# Dynamics of genomic innovation in the unicellular ancestry of animals

**Xavier Grau-Bové[1,2]*, Guifré Torruella[3], Stuart Donachie[4,5], Hiroshi Suga[6], Guy Leonard[7], Thomas A Richards[7], Iñaki Ruiz-Trillo[1,2,8]***

[1]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Catalonia, Spain; [2]Departament de Genètica, Microbiologia i Estadística, Universitat de Barelona, Barcelona, Catalonia, Spain; [3]Unité d'Ecologie, Systématique et Evolution, Université Paris-Sud/Paris-Saclay, AgroParisTech, Orsay, France; [4]Department of Microbiology, University of Hawai'i at Mānoa, Honolulu, United States; [5]Advanced Studies in Genomics, Proteomics and Bioinformatics, University of Hawai'i at Mānoa, Honolulu, United States; [6]Faculty of Life and Environmental Sciences, Prefectural University of Hiroshima, Hiroshima, Japan; [7]Department of Biosciences, University of Exeter, Exeter, United Kingdom; [8]ICREA, Passeig Lluís Companys, Barcelona, Catalonia, Spain

**Abstract** Which genomic innovations underpinned the origin of multicellular animals is still an open debate. Here, we investigate this question by reconstructing the genome architecture and gene family diversity of ancestral premetazoans, aiming to date the emergence of animal-like traits. Our comparative analysis involves genomes from animals and their closest unicellular relatives (the Holozoa), including four new genomes: three Ichthyosporea and *Corallochytrium limacisporum*. Here, we show that the earliest animals were shaped by dynamic changes in genome architecture before the emergence of multicellularity: an early burst of gene diversity in the ancestor of Holozoa, enriched in transcription factors and cell adhesion machinery, was followed by multiple and differently-timed episodes of synteny disruption, intron gain and genome expansions. Thus, the foundations of animal genome architecture were laid before the origin of complex multicellularity – highlighting the necessity of a unicellular perspective to understand early animal evolution.

DOI: 10.7554/eLife.26036.001

***For correspondence:** xavier.graubove@gmail.com (XG-B); inaki.ruiz@multicellgenome.org (IR-T)

**Competing interests:** The authors declare that no competing interests exist.

**Funding:** See page 28

**Received:** 21 February 2017
**Accepted:** 11 July 2017
**Published:** 20 July 2017

**Reviewing editor:** Diethard Tautz, Max-Planck Institute for Evolutionary Biology, Germany

## Introduction

The transition from a unicellular organism to the first multicellular animal, more than 600 million years ago (**Budd and Jensen, 2017**; **dos Reis et al., 2015**), marks one of the most radical evolutionary innovations within the eukaryotes. Although multicellularity has independently evolved multiple times in the eukaryotic lineage, the highest levels of organismal complexity, body plan diversity and developmental regulation are found in the Metazoa (**Grosberg and Strathmann, 2007**). Key advances in the study of animal origins have been made by comparing the genomes of early branching metazoa, such as cnidarians, ctenophores or sponges (**Putnam et al., 2007**; **Srivastava et al., 2010a**; **Moroz et al., 2014**; **Srivastava et al., 2008**; **Fortunato et al., 2014**), with their closest unicellular relatives in the Holozoa clade, such as the choanoflagellates *Monosiga brevicollis* and *Salpingoeca rosetta* (**King et al., 2008**; **Fairclough et al., 2013**), and the filasterean *Capsaspora owczarzaki* (**Suga et al., 2013**) (**Figure 1**). By focusing on the transition, it is possible to determine which genomic innovations occurred at the origin of metazoa, and whether it required the invention of novel genes or structural features.

**eLife digest** Hundreds of millions of years ago, some single-celled organisms gained the ability to work together and form multicellular organisms. This transition was a major step in evolution and took place at separate times in several parts of the tree of life, including in animals, plants, fungi and algae.

Animals are some of the most complex organisms on Earth. Their single-celled ancestors were also quite genetically complex themselves and their genomes (the complete set of the organism's DNA) already contained many genes that now coordinate the activity of the cells in a multicellular organism.

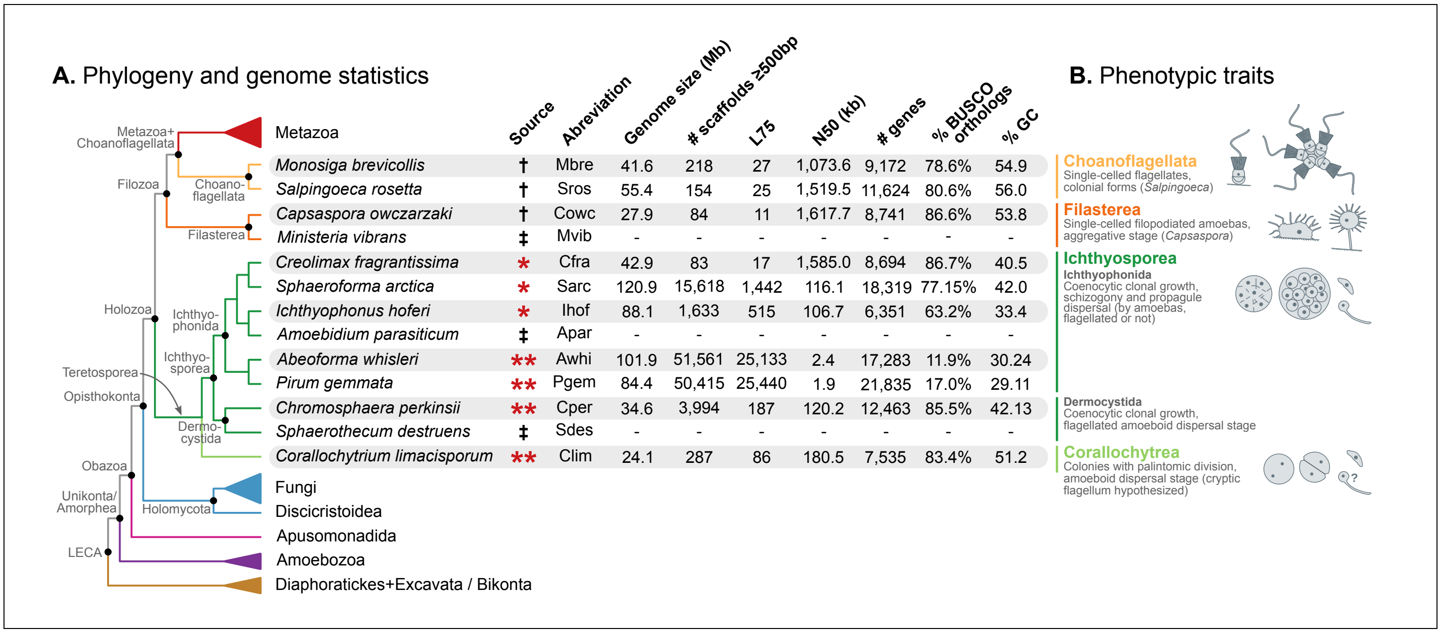
The genome of an animal typically has certain features: it is large, diverse and contains many segments (called introns) that are not genes. By seeing if the single-celled relatives of animals share these traits, it is possible to learn more about when specific genetic features first evolved, and whether they are linked to the origin of animals.

Now, Grau-Bové et al. have studied the genomes of several of the animal kingdom's closest single-celled relatives using a technique called whole genome sequencing. This revealed that there was a period of rapid genetic change in the single-celled ancestors of animals during which their genes became much more diverse. Another 'explosion' of diversity happened after animals had evolved. Furthermore, the overall amount of the genomic content inside cells and the number of introns found in the genome rapidly increased in separate, independent events in both animals and their single-celled ancestors.

Future research is needed to investigate whether other multicellular life forms – such as plants, fungi and algae – originated in the same way as animal life. Understanding how the genetic material of animals evolved also helps us to understand the genetic structures that affect our health. For example, genes that coordinate the behavior of cells (and so are important for multicellular organisms) also play a role in cancer, where cells break free of this regulation to divide uncontrollably.

DOI: 10.7554/eLife.26036.002

We now know that the animal ancestor was already a genomically complex organism, with a rich complement of genes encoding proteins related to a multicellularity. These include transcription factors, extracellular matrix components and intricate signaling pathways that were previously considered animal-specific, but were already poised to be co-opted for multicellularity when animals emerged (***Fairclough et al., 2013***; ***Suga et al., 2013***; ***Richter and King, 2013***; ***Manning et al., 2008***; ***Suga et al., 2012***; ***de Mendoza et al., 2013***; ***Sebé-Pedrós et al., 2017***). Suggestively, detailed analyses of the transcriptomic and proteomic regulatory dynamics of *Capsaspora* and *Salpingoeca* showed that these genes are frequently implicated in the transition to life stages reminiscent of multicellularity – aggregative in *Capsaspora* (***Sebé-Pedrós et al., 2013***, ***Sebé-Pedrós et al., 2016a***), and clonal colonies in *Salpingoeca* (***Fairclough et al., 2013***). Furthermore, the genome architectures of extant Metazoa are, in many aspects, markedly different from most other eukaryotes: they have larger genomes (***Elliott and Gregory, 2015a***), containing more (***Csuros et al., 2011***) and longer introns (***Elliott and Gregory, 2015a***) that can sustain alternative splicing-rich transcriptomes (***McGuire et al., 2008***; ***Irimia and Roy, 2014***), have richer complements of repetitive sequences such as transposable elements (***Elliott and Gregory, 2015b***) and are structured in ancient patterns of gene linkage associated with transcriptional co-regulation (***Irimia et al., 2012***; ***Simakov et al., 2013***) – e.g., the Homeobox clusters (***Ferrier, 2016***). The relationship between these patterns of genome evolution and multicellularity is, however, unclear: these traits are not exclusive of animals (*cf.* (***Curtis et al., 2012***; ***Shoguchi et al., 2013***; ***Michael, 2014***; ***de Mendoza et al., 2015***; ***French-Italian Public Consortium for Grapevine Genome Characterization et al., 2007***)); the existence of secondarily reduced genomes in animals (smaller, gene-compact, less repetitive) in animals blurs their link with organismal complexity (***Simakov and Kawashima, 2017***; ***Seo et al., 2001***; ***Petrov et al., 1996***); and non-adaptive scenarios can explain the emergence of genomic complexities as a consequence drift-enhancing population-genetic environments (***Lynch and Conery, 2003***; ***Lynch, 2002***, ***Lynch, 2007***). Establishing the timeline of genome architecture evolution in the


A. Phylogeny and genome statistics
Metazoa+
Choanoflagellata
Filozoa
Choano-
flagellata
Filasterea
Holozoa
Ichthyo-
phonida
Ichthyo-
sporea
Teretosporea
Opisthokonta
Dermo-
cystida
Obazoa
Unikonta/
Amorphea
Holomycota
LECA
Metazoa
Fungi
Discicristoidea
Apusomonadida
Amoebozoa
Diaphoratickes+Excavata / Bikonta
B. Phenotypic traits
Choanoflagellata
Single-celled flagellates, colonial forms (*Salpingoeca*)
Filasterea
Single-celled filopodiated amoebas, aggregative stage (*Capsaspora*)
Ichthyosporea
Ichthyophonida
Coenocytic clonal growth, schizogony and propagule dispersal (by amoebas, flagellated or not)
Dermocystida
Coenocytic clonal growth, flagellated amoeboid dispersal stage
Corallochytrea
Colonies with palintomic division, amoeboid dispersal stage (cryptic flagellum hypothesized)
?

| | Source | Abreviation | Genome size (Mb) | # scaffolds ≥500bp | L75 | N50 (kb) | # genes | % BUSCO orthologs | % GC |
|---|---|---|---|---|---|---|---|---|---|
| *Monosiga brevicollis* | † | Mbre | 41.6 | 218 | 27 | 1,073.6 | 9,172 | 78.6% | 54.9 |
| *Salpingoeca rosetta* | † | Sros | 55.4 | 154 | 25 | 1,519.5 | 11,624 | 80.6% | 56.0 |
| *Capsaspora owczarzaki* | † | Cowc | 27.9 | 84 | 11 | 1,617.7 | 8,741 | 86.6% | 53.8 |
| *Ministeria vibrans* | ‡ | Mvib | - | - | - | - | - | - | - |
| *Creolimax fragrantissima* | * | Cfra | 42.9 | 83 | 17 | 1,585.0 | 8,694 | 86.7% | 40.5 |
| *Sphaeroforma arctica* | * | Sarc | 120.9 | 15,618 | 1,442 | 116.1 | 18,319 | 77.15% | 42.0 |
| *Ichthyophonus hoferi* | * | Ihof | 88.1 | 1,633 | 515 | 106.7 | 6,351 | 63.2% | 33.4 |
| *Amoebidium parasiticum* | ‡ | Apar | - | - | - | - | - | - | - |
| *Abeoforma whisleri* | ** | Awhi | 101.9 | 51,561 | 25,133 | 2.4 | 17,283 | 11.9% | 30.24 |
| *Pirum gemmata* | ** | Pgem | 84.4 | 50,415 | 25,440 | 1.9 | 21,835 | 17.0% | 29.11 |
| *Chromosphaera perkinsii* | ** | Cper | 34.6 | 3,994 | 187 | 120.2 | 12,463 | 85.5% | 42.13 |
| *Sphaerothecum destruens* | ‡ | Sdes | - | - | - | - | - | - | - |
| *Corallochytrium limacisporum* | ** | Clim | 24.1 | 287 | 86 | 180.5 | 7,535 | 83.4% | 51.2 |

**Figure 1.** Evolutionary framework and genome statistics of the study. (A) Schematic phylogenetic tree of eukaryotes, with a focus on the Holozoa. The adjacent table summarizes genome assembly/annotation statistics. Data sources: red asterisks denote Teretosporea genomes reported here; double asterisks denote organisms sequenced for this study; † previously sequenced genomes (**King et al., 2008**; **Fairclough et al., 2013**; **Suga et al., 2013**); ‡ organisms for which transcriptomic data exists but no genome is available (**Torruella et al., 2015**). (B) Overview of the phenotypic traits of each group of unicellular Holozoa, focusing on their multicellular-like characteristics. For further details, see (**Torruella et al., 2015**; **Mendoza et al., 2002**; **Marshall et al., 2008**; **Glockling et al., 2013**). ***Figure 1—source data 1*** and **2**.

DOI: 10.7554/eLife.26036.003

The following source data and figure supplement are available for figure 1:

**Source data 1.** Table of genome structure statistics, from the data-set of eukaryotic genomes used in the study.

DOI: 10.7554/eLife.26036.004

**Source data 2.** Rates of gain and loss of orthogroups for extant and ancestral eukaryotes, using a phylogenetic birth-and-death probabilistic model that accounts for gains, losses and duplications.

DOI: 10.7554/eLife.26036.005

**Figure supplement 1.** Comparisons of gene length of one-to-one orthologs from pair-wise comparisons of all 10 unicellular Holozoa.

DOI: 10.7554/eLife.26036.006

ancestry of Metazoa is thus essential to understand to which extent genomic complexity is linked to multicellularity.

Overall, gene content has been extensively studied in the unicellular ancestry of animals, but less attention has been devoted to the evolution of genome architecture in this period – covering features such as the repetitive content, intron creation and synteny conservation (although *cf.* (**King et al., 2008**; **Irimia et al., 2012**)). This bias is partly due to the multi-million year gap separating animals from their unicellular relatives and the limited genome sampling of unicellular holozoans. We now know several examples of the effects of such limitations. For instance, our view of the transcription factor repertoire of the animal ancestor was confounded by the gene losses of *Monosiga*, which only became evident when *Capsaspora* genome was analyzed (**SebeSebé-PedrosPedrós et al., 2011**); and the same happened with the ancestral animal diversity of cadherin and integrin adhesion systems before genomes from choanoflagellates and *Capsaspora* were analyzed (**Nichols et al., 2012**; **Sebé-Pedrós et al., 2010**). Therefore, comparative genomics studies are highly sensitive to taxonomic biases, meaning that rare genomic changes can remain elusive, and more frequent events can manifest saturated evolutionary signals. To overcome these limitations, we analyze the genomes of the third lineage of close unicellular relatives of animals, the Teretosporea, composed of Ichthyosporea and *Corallochytrium limacisporum* (**Torruella et al., 2015**).

As the earliest-branching holozoan clade, Teretosporea are in a key phylogenetic position to complement our current view of premetazoan evolution. Interestingly, they display a developmental mode that radically differs from choanoflagellates and filastereans: many ichthyosporeans have a multinucleate coenocytic stage (**Mendoza et al., 2002**; **Marshall et al., 2008**), and *Corallochytrium* develops colonies by binary, palintomic, cell division (**Raghukumar, 1987**). In both cases, completion of the life cycle frequently involves release of propagules that restart the clonal proliferation (**Mendoza et al., 2002**; **Marshall et al., 2008**). In addition, the ichthyosporean *Creolimax fragrantissima* exhibits many features reminiscent of animals, such as transcriptional regulation of cell type differentiation or synchronized nuclei division during its development (**de Mendoza et al., 2015**; **Suga and Ruiz-Trillo, 2013**).

Here, we present the complete genomes of four newly sequenced organisms: *Corallochytrium limacisporum* and the ichthyosporeans *Chromosphaera perkinsii* (gen. nov., sp. nov.), *Pirum gemmata* and *Abeoforma whisleri.* These are added to the already available *Creolimax fragrantissima, Ichthyophonus hoferi* and *Sphaeroforma arctica* (**de Mendoza et al., 2015**; **Torruella et al., 2015**) (Ichthyosporea), and to the afore-mentioned *Salpingoeca rosetta*, *Monosiga brevicollis* (choanoflagellates) and *Capsaspora owczarzaki* (Filasterea), totaling 10 unicellular holozoan genomes (**Figure 1**).

Our aim is to provide new insights into the evolutionary dynamics of the genome in the ancestral unicellular lineage leading to animals, at two broad levels: gene family origin and diversification, and conservation of genome architectural features. We address the origin of the large and intron-rich animal genomes, changes in gene linkage (microsynteny), and ancient patterns of gene family diversification. The leitmotiv of these analyses is to identify and date genomic novelties along the ancestry of Metazoa, aiming to understand the foundations of the transition to multicellularity. The emerging picture from this comparative study is one of punctuated, differently-timed bursts of innovation in genome content and structure, occurring in the unicellular ancestry of animals.

# Results

## Four new genomes of unicellular relatives of animals

We obtained the complete nuclear genome sequences of *Corallochytrium limacisporum* and the ichthyosporeans *Chromosphaera perkinsii*, *Pirum gemmata* and *Abeoforma whisleri.* For all these taxa, we sequenced genomic DNA from axenic cultures using Illumina paired-end and mate-pair reads, which were assembled using Spades (**Nurk et al., 2013**). Gene annotation was performed using a combination of de novo gene predictions and transcriptomic evidence derived from RNA sequencing experiments (see Methods). Of the four genomes presented here, *Corallochytrium* (24.1 Mb) and *Chromosphaera* (34.6 Mb) have the highest completeness and contiguity (**Figure 1**). Specifically, *Corallochytrium* has 7535 genes and 83.4% of the BUSCO paneukaryotic gene set (a proxy to genome completeness (**Simão et al., 2015**)), and 75% of the assembly length is covered by 86 scaffolds (L75 statistic). *Chromosphaera* has 12,463 annotated genes comprising 85.5% of the BUSCO set, and its L75 statistic is 187 scaffolds. In contrast, *Abeoforma* and *Pirum* have larger genome assemblies (101.9 and 84.4 Mb), but these are fragmented (L75 = 25,133 and 25,440 scaffolds) and incomplete (11.9% and 17.0% of BUSCO). These lower contiguities are reflected in their partial gene predictions (**Figure 1A**, **Figure 1—figure supplement 1**), which consequently hindered the detection of BUSCO orthologs.

Overall, together with *Capsaspora*, the two choanoflagellates and three already available ichthyosporeans, our expanded dataset now comprises 10 genomes from all unicellular Holozoa lineages – eight more than in previous genome analyses (**Fairclough et al., 2013**; **Suga et al., 2013**).

## The new *Chromosphaera* (gen. nov.) helps resolve the phylogeny of Holozoa

To have a robust phylogenetic framework for our comparative analyses, we investigated the phylogenetic relationships between holozoans with a phylogenomic analysis based on the dataset developed in **Torruella et al. (2015)**. We classified the newly identified *Chromosphaera perkinsii* (gen. nov., sp. nov.) as a member of Ichthyosporea, in the order Dermocystida, as it clusters with *Sphaerothecum destruens* in our phylogenomic analysis (**Figure 2**; BS = 100%, BPP = 1). Therefore, *Chromosphaera,* isolated from shallow marine sediments in Hawai'i, is the first described putatively free-living dermocystid Ichthyosporea. Indeed, all described dermocystids are strict vertebrate parasites,

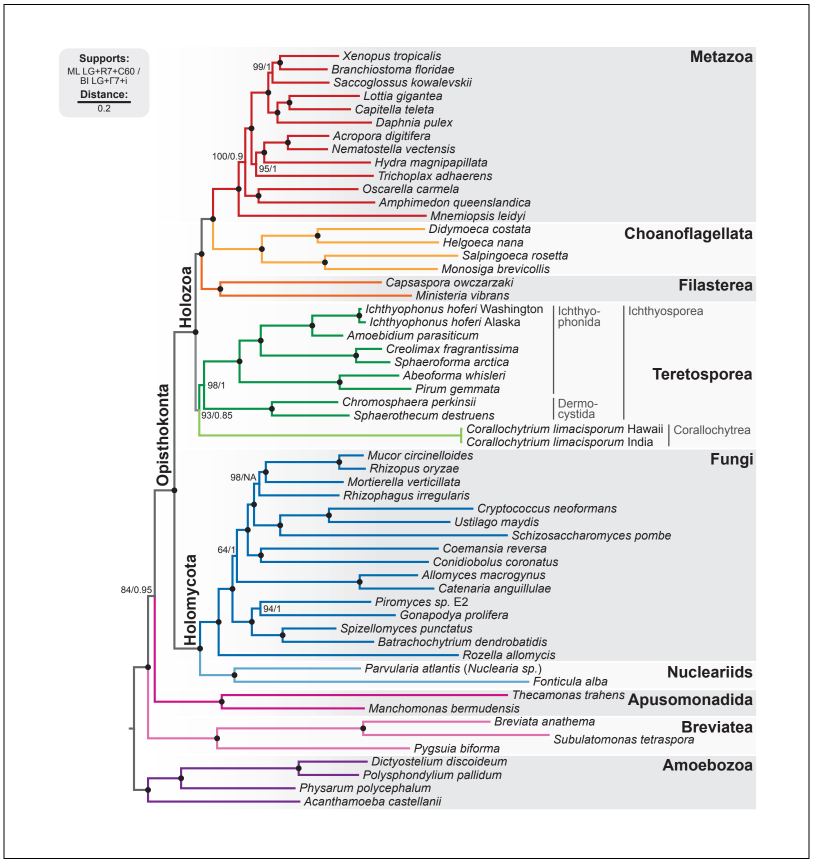


**Figure 2.** Phylogenomic tree of Unikonta/Amorphea. Phylogenomic analysis of the BVD57 taxa matrix. Tree topology is the consensus of two Markov chain Monte Carlo chains run for 1231 generations, saving every 20 trees and after a burn-in of 32%. Statistical supports are indicated at each node: (i) non-parametric maximum likelihood ultrafast-bootstrap (UFBS) values obtained from 1000 replicates using IQ-TREE and the LG + R7+C60 model; (ii) 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 by a black bullet. See ***Figure 2—figure supplement 1*** for raw trees with complete statistical supports. ***Figure 2—source data 2***.

