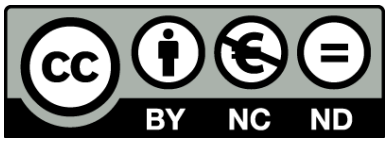
UNIVERSITAT DE BARCELONA

L´origen de la multicel·lularitat en animals: una aproximació genómica

The origin of multicellularity in animals: a genomic approach

Francesc Xavier Grau Bové



Programa de Doctorat de Genètica
Departament de Genètica
Facultat de Biologia
Universitat de Barcelona

L'origen de la multicel·lularitat en animals: una aproximació genòmica

The origin of multicellularity in animals: a genomic approach

Memòria de tesi realitzada per Francesc Xavier Grau Bové a l'Institut de Biologia Evolutiva, per tal d'optar al títol de Doctor per la Universitat de Barcelona.

Dr. Iñaki Ruiz-Trillo
Director de la tesi

Dra. Marta Riutort
Tutora

Francesc Xavier Grau Bové
Autor

Barcelona, 27 d'abril de 2017

Cover picture: electron micrographs of *Corallochytrium limacisporum*, Raghukumar 1989

Pel padrí i per la iaia.

Table of contents

List of figures

List of tables

1. Introduction

A genomic perspective of the transition to multicellularity in Metazoa

We used to think that if we knew one, we knew two, because one and one are two. We are finding that we must learn a great deal more about 'and'.

Arthur Eddington

Multicellularity is a recurrent theme in evolution: it has independently appeared more than 25 times in Eukaryota, Bacteria and Archaea. The most widely known and studied multicellular organisms are animals, fungi and plants, but the list goes on to include organisms such as the mat structures formed by cyanobacteria, the large brown and red algae, the volvocine green algae, and a myriad of aggregative multicellular amoebas and bacteria. The evolutionary consequences of multicellularity contribute dearly to the diversity of life forms that inhabit the Earth.

In the present introduction I will start by defining what multicellularity is, when and how it appeared, and which evolutionary properties it entails. Secondly, I will focus on the specific advent of multicellularity in Metazoa, and introduce the closest unicellular relatives of animals on which my work has been based: choanoflagellates, filastereans, ichthyosporeans and *Corallochytrium*. Finally, I will examine the current state of research in genomics concerning the origin of animals. In this last section, I will put particular stress on the benefits of studying genomes from unicellular animal relatives: such comparative analyses allow us to reconstruct the genome content and dynamics of the unicellular premetazoan that ultimately underwent the transition to multicellularity.

1.1. What is multicellularity?

Multicellularity is a property of living systems in which single cells assemble and form a physically integrated and functionally coordinated organism. Such collectives of cells behave as an individual that can be described both by the properties of its components and the properties of the multicellular entity (Michod 2007), as the collective acquires the properties of a Darwinian individual that participates in the evolutionary process as a whole (Rainey and De Monte 2014).

The origin of multicellularity falls within the ‘major evolutionary transitions’ as characterized by John Maynard Smith and Eörs Szathmáry: a suite of radical innovations that lead from pre-existing replicating entities to complex, higher-level living entities, which collectively assume functions of their lower-level building blocks (Szathmáry and Smith 1995). For example, the emergence of cells from replicating molecules underlies the origin of life; independent replicating molecules were joined in chromosomes; endosymbiosis between prokaryotes lead to the origin of eukaryotes; acquisition of plastids lead to synergetic energy production in eukaryotes; and multicellular organisms can appear from cell populations if they acquire regulated differentiation and cell division (Szathmáry 2015). Therefore, all life forms are a hierarchy of organization levels corresponding to the major transitions they have undergone, which have as a common theme some sort of transition to a higher level of individuality (Michod 2007; Rainey and De Monte 2014).

Major evolutionary transitions are contingent: not all living systems have been subject to all of them as there is no reason to expect continuous increases in complexity (Szathmáry and Smith 1995). But at the same time, some transitions are recurrent: they occur multiple times in evolution and their effects are similar. Multicellularity is a clear example of both these properties. In the section *1.1.1. The plurality of multicellularity*, I will examine the variety of multicellular organizations that has evolved since the origin of life on Earth. Whenever multicellularity appears in any evolutionary lineage, it entails a variety of selective advantages, from enhanced predatory

capacities for heterotrophs to larger lit surfaces for autotrophs. At the same time, as multicellularity represents a transition to a new form of individuality, they are bound to be affected by conflicting genetic pressures between constituent cells and the collective. The selective consequences and the adaptive value of multicellular life are considered in the section *1.1.2. Evolutionary consequences of multicellularity*.

1.1.1. The plurality of multicellularity

1.1.1.1. A palaeontological perspective of multicellular transitions

Multicellular organisms have independently appeared at least 25 times during the course of evolution of life on Earth, both within eukaryotes (Figure 1) and prokaryotes (Grosberg and Strathmann 2007). From a historical viewpoint, the first evidence of multicellular assemblies are the mats of cyanobacteria-like prokaryotes dated at 3 to 3.5 billion years ago (Gya) (Schopf 1993), for which cell differentiation evidence exists from ~2 Gya (Tomitani *et al.* 2006). However, it is after the origin of Eukaryota (1-1.9 Gya according to molecular clocks, *cf. Eme et al. 2014*; or 2.1 Gya according to the fossil record, *cf. Knoll 2014*) that most of the known multicellular lineages appear within this group (Grosberg and Strathmann 2007; Knoll and Hewitt 2011).

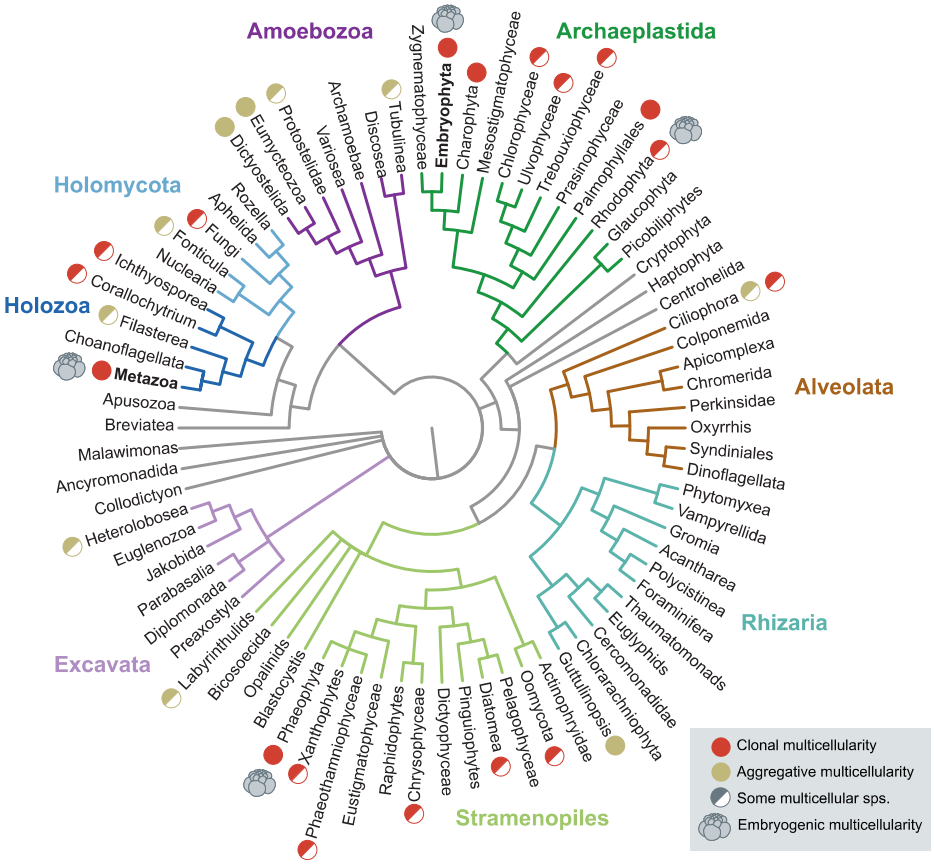


Figure 1. Overview of the eukaryotic tree of life, highlighting the multiple origins of multicellularity and their modalities: aggregative or clonal, and embryogenic/complex.

Various explanations have been proposed to account for the more frequent occurrence of multicellularity in eukaryotes compared to prokaryotes, ranging to bioenergetics to cell and genome biology. For example, energy production by synergistic mitochondria in eukaryotes has been proposed to be key to sustain complex life traits like multicellularity (Lane and Martin 2010) – although this hypothesis has been recently challenged: Booth and Doolittle (2015) emphasized the lack of a clear correlation between phenotypic complexity, genome content and and energy output in eukaryotes; and Lynch and Marinov (2017) have cast doubt on the alleged higher energetic efficiency of eukaryotes. Another possible explanation is the presence of cytoskeleton and membrane systems in eukaryotes, which allow the cell to have a more plastic and dynamic morphology, responsive to external stimuli, and acquire sizes and structures absent in prokaryotes – an ultimately key trait to sustain cell differentiation (Knoll 2011). Finally, the selective pressures on a fast genome replication in prokaryotes leaves very little evolutionary space for innovation in terms of complex genome regulation, thus promoting stream-lined genomes (Lynch 2006b). In contrast, eukaryotic genomes have multiple replication sites, which decreases the burden of accumulating regulatory elements within the genome, both at the coding and non-coding levels (Knoll 2011).

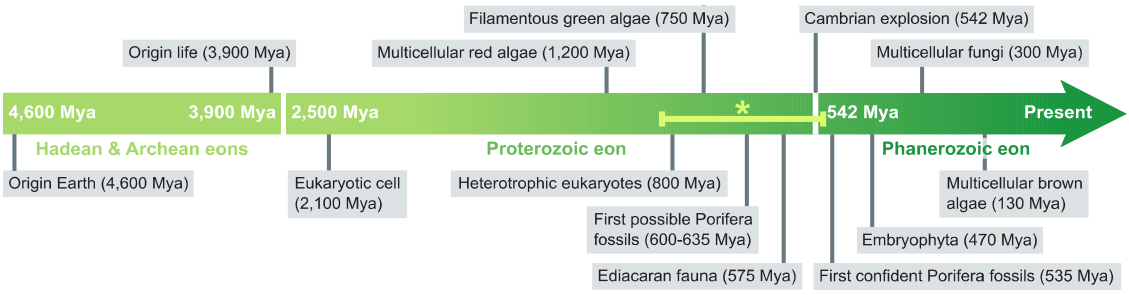


Figure 2. Time-line of origin of major multicellular eukaryotic lineages. The asterisk and bar within the time-line mark the period spanned in Figure 3.

Whichever were the reasons for the relative success of multicellularity in eukaryotes, the fossil record holds multiple evidences of multicellularity in 'crown group' eukaryotes (*i.e.*, belonging to extant eukaryotic lineages) in the Proterozoic eon (Knoll and Hewitt 2011; Eme *et al.* 2014; Knoll 2014). The first fossil of an unambiguous crown multicellular eukaryote is the bangiophyte red algae *Bangiomorpha pubescens* (Figure 2), 1,200 million years ago (Mya), which presented simple filamentous multicellularity and cell differentiation (Butterfield 2000). Other putative crown rhodophytes have been described to pre-date *Bangiophorma* by at least 400 million years, but could not be confidently classified into any extant group (Bengtson *et al.* 2017). The next clear multicellular fossils are florideophyte red algae, with more complex three-dimensional morphologies, from 600-560 Mya (from Doushantuo); simple filamentous green algae 750-800 Mya (including); and larger green algae fossils 550-570 Mya (Knoll and Hewitt 2011). The first fossils of large, complex green algae emerged later, in the Phanerozoic eon: three-dimensional coenocytic green algae during the Cambrian and 'stem' embryophytes (land plants) in the Ordovician (470 Mya; Wellman and Gray 2000). The radiation of 'crown' embryophytes had already occurred 400 Mya (Knoll 2011; Becker 2013). In contrast, multicellular phaeophytes (brown algae), Volvocales

and fungi appeared in more recent times: ~130 Mya for brown algae, according to molecular clock estimates (Silberfeld *et al.* 2010), ~200 Mya for volvocine algae (Herron *et al.* 2009); and ~300 Mya for basidiomycete and ascomycete Fungi (Lücking *et al.* 2009; Stajich *et al.* 2009).

The specific moment when multicellular Metazoa evolved is subject to intense debate (Knoll and Hewitt 2011; Antcliffe *et al.* 2014; Budd and Jensen 2015). The earliest animal fossils (Figure 3) date back to the Ediacaran period (at the end of the Neoproterozoic era), in a series of fossils dated around 600 Ma. This early fossil record includes a well-established crown Porifera, *Protohertzina anabarica* (Soltanieh, 535 Mya; Hamdi et al. 1989) and some earlier but unclear or disputed sponges, like *Eocyathispongia qiania* (Doushantuo, 600 Mya; Yin et al. (2015) and pre-Marinoan specimens from South Australia (635 Mya; Maloof et al. 2010; Antcliffe et al. 2014). The record of stem metazoans is slightly older than the sponge described by Hamdi et al. (1989), including specimens from the Avalonian assemblage (579 Mya), just after the Gaskiers glacial period (583 Mya; Narbonne and Gehling 2003; Narbonne 2005). The interpretation of the early metazoan fossil record is deeply intermingled with controversies regarding extant metazoan phylogenomics (Telford *et al.* 2015). For example, even if phylogenetic analyses often place Porifera as the earliest-branching metazoan lineage (Nosenko *et al.* 2013; Pisani *et al.* 2015; Simion *et al.* 2017), clear sponge fossils are scarce in the Ediacaran and Cryogenian (pre-Ediacaran) periods (Knoll 2011; Antcliffe *et al.* 2014). Ctenophora, the other contentious earliest-branching metazoan group (Nosenko *et al.* 2013; Ryan *et al.* 2013; Moroz *et al.* 2014; Whelan *et al.* 2015; Shen *et al.* 2017), are also relatively absent from the early palaeontological records, with only a few confident fossils dated 540-580 Mya, *e.g.* the *Eoandromeda* genus (Tang *et al.* 2011; Ou *et al.* 2015).

However, molecular clock estimates of the early animal origins, obtained from phylogenomic and fossil data of 'crown' Metazoa, have frequently yielded far older estimates than any individual fossil, at 700-800 (Erwin *et al.* 2011), 755-838 (Sperling *et al.* 2010) or 650-833 Mya (dos Reis *et al.* 2015). If that were the case, 'crown' animals would have originated in the Cryogenian, before the Ediacaran and Cambrian fossil fauna; and each of the extant lineages would have radiated around this period as well (Eumetazoa: 626-746 Mya; Bilateria 596-688 Mya; Deuterostomia: 587-662 Mya; Protostomia: 578-653 Mya; dos Reis et al. 2015). This early-origin scenario would unlink the palaeontological and phylogenomic discussion about the nature of the first animals inasmuch the occurrence of specific fossilized clades is concerned (their usefulness for molecular clock calibration would be, of course, still of great importance). In parallel, it would emphasize the evolutionary-ecological relationships between 'stem' and 'crown' metazoans in shaping extant animal diversity (Budd and Jensen 2015; Telford *et al.* 2015).

