and annotation

Raw reads obtained were trimmed with Trimmmomatic v. 3.0 [44] using the following options: ILLUMINACLIP:/adapters/NexteraPE-PE.fa:2:40:15 HEADCROP:10 CROP:240 SLIDING WINDOW:6:20 MINLEN:50. A range of between 42 and 45 million reads was obtained from the low-quality SAGs UC2 and UC3, and for the SAGs with more sequencing applied—UC1 and UC4—the range moves between 110 and 125 million reads (electronic supplementary material, table S2). Next, we performed the genome assembly with SPAdes v3.6.1 [45] with the options -sc, -careful and -k 21,33,55,77,99. The final genome statistics were obtained with QUAST [46]. The percentage of core eukaryotic conserved proteins was calculated with CEGMA [47] and BUSCO [48]. We screened for mitochondrial genomic sequences in our SAGs by performing a bLASTn v. 2.2.31+ [49] using as query the mitochondrial proteins of Andalucia godoi [50] with a cut-off e-value of 1e-04. Only the SAG UC2 presented three scaffolds with mitochondrial proteins. We mapped SAG UC2 reads over the scaffolds selected using the program Bowtie2 v. 21.0 [51] in order to perform a re-assembly with SPAdes v. 3.6.1 [45], and to see if we could recover in one scaffold the full mitochondrial genomic sequence. The assembly yielded a more fragmented output, and two of the previously selected scaffolds contained proteins that came from bacterial contamination. Thus, we decided to keep the first scaffolds obtained (32 Kb length). UC2 partial mitochondrial genome is available at Figshare (https://doi.org/10.6084/m9.figshare.7819571.v1). We annotated the mitochondrial genes with Mfannot [52] and these are available in electronic supplementary material, table S3.

As the genome completeness of the SAGs UC2 and UC3 was very low, we decided to continue the genome annotation only for the SAGs UC1 and UC4. We annotated the genome with AUGUSTUS [53] trained with CEGMA proteins [47], as explained in [34]. To predict the number of genes that may contain the full genome sequence of UC1 and UC4, we first performed a bLASTp v. 2.2.31+ [49] using as a query our predicted proteins against a database that includes all non-redundant proteins from UniProt [54] with a cut-off e-value of 1e-04, to identify the potential contaminant proteins. We removed the proteins that had the first hit from a bacterial or archaeal origin. However, as there were many genes without blast match and the annotation process can overpredict gene content [34], we decided to take into account only proteins in which, using a Pfam scan, we could find protein domains. The proteomes of M. brevicollis and S. rosetta contain approximately 70% of their proteins with a described Pfam protein domain. Therefore, we considered this fact in our calculations together with the average between the genomic completeness obtained by BUSCO and CEGMA (taking into account complete and fragmented proteins detected), to infer the total number of genes that UC1 and UC4 may have in their complete genomes. SAGs assembly and annotation are available at Figshare (https://doi.org/10.6084/m9.figshare.7819571.v1), including the list of protein classification.

(d) Genome size estimations UC1 and UC4

To estimate the genome sizes of UC1 and UC4, we took the average of the percentage of presence of BUSCO and CEGMA proteins (complete and fragmented), and the length of each assembly. Then, we inferred the putative genomic size of UC1 and UC4 if the percentage was 100%.

(e) Ecological distribution of our SAGs

We performed a bLASTn v. 2.2.31 [49] using as query the 18S sequences of our SAGs against the operational taxonomic units (OTUs) from the TARA Oceans database [55]. We found four OTUs that correspond to our SAGs with 100% or 99.2% identity (only one mismatch) (electronic supplementary material, table S4). We plotted the read distribution according to geographical locations using R [56].

(f) 18S ribosomal gene phylogeny

We collected 18S rDNA ribosomal sequences from representatives of all known 18S rDNA molecular diversity available at public repositories of unicellular holozoans, including uncultured lineages Clade L [29], FRESHChOs, MACHOs and MAoPs [30] (electronic supplementary material, table S5). We ended up with a dataset of 117 18S rDNA sequences. Next, we aligned them using MAFFT [57] with the E-INS-i algorithm. After manual trimming sequence ends, indels and spuriously aligned sites we ended up with a total of 1754 sites, as we did in other 18S rDNA phylogenetic reconstructions [58,59]. As a control, we also aligned the sequences with MAFFT Q-INSI algorithm, which takes into account the structural RNA information [60] and the alignment was trimmed with trimAl (automated1 argument) [61]. We inferred both phylogenetic trees from these alignments using a maximum-likelihood (ML) inference. The best substitution model for phylogenetic inference was selected using IQ-TREE [62], using the TESTNEW model selection procedure and following the Bayesian information criterion (BIC). In all four cases, the GTR substitution matrix with a 5-categories free-rate distribution [63] (a modification of the standard Γ distribution) was selected as the best-fitting model. ML inferences were performed with IQ-TREE, and statistical supports were drawn from 1000 ultrafast bootstrap values with a 0.99 minimum correlation as convergence criterion [64], and 1000 replicates of the SH-like approximate likelihood ratio test [65]. Both phylogenetic trees in nexus file and the alignments before and after the trimming are available at Figshare (https://doi.org/10.6084/m9.figshare.7819571.v1). Both trees show the same position of our SAGs, although the trimmed tree is more consistent with previous phylogenetic reconstructions of the 18S gene in choanoflagellates in which Casperdida appears monophyletic [27,30]. Thus, it is the one used to prepare the figure 1.

(g) Eight-gene phylogeny

Similar to the recently published choanoflagellate phylogeny [27], we built a phylogenetic matrix with the nucleotide sequences of eight house-keeping genes, to infer the choanoflagellate phylogeny with a broader diversity than our phylogenomics approach. The genes used are: the ribosomal SSU (18S) and LSU (28S) genes, actin, beta tubulin, hsp90, hsp70, EF, and EFlA. Electronic supplementary material, table S6 summarizes the presence of each gene in each taxon. All the sequences used in the analysis
are available at Figshare (https://doi.org/10.6084/m9.figshare.7819571.v1). The analysis was performed with 66 taxa, 57 of them being choanoflagellates. To build the final matrix, we aligned each gene separately with MAFFT [37] using E-INS-i algorithm, and we next trimmed the spurious positions manually. Finally, we concatenated the trimmed alignments for each gene, building a phylogenetic matrix composed of 12884 nucleotide positions. To run the phylogenetic analysis we partitioned our dataset into three parts, and in each of them we ran an evolutionary model with different rate distributions separating the ribosomal genes (partition 1), the 1st and 2nd codon positions of the non-ribosomal genes (partition 2), and the third codon position of the non-ribosomal genes (partition 3). The best substitution model for each partition was selected, again, using IQ-TREE v.1.5.1, under the best-fitting model in the partition 1, with 3-categories in the partition 2 and with 4-categories in the partition 3. Statistical supports were drawn from 1000 ultrafast bootstrap values with a 0.99 minimum correlation as a convergence criterion [64]. Bayesian inference (BI) was performed with MrBayes 3.2.6 [66] using the GTR+Γ model of nucleotide substitution in all partitions, running at different distributions according to the model given by IQ-TREE (Γ5, Γ3, Γ4 respectively for each partition). Four chains ran for 4000000 generations and converged (standard deviation of split frequencies = 0.02) and were analysed after a burn-in of 25%. The trimmed concatenated alignment, the partition information and the phylogenetic trees from ML and BI are available at Figshare (https://doi.org/10.6084/m9.figshare.7819571.v1).