DOI: 10.7554/eLife.26036.007

The following source data and figure supplement are available for figure 2:

**Source data 1.** List of genome and transcriptome assemblies and annotations, including abbreviations, taxonomic classification and data sources.

DOI: 10.7554/eLife.26036.008

**Source data 2.** BVD57 phylogenomic dataset (see (***Torruella et al., 2015***)), including 87 unaligned protein domains (with PFAM accession number) per species.

DOI: 10.7554/eLife.26036.009

**Figure supplement 1.** Phylogenomic analysis of the BVD57 matrix using (**A**) IQ-TREE maximum likelihood and the LG + R7+C60 model (supports are SH-like approximate likelihood ratio test/UFBS, respectively); (**B**) IQ-TREE maximum likelihood and the LG + R7+PMSF model (fast CAT approximation; non-parametric bootstrap supports); and (**C**) Phylobayes Bayesian inference under the LG+Γ7 + CAT model (BPP supports).

DOI: 10.7554/eLife.26036.010

whereas ichthyophonids are typical animal commensals or parasites (although free-living species have been described and some have been identified in environmental surveys of marine microbial eukaryotic diversity) (***del Campo and Ruiz-Trillo, 2013***; ***Glockling et al., 2013***).

Our analysis confirms our previous results with regards to the phylogenetic relationships within Holozoa: the Teretosporea, comprising Ichthyosporea and the small free-living osmotroph *Corallochytrium* (***Raghukumar, 1987***), are a sister-group to all the other holozoans (filastereans, choanoflagellates and animals) with improved statistical support (***Figure 2***; BS = 93%, BPP = 0.85). The monophyly of Teretosporea rejects alternative scenarios such as the 'Filasporea' hypothesis (a grouping of Filasterea + Ichthyosporea) (**Ruiz-Trillo et al., 2008**; **Liu et al., 2009**) or the status of *Corallochytrium* as an independent opisthokont lineage.

## Trends in the evolution of genome size, synteny and gene conservation across Holozoa

### Independent increases in genome size in Metazoa and unicellular holozoans

We found that Metazoa typically have larger genomes than their unicellular relatives: early-branching animals are within the 300–500 Mb range (**Elliott and Gregory, 2015a**; **Simakov and Kawashima, 2017**) and most unicellular holozoans have relatively compact genomes, like *Corallochytrium*, *Capsaspora* or *Chromosphaera* (24.1, 27.9 and 34.6 Mb, respectively; **Figure 3A**). There are, however, a few exceptions in the Ichthyosporea: *Sphaeroforma*, *Abeoforma*, *Pirum* and *Ichthyophonus* have genomes in the 84.4–120.9 Mb range (using assembly length as a proxy to genome size), sometimes larger than some secondarily simplified early-branching animals like *Trichoplax adhaerens* (~100 Mb) or *Oscarella carmela* (57 Mb; **Figure 3A**) (**Srivastava et al., 2008**; **Simakov and Kawashima, 2017**).

A parsimonious scenario for genome size evolution would imply an holozoan ancestor with a fairly compact genome, in line with the values of *Corallochytrium*, *Capsaspora* and *Chromosphaera* (24.1–34.6 Mb), followed by secondary genome expansions in ichthyosporeans (the stem lineage of ichthyophonids, and then again in individual species) and possibly *Salpingoeca* (55.4 Mb). The largest unicellular holozoan assembled genomes fall short of the inferred C-values of ancestral Metazoa (~300 Mb) (**Simakov and Kawashima, 2017**), thus indicating another genome expansion at the origin of multicellularity.

Transposable element (TE) invasions partially explain the inflations in genome size and can carry the signal of the independent expansions (**Elliott and Gregory, 2015b**). Indeed, 5–9% of the genome of *Salpingoeca, Sphaeroforma*, *Abeoforma* and *Pirum* are covered by TEs, whereas other holozoans are below 2.5% (**Figure 3A**). Unicellular holozoan have diverse TE complements, ranging between 42 families in *Corallochytrium* to >400 in *Pirum* or *Abeoforma* (**Figure 3—figure supplement 1**; [**Carr and Suga, 2014**]); and ~31% of these families are shared with metazoan genomes (**Figure 3—figure supplement 2**). In *Salpingoeca, Pirum* and *Abeoforma*, we found species-specific small sets of TE families, sharing high sequence identity, that accounted for the vast majority of copies (**Figure 3B**). This signaled recent TE invasions, and, therefore, independent contributions to genome expansion. There were hints of older TE propagation events in *Sphaeroforma* and *Pirum*, with a long tail of low-similarity TE copies (**Figure 3B**). In *Abeoforma* and *Pirum*, TEs and other simple repeats comprised up to 17% and 34% of the genome, accompanied by unusually AT-biased nucleotide compositions (**Figure 1A**). However, the exact repetitive fraction of *Abeoforma* and *Pirum* genomes cannot be exactly quantified: their highly repetitive nature has contributed to their fragmented and incomplete assemblies (**Figure 1A**, **Figure 1—figure supplement 1**) (**Treangen and Salzberg, 2011**), which hinders the annotation of TEs and simple repeats. Finally, the smaller genomes of *Corallochytrium* and *Chromosphaera* were largely depleted of repetitive/satellite regions and TEs (1.8% and 3.8% of their genomes). This finding, together with their reduced intron content (see below, **Figure 4**) suggests a secondary streamlining process.

### Synteny conservation across holozoan lineages is rare, except in *Capsaspora*

Ancestral conservation of gene linkage at the local level (microsynteny) is common in Metazoa, frequently due to coordinated *cis*-regulation (**Irimia et al., 2012**; **Simakov et al., 2013**). Following this reasoning, we analyzed the microsyntenic gene pairs of unicellular holozoan genomes (**Figure 3C**), expecting higher degrees of conservation within lineages than across them. This hypothesis held true for the *Salpingoeca-Monosiga* genome pair, but we found little or no conservation in almost all inter-specific comparisons of ichthyosporeans and *Corallochytrium*. There were, however, two exceptions: *Creolimax-Sphaeroforma* (sibling species; 907 syntenic orthologous genes) and, to a lesser extent, *Chromosphaera-Corallochytrium* (72 genes). In the case of the closely-related *Pirum* and *Abeoforma,* their fragmented genomes hindered the gene order analyses and yielded low synteny conservation values.

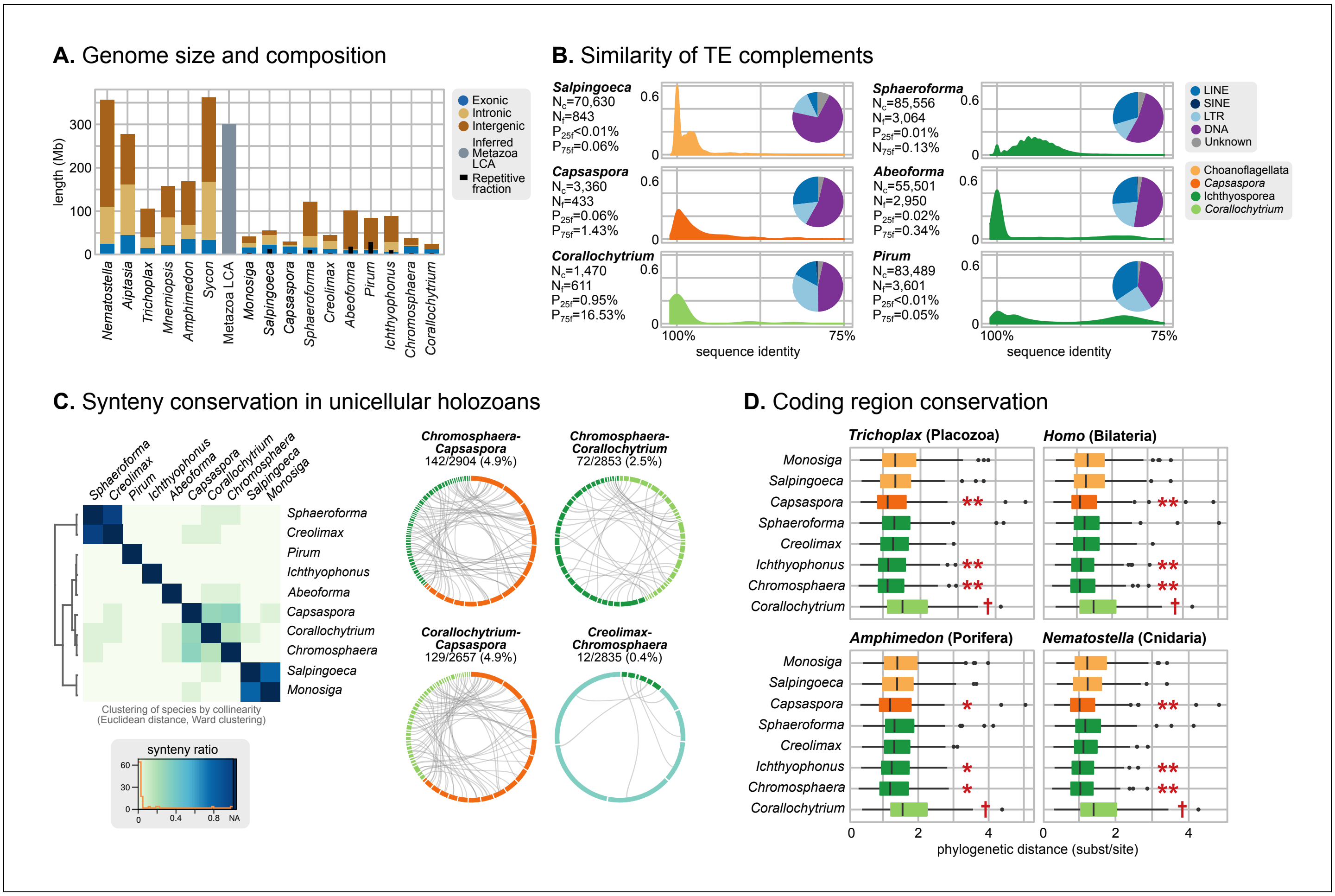


**Figure 3.** Patterns of genome evolution across unicellular Holozoa. (A) Genome size and composition in terms of coding exonic, intronic and intergenic sequences of unicellular holozoan and selected metazoans. Percentage of repetitive sequences shown as black bars. Genome size of the Metazoa LCA (gray bar) from (***Simakov and Kawashima, 2017***) (exonic, intronic and intergenic composition not known). (B) Profile of TE composition for selected organisms. Density plots indicate the sequence similarity profile of the TE complement in each organism. Embedded pie-charts denote the relative abundance, in nucleotides, of the main TE superclasses in each genome: retrotransposons (SINE, LINE and LTR), DNA transposons (DNA) and unknown. $N_c$: total number TE copies in the genome; $N_f$: number of families to which these belong; $P_{25f}$ and $P_{75f}$: percentage of most-frequent TE families that account for 25% and 75% of the total number of TE copies, respectively. (C) Heatmap of pairwise microsynteny conservation between 10 unicellular holozoan genomes. Species ordered according the number of shared syntenic genes (Euclidean distances, Ward clustering). At the right: selected pairwise comparisons of syntenic single-copy orthologs between unicellular holozoan genomes. Numbers denote number of syntenic genes, total number of single-copy orthologs, and proportions (%) of syntenic genes per the compared orthologs. Circle segments are scaffolds sharing ortholog pairs, connected by gray lines. (D) Phylogenetic distances between unicellular holozoans and four selected animals: *Homo sapiens*, *Nematostella vectensis*, *Trichoplax adhaerens* and *Amphimedon queenslandica*. Red asterisks denote organisms that have lower phylogenetic distances to metazoans than one (single asterisk) or both choanoflagellates (double asterisks) ($p$ value < 0.05 in Wilcoxon rank sum test). † indicates significantly higher distances between *Corallochytrium* and metazoans. ***Figure 1—source data 1***, ***Figure 3—source data 1***, ***2*** and ***3***.

DOI: 10.7554/eLife.26036.011

The following source data and figure supplements are available for figure 3:

**Source data 1.** Annotated repetitive sequences from 10 unicellular Holozoa genomes.

DOI: 10.7554/eLife.26036.012

**Source data 2.** List of annotated transposable element families in 10 unicellular Holozoa genomes, with copy counts.

DOI: 10.7554/eLife.26036.013

**Source data 3.** List of annotated transposable element families shared between the genomes of 10 unicellular holozoans and 11 animals, including the number of species where the TE family is present.

DOI: 10.7554/eLife.26036.014

**Figure supplement 1.** Profile of TE composition of unicellular Holozoa.

DOI: 10.7554/eLife.26036.015

**Figure supplement 2.** Shared TEs between unicellular Holozoa and animal genomes.

*Figure 3 continued on next page*

*Figure 3 continued*

DOI: 10.7554/eLife.26036.016

**Figure supplement 3.** Heatmap of pairwise ratios of ortholog collinearity between 10 unicellular holozoan genomes.

DOI: 10.7554/eLife.26036.017

In contrast, the analysis of microsynteny in *Capsaspora* revealed remarkable across-lineage conservation with the distant teretosporeans *Chromosphaera* and *Corallochytrium* (142 and 129 genes, respectively). Moreover, and to a lesser degree, *Capsaspora* also retains a few shared linked gene pairs with *Trichoplax*, the cnidarians *Aiptasia* sp., *Nematostella vectensis*, and the sponges *Amphimedon queenslandica* and *Oscarella carmela* (***Figure 3C***, ***Figure 3—figure supplement 3***). A notable example of ancestral microsynteny is that of integrins: heterodimeric transmembrane proteins involved in cell-to-matrix adhesion and signaling in animals that are also present in unicellular Holozoa (***de Mendoza et al., 2015***; ***Sebé-Pedrós et al., 2010***). Indeed, integrin-α and integrin-*β* genes from *Corallochytrium* (one pair) and *Capsaspora* (four pairs) are in a conserved head-to-head arrangement of likely holozoan origin. Incidentally, *Capsaspora*'s pairs of collinear α/*β* integrins co-express during its life cycle (***Sebé-Pedrós et al., 2013***), a typical cause of microsynteny conservation in animals (***Irimia et al., 2012***). Overall, gene linkage of most extant holozoans appears to be markedly different from their common ancestor, with specific gene pairings arising in Metazoa (***Irimia et al., 2012***; ***Simakov et al., 2013***), choanoflagellates and some ichthyophonids. In contrast, *Capsaspora* harbors a relatively slow-evolving genome in terms of synteny conservation.

## Coding sequence conservation patterns vary across holozoan lineages

Finally, we examined the level of coding sequence conservation between unicellular holozoans and animals. We aimed to contrast the patterns of conservation at the structural level (outlined above) with those of the genic regions. Using 143 phylogenies of paneukaryotic orthologous genes, we examined the pairwise distances between unicellular holozoans and *Homo sapiens* (bilaterian), *Amphimedon* (sponge), *Nematostella* (sea anemone) and *Trichoplax* (placozoan) (***Figure 3D***). In all comparisons, *Capsaspora*, *Chromosphaera* and *Ichthyophonus* accumulated fewer amino-acidic substitutions per alignment position than choanoflagellates since their divergence from animals ($p<0.05$ in Wilcoxon rank sum test). Conversely, *Corallochytrium* was singled out as the taxon with more cumulative amino acid differences with animals. Thus, the analysis of coding sequence conservation across holozoans—a genomic trait fundamentally unrelated to synteny—also attests to *Capsaspora*'s slower pace of genome change.

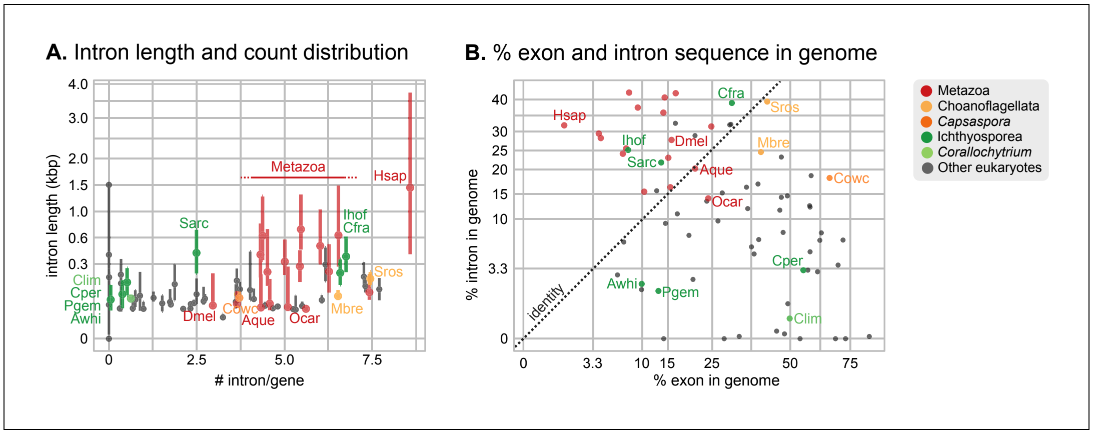


**Figure 4.** Intron abundance in eukaryotes. (A) Distribution of intron lengths and number of introns per gene in selected eukaryote genomes. Dots represent median intron lengths and vertical lines delimit the first and third quartiles. Color code denotes taxonomic assignment. Species abbreviation as in ***Figure 1*** and ***Figure 2—source data 1***. (B) Fraction of the genome covered by introns and exons in selected eukaryotes. Dotted line represents the identity between both values. Color code denotes taxonomic assignment. ***Figure 1—source data 1***.

DOI: 10.7554/eLife.26036.018

## Intron evolution in Holozoa: two independent 'great intronization events'

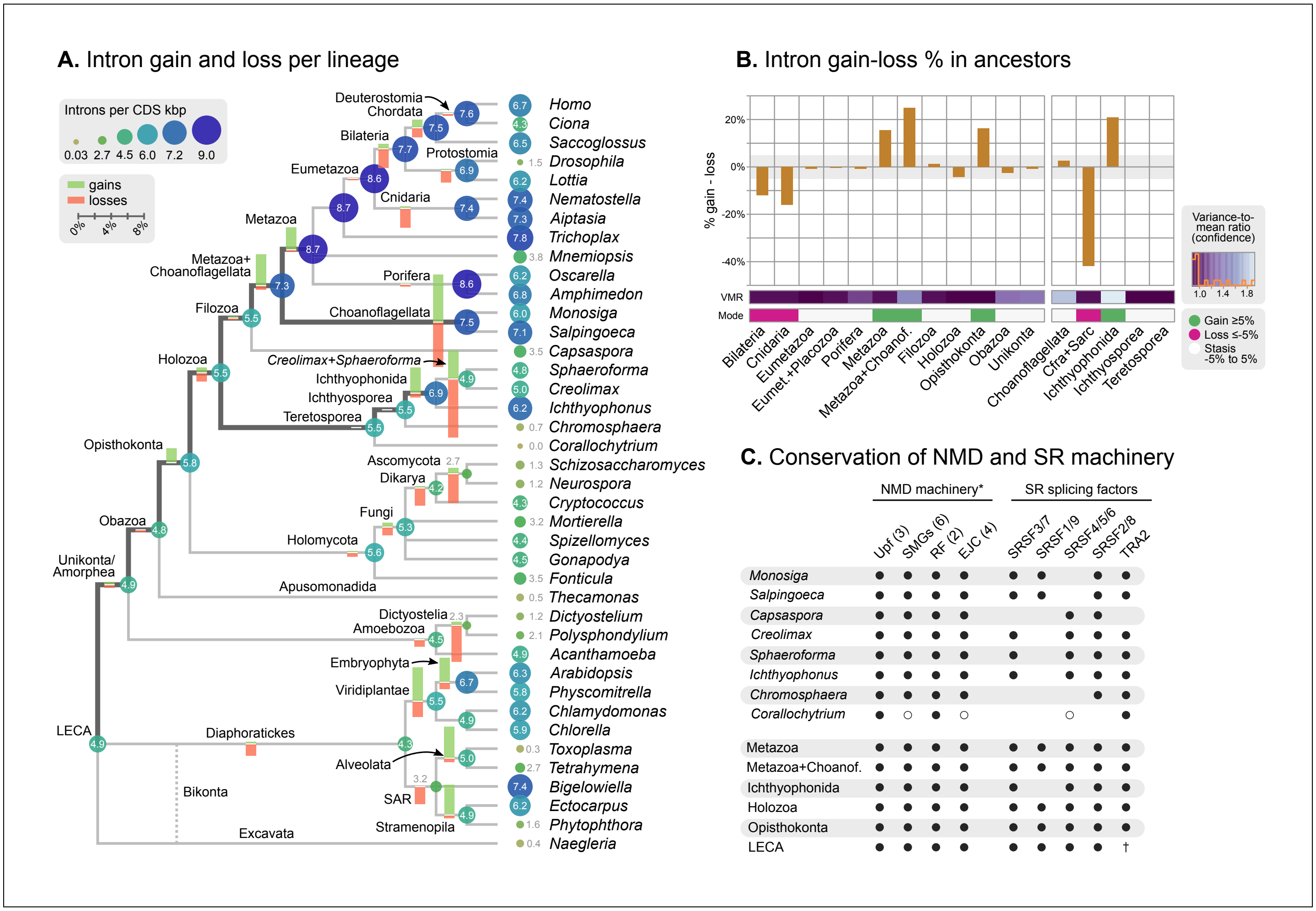
### Evolution of intron structure

Intron-rich genomes are a hallmark of Metazoa. Indeed, the last common ancestor (LCA) of Metazoa is inferred to have had the highest intron density among eukaryotes, due to a process of continuous intron gain starting in the last eukaryotic common ancestor (LECA) (**Csuros et al., 2011**; **Carmel et al., 2007**). The high intron density of multicellular animals has been linked to their higher organismal complexity, as it enables frequent alternative splicing (AS) and richer transcriptomes (**Rogozin et al., 2012**; **Barbosa-Morais et al., 2012**; **Irimia et al., 2009**; **Nilsen and Graveley, 2010**), provides physical space for transcription regulatory sites (**Le Hir et al., 2003**; **Sebé-Pedrós et al., 2016b**), and facilitates the diversification of gene families by exon shuffling (**Liu et al., 2005**). The dominance of weak splice sites inferred at the intron-rich ancestral Metazoa reinforces the proposed role of alternative splicing as an important source of transcriptomic innovation at the dawn of animal multicellularity (**Csuros et al., 2011**; **Irimia et al., 2007**).