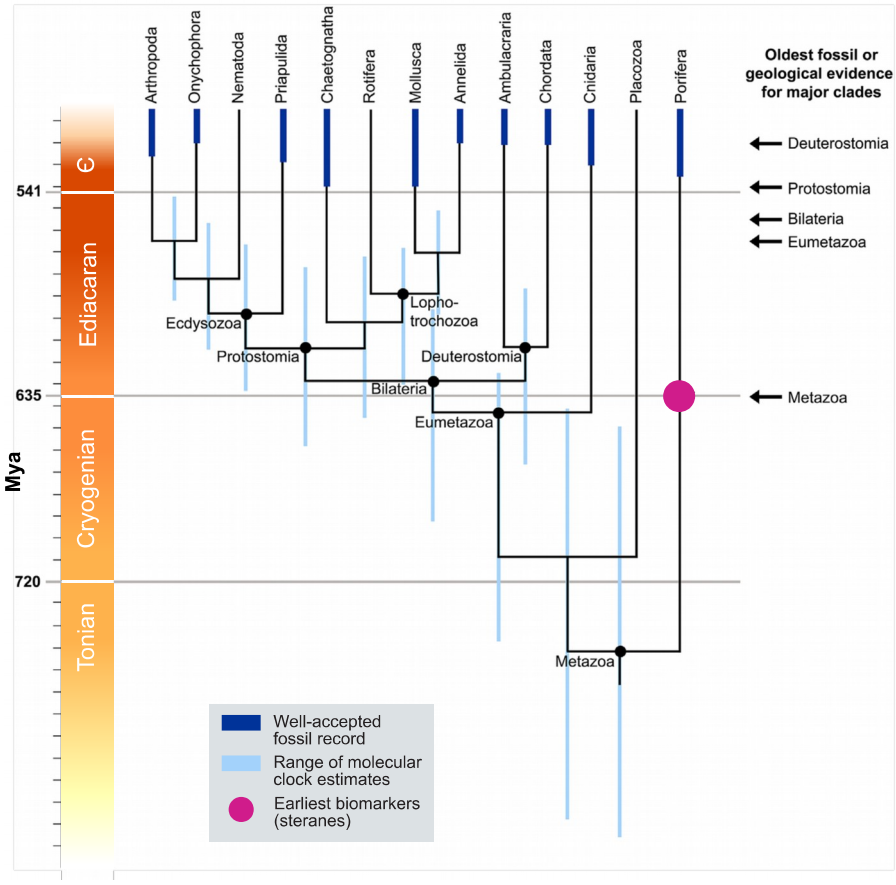


Figure 3. Summary of the incongruence between the molecular clock estimates, accepted fossil record and other biomarkers regarding animal origins. From Cunningham et al. (2016).

1.1.1.2. Types of multicellularity and their occurrence in nature

Key to the understanding of this transition is to discern the different types of multicellularity. First, it is useful to distinguish 'simple' from 'complex' multicellular organisms, as defined by Andrew H. Knoll (Knoll and Hewitt 2011; Knoll 2011). In this regard, 'simple multicellularity' consists of any persistent assemble of cells (filaments, clusters, sheets, mats) that arise via mitotic cell division. It can include germ-soma differentiation, cell-to-cell communication and give rise to reproducible morphologies. Examples of simple multicellular organisms can be found both in eukaryotes (Figure 1) and prokaryotes: the cyanobacterial mats, the colonies of choanoflagellates, the various green algae structures like Volvocales (spherical colonies) or Charales (filamentous), the aggregative structures of amoebozoan slime molds, *Fonticula* or the ciliate *Sorogena* (Bonner 2000a; Knoll and Hewitt 2011). It is often the case that simple multicellular structures take the form of undifferentiated cell assemblages, like in choanoflagellates (Fairclough *et al.* 2010); but *Volvox* exhibits a complete germ-soma differentiation with division of labour (Umen 2014).

'Complex' multicellularity, in contrast, is a relatively rarer evolutionary event only found in eukaryotes, where this feature has appeared 6-7 times: in metazoans, embryophytes (land plants), florideophytes (red algae), laminarialeans (brown algae), and 2-3 times in basidiomycetes and ascomycetes (Knoll 2011). All these lineages exhibit cell adhesion and communication, as well as a

regulated developmental program that leads to a three-dimensional body organisation with cell types and, frequently, tissues. In addition, animals, embryophytes, florideophytes and laminarialeans have embryogenic developmental processes (Figure 1) underlying their complex body plans (Mshigeni and Lorri 1977; Bouget *et al.* 1998; Xie *et al.* 2010). The distinction between internal and external environments implies the necessity of establishing transport mechanisms for nutrients, oxygen and cellular signals (Knoll and Hewitt 2011), as well as facilitating stable conditions in which morphogenetic cell differentiation programs can operate (Bonner 2000b; c; Sebé-Pedrós *et al.* 2017).

Another relevant distinction is that of clonal *versus* aggregative multicellularity (Bonner 2000a). In clonal organisms, a single progenitor cell divides, and its progeny remains attached in a cluster of genetically identical cells. Plants, animals and all other complex multicellular lineages fall within this category; which also includes simpler green algae like *Volvox*, *Ulva* and some diatoms (Bonner 2000a). In aggregative multicellularity, single cells associate to form a cluster that can contain genetically disparate individuals. Despite being more transient in nature than clonal organisms, aggregates can also exhibit both temporal and spatial cell type differentiation, as it is the case in one of the most common forms of aggregative multicellularity: the formation of spore-bearing fruiting bodies. Chief among these are the cellular slime molds like *Dictyostelium* (Amoebozoa) with up to five different cell types. But some sort of cell specialisation also occurs in the aggregates formed by the amoebas *Fonticula* (Nucleariida), *Sorogena* (Ciliata) or *Guttulinopsis* (Rhizaria); or the myxobacteria *Chondromyces* (Bonner 2000a; Du *et al.* 2015). In other cases, such as the amoeba *Capsaspora owczarzaki* (Filasterea), spatial cell differentiation has not been demonstrated despite having clear temporally regulated cell types (Sebé-Pedrós *et al.* 2013a, 2016a; b).

1.1.2. Evolutionary consequences of multicellularity

1.1.2.1. Adaptive scenarios

As multicellularity has appeared multiple times along evolution, it is reasonable to assume that selective pressures favoring its establishment are relatively frequent, and its disadvantages possible to overcome (Bonner 2000a; Grosberg and Strathmann 2007).

The ability to sustain larger organismic sizes is a frequently invoked advantage to becoming multicellular (Bonner 2000d; King 2004): as Bonner simply put it, '[since all organisms evolved from small unicellular forms] there is always an open niche at the top of the size spectrum; it is the realm that is ever available to escape competition'. It is the type of competition what defines the effects that size selection has in the resulting multicellular organism. For example, size can be a defense mechanism against heterotrophic predators, which can lead to a prey-predator arms race that further enhances diversification (Stanley 1973). An interesting study in this regard is the experimental evolution of clonal multicellularity by the chlorophyte *Chlorella vulgaris.* After few generations of culture with the predatory chrysophyte *Ochromonas vallescia*, the green algae started forming self-replicating eight-celled colonies that were virtually immune to predation (Boraas *et al.* 1998). Likewise, larger Volvocales species (a group that contains organisms ranging from the single-celled *Chlamydomonas reinhardtii* to the large colonies of *Volvox* species, including a wide

range of phylogenetically scattered intermediate forms) can swim faster and therefore escape predators more efficiently. They are also more efficient in nutrient storage (Bonner 2000d; Kirk 2003). Larger sizes can also facilitate the emergence of an internal, stable chemical milieu that shields an organism from environmental changes; *e.g.* the aggregates of the anaerobic archaea *Methanosarcina* maintain an internal anoxic environment (Bonner 2000a).

Division of labour and resource pooling are other adaptations typical of multicellular organisms (Szathmáry and Smith 1995; Ispolatov *et al.* 2012), a direct consequence of the emergence of individuality from low-level groups of cooperating cells (Michod 2007). In multicellular organisms, for example, different cells can be responsible for mutually exclusive processes, such as metabolic pathways (*e.g.*, the biochemical incompatibility of photosynthesis and nitrogen fixation is overcome in cyanobacterial colonies (Rossetti *et al.* 2010)) or simultaneous flagellar motility and mitosis (which otherwise compete for the same cellular machinery, the microtubule organizing center; Buss 1987). This latter case has been proposed as a case-in-point example of 'path-dependence' in the emergence of multicellularity: the cell biology of the unicellular ancestor (which imposes the mechanistic incompatibility of both processes) impacts the early evolution, and the developmental constraints, of the early multicellular lineage (Grosberg and Strathmann 2007). This effect has been proposed to shape cell differentiation patterns of animals (Buss 1987; King 2004) and Volvocales (Michod 2003; Kirk 2003), as both fall within a model comprising 1) somatic external cells, flagellated and motile; and 2) internal proliferative and unflagellated cells (which, earlier in evolution, were poised to become the germ line; Grosberg and Strathmann 2007).

1.1.2.2. Non-adaptive scenarios

In contrast, other authors have argued that the emergence of many traits associated with complex multicellularity can be more readily explained by non-adaptive population-genetic processes than by the adaptationist paradigm (Lynch and Conery 2003; Koonin 2004, 2016; Lynch 2007), particularly at the genome level. This is a three-fold argument. First, multicellular species such as Metazoa have reduced effective populations and recombination rates and higher deleterious mutation rates, which reduces the efficiency of selection (Lynch 2003). Second, the genomic complexities that characterize multicellular species (a tractable proxy for organismic complexity examined later in the section *1.3. The genomic foundations of Metazoa origins*; Koonin 2011; Wolf and Koonin 2013) appeared not because of their adaptive value, but because of the lower efficiency of purifying selection (Lynch 2007). Third, Lynch proposes that, as the phenotype depends on the underlying complexity of the genome, the non-adaptive processes governing genome evolution are essential to understand the adaptations developed at the phenotypic level by multicellular organisms – namely, morphology and development, complex patterns of transcriptome regulation, cell types, or cell signaling (Lynch 2007).

1.1.2.3. Genetic conflicts in multicellular individuals

Central to the ideas of division of labour and cell type differentiation is the fact that multicellular organisms behave like a unit of selection *per se*, *i.e.* a 'Darwinian' individual, equally composed of other individuals also subject to selection (Michod and Roze 2001; Michod 2003). Because

multicellular organisms frequently contain self-sacrificial cell lines, *e.g.* somatic cells forgoing reproduction to promote the fitness of the germ line, multicellularity bears an implicit conflict due to overlapping levels of selection (Michod 2007; Rainey and De Monte 2014).

This conflict has been modeled using the multilevel selection theory, which proposes a series of analytically tractable stages during the transition to multicellular individuality, and offers a way out of the conflict conundrum (Heisler and Damuth 1987; Damuth and Heisler 1988). First, undifferentiated cell groups are dominated by within-level conflict because of the coexistence of cooperating and parasitic cells; which is tractable to model from the point of view of kin/group selection (Rainey and De Monte 2014). Second, after the transition to multicellularity is completed, the existence of germ-soma differentiation creates between-level conflicts involving the differentiated cell populations and the organism (as they reproduce at different timescales). The emergence of lineage selection, conferring Darwinian individuality to cell types within the multicellular organism, is a favored model to overcome the between-level conflict (Michod 2006; Rainey and De Monte 2014). In between both stages, however, stands the necessity of reproduction of the collective without inheriting the within- and between-level genetic conflicts – a process often achieved via single-celled propagules that found new, genetically related organisms (Szathmáry and Smith 1995; Grosberg and Strathmann 2007; Godfrey-Smith 2009).

Interestingly, the transitions to multicellular individuality based on single-celled propagules have been shown to be able to occur suddenly, even with minimal genetic changes (Ratcliff *et al.* 2013; Hammerschmidt *et al.* 2014; Rainey and De Monte 2014). Therefore, abrupt transitions implicitly acknowledge the essential contribution of pre-existing genetic traits to multicellularity: the recruitment of gene tool-kits, regulatory networks and functional modules that had evolved in unicellular organisms can have a pre-adaptive value at the collective level, and even support the emergence of life cycles. This pre-adaptive value of the genome content has been well studied in Metazoa and their closest unicellular relatives, as I shall explain in the following section.

Finally, it is worth mentioning that multicellular aggregates are fundamentally different from clonal organisms in the type of evolutionary trade-offs they are subject to. The existence of higher intra-organism genetic variance in aggregates can result in a lower efficiency of selection at the whole-organism level (for example, because of opportunists or free-loaders) (Michod and Roze 2001; Michod 2003). Therefore, selection for traits of the emergent individuals coexists with selection at the single-cell level (Michod 2007; Grosberg and Strathmann 2007). A possible consequence of this decreased integration is the lower frequency of complex cell differentiation in aggregates when compared to clonal organisms, often limited to simple reproductive foraging propagules, as in *Dictyostelium discoideum* (Bonner 2000e).

1.2. Multicellularity in Metazoa

Under the previous section *1.1.1. The plurality of multicellularity*, I have reviewed the time-line of the emergence of Metazoa as inferred from the fossil record of stem and crown animals, and molecular clock estimates. In summary, the first *bona fide* animal fossils date back to the Ediacaran period, circa 600 Mya (Narbonne and Gehling 2003; Narbonne 2005; Yin *et al.* 2015); while the divergence of

extant animal lineages is proposed to have occurred earlier, during the Cryogenian, 700-800 Mya (Erwin *et al.* 2011; dos Reis *et al.* 2015). In this section I will present the geological and ecological circumstances of this event and provide insights into the phylogenetic context in which the transition to multicellularity occurred by presenting the closest unicellular relatives of Metazoa: choanoflagellates, filastereans, ichthyosporeans and *Corallochytrium limacisporum*.

1.2.1. Evolutionary perspectives of animal origins

1.2.1.1. Geochemical environment

More than 2 billion years passed between the origin of prokaryotic life on Earth (~3.5 Gya; Allwood et al. 2007) and the emergence of Metazoa during the Neoproterozoic era (750-800 Mya), which also lagged way behind the origin of eukaryotes (~2.1 Gya; Figure 2; El Albani et al. 2010). Similarly, it is around 800 Mya when molecular clock estimates and microfossil evidence suggest that the other major eukaryotic lineages began to diversify: fungi, red and green algae, rhizarians, stramenopiles and alveolates (Berney and Pawlowski 2006; Knoll 2011). The explanation behind this protracted diversification of eukaryotes, specifically in Metazoa, has been based on changes in the Earth geochemical environment. In particular, a major shift in ocean chemistry occurred at the end of the Neoproterozoic (~800 Mya): the anoxic ocean sub-surface zone gradually became less sulfidic and more ferruginous, switching from a eukaryote-toxic environment to one that enabled the spread of eukaryotes to new environments (Knoll 2011).

In parallel, this process was accompanied by an increase in atmospheric and oceanic oxygen levels that is deemed crucial to the emergence of Metazoa (Knoll and Carroll 1999; Budd and Jensen 2007). During its earliest history, the Earth atmosphere was deprived of oxygen. At the beginning of the Proterozoic (~2.4 Gya), the Great Oxygenation Event increased oxygen concentration to ~1% of present atmospheric levels (Sahoo *et al.* 2012), a value that is at the threshold of the minimum requirements for metazoan life (Towe 1970; Knoll and Carroll 1999; Budd and Jensen 2007). A second profound oxygenation event occurred during the transition from the Cryogenian to the Ediacaran period ~635 Mya, at the end of the Proterozoic (Sahoo *et al.* 2012). It is the latter increase that is considered to be an enabling factor in the emergence of multicellularity. Firstly, because high oxygen concentrations are needed in order to synthesize the collagen-based extracellular matrices that sustain multicellular tissues in Metazoa (Towe 1970). Secondly, because multicellular life is constrained by its ability to distribute oxygen within the organism: before active transport evolved, organismic size was limited by the efficiency of diffusion (Erwin 1993; Budd and Jensen 2007). All Metazoa have methods to circumvent the oxygen diffusion constraints imposed by their large sizes: poriferans are porous organisms embedded in water and promote oxygen exchange by coordinated flagellar movement; cnidarian bodies typically comprise extensive, thin sheets with little impediment to diffusion; and bilaterians have complex respiratory and circulatory systems to support oxygen active distribution and exchange (Knoll 2011).