\[(i)\] Comparative genomics by protein domain gains and losses

116 different eukaryotic taxa with proteomic information available were selected to perform an analysis of protein gains and losses over the eukaryotic tree of life focusing on choanoflagellates (56 taxa) (electronic supplementary material, table S7, and supplementary information at Figshare https://doi.org/10.6084/m9.figshare.7819571.v1). Protein domain annotations of each proteome were computed using Pfamscan and the 29th release of the Pfam database [70]. We used a custom script to build a matrix containing the eukaryotic taxa and the number of copies of each protein domain. To reduce noise and eliminate possible contaminants, we removed all the protein domains that were found in 95% (or more) of cases within prokaryotic species (Bacteria and Archaea) according to Pfam database. We ended up with a matrix of 116 taxa and 8920 protein domains. Next, we produced a tree nexus file according to the topology of eukaryotes [71]. For unicellular choanoflagellates we incorporated the topology of our phylogenomic analysis. With the protein domain matrix and the consensus taxa tree we used Count [72] to infer the gains and losses for each node of the tree using Dollo parsimony. Using Count, the domains gained at the different ancestral nodes of choanoflagellates could be retrieved. The functional annotation of the 120 protein domains gains at Choanozoa was done manually by checking the literature available for each protein domain. The list of different protein domains across Opisthokonta ancestors (Opisthokonta, Holozoas, Filozoas, Choanozoa, and Metazoas) is available at Figshare (https://doi.org/10.6084/m9.figshare.7819571.v1) together with the protein domain matrix used.

\[(j)\] The probability of retention in extant species

With the list of proteins domains gained at the ancestral nodes, we calculated the probability of conservation of a given protein domain in a phylogenetic group by dividing the number of species of this group (i.e. animals) that have maintained this protein domain by the total number of species of this group (i.e. all the animals present in our analysis). The list of probabilities of retention is joined to the list of proteins domains gained at the different ancestral nodes within Figshare (https://doi.org/10.6084/m9.figshare.7819571.v1). The distribution of these probabilities has been plotted in R [56]. SAGs UC1 and UC4 were not included in this probabilistic analysis given that their fragmented genomes will underestimate the probability of retention in choanoflagellates.

\[(k)\] Pou, Plexin and Nucleophosmin phylogenies

We studied the phylogenetic history of Pou transcription factor, Plexin proteins and the C-terminal domain of Nucleophosmin in detail by following similar approaches. We selected the protein or the protein domains present in choanoflagellates species and also in a set of metazoan species in which most of animal phyla are represented (electronic supplementary material, table S7). We aligned the sequences using MAFFT [60] and trimmed the alignment with trimAl [61]. The phylogenetic inferences were done with IQ-TREE v. 1.5.1, under the best-fitting LG
model. The alignments and the trees can be found in the electronic supplementary material (https://doi.org/10.6084/m9.figshare.7819571.v1).

3. Results

(a) Expanding the genomic diversity of choanoflagellates

To broaden our understanding of the genomic diversity among choanoflagellates, we sequenced four single-cell amplified genomes corresponding to uncultured choanoflagellate cells collected during the TARA oceans expedition [55] (see electronic supplementary material, table S1 for collection environmental details). The four cells belonged to different choanoflagellate taxa, and they did not appear to be related to any previously described species with transcriptomic or genomic information available (figure 1).

To place the different SAGs within the choanoflagellate tree, we first performed a phylogeny of the 18S ribosomal subunit that includes the SAGs and the known 18S molecular diversity of unicellular holozoans, including environmental sequences [30]. UC1 appears as an early-branching clade 1 craspedidan that groups with *Lagenoea antarctica* [73] (figure 1). Its 18S sequence is identical to the environmental NCBI sequence AY426842 (100% of pairwise identity). UC2 forms a monophyletic clade with the rest of Acanthoeidae (nudiform loricates) (figure 1) appearing as sister-group to the rest of acanthoecids. UC3 clusters with the tectiform loricates *Stephanoea paucicostata* and *Stephanoea caudulata*. Finally, UC4 falls as the earliest-branching sister to Acanthoecidae, together with the environmental sequence JQ223245. Thus, the four cells belonged to different choanoflagellate taxa and were not related to any previously described species (figure 1). Moreover, they appeared distantly related to the two choanoflagellate species with a whole genome sequence (*Monosiga brevicollis* and *Salpingoea rosetta*), thus expanding the genomic information currently available for choanoflagellates.

We then analysed the geographical distribution of these uncultured choanoflagellates using the metabarcoding data from the TARA Oceans database [55]. We found that the craspedidan UC1 is mainly present in Mediterranean samples, although not exclusively (electronic supplementary material, figure S2A). Interestingly, the environmental sequence AY426842 (which is identical to UC1) was also sampled in the Mediterranean [74]. Acanthoecidae sister UC4 is the third most abundant choanoflagellate in TARA Oceans, and it has a cosmopolitan distribution (present in 46 sampling stations out of 47) (electronic supplementary material, figure S2A). The nudiform UC2 and the tectiform UC3 are also widely distributed (45 samples out of 47), albeit less abundant than UC4 (electronic supplementary material, figure S2B). Since most of the TARA Oceans reads associated to our SAGs appear in the picoplanktonic fraction (electronic supplementary material, figure S2B) our SAGs cell size likely ranges between 0.8 and 5 µm, in agreement with the typical size range of described choanoflagellate species [25]. Furthermore, our four SAGs are relatively more abundant in surface waters than in deeper sampling points such as the deep chlorophyl maximum (electronic supplementary material, figure S2B).

(b) Genome completeness and statistics of the SAGs

Once we had deciphered the taxonomy and the ecological distributions of our SAGs, we sequenced their genomes using Illumina MiSeq as in [34] (see §2 for further details). We then checked the genome recovery and the genome statistics of our final assemblies, including the estimation of genome completeness by BUSCO [48]. UC1 and UC4 presented a significant genome recovery (7.74 MB and 31.68% BUSCO for UC1; 7.25 Mb and 13.53% of BUSCO for UC4) (table 1; electronic supplementary material, table S2). However, UC2 and UC3 were mostly incomplete and fragmented (table 1), and for this reason they were not included in most of the subsequent analyses, except for the eight gene-based phylogeny (electronic supplementary material, figure S3). Interestingly, we were able to recover the mitochondrial genome of UC2 (table 1), which is the first available mitochondrial genome of an acanthoecid choanoflagellate. We could annotate 59 mitochondrial genes (electronic supplementary material, table S3), which revealed a high degree of conservation with the mitochondrial genome of *M. brevicollis* [75].