Our expanded set of unicellular holozoan genomes provides an ideal framework to investigate the emergence of the high intron densities found in animal genomes. Our survey of intron richness across eukaryotes identifies a high number of introns per gene in many ichthyosporeans, choanoflagellates and animals (**Figure 4A**). Moreover, *Creolimax* and *Ichthyophonus* harbor longer introns than most protistan eukaryotes, similar in length to those of some animals (**Figure 4B**). These similarities between ichthyosporeans and animals suggest two possible scenarios: (1) an early intronization event at the origin of Holozoa followed by reduction in some unicellular lineages (e.g., *Capsaspora* or *Corallochytrium*); or (2) independent episodes of intron proliferation in Metazoa, Choanoflagellata and Ichthyosporea. To test these hypotheses, we assembled a set of 342 paneukaryotic orthologs from 40 complete genomes and analyzed the conservation of their intron sites according to the maximum likelihood method developed by **Csűrös and Miklós (2006)** (**Figure 5—figure supplement 1**). This analysis supports the second hypothesis and reveals two independent periods of intense intron gain in unicellular holozoans: at LCA of Metazoa and Choanoflagellata, and in the branch leading to ichthyophonid Ichthyosporea (**Figure 5A–B**). After animals and choanoflagellates diverged, intron gains independently persisted in both lineages.

Our reconstruction shows that, since the origin of introns in the LECA, most ancestors were dominated by intron loss while a few remain in an equilibrium, static or dynamic (consistent with previous studies [**Csuros et al., 2011**; **Rogozin et al., 2012**]) (**Figure 5B**). A prolonged process of intron gain can be observed, however, in the lines of descent from the LECA (4.9–5.5 introns per kbp of coding sequence) to Ichthyophonida (6.9 introns/CDS kbp) and Metazoa LCAs (8.7 introns/CDS kbp), interrupted by phases of stasis with slight intron loss, such as in the Filozoa or Holozoa LCAs (**Figure 5A–B**).

The existence of independent intronization events in ancestral holozoans is supported by a hierarchical clustering analysis of the intron presence/absence profile across extant and ancestral genomes (**Figure 6A**; Ward clustering from Spearman correlation-based distances). First, most intron-rich animals form a cluster with *Salpingoeca* and *Monosiga* that also includes the LCAs of Metazoa and Metazoa + Choanoflagellata. Second, ichthyosporeans and *Corallochytrium*, although phylogenetically closely-related to each other, are highly divergent in their pattern of shared introns: the intron-dense *Creolimax* and *Sphaeroforma* form an independent cluster that differs from the Holozoa LCA; whereas *Corallochytrium* and *Chromosphaera* undergo independent secondary simplifications (from 5.5 introns/CDS kbp in the Teretosporea LCA, to 0.0 and 0.7, respectively). In contrast, *Ichthyophonus* (intron-rich) and *Capsaspora* have lower intron loss rates and are more similar to older eukaryotic ancestors, from Holozoa to the LECA (**Figure 6A**). In *Ichthyophonus*, retention is accompanied by a high gain rate, giving intron densities similar to some modern animals (7.1 intron/CDS kbp). In contrast, *Capsaspora* (3.5 intron/CDS kbp) appears to have undergone little ancestral reconfiguration of its gene architecture: there is an equilibrium between few losses and gains at the root of Filozoa (**Figure 5A**), and 85.5% of its introns are of holozoan or earlier origin (**Figure 6B**). Interestingly, introns with regulatory sites from *Capsaspora* (identified in [**Sebé-Pedrós et al., 2016b**]) have a similar, ancestral-biased, age distribution (Fisher's exact test, p-value=1; **Figure 6B**). This hints at a decoupling between the evolutionary dynamics of introns and regulatory sites, despite sharing physical space in the genome.



|  | NMD machinery* |  |  |  | SR splicing factors |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  | Upf (3) | SMGs (6) | RF (2) | EJC (4) | SRSF3/7 | SRSF1/9 | SRSF4/5/6 | SRSF2/8 | TRA2 |
| *Monosiga* | ● | ● | ● | ● | ● | ● |  | ● | ● |
| *Salpingoeca* | ● | ● | ● | ● | ● | ● |  | ● | ● |
| *Capsaspora* | ● | ● | ● | ● |  |  | ● | ● |  |
| *Creolimax* | ● | ● | ● | ● | ● |  | ● | ● | ● |
| *Sphaeroforma* | ● | ● | ● | ● | ● |  | ● | ● | ● |
| *Ichthyophonus* | ● | ● | ● | ● | ● |  | ● | ● | ● |
| *Chromosphaera* | ● | ● | ● | ● |  |  |  | ● | ● |
| *Corallochytrium* | ● | ○ | ● | ○ |  |  | ○ |  | ● |
| Metazoa | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| Metazoa+Choanof. | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| Ichthyophonida | ● | ● | ● | ● | ● |  | ● | ● | ● |
| Holozoa | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| Opisthokonta | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| LECA | ● | ● | ● | ● | ● | ● | ● | ● | † |

**Figure 5.** Intron evolution. (A) Rates of intron gain and loss per lineage, including extant genomes and ancestral reconstructed nodes. Diameter and color of circles denote the number of introns per kbp of coding sequence at each ancestral node. Bolder edges mark the lines of descent between the LECA and Metazoa/Ichthyophonida, which were characterized by continued high intron densities (see text). Red and green bars represent the inferred number of intron gains (green) and losses (red) in ancestral nodes. (B) Difference between intron site gains and losses in selected ancestors, including animals (left; from Metazoa to Unikonta/Amorphea) and unicellular holozoans (right). For each ancestor, we specify the variance-to-mean ratio of the inferred number of introns from 100 bootstrap replicates (higher values, denoted by lighter purple, indicate less reliable inferences; see Methods. The color code denotes modes of intron evolution: dominance of gains (green), losses (pink) and stasis (light gray). (C) Conservation of the NMD machinery and SR splicing factors in unicellular holozoans (up) and selected ancestors (down). Black dots indicate the presence of an ortholog, and empty dots partial conservation. For the NMD machinery, each column summarizes the presence of multiple gene families (number between brackets). † denotes the ancestral eukaryotic origin of TRA2 according to (***Plass et al., 2008***). Complete survey at the species and gene levels available as Figure 4—figure supplements 2 and 3. ***Figure 5—source data 3***.

DOI: 10.7554/eLife.26036.019

The following source data and figure supplements are available for figure 5:

**Source data 1.** Rates of gain and loss of intron sites for extant and ancestral eukaryotes, calculated for a rates-across-sites Markov model for intron evolution with branch-specific gain and loss rates (***Csurös, 2008***).

DOI: 10.7554/eLife.26036.020

**Source data 2.** Reconstruction of intron site evolutionary histories, using a rates-across-sites Markov model for intron evolution, with branch-specific gain and loss rates (***Csurös, 2008***).

DOI: 10.7554/eLife.26036.021

**Source data 3.** Reconstruction of the evolution of the NMD machinery (***He and Jacobson, 2015***) and key SR splicing factors (***Plass et al., 2008***).

DOI: 10.7554/eLife.26036.022

**Figure supplement 1.** Classification of intron sites by conservation in protein alignments, as used in (***Csűrös and Miklós, 2006***; ***Csurös, 2008***).

DOI: 10.7554/eLife.26036.023

**Figure supplement 2.** Phylogenetic distribution of the NMD machinery, SR splicing factors and RNA-binding domains in eukaryotes.

DOI: 10.7554/eLife.26036.024

*Figure 5 continued on next page*

*Figure 5 continued*

**Figure supplement 3.** Phylogenetic analysis of (A) eIF4A3, (B) Smg5/6/7, and (C) eRF3, using Maximum likelihood in IQ-TREE (supports are SH-like approximate likelihood ratio test/UFBS, respectively); including Bayesian inference supports for the ortologous groups of interest (BPP statistical supports, in red).

DOI: 10.7554/eLife.26036.025

## Consequences of intron gains in early holozoan evolution

The evolutionary implications of intron gain episodes in Holozoa remain an open question. High intron densities have been linked to inefficient purifying selection: according to the mutational-hazard hypothesis, the lower effective population sizes of animals preclude the loss of slightly deleterious intronic sequence – which can constitute an impediment to genome replication or precise transcription (***Csuros et al., 2011***; ***Lynch and Conery, 2003***; ***Lynch, 2002***, ***Lynch, 2006***). Whether this population-genetic effect is also connected with the intron gains in *Creolimax*, *Sphaeroforma* and *Ichthyophonus*, however, is unclear: their specific effective population sizes are not known, but estimates from their close relative *Sphaeroforma tapetis* are in line with typical unicellular eukaryotes (in the $10^6$ to $10^7$ range [***Marshall and Berbee, 2010***]) and thus higher than most animals (***Lynch, 2006***).

Alternatively, holozoans' intron gains could be linked to adaptive roles related to alternative splicing (AS): intron-dense genomes exhibit AS-rich transcriptomes (***Irimia and Roy, 2014***), which can

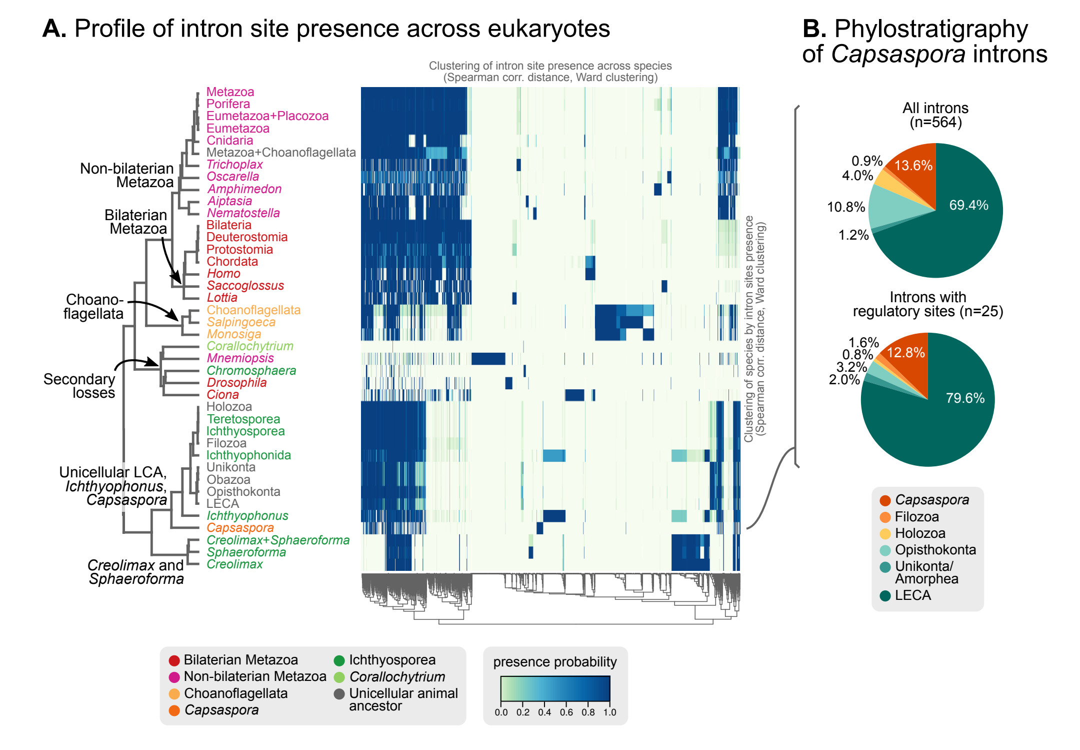


**Figure 6.** Profile of intron site presence across eukaryotes. (A) Heatmap representing presence/absence of 4312 intron sites (columns) from extant and ancestral holozoan genomes, plus the line of ascent to the LECA (rows). Intron sites and genomes have been grouped according to their respective patterns of co-occurrence (dendrogram based on Spearman correlation distances and Ward clustering algorithm; see Methods). The dendrogram of genome clusterings is shown to the left. ***Figure 5—source data 2***. (B) Phylostratigraphic analysis of the origin of *Capsaspora* introns, considering all sites (left) and those with putative regulatory sites (right; after [***Sebé-Pedrós et al., 2016b***]).

DOI: 10.7554/eLife.26036.026

increase proteomic diversity (**Barbosa-Morais et al., 2012**; **Nilsen and Graveley, 2010**; **Bush et al., 2017**) or fine-tune gene expression regulation (**Lareau et al., 2007**; **He and Jacobson, 2015**). Transcriptomes of complex animals frequently feature exon skipping events that conduce to multiple protein isoforms per gene (**McGuire et al., 2008**; **Irimia and Roy, 2014**; **Bush et al., 2017**). In contrast, the AS profiles of *Creolimax* and *Capsaspora* are dominated by intron retention (affecting 24.9% and ~33% of their genes, respectively), which can disrupt the transcripts' open reading frames (**Sebé-Pedrós et al., 2013**; **de Mendoza et al., 2015**). Intron retention is present in virtually all intron-bearing eukaryotes, pointing at an early origin in evolution (**Irimia and Roy, 2014**). Consequently, AS events in the intron-rich *Creolimax* were proposed to be involved in down-regulation of gene expression (**de Mendoza et al., 2015**) by a mechanism akin to the nonsense-mediated decay (NMD) pathway that operates in other eukaryotes (**Lareau et al., 2007**; **He and Jacobson, 2015**; **Braunschweig et al., 2014**; **Kerényi et al., 2008**).

In order to explore the relationship between intron evolution and AS-based transcriptome regulation, we surveyed the conservation in unicellular holozoans of the NMD protein complex and key splicing factors involved in AS (**Figure 5C**, **Figure 5—figure supplement 2** and **3**). The core NMD toolkit (consisting of the Upf1-3, Smg1, Smg5/6/7 and Smg8/9 genes; the release factors 1 and 3; and the exon-junction complex [EJC] [**He and Jacobson, 2015**]) has a pan-eukaryotic distribution (**Figure 5C-Figure 5—figure supplement 2A**), as previously reported for the wider spliceosomal molecular machinery (**Collins and Penny, 2005**). The NMD toolkit was also fully conserved in the LCAs of Ichthyophonida, Metazoa and Metazoa + Choanoflagellata – which underwent the above-reported intron gain episodes (**Figure 5A**). Similarly, the SR splicing factors (serine/arginine-rich proteins, termed SRSF1-9 and TRA2A/B in humans), which are involved in splice site recognition in metazoan AS (**Plass et al., 2008**; **Sanford et al., 2005**), also appeared early in eukaryotic evolution and were conserved in LCAs ranging from Opisthokonta to Metazoa (**Figure 5C**, **Figure 5—figure supplement 2B**). Interestingly, *Corallochytrium* secondarily lost part of its NMD machinery and SR splicing factors (e.g., it lacks three out of four EJC components, and only possesses one canonical SR gene) concomitantly with its acute intron losses – a process that mirrors the depletion of splicing factors in the intron-depleted ascomycete *Saccharomyces cerevisiae* (**Plass et al., 2008**). Thus, we found that the intron gain episodes of the LCAs of ichthyophonids and animals occurred in ancestral holozoans that were potentially able to perform NMD of aberrant transcripts.

## Timing of gene family diversification in holozoa

The *Monosiga* genome paper by **King et al. (2008)** revealed that much of the innovation in gene content seen in the transition to multicellularity is rooted in pervasive 'tinkering' with preexisting gene families, notably by rearrangements of protein domains. This mechanism, combined with gene duplication, allows for a functional diversification of gene families by tuning the interactions with other components of the cell—its substrate specificities, sub-cellular localization or partnerships with other proteins within larger complexes. Albeit protein domain rearrangements are not uncommon in eukaryotes (**Basu et al., 2008**, **Basu et al., 2009**; **Leonard and Richards, 2012**), this process is specifically credited with the diversification of many gene families involved in complex signaling and/or multicellular integrated lifestyle in Metazoa (**Suga et al., 2012**; **Simakov and Kawashima, 2017**; **Sebé-Pedrós et al., 2010**; **Tordai et al., 2005**; **Ekman et al., 2007**; **Hynes, 2012**; **Deshmukh et al., 2010**; **Grau-Bové et al., 2015**).

Here, we present a comprehensive study of gene diversification in Holozoa, using our taxon-rich genomic dataset to reconstruct its effect in the animal ancestry. We thus performed a comparative analysis of protein domain architectures across eukaryotes, using the rates of domain rearrangement (or shuffling) as a proxy for gene family diversification. We compared the phylogenetic distribution of protein domain co-occurrences across species and gene families (using a dataset comprising 26,377 gene families or clusters of orthologs derived from 40 eukaryotic species (see Methods). We inferred rates of domain rearrangement at ancestral nodes of the eukaryotic tree using a probabilistic birth-and-death model (**Csűrös and Miklós, 2006**) to reconstruct the content of specific protein domain architectures in ancestral genomes (available as **Figure 7—source data 2**). In our approach, pairs of domains can create novel combinations ('gain') that diversify existing gene families, or dissociate domains ('loss'), which results in decreased diversity of multi-domain proteins.

## Shuffling of protein domain architectures is common in the holozoan ancestors

We assessed the frequency of protein domain rearrangements by quantifying the rates of domain pair gain and loss at each node of the eukaryotic tree (number of gained or lost domain pairs relative to the total number of pairs in that node) (**Figure 7A–B**). Gains and losses are frequent but unequally distributed across organisms and over time, with a majority of nodes showing a tendency towards destruction or creation of domain combinations. Out of 73 analyzed organisms, 20 show a strong bias towards gains, 32 a bias towards losses (>5% difference in either sense), and 64 show combined rates of gain and loss of >10% (**Figure 7A**). In contrast, the ancestral reconstruction of individual protein domain evolution (based on Dollo parsimony) showed that losses dominate in most nodes, both extant and ancestral – with the exception of animals and their ancestors (**Figure 7—figure supplement 1**) (**Zmasek and Godzik, 2011**).

In this scenario of pervasive domain rearrangements, we identified a consistent pattern of creation of protein domain architectures in the lineage leading to Metazoa – specifically, the line of descent from the opisthokont to the bilaterian LCA (**Figure 7A–B**). This tendency was most acute at three points in animal prehistory: the Holozoa LCA, the Filozoa LCA (*Capsaspora*, animals and choanoflagellates) and the Metazoa LCA. Conversely, unicellular holozoans outside the animal lineage were dominated by secondary simplification (e.g., the LCAs of choanoflagellates or ichthyosporeans, as well as some individual species such as *Sphaeroforma*, *Ichthyophonus* or *Corallochytrium*) or by dynamic stasis (e.g., *Capsaspora*, *Creolimax* or *Chromosphaera*). Our analysis thus shows that the increased diversity of protein organizations in animals has its roots in successive events of domain shuffling during their unicellular holozoan prehistory, even if this period was dominated by a relative stasis in terms of the emergence of new protein domain families (**Figure 7A** and **Figure 7—figure supplement 1**).

Then, we questioned whether these expansions were more frequent in protein domains related to typical multicellular functions, such as the extracellular matrix (ECM), transcription factors (TF) or signaling pathways (**Suga et al., 2013**; **Richter and King, 2013**; **de Mendoza et al., 2013**; **Hynes, 2012**; **de Mendoza et al., 2014**). We found that gene families carrying TF- and ECM-related domains had consistently higher diversification rates not only in Metazoa but also in their unicellular ancestors (**Figure 7B**, right panel; asterisks indicate two-fold differences). We thus identify a continuous process of protein diversity gain involving multicellularity-related genes in animal ancestors ranging from the LCA of Obazoa (Opisthokonta + Apusomonadida) to the LCA of Metazoa.