However, evolutionary hypotheses drawn from geochemistry alone have limited explanatory power: they can point at the specific timing of animal origins, but cannot fully explain the diversity of species and developmental modes that animal multicellularity entailed (Knoll and

Carroll 1999; Sperling *et al.* 2013). In addition, they can overlook critical events. For example, both the sedimentary record and geochemical data are biased towards marine environments, which excludes possible diversifications outside the oceans (Knoll 2011). In addition, it has been shown that modern animals can thrive under low oxygen conditions that had already been attained almost 2 billion years before the Cryogenian-Ediacaran oxygenation event (Sahoo *et al.* 2012): the demosponge *Halichondria panicea* can grow at 0.5-4% of the present atmospheric oxygen levels (Mills *et al.* 2014); some bilaterians as low as 0.3% (Levin *et al.* 2002; Mills and Canfield 2014); and collagen synthesis can occur at low oxygen concentrations, albeit with lower efficiency (Mills and Canfield 2014).

1.2.1.2. Ecological landscape

Combined with inputs from palaeogeochemistry, ecological hypotheses can unravel the reasons behind the specific organismic innovations seen in Metazoa, including feeding modes and morphology (Knoll and Carroll 1999; Sperling *et al.* 2013). In this sense, prey-predator dynamics between eukaryotes have been proposed to set off animal diversification by initiating an 'arms race' circa 800 Mya, during the Cambrian period (Butterfield 2007, 2011). An early, purely ecological theory (Stanley 1973) suggested that the emergence of eukaryotic heterotrophs (predators) in the premetazoan, autotroph-dominated oceans triggered a positive feedback loop of diversification that led to the emergence of Metazoa. This line of reasoning affects both micro- and macro-eukaryotes: the predatory pressure from novel heterotrophs favored the diversification observed in the protist fossil record, including the rise of phytoplankton to ecological prominence (Trommer *et al.* 2012); biomineralized, shelled eukaryotes (Porter 2011); and multicellular or coenocytic eukaryotes as a defense mechanism (Boraas *et al.* 1998; Knoll 2014).

Given the protracted radiation of eukaryotes outlined above (Berney and Pawlowski 2006), it seems likely that the biosphere from which animals appeared was dominated by prokaryotes (McFall-Ngai *et al.* 2013; Alegado and King 2014). This prokaryote-dominated world shaped metazoan evolution in different aspects. For example, and in connection to geochemistry, the serial oxygenation events that enabled the rise of multicellularity were partly driven by the photosynthetic activity of marine cyanobacteria (Kasting and Siefert 2002; Alegado and King 2014). In parallel, it has been suggested that early metazoans preyed on bacteria-dense stromatolite communities, which points at a close relationship between animals and prokaryotes (Dornbos *et al.* 2004). In this sense, it is relevant to note that the colonial development of the choanoflagellate *Salpingoeca rosetta* (a phagotrophic bacterivore) is triggered by a bacteria-derived sulfonolipid (Alegado et al. 2012; Woznica et al. 2016; see below). As *S. rosetta* colonies exhibit enhanced prey capture abilities, this simple multicellular structure could be a consequence of adaptation to environmental changes in food availability (Alegado and King 2014; Dayel and King 2014). Tantalizingly, regulation of development by disparate chemical cues from environmental bacteria is a common event in Metazoa and even multicellular green algae (McFall-Ngai *et al.* 2013; Alegado and King 2014).

1.2.2. Holozoa: animals and their unicellular relatives

The study of the taphonomic, ecological and geochemical circumstances of early animal evolution reveals that the first Metazoa likely emerged from a group of heterotrophic protists. However, in order to fully understand the biology of the unicellular ancestor of Metazoa, we need to draw information from comparisons with the animals' phylogenetic vicinity: the Holozoa clade, which comprise animals and their closest unicellular relatives (Figure 4).

The Holozoa group was erected by Lang et al. (2002) to formalize the close relationship between Metazoa and two clades of unicellular opisthokonts: Choanoflagellata and Ichthyosporea. Later phylogenomic studies have included three additional species in the Holozoa: *Capsaspora owczarzaki* (Owczarzak *et al.* 1980; Hertel *et al.* 2002) and *Ministeria vibrans* (Cavalier-Smith and Chao 2003), which belong to the Filasterea; and *Corallochytrium limacisporum* (Raghukumar 1987). In this section I will examine the basic characteristics of each of these groups of protists, focusing on their morphological features and multicellular-like behaviours. In a later section (*1.3.1. Phylogenomics unveils the unicellular relatives of animals*), I will expand on the contribution of phylogenomic analyses to the systematics of Holozoa.

1.2.2.1. Choanoflagellata

Choanoflagellates, also known as Choanomonada, were first associated with Metazoa in the mid-19th century, based on their striking similarity with the choanocytes, a cell type of present in Porifera (James-Clark 1866, 1871)[1]. Together with animals, they have recently proposed to belong to the monophyletic Apoikozoa clade (Budd and Jensen 2015), a synonym of informal terms such as 'choanimals' (Fairclough *et al.* 2013).

The choanoflagellate lineage includes ~250 species of spherical/ovoid protists that display a collar of microvilli (a specialized actin-based filopodial structure) surrounding a single apical flagellum (Adl *et al.* 2012). The microvilli-flagellum complex is often involved in their lifestyle as phagotrophic bacterivores: they use the flagellar whipping to create currents leading to the collar, which facilitates phagocytosis (Dayel and King 2014). Choanoflagellate species can be colonial or solitary, and have been found as free-living in a range of aquatic environments: the marine water column, abyssal plains, freshwater and anoxic/hypoxic brackish waters (Nitsche *et al.* 2007; del Campo and Massana 2011; Wylezich *et al.* 2012; Del Campo and Ruiz-Trillo 2013).

Choanoflagellates have been recently re-classified in order to overcome the limitations of morphology-based taxonomy at the species/genera-level (Nitsche *et al.* 2011; Carr *et al.* 2017). Two main monophyletic clades are recognized, in overall agreement with traditional classifications: Acanthoecida (including Acanthoecidae and Stephanoecidae) and Craspedida. Acanthoecida have a distinctive siliceous structure known as lorica that surrounds the cell and facilitates a pelagic lifestyle; Craspedida have a vegetative sedentary stage in which the cell is bound to the substrate, but can also swim (Carr *et al.* 2008, 2017; Nitsche *et al.* 2011). Currently, there are two available choanoflagellate genomes, both of them craspedids: the solitary *Monosiga brevicollis* (King *et al.*

1. The animal status of sponges was established shortly after, based on analyses of their ontogeny (Schulze 1885).

2008) and the colony-forming *Salpingoeca rosetta* (Fairclough *et al.* 2013). *S. rosetta* has been thoroughly characterized from the point of view of transcriptomics and cell biology (see below).

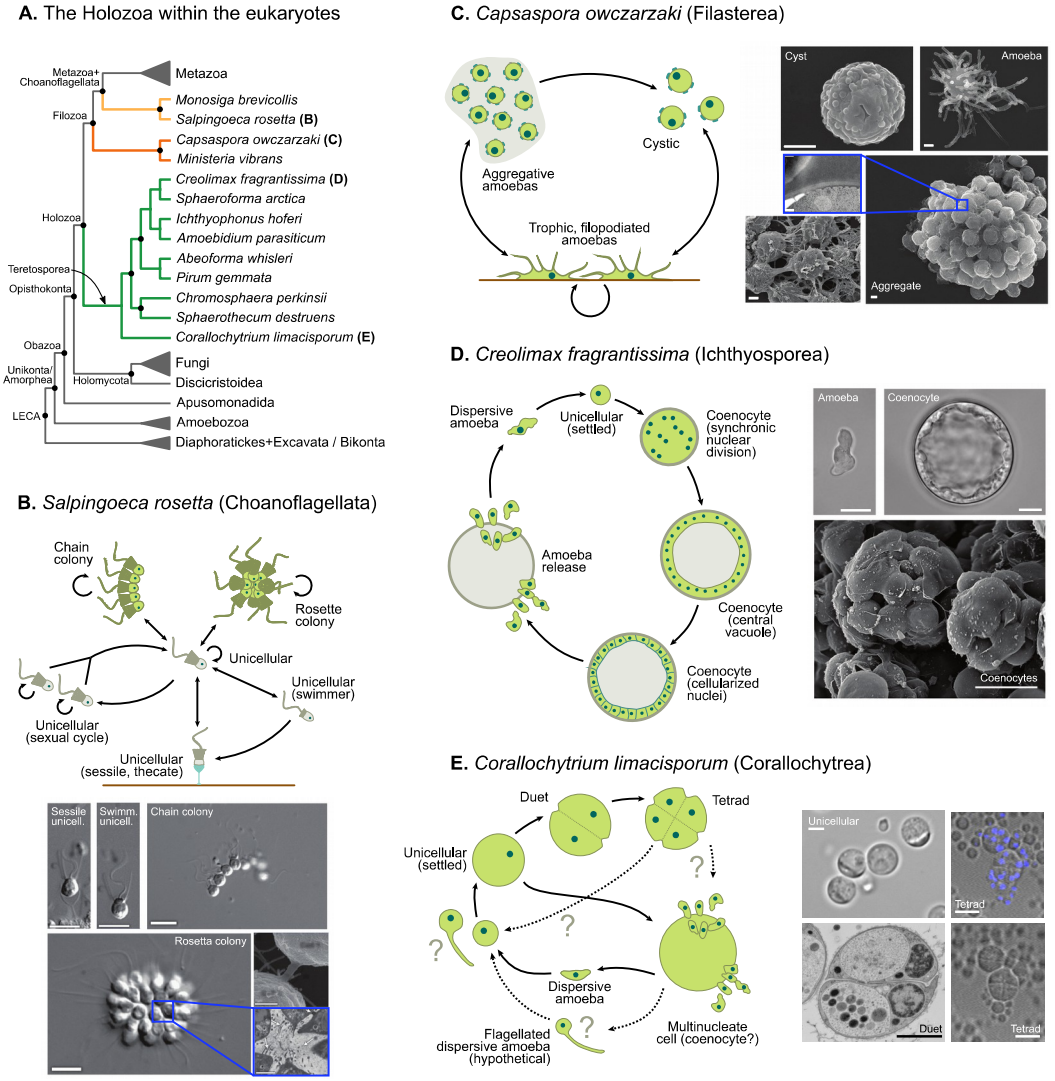


Figure 4. A) Phylogenetic classification of Holozoa within the eukaryotes. **B)** Life cycle of *Salpingoeca rosetta*, a colonial choanoflagellate. Circular arrows indicate the proliferative stages. It comprises two kinds of colonies (chain and *rosette*-like), three unicellular stages (fast- and slow-swimming and a thecate, sessile flagellate) and a unicellular sexual cycle (meiosis). Scale bars are 5µm, except in the 'rosette' right panel, where it is 1µm. **C)** Life cycle of *Capsaspora owczarzaki*, an aggregative amoeba within the Filasterea. The proliferative stage consists of filopodiated, surface-adherent amoebas that can form aggregates by extracellular matrix segregation (composition unknown). Amoebas can encyst (resistance form). Scale bars are 1µm, except in the aggregate panel, where it is 200nm. **D)** Life cycle of *Creolimax fragrantissima*, an ichthyosporean. Single-celled motile amoebas settle and start a coenocytic outgrowth with synchronized nuclear division. The nuclei are gradually displaced towards the cell periphery as a central vacuole grows. Then, individual nuclei are cellularized and released as dispersive amoebas. Scale bars are 10µm, except in lower picture, where it is 50µm. **E)** Life cycle of *Corallochytrium limacisporum*, sister group to ichthyosporeans. Clonal outgrowths from settled amoebas are similar to *C. fragrantissima*'s, but the existence of a multinucleate, vacuolated coenocyte is unclear. Sometimes, individual cells undergo (confocal microscopy) serial binary palintomic division to form cell duets (TEM picture), tetrads (pictured with confocal microscopy and DAPI nuclear staining; upper right), etc. A flagellated stage (possibly dispersive) has been hypothesized. Scale bars are 1µm. Adapted from Sebé-Pedrós et al. (2017) and references therein.

Homology and homoplasy of cell types between choanoflagellates and animals

Since the identification of choanoflagellates as sister group to Metazoa (James-Clark 1866, 1871), the proposed homology of the classical choanoflagellate cell with poriferan choanocytes has been widely discussed in the context of the origin of multicellularity (Maldonado 2004; Mikhailov *et al.* 2009; Budd and Jensen 2015). However, this apparent similarity must be treated with caution. First, choanocyte-like cells are not exclusive to choanoflagellates and Porifera, being present in scattered hemichordates, echinoderms, ascidians, cnidarians and also other protists (Maldonado 2004; Alegado and King 2014). Second, a recent cytological analysis downplays their possible homology based on fundamental differences in morphology and mechanistics between the microvilli-flagellum complexes of the sponge *Spongilla lacustris* and the choanoflagellate *M. brevicollis* (Mah *et al.* 2014).

Likewise, a link between choanoflagellate coloniality and animal multicellularity has also been long debated, *e.g.* the proposal that a primary colony of protists would be the first step towards the evolution of more complex multicellular Metazoa (Metchnikoff and Metschnikoff 1886; Nielsen 2008; Carr *et al.* 2008; Mikhailov *et al.* 2009). The similarity between the cytoplasmic bridges established by animal cells and choanoflagellate colonies (Fairclough *et al.* 2013), for example, would fit into this hypothesis. However, since almost all colonial choanoflagellates are craspedids, this morphology could be a secondary innovation after the divergence of choanoflagellates from other Holozoa, although the presence of coloniality in the acanthoecid *Diaphanoeca sphaerica* could push the origin of coloniality to the choanoflagellate ancestor (Carr *et al.* 2017). However, the analysis of cell type-specific transcriptomic profiles does not appear to support a scenario in which choanoflagellate-style coloniality existed at their shared ancestor with Metazoa, as explained below (Fairclough *et al.* 2013; de Mendoza *et al.* 2015)).

Transcriptomic characterization of Salpingoeca *life cycle*

The colony-forming *S. rosetta* (Figure 4B) is an emerging model for the study of premetazoan multicellular-like cellular functions. Its most distinctive morphology is the formation of spherical 'rosette' colonies by incomplete cell division (Fairclough *et al.* 2010), a process triggered and enhanced by bacterial signals (Alegado *et al.* 2012; Woznica *et al.* 2016). A comparative analysis of multiple cell stages revealed that the transcriptomic profile of *S. rosetta* rosette colonies and the solitary swimming cells from which they develop is enriched in genes exclusively shared by choanoflagellates and Metazoa; whereas colonies have a *Salpingoeca*-specific profile (Fairclough *et al.* 2013). The authors thus hypothesized that early colony development is based on genomic features that originated at the shared ancestor of choanoflagellates and Metazoa, while the specific colonial cell type would be a choanoflagellate innovation. This result is supported by the identification of up-regulated septin genes in the rosette colonies (GTPases that regulate cytokinesis in Fungi and Metazoa), which the authors linked to a mode of incomplete cell division also found in Metazoa (Fairclough *et al.* 2010, 2013). A recent forward genetics study has identified a C-type lectin gene, termed *rosetteless*, which is essential for the establishment of rosette colonies (Levin *et al.* 2014). C-type lectins are an exclusive choanoflagellate and metazoan gene family, but the specific orthology of *rosetteless* is currently unknown – adding further speculation to the

question of whether the colony-forming behaviour of choanoflagellates is an ancestral feature shared with 'stem' Metazoa.