In order to extrapolate the putative genome size of our SAGs, we performed an estimation using BUSCO end CEGMA values (see §2). The results showed that the craspedidan UC1 (29.4 Mb, see table 2) would potentially contain the smallest genome among the so far sequenced choanoflagellates; *S. rosetta* (55.4 Mb) and *M. brevicollis* (41.6 Mb). The predicted genome length of the early-branching UC4 (52.5 Mb) is similar to that of *S. rosetta*. However, it is worth mentioning that this approach can yield biased results, because it assumes that core eukaryotic genes are homogeneously distributed along the genome, it does not differentially account for fragmented and complete genes, and it does not account for the lower detection rate of fragmented genes inherent to all gene prediction algorithms [53]. Thus, the results have to be interpreted cautiously. Using this approach with previously published SAG of the choanoflagellate *M. brevicollis* [34] with a BUSCO values above 12% (similar to UC1 and UC4) results in an approximately 10% over-estimation of its genome size relative to the reference genome [9]. Finally, it is worth mentioning that alternative methods of genome size estimation, such as

<table>
<thead>
<tr>
<th>SAG</th>
<th>taxonomy</th>
<th>scaffolds</th>
<th>largest scaffold (bp)</th>
<th>N50</th>
<th>total length (Mb)</th>
<th>GC (%)</th>
<th>CEGMA (%)</th>
<th>BUSCO (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC1</td>
<td>Craspedida clade 1</td>
<td>3276</td>
<td>41 637</td>
<td>4928</td>
<td>7.74</td>
<td>49.8</td>
<td>20.1</td>
<td>31.7</td>
</tr>
<tr>
<td>UC2</td>
<td>Acanthoecidae</td>
<td>746</td>
<td>32 186</td>
<td>1499</td>
<td>1.00</td>
<td>30.8</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>UC3</td>
<td>Stephanoeidae</td>
<td>819</td>
<td>11 187</td>
<td>2197</td>
<td>1.31</td>
<td>33.5</td>
<td>—</td>
<td>0.3</td>
</tr>
<tr>
<td>UC4</td>
<td>Basal Acanthoecida</td>
<td>2527</td>
<td>72 672</td>
<td>11360</td>
<td>7.25</td>
<td>40.0</td>
<td>14.1</td>
<td>13.5</td>
</tr>
</tbody>
</table>

*Scaffolds bigger than 500 bp.*

Table 1. Summary of the genome statistics of each SAG assembly.

---

---
k-mer frequency distribution, are unfortunately not adequate for SAG assemblies, as they require that sequencing coverage along the genome be unbiased [76], a condition not fulfilled by SAG-derived short reads [34]. Altogether, these observations suggest that genome size estimates in SAG data suffer from inherent biases and have to be interpreted with caution.

Next, we predicted the number of genes that might contain the full genomic sequences of UC1 and UC4 taxa, by extrapolating the numbers of genes annotated with the BUSCO/CEGMA values, removing the potential contamination and taking into account Pfam protein domain predictions (see §2). The difference in estimated size is proportional to the number of estimated genes. UC1 has a smaller number of genes (6039) according to the predicted reduced genome size, and UC4 would present a more similar number of genes than the previous choanoflagellate genomes (10 075) (table 2).

We then characterized our SAGs by screening for genes linked to morphological structures, such as the microvilli or the lorica of acanthoecids, in order to speculate on the potential morphology of these two choanoflagellate taxa. Therefore, we searched for the presence of Ezrin/Radixin/Moesin (ERM) protein [77,78], which is known to be involved in microvilli elongation processes; as well as for the presence of Si transporters (SITs) [79], needed for the lorica formation in Acanthoeca. The microvilli-related ERM protein was only found in the UC4 genome, and not detected in UC1. We also failed to identify any SIT in any of those taxa, including the sister Achantoecida UC4 [79]. Thus, with these results and without complete genomic data we cannot speculate whether these choanoflagellates present Lorica or not.

(c) Phylogenomics confidently reconstruct the phylogenetic position of UC4 and UC1 and raise questions regarding deep phylogenetic branches of choanoflagellates

To reconstruct the evolutionary history of the different protein domains from the LECA to extant animals, we need a proper phylogenetic framework. Moreover, we were also interested in confidently placing our SAGs within the phylogeny of choanoflagellates. We, thus, built a phylogenomic matrix based on 87 single-copy protein domains [14] over 79 taxa including animals and all their unicellular relatives with available transcriptomic or genomic data. We therefore included the SAGs UC1 and UC4, the recent transcriptomic data of 19 choanoflagellates [23] plus the two choanoflagellates with complete genome sequences, *M. brevicollis* and *S. rosetta* [9,10]. In addition, we included the four known filasterean taxa, including the recently described genera *Pigomantor*. We also included all ichthyosporean taxa with genome information, the early branching ichthyosporean *Chromosphaira perkinsii*, as well as the plurimorfean (or corallochotryeans) *Corallochytrium* and *Syssomonas* [14,41] (figure 2). Finally, we included an extensive outgroup composed by holomycotans (18 taxa), apusomonads (2 taxa), breviates (3 taxa) and amoebzoans (4 taxa).

Our tree recovers monophyly of choanoflagellates with maximum support (figure 2). Our SAGs UC1 and UC4 were confidently placed within choanoflagellates. UC4 is confirmed to be an early-branching sister to Acanthoeca. Thus, UC4 falls in a key phylogenetic position to better understand Choanoflagellata and Acanthoeca evolution. Given its high abundance, it might also be an important ecological player. On the other hand, UC1 is confirmed to be a clade 1 craspedidan, as in the 18S rRNA tree (figure 1), appearing as sister-group to the previously described craspedidan clade 1 [27].

However, and somehow unexpectedly, our tree recovered some important topological differences compared to previous choanoflagellate phylogenies based on a few genes [24,27]. The main one is that in our tree Craspedida appears paraphyletic (figure 2). Interestingly, *Codosiga hollandica* appears as sister-group to the rest of the choanoflagellates prior to the split of *Salpingoeca dolicothecata* and the divergence of Craspedida and Acanthoeca. This position for *C. hollandica* is relatively well supported (75% ML UFBS/0.99pp BI), although the nodal supports increase a bit when the fastest 2500 evolving sites are removed (electronic supplementary material, figure S4); however, nodal support decreases again if more positions are removed from the alignment (electronic supplementary material, figure S4). Hence, our extended analyses show that some relationships are unstable. For instance, the next branching after *C. hollandica*, in our Bayesian reconstruction is *S. dolicothecata* prior to the divergence of Craspedida and Acanthoeca plus UC4, but in the ML inference, *S. dolicothecata* falls sister to Acanthoeca and UC4 (figure 2 and electronic supplementary material, figure S1). Moreover, different topological tests were not able to discard any of the following alternative hypotheses: (1) *C. hollandica* as the earliest branching lineage and *S. dolicothecata* sister to the clade formed by the rest of craspedidans and Acanthoeca; (2) *S. dolicothecata* branching as sister to Acanthoeca and UC4 clade and *C. hollandica* remaining early branching; and (3) the classical view in which Craspedida and Acanthoeca are monophyletic (electronic supplementary material, figure S5). In addition, our eight-gene based phylogeny using a broader taxon sampling, similar to the most recent, and taxon-rich, phylogenetic reconstruction of choanoflagellates [27], brings *C. hollandica* together with other codosiga species within clade 2 of Craspedida. But still, before the split Acanthoeca–Craspedida there is the craspedidan clade 3 composed by *S. dolicothecata* and *Salpingoeca tuba* appearing as the earliest branching clade (electronic supplementary material, figure S3).