## A unique mode of transcription factor diversification in premetazoan ancestors

Next, we analyzed the dynamics of the bursts of innovation in protein domain architectures in the unicellular ancestry of Metazoa, particularly regarding TFs and ECM-related genes. Specifically, we examined the degree of protein domain promiscuity across gene families (i.e., whether a specific domain combination is re-used in multiple gene families) in different ancestors, to measure changes in the specificity of protein domain architecture diversity.

We measured domain promiscuity by modeling each proteome as a network graph, where vertices represented protein domains that were linked by edges if they co-occurred in a given gene family (with $\geq$90% probability for the ancestral reconstructions; Methods and **Figure 8**). In this context, highly promiscuous domains would join multiple gene families within the network, whereas gene family-specific domains would form independent clusters. This effect can be investigated by computing the network modularity: a parameter describing the degree of isolation of 'modules' (here, groups of co-occurring domains) within a network given their connections to other 'modules' (**Figure 8C**).

We identified a general tendency for multi-domain protein families to diversify by acquisition of highly promiscuous domains also present in other families. This result was based on two observations. First, network modularities were high in most analyzed genomes (within the 0.7–1 range; consistent with previous observations (**Itoh et al., 2007**; **Xie et al., 2011**)) but they were generally lower in animals than in their unicellular relatives and ancestors (**Figure 8A**). Second, there was a strong negative relationship between modularity and the number of protein domains per gene family (Spearman's rank correlation coefficient, $\rho_s$=$-$0.96, p<0.001, **Figure 8B**). Therefore, at the genome

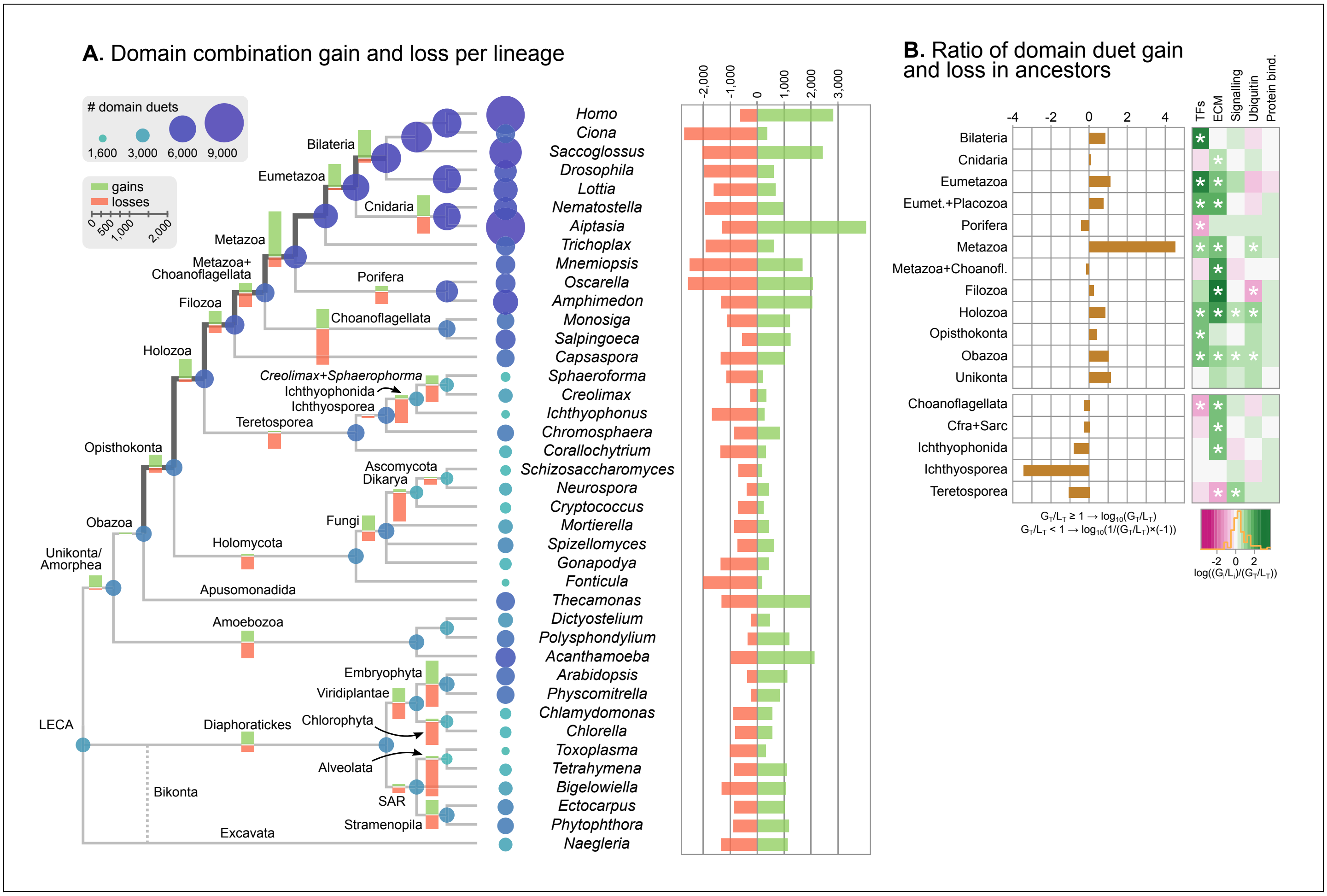


**Figure 7.** Evolution of protein domain architectures. (**A**) Protein domain combination gain and loss per lineage, including extant genomes and ancestral reconstructed nodes. Diameter and color of circles denote the number of different domain combinations (in different gene families) in that node of the tree. Bolder edges mark the line of descent between the LCAs of Opisthokonta and Bilateria, which was generally dominated by gains of protein domain combinations (see text). Red and green bars represent the inferred number of gains and losses, respectively. (**B**) Gain/loss ratio of protein domain diversity in selected ancestors, including animals (upper chart; from Metazoa to Unikonta/Amorphea) and unicellular holozoans (lower). Heatmap to the right represents the log-ratio value of the diversification rate for selected sub-sets of functionally-related protein domains relevant to multicellularity: green indicates higher-than-average diversification; pink less; white asterisks indicate two-fold or more increases or decreases (all comparisons relative to the whole set of protein domains). Source Data ***Figure 7—source data 1*** and ***2***, ***Figure 1—source data 2***.

DOI: 10.7554/eLife.26036.027

The following source data and figure supplement are available for figure 7:

**Source data 1.** Rates of gain and loss of protein domain pairs within a given orthogroup for extant and ancestral eukaryotes, calculated for a phylogenetic birth-and-death probabilistic model that accounts for gains, losses and duplications (***Csurös, 2010***).

DOI: 10.7554/eLife.26036.028

**Source data 2.** Reconstruction of the evolutionary histories of protein domain pairs gains within orthogroups, using a phylogenetic birth-and-death probabilistic model that accounts for gains, losses and duplications (***Csurös, 2010***).

DOI: 10.7554/eLife.26036.029

**Source data 3.** Reconstruction of the evolutionary histories of individual protein domains, using Dollo parsimony and accounting for gains and losses (***Csurös, 2010***).

DOI: 10.7554/eLife.26036.030

**Figure supplement 1.** Gains and losses of individual protein domains across eukaryotes.

DOI: 10.7554/eLife.26036.031

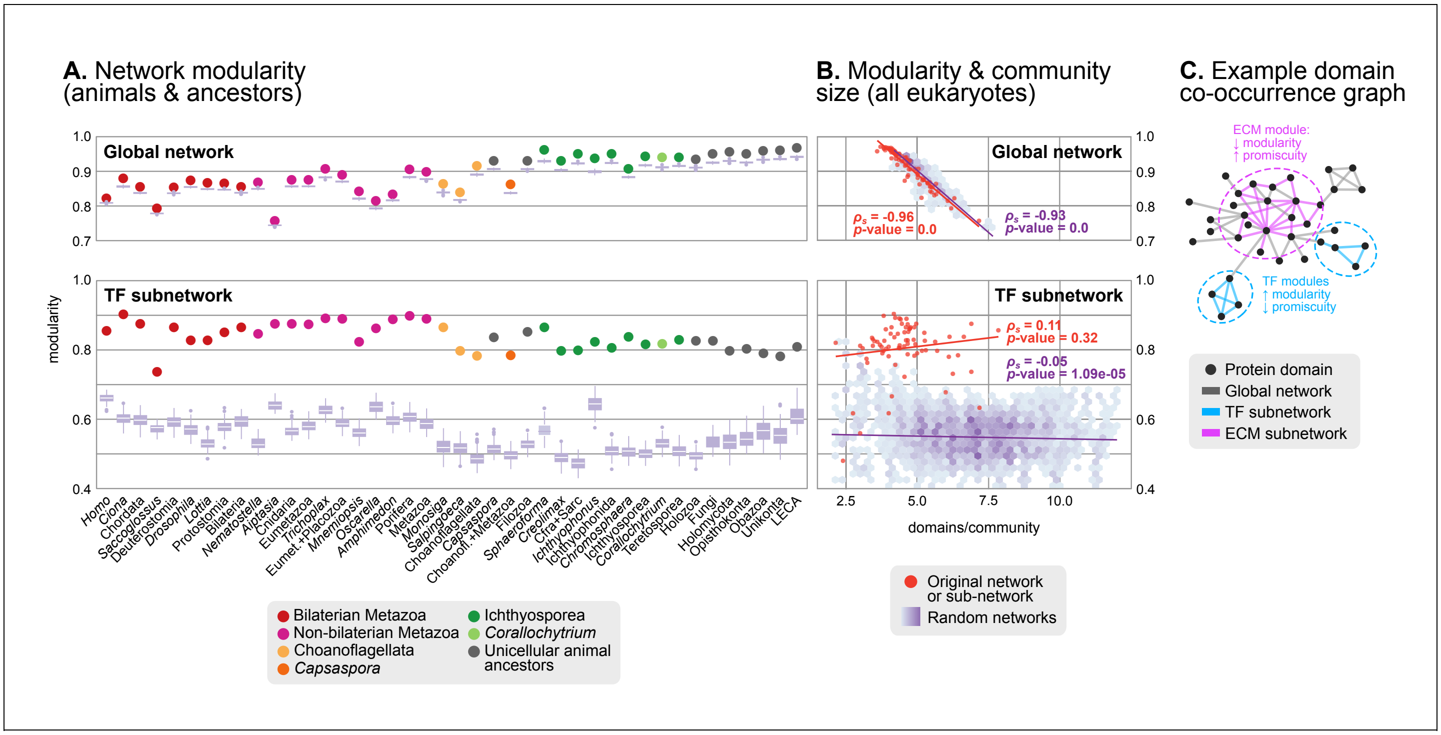


**Figure 8.** Protein domain architecture networks. (**A and B**) Modularity and community size of the global network of domain pairs (upper panels) and the TF subnetwork (lower panels), with ≥90% probability. The modularity parameter measures the fraction of the intra-community edges in the network, minus the expected value in a random network (takes values from 0 to 1; see Materials and methods and [***Newman and Girvan, 2004***]). Panels at the left show the observed modularity of the protein domain (sub)networks of various genomes (Holozoa and selected ancestors; dots are taxa-colored). Purple box plots represent the distribution of simulated modularities from 100 rewirings of the original organism-specific networks, while keeping a constant vertex degree distribution. Panels to the right show the relationship between modularities and the number of domains/community, both for actual genomes (orange) and simulated rewired networks (purple density plot, see Methods). Monotonic dependence between modularity and domains/community was tested for each set of data (global, TF and their respective simulations) using Spearman's rank correlation coefficient ($\rho_s$), and linear regression fits are included for clarity. Note that simulated TF subnetworks are less modular and have more domains/community than the original ones, signaling their higher-than-expected modularities. Note that the scales of the vertical axes change between upper and lower panels. (**C**) Example of protein domain co-occurrence network. Vertices represent domains, linked by edges if they co-occur within the same gene family. Two subnetworks are highlighted in yellow (domain pairs occurring in TF genes) or green (same for signaling genes). ***Figure 7—source data 1*** and ***2***, ***Figure 1—source data 2***, ***Figure 10—source data 1***.

DOI: 10.7554/eLife.26036.032

The following figure supplement is available for figure 8:

**Figure supplement 1.** Mo dularity of protein domain co-occurrence networks of multicellularity-related gene sets across eukaryotes.

DOI: 10.7554/eLife.26036.033

level, gene family diversification tends to reduce modularity due to the use of highly promiscuous protein domains, as it has been frequently reported in animals (***Simakov and Kawashima, 2017***; ***Basu et al., 2008***). This same effect was observed when we analyzed subsets of the proteome networks sharing a common function: the diversification of gene families with domains related to the ECM, signaling, ubiquitination or protein–protein interactions occurs by acquisition of promiscuous domains that reduce their modularity (with $\rho_s$ in the range −0.32 to −0.84 and p<0.001; ***Figure 8—figure supplement 1A–D***), and this reduction is frequently stronger in animals than in their unicellular relatives and ancestors (***Figure 8—figure supplement 1E–H***). The high promiscuity of domains mediating protein-protein interactions has already been reported in previous analyses (***Basu et al., 2008***; ***Zmasek and Godzik, 2012***), thus confirming the validity of our approach.

However, the analysis of the transcription factor domain sub-networks exhibited an opposite signal: animal TF genes have more exclusive domains than their unicellular ancestors or relatives (reflected by higher modularities; ***Figure 8A***, lower panel). Also, there was no negative relationship between the number of domains per community and the network modularity ($\rho_s$=0.12,

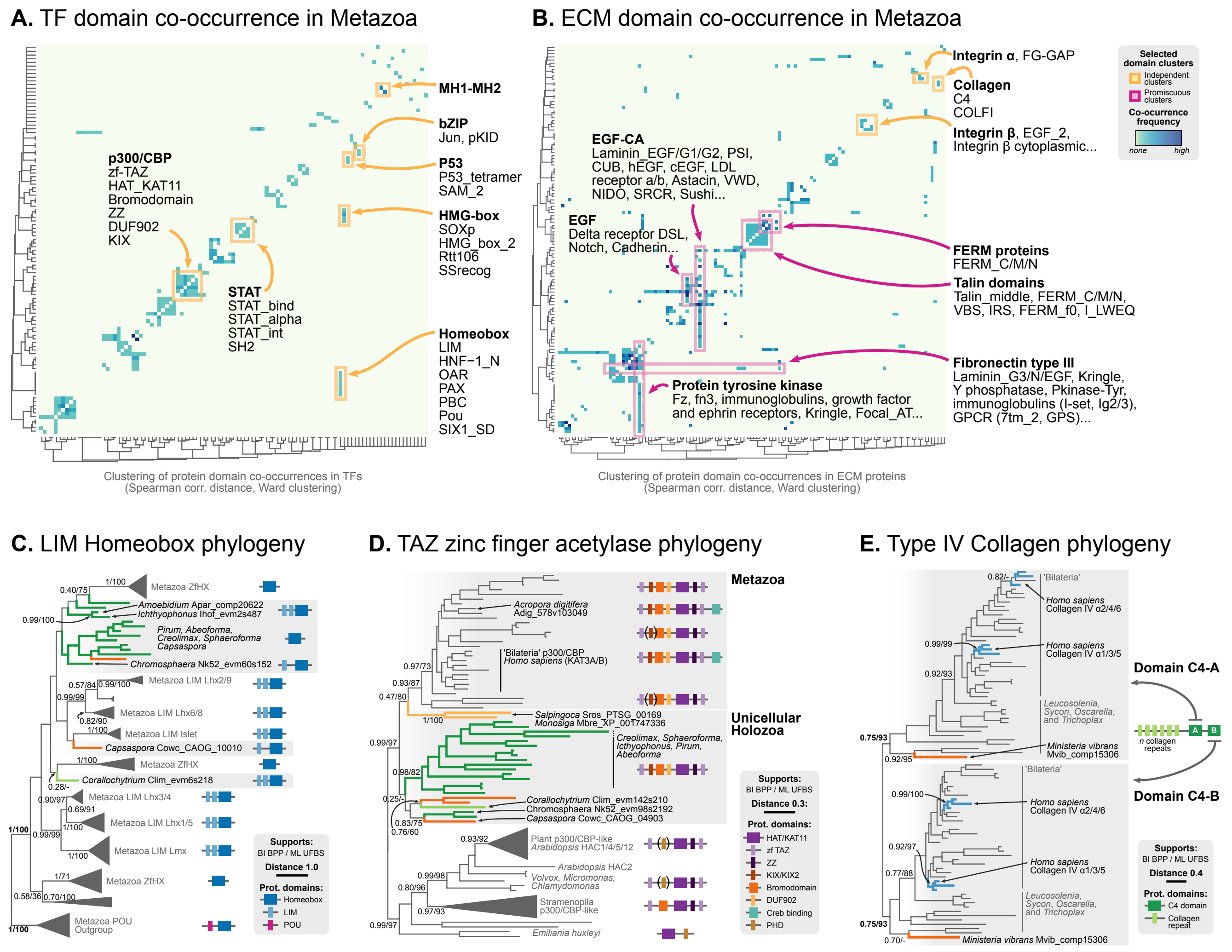


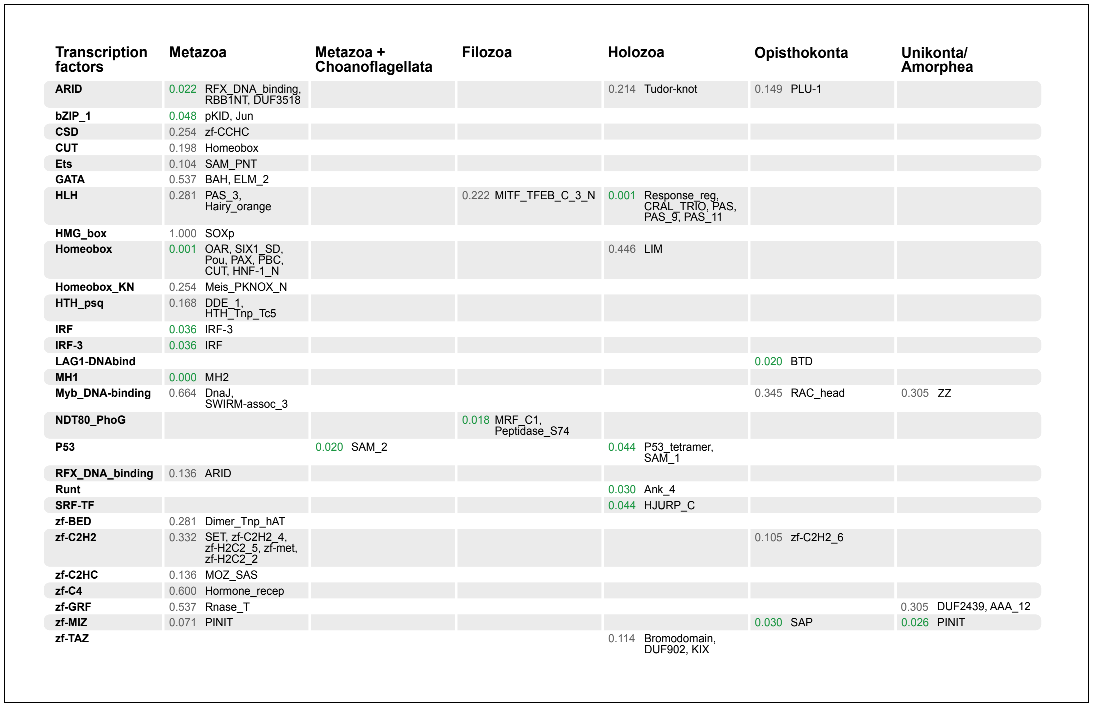
**Figure 9.** Phylogenetic analysis of the premetazoan gene families LIM Homeobox, CBP/p300 and type IV collagen. (**A and B**) Protein domain co-occurrence matrices of transcription factor (TF) (**A**) or extracellular matrix (ECM)-related gene families (**B**), inferred at the LCA of Metazoa (≥90% probability). Horizontal and vertical axes of the heatmap represent individual protein domains and their mutual co-occurrence frequency, and have been clustered according to the number of shared domains (dendrogram based on Spearman correlation distances and Ward clustering algorithm). Note that, for TFs, most co-occurrence clusters are located along the diagonal, indicating isolated domain communities; whereas ECM genes tend to contain promiscuous domains shared in multiple domain co-occurrence communities. Representative examples of independent and promiscuous domain clusters have been highlighted in both heat maps (orange and pink, respectively). (**C**) Phylogenetic tree of LIM Homeobox TFs, with mapped protein domains architectures. (**D**) Phylogenetic tree of CBP/p300 TFs based on HAT/KAT11 domain, with mapped consensus protein domain architectures. (**E**) Phylogeny of type IV collagen genes based on the C4 domain. All extant homologs, from *Ministeria* to animals, have a C4-C4 dual arrangement of filozoan origin (reflected in the phylogeny by two parallel clades representing the first and second domains within each gene). *Ministeria* (orange) and human (blue) homologs are highlighted. In **C**, **D** and **E** panels, bold branches represent unicellular holozoan genes and are color-coded by taxonomic assignment. All trees are Bayesian inferences (BI). Protein domain architectures and statistical supports (BPP/UFBS) are shown for selected nodes (see ***Figure 7—figure supplement 1*** for the complete BI and ML trees with statistical supports). Species abbreviation as in ***Figure 2—source data 2***. ***Figure 5—source data 2***, ***Figure 7—source data 1*** and **2**, ***Figure 2—source data 1***.

DOI: 10.7554/eLife.26036.034

The following figure supplement is available for figure 9:

**Figure supplement 1.** Phylogenetic analysis of the (**A**) LIM-Homeobox, (**B**) p300/CBP, and (**C**) Collagen Type IV, using Maximum likelihood in IQ-TREE (supports are SH-like approximate likelihood ratio test/UFBS, respectively) and Bayesian inference in Mr. Bayes (BPP statistical supports).