1.2.2.2. Filasterea

Filasterea is an holozoan lineage composed of only two known species: *Capsaspora owczarzaki* and *Ministeria vibrans* (Shalchian-Tabrizi *et al.* 2008). Both of them are small (3-7 µm) naked filopodiated amoebas with a single nuclei. Strikingly, environmental surveys of eukaryotic genetic diversity have failed to detect putative new filasterean taxa (Del Campo and Ruiz-Trillo 2013), although two putative new species have been recently reported (Tikhonenkov et al. 2016, and personal communication by Elisabeth Hehenberger). Together with animals and choanoflagellates, they conform the Filozoa clade (Shalchian-Tabrizi *et al.* 2008; Torruella *et al.* 2012).

*M. vibran*s was characterized as a heterotrophic, flagellated amoeba, of probable cosmopolitan distribution as a free-living bacterivore (Cavalier-Smith and Chao 2003). *M. vibrans* has a reported sibling species of the same genera, *M. marisola* (Patterson *et al.* 1993), which has never been molecularly characterized. *M. vibrans* has a single posterior flagellum that is not present in *C. owczarzaki or M. marisola*. The characterization of its flagellum remained elusive until its recent confirmation using transcriptomic data, transmission electron microscopy and immunological staining of the structural α-tubulin (Torruella *et al.* 2015).

C. owczarzaki was isolated from the hemolymph of the freshwater snail *Biomphalaria glabrata*, and was thus considered to be a symbiont (Owczarzak *et al.* 1980; Hertel *et al.* 2002). It was first associated to nuclearids, a group of amoeboid fungal relatives, but later phylogenetic analysis placed it within the Holozoa (Ruiz-Trillo *et al.* 2004, 2006; Steenkamp *et al.* 2006). Most notably, its life cycle includes a transient stage as a multicellular aggregate that can be induced in culture, starting from filopodiated, crawling amoebas that attach to the substrate. Additionally, it presents a cystic stage during which it exhibits many features of a dormant/resistance cell type. These cell types have been characterized using transcriptomic, proteomic and epigenomic analyses (Sebé-Pedrós *et al.* 2013a, 2016a; b).

Transcriptomic, proteomic and epigenomic characterization of Capsaspora *life cycle*

The multicellular stage of *C. owczarzaki* develops from non-clonal aggregation of single amoeboid cells (Figure 4C), and it was first characterized using comparative transcriptomic analysis of different temporal cell types (Sebé-Pedrós *et al.* 2013a). This study showed that cystic, aggregative and single amoebas have distinct transcriptomic profiles. In particular, multicellular aggregates up-regulate genes related to multicellular behaviour in Metazoa such as integrin adhesome components, laminins, and tyrosine kinases, which leads the authors to argue that the molecular tool-kit associated with animal multicellularity can function in both aggregative and clonal contexts. Furthermore, they identified a small but significant contribution of regulated alternative splicing to the cell type-specific transcriptomic profiles, including exon skipping of kinase-mediated signaling genes (Sebé-Pedrós *et al.* 2013a).

A later study characterized *C. owczarzaki* protein expression patterns, which also revealed temporally regulated profiles that correlated with transcriptomic data (Sebé-Pedrós *et al.* 2016a). The authors found a significant enrichment of genes shared with Metazoa and choanoflagellates in the proteomic profile of aggregates, which, they argue, supports the presence of aggregative behaviour in ancestral holozoans. Most interestingly, Sebé-Pedrós *et al.* also revealed a dynamic pattern of protein phosphorylation across *C. owczarzaki* cell types, which they argue to be an holozoan distinctive feature on the basis of the emergence of tyrosine kinase signalling genes at the ancestor of Holozoa (Suga *et al.* 2012).

Finally, the regulatory functions of *C. owczarzaki*'s genome have also been characterized in a cell type-specific manner, which revealed a dynamic regulation of the chromatin states (measured by histone post-translational modifications) that also correlated with gene expression (Sebé-Pedrós *et al.* 2016b). Furthermore, a characterization of transcription factor (TF) binding sites revealed, for example, that the *Capsaspora* ortholog of *Brachyury*, a TF involved in animal gastrulation, also controls genes related to cell migration in the aggregative and filopodial stages. This implies that some animal TF networks were already present in the unicellular ancestor of animals.

1.2.2.3. Ichthyosporea

Ichthyosporea are also sometimes known as Mesomycetozoea (Mendoza *et al.* 2002), and were formerly referred to as the DRIP clade (an acronym of the original species it included: *Dermocystidium*, the 'rosette agent', *Ichthyophonus*, and *Psorospermium*; *Ragan et al. 1996; Cavalier-Smith 1998a*). They are a group of osmotrophic/saprotrophic protists, frequently multinucleated and sometimes with a single posterior flagellum. Almost all ichthyosporeans have been isolated from animal tissues, where they live either as parasites, mutualists or commensals (Glockling *et al.* 2013); but a few free-living species have been identified as well (Hassett *et al.* 2015) and unsampled lineages have been identified in environmental surveys of ocean eukaryotic diversity (Del Campo and Ruiz-Trillo 2013; del Campo *et al.* 2015).

Ichthyosporea are divided in two groups that include about 40 characterized species: Ichthyophonida and Dermocystida (Cavalier-Smith 1998b; Mendoza *et al.* 2001, 2002; Adl *et al.* 2012; Glockling *et al.* 2013). This division is supported by phylogenetic analyses, according to which both groups are monophyletic (Marshall and Berbee 2011), and is consistent with a number of phenotypic traits related to morphology and life cycle (Mendoza *et al.* 2002; Glockling *et al.* 2013).

The Ichthyophonida is the most species-rich clade according to environmental surveys (Del Campo and Ruiz-Trillo 2013) and includes organisms such as *Amoebidium parasiticum*, *Ichthyophonus hoferi*, *Creolimax fragrantissima*, *Pirum gemmata*, *Abeoforma whisleri*, *Sphaeroforma tapetis*, *Sphaeroforma arctica*, *Sphaeroforma sirkka*, *Sphaeroforma napiecek* (only the latter two have been described as free-living; Mendoza et al. 2002; Marshall et al. 2008; Marshall and Berbee 2011; Glockling et al. 2013; Hassett et al. 2015). Many ichthyophonids have a broadly conserved developmental mode consisting of large, multinucleated, spherical coenocytes with a central vacuole (also known as sporangia or sporocyst), that release a dispersive amoeboid stage (sometimes referred to as spores, zoospores, endospores or schizonts) by cellularization of the internal nuclei; amoebas will then typically disperse and establish a new colony (Mendoza *et al.* 2002). Ichthyophonid amoebas are

frequently spherical or limax-shaped and lack a flagellum. However, some species exhibit fungal-like features: *A. parasiticum* has thalli that release elongated amoebas with chitin walls (Mendoza *et al.* 2002; Torruella *et al.* 2015); and *I. hoferi* can develop hyphal structures (Mendoza *et al.* 2002). Others, like *A. whisleri*, exhibit a wide range of phenotypes: cells with pseudopodia, hyphal and plasmodial structures, and amoeboid cell types that can divide without reaching the coenocytic stage (Marshall and Berbee 2011).

The order Dermocystida (sometimes known as Rhinosporideacae) is historically composed of strictly parasitic species, a notable example being the 'rosette agent' *Sphaerothecum destruens*, a well-known fish pathogen (Mendoza *et al.* 2002; Glockling *et al.* 2013). Their developmental mode is roughly conserved with ichthyophonids: a spherical sporangium that releases dispersive zoospores. However, the zoospores are frequently uni-flagellated; and the sporangia can lack the central vacuole. Due to their strictly parasitic nature and difficulties in establishing monoaxenic cultures, they are less well characterized than ichthyophonids from the molecular point of view (Glockling *et al.* 2013).

Currently, there are three available ichthyosporean genomes: *C. fragrantissima* (de Mendoza *et al.* 2015), *I. hoferi* (Torruella *et al.* 2015) and *S. arctica* (Ruiz-Trillo *et al.* 2007); all of them belonging to the Ichthyophonida. Transcriptomic data exists for three additional species (Torruella *et al.* 2015), including *Sphaerothecum destruens*, a parasitic dermocystid also known as 'rosette agent'.

Transcriptomic and cell biology insights into the life cycle of Creolimax fragrantissima

C. fragrantissima has been isolated multiple times from a range of invertebrates (Marshall *et al.* 2008). Its life cycle follows the prototypical developmental mode of ichthyophonids (Figure 4D): a small, spherical zoospore (~6-8 µm) develops a central vacuole and grows in size. After multiple rounds of coenocytic nuclei division, it reaches maturation (25-60 µm), cellularizes and releases a number of motile amoebas through its cell wall, which then disperse, encyst and restart the cycle (Marshall *et al.* 2008). This process has been studied in the recent years due to its close resemblance with the coenocytes and/or syncytia exhibited by some animal embryos, slime molds and Fungi (Bonner 2000a; Chen *et al.* 2007; Grosberg and Strathmann 2007; Suga and Ruiz-Trillo 2013), as it provides interesting insights into the strategies for multicellular development in a wide range of eukaryotes. For example, Suga and Ruiz-Trillo (2013) described that nuclei division is synchronized within the coenocytic cell, and that the nuclei are arranged beneath the cell surface of the colony. Interestingly, they also provided evidence that the same process also occurs in *S. arctica*, another ichthyophonid.

Recently, de Mendoza et al. (2015) investigated the transcriptomic profile of *C. fragrantissima* developmental cell types in a comparative analysis with other holozoans, and demonstrated that it has a program of transcriptionally regulated cell type specification. Unexpectedly, they identified an up-regulation of animal-like gene tool-kits in the amoeboid dispersive stage, and not in the coenocytic growth phase: this pattern includes developmental transcription factors and adhesion genes involved in the integrin adhesome. The multinucleated coenocytes, instead, appear to have transcriptomic profiles analogous to the proliferative, undifferentiated animal cell types, like stem

cells. In parallel, they also demonstrated that *C. fragrantissima* has co-opted ancestral gene regulatory programs to develop a novel osmotrophic feeding mode (absent in non-ichthyosporean holozoans). Overall, they provide direct evidence of the plasticity of cell type evolution across holozoan lineages, supporting a scenario of recurrent recruitment of co-regulated expression programs to support the emergence of novel cell types and developmental programs (Newman 2012).

1.2.2.4. Corallochytrea

The Corallochytrea clade includes a single described species, *Corallochytrium limacisporum*. It is a small free-living osmotroph, first isolated from marine coral reef lagoons in the Arabian Sea and, more recently, Hawaii (Raghukumar 1987; Torruella *et al.* 2015). Its taxonomic affiliation has long been elusive: it was first classified as a thraustochytrid due to its morphology (Raghukumar 1987), as a fungus do to its lysine catabolism (Sumathi *et al.* 2006) or as sister to choanoflagellates based on phylogenetic analysis of the small ribosomal subunit (Cavalier-Smith and Paula Allsopp 1996). *C. limacisporum* was finally classified as sister group to Ichthyosporea within Holozoa in more recent and taxon-rich phylogenomic analyses (Torruella *et al.* 2015).

From a morphological point of view, *C. limacisporum* is a small (4.5-20 µm) spherical protist. It has been proposed to have lost its flagellum secondarily (Cavalier-Smith 1998b), but a recent comparative transcriptomic analysis revealed that it nevertheless expresses most of the required flagellar genetic tool-kit (Torruella *et al.* 2015). Its life cycle bears some similarities with that of many ichthyosporeans (Figure 4E): it starts with a uninucleated cell that undergoes a number of rounds of binary cell division during which the daughter cells remain attached to each other, until the release of amoeboid limax-like cells that settle and form new colonies (Raghukumar 1987). Unlike ichthyosporeans, however, it is not clear whether it goes through a coenocytic stage. Most interestingly, cell division sometimes occur by palintomic cleavage (*i.e.*, originating Y-shaped junctions and without/little cytoplasmic growth between divisions), a feature that has otherwise been used to classify unclear micro-fossils as animals (Xiao *et al.* 2012; Chen *et al.* 2014b). The presence of this division mode in *C. limacisporum* forestalls such interpretations (Huldtgren *et al.* 2011, 2012; Cunningham *et al.* 2016), although it can fuel further speculation as to a possible homologous occurrence of this trait in animals.

1.3. The genomic foundations of Metazoa origins

The sequencing of the genomes of key early-branching animals and their unicellular relatives over the last decade (Putnam *et al.* 2007; King *et al.* 2008; Srivastava *et al.* 2010; Suga *et al.* 2013; Simakov *et al.* 2013, 2015; Fairclough *et al.* 2013; Ryan *et al.* 2013; Moroz *et al.* 2014; Francis *et al.* 2017) has enabled a novel approach to the study of early metazoan evolution: comparative genomic analyses aimed at reconstructing the genome of the last common ancestor (LCA) of Metazoa. This endeavor involves methodologies drawn from phylogenetic inference and genome structural analyses, and

aims to reveal the genomic changes that underpinned the transition from a solitary unicellular protist to the first multicellular ancestor of animals.

In the present section I will examine the impact of comparative genomics on the study of early animal evolution. First, how phylogenomic analyses have enabled the identification of the relevant animal unicellular relatives. Second, I will elaborate on the nature of genomic novelties in Metazoa, from gene content to genome structure. In the former, I aim to reconcile the influence of exclusive animal innovations, co-option of pre-existing genes and tinkering with ancient gene families at the protein level. Third, I will address the evolutionary basis of an alternative method of gene innovation in Metazoa: alternative splicing of transcript RNAs.

1.3.1. Phylogenomics unveils the unicellular relatives of animals

Phylogenomics refers to the use of phylogenetic analysis based on genome-scale markers, be it multiple genes or other changing genomic features, in order to classify species according to their evolutionary relationships (Eisen and Fraser 2003). The use of phylogenomic approaches, combined with improved inference methods and emergence of cheaper whole-genome or whole-transcriptome sequencing techniques, enabled an unprecedented precision in the study of deep phylogenetic relationships in the tree of life.

By combining multiple genomic markers such as sets of orthologous genes, phylogenomics allows to extract a congruent phylogenetic signal that overcomes the noise of individual markers (Delsuc *et al.* 2005). In contrast, a phylogenetic analysis based on a single gene, for example, is prone to a number of errors. For example, the choice of the individual gene marker is no trivial, as some universal genes like the small ribosomal subunit (SSU rDNA or 18S) can easily exhibit saturated rates of nucleotide substitutions due to homoplasy, which can in turn produce long-branch attraction artifacts. Also, other frequently used gene markers have turned out to be affected by orthology mis-assignment, as their specific evolutionary histories differ from that of their species. This is the case of the horizontal gene transfer events detected in α-tubulin (Kim *et al.* 2006); as well as the cryptic paralogy of β-tubulin in Opisthokonta (Steenkamp *et al.* 2006) and elongation factor-1α in eukaryotes (Keeling and Inagaki 2004). Finally, single-gene phylogenies cannot deal with secondary losses of the marker. Phylogenomics overcomes these limitations by taking advantage of informative sites from multiple genes at the same time, and allowing the analysis of truly orthologous genes even if some are missing in individual species.