<table>
<thead>
<tr>
<th>genome</th>
<th>assembly size (Mb)</th>
<th>genome size (Mb)</th>
<th>no. of annotated genes</th>
<th>total no. of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC1</td>
<td>7.74</td>
<td>29.4</td>
<td>3025</td>
<td>6039</td>
</tr>
<tr>
<td>UC4</td>
<td>7.25</td>
<td>52.5</td>
<td>2518</td>
<td>10 075</td>
</tr>
<tr>
<td><em>Salpingoeca rosetta</em></td>
<td>—</td>
<td>55.4</td>
<td>—</td>
<td>11 624</td>
</tr>
<tr>
<td><em>Monosiga brevicollis</em></td>
<td>—</td>
<td>41.6</td>
<td>—</td>
<td>9172</td>
</tr>
</tbody>
</table>

Table 2. Genome estimation of our SAGs within choanoflagellate context.
**Figure 2.** Phylogenomic tree of holozoans. Phylogenomic analysis of 87 single-copy protein domains [14] accounting for 23 364 amino acid positions. Tree topology is the consensus of two Markov chain Monte Carlo chains run for 5660 and 5685 generations, after a burn-in of 13%. Statistical supports are indicated at each node: on the left, non-parametric ML ultrafast-bootstrap (UFBS) values obtained from 1000 replicates using IQ-TREE and the LG+R7+C60 model; on the right, Bayesian posterior probabilities (BPP) under the LG_+Γ7+CAT model as implemented in Phylobayes. Nodes with maximum support values (BPP = 1 and UFBS = 100) are indicated with a black bullet. Raw trees are available on Figshare (https://doi.org/10.6084/m9.figshare.7819571.v1) and electronic supplementary material, figure S1 shows the topology and the supports of the ML inference.
Thus, our data suggest that with the current information we cannot properly tackle deep choanoflagellate relationships. We need more genomic information from broader taxon sampling, as well as further understanding of choanoflagellate diversity. Previously described environmental clades like Clade L [29] or FRESCHO3-4 groups [30] that branch in early positions in the 18S rRNA phylogeny (figure 1), together with the species S. tuba and others from the genera Codosiga and Sphaeroeca (related to Codosiga in the 18S rRNA (figure 1) and in the eight-gene phylogenies; electronic supplementary material, figure S3) will help to solve these deep choanoflagellate relationships by increasing our knowledge of choanoflagellate diversity in missing (and key) phylogenetic positions.

Another interesting result from our phylogenetic reconstruction is that Salpingoeca urceolata and Salpingoeca keweri appear as sister-group to clade 1 and not within clade 2, as previously described in [27]. Therefore, our results might redefine these two groups of craspedidans (figure 2). On the other hand, our data show with high nodal support that nudiforms cluster within tectiforms, meaning that nudiforms and tectiforms are not two independent lineages within Acanthoecida.

Finally, our results recovered, within unicellular holozoans, the monophyly of Teretosporea [13] with high nodal support (99% of ultra-fast bootstrap from ML (UFBS) and 1 of posterior probability of Bayesian inference (BI) (figure 2). Apparently, the addition of Sysposomonas and Chromosphaerae together in the same phylogeny allows better statistical support than obtained in previous studies [14,41]. Therefore, this is the phylogenetic framework we used in our subsequent protein domain evolution reconstruction analysis.

(d) The Urmetazoan genome did not experience an increase of innovation at the level of protein domains

After establishing the taxonomical framework, we analysed the evolutionary history of protein domains from LECA to animals. To perform this analysis, we built a database of 116 proteomes of different eukaryotic taxa (electronic supplementary material, table S7), representing the entire eukaryotic diversity. We then predicted the protein domain architectures and produced a matrix of presence/absence of each protein domain across all the eukaryotic taxa (see §2). Finally, we inferred the gains and losses among each node of the eukaryotic tree of life by Dollo parsimony. The results shown are focused on opisthokonts (figure 3), using the taxonomical framework obtained in our phylogenomic analysis; except for Ctenophora, in which we assumed Porifera as the earliest branching animal lineage. This phylogenetic framework is in accordance with recent phylogenomic analyses [67] and previous studies of animal genomic reconstructions [14,22,23] and it reduces the effect of the excess of gene losses in the fast-evolving ctenophore Mnemiopsis leidyi [22].

The first observation is that the number of novel protein domains gained in the last common ancestor (LCA) of Metazoa (181) is in line with the number of gains in previous LCAs since the origin of Opisthokonta (domains gained: 248 Opisthokonta LCA, 140 Holozoa LCA, 69 Filozoa LCA, 120 Choanozoza LCA; figure 3), without an important increase of new protein domain acquisitions at the stem of Metazoa compared with its unicellular ancestors. On the other hand, these gains are offset by a high number of protein domain losses compared with the Choanozoan ancestor (animals+choanoflagellates) (from 5715 to 5497 protein domains).

Among the losses at the stem of metazoans, we found protein domains involved in the biosynthesis of essential amino acids (for instance: Anth_synth_I_N, Shikimate_dh_N, PDT) and carbohydrate metabolism (Pantoate_ligase, Glyco_transf_34) as previously suggested [23,80–82].

In addition, we observed an important fraction of protein domains lost that had a bacterial ancestry (i.e. ion transporters; see supplementary material at Figshare https://doi.org/10.6084/m9.figshare.7819571.v1). Thus, with our analysis it is difficult to disentangle which is the real origin of these protein domains. We can expect that most of the domains were already present in the LECA inherited from vertical transmission from prokaryotic ancestors, but others might have been acquired in extinct eukaryotic lineages by horizontal gene transfer; for example, bacterial rhodopsins (domain Bac_rhodopsin) are described as an horizontal gene transfer from bacteria to eukaryotes [83]. We discard the possibility of these domains coming from contaminants because the genomes/transcriptomes used in our dataset are of good quality and we were very strict in the treatment of possible contaminations in our SAGs data (see §2).

Thus, in order to have more data to interpret these results, we compared the retention of these Metazoan-lost protein domains with the retention of all protein domains present at the LECA, in the rest of eukaryotic species that were not animals. The results show that the domains lost in Metazoa are less retained in the rest of eukaryotic species than all protein domains present at LECA (a median of 18% of retention for metazoan lost domains compared with a 60% of retention of all domains present at LECA; electronic supplementary material, figure S6). Actually, only 40 out of the 399 protein domains lost at Metazoa are retained in half or more of the rest of eukaryotic species (electronic supplementary material, table S8). Among these 40 clear Metazoan losses are included the protein domains mentioned above that are related in amino acid biosynthesis and carbohydrate metabolism (electronic supplementary material, table S8).