DOI: 10.7554/eLife.26036.035



| Transcription factors | Metazoa | Metazoa + Choanoflagellata | Filozoa | Holozoa | Opisthokonta | Unikonta/ Amorphea |
|---|---|---|---|---|---|---|
| ARID | 0.022 RFX_DNA_binding, RBB1NT, DUF3518 | | | 0.214 Tudor-knot | 0.149 PLU-1 | |
| bZIP_1 | 0.048 pKID, Jun | | | | | |
| CSD | 0.254 zf-CCHC | | | | | |
| CUT | 0.198 Homeobox | | | | | |
| Ets | 0.104 SAM_PNT | | | | | |
| GATA | 0.537 BAH, ELM_2 | | | | | |
| HLH | 0.281 PAS_3, Hairy_orange | | 0.222 MITF_TFEB_C_3_N | 0.001 Response_reg, CRAL_TRIO, PAS, PAS_9, PAS_11 | | |
| HMG_box | 1.000 SOXp | | | | | |
| Homeobox | 0.001 OAR, SIX1_SD, Pou, PAX, PBC, CUT, HNF-1_N | | | 0.446 LIM | | |
| Homeobox_KN | 0.254 Meis_PKNOX_N | | | | | |
| HTH_psq | 0.168 DDE_1, HTH_Tnp_Tc5 | | | | | |
| IRF | 0.036 IRF-3 | | | | | |
| IRF-3 | 0.036 IRF | | | | | |
| LAG1-DNAbind | | | | | 0.020 BTD | |
| MH1 | 0.000 MH2 | | | | | |
| Myb_DNA-binding | 0.664 DnaJ, SWIRM-assoc_3 | | | | 0.345 RAC_head | 0.305 ZZ |
| NDT80_PhoG | | | 0.018 MRF_C1, Peptidase_S74 | | | |
| P53 | | 0.020 SAM_2 | | 0.044 P53_tetramer, SAM_1 | | |
| RFX_DNA_binding | 0.136 ARID | | | | | |
| Runt | | | | 0.030 Ank_4 | | |
| SRF-TF | | | | 0.044 HJURP_C | | |
| zf-BED | 0.281 Dimer_Tnp_hAT | | | | | |
| zf-C2H2 | 0.332 SET, zf-C2H2_4, zf-H2C2_5, zf-met, zf-H2C2_2 | | | | 0.105 zf-C2H2_6 | |
| zf-C2HC | 0.136 MOZ_SAS | | | | | |
| zf-C4 | 0.600 Hormone_recep | | | | | |
| zf-GRF | 0.537 Rnase_T | | | | | 0.305 DUF2439, AAA_12 |
| zf-MIZ | 0.071 PINIT | | | | 0.030 SAP | 0.026 PINIT |
| zf-TAZ | | | | 0.114 Bromodomain, DUF902, KIX | | |

**Figure 10.** Domain combinations that appear in transcription factor (TF) families in unicellular premetazoans, from the LCA of Unikonta/Amorphea to the LCA of Metazoa. First and second columns indicate the TF family and its inferred evolutionary origin, respectively (from [***de Mendoza et al., 2013***]). Subsequent columns list (i) the *p*-value of a Fisher's exact test for the relative enrichment of that TF family in that node of the tree (compared to other domains that rearrange there; p-values<0.05 in green); and (ii) the accessory domains that appear within each TF family. ***Figure 7—source data 2***, ***Figure 10—source data 1***.

DOI: 10.7554/eLife.26036.036

The following source data is available for figure 10:

**Source data 1.** Probability of emergence of protein domain combinations present in the LCA of Metazoa in previous ancestral nodes (from LCA of Metazoa to LCA of Unikonta/Amorphea).

DOI: 10.7554/eLife.26036.037

p-value=0.32), meaning that the addition of new domains to TF genes occurred in a gene family-specific manner (***Figure 8B***). This implies that the expanded TF repertoires of animal genomes (***de Mendoza et al., 2013***) preferentially diversify their protein domain architectures by acquiring new, not promiscuous, domains.

In summary, we identify a distinct dynamics of protein domain rearrangements for TF families in the LCA of Metazoa: new domains tend to be acquired in a family-specific manner (as opposed to reuse of promiscuous domains), contributing to the functional specialization of the animal TF repertoire.

## Gene family-specific protein domain diversification: TFs and collagen IV

Our ancestral reconstruction of protein domain architectures (***Figure 7—source data 2***) allowed us to investigate the evolutionary origin of specific domain organizations within gene families and examine their diversification pattern in the ancestry of animals (***Figure 10—source data 1***). For example, we recovered many examples of gene family-specific domain diversification in novel animal TFs (***Figure 10***): Homeobox families (OAR, PBC/X, SIX, CUT, Pou, HNF or PAX families), TALE Homeobox (Homeobox_KN domain; Meis/Knox families), MH (MH1 and MH2 domains), bZIPs (Jun), C4 zinc finger (nuclear hormone receptors), Ets (Ets with modified SAM motifs) and HMG-box (SOX). Interestingly, the functions of accessory domains were often related to regulation of TF multimerisation or the DNA-binding affinities of the protein (***de Mendoza et al., 2013***; ***SebeSebé-PedrosPedrós et al., 2011***; ***Holland et al., 2007***; ***Holland, 2013***). These TF families appeared as isolated clusters when we sorted protein domains by their pattern of co-occurrence in the reconstructed Metazoa LCA (***Figure 9A***). Furthermore, we detected an unexpected premetazoan origin for some TF classes as per their domain combinations (**Figure 10**). We validated two case-in-point examples by phylogenetic analysis, in order to illustrate the distinct pattern of TF domain diversification: the LIM Homeobox (LIM-HD) and p300/CBP transcriptional coactivators.

LIM homeobox genes have been classified as an animal-specific non-TALE family (**Srivastava et al., 2010b**). However, we identified LIM-associated homeobox genes in multiple ichthyosporeans, *Corallochytrium* and *Capsaspora*. We classified these candidate genes according to HomeoDB (**Zhong and Holland, 2011**) using (**Holland et al., 2007**) as a phylogenetic reference. Our analysis identified *bona fide* LIM-HD homologs with 1–2 LIM domains in *Corallochytrium*, *Chromosphaera*, *Ichthyophonus*, *Amoebidium* and *Capsaspora* (which had 1–2 LIM domains and a homeodomain); together with many LIM-devoid homologs in *Creolimax*, *Sphaeroforma*, *Pirum* and *Abeoforma* (**Figure 9C**). None of the unicellular holozoan LIM-HD genes could be confidently assigned to animal LIM homeodomain subfamilies (*Lhx1/5*, *Lhx3/4*, *Lmx*, *Islet*, *Lhx2/9*, *Lhx6/8*), probably because they emerged before LIM-HD radiation in animals. As such, they also predate the establishment of the LIM code of cell type specification, which has been shown to control neuronal differentiation via combinatorial expression of LIM-HD subfamilies, in animals from *Caenorhabditis elegans* to mammals or the sea walnut *Mnemiopsis* (**Simmons et al., 2012**; **Thor et al., 1999**; **Gadd et al., 2011**). Given that transcriptionally regulated cell type specification has already been demonstrated in *Creolimax* (**de Mendoza et al., 2015**), the presence of LIM-HD paralogs in ichthyosporeans will require further examination, as it raises the possibility of a conserved or convergent regulatory role in cell differentiation.

The p300/CBP TF is a transcriptional activator that contributes to distal enhancer demarcation by histone acetylation in bilaterian animals and *Nematostella* (**Gaiti et al., 2017a**). Most eukaryotes have a consensus architecture composed of a central HAT/KAT11 domain (acetylase) flanked by three zinc fingers of TAZ (2) and ZZ (1) types (DNA-binding motifs) (**Figure 9D**). Animal p300/CBP homologs typically include an additional 3-domain structure, N-terminal to the acetylase domain, composed of KIX-Bromodomain-DUF902. KIX recognizes and binds to CREB in animals (a cAMP-responsibe bZIP TF), and the Bromodomain is responsible for interaction with acetylated histones. We identified this protein domain architecture in both *Capsaspora* and ichthyosporeans, which also have the CREB gene (**SebeSebé-PedrosPedrós et al., 2011**). Intriguingly, *Capsaspora*'s epigenome contains p300/CBP-specific histone acetylation marks, but its relatively compact genome lacks distal enhancers (**Sebé-Pedrós et al., 2016b**).

Finally, in stark contrast to TF domain-specific diversifications, clusters of co-occurring protein domains in ECM-related genes were dominated by highly promiscuous domains shared between different gene families (**Figure 9B**). This pattern explains the lower network modularity of animal ECM genes (**Figure 8—figure supplement 1**). Among the most promiscuous domains, we found epidermal growth factor-related domains (EGF-CA, EGF), type III fibronectin or protein tyrosine kinase motifs, consistent with previous observations (**Cromar et al., 2014**). These domains are part of multiple, functionally different gene families: structural laminins, immunoglobulins, the Notch/Delta signaling system, LDL receptors or GPCR signaling genes (pink highlight, **Figure 9B**).

The diversification of collagen genes, however, is a counterexample to the promiscuous domain shuffling at the ECM: like many TFs, collagens typically contain repetitive motifs with unique domains conferring functional specificity (**Hynes, 2012**). This includes, for example, structural fibrillar collagens (COLFI domains and further specialization within metazoans), type XV/XVIII (endostatin/NC10 domains), type IV collagen or type IV-like spongins (specific of invertebrate metazoans); there are also non-structural genes like collectin receptors (Lectin-C) or the C1q complement subcomponent (C1q) (**Hynes, 2012**; **Aouacheria et al., 2006**; **Heino, 2007**; **Fahey and Degnan, 2012**; **Exposito et al., 2008**). Most collagen genes appeared and expanded in Metazoa, concomitantly with the ECM structures they associate with (**Hynes, 2012**; **Fidler et al., 2017**). We found, however, a remarkable exception: a canonical type IV collagen gene in the filasterean *Ministeria vibrans*, a naked filose amoeba devoid of basement membrane or ECM (**Patterson et al., 1993**; **Cavalier-Smith and Chao, 2003**). Cross-linked type IV collagens are part of the structural core of animal basement membranes (to date, all of its components had been described as exclusive to animals) (**Hynes, 2012**; **Fidler et al., 2017**). This *Ministeria* ortholog is composed of a pair of C4 domains at the C-terminus and multiple collagen Gly-X-Y repeats. Phylogenetic analysis of C4 showed that this domain arrangement appeared from two duplicated motifs within the same protein, and its order is thoroughly conserved in animals and *Ministeria* (**Figure 9E**). Thus, a canonical type IV collagen was already present in the common ancestor of filastereans, choanoflagellates and animals – which was unicellular and most likely lacked ECM or basement membrane-like structures. The essential role of collagen IV in the organization of extant metazoans' tissues (**Fidler et al., 2017**) would therefore require a co-option from an earlier function in a unicellular context, as it has been previously proposed for other ECM components such as the integrin adhesome (**Sebé-Pedrós et al., 2010**) or cadherins (**Abedin and King, 2008**).

# Discussion

We have investigated the evolutionary dynamics of key genomic traits in the unicellular ancestry of Metazoa, in the first comparative genomic study that simultaneously includes all unicellular holozoan lineages, and more than one species per lineage: animals, seven Teretosporea genomes (six ichthyosporeans and *Corallochytrium*), *Capsaspora*, and two choanoflagellates (*Salpingeoca* and *Monosiga*). Our enhanced taxon sampling, including four newly sequenced genomes, allows us to perform both within- and across-lineage comparisons, thus covering the different time scales at which the evolution of coding and non-coding genome features occurred.

## Dating the origin of animal-like protein domain architectures, intron architecture and genome size

We have identified continued process of gene innovation in terms of protein domain architectures in the animal ancestry, peaking at the LCA of Holozoa. This burst of diversification, enriched in TFs and ECM-related domains (**Figure 7B**), set the foundations of the animal-like gene tool-kits of unicellular holozoans that have been reported in previous studies of gene family evolution regarding signaling pathways (**Suga et al., 2012**; **Grau-Bové et al., 2015**, **2013**), cell adhesion systems (**de Mendoza et al., 2015**; **Nichols et al., 2012**; **Sebé-Pedrós et al., 2010**) and transcription factors, often involved in developmental processes (**de Mendoza et al., 2013**; **SebeSebé-PedrosPedrós et al., 2011**). The expansion of protein diversity in early holozoans provided fertile ground for the frequent co-option of ancestral genes for multicellular functions in Metazoa (**Richter and King, 2013**). Overall, our probabilistic reconstruction of the genome content of unicellular animal ancestors (available as **Figure 7—source data 2**) provides a useful framework for targeted analysis of gene evolution and protein domain architecture evolution. As case-in-point examples of our approach, we have established the premetazoan origin of the transcription factors LIM Homeobox (present in Ichthyopsorea and *Capsapsora*) and p300/CBP-like (all unicellular Holozoa) (**Figure 9C–E**), and canonical Type IV collagens, a key element of the animal ECM (**Hynes, 2012**) (present in the filasterean amoeba *Ministeria vibrans*).

We have also investigated the time of origin of intron-rich genomes in Holozoa. We detect three independent episodes of massive intron gain: (1) at the root of Metazoa, (2) the shared LCA between Metazoa and Choanoflagellata, and (3) the root of ichthyophonid Ichthyosporea (*Creolimax*, *Sphaeroforma* and *Ichthyophonus*). Furthermore, since the early origin of introns in the earliest eukaryotes (**Irimia and Roy, 2014**), the ancestry of both animals and ichthyophonids maintained a state of high intron density. The evolutionary implications of this circumstance, however, remain an open question. First, the independent intron gain episodes of animals and unicellular holozoans are mirrored by two different modes of alternative splicing dominating in each clade: animal transcriptomes are rich in isoform-producing exon skipping (**McGuire et al., 2008**; **Irimia and Roy, 2014**), whereas most of the alternatively spliced transcripts of *Capsaspora* (**Sebé-Pedrós et al., 2013**) and *Creolimax* (**de Mendoza et al., 2015**) originate by intron retention and are thus more similar to the putative ancestral eukaryote than to Metazoa (**Irimia and Roy, 2014**). Second, we here show that the holozoan LCAs that underwent intron invasions (Ichthyophonida, Metazoa and Metazoa + Choanoflagellata) all possessed the essential NMD machinery and a rich complement of assisting splicing factors (**Figure 5C**). Thus, they were in principle able to reduce the costs imposed by slightly deleterious intron invasions, as predicted by the mutational-hazard hypothesis (**Lynch and Conery, 2003**; **Lynch, 2002**, **Lynch, 2006**). And third, the protracted state of high intron density in the ancestry of Metazoa and Ichthyophonida could have contributed to maintaining high levels of transcriptome variability that could in turn be co-opted for potentially adaptive, regulated AS events (**Irimia and Roy, 2014**; **Koonin et al., 2013**). However, we cannot at present elucidate the relative importance of adaptation and population-genetic effects in the holozoans' intron gain episodes: further transcriptomic analyses of unicellular holozoans are required to confirm that intron retention is their ancestrally prevalent AS mode (**Sebé-Pedrós et al., 2013**; **Irimia and Roy, 2014**; **de Mendoza et al., 2015**); and the scant data on unicellular holozoans' population genetics hampers the interpretation of genome architecture evolution under the light of the mutational-hazard hypothesis (**Lynch and Conery, 2003**; **Lynch, 2002**).

We also addressed the evolution of genome size across holozoans. The emergence of larger genomes in Metazoa cannot be explained solely by intron gain and gene family expansion (**Elliott and Gregory, 2015a**). Unfortunately, other factors such as the contribution of TE invasions (**Figure 3B**) or the extension of intron sites are not possible to date at the holozoan-wide evolutionary scale due to the lack of conserved signals. A possible way out of the conundrum is to study the conserved functions in the non-coding parts of the genome. For example, the compact genome of *Capsaspora* (median intergenic regions: 373 bp) has intragenic *cis*-regulatory elements key to its temporal regulation of cell differentiation (**Sebé-Pedrós et al., 2016b**), but the putative regulatory functions in the larger intergenic regions of *Creolimax*, *Sphaeroforma* and *Salpingoeca* (median intergenic 900–1200 bp) remain uncharacterized. It is tantalizing to note that (1) *Creolimax* and *Salpingoeca* exhibit temporal differentiation of cell types (**Fairclough et al., 2013**; **de Mendoza et al., 2015**), and (2) their intergenic median sizes are in line with those of *Amphimedon* (885 bp) (**Figure 1—source data 1**), a demosponge with bilaterian-like promoters and enhancers, including distal regulation (**Gaiti et al., 2017a**, **Gaiti et al., 2017b**). However, the ancestral gene linkages conserved across Metazoa, frequently due to common *cis*-regulation (**Irimia et al., 2012**), appear to be animal innovations absent in unicellular holozoans (**Figure 3—figure supplement 1**). We thus propose that homologous regulatory regions would be rarely conserved between animals and unicellular holozoans; and only common *types* of regulatory elements could be expected, e.g. distal enhancers or developmental promoters.

## Independence of genome features in premetazoan evolution

Overall, our results show that extant holozoan genomes have been shaped by both differential retention of ancestral states and secondary innovations, for the multiple genomic traits analyzed here, namely genome size, intron density, synteny conservation, protein domain diversity and gene content (reviewed in (**Richter and King, 2013**)). We can thus conclude that the genomes of unicellular premetazoans were shaped by independent evolutionary pressures on different traits, as has been seen in Metazoa (**Simakov and Kawashima, 2017**).

Our findings can help to delimit the implicit trade-offs of choosing a unicellular model organism for functional and comparative studies with Metazoa, taking into account the loss of animal-like genomic traits relevant to different analyses. For example, phylogenetic distances between orthologous genes are shorter between some ichthyosporeans and animals than between choanoflagellates and animals (**Figure 3D**), yet choanoflagellates are more similar to the animal ancestor in terms of intron structure (**Figure 6A**) and have lower rates of protein domain diversity loss (**Figure 7B**). Interestingly, *Capsaspora* emerges as a well-suited model with a slow pace of genomic change attested for multiple traits: intron evolution, coding sequence conservation, gene order and (possibly) genome size. Its remarkable microsynteny conservation with *Corallochytrium* and *Chromosphaera* indicates the existence of ancestral holozoan gene linkages that have been disrupted, and rewired, in extant choanoflagellates, ichthyosporeans and animals (**Figure 3C**). However, *Capsaspora*'s lack of close sister groups hampers comparative studies of faster-evolving genomic features, be it the regulatory circuitry (**Sebé-Pedrós et al., 2016b**), or co-option of genetic tool-kits for its unique aggregative development (**Sebé-Pedrós et al., 2013**).

The new genomes from Ichthyosporea and *Corallochytrium* analyzed here provide novel insights into the reconstruction of premetazoan genomes. The Teretosporea clade has a deeper sampling than other unicellular holozoans and exhibit a mixture of slow- and fast-evolving genomic traits, which provides novel insights into the independence of genomic characters during premetazoan evolution. For example, *Ichthyophonus* tends to retain the ancestral intron/exon structure (**Figure 6A**) and is relatively similar to animals in terms of coding sequence conservation (**Figure 3D**), but it harbors a secondarily expanded genome with disrupted gene linkage (**Figure 3A, C**). Another example is *Corallochytrium* and *Chromosphaera*, both with massive simplifications of intron content (**Figure 5A**), but higher synteny conservation with the inferred ancestral Holozoa (**Figure 3C**). Also, the diversity of protein domain combinations of *Chromosphaera* is the highest among ichthyosporeans (in line with values of animals and holozoan ancestors; ***Figure 7A***) and phylogenetic distances to animal orthologs are comparatively low (***Figure 3D***). These studies of genome history in holozoans are key to our interpretation of functional genomics analyses. For example, *Creolimax* and *Sphaeroforma* are close species with a broadly conserved life cycle (***Glockling et al., 2013***), and they could therefore be an apt model to test hypotheses of cell type evolution in Holozoa – for example, whether new cell types emerge as lineage-specific transcriptomic specializations, as proposed by (***de Mendoza et al., 2015***). This investigation would benefit from taking into account their high microsynteny when analyzing co-regulated gene modules, while considering that *Sphaeroforma*'s multiple TE invasions could blur the conservation of non-coding regulatory elements in the intergenic regions (***Figure 3A–C***).

## Genomic innovation in the animal ancestry

Our analysis of ten unicellular holozoans has uncovered the timing of genome evolution in the ancestry of Metazoa, at both the architectural and gene content levels. In particular, we have observed that holozoan genomes evolved under temporally uncoupled dynamics for synteny reorganization, intron gains, TE propagation, coding sequence conservation and gene family diversification. Some of these traits have independent effects in extant holozoans, e.g., different episodes of intron gain or genome expansion in ichthyosporeans and animals. Yet, other traits exhibit conserved dynamics across the unicellularity/multicellularity divide: the diversification of ECM and TF gene families—including molecular tool-kits essential for multicellularity—extends back to the LCA of Holozoa; and the high intron densities in premetazoans suggest a continued state of transcriptome variability, co-optable for regulation or protein innovation, in the unicellular prehistory of Metazoa. Overall, our timeline of holozoan genome evolution offers a framework to investigate when and how premetazoan genomic elements—gene tool-kits, linkages and structure, and the non-coding sequences that harbor epigenomic regulatory elements—were functionally co-opted in multicellular animals.

# Materials and methods

## Cell cultures

*Corallochytrium limacisporum*, *Abeoforma whisleri* and *Pirum gemmata* were grown in axenic culture in marine broth medium (Difco 2216) at 18°C (*Abeoforma* and *Pirum*) or 23°C (*Corallochytrium*). *Chromosphaera* was grown in axenic culture at 18°C in YM medium (containing 3 g yeast extract, 3 g malt extract, 5 g bacto peptone, 10 g dextrose, 14.5 g Difco agar, and 25 g sodium chloride, per liter of distilled water).