Metazoa and Fungi were classified as members of the Opisthokonta eukaryotic supergroup by *Cavalier-Smith (1986)*. Choanoflagellates were also included in the opisthokonts as they share a synapomorphic single flagellum emerging from the posterior part of the cell with animals and fungi, and had long been associated with Metazoa on morphological grounds (James-Clark 1866, 1871). It was not until the first molecular phylogenies of opisthokonts that it became clear that other protistan lineages were also close relatives of Metazoa, in what was to become the Holozoa lineage. First, the 'fungus-like' Ichthyosporea were found to be an early-branching group of unicellular holozoans, closer to Metazoa than to Fungi (Lang *et al.* 2002). Shortly after, the amoeba

C. owczarzaki, isolated from the haemolymph of the snail *B. glabrata* (Hertel *et al.* 2002), was classified as an independent opisthokont lineage together with ichthyosporeans and choanoflagellates (Ruiz-Trillo *et al.* 2004), in results similar to the Choanozoa paraphyletic grouping, proposed by Cavalier-Smith to include all unicellular animal relatives (Cavalier-Smith and Chao 2003). *C. owczarzaki* turned out to be closely associated to the solitary amoeba *M. vibrans*, which had also been associated to Metazoa (Steenkamp *et al.* 2006), in the henceforth named Filasterea class (Shalchian-Tabrizi *et al.* 2008). Two parallel contentious issues remained. First, it was not clear whether Filasterea and Ichthyosporea conformed a monophyletic grouping as the earliest branching holozoans ('Filasporea' hypothesis) (Ruiz-Trillo *et al.* 2004, 2008; Liu *et al.* 2009); or if they constituted two independent lineages instead, with filastereans being closer to Metazoa + Choanoflagellata ('Filozoa' hypothesis) (Ruiz-Trillo *et al.* 2008; Shalchian-Tabrizi *et al.* 2008; Torruella *et al.* 2012). Second, the affiliation of the enigmatic osmotrophic protist *C. limacisporum*: it was first classified as a choanozoan (Cavalier-Smith and Chao 2003), but it was unclear whether it was closer to Ichthyosporea (Steenkamp *et al.* 2006) or to choanoflagellates (Cavalier-Smith and Paula Allsopp 1996; Ruiz-Trillo *et al.* 2006). Both controversies were largely solved by a recent phylogenomic investigation of Holozoa (Torruella *et al.* 2015): *C. limacisporum* is the sister group of Ichthyosporea within the newly defined Teretosporea clade, which is the earliest-branching holozoan lineage. Therefore, the current scenario includes four independent clades in Holozoa: multicellular Metazoa, and the unicellular Choanoflagellata, Filasterea (*C. owczarzaki* and *M. vibrans*) and Teretosporea (comprising Ichthyosporea and *C. limacisporum*).

Having a clear and robust phylogenetic framework is key to the interpretation of comparative genomic analyses that take advantage of the accumulating data from Metazoa and their unicellular relatives. Therefore, the study of new holozoan genomes requires a continued sampling effort coupled with phylogenomic investigations in order to illuminate the phylogenetic and taxonomic neighborhood of Metazoa.

1.3.2. Reconstruction of ancestral genomes by comparative genomics

Comparative genomics allows a unique view of the nature of the metazoan LCA, uncovering traits that cannot be studied using palaeoecology or the scant fossil record of premetazoans. For example, it allows the reconstruction of the essential tool-kit for multicellularity, including gene content and genome structure, and how did it evolve before and after the emergence of Metazoa. Table 1 summarizes various statistics regarding the genome composition of the reconstructed last common ancestor of animals, derived from the study of extant genomes (Putnam *et al.* 2007; Srivastava *et al.* 2008, 2010; Csűrös *et al.* 2011; Simakov *et al.* 2013; Simakov and Kawashima 2016). These studies can be extended in order to reconstruct the genome contents of both animals and their unicellular ancestors, thus elucidating the amount and types of innovation entailed by the transition to multicellularity.

Table 1. Reconstructed genome content of the LCA of Metazoa. Data from Csűrös et al. (2011); Simakov and Kawashima (2016).

Genomic feature	Inferred ancestral values
Genome size	~ 300 Mb
Gene family number	7,000-8,000
Total gene number	> 20,000
Intron density	8.8 introns/CDS kbp
Repetitive regions content	~ 30%
Macro-syntenic linkage groups	~ 10-17
Micro-syntenic linkage groups	~ 400

1.3.2.1. Gene content analyses: expansion, co-option, innovation

To date, genomic analyses have been performed comparing animals and *S. rosetta*, *M. brevicollis* (King *et al.* 2008; Fairclough *et al.* 2013), *C. owczarzaki* (Suga *et al.* 2013) and *C. fragrantissima* (de Mendoza *et al.* 2015). Reconstruction of gene family evolution has shown a mixed contribution of old and new genes to the origin of multicellularity. First, the unicellular ancestor of Metazoa was already equipped with a rich repertoire of genes involved in multicellular functions, including developmental transcription factors, cell adhesion and cell signaling. These mechanisms were later co-opted for multicellularity-related functions (King *et al.* 2008; Sebé-Pedrós *et al.* 2010, 2011; Nichols *et al.* 2012; Suga *et al.* 2013; Fairclough *et al.* 2013). Second, there was a process of novel gene evolution concomitant with animal origins that sets multicellular genomes apart from premetazoans, with the invention of 300-400 novel genes (Srivastava *et al.* 2010; Tautz and Domazet-Lošo 2011; Richter and King 2013; Simakov and Kawashima 2016). A common theme in both animal-specific and premetazoan gene families was the marked increase of protein diversity by the combined means of paralogy and shuffling of protein domains (King *et al.* 2008; Basu *et al.* 2008, 2009; Suga *et al.* 2012; Nichols *et al.* 2012). This expanded gene content in the LCA of Metazoa was poised to have played a role in its increased organismic complexity. Figure 5 offers an overview of the gene content in the ancestral animal genome (Richter and King 2013), whose origin and evolution I will examine in the following pages.

Co-option of ancient genes for multicellularity

One of the most surprising outcomes of early studies of premetazoan gene evolution was the high share of genes, classically considered metazoan-specific, that also existed in unicellular Holozoa, thus suggesting that functional co-option could have played an important role in the dawn of animal multicellularity (King *et al.* 2008; Sebé-Pedrós *et al.* 2011; Suga *et al.* 2013; Fairclough *et al.* 2013).

The most successful models to understand the role of gene co-option in multicellularity are volvocine chlorophytes, where a number of relatively simple changes in gene function have been found to be related to its simple multicellular functions (Olson and Nedelcu 2016). For example, paralogy and shifts in the regulatory network of a nearly-paneukaryotic cell cycle controller, the retinoblastoma gene, correlate with the occurrence of colonies in *Gonium pectorale* and *Volvox carteri*, in opposition to the unicellular *Chlamydomonas reinhardtii* (Hanschen *et al.* 2016). This kind

of studies have the potential to unravel the logic and necessities behind the frequent co-option occurred at the transition to Metazoa. In particular, this phenomenon is involved in multiple animal innovations: 1) cell adhesion, 2) spatial-temporal regulation of transcription, and 3) spatial-temporal regulation of signal transduction (Richter and King 2013).

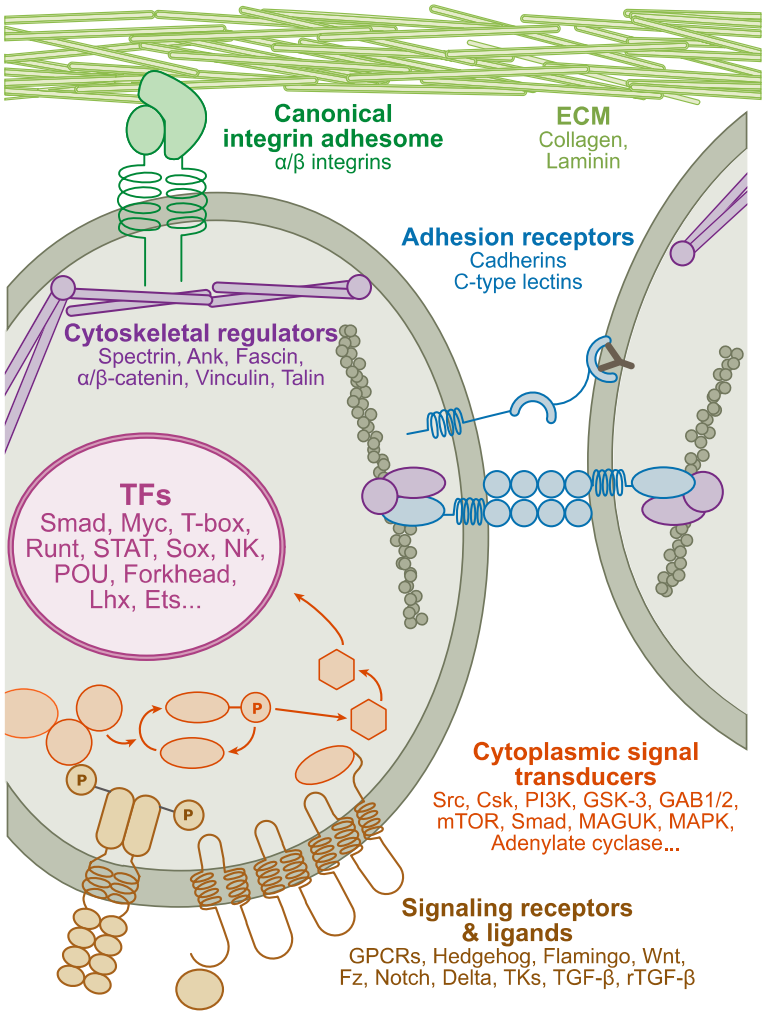


Figure 5. Genome content of the reconstructed LCA of Metazoa. Based on (Richter and King 2013).

Cell adhesion molecules involve both direct cell-to-cell contacts and indirect adhesion to an extracellular matrix (ECM), a highly specialized, multi-protein structure that provides a physical support for multicellular tissues (Hynes 2012). Cadherins and C-type lectins are proteins involved in animal cell adhesion that are also found in *C. owczarzaki* and choanoflagellates, although these unicellular and multicellular homologs are not orthologous (Nichols *et al.* 2012; Levin *et al.* 2014). Interestingly, the colony-forming factor of *S. rosetta*, the *rosetteless* gene, is a C-type lectin, thus hinting at some sort of two-way co-option at the respective origins of metazoan and choanoflagellate multicellularity (Levin *et al.* 2014). With respect to adhesion to the ECM, the complete gene tool-kits to build an integrin adhesome and its associated signal-transducing complex has been identified in the filasterean *C. owczarzaki* (Sebé-Pedrós *et al.* 2010) and the ichthyosporean *C. fragrantissima* (de Mendoza *et al.* 2015). Many of the protein modules involved in

building the animal ECM, like fibrillar collagens and specific laminins, are of ancient eukaryotic, opisthokont or holozoan origin, but the fully formed ECM is a uniquely metazoan structure (Exposito *et al.* 2008; Hynes 2012; Richter and King 2013; Cromar *et al.* 2014). Therefore, the presence of ECM adhesion molecules in the ECM-devoid unicellular holozoans has been proposed to be linked with an ancestral role in extracellular signaling, later co-opted for regulated adhesion to the ECM substrate (Sebé-Pedrós and Ruiz-Trillo 2010). This hypothesis fits with the up-regulation of the integrin adhesome in the dispersive amoeboid stage of *C. fragrantissima* but not in its coenocytic, multinucleate phase (de Mendoza *et al.* 2015). Incidentally, choanoflagellate cadherins have also been proposed to have extracellular sensing functions, *e.g.* in interactions with bacterial preys (Abedin and King 2008).

Pre-metazoan gene origin also shapes the evolution of the animal complement of transcription factors (TFs), frequently including developmental regulators (Sebé-Pedrós *et al.* 2011; de Mendoza *et al.* 2013). For example, choanoflagellates, *C. owczarzaki* and ichthyosporeans have homologs of the animal *p53* tumor repressor (Sebé-Pedrós *et al.* 2011), and, although its unicellular functions are unknown, it appears to have been co-opted as a DNA damage control switch in Metazoa (Srivastava *et al.* 2010). However, the function of *C. owczarzaki* p53 appears to be regulated by dynamic phoshporylation mechanisms similar to those acting in Metazoa (Sebé-Pedrós *et al.* 2016a). Another example of premetazoan TF origin is the animal T-box family *Brachyury.* In rescue experiments in *Xenopus* embryos, *Brachyury*'s function in gastrulation could be recapitulated using the *C. owczarzaki* ortholog due to a conserved DNA-binding motif and cofactors (Sebé-Pedrós *et al.* 2013b). Moreover, a survey of *C. owczarzaki*'s regulatory landscape showed that the downstream targets of its *Brachyury* homolog include genes involved in cell migration, a basic cellular process common to gastrulating embryos and *C. owczarzaki*'s crawling amoebas (Keller 2005; Sebé-Pedrós *et al.* 2016b), suggesting a co-option process followed by further elaboration of the TF downstream network. Overall, the presence of rich TF repertoires in the immediate unicellular ancestry of Metazoa points at a 'pre-adaptive' expansion: phylogenetic inertia drives the diversification and *de novo* origin of TF families before the emergence of multicellularity, laying the basic regulatory switches that are later co-opted for developmental processes (de Mendoza *et al.* 2013).

The establishment of diverse signal transduction pathways is also essential in order to coordinate the functions of a multicellular body (Bonner 2000b; King *et al.* 2003). One of the major signaling systems of eukaryotes is protein phosphorylation, by which protein products can be labeled with phosphate groups in specific residues. These phosphorylation systems are involved in regulating a myriad of cellular processes: cell-to-cell and cell-to-matrix adhesion, proliferation, development, or differentiation (King 2004; Hunter 2009). Tyrosine-specific kinases, together with serine/threonine kinases, are the dominant phosphorylation systems of eukaryotes (Choi *et al.* 2008) and consist of a wide array of highly diverse gene families that are thoroughly conserved in Metazoa. However, recent studies have also identified important enrichments in their closest unicellular relatives, like *M. brevicollis* (Manning *et al.* 2008), *C. owczarzaki*, *M. vibrans* (Suga *et al.* 2012), and ichthyosporeans (Suga *et al.* 2014). Interestingly, this holozoan-wide expansion of phosphotyrosine signaling was due to a dual evolutionary trend by which the cytoplasmic enzymes tend to be conserved across holozoan genomes, but the membrane-bound receptor enzymes are largely lineage- or species-specific (Suga *et al.* 2014). For example, only 20% of *M.*

brevicollis receptor tyrosine kinases are conserved in its close choanoflagellate relative *S. rosetta* (Fairclough *et al.* 2013). However, these studies identified a basic set of receptor kinases that are common in unicellular holozoans and animals, which have been proposed to have been co-opted from extracellular sensing to intracellular communication with the emergence of multicellularity (Suga *et al.* 2012; Richter and King 2013). This hypothesis is supported by various observations. First, the tyrosine kinases from *M. brevicollis* exhibit variable expression with changing environmental cues (King *et al.* 2003). Second, some pan-holozoan tyrosine kinases involved in the integrin adhesome-mediated signaling processes, *e.g.* the cytoplasmic enzymes Src, Tec or FAK (Sebé-Pedrós *et al.* 2010; Suga *et al.* 2012, 2014), are co-regulated with the rest of the integrin adhesome in *C. owczarzaki* aggregative cells (Sebé-Pedrós *et al.* 2013a). Further indirect evidence of an early holozoan role of tyrosine kinase signaling in adhesion can be drawn from their pattern of phosphorylation in *C. owczarzaki*: FAK and Tec are phosphorylated and active in the aggregative stage; whereas Src, which has a conserved auto-phosphorylation activity and controls cell proliferation in animals (Schultheiss *et al.* 2012), is equally activate in the proliferative amoeboid stage of *C. owczarzaki* (Sebé-Pedrós *et al.* 2016a).