Thus, either most of these protein domains lost in metazoans were vertically acquired from the LECA to the Choanozoan ancestor and subsequently lost in animals, which will imply major losses in other eukaryotic lineages, or most of these protein domains are recent acquisitions by horizontal gene transfer events from bacteria to some extinct eukaryotic lineages, including Choanoflagellates. Probably both scenarios are occurring in our metazoan losses, but without a detailed phylogenetic analysis of these domains we cannot discover to what extent horizontal gene transfer events are affecting our results.

Thus overall, our data show that, at the stem of animals, there was not an important increase in new protein domains, and at the same time, many important metabolic functions were lost. Therefore, if we compare our results with previous studies that had shown an important acquisition of new genes in the origins of animals [14,22,23]—and that some of these acquisitions were the product of domain shuffling events [9,10,12,14]—we might conclude that such appearance of new gene families at the onset of Metazoa was facilitated by a massive rearrangement and duplication of pre-existent protein domains and not by the gain of new protein domains.

Finally, it is worth mentioning that an important fraction of the losses observed might correspond to horizontal gene transfer events to the unicellular relatives of animals, thus it
**Figure 3.** Summary of proteins gains and losses in Opisthokonta, focusing on Choanozoa gains. (a) Schematic of the choanoflagellate phylogeny obtained, including the numbers of protein domain gains and losses in each Opisthokonta clade (depicted in green and red respectively). Light green numbers represent the protein domain gains that have retention of over 70% in extant metazoan species, in a given ancestor. Protein domains from potential bacterial or archeal contamination were excluded from the analysis (see §2). The ability to form colonies (marked with a colony drawing) is shown on the right, and has been adapted from [27]. Our SAGs (UC1 and UC4) are marked in italic. Next to the tree, there is a bar chart indicating the percentage of protein domains gained at Choanozoa and judged to be involved in animal multicellular processes (a total of 69 domains out of 120), retained in each choanoflagellate taxa. Animal data are displayed by phylum instead of species, thus what it is shown is the average and the distribution of domains kept by all the analyzed species of each animal phylum. As a control, the retention of all protein domains that have the Choanozoan ancestor among extant species is shown in grey. Further to the right, the POU protein domain distribution and the protein domains gained at Choanozoa are shown, which are present in our sequenced SAGs UC1 and UC4. A black dot indicates the presence of each domain in the different taxa/clade. (b) Function of the protein domains gained at Choanozoa. In green, the biochemical roles in which the protein domains are involved. In blue, the biological processes in which the domains have been shown to participate. These two classifications are not exclusive; one protein domain can appear in one or multiple categories. In grey, protein domains with unknown function, or contaminants or a product of an horizontal gene transfer (HGT) event.
might be the case that there was not a net loss of protein domains at the stem of Metazoa.

(e) Pre-metazoan protein domains essential for animal multicellularity are not retained in their unicellular relatives

Next, we questioned which of these pre-existing protein domains were more important in the transition towards animal multicellularity, and when they appeared. To analyse this, for each domain acquired in the successive ancestors from the last Opisthokonta common ancestor (LOCA) to the Urmetazoa, we plotted the probability of retention on extant metazoan species and their unicellular relatives (figure 4). SAGs UC1 and UC4 were not taken into account for this probabilistic analysis because their genomes are not complete, and therefore the lack of data would underestimate the probability of protein domain retention in choanoflagellates. The rest of the eukaryotic species used in the analysis have high BUSCO values (greater than 90% in most cases) and present similar numbers of total protein domains (mostly between 3000 and 4500); electronic supplementary material, table S9), except for vertebrate species (3 out of 21 Metazoans), which are better studied and present around 5500 protein domains, and the parasitic species of Microsporidia (around 1000). Since we have 22 other holomycotan species, we do not consider that the low number of genes in Microsporidia would significantly affect our results (electronic supplementary material, table S9).

The results show that the protein domains acquired at the base of opisthokonts were proportionally less retained in the extant metazoan species than the protein domains acquired in the successive Holozoan ancestors. This implies that an important fraction of the 248 protein domains gains occurred at LOCA that are not crucial for animal functions. On the other hand, the protein domains gained at the stem of Holozoa and the successive unicellular ancestors (Filozoa and Choanozoa) are likely to be retained in extant animal species, but not in their unicellular relatives (figure 4). Figure 3a shows the number of protein domain gains that have retention of over 70% in extant metazoan species. We can observe that unicellular holozoan ancestors contributed significantly in the acquisition of these conserved protein domains (139 in total; 62 Holozoa, 27 Filozoa, and 50 Choanozoa). Finally, protein domains acquired at the origins of animals have the highest rates of retention in extant animals (87 protein domain innovations were retained by more than 70% of metazoan taxa).

Therefore, the protein domains acquired at the origins of animals seems to be more essential for animal functions because they tend to be more conserved. However, proportionally, the protein domains gained in the different unicellular holozoan ancestors are also key for maintaining such functions, presenting a high degree of conservation among extant animal species compared with their unicellular counterparts.
(f) Key protein domains for animal multicellularity
To get a better understanding of the set of conserved protein domains from LOCA to the Urmetazoa, we list in table 3 the protein domains that were retained in all animal species here analysed (21 species). Those protein domains are related to biological processes that have been mainly hypothesized to be required for the evolution of animal multicellularity. These processes include gene expression regulation, cell-to-cell communications and cell adhesion [9,12,20,84,85]. Most of the domains listed predate the origins of animal multicellularity. Protein domains like T-Box, runt, Integrin_beta_2 and Laminin_N were already described in proteins shown to be of pre-metazoan origin [16,19,85,86]. However, our analysis also shows highly conserved protein domains that predate the origins of animals, like LEM or NUDE_C. LEM acts in a protein of the inner nuclear membrane involved in the chromatin organization and the post-mitotic re-assembly of the nucleus [87]. NUDE_C is the C terminal protein domain of the NDE1 protein, which is required for centrosome duplications and formation and the correct functioning of the mitotic spindle. It has also been described as essential for the development of the cerebral cortex, by controlling the orientation of the mitotic spindle in cortical neuronal progenitors [88]. Thus, besides cell-to-cell communications, cell adhesion and gene expression regulation, our results suggest that other more general eukaryotic functions, such as chromatin organization, cell division, ubiquitination (beta-TrCP_D domain) or translational regulation at ribosomal scale (RAC_·head domain) (table 3) were more specialized in metazoans mediated by these highly conserved proteinic domains.
Finally, we also detected among these `essential domains' that predate the origins of animals, protein domains of unknown function, such as Calpolin (table 3).

On the other hand, we found seven protein domains that were gained at the origin of animals and conserved in all of the 21 metazoan species analysed. Most of these protein domains have already been described as animal-specific, like two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families, two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families, two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families, two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families, two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families, two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families, two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families, two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families, two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families, two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families, two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families, two components of major signalling pathways (Wnt and TGF-beta); and two essential transcription factor families. These include the protein domains POU and zf-C4, both previously thought to be animal-specific [22,85]. POU is the N-terminal protein domain of the POU homeobox gene family. POU genes are known for their roles in cell-type specification and developmental regulation in animals [89]; hence they are essential genes for animal multicellularity (table 3). Among choanoflagellates, we only identified the POU protein domain in the choanoflagellate species Mylnosiga fluctuans, appearing together with a homeobox domain, adopting the animal-like structure of POU genes.