## DNA and RNA extraction and sequencing

DNA-seq data was produced for *Pirum*, *Abeoforma*, *Chromosphaera* and *Corallochytrium*, by sequencing paired-end (PE) and Nextera mate-pair (MP) libraries. DNA extractions were performed from confluent axenic cultures, grown in three flasks of 25 ml for 5 days. DNA was extracted using a standard protocol by which cells were lysed in the extraction buffer composed of Tris-HCL, 50 mM EDTA, 500 mM NaCl and 10 mM ß-mercaptoethanol. DNA was purified with phenol:chloroform:isoamyl alcohol (25:24:1) and treated with of Rnase A (Sigma Aldrich, Saint Louis, MO, USA). For each library, the read numbers, lengths and insert/fragment sizes were as follows: *Pirum,* PE 125 bp ($250{\cdot}10^6$ reads, 250 bp insert size), MP 50 bp ($108{\cdot}10^6$ reads, 6 kb fragment size); *Abeoforma*, PE 100 bp ($73{\cdot}10^6$ reads, 600 bp insert size), MP 100 bp ($41{\cdot}10^6$ reads, 6 kb fragment size); *Chromosphaera*, PE 125 bp ($143{\cdot}10^6$ reads, 250 bp insert size), MP 50 bp ($114{\cdot}10^6$ reads, 5 kb fragment size); and *Corallochytrium*, PE 100 bp ($150{\cdot}10^6$ reads, 420 bp insert size), MP 100 bp ($47{\cdot}10^6$ reads, 3 kb fragment size). All PE and MP libraries were prepared and sequenced at the CRG Genomics Unit (Barcelona), using Illumina HiSeq 2000 and the Trueseq Sequencing Kit v3 (*Abeoforma* and *Corallochytrium*) or v4 (*Pirum* and *Chromosphaera*). The only exception was *Corallochytrium* PE libraries, which were sequenced at the Earlham Institute Genomics Unit (Norwich, UK) using Illumina MiSeq and the Trueseq protocol v2. Genome sequencing data has been deposited in NCBI SRA under the BioProject accession PRJNA360047.

RNA-seq data was produced for *Chromosphaera* and *Abeoforma*. RNA extractions were performed from confluent axenic cultures grown in three 25 ml flasks for 5 days. RNA was extracted using Trizol reagent (Life Technologies, Carlsbad, CA, USA) with a further step of Dnase I (Roche) to avoid contamination by genomic DNA, then purified using RNeasy columns (Qiagen). We sequenced PE libraries of 125 bp with an insert size of 250 bp, yielding $168{\cdot}10^6$ reads for *Chromosphaera* and $178{\cdot}10^6$ for *Abeooforma*; which were constructed using the Trueseq Sequencing Kit v4 (Illumina, San Diego, CA). The libraries were sequenced in one lane of an Illumina HiSeq 2000 at the CRG genomics unit (Barcelona). All transcriptome sequencing data has been deposited in NCBI SRA using the BioProject accession PRJNA360056.

## Genome assembly

Genomic PE and MP libraries were quality-checked using FastQC v0.11.2 (**Andrews, 2014**) and trimmed accordingly with Trimmomatic v0.33 (**Bolger et al., 2014**) to remove remnant adapter sequences (*ad hoc*) and the low-quality 5' read ends (sliding window = 4 and requiring a minimum Phred quality = 30). A minimum length equal to the original read length was required. During the quality-trimming process, libraries of unpaired forward reads were kept as single-end reads (SE). After trimming, the read survival rate for each DNA library was as follows: *Pirum*, PE 30.2%, MP 91.2%; *Abeoforma*, PE 75.5%, MP 31.0%; *Chromosphaera*, PE 81.1%, MP 89.9%; and *Corallochytrium*, PE 94.7%, MP 73.1%.

Genome assemblies were performed using Spades v3.6.2 (**Nurk et al., 2013**) with the BayesHammer error correction algorithm (**Nikolenko et al., 2013**). For each organism, PE data were analyzed using Kmergenie (**Chikhi and Medvedev, 2014**) to determine the optimal k-mer length for the assembly process, which was used in the Spades assembly in combination with smaller and larger values, including the maximum possible odd length below the maximum read length after trimming. The optimized assemble parameters for each genome were as follows: *Pirum*, max. read length = 125, k = 55,123; *Abeoforma*, max. read length = 100, k = 47,91; *Chromosphaera*, max. read length = 125, k = 91,121; *Corallochytrium*, max. read length = 100, k = 41,63,91. In the cases of *Corallochytrium* and *Chromosphaera* genomes, Spades was run in *careful* mode, taking into account PE, SE and MP data in the same run. In the cases of the highly repetitive *Abeoforma* and *Pirum* genomes, an initial Spades assembly of PE and SE libraries was combined with MP libraries using the Platanus v1.2.1 scaffolding module (**Kajitani et al., 2014**). Each assembly was later processed using the GapCloser module from SOAPdenovo assembler with PE data, in order to extend the scaffolded contigs by shortening N stretches (**Luo et al., 2012**). Genome assembly statistics (genome size, N50, L75) were calculated using Quast v2.3 (**Gurevich et al., 2013**), and completeness was assessed using the BUSCO v1.1 (**Simão et al., 2015**) database of universal eukaryotic genes, based on the predicted transcripts.

## Genome annotation

Genome feature annotations were produced for *Corallochytrium*, *Chromosphaera*, *Abeoforma*, *Pirum* and *Ichthyophonus*. We used evidence-based gene finders (relying on transcript/peptide mapping: Augustus v3.1 (**Keller et al., 2011**) and PASA v2.0.2 [**Haas et al., 2003**, **2008**]), plus complementary *ab initio* predictors (based on hidden Markov models for gene structure: GeneMark-ES v4.21 (**Lomsadze et al., 2005**) and SNAP [**Korf, 2004**]). These results were combined to produce a consolidated gene annotation using Evidence Modeler v1.1.1 (**Haas et al., 2008**).

SNAP and GeneMark-ES annotations were iterated for three times on the final genome assemblies, using the output of each step as a training set for the next one (the first SNAP prediction was done using the standard minimal HMM; GeneMark-ES was omitted for *Abeoforma* and *Pirum* due to its highly fragmented gene bodies, which impaired intron delimitation).

Transcriptome assemblies were produced to support PASA and Augustus gene predictions. RNA-seq PE libraries were assembled using genome-guided Trinity v2.0.6 and STAR v2.5 (for *Corallochytrium*, *Chromosphaera* and *Ichthyophonus*) or de novo Trinity (*Pirum* and *Abeoforma*, assemblies from (**Torruella et al., 2015**; **Grabherr et al., 2011**; **Dobin and Gingeras, 2015**). In the case of the *Corallochytrium*, *Chromosphaera* and *Ichthyophonus* genome-guided assemblies, quality control was performed as indicated above for the genomic libraries, using the RNA-seq data generated for this study (*Chromosphaera*) or in (**Torruella et al., 2015**) (*Ichthyophonus* accession: PRJNA264423; *Corallochytrium* accession: PRJNA262632). A minimum k-mer coverage = 2 was used in all Trinity assemblies. Transcriptome assemblies were annotated with Transdecoder using Pfam release 29 protein domain database, in order to obtain mRNA and translated peptides. Next, PASA annotations were obtained from assembled transcripts, mapped to the genome using GMAP and BLAT v35 (**Kent, 2002**; **Wu et al., 2016**). Only high quality mapping was accepted, with a minimum of 95% identity and 75% transcript coverage. We then trained Augustus independently, using protein and mRNA predictions (mapped to the genome with Scipio 1.4 (**Keller et al., 2008**), BLAT and GMAP), followed by an optimization round of the species-specific parameters. After the training, an Augustus prediction was performed using the optimized parameters.

Finally, all annotations were consolidated using Evidence Modeler. In this step, gene models from PASA and Augustus were given higher relative weights than *ab initio*-predicted models (10 and 5 times more reliability, respectively).

## Phylogenomic analysis

We used an improved version of the dataset published by Torruella *et al.* (**Torruella et al., 2015**), adding nine single-copy protein domains to the previous version (which included 78 alignments) according to the methodology developed in (**Torruella et al., 2012**). Since *Abeoforma* and *Pirum* genome assemblies were fragmented and contained partial gene models, we used transcriptome assemblies from (**Torruella et al., 2015**) instead. We compiled a 57-taxa dataset of Unikonta/Amorphea species (hereby termed BVD57 taxa matrix; including Holozoa, Holomycota, Breviatea, Apusomonadida and Amoebozoa; 24,021 amino acid positions). This dataset represents a ~ 10% increase in the number of aligned positions, compared to the original S70 dataset from (**Torruella et al., 2015**).

We used the BVD57 dataset to build ML phylogenetic trees using IQ-TREE v1.5.1 (**Nguyen et al., 2015**), under the LG model with a 7-categories free-rate distribution, and a frequency mixture model with 60 frequency component profiles based on CAT (LG + R7+C60) (**Quang et al., 2008**). LG + R7 was selected as the best-fitting model according to the IQ-TREE *TESTNEW* algorithm as per the Bayesian information criterion (BIC), and the C60 CAT approximation was added because of its higher rate of true topology inference (**Quang et al., 2008**). Statistical supports were drawn from 1000 ultrafast bootstrap values with a 0.99 minimum correlation as convergence criterion (**Minh et al., 2013**) and 1000 replicates of the SH-like approximate likelihood ratio test (**Guindon et al., 2010**), for all models stated above. Furthermore, 500 non-parametric bootstrap replicates were computed for the LG + R7+PMSF CAT approximation (as this was the only CAT approximation for which non-parametric bootstraps could be calculated in a feasible computation time).

We then used the same alignment to build a Bayesian inference tree with Phylobayes MPI v1.5 (**Lartillot et al., 2013**), using the LG exchange rate matrix with a 7-categories gamma distribution and the non-parametric CAT model (**Lartillot and Philippe, 2004**) (LG+$\Gamma$7 + CAT). A $\Gamma$7 distribution was considered to be the closest approximation to the free-rates R7 distribution of the IQ-TREE ML analysis (as free-rates distributions are not implemented in Phylobayes). We removed constant sites to reduce computation time. We ran two independent chains for 1231 generations until convergence was achieved (maximum discrepancy <0.1) with a burn-in value of 32% (381 trees). The adequate burn-in value was selected by sequentially increasing the number of burn-in trees, until we achieved (1) a minimum value of the maximum discrepancy statistic, and (2) the highest possible effective size for the log-likelihood parameter. The *bpcomp* analysis of the sampled trees yielded a maximum discrepancy = 0.095 and a mean discrepancy = 0.001. The *tracecomp* parameter analysis gave an effective size for the log-likelihood parameter = 37; and the minimum effective size = 11 (for the alpha statistic).

## Generation of a species tree and ortholog datasets for comparative analyses

Our comparative genomics analyses are based on a dataset of 42 complete eukaryotic genomes, with a focus on unicellular and multicellular Holozoa, and using relevant outgroups from the Holomycota, Apusomonadida, Amoebozoa, Viridiplantae, Stramenopila, Alveolata, Rhizaria and Excavata groups. The complete list of species, abbreviations and data sources is available as ***Figure 2—source data 2***.

Since ancestral state reconstruction requires the assumption of an explicit species tree, we classified the 42 genomes in our dataset according to a consensus of phylogenomic studies (***Torruella et al., 2015***; ***Derelle et al., 2015***; ***He et al., 2014***) and our own results. We remained agnostic about the internal topology of SAR (***Burki et al., 2016***), Fungi (***Torruella et al., 2015***), the contentious hypotheses for the root of eukaryotes (namely, 'Opimoda-Diphoda' or 'Excavata-first') (***Derelle et al., 2015***; ***He et al., 2014***) and the earliest-branching animal group (Porifera or Ctenophora) (***Whelan et al., 2015***; ***Simion et al., 2017***). All these cases were recorded in our species tree as polytomic branchings.

We inferred two different ortholog datasets using the predicted proteins from the afore-mentioned genomes, using Orthofinder v0.4.0 with a MCL inflation = 2.1 (***Emms and Kelly, 2015***). The first database included 40 eukaryotic species (excluding the low-quality gene models of *Pirum* and *Abeoforma*), whose genes were classified in 162,559 clusters of orthologs, 26,377 of which contained >1 gene (henceforth, 'orthocluster'). The second database included all available unicellular holozoan genomes (*i.e.*, six ichthyosporeans, two choanoflagellates, *Corallochytrium* and *Capsaspora*) and yielded 58,516 orthoclusters, 11,925 of which contained >1 gene.

## Gene family evolution analyses

### Retrieval of homologous sequences

Retrieval of homologous protein sequences was performed by querying orthologs or protein domain HMM profiles (depending on the gene family; see below) against a database of protein sequences from 69 selected eukaryotic genomes and transcriptomes (***Figure 2—source data 2***). Since *Abeoforma* and *Pirum* genome assemblies were fragmented and contained broken gene models, we used transcriptome assemblies from (***Torruella et al., 2015***) instead.

The following gene families were defined by its catalytic/representative protein domain: type IV collagen (PF01413), TAZ zinc finger TFs with HAT/KAT11 domains (PF08214), Upf1 (PF09416), Upf2 (PF04050), Upf3 (PF03467), Smg1 (PF15785), Smg8/9 (PF10220), eRF1 (combination of PF03463 +PF03464+PF03465), Y14 (PF09282), Magoh (PF02792) and MLN51/CASC3 (PF09405). Homologs were thus retrieved by querying Pfam protein domains (29th Pfam release (***Punta et al., 2012***)), using HMMER v3.1b2 (***HMMER, 2015***) searches with *hmmersearch*, using the profile-specific gathering threshold cut-off.

In the case of LIM homeodomain genes, we queried the genomes/transcriptomes of all available unicellular holozoans (see taxon sampling above) using the homeobox HMM (PF00046), and restricted the subsequent phylogenetic analysis (see below) to sequences that clustered with known LIM-HD genes from the HomeoDB database in *blastp* searches (***Zhong and Holland, 2011***).

In the case of the eRF3, eIF4A3, Smg5/6/7 and SRSF1-9 gene families, we queried the genomes/transcriptomes mentioned above using *blastp* searches of the human orthologs of these gene families (Uniprot accession numbers: eRF3 is P15170; eIF4A3 is P38919; Smg5/6/7 are Q9UPR3/Q86US8/Q92540; SRSF1-9/TRA2 (***Plass et al., 2008***) are Q07955, Q01130, P84103, Q08170, Q13243, Q13247, Q16629, Q9BRL6, Q13242 and P62995). For eRF3 and eIF4A3 searches, we also included a selection of orthologs from the nearest outgroup gene families: EF1-alpha and HBS1L genes for eRF3 (human accessions: Q05639/P68104 and Q9Y450); and eIF4A1/2 for eIF4A3e (human accessions: P60842/Q14240).

### Protein alignments and phylogenetic analyses (LIM homeobox, type IV collagen and CBP/p300, Smg5/6/7, eIF4A3, eRF3, SRSF1-9/TRA2 splicing factors)

Protein alignments were built with MAFFT v7.245 (***Katoh and Standley, 2013***), using the G-INS-i algorithm optimized for global homology for single-domain alignments (LIM homeobox, type IV collagen, CBP/p300 and SRSF1-9/TRA2) or the E-INS-i for multiple local homology for whole-protein alignments (Smg5/6/7, eIF4A3 and eRF3). All alignments were run for up to $10^6$ cycles of iterative refinement. Then, the resulting alignments were manually examined, curated and trimmed (a process that included the removal of non-homologous amino acid positions and, eventually, non-essential sequences containing too few aligned positions that could disrupt the subsequent phylogenetic analysis). If necessary, the alignment and trimming process was repeated to incorporate the changes from manual curation.

Phylogenetic analyses were performed in the final, trimmed alignments using two independent approaches: maximum likelihood using IQ-TREE v1.5.1 (**Nguyen et al., 2015**) and Bayesian inference using MrBayes v3.2.6 (except in the case of SRSF1-9/TRA2, in which Bayesian inference it was omitted due to the large number of retrieved sequences) (**Ronquist and Huelsenbeck, 2003**). The optimal evolutionary models for each alignment were selected using ProtTest v3.4's BIC criterion (**Darriba et al., 2011**), yielding LG+$\Gamma$4 + i as the best model for the Collagen IV, HAT/KAT11, LIM Homeobox, eRF3 and eIF4A3 phylogenies; LG+$\Gamma$4 for SRSF1-9/TRA2; and LG+$\Gamma$4 + F + i for Smg5/6/7.

For IQ-TREE (**Nguyen et al., 2015**) analyses, the best-scoring ML tree was searched for up to 100 iterations, starting from 100 initial parsimonious trees; statistical supports for the bipartitions were drawn from 1000 ultra-fast bootstrap (**Minh et al., 2013**) replicates with a 0.99 minimum correlation as convergence criterion, and 1000 replicates of the SH-like approximate likelihood ratio test. For MrBayes analyses, we ran two independent runs of four chains each (three cold, one heated) for a variable number of generations until run convergence was achieved (at different values depending on the gene family), sampling every 100 steps and running a diagnostic convergence analysis every 1000 steps. Convergence was deemed to occur when, using a 25% relative burn-in value, the average standard deviation of split frequencies was <0.01. Final number of generations for each gene family: $7.2{\cdot}10^7$ generations for Collagen IV; $1.2{\cdot}10^7$ for LIM Homeobox; and $9.9{\cdot}10^6$ for HAT/KAT11; $6.4{\cdot}10^7$ for Smg5/6/7; $1.6{\cdot}10^7$ for eRF3; $7{\cdot}10^6$ for eIF4A3.

### Other ortholog searches (Upf1, Upf2, Upf3, Smg1, Smg8/9, eRF1, Y14, Magoh and MLN51)

The following gene families, part of the NMD machinery, are unambiguously defined by the presence of their defining protein domains (see above): Upf1, Upf2, Upf3, Smg1, Smg8/9, eRF1, Y14, Magoh and MLN51. Thus, presence of the protein domain in a given species was used to establish the presence of the corresponding ortholog.

## Analysis of repetitive elements

Repetitive regions were annotated in Holozoa genomes using RepeatMasker open-4.0.5 (**Smit et al., 2015**) and annotations from the 20150807 release of GIRI RepBase database (**Bao et al., 2015**). We used the Eukaryota-specific database, with either the slow high-sensitivity search mode (unicellular holozoans) or the default search mode (metazoans); and stored the genome coordinates of TEs, low complexity repeats, tRNA genes, simple repeats and satellite regions. Internal similarity of each genome's TE complements was analyzed with *blastn* self-alignments of all TEs (considering a minimum 70% identity and 80 bp alignment length), and the distribution of percentage identity values was plotted using R.

## Analysis of gene microsynteny by ortholog pair collinearity

We used the frequency of collinear ortholog pairs as a proxy to estimate microsynteny across holozoans. Specifically, we retrieved all sets of single-copy orthologs for each pairwise species comparison within our set of 10 unicellular holozoan genomes. We then defined collinear gene pairs for each species pairs if the same two orthologs were adjacent in both genomes (irrespective of individual gene orientation to account for possible local inversions, as in (**Putnam et al., 2007**)). To account for spurious conservation of gene order, we assigned random positions to each gene using the bedtools v2.24.0 *shuffle* utility (**Quinlan and Hall, 2010**) in 100 independent rounds, for which the number of spurious conserved syntenic pairs was recorded. Then, we calculated the gene synteny ratio *r* of each species pair *i-j* as follows:

$$r_{ij} = \frac{\left(\frac{c_{ij} - s_{ij}}{N_{ij}}\right)}{\left(\frac{c_{max} - s_{max}}{N_{max}}\right)}$$

where *c* denotes the number of syntenic orthologs between *i* and *j; s* is the number of spurious syntenic orthologs averaged over 100 random replicates; and *N* is the number of comparable ortholog pairs between *i* and *j*. Values are normalized to the 0–1 interval using the maximum values of the dataset as a reference, i.e. *Sphaeroforma* and *Creolimax*. A heatmap representing the degree of similarity in pairwise species comparisons was produced using the synteny ratio (R gplots library (**Warnes et al., 2016**)). Species were clustered according to their mean synteny. The same analysis was performed using the database of 40 eukaryotic genomes, which excluded *Abeoforma* and *Pirum*. In this case, the maximum values used as a reference were the *Nematostella-Aiptasia* pair.

For specific selected species comparisons, syntenic pairs were plotted onto the genome scaffolds using Circos v0.67 (**Krzywinski et al., 2009**).

## Analysis of coding sequence conservation

From our ortholog database using 40 eukaryotic genomes (excluding *Pirum* and *Abeoforma*, which had lower-quality gene annotations due to their fragmented assemblies), we selected 143 orthoclusters present in all unicellular holozoans, plus *Amphimedon queenslandica, Trichoplax adhaerens, Homo sapiens* and *Nematostella vectensis* (as representative animal genomes). We aligned each group of orthologs using MAFFT G-INS-i (**Katoh and Standley, 2013**), trimmed the alignments using trimAL automated algorithm (**Capella-Gutiérrez et al., 2009**), and inferred maximum likelihood trees for each ortholog group using RAxML v8.2.0 (**Stamatakis, 2014**) and the LG amino acid substitution model. Then, for each tree, we recorded all pairwise phylogenetic distances between species as measured by substitutions per alignment position using the cophenetic module of the ape v3.5 R library (**Paradis et al., 2004**; **Core Team, 2015**). We retrieved distances between each unicellular holozoan ortholog and, separately, *Amphimedon*, *Trichoplax*, *Homo* and *Nematostella* orthologs. For each inter-species comparison, we tested the significance of differences in phylogenetic distances between unicellular holozoans, using the non-parametric Wilcoxon rank sum test from the R stats library (**Core Team, 2015**).