The study of the protein tyrosine kinase evolution in metazoans and their unicellular relatives has thus offered essential insights into the evolutionary dynamics of a genetically diverse signaling pathway, in which co-option and *de novo* origin are combined. A common theme in such genetic tool-kits is the widespread conservation of an element of the pathway (cytoplasmic tyrosine kinases), while upstream receptors and ligands (receptor tyrosine kinases) have a higher evolvability and develop lineage-specific adaptations. This theme is paralleled in the two-component system (Capra and Laub 2012) or the Hippo pathway (Sebé-Pedrós *et al.* 2012), but also in other typical metazoan signaling networks of later origin (see below).

Metazoa-specific gene innovations

As mentioned before, the emergence of Metazoa was accompanied by the origin of 300-400 completely novel gene families with no homology in their unicellular relatives (Srivastava *et al.* 2010; Tautz and Domazet-Lošo 2011; Simakov and Kawashima 2016). Notable examples are the animal-exclusive Wnt, TGF-β, Notch JAK/Stat and Hedgehog signaling pathways, all of which are involved in development regulation and cell type specification (Richards and Degnan 2009; Richter and King 2013). Crux to all these pathways is the emergence of novel genes that function as specific ligands or receptors, even if they often re-use ancient components and signal transducers. Take, for example, the Wnt pathway, that controls proliferation, cell differentiation, co-ordinated movement and polarity in both bilaterian and non-bilaterian Metazoa. The Wnt ligand has no distinguishable homology outside of Metazoa, but it interacts with a set of cytosolic/membrane proteins (Frizzled receptors, β-catenin or GSK3) that are also present in *Dictyostelium* and unicellular holozoans (Holstein 2012; Suga *et al.* 2013; Fairclough *et al.* 2013). Other essential transducers of the pathway, like the Tcf/Lef TFs or Dishevelled, are animal innovations as well (Sebé-Pedrós *et al.* 2011; Holstein 2012; Simakov and Kawashima 2016).

In addition to *de novo* gene origin, serial paralogy of preexistent genes also led to an increased diversity of certain genetic tool-kits. A classical example is that of the homeobox genes, a paneukaryotic TF family that underwent a massive expansion in Metazoa (de Mendoza *et al.*

2013) and is credited with having contributed to the explosion of new developmental body plans during the pre- and post-Cambrian periods (Peterson *et al.* 2005; Holland 2015). During early metazoan evolution, a handful of homeobox genes underwent serial tandem duplications, sometimes accompanied by novel protein domains, giving rise to 11 new transcription factor multi-gene classes: ANTP, PRD, LIM, POU, HNF, SINE, TALE, CUT, PROS, ZF and CERS; which contain over a hundred sub-classes (Holland *et al.* 2007). Although not all of these 11 classes are animal-exclusive (*e.g.*, a PRD-like homolog exists in *C. owczarzaki*; Sebé-Pedrós et al. 2011), the expansions at the sub-class level generally coincide with the Metazoa root. In addition, animal homeoboxes are typically associated in multiple syntenic clusters with common functions in development, which are thoroughly conserved in diverse extant Metazoa (Ferrier 2016).

Another example of genomic innovation that builds on preexisting genes is the myriad of new gene families based on the collagen domain that appeared at the Metazoa root (Simakov and Kawashima 2016). These genes encode a diverse array of proteins with repetitive motifs and multiple functions related to the ECM organization (structural fibrils like collagens of type XV/XVIII; or non-fibrillar type IV and IV-like/spongins), or to signaling functions (collectin receptors, C-type lectin sub-families, etc.; Aouacheria et al. 2006; Heino 2007; Exposito et al. 2008; Hynes 2012; Fahey and Degnan 2012). While the basic building blocks (collagen domains) exist outside of Metazoa, no unicellular animal relative is known to have an homologous collagen-based ECM (Richter and King 2013).

1.3.2.2. Genome structure and dynamics

In order to fully understand the genomic changes underlying the transition to multicellularity, a picture drawn from gene content analysis alone is forcibly incomplete: the contribution of non-coding genomic traits to shaping Metazoa genomes is key to pinpoint differences with their unicellular ancestors. Indeed, when compared to most eukaryotes, animal genomes appear to be distinctly larger (Elliott and Gregory 2015a), contain more (Csűrös *et al.* 2011) and longer introns (Elliott and Gregory 2015a), more transposable elements and repetitive sequences (Elliott and Gregory 2015a; b), are structured in a conserved patterns of gene linkage (Nakatani *et al.* 2007; Putnam *et al.* 2008; Irimia *et al.* 2012; Simakov *et al.* 2013; Smith and Keinath 2015). Animal genomes also harbour exclusive regulatory elements such as distal and developmental transcriptional enhancers (Sebé-Pedrós *et al.* 2016b; Gaiti *et al.* 2016, 2017), and are thought to contain topologically associated genomic domains (TADs) that physically enable distal transcriptional regulation (proved in Bilateria, Lee and Iyer 2012 and Seitan et al. 2013; hypothesized in Porifera, Gaiti et al. 2016). Some of these features are assuredly not exclusive to Metazoa: see, for example, the long introns in *Vitis vinifera* genome (Jaillon *et al.* 2007), the high intron densities in the chlorarachniophyte *Bigelowiella natans* (Curtis *et al.* 2012) or the dinoflagellate *Symbiodinium minutum* (Shoguchi *et al.* 2013), or the frequent repetitive element expansions of plants (Michael 2014)). However, the apparent coordination of their emergence in Metazoa deserves scrutiny in order to understand the transition to multicellularity (Table 1).

One of the hallmarks of animal genomes is the existence of conserved physical arrangements of gene order, a phenomenon known as synteny. Synteny conservation can occur at different levels,

encompassing a few dozen genes (microsynteny) or up to whole chromosomes (macrosynteny). It is the former that has received most of the attention, as it has been linked to co-regulation of the associated genes, or to commonalities in their functions (Simakov and Kawashima 2016). How and why do syntenic gene pairs appear and establish, however, is still a matter of debate. Across long evolutionary distances, conservation of gene linkage is expected to be low as genes randomly 'drift' through the genome due to recombination (Srivastava *et al.* 2008; Koonin and Wolf 2010). Indeed, gene linkage levels across eukaryotic lineages has been found to be nearly absent (Koonin 2009; Koonin and Wolf 2010). However, estimations of the number of conserved microsyntenic regions across animal genomes yield ~400 blocks at the animal LCA, and a later process of consolidation of additional blocks in the Bilateria root (Irimia *et al.* 2012; Simakov *et al.* 2013). The functional significance of these conserved gene architectures is frequently unclear, but some have been found to be co-regulated at the transcription level by common proximal promoters; or to be part of regulatory blocks involving *trans-dev* genes (involved in transcriptional regulation and/or development) and other by-stander genes that harbour regulatory sites for the *trans-dev* gene (for example, within their introns) (Irimia *et al.* 2012). Some well-characterized examples of *trans-dev* regulatory blocks are the homeobox clusters that emerged in early Metazoa, like Hox and ParaHox (Brooke *et al.* 1998; Duboule 2007; Irimia *et al.* 2012; Fortunato *et al.* 2014; Ferrier 2016), which have conserved functions in animal development (Holland 2013; Hudry *et al.* 2014). Interestingly, it has been recently reported that the Hox cluster is embedded within topologically associated domains (tri-dimensional genomic regions) in different vertebrate genomes (Dixon *et al.* 2012; Pope *et al.* 2014), and that the domain borders have shifted between amphioxus and vertebrates in parallel with changes in the transcriptomic regulation of the Hox cluster (Lonfat and Duboule 2015; Acemel *et al.* 2016). Overall, current studies support the view that the earliest establishment of microsyntenic blocks in Metazoa occurred at the root of the clade, with further refinement particularly in Bilateria. Indeed, microsyntenic blocks have not been reported between animals and earlier-branching unicellular holozoans (Suga *et al.* 2013), except for a handful of isolated gene pairs sharing collinearity between animals and *C. owczarzaki* (Irimia *et al.* 2012).

With respect to the evolution of genome size in Metazoa, it is inferred that the size of the animal LCA genome was larger (~300 Mb) than all currently known unicellular holozoans (most below ~100 Mb; King et al. 2008; Suga et al. 2013; Fairclough et al. 2013; Simakov and Kawashima 2016). This relationship, however, does not have a clear interpretation in terms of organismal complexity, as many disparate genomic traits can drive changes in genome size: repetitive elements, and chiefly transposable element propagation; the intron density and length; or also more 'classical' figures like the total number of genes or the percentage of coding sequence in the genome (Elliott and Gregory 2015a). For example, it is in the animal root where the most pronounced process of intron gain is inferred to have occurred, reaching the highest intron density of all eukaryotes (Carmel *et al.* 2007b; Csűrös *et al.* 2011). Similarly, Metazoa have higher rates of gene family gain with respect to loss, *e.g.* when measured by gene family (Borenstein *et al.* 2006; Tautz and Domazet-Lošo 2011) or presence of protein domains (Suga *et al.* 2013).

It must be noted that the above-mentioned structural genomic traits are frequently difficult to associate with direct adaptive advantages. It has been argued that they accumulate in metazoan genomes as a consequence of their sustained low effective population sizes, which diminishes the

effect of purifying selection (Lynch and Conery 2003), in a population-genetic effect termed mutational-hazard hypothesis (Lynch 2007; Lynch *et al.* 2011). This hypothesis aims to explain the high rate of intron insertion and lengthening (Lynch 2002; Csűrös *et al.* 2011), the transposable element invasions (Rho *et al.* 2010)—both phenomena are direct drivers of genome size change (Elliott and Gregory 2015a) while lacking an immediate adaptive effect—, and the rate of changes in gene order (Koonin and Wolf 2010). One must note, however, that the neutral emergence of a given structural genomic trait does not preclude a functional/adaptive role after the trait has been established (Lynch and Conery 2003; Lynch 2006a). Further evidence compromising the mutational-hazard hypothesis has recently emerged from intra-genome comparisons: the intron creation rates do not change when comparing genomic regions with disparate selective efficiencies (Roy 2016).

It is worth noting that a persistent dominance of drift and neutral changes in the structural evolution of animal genomes could be a problem in comparative studies with their unicellular relatives: protists typically have much higher effective population sizes than Metazoa (Lynch 2006b), and are therefore expected to be subject to more efficient purifying selection (Lynch and Conery 2003; Koonin 2011). This can preclude the conservation of ancestral genomic traits across the unicellular-to-multicellular divide, as some complex traits would only become (or remain) selectively sustainable in Metazoa (Koonin 2004; Lynch 2007), while some others may secondarily disappear in extant, stream-lined protist genomes (Lynch 2006b; Wolf and Koonin 2013). Comparative genomic analyses of such traits therefore need to consider the effect of such limitations.

2. Objectives

Expanding the outlook of holozoan comparative genomics

Stars! You know your place in the sky
your hold your course and your aim
and each in your season
returns and returns,
and is always the same.
And if you fall as Lucifer fell,
you fall in flames!

Javert– Les Misérables 1980, *Stars*

The general framework of my thesis is the reconstruction of ancestral genomes of metazoan and premetazoan lineages, by the means of comparative genomic analyses. To this end, I have focused on three main objectives:

1. Sequencing, assembly and annotation of new genomes of unicellular relatives of animals. This includes *Corallochytrium limacisporum*, and the ichthyosporeans *Pirum gemmata*, *Abeoforma whisleri* and *Chromosphaera perkinsii*.

2. Resolution of the phylogenetic relationships among species with newly sequenced genomes and/or transcriptomes, using the tools of phylogenomics.

3. Comparative genomics of unicellular and multicellular holozoans (Metazoa), plus additional eukaryotic genomes, in order to elucidate the genomic landscape of the metazoan ancestors. This approach consists in inferring the gene content and the genomic structure/architecture of various ancestors, from the last common ancestor of Metazoa to the last common ancestor of eukaryotes (LECA). It aims to illustrate relevant phenotypes, including cell biology and lifestyle, of these ancestors.

3. Results

New insights from comparative genomics between Metazoa and their unicellular relatives

Our little systems have their day;
they have their day and cease to be:
they are but broken lights of thee,
and thou, O Lord, art more than they.

Alfred Tennyson, In Memoriam A.H.H., 1849

Impact and authorship report of the publications

Director: Dr. Iñaki Ruiz-Trillo

Tutor: Dr. Marta Riutort León

Five out of the seven articles that conform this thesis dissertation (henceforth, Results R1-7) been published in high impact journals covering the fields of Evolutionary Biology, Genetics and Genomics, including both field-specific and multidisciplinary journals. The other two are unpublished manuscripts, one of which is under revision in *eLife*. The five published articles have been indexed in bibliographic databases (PubMed, ISI).

Among the eight manuscripts here presented, Xavier Grau-Bové has been the sole first author of four of them, and co-first author of two more. The two remaining articles are collaborations, either with fellow researchers in the laboratory of Iñaki Ruiz-Trillo, or the Department of Genetics, Microbiology and Statistics (Universitat de Barcelona).

The specific contributions of Xavier Grau-Bové to each publication are indicated in the following pages, together with the yearly impact factor and ranking of each journal (as per the ISI proprietary ranking).

Publication R1 – HECT evolution

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 Biol Evol*** 5: 833–47.

Impact Factor (2013): 4.532

Journal ranking: Evolutionary Biology Q1 (11/46); Genetics & Heredity Q1 (33/164)

Authorship: The project was conceived jointly by IRT and ASP. XGB performed, under the supervision of ASP, the phylogenetic analysis of HECT and the ancestral gene content reconstructions and protein domain architecture analyses. Data discussion and manuscript writing were carried out by XGB, ASP and IRT.

Publication R2 – Myosin evolution

Sebé-Pedrós A, Grau-Bové X, Richards TA, Ruiz-Trillo I. 2014. Evolution and classification of myosins, a paneukaryotic whole genome approach. ***Genome Biol Evol*** 6: 290–305.

Impact Factor (2014): 4.229

Journal ranking: Evolutionary Biology Q1 (9/46); Genetics & Heredity Q1 (37/167)

Authorship: co-first authorship with equal contributions with ASP. Project conception, experimental design, phylogenetic analyses, ancestral reconstructions and protein domain architecture analyses carried out by ASP and XGB. Data discussion and manuscript writing were carried out by XGB, ASP, TAR and IRT.

Publication R3 – Opisthokonta phylogenomics

Torruella G, De Mendoza A, Grau-Bové X, Antó M, Chaplin MA, Campo J Del, Eme L, Pérez-Cordó G, Whipps CM, Nichols KM, *et al.* 2015. Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. ***Curr Biol*** 25: 1–7.