In order to confirm that POU homeobox transcription factors were already present in the Choanozoan ancestor, we performed a phylogenetic analysis of the adjacent homeobox domain (which is conserved at the pan-eukaryotic level [90], using LIM-associated homeobox as outgroup [90] (figure 5). The phylogenetic reconstruction places Mylnosiga’s homeobox domain within the animal POU family with high nodal support (98% UFBS; figure 5a). It falls in an early-branching position, before the expansion into different paralogs in animal species [89]. Thus, according to the results, we consider Mylnosiga’s POU homeobox domain as a bona fide POU orthologue. While sequence contamination could potentially explain the presence of this POU homeobox in Mylnosiga, two observations render this possibility unlikely: first, POU domains were previously considered to be animal-specific, thus ruling out non-animal contaminants [86]; second, the peptide sequence of the homeobox domain is clearly different from canonical animal POU domains and consistent with the phylogenetic relationships between choanoflagellates and animals (early-branching; figure 5a). Furthermore, secondary losses of POU N-terminal domains have already been reported within bona fide animal homologues of the POU homeobox class [90], which can explain the low level of conservation of the associated POU domain in Mylnosiga (see alignment in supplementary material, Figshare https://doi.org/10.6084/m9.figshare.7819571.v1).

Finally, in order to clarify the POU phylogeny, we performed a joint phylogenetic analysis of the POU and homeobox domains. This analysis of the whole protein (POU plus homeobox domain, figure 5b) revealed a moderate improvement in the nodal support in most POU classes nodes. Surprisingly, Mylnosiga POU protein was associated to POU class 2 proteins, although with low bootstrap support (65% UFBS; figure 5b). Given that POU class 2 is reported to have appeared at the stem of Bilateria [86], this scenario would imply a rampant gene loss of POU classes in non-bilaterian animals and in choanoflagellates. This scenario is not very parsimonious, and the association of Mylnosiga to POU class 2 proteins might also be explained as an artefact owing to the lack of resolution in our phylogenetic inference. More genomic sequences of non-bilaterian and choanoflagellate organisms could help resolve this issue.

On the other hand, we also detected the zf-C4 domain in choanoflagellate species. The zf-C4 domain is the DNA binding domain of the nuclear hormone receptors. We identified the domain alone in choanoflagellate protein sequences, without the animal-like structure, in which the zf-C4 domain is accompanied by the hormone_receptor Pfam domain.

The hormone_receptor domain is animal-specific and highly conserved in extant metazoans (table 3), having been lost only in the ctenophore M. leali and in the poriferan Oscar ella carnea. Thus, nuclear hormone receptors remain animal innovations under the current taxon sampling, but the DNA binding domain zf-C4 predate the origin of animals.

Finally, we also detected the presence of two protein domains, MH1 and MH2, that are highly conserved among
<table>
<thead>
<tr>
<th>Origin</th>
<th>Protein Domains Retained</th>
<th>Protein Domain Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opisthokonta</td>
<td>Transcription factors and DNA binding domains</td>
<td>Transcription factor involved in animal development</td>
</tr>
<tr>
<td></td>
<td>T-box</td>
<td>Nuclear effector of Notch signalling</td>
</tr>
<tr>
<td></td>
<td>BTD</td>
<td>Related to BTD, nuclear effector of Notch signalling</td>
</tr>
<tr>
<td></td>
<td>LAG1-DNAbind</td>
<td>Involved in chromosome migration</td>
</tr>
<tr>
<td></td>
<td>NUDE_C</td>
<td>Interacts with STAT6 transcription factor</td>
</tr>
<tr>
<td></td>
<td>PAS_11</td>
<td>Signalling and GTPase interactors</td>
</tr>
<tr>
<td></td>
<td>Arfaptin</td>
<td>Involved in the vesicle budding at Golgi apparatus</td>
</tr>
<tr>
<td></td>
<td>GIT_SHD</td>
<td>Signalling integrators with GTPase activity</td>
</tr>
<tr>
<td></td>
<td>RAC_head</td>
<td>Involved in ribosomal binding</td>
</tr>
<tr>
<td></td>
<td>Calpolin</td>
<td>Translational regulator</td>
</tr>
<tr>
<td></td>
<td>HS1_rep</td>
<td>Unknown function</td>
</tr>
<tr>
<td></td>
<td>DUF3518</td>
<td>Unknown function</td>
</tr>
<tr>
<td></td>
<td>DUF3585</td>
<td>Unknown function</td>
</tr>
<tr>
<td>Holozoa</td>
<td>Signalling binding-related domains</td>
<td>Interacts with guanylate kinase-like domain</td>
</tr>
<tr>
<td></td>
<td>GKAP</td>
<td>Phosphotyrosine interacting domain</td>
</tr>
<tr>
<td></td>
<td>PID</td>
<td>Known to be present in syntaxin-binding proteins</td>
</tr>
<tr>
<td></td>
<td>LLGL</td>
<td>Found in inner nuclear membranes</td>
</tr>
<tr>
<td></td>
<td>LEM</td>
<td>Nuclear membrane protein</td>
</tr>
<tr>
<td></td>
<td>Runn</td>
<td>Transcription factor related to animal development</td>
</tr>
<tr>
<td></td>
<td>DUF1908</td>
<td>Unknown function</td>
</tr>
<tr>
<td>Filozoa</td>
<td>Signalling and adhesion</td>
<td>Extracellular domain of integrins</td>
</tr>
<tr>
<td></td>
<td>Integrin_alpha2</td>
<td>Ubiquitination</td>
</tr>
<tr>
<td></td>
<td>Beta-TrCP_D</td>
<td>D domain of beta-TrCP that acts as ubiquitin ligase</td>
</tr>
<tr>
<td>Choanozoa</td>
<td>Transcription factors</td>
<td>DNA binding domain of Smad TF</td>
</tr>
<tr>
<td></td>
<td>MH1</td>
<td>Domain that interacts with Smad TF regulators</td>
</tr>
<tr>
<td></td>
<td>MH2</td>
<td>Domain related with Homeobox superfamily</td>
</tr>
<tr>
<td></td>
<td>Pou</td>
<td>Extracellular matrix protein domains</td>
</tr>
<tr>
<td></td>
<td>Laminin_N</td>
<td>N terminal domain of laminins, extracellular proteins related to cell adhesion</td>
</tr>
<tr>
<td></td>
<td>P4Ha_N</td>
<td>Domain from prolyl 4-hydroxylase that is important in the post-translational modification of collagen</td>
</tr>
<tr>
<td></td>
<td>TSP_C</td>
<td>C terminal domain of Thrombospondin, an adhesive glycoprotein that mediates cell-to-cell and cell-to-ECM interactions</td>
</tr>
<tr>
<td></td>
<td>PET</td>
<td>Suggested to be involved in protein–protein interactions</td>
</tr>
<tr>
<td></td>
<td>Lamp</td>
<td>Integral membrane proteins of the lysosome with unclear functions</td>
</tr>
</tbody>
</table>

(Continued.)
metazoans (table 3) and were recently found as Choanozoan innovations [23]. MH1 is the DNA binding protein of Smad transcription factors that together with MH2 form the canonical Smad proteins [91]. Both MH1 and MH2 have a Choanozoan origin even though the canonical Smad architecture remains metazoan-specific, as previously described [20,85]. Thus, Smad proteins might be the result of a domain shuffling event at the stem of Metazoa, as already described [23].