## Comparative analysis of intron content

Intron content of a subset of 40 eukaryotic genomes (excluding *Abeoforma* and *Pirum*, which had lower-quality gene annotations due to their fragmented assemblies) was analyzed using a set of single-copy orthologous genes, and used to reconstruct ancestral states as described by Csűrös *et al.* (**Csuros et al., 2011**; **Csurös et al., 2007**, **Csurös et al., 2008**). We then selected orthocluster present as single-copies in 80% of our species dataset, allowing for paralog genes to occur in just one species per group (if that was the case, the best-scoring copy in BLAST alignments was kept). This yielded a group of 342 nearly paneukaryotic genes, whose protein translations were then aligned using MAFFT v7.245 G-INS-i algorithm (**Katoh and Standley, 2013**) and annotated with their intron coordinates (retrieved from their respective genome annotations). With this information, we reconstructed the ancestral states of each intron using the Malin implementation of the probabilistic model of intron evolution developed by Csűrös *et al.* (**Csűrös and Miklós, 2006**; **Csurös, 2008**), starting from the standard null model, running 1000 optimization rounds (likelihood convergence threshold = 0.001) and assuming a consensus eukaryotic phylogeny (see *Generation of a species tree for comparative analyses*).

Conserved intron sites (defined as unambiguously aligned in 80% of the orthologs, maximum of 10% of gap positions) were used to calculate the rates of intron loss and gain for each node of the tree. These rates were used to calculate a table of intron sites with a certain probability of presence, gain or loss at every node of the tree (which, when summed, give the number of introns that are present, gained or lost at that node (**Csűrös and Miklós, 2006**)). We computed 100 bootstrap replicates in Malin to assess uncertainty about inferred rate parameters and evolutionary history. In particular, we calculated the variance-to-mean ratio of the inferred number of introns in each ancestor with 100 bootstrap replicates (with values higher than one indicating more dispersed results and less reliable inferences).

For each node *i*, we calculate the percentage of introns gained ($p_{G,i}$)or lost ($p_{L,i}$) as a percentage of the total number of introns at that node. Then, the gain/loss ratio of a node, $r_i$, was calculated as follows:

$$p_{G,i} > p_{L,i} \rightarrow r_i = log_{10}\left(\frac{p_{G,i}}{p_{L,i}}\right) p_{L,i} < p_{L.i} \rightarrow r_i = log_{10}\left(\left(\frac{p_{G,i}}{p_{L,i}}\right)^{-1}\right) \times -1$$

We represented the presence and absence of intron sites at each lineage (extant and ancestral), and the number of introns shared between species (only extant), using heatmaps (R gplots library (**Warnes et al., 2016**)). Inter-species distances were calculated using the pairwise counts of shared introns and the Spearman correlation algorithm, which was used to perform Ward hierarchical clustering as implemented in R stats library (**Core Team, 2015**). We used the same algorithms to calculate distances of intron presence probability profiles, and subsequent clustering.

For *Capsaspora*, the phylostratigraphy of intron sites was combined with the nucleosome-free sites identified by ATAC-seq analysis in (**Sebé-Pedrós et al., 2016b**), which were assumed to be putative regulatory sites. Then, we compared phylostrata distribution ('ancestral' *versus* 'recent' *Capsaspora*-specific sites) for introns with and without regulatory sites, using a Fisher's exact test: 74 recent introns and 465 ancestral introns lacked putative regulatory sites ($\geq$50% ATAC site overlap with the intron sequence, calculated using bedtools v2.24.0 *intersect* utility (**Quinlan and Hall, 2010**)), while 3 and 22 recent and ancestral introns had regulatory sites.

## Comparative analysis of protein domain architecture evolution

Protein domain architectures of the 40 eukaryotic species subset (excluding *Abeoforma* and *Pirum*, which had lower-quality gene annotations due to their fragmented assemblies) were computed using Pfamscan and the 29th release of the Pfam database (**Punta et al., 2012**). For each protein, the domain architecture was decomposed into all possible directed binary domain pairs (ignoring repeated consecutive domains; i.e. from protein A-B-B-C, the pairs A-B, A-C and B-C were built), and linked to its presence in its corresponding orthocluster (see *Generation of a species tree and ortholog datasets for comparative analyses* section). The final output was a numerical profile of species distribution for each combination of domain pairs in orthoclusters (considering that a cluster can contain more than one pair, and a pair can be present in more than one cluster, and thence the number of occurrences is recorded).

The numerical profile was analyzed using the general phylogenetic birth-and-death model developed by Csűrös and Miklós (**Csűrös and Miklós, 2006**) as implemented in Count (**Csurös, 2010**). This allows the comparative analysis and ancestral reconstruction of discretized quantitative properties of genomes, assuming a specific species tree (see *Comparative analysis of intron content*). We used a gain-loss-duplication model with unconstrained gain/loss and duplication/loss ratios in all lineages, assuming a Poisson distribution of orthocluster size at the LECA (root) and no rate variation categories. In this context, 'gain' was defined as the acquisition of a new pairwise domain combination in an orthocluster; a 'duplication' as the propagation of the combination (by gene duplication or convergent domain rearrangements); and 'loss' as pair dissociation. Starting from the standard null model, we ran 100 optimization rounds (convergence threshold = 0.1).

To analyze the modularity of the protein domain networks (and subnetworks) for each genome, we 1) calculated the community structure of each network using Louvain iterative clustering to obtain communities of domain pairs (undirected graphs), and 2) calculated the global network modularity according to these communities. The modularity parameter measures the fraction within-community edges minus the expected value obtained from a network with the same communities but random vertex connections (**Newman and Girvan, 2004**). A maximum value of 1 indicates a strong community structure, while a minimum value of 0 indicates that within-community edges are as frequent as expected in a random network. For these analyses we used the relevant algorithms from the igraph R library v1.0.1 (**Core Team, 2015**; **Csárdi and Nepusz, 2006**). Function-oriented domain subnetworks were obtained by retrieving orthologous groups that contained relevant domains, which were obtained from previous studies (transcription factors from (**de Mendoza et al., 2013**; **Weirauch and Hughes, 2011**), signaling domains from (**Richter and King, 2013**), ECM-related domains from (**Richter and King, 2013**; **Sebé-Pedrós et al., 2010**; **Hynes, 2012**), ubiquitination from (**Grau-Bové et al., 2015**)) and pfam2go annotations (for the subsets mentioned above, and also for protein-binding domains) (**Mitchell et al., 2015**). Monotonic statistical dependence between modularity and the number of domains per community was tested using Spearman's rank correlation coefficient ($\rho_s$) for all network or subnetwork (for original and simulated data).

## Comparative analysis of individual protein domain evolution

We mapped the presence of individual protein domains across our dataset of 40 eukaryotic species (excluding *Abeoforma* and *Pirum*), as predicted by Pfamscan and the 29th release of the Pfam database (***Punta et al., 2012***). Using this numerical profile of domain presence in extant genomes, we computed the gains and losses at ancestral nodes using the Dollo parsimony algorithm as implemented in Count (***Csurös, 2010***).

## Accession numbers

Genome sequencing and assembly data from *Corallochytrium*, *Abeoforma*, *Pirum* and *Chromosphaera* has been deposited in NCBI using the BioProject accession PRJNA360047. Transcriptome sequencing data from *Abeoforma* and *Chromosphaera* has been deposited in NCBI using the BioProject accession PRJNA360056.

# Acknowledgements

This work was supported by an ERC Consolidator Grant (ERC-2012-Co-616960), support from the Secretary's Office for Universities and Research of the Generalitat de Catalunya (project 2014 SGR 619) and two grants from the Spanish Ministry for Economy and Competitiveness (MINECO; BFU2011-23434 and BFU2014-57779-P, the latter with European Regional Development Fund support), all to IRT. XGB was supported by a pre-doctoral FPI grant from MINECO (except for the January-March 2015 period). GT was funded by a European Marie Skłodowska-Curie Action (704566 AlgDates). HS was supported by JSPS KAKENHI 16K07468 and research grants from the NOVARTIS foundation for the Promotion of Science, ITOH Science Foundation, and JUTEN grant from the Prefectural University of Hiroshima. We thank Krista M. Nichols (NOAA Fisheries) and Chris Whipps (SUNY-ESF) for sharing their assembly of *Ichthyophonus hoferi*. We warmly acknowledge the help of Arnau Sebé-Pedrós and Alex de Mendoza for their invaluable comments on the manuscript; and thank Meritxell Antó, Elisabeth Hehenberger, Manuel Irimia, David López-Escardó, Jordi Paps, Dan Richter, and Valèria Romero-Soriano for their technical assistance and insightful remarks on the study.

# Additional information

## Funding

| Funder | Grant reference number | Author |
|---|---|---|
| Ministerio de Economía y Competitividad | BFU2014-57779-P | Iñaki Ruiz-Trillo |
| European Commission | ERC-2012-Co -616960 | Iñaki Ruiz-Trillo |
| Ministerio de Economía y Competitividad | BFU-2011-23434 | Iñaki Ruiz-Trillo |
| Generalitat de Catalunya | 2014 SGR 619 | Iñaki Ruiz-Trillo |

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

## Author contributions

XG-B, Conceptualization, Resources, Data curation, Formal analysis, Validation, Investigation, Methodology, Writing—original draft, Writing—review and editing; GT, Resources, Data curation, Formal analysis, Writing—review and editing; SD, Resources, Writing—review and editing; HS, GL, TAR, Resources, Data curation, Writing—review and editing; IR-T, Conceptualization, Supervision, Funding acquisition, Investigation, Project administration, Writing—review and editing

## Author ORCIDs

Xavier Grau-Bové, http://orcid.org/0000-0003-1978-5824
Guifré Torruella, http://orcid.org/0000-0002-6534-4758
Guy Leonard, http://orcid.org/0000-0002-4607-2064
Iñaki Ruiz-Trillo, http://orcid.org/0000-0001-6547-5304

# Additional files

## Major datasets

The following datasets were generated:

| Author(s) | Year | Dataset title | Dataset URL | Database, license, and accessibility information |
|---|---|---|---|---|
| Xavier Grau-Bové, Meritxell Antó, Iñaki Ruiz-Trillo | 2017 | Genome sequencing and assembly data from Corallochytrium, Abeoforma, Pirum and Chromosphaera | https://www.ncbi.nlm.nih.gov/bioproject/PRJNA360047 | Publicly available at the NCBI BioProject database (accession no: PRJNA360047) |
| Xavier Grau-Bové, Meritxell Antó, Iñaki Ruiz-Trillo | 2017 | Transcriptome sequencing data from Abeoforma and Chromosphaera | https://www.ncbi.nlm.nih.gov/bioproject/PRJNA360056 | Publicly available at the NCBI BioProject database (accession no: PRJNA360056) |

# References

**Abedin M**, King N. 2008. The premetazoan ancestry of cadherins. *Science* **319**:946–948. doi: 10.1126/science.1151084, PMID: 18276888

**Andrews S**. 2014. FastQC. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc [Accessed 3 May, 2016].

**Aouacheria A**, Geourjon C, Aghajari N, Navratil V, Deléage G, Lethias C, Exposito JY. 2006. 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. *Molecular Biology and Evolution* **23**:2288–2302. doi: 10.1093/molbev/msl100, PMID: 16945979

**Bao W**, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**:4–9. doi: 10.1186/s13100-015-0041-9, PMID: 26045719

**Barbosa-Morais NL**, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**:1587–1593. doi: 10.1126/science.1230612, PMID: 23258890

**Basu MK**, Carmel L, Rogozin IB, Koonin EV. 2008. Evolution of protein domain promiscuity in eukaryotes. *Genome Research* **18**:449–461. doi: 10.1101/gr.6943508, PMID: 18230802

**Basu MK**, Poliakov E, Rogozin IB. 2009. Domain mobility in proteins: functional and evolutionary implications. *Briefings in Bioinformatics* **10**:205–216. doi: 10.1093/bib/bbn057, PMID: 19151098

**Bolger AM**, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120. doi: 10.1093/bioinformatics/btu170, PMID: 24695404

**Braunschweig U**, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. 2014. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research* **24**:1774–1786. doi: 10.1101/gr.177790.114, PMID: 25258385

**Budd GE**, Jensen S. 2017. The origin of the animals and a 'Savannah' hypothesis for early bilaterian evolution. *Biological Reviews* **92**:446–473. doi: 10.1111/brv.12239, PMID: 26588818

**Burki F**, Kaplan M, Tikhonenkov DV, Zlatogursky V, Minh BQ, Radaykina LV, Smirnov A, Mylnikov AP, Keeling PJ. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proceedings of the Royal Society B: Biological Sciences* **283**: 20152802. doi: 10.1098/rspb.2015.2802, PMID: 26817772

**Bush SJ**, Chen L, Tovar-Corona JM, Urrutia AO. 2017. Alternative splicing and the evolution of phenotypic novelty. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**:20150474. doi: 10.1098/rstb.2015.0474, PMID: 27994117

**Capella-Gutiérrez S**, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973. doi: 10.1093/bioinformatics/btp348, PMID: 19505945

**Carmel L**, Wolf YI, Rogozin IB, Koonin EV. 2007. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Research* **17**:1034–1044. doi: 10.1101/gr.6438607, PMID: 17495008

**Carr M**, Suga H. 2014. The holozoan Capsaspora owczarzaki possesses a diverse complement of active transposable element families. *Genome Biology and Evolution* **6**:949–963. doi: 10.1093/gbe/evu068, PMID: 24696401

**Cavalier-Smith T**, Chao EE. 2003. Phylogeny of choanozoa, apusozoa, and other protozoa and early eukaryote megaevolution. *Journal of Molecular Evolution* **56**:540–563. doi: 10.1007/s00239-002-2424-z, PMID: 12698292

**Chikhi R**, Medvedev P. 2014. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**:31–37. doi: 10.1093/bioinformatics/btt310, PMID: 23732276

**Collins L**, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Molecular Biology and Evolution* **22**:1053–1066. doi: 10.1093/molbev/msi091, PMID: 15659557

**Core Team R**. 2015. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing.

**Cromar G**, Wong KC, Loughran N, On T, Song H, Xiong X, Zhang Z, Parkinson J. 2014. New tricks for "old" domains: how novel architectures and promiscuous hubs contributed to the organization and evolution of the ECM. *Genome Biology and Evolution* **6**:2897–2917. doi: 10.1093/gbe/evu228, PMID: 25323955

**Csárdi G**, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal Complex Sy*: 1695.

**Csuros M**, Rogozin IB, Koonin EV. 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Computational Biology* **7**:e1002150. doi: 10.1371/journal.pcbi.1002150, PMID: 21935348

**Csurös M**, Holey JA, Rogozin IB. 2007. In search of lost introns. *Bioinformatics* **23**:i87–i96. doi: 10.1093/bioinformatics/btm190, PMID: 17646350

**Csurös M**, Rogozin IB, Koonin EV. 2008. Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach. *Molecular Biology and Evolution* **25**:903–911. doi: 10.1093/molbev/msn039, PMID: 18296415

**Csurös M**. 2008. Malin: maximum likelihood analysis of intron evolution in eukaryotes. *Bioinformatics* **24**:1538–1539. doi: 10.1093/bioinformatics/btn226, PMID: 18474506

**Csurös M**. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**:1910–1912. doi: 10.1093/bioinformatics/btq315, PMID: 20551134

**Csűrös M**, Miklós I. 2006. A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. In: Apostolico A, Guerra C, Istrail S, Pevzner P. A, Waterman M (Eds). *Research in Computational Molecular Biology*. Venice: Springer. p. 206–220.

**Curtis BA**, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH, Hirakawa Y, Hopkins JF, Kuo A, Rensing SA, Schmutz J, Symeonidi A, Elias M, Eveleigh RJ, Herman EK, Klute MJ, Nakayama T, et al. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492**:59–65. doi: 10.1038/nature11681, PMID: 23201678

**Darriba D**, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**:1164–1165. doi: 10.1093/bioinformatics/btr088, PMID: 21335321

**de Mendoza A**, Sebé-Pedrós A, Ruiz-Trillo I. 2014. The evolution of the GPCR signaling system in eukaryotes: modularity, conservation, and the transition to metazoan multicellularity. *Genome Biology and Evolution* **6**:606–619. doi: 10.1093/gbe/evu038, PMID: 24567306

**de Mendoza A**, Sebé-Pedrós A, Šestak MS, Matejcic M, Torruella G, Domazet-Loso T, Ruiz-Trillo I. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *PNAS* **110**:E4858–E4866. doi: 10.1073/pnas.1311818110, PMID: 24277850

**de Mendoza A**, Suga H, Permanyer J, Irimia M, Ruiz-Trillo I. 2015. Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. *eLife* **4**:e08904. doi: 10.7554/eLife.08904, PMID: 26465111

**del Campo J**, Ruiz-Trillo I. 2013. Environmental survey meta-analysis reveals hidden diversity among unicellular opisthokonts. *Molecular Biology and Evolution* **30**:802–805. doi: 10.1093/molbev/mst006, PMID: 23329685

**Derelle R**, Torruella G, Klimeš V, Brinkmann H, Kim E, Vlček Č, Lang BF, Eliáš M. 2015. Bacterial proteins pinpoint a single eukaryotic root. *PNAS* **112**:E693–E699. doi: 10.1073/pnas.1420657112, PMID: 25646484

**Deshmukh K**, Anamika K, Srinivasan N. 2010. Evolution of domain combinations in protein kinases and its implications for functional diversity. *Progress in Biophysics and Molecular Biology* **102**:1–15. doi: 10.1016/j.pbiomolbio.2009.12.009, PMID: 20026163

**Dobin A**, Gingeras TR. 2015. Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics* **51**:11.14.1-19. doi: 10.1002/0471250953.bi1114s51, PMID: 26334920

**dos Reis M**, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PC, Yang Z. 2015. Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. *Current Biology* **25**:2939–2950. doi: 10.1016/j.cub.2015.09.066, PMID: 26603774

**Ekman D**, Björklund AK, Elofsson A. 2007. Quantification of the elevated rate of domain rearrangements in metazoa. *Journal of Molecular Biology* **372**:1337–1348. doi: 10.1016/j.jmb.2007.06.022, PMID: 17689563

**Elliott TA**, Gregory TR. 2015a. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**:20140331. doi: 10.1098/rstb.2014.0331, PMID: 26323762

**Elliott TA**, Gregory TR. 2015b. Do larger genomes contain more diverse transposable elements? *BMC* **15**:69. doi: 10.1186/s12862-015-0339-8, PMID: 25896861

**Emms DM**, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome* **16**:157. doi: 10.1186/s13059-015-0721-2, PMID: 26243257

**Exposito JY**, Larroux C, Cluzel C, Valcourt U, Lethias C, Degnan BM. 2008. Demosponge and sea anemone fibrillar collagen diversity reveals the early emergence of A/C clades and the maintenance of the modular structure of type V/XI collagens from sponge to human. *Journal of Biological Chemistry* **283**:28226–28235. doi: 10.1074/jbc.M804573200, PMID: 18697744

**Fahey B**, Degnan BM. 2012. Origin and evolution of laminin gene family diversity. *Molecular Biology and Evolution* **29**:1823–1836. doi: 10.1093/molbev/mss060, PMID: 22319142

**Fairclough SR**, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, Begovic E, Richter DJ, Russ C, Westbrook MJ, Manning G, Lang BF, Haas B, Nusbaum C, King N. 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. *Genome Biology* **14**:R15. doi: 10.1186/gb-2013-14-2-r15, PMID: 23419129

**Ferrier DEK**. 2016. Evolution of Homeobox Gene clusters in animals: the Giga-Cluster and primary vs. secondary clustering. *Frontiers in Ecology and Evolution* **4**:1–13. doi: 10.3389/fevo.2016.00036

**Fidler AL**, Darris CE, Chetyrkin SV, Pedchenko VK, Boudko SP, Brown KL, Gray Jerome W, Hudson JK, Rokas A, Hudson BG. 2017. Collagen IV and basement membrane at the evolutionary dawn of metazoan tissues. *eLife* **6**: e24176. doi: 10.7554/eLife.24176, PMID: 28418331

**Fortunato SA**, Adamski M, Ramos OM, Leininger S, Liu J, Ferrier , Adamska M. 2014. Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. *Nature* **514**:620–623. doi: 10.1038/nature13881, PMID: 25355364

**Gadd MS**, Bhati M, Jeffries CM, Langley DB, Trewhella J, Guss JM, Matthews JM. 2011. Structural basis for partial redundancy in a class of transcription factors, the LIM homeodomain proteins, in neural cell type specification. *Journal of Biological Chemistry* **286**:42971–42980. doi: 10.1074/jbc.M111.248559, PMID: 22025611

**Gaiti F**, Calcino AD, Tanurdžić M, Degnan BM. 2017a. Origin and evolution of the metazoan non-coding regulatory genome. *Developmental Biology* **427**:193–202. doi: 10.1016/j.ydbio.2016.11.013, PMID: 27880868

**Gaiti F**, Jindrich K, Fernandez-Valverde SL, Roper KE, Degnan BM, Tanurdžić M. 2017b. Landscape of histone modifications in a sponge reveals the origin of animal *cis*-regulatory complexity. *eLife* **6**:e22194. doi: 10.7554/eLife.22194, PMID: 28395144

**Glockling SL**, Marshall WL, Gleason FH. 2013. Phylogenetic interpretations and ecological potentials of the Mesomycetozoea (Ichthyosporea). *Fungal Ecology* **6**:237–247. doi: 10.1016/j.funeco.2013.03.005

**Grabherr MG**, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**:644–652. doi: 10.1038/nbt.1883, PMID: 21572440

**Grau-Bové X**, Sebé-Pedrós A, Ruiz-Trillo I. 2013. A genomic survey of HECT ubiquitin ligases in eukaryotes reveals independent expansions of the HECT system in several lineages. *Genome* **5**:833–847. doi: 10.1093/gbe/evt052, PMID: 23563970

**Grau-Bové X**, Sebé-Pedrós A, Ruiz-Trillo I. 2015. The eukaryotic ancestor had a complex ubiquitin signaling system of archaeal origin. *Molecular Biology and Evolution* **32**:726–739. doi: 10.1093/molbev/msu334, PMID: 25525215