Impact Factor (2015): 8.983

Journal ranking: Biochemistry and Molecular Biology Q1 (19/289); Cell Biology Q1 (20/187)

Authorship: collaboration in an article co-authored by GT and AdM, who designed and coordinated the study jointly with IRT. This article was part of the thesis dissertation of GT. XGB was in charge of performing microscopy analyses to demonstrate the existence of flagellar structures in *C. limacisporum* and *M. vibrans*, using both immunostaining coupled with high-resolution optical microscopy, and transmission electron microscopy, jointly with MA, ASB, and GT. In parallel, XGB also contributed to the assembly of the manuscript's data and figures. Phylogenomic analyses carried out by GT, LE and AJR. Gene phylogeny analyses by AdM. Data generation (culture, extractions and sequencing) by MAC, SD, KMN, CMW, RP, GT, JdC, MA, GPC and XGB. Manuscript written by AdM, GT and IRT.

Publication R4 – Ubiquitin signaling evolution

Grau-Bové X, Sebé-Pedrós A, Ruiz-Trillo I. 2015. The Eukaryotic Ancestor Had a Complex Ubiquitin Signaling System of Archaeal Origin. ***Mol Biol Evol*** 32: 728-739.

Impact Factor (2015): 13.649

Journal ranking: Biochemistry and Molecular Biology Q1 (5/289); Evolutionary Biology Q1 (2/46); Genetics & Heredity Q1 (4/166)

Authorship: co-first authorship with equal contributions with ASP. Project conception, experimental design, phylogenetic analyses, ancestral reconstructions, protein domain architecture analyses and data discussion carried out by ASP and XGB. The manuscript was written by XGB, ASP and IRT.

Publication R5 – LOX evolution

Grau-Bové X, Ruiz-Trillo I, Rodriguez-Pascual F. 2015. Origin and evolution of lysyl oxidases. ***Sci Rep*** 5: 10568.

Impact Factor (2015): 5.228

Journal ranking: Multidisciplinary Sciences Q1 (7/63)

Authorship: XGB was in charge of all the phylogenetic analyses and preparation of the data. Project conceived by FRP. Experimental design, analyses and manuscript writing carried out jointly by XGB, IRT and FRP.

Unpublished result R6 (under revision in *eLife*) – Teretosporea genomes

Grau-Bové X, Torruella G, Donachie S, Suga H, Leonard G, Toulis V, Richards TA, Ruiz-Trillo I. 2017. Dynamics of genomic innovation in the unicellular ancestry of animals. ***eLife*** (under revision).

Impact Factor (2017): NA – previous available year: 8.282

Journal ranking: NA – previous available year: Biology Q1 (4/86)

Authorship: XGB and IRT designed and coordinated the study. Experimental design, data analysis, preparation and discussion by XGB. Phylogenomic analyses carried out jointly by GT and XGB. Genome assembly by XGB, TAR, HS and GL. Genome annotation by XGB. Comparative genomic analyses, gene phylogenies and ancestral reconstructions by XGB. Manuscript written by XGB.

Unpublished result R7 (not yet submitted) – Alternative Splicing evolution

Grau-Bové X, Ruiz-Trillo I, Irimia M. 2017. Correlated evolution of alternative splicing and gene architecture across eukaryotes. Unpublished.

Impact Factor (2017): NA

Journal ranking: NA

Authorship: MI conceived the study and the analytical framework for RNA-seq data. Experimental design by XGB and MI. Data analysis, preparation and discussion by XGB. Manuscript written by XGB.

3.1. A genomic survey of HECT ubiquitin ligases in eukaryotes reveals independent expansions of the HECT system in several lineages

Abstract – The posttranslational modification of proteins by the ubiquitination pathway is an important regulatory mechanism in eukaryotes. To date, however, studies on the evolutionary history of the proteins involved in this pathway have been restricted to E1 and E2 enzymes, whereas E3 studies have been focused mainly in metazoans and plants. To have a wider perspective, here we perform a genomic survey of the HECT family of E3 ubiquitin-protein ligases, an important part of this posttranslational pathway, in genomes from representatives of all major eukaryotic lineages. We classify eukaryotic HECTs and reconstruct, by phylogenetic analysis, the putative repertoire of these proteins in the last eukaryotic common ancestor (LECA). Furthermore, we analyze the diversity and complexity of protein domain architectures of HECTs along the different extant eukaryotic lineages. Our data show that LECA had six different HECTs and that protein expansion and N-terminal domain diversification shaped HECT evolution. Our data reveal that the genomes of animals and unicellular holozoans considerably increased the molecular and functional diversity of their HECT system compared with other eukaryotes. Other eukaryotes, such as the Apusozoa *Thecanomas trahens* or the Heterokonta *Phytophthora infestans*, independently expanded their HECT repertoire. In contrast, plant, excavate, rhodophyte, chlorophyte, and fungal genomes have a more limited enzymatic repertoire. Our genomic survey and phylogenetic analysis clarifies the origin and evolution of different HECT families among eukaryotes and provides a useful phylogenetic framework for future evolutionary studies of this regulatory pathway.

A Genomic Survey of HECT Ubiquitin Ligases in Eukaryotes Reveals Independent Expansions of the HECT System in Several Lineages

Xavier Grau-Bové[1], Arnau Sebé-Pedrós[1,*], and Iñaki Ruiz-Trillo[1,2,3]

[1]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Catalonia, Spain

[2]Departament de Genètica, Universitat de Barcelona, Catalonia, Spain

[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

*Corresponding author: E-mail: arnau.sebe@ibe.upf-csic.es.

Accepted: March 29, 2013

Data deposition: Protein alignments from this project have been deposited at http://www.datadryad.org under the accession doi:10.5061/dryad.mt620.

Abstract

The posttranslational modification of proteins by the ubiquitination pathway is an important regulatory mechanism in eukaryotes. To date, however, studies on the evolutionary history of the proteins involved in this pathway have been restricted to E1 and E2 enzymes, whereas E3 studies have been focused mainly in metazoans and plants. To have a wider perspective, here we perform a genomic survey of the HECT family of E3 ubiquitin-protein ligases, an important part of this posttranslational pathway, in genomes from representatives of all major eukaryotic lineages. We classify eukaryotic HECTs and reconstruct, by phylogenetic analysis, the putative repertoire of these proteins in the last eukaryotic common ancestor (LECA). Furthermore, we analyze the diversity and complexity of protein domain architectures of HECTs along the different extant eukaryotic lineages. Our data show that LECA had six different HECTs and that protein expansion and N-terminal domain diversification shaped HECT evolution. Our data reveal that the genomes of animals and unicellular holozoans considerably increased the molecular and functional diversity of their HECT system compared with other eukaryotes. Other eukaryotes, such as the Apusozoa *Thecanomas trahens* or the Heterokonta *Phytophthora infestans*, independently expanded their HECT repertoire. In contrast, plant, excavate, rhodophyte, chlorophyte, and fungal genomes have a more limited enzymatic repertoire. Our genomic survey and phylogenetic analysis clarifies the origin and evolution of different HECT families among eukaryotes and provides a useful phylogenetic framework for future evolutionary studies of this regulatory pathway.

Key words: ubiquitination pathway, posttranslational regulation, multicellularity, last common ancestor of eukaryotes, Holozoa.

Introduction

Proteins are the main structural and functional components of all cells. To efficiently respond to different environmental conditions, the protein levels need to be constantly regulated. The ubiquitination pathway is one of the most important posttranslational mechanisms for regulating protein turnover and molecular cell dynamics (Rotin and Kumar 2009). It is based on the posttranslational modification of proteins by the ligation of ubiquitin, a 76 amino acid signaling peptide that is conserved across eukaryotes. This ubiquitin flag targets the proteins to a number of different outcomes, such as protein degradation, membrane sorting, and signaling functions (Rotin and Kumar 2009). The ubiquitination pathway involves the sequential transfer of activated ubiquitin (Ub) from E1 (ubiquitin activating enzyme) to E2 (ubiquitin conjugating enzyme), and subsequently from E2 to E3 (ubiquitin ligase), which binds Ub to the protein of interest. E3 ubiquitin ligases transfer Ub to one or more Lys residues in the substrate by linking the C-terminal Gly of Ub with a Lys of the target protein (and/or a Lys of the Ub itself). Ubiquitination can occur in different forms (Mukhopadhyay and Riezman 2007): mono-ubiquitination (attachment of a single Ub to a single Lys), multi-ubiquitination (several Lys residues tagged with Ub) and polyubiquitination (addition of a Ub chain to a single Lys of the target protein). Typically, mono- and multi-ubiquitination are related to subcellular localization processes such as the secretory and endocytic pathways (Hicke 2001). Polyubiquitination, on the other

hand, directs proteins to the 26S proteasome (a multiprotein complex consisting of 19S regulatory and 20S catalytic subcomplexes), which recognizes ubiquitinated proteins and degrades them; a common fate for misfolded or damaged proteins (Pickart and Fushman 2004).

To date, several studies have been carried out to resolve the evolutionary history of the ubiquitination pathway from a pan-eukaryotic point of view. These studies have, however, focused on the most conserved elements of the system, that is, the E1 (Burroughs et al. 2009) and E2 enzymes (Burroughs et al. 2008; Michelle et al. 2009; Ying et al. 2009), revealing that this pathway is ancient and widely distributed in all the considered eukaryotic lineages—as it is also the case for the ubiquitin proteins themselves (Burroughs et al. 2007).

Conversely, most studies on E3 ubiquitin ligases have focused mainly on animals (Rotin and Kumar 2009; Marín 2010) and plants (Downes et al. 2003); and so little is known about the origin and evolution of these ligases within eukaryotes, and their relative importance in different eukaryotic lineages.

E3 ubiquitin ligases are of particular interest in evolutionary studies of the ubiquitination system, because they are way more diversified than E1 and E2 enzymes. The reason for this is that they are responsible for the specificity of the ubiquitination system, that is, they recognize, discriminate, and interact with the proper protein substrate (Rotin and Kumar 2009), and therefore are more functionally specialized. In fact, there are various groups of E3 enzymes according to their quaternary structure, their specific domain arrangements and the way in which they interact with E2 and the target protein. This includes, for instance, the HECT and RING ligases, and the CRL complexes. These proteins typically have a wide range of domain architectures involving specific protein–protein interaction motifs.

Indeed, the few eukaryotic genomes so far analyzed often encode many more E3 enzymes than E1 or E2. For example, there are more than 600 types of E3 in the human genome, whereas there are only two E1 proteins and approximately 30 E2 proteins (Schwartz and Ciechanover 2009).

HECT proteins are defined by the specific HECT domain, a C-terminal domain of approximately 350 amino acids that is essential for their Ub-ligase activity. The HECT domain is exclusive to HECT E3 ligases and is widespread among eukaryotes (Punta et al. 2012). HECT proteins directly intervene in the ligation process by forming an intermediate thioester bond between a highly conserved cysteine residue and Ub that binds Ub to the substrate (fig. 1) (Rotin and Kumar 2009).

Previous studies have devised a phylogenetic classification of animal HECTs (Marín 2010); however, there is little knowledge on the diversity of HECTs among all eukaryotes. Here, we perform a genomic survey of HECT ligases in eukaryotes and provide a useful evolutionary framework for future analyses. We also analyze the diversity of protein domain architectures of HECTs along the different eukaryotic lineages, as well as the putative relationship between the expansion of the HECT-dependent ubiquitination system and the origin of multicellularity in several eukaryotic clades.

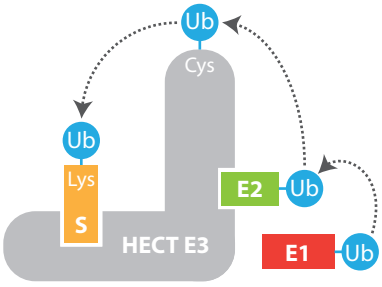


Fig. 1.—Schematic representation of Ub ligation to a protein substrate with a HECT ligase. The ligation process involves transferring the Ub from an activating enzyme (E1) to a transferase (E2) and then to the HECT ligase (E3). The E3 then ligates the Ub to a Lys residue of the substrate (S) with a thioester bond, involving a Cys residue in the HECT enzyme itself.

Materials and Methods

Taxon Sampling and Sequence Retrieval

HECT sequences were obtained from sequence data from complete genome sequences of 44 taxa, which represented all the recognized eukaryotic supergroups. Taxon sampling included 9 animals, 5 unicellular Holozoa, 8 Fungi, 1 Apusozoa, 3 Amoebozoa, 3 plants, 5 unicellular algae, 3 Heterokonta (1 being multicellular), and 11 other unicellular Bikonta (see supplementary table S1, Supplementary Material online). HECT amino acid sequences were retrieved with a HMMER search, using the HMM profile of the Pfam HECT domain entry (PF00632) as a query, the default parameters and an inclusive *E* value of 0.05. The search yielded 744 sequences (see supplementary fig. S3, Supplementary Material online).

Protein Alignment, Manual Edition, and Data Curation

The retrieved sequences were aligned using Mafft (Katoh et al. 2002) L-INS-i algorithm (optimized for local sequence homology [Katoh et al. 2005]). The alignment was further edited manually and hits fulfilling one of the following conditions were removed: 1) incomplete sequences with more than 99% of sequence similarity with a complete sequence from the same taxa, and 2) sequences that showed extreme long branches in the preliminary maximum likelihood (ML) trees. The final alignment was carried out based on the HECT domain alone using the Mafft G-INS-i algorithm (for global homology).

Phylogenetic Analyses

The phylogenetic trees of eukaryotic HECTs were inferred from both ML and Bayesian inference (BI) analyses, using

the LG evolutionary model with a discrete gamma distribution of among-site variation rates (four categories) and a proportion of invariable sites, which constituted the best model for this data set, according to Prottest (Abascal et al. 2005).

ML trees were estimated with RAxML 7.2.6 (Pthreads version [Stamatakis 2006]) and the best tree from 100 replicates was selected. Bootstrap support (BS) was calculated from 500 replicates. BI trees were estimated with Phylobayes 3.3 (Lartillot et al. 2009), using two parallel runs for 500,000 generations and sampling every 100. Bayesian posterior probabilities (BPPs) were used for assessing the statistical support of each bipartition.

Domain Architecture Analysis

The N-terminal domain architecture of all retrieved sequences was inferred by performing a Pfam scan (Punta et al. 2012), using the gathering threshold as cut-off value. The domain information of each protein was used to 1) assess the reliability of each sequence of the initial data set, 2) help define protein families according to its architectural coherence, and 3) assess the level of functional and architectural diversification of HECT proteins across the eukaryote lineages. Additional information about some previously uncharacterized domain architectures was obtained from the bibliography and verified using manual protein alignments. The pattern of acquisition of new domains at the N-terminus of HECT proteins across the eukaryote tree of life was inferred using a strict parsimony approach based on phylogenetic information from BI and ML trees.

Classification Criteria

The classification of the HECT proteins is based on two hierarchical categories: 1) protein families, which contain all proteins from orthologous genes with high nodal support, and 2) protein classes with one or more families, which are wider groups of phylogenetically related families that descend from one of the HECT proteins that have been inferred to exist in the last eukaryote common ancestor (LECA). Protein families sometimes share a common domain architecture, and therefore the domain content of each protein was used as an additional, conditional criterion to define some families. The pattern of gain and loss of families was inferred by strict parsimony based on phylogenetic information from BI and ML trees.