(h) Colonial choanoflagellates do not specially retain protein domains related in multicellular-like functions

We identified that protein domains that appeared at the origin of animals and choanoflagellates are involved in crucial functions to maintain animal multicellularity (figure 4b). These are protein domains involved not only in development, cell-to-cell signalling or adhesion, but also in other multicellular functions [19] such as neural functions, immunologic response [9,10,23], cell cycle control, or the control of cell polarity and division. Those ‘multicellular’ protein domains appeared to be more frequently retained in animals than in choanoflagellates (figure 3c), showing a very different pattern of conservation of the 5715 protein domains present at the Choanozoan ancestor (figure 3a, grey line behind the bar chart). Thus, we interrogated whether these ‘multicellular-animal-like’ protein domains are involved in colony formation in choanoflagellates. Our data show that colonial choanoflagellates have not kept more (16.5 on average) of those protein domains related in multicellular functions than non-colonial taxa (18 on average). This may suggest that the molecular mechanisms involved in animal and choanoflagellate multicellularity require different protein players. One clear example of this is the POU gene, which is present in all animal taxa, but only found in the choanoflagellate M. fluctuans.

(i) UC1 and UC4 SAGs contain protein domains involved in key animal functions

Among the protein domains that originated in Chaozna, nine protein domains were recovered in our SAGs and also in other choanoflagellate taxa (figure 3a). In particular, our UC1 SAG recovered two of the most conserved protein domains in extant animal species: Lamp and TSP_C. Lamp proteins are uni-domain lysosome-associated membrane glycoproteins, which seem to be required for phagocytic processes and are related to immunogenic responses [92]. Besides UC1, Lamp proteins are more conserved among Craspedida species within choanoflagellates (figure 3). TSP_C is the C-terminal domain of thrombospondin proteins, which are secreted proteins that interact with the extracellular matrix and plasma proteins. Thrombospondins are related to embryonic development, tissue differentiation, tumour growth and angiogenesis [93]. Choanoflagellates contain only the C-terminal domain; the full protein architecture is animal-specific. We also found other protein domains on UC1, which are highly conserved among animal species. Those are Sema and NKAIN (Sema, 20 out of 21 species; NKAIN, 18 out 21 species), both involved in extracellular receptors or transmembrane proteins that participate in neural functions in animals. NKAIN is a sodium-dependent ATPase interacting protein [94], while Sema is the core domain of semaphorins and their binding receptors, Plexin proteins. Semaphorins and Plexins are signalling extracellular proteins involved in the guidance of axon formation in neural development [95]. A recent phylogenetic reconstruction of Semaphorins and Plexins showed the pre-metazoan origin of this domain, which was duplicated before the origin of Choanozoa [96]. Interestingly, the Sema domain found in the SAG UC1 belongs to a Plexin-like protein (electronic supplementary material, figure S7). Therefore, Plexins, semaphorins, and NKAIN proteins predate the origins of animals,
together with other genes related in neural functions such as sodium [97] and calcium channels [98], neuroglobulins [99], and proteins related in synapsis [10,100] and neural secretion [101].

Finally, our SAGs contain two protein domains gained at Choanozoa that are less conserved in all metazoan species: NPM1-C and Fanconi_A_N (figure 3). NPM1-C is the C-terminal protein domain of the transcription factor Nucleophosmin, previously thought to be vertebrate-specific [102]. NPM1-C is essential in the regulation of DNA replication malfunctions, a n di ti si n v o l v e di np 5 3 - medi at ed pa thw a y st op r o mo te apo pt o si s in case of DNA damage [102]. Nucleophosmin protein domain architecture consists of the Nucleoplasmin protein domain followed by NPM1-C in all animals. The domain Nucleoplasmin is paneukaryotic, and the protein domain architecture consisting of Nucleoplasmin plus NPM1-C is animal-specific according to our results. In the SAG UC4 we identified the domain NPM1-C, which we confirmed by phylogeny not to be a contamination (electronic supplementary material, figure S8), while the Nucleoplasmin domain was missing. As SAGs genomes are partial, we cannot rule out the possibility that the Nucleoplasmin domain is indeed present in this taxa. However, we believe the most likely explanation is that the animal Nucleophosmin, as Smad proteins, appeared as a product of a domain shuffling between the two more ancient Nucleophosmin domains (Nucleoplasmin and NPM1-C) at the origin of animals. It may not be essential in maintaining animal functions, given its low conservation among extant animal species (30%), but it might be crucial to perform vertebrate-specific functions because it is conserved among all analysed vertebrates.

Fanconi_A_N is the N-terminal domain of the Fanconi anaemia complementation group A protein (FANCA human protein) that acts in DNA damage-repair processes and also in the differentiation of blood cells [103]. Mutations in these gene cause Fanconi anaemia (FA) in humans [103]. Our data show a pre-metazoan origin for the Fanconi_A_N domain, and a vertebrate acquisition for the Fanconi_A domain that, together with Fanconi_A_N, conform the canonical FANCA protein. This is probably the reason why this protein domain, like Nucleophosmin, is conserved in all vertebrate species.

Thus, our SAGs from uncultured choanoflagellate species, UC1 and UC4, together with the new choanoflagellate transcriptomic data [23] have allowed us to show the pre-metazoan origin of protein domains, present in overlooked proteins that could have been essential in the origins of animals and, therefore, in the transition towards animal multicellularity. These include Lamp, thrombospondins, NKAIN, Semaphorins and Plexin proteins. Finally, we also identified proteins that are particularly conserved among vertebrate species but not in other metazoans, like Nucleophosmin and FANCA protein.

Figure 5. POU phylogenetic tree. (a) ML inference of Homeobox domain using LIM homeobox as an outgroup. LIM domains were downloaded from (http://homeodb.zoo.ox.ac.uk/download.get). Pou domains were selected from a wide range of metazoans available in our dataset, plus the choanoflagellate sequence (marked in red). Supports are SH-like approximate likelihood ratio test (left) and UFBS, respectively (right) calculated with IQ-TREE v. 1.5.1. (b) Maximum-likelihood inference of Pou transcription factors using the whole protein, the Pou domain and the Homebox domain. M. fluctuans Pou sequence falls within POU-2 group (marked in red).
4. Discussion

In this work, we took advantage of the single-cell genomics technique to expand the extant genomic diversity of choanoflagellates by recovering a substantial proportion of the genomes of two uncultured choanoflagellate species (UC1 and UC4). Our data, together with the recent new genomic/transcriptomic information from 19 choanoflagellates and other unicellular holozoans [14,23,41], have allowed us to perform, with an unprecedented level of detail, the reconstruction of the protein domains’ gains and losses from the LECA to the Urmetazoan, improving our understanding of the genomic changes that allowed the transition towards multicellularity in animals.