**Grosberg RK**, Strathmann RR. 2007. The evolution of Multicellularity: a Minor Major transition? *Annual Review of Ecology, Evolution, and Systematics* **38**:621–654. doi: 10.1146/annurev.ecolsys.36.102403.114735

**Guindon S**, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**: 307–321. doi: 10.1093/sysbio/syq010, PMID: 20525638

**Gurevich A**, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**:1072–1075. doi: 10.1093/bioinformatics/btt086, PMID: 23422339

**Haas BJ**, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**:5654–5666. doi: 10.1093/nar/gkg770, PMID: 14500829

**Haas BJ**, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**:R7. doi: 10.1186/gb-2008-9-1-r7, PMID: 18190707

**He D**, Fiz-Palacios O, Fu CJ, Fehling J, Tsai CC, Baldauf SL. 2014. An alternative root for the eukaryote tree of life. *Current Biology* **24**:465–470. doi: 10.1016/j.cub.2014.01.036, PMID: 24508168

**He F**, Jacobson A. 2015. Nonsense-mediated mRNA decay: degradation of defective transcripts is only part of the story. *Annual Review of Genetics* **49**:339–366. doi: 10.1146/annurev-genet-112414-054639, PMID: 26436458

**Heino J**. 2007. The collagen family members as cell adhesion proteins. *BioEssays* **29**:1001–1010. doi: 10.1002/bies.20636, PMID: 17876790

**HMMER**. 2015. HMMER. http://hmmer.org/ [Accessed May 24, 2017 ].

**Holland** , Booth , Bruford EA. 2007. Classification and nomenclature of all human homeobox genes. *BMC* **5**:47. doi: 10.1186/1741-7007-5-47, PMID: 17963489

**Holland** . 2013. Evolution of homeobox genes. *Wiley* **2**:31–45. doi: 10.1002/wdev.78, PMID: 23799629

**Hynes RO**. 2012. The evolution of metazoan extracellular matrix. *Cell* **196**:671–679. doi: 10.1083/jcb.201109041, PMID: 22431747

**Irimia M**, Penny D, Roy SW. 2007. Coevolution of genomic intron number and splice sites. *Trends* **23**:321–325. doi: 10.1016/j.tig.2007.04.001, PMID: 17442445

**Irimia M**, Roy SW. 2014. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harbor Perspectives in Biology* **6**:a016071. doi: 10.1101/cshperspect.a016071, PMID: 24890509

**Irimia M**, Rukov JL, Roy SW, Vinther J, Garcia-Fernandez J. 2009. Quantitative regulation of alternative splicing in evolution and development. *BioEssays* **31**:40–50. doi: 10.1002/bies.080092, PMID: 19154001

**Irimia M**, Tena JJ, Alexis MS, Fernandez- A, Maeso I, Bogdanovic O, Calle-Mustienes E, Roy SW, -Skarmeta JL, Fraser HB. 2012. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome* **22**:2356–2367. doi: 10.1101/gr.139725.112, PMID: 22722344

**Itoh M**, Nacher JC, Kuma K, Goto S, Kanehisa M. 2007. Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome* **8**:R121. doi: 10.1186/gb-2007-8-6-r121, PMID: 17588271

**French-Italian Public Consortium for Grapevine Genome Characterization**, Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**:463–467. doi: 10.1038/nature06148, PMID: 17721507

**Kajitani R**, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* **24**:1384–1395. doi: 10.1101/gr.170720.113, PMID: 24755901

**Katoh K**, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–780. doi: 10.1093/molbev/mst010, PMID: 23329690

**Keller O**, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**:757–763. doi: 10.1093/bioinformatics/btr010, PMID: 21216780

**Keller O**, Odronitz F, Stanke M, Kollmar M, Waack S. 2008. Scipio: protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics* **9**:278. doi: 10.1186/1471-2105-9-278, PMID: 18554390

**Kent WJ**. 2002. BLAT-the BLAST-like alignment tool. *Genome* **12**:656–664. doi: 10.1101/gr.229202, PMID: 11932250

**Kerényi Z**, Mérai Z, Hiripi L, Benkovics A, Gyula P, Lacomme C, Barta E, Nagy F, Silhavy D. 2008. Inter-kingdom conservation of mechanism of nonsense-mediated mRNA decay. *EMBO* **27**:1585–1595. doi: 10.1038/emboj.2008.88, PMID: 18451801

**King N**, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, Marr M, Pincus D, Putnam N, Rokas A, Wright KJ, Zuzow R, Dirks W, Good M, Goodstein D, Lemons D, et al. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* **451**:783–788. doi: 10.1038/nature06617, PMID: 18273011

**Koonin EV**, Csuros M, Rogozin IB. 2013. Whence genes in pieces: of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. *Wiley RNA* **4**:93–105. doi: 10.1002/wrna.1143, PMID: 23139082

**Korf I**. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**:59. doi: 10.1186/1471-2105-5-59, PMID: 15144565

**Krzywinski M**, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA, Spring C. 2009. Circos: information aesthetic for comparative genomics. *Genome* **19**:1639–1645. doi: 10.1101/gr.092759.109, PMID: 19541911

**Lareau LF**, Brooks AN, Soergel DA, Meng Q, Brenner SE. 2007. The coupling of alternative splicing and nonsense-mediated mRNA decay. *Advances in Experimental Medicine and Biology* **623**:190–211. PMID: 18380348

**Lartillot N**, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21**:1095–1109. doi: 10.1093/molbev/msh112, PMID: 15014145

**Lartillot N**, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology* **62**:611–615. doi: 10.1093/sysbio/syt022, PMID: 23564032

**Le Hir H**, Nott A, Moore MJ. 2003. How introns influence and enhance eukaryotic gene expression. *Trends* **28**:215–220. doi: 10.1016/S0968-0004(03)00052-5, PMID: 12713906

**Leonard G**, Richards TA. 2012. Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. *PNAS* **109**:21402–21407. doi: 10.1073/pnas.1210909110, PMID: 23236161

**Liu M**, Walch H, Wu S, Grigoriev A. 2005. Significant expansion of exon-bordering protein domains during animal proteome evolution. *Nucleic Acids* **33**:95–105. doi: 10.1093/nar/gki152, PMID: 15640447

**Liu Y**, Steenkamp ET, Brinkmann H, Forget L, Philippe H, Lang BF. 2009. Phylogenomic analyses predict sistergroup relationship of nucleariids and fungi and paraphyly of zygomycetes with significant support. *BMC* **9**:272. doi: 10.1186/1471-2148-9-272, PMID: 19939264

**Lomsadze A**, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids* **33**:6494–6506. doi: 10.1093/nar/gki937, PMID: 16314312

**Luo R**, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**:18. doi: 10.1186/2047-217X-1-18, PMID: 23587118

**Lynch M**, Conery JS. 2003. The origins of genome complexity. *Science* **302**:1401–1404. doi: 10.1126/science.1089370, PMID: 14631042

**Lynch M**. 2002. Intron evolution as a population-genetic process. *PNAS* **99**:6118–6123. doi: 10.1073/pnas.092595699, PMID: 11983904

**Lynch M**. 2006. The origins of eukaryotic gene structure. *Molecular Biology and Evolution* **23**:450–468. doi: 10.1093/molbev/msj050, PMID: 16280547

**Lynch M**. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *PNAS* **104 Suppl 1**: 8597–8604. doi: 10.1073/pnas.0702207104, PMID: 17494740

**Manning G**, Young SL, Miller WT, Zhai Y. 2008. The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. *PNAS* **105**:9674–9679. doi: 10.1073/pnas.0801314105, PMID: 18621719

**Marshall WL**, Berbee ML. 2010. Population-level analyses indirectly reveal cryptic sex and life history traits of *Pseudoperkinsus tapetis* (Ichthyosporea, Opisthokonta): a unicellular relative of the animals. *Molecular Biology and Evolution* **27**:2014–2026. doi: 10.1093/molbev/msq078, PMID: 20360212

**Marshall WL**, Celio G, McLaughlin DJ, Berbee ML. 2008. Multiple isolations of a culturable, motile Ichthyosporean (Mesomycetozoa, Opisthokonta), Creolimax fragrantissima n. gen., n. sp., from marine invertebrate digestive tracts. *Protist* **159**:415–433. doi: 10.1016/j.protis.2008.03.003, PMID: 18539526

**McGuire AM**, Pearson MD, Neafsey DE, Galagan JE. 2008. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome* **9**:R50. doi: 10.1186/gb-2008-9-3-r50, PMID: 18321378

**Mendoza L**, Taylor JW, Ajello L. 2002. The class mesomycetozoea: a heterogeneous group of microorganisms at the animal-fungal boundary. *Annual Review of Microbiology* **56**:315–344. doi: 10.1146/annurev.micro.56.012302.160950, PMID: 12142489

**Michael TP**. 2014. Plant genome size variation: bloating and purging DNA. *Genomics* **13**:308–317. doi: 10.1093/bfgp/elu005, PMID: 24651721

**Minh BQ**, Nguyen MA, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution* **30**:1188–1195. doi: 10.1093/molbev/mst024, PMID: 23418397

**Mitchell A**, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJ, Redaschi N, Rivoire C, et al. 2015. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research* **43**:D213–D221. doi: 10.1093/nar/gku1243, PMID: 25428371

**Moroz LL**, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov E, Buckley KM, Ptitsyn A, Reshetov D, Mukherjee K, Moroz TP, Bobkova Y, Yu F, Kapitonov VV, Jurka J, Bobkov YV, Swore JJ, et al. 2014. The ctenophore genome and the evolutionary origins of neural systems. *Nature* **510**: 109–114. doi: 10.1038/nature13400, PMID: 24847885

**Newman ME**, Girvan M. 2004. Finding and evaluating community structure in networks. *Physical Review E* **69**: 26113. doi: 10.1103/PhysRevE.69.026113, PMID: 14995526

**Nguyen LT**, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**:268–274. doi: 10.1093/molbev/msu300, PMID: 25371430

**Nichols SA**, Roberts BW, Richter DJ, Fairclough SR, King N. 2012. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/$\beta$-catenin complex. *PNAS* **109**:13046–13051. doi: 10.1073/pnas.1120685109, PMID: 22837400

**Nikolenko SI**, Korobeynikov AI, Alekseyev MA. 2013. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* **14 Suppl 1**:S7–11. doi: 10.1186/1471-2164-14-S1-S7, PMID: 23368723

**Nilsen TW**, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**:457–463. doi: 10.1038/nature08909, PMID: 20110989

**Nurk S**, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelsky A, Pyshkin A, Sirotkin A, Sirotkin Y. 2013. Assembling genomes and mini-metagenomes from highly chimeric reads. *Lect. Notes Comput Sci* **7821**:158–170. doi: 10.1007/978-3-642-37195-0_13

**Paradis E**, Claude J, Strimmer K. 2004. APE: of and in R language. *Bioinformatics* **20**:289–290. doi: 10.1093/bioinformatics/btg412, PMID: 14734327

**Patterson DJ**, Nygaard K, Steinberg G, Turley CM. 1993. Heterotrophic flagellates and other protists associated with oceanic detritus throughout the water column in the mid North Atlantic. *Journal of the Marine Biological Association of the United Kingdom* **73**:67. doi: 10.1017/S0025315400032653

**Petrov DA**, Lozovskaya ER, Hartl DL. 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**:346–349. doi: 10.1038/384346a0, PMID: 8934517

**Plass M**, Agirre E, Reyes D, Camara F, Eyras E. 2008. Co-evolution of the branch site and SR proteins in eukaryotes. *Trends* **24**:590–594. doi: 10.1016/j.tig.2008.10.004, PMID: 18992956

**Punta M**, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. 2012. The Pfam protein families database. *Nucleic Acids Research* **40**:D290–D301. doi: 10.1093/nar/gkr1065, PMID: 22127870

**Putnam NH**, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**:86–94. doi: 10.1126/science.1139158, PMID: 17615350

**Quang leS**, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**:2317–2323. doi: 10.1093/bioinformatics/btn445, PMID: 18718941

**Quinlan AR**, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842. doi: 10.1093/bioinformatics/btq033, PMID: 20110278

**Raghukumar S**. 1987. Occurrence of the Thraustochytrid, *Corallochytrium limacisporum* gen. et sp. nov. in the Coral reef Lagoons of the Lakshadweep Islands in the Arabian Sea. *Botanica Marina* **30**:83–89. doi: 10.1515/botm.1987.30.1.83

Richter DJ, King N. 2013. The genomic and cellular foundations of animal origins. *Annual Review of Genetics* **47**: 509–537. doi: 10.1146/annurev-genet-111212-133456, PMID: 24050174

Rogozin IB, Carmel L, Csuros M, Koonin EV. 2012. Origin and evolution of spliceosomal introns. *Biology Direct* **7**: 11. doi: 10.1186/1745-6150-7-11, PMID: 22507701

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574. doi: 10.1093/bioinformatics/btg180, PMID: 12912839

Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF. 2008. A phylogenomic investigation into the origin of metazoa. *Molecular Biology and Evolution* **25**:664–672. doi: 10.1093/molbev/msn006, PMID: 18184723

Sanford JR, Ellis J, Cáceres JF. 2005. Multiple roles of arginine/serine-rich splicing factors in RNA processing. *Biochemical Society Transactions* **33**:443–446. doi: 10.1042/BST0330443, PMID: 15916537

Sebé-Pedrós A, Ballaré C, Parra-Acero H, Chiva C, Tena JJ, Sabidó E, Gómez-Skarmeta JL, Di Croce L, Ruiz-Trillo I. 2016b. The Dynamic Regulatory Genome of *Capsaspora* and the Origin of Animal Multicellularity. *Cell* **165**:1224–1237. doi: 10.1016/j.cell.2016.03.034, PMID: 27114036

Sebé-Pedrós A, de Mendoza A, Lang BF, Degnan BM, Ruiz-Trillo I. 2011. Unexpected repertoire of metazoan transcription factors in the unicellular holozoan Capsaspora owczarzaki. *Molecular Biology and Evolution* **28**: 1241–1254. doi: 10.1093/molbev/msq309, PMID: 21087945

Sebé-Pedrós A, Degnan BM, Ruiz-Trillo I. 2017. The origin of Metazoa: a unicellular perspective. *Nature Reviews Genetics* **18**:498–512. doi: 10.1038/nrg.2017.21, PMID: 28479598

Sebé-Pedrós A, Irimia M, Del Campo J, Parra-Acero H, Russ C, Nusbaum C, Blencowe BJ, Ruiz-Trillo I. 2013. Regulated aggregative multicellularity in a close unicellular relative of metazoa. *eLife* **2**:e01287. doi: 10.7554/eLife.01287, PMID: 24368732

Sebé-Pedrós A, Peña MI, Capella-Gutiérrez S, Antó M, Gabaldón T, Ruiz-Trillo I, Sabidó E. 2016a. High-Throughput Proteomics Reveals the Unicellular Roots of Animal Phosphosignaling and Cell Differentiation. *Developmental Cell* **39**:186–197. doi: 10.1016/j.devcel.2016.09.019, PMID: 27746046

Sebé-Pedrós A, Roger AJ, Lang FB, King N, Ruiz-Trillo I. 2010. Ancient origin of the integrin-mediated adhesion and signaling machinery. *PNAS* **107**:10142–10147. doi: 10.1073/pnas.1002257107, PMID: 20479219

Seo HC, Kube M, Edvardsen RB, Jensen MF, Beck A, Spriet E, Gorsky G, Thompson EM, Lehrach H, Reinhardt R, Chourrout D. 2001. Miniature genome in the marine chordate *Oikopleura dioica*. *Science* **294**:2506. doi: 10.1126/science.294.5551.2506, PMID: 11752568

Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, Takeuchi T, Hisata K, Tanaka M, Fujiwara M, Hamada M, Seidi A, Fujie M, Usami T, Goto H, Yamasaki S, Arakaki N, Suzuki Y, Sugano S, Toyoda A, et al. 2013. Draft assembly of the Symbiodinium minutum nuclear genome reveals dinoflagellate gene structure. *Current Biology* **23**:1399–1408. doi: 10.1016/j.cub.2013.05.062, PMID: 23850284

Simakov O, Kawashima T. 2017. Independent evolution of genomic characters during major metazoan transitions. *Developmental Biology* **427**:179–192. doi: 10.1016/j.ydbio.2016.11.012, PMID: 27890449

Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo DH, Larsson T, Lv J, Arendt D, Savage R, Osoegawa K, de Jong P, Grimwood J, Chapman JA, Shapiro H, Aerts A, Otillar RP, Terry AY, Boore JL, et al. 2013. Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**:526–531. doi: 10.1038/nature11696, PMID: 23254933

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:3210–3212. doi: 10.1093/bioinformatics/btv351, PMID: 26059717

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, Lapébie P, Corre E, Delsuc F, King N, Wörheide G, Manuel M. 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Current Biology* **27**:958–967. doi: 10.1016/j.cub.2017.02.031, PMID: 28318975

Simmons DK, Pang K, Martindale MQ. 2012. Lim homeobox genes in the Ctenophore *Mnemiopsis leidyi*: the evolution of neural cell type specification. *EvoDevo* **3**:2. doi: 10.1186/2041-9139-3-2, PMID: 22239757

Smit A, Hubley R, Green P. 2015. RepatMasker. *Repeat Masker*. 4.0.

Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, Signorovitch AY, Moreno MA, Kamm K, Grimwood J, Schmutz J, Shapiro H, Grigoriev IV, Buss LW, Schierwater B, Dellaporta SL, et al. 2008. The *Trichoplax* genome and the nature of placozoans. *Nature* **454**:955–960. doi: 10.1038/nature07191, PMID: 18719581

Srivastava M, Larroux C, Lu DR, Mohanty K, Chapman J, Degnan BM, Rokhsar DS. 2010b. Early evolution of the LIM homeobox gene family. *BMC* **8**:4. doi: 10.1186/1741-7007-8-4, PMID: 20082688

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier ME, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U, Larroux C, Putnam NH, Stanke M, Adamska M, Darling A, Degnan SM, Oakley TH, Plachetzki DC, Zhai Y, Adamski M, et al. 2010a. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* **466**:720–726. doi: 10.1038/nature09201, PMID: 20686567

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313. doi: 10.1093/bioinformatics/btu033, PMID: 24451623

Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E, Carr M, Kerner P, Vervoort M, Sánchez-Pons N, Torruella G, Derelle R, Manning G, Lang BF, Russ C, Haas BJ, Roger AJ, Nusbaum C, Ruiz-Trillo I. 2013. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. *Nature Communications* **4**: 2325. doi: 10.1038/ncomms3325

**Suga H**, Dacre M, de Mendoza A, Shalchian-Tabrizi K, Manning G, Ruiz-Trillo I. 2012. Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. *Science Signaling* **5**:ra35. doi: 10.1126/scisignal.2002733, PMID: 22550341

**Suga H**, Ruiz-Trillo I. 2013. Development of ichthyosporeans sheds light on the origin of metazoan multicellularity. *Developmental Biology* **377**:284–292. doi: 10.1016/j.ydbio.2013.01.009, PMID: 23333946

**Thor S**, Andersson SG, Tomlinson A, Thomas JB. 1999. A LIM-homeodomain combinatorial code for motor-neuron pathway selection. *Nature* **397**:76–80. doi: 10.1038/16275, PMID: 9892357

**Tordai H**, Nagy A, Farkas K, Bányai L, Patthy L. 2005. Modules, multidomain proteins and organismic complexity. *FEBS* **272**:5064–5078. doi: 10.1111/j.1742-4658.2005.04917.x, PMID: 16176277

**Torruella G**, de Mendoza A, Grau-Bové X, Antó M, Chaplin MA, del Campo J, Eme L, Pérez-Cordón G, Whipps CM, Nichols KM, Paley R, Roger AJ, Sitjà-Bobadilla A, Donachie S, Ruiz-Trillo I. 2015. Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Current Biology* **25**:2404–2410. doi: 10.1016/j.cub.2015.07.053, PMID: 26365255

**Torruella G**, Derelle R, Paps J, Lang BF, Roger AJ, Shalchian-Tabrizi K, Ruiz-Trillo I. 2012. Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Molecular Biology and Evolution* **29**:531–544. doi: 10.1093/molbev/msr185, PMID: 21771718

**Treangen TJ**, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13**:36–46. doi: 10.1038/nrg3117, PMID: 22124482

**Warnes GR**, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M. 2016. Gplots: various R programming tools for Plotting Data. https://CRAN.R-project.org/package=gplots

**Weirauch MT**, Hughes TR. 2011. Hughes T. R (Ed). *A Handbook of Transcription Factors*. Dordrecht: Springer.

**Whelan NV**, Kocot KM, Moroz LL, Halanych KM. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. *PNAS* **112**:5773–5778. doi: 10.1073/pnas.1503453112, PMID: 25902535

**Wu TD**, Reeder J, Lawrence M, Becker G, Brauer MJ. 2016. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, Accuracy, and functionality. *Methods in Molecular Biology* **1418**:283–334. doi: 10.1007/978-1-4939-3578-9_15, PMID: 27008021

**Xie X**, Jin J, Mao Y. 2011. Evolutionary versatility of eukaryotic protein domains revealed by their bigram networks. *BMC Evolutionary Biology* **11**:242. doi: 10.1186/1471-2148-11-242, PMID: 21849086

**Zhong YF**, Holland PW. 2011. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evolution & Development* **13**:567–568. doi: 10.1111/j.1525-142X.2011.00513.x, PMID: 23016940

**Zmasek CM**, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome Biology* **12**:R4. doi: 10.1186/gb-2011-12-1-r4, PMID: 21241503

**Zmasek CM**, Godzik A. 2012. This Déjà vu feeling–analysis of multidomain protein evolution in eukaryotic genomes. *PLoS Computational Biology* **8**:e1002701. doi: 10.1371/journal.pcbi.1002701, PMID: 23166479