Results and Discussion

The Evolutionary Origin of HECT E3 Protein Family

Our phylogenetic analyses recovered six pan-eukaryotic clades of HECT proteins, defined as classes I to VI (figs. 2 and 3). Assuming the leading hypothesis that the root of eukaryotes lies between Unikonta and Bikonta (Stechmann and Cavalier-Smith 2002; Derelle and Lang 2012), our data imply that the last eukaryotic common ancestor had at least six HECTs that remain present in diverse eukaryotic lineages. In turn, these six main classes are divided into 35 distinct HECT families that are specific to certain eukaryotic lineages (fig. 3). This scenario remains the same if the alternative "Excavate-first" hypothesis of the root of the eukaryotes is considered (Rodríguez-Ezpeleta et al. 2007).

The diversification of each class involves many gene duplication events and secondary losses (fig. 4), as well as the acquisition of new accessory domains. Our data show that the protein domain architecture is quite diverse as a result of domain rearrangements and the acquisition of new domains at the N-terminal region (fig. 3).

Remarkably, domain fusions at the C terminus have not been detected in any of the analyzed organisms. This might be explained by the fact that the catalytic activity of the HECT domain strongly depends on its tertiary structure: all HECTs are organized in two structurally distinct lobes (N-lobe and C-lobe, where HECT is located) that can adopt a limited range of three-dimensional conformations (Huang et al. 1999; Verdecia et al. 2003; Rotin and Kumar 2009). This tertiary structure is functionally relevant (and therefore constrained) because it defines the position of the catalytic cysteine residue with respect to the E2 enzyme and the ubiquitination substrate during the ligation process (Verdecia et al. 2003). It also determines the way in which the ubiquitin chain elongation occurs (Maspero et al. 2011).

Assuming the "Unikont–Bikont split" hypothesis on the root of eukaryotes (Stechmann and Cavalier-Smith 2002; Derelle and Lang 2012), the analysis of protein domain architectures reveals class-specific N-terminal domain arrangements that are pan-eukaryotically distributed in classes I (SPRY), V (IQ), and VI (DUF908, DUF913, UBA, and DUF4414), whereas the founding proteins of classes II, III, and IV (Rodríguez-Ezpeleta et al. 2007), a similar scenario emerges, except for the ancestral IQ (class V), DUF908, DUF913, and UBA domains (class VI), which are not recovered. However, DUF4414 (class VI) still appears to be present in the LECA.

The syntax of N-terminal domain architectures in HECTs is mainly based on protein recognition motifs (IQ, WW, Ankyrin repeats, zinc fingers, etc.) that enable HECTs to specifically ubiquitinate certain substrates. Domains involved in targeting the HECT enzyme to certain molecules are also common, such as C2 (lipid binding), Laminin-G3 (complex sugar binding), and PABP (mRNA polyadenylate binding). Some of these motifs are especially "promiscuous" and have been independently gained several times thorough HECT evolution (for instance, ubiquitin-binding UBA and protein-binding domains such as WWE, SPRY, RCC1-like domain [RLD], Ankyrin, and MIB-HERC2) (fig. 5; details discussed later). Despite the generally conserved syntax of HECT N-terminal architectures, rare domains with no clear function exist on some uncharacterized HECTs. It is expected that the discovery of such unusual HECTs

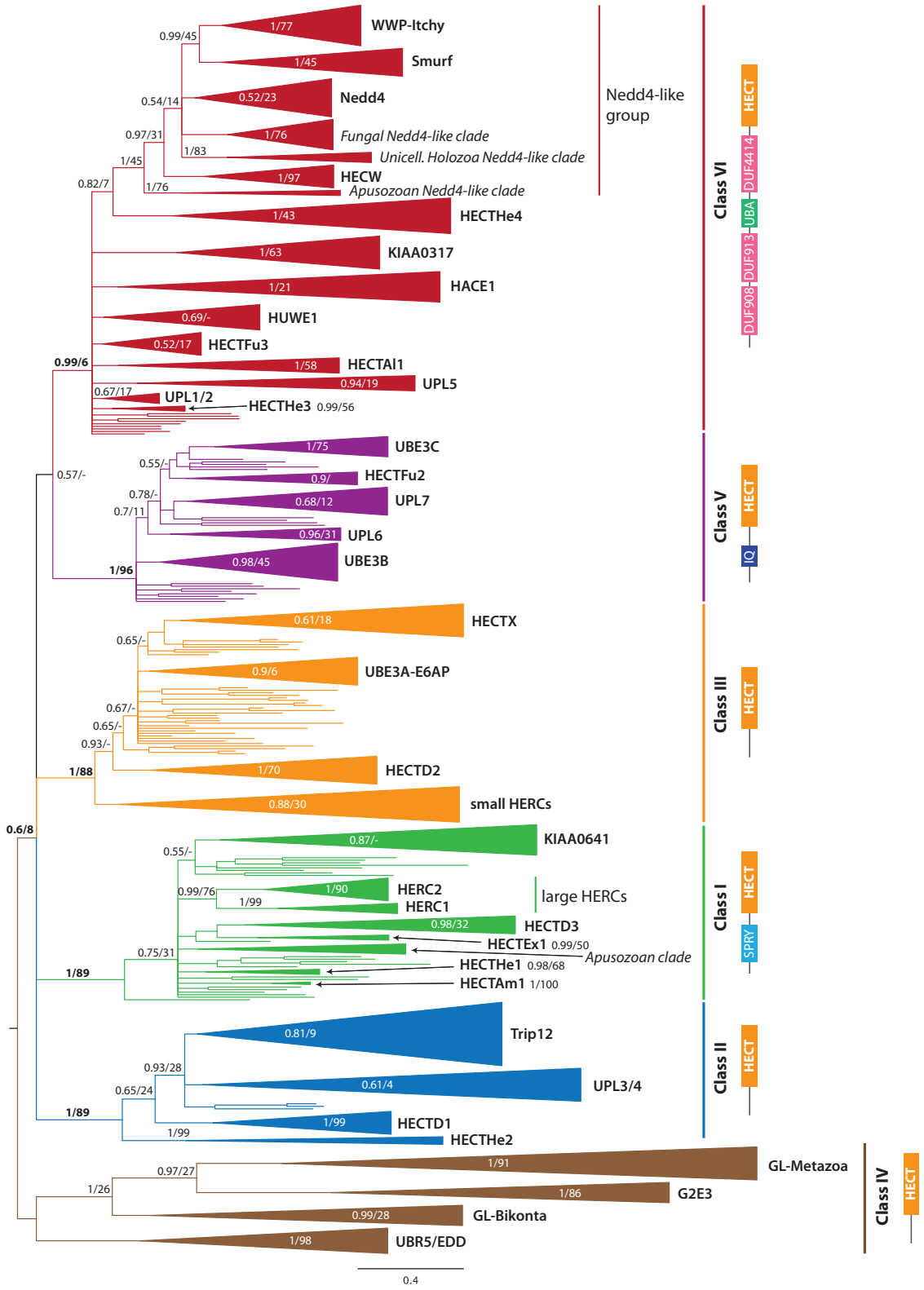


Fig. 2.—BI phylogenetic tree of HECT proteins inferred from an alignment of the HECT domain (220 amino acidic positions). Colored clades indicate classes; collapsed clades indicate families (in regular text) and other clades of interest (italics). Nodal labels indicate BPP and 500-replicate ML BS values, respectively. Dashes indicate that the node is not recovered. Six pan-eukaryotic classes can be distinguished, with 35 families within these. For each class, the putative ancestral N-terminal architecture is shown. Complete BI and ML trees are shown in supplementary figures S1 and S2, Supplementary Material online.

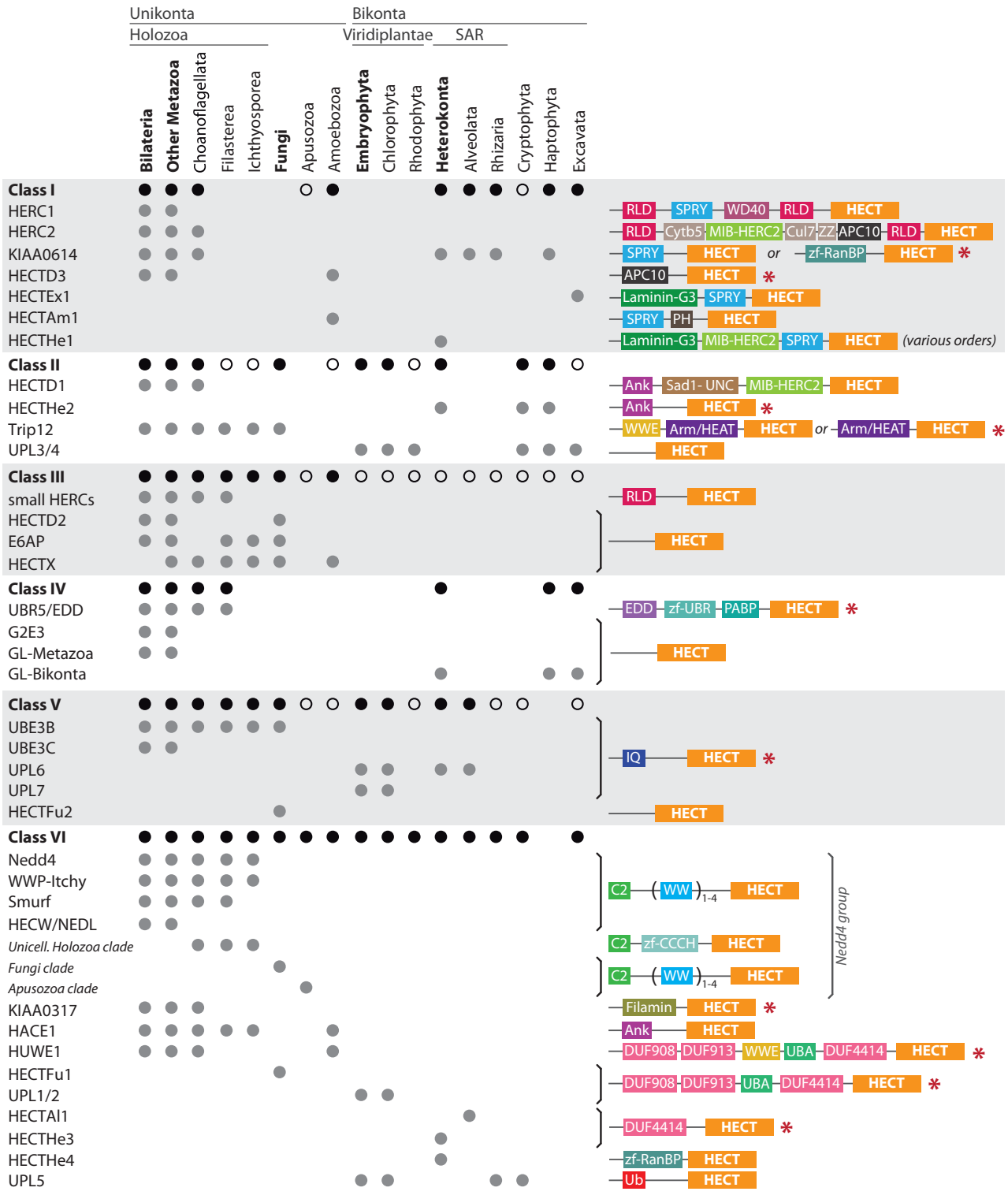


Fig. 3.—Schematic representation of HECT E3 classes and families with their archetypical domain architectures and presence/absence information in the sampled taxonomic groups. Filled circles signify that a given taxon has representatives in a certain family; empty circles signify that proteins of a given taxon are part of a class but cannot be classified into any described family. For each family, its typical architecture is shown on the right. Red asterisks indicate that a single HECT domain is also found in some sequences of that family. For more information on the architecture of each analyzed protein (supplementary figs. S1 and S2, Supplementary Material online). Taxa containing multicellular organisms are shown in bold (namely, Bilateria, other Metazoa, Fungi, Embryophyta, and Heterokonta).

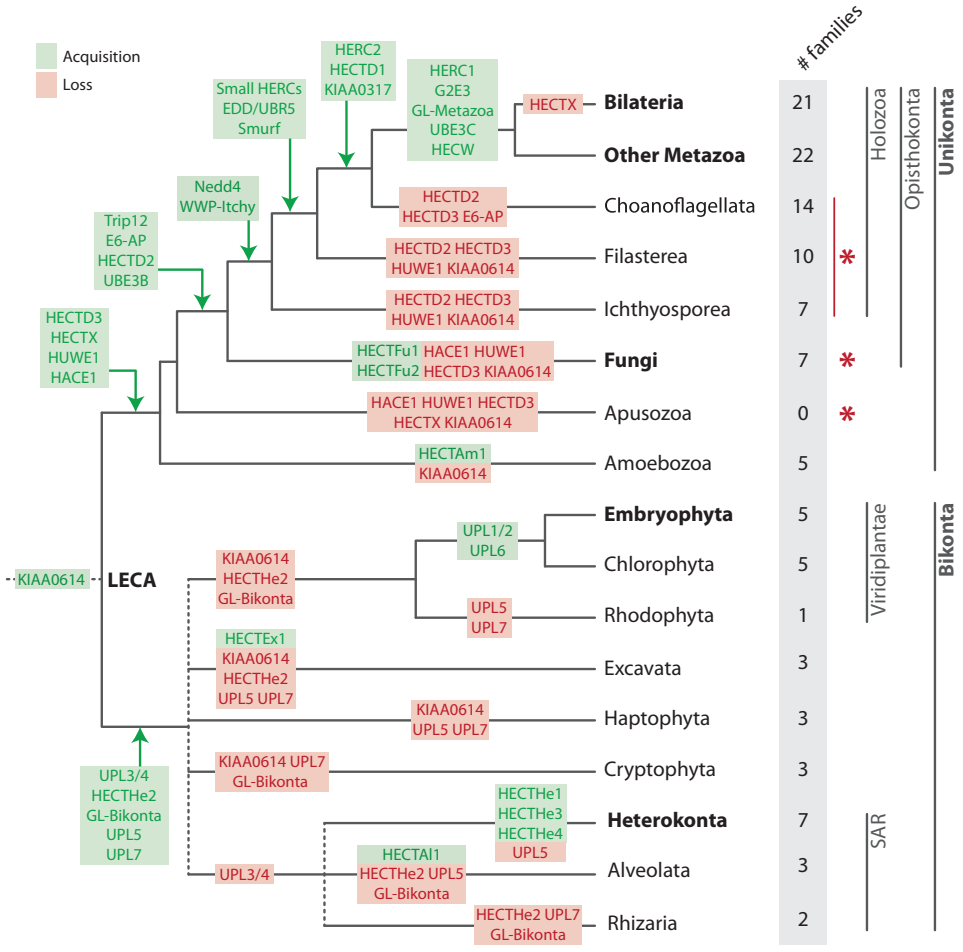


Fig. 4.—Pattern of acquisition and loss of HECT families across the eukaryote tree of life, assuming that LECA is placed at the Unikonta–Bikonta split. Green and red boxes indicate acquisition and loss of families, respectively. The number of families with representatives from each clade is shown at the right. Red asterisks indicate taxonomic groups that have one additional HECT clade that cannot be classified as part of any family (but are still known to be part of the Nedd4 group). Taxa containing multicellular organisms are shown in bold.

will increase when more and more genomes are taken into account in future similar surveys.

Classification of Eukaryotic HECT E3 Ligases

We have classified the different eukaryotic HECTs in different classes and families, according to the topology obtained by the phylogenetic analyses. A description of the main characteristics of each class and family is given in the following section.

Class I: Large HERCs and Related Families

Class I contains seven protein families: HERC1, HERC2 (both known as large HERCs