As commented in previous studies, the current state of single-cell genomics techniques based on cell isolation, cell lysis and whole genome amplification with MDA, has important limitations in terms of genome recovery [34,40]. Our results are in agreement with the literature: only two out of our four SAG assemblies (UC1 and UC4) contained enough genomic information (31.7–13.5% BUSCO completeness values) to perform gene and protein domain annotation. These limitations have a strong effect on genome-wide analyses of gene evolution based on SAG data. For instance, gene-level comparative analyses are strongly affected by the incompleteness of SAG assemblies. Similarly, SAG-derived gene predictions cannot be used in probabilistic ancestral reconstructions of gene content that rely on estimations of gain and loss rates at different branches (e.g. [72]) like, for instance, our estimations of gene retention probabilities in various clades (figure 4). In this case, the inclusion of UC1 and UC4 in this analysis would have had underestimated the probability of gene retention in the wider choanoflagellate clade.

Thus, genome-wide comparative analyses using SAG data need to be designed aiming to minimize the effects of incompleteness biases. For example, gene family evolutionary studies can be based on protein domains instead of full-length genes (to take advantage of the fact that the recovery rate of individual protein domains from SAG assemblies is higher than for genes [34]). In addition, our comparative framework complemented our SAGs with a wide array of complete genomic and transcriptomic data from other choanoflagellates (21 in total) [9,10,23] and other eukaryotes. By taking these limitations into account in our experimental design, we were able to take leverage of the limited genomic information contained in UC1 and UC4 in two key analyses: our phylogenomic analysis, which includes the widest sampling of choanoflagellates to date (figure 2), and the reconstruction of ancestral protein domain evolution from the LECA to the origin of animals (figure 3). This ancestral reconstruction, based on Dollo parsimony, led to the identification of a Nucleophosmin-linked domain as evolving in the Choanozoan ancestor, which could be confirmed using phylogenetic analysis (electronic supplementary material, figure S7). Thus, single-cell genomics data, with its limitations, can be a good resource for phylogenomics and also for gene family evolutionary studies.

From the protein domain evolutionary analysis, we show that the high degree of genic innovation that occurred at the stem of Metazoan did not coincide with an important increase in protein domain richness. Instead, while the Urmetazoan ancestor acquired a large number of new gene families, it contained fewer protein domains than its unicellular Choanozoan ancestor. There are two reasons for this. First, because animals lost many genes related to metabolic functions mainly involved in amino acid biosynthesis and carbohydrate metabolism, implying an important change in their metabolic and ecological niche capabilities [23]. Second, many new gene families were the result of the combined action of gene duplications and domain shuffling events [14], which originated the important increase of around 1500–1700 new gene families at stem of Metazoan [22,23].

We have also described many examples of protein domains that appeared before the transition to animal multicellularity and that are highly conserved among animals. However, when encoded in pre-metazoan proteins, they do not present the animal-like protein domain architectures. This is the case, for example, for Nucleophosmin or the nuclear hormone receptor. Thus, domain shuffling events that were already described to explain the appearance of Notch [9,12], or more recently, Smad proteins [23], seem to be the mechanism by which the unicellular ancestor of animals rearranged its genome as it progressed towards animal multicellularity, together with domain duplications. However, we cannot discard the less parsimonious possibility that the Choanozoan ancestor had a much more complete set of protein architectures than animals that has been lost in the lineage leading to extant choanoflagellates. Interestingly, the choanoflagellate M. fluctuans has a clear Notch homologue with the prototypical EGF, Notch transmembrane and Ank domains in canonical order [23].

Regarding our analysis of the evolutionary retention of protein domains, we show that the gains acquired in the unicellular holozoan ancestors (Holozoa, Filozoa, Choanozoa) were, proportionally, more retained by animals than by their unicellular counterparts (figure 4). However, this can be biased by the poor knowledge that we have of the unicellular relatives of animals, given that specific protein domains acquired in the unicellular holozoan ancestors, which were subsequently lost or poorly conserved in animals, would be very unlikely to appear in the Pfam database. As it has been shown that choanoflagellates contain a rich set of specific genes [23], this situation might be affecting our results.

Following this rationale, the coloniality in choanoflagellates would be largely based on genetic innovations without direct homology to animal genes, as both multicellular stages (choanoflagellate colonies and clonal multicellularity of animals) might have been facilitated by different genetic players. This comparative genomic approach is in agreement with previous transcriptomic analyses that showed that the colonial stage of S. rosetta is enriched in evolutionarily recent choanoflagellate-specific genes [10]. Thus a deeper, comprehension of choanoflagellate gene functions is required to understand the subtle homology relationships between animal multicellularity and choanoflagellate coloniality.

Finally, our analysis has shown that POU genes were already present before the transition towards animal multicellularity. POU is a transcription factor involved in development and the maintenance of undifferentiated cells [89]. Thus, it is unknown which functions were involved in the Choanozoan ancestor. Together with POU, we revealed other overlooked proteins whose canonical protein domain architecture is highly conserved among metazoans, like Lamp or thrombospondin proteins. One could argue that Lamp proteins (lysosomal membrane proteins that have been related with phagocytosis [92]) could have been key in improving the predatory capabilities of the unicellular metazoan ancestor. These new
prey, the tissue and cellular relationships of the choanoflagellates [104,105].

5. Conclusion

Overall, in this work we have revealed the genomic changes that facilitated the origins of animal multicellularity. These include massive domain shuffling events and duplications of pre-existing protein components, discrete acquisition of new domains and the loss of metabolic proteins. We have also described the composition of the Urmetazoan and several unicellular ancestors of animals, showing—and in some cases redefining—the evolutionary origin of the most essential protein domains required for animal multicellularity. Finally, we revealed that proteins extremely conserved in all animal species are not limited to the usual suspects involved in functions such as signalling, adhesion or gene expression regulation (like wnt, integrins or Ets transcription factors). Indeed, there is a wider spectrum of overlooked protein domains involved in neural functions, nuclear organization, phagocytosis or without a known-associated function, which are highly conserved among extant animals species. Thus, we here have expanded the knowledge of new putative genetic players required for the emergence of animal multicellularity.

Data accessibility. Raw reads can be found at ENA, project accession number PRJEB31614. SAGs assemblies and annotations, protein domain matrix, summary of protein domain gains in different Opisthokonta ancestors, metazoan losses information and all the supplementary information of the phylogenetic analysis can be found at Figshare (https://doi.org/10.6084/m9.figshare.7819571.v1).

Authors’ contributions. D.L.-E. and I.R.-T. designed the study. M.E.S. collected the cells from the environment. A.G.-A. and M.G. developed protocols for library preparation and sequencing. D.L.-E. generated the genome assemblies and annotations and performed all the analysis. Comparative genomics analysis and phylogenetic reconstructions were performed with the assistance of X.G.-B. D.L.-E. and I.R.-T. wrote the manuscript, which was critically reviewed by all the authors.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by a European Research Council Consolidator grant no. (ERC-2012-Co-616960) to I.R-T.

References


Pesan S et al. 2015 Open science resources for the discovery and analysis of Tara Oceans data. Sci. Data 2, 150023. (doi:10.1038/sdata.2015.23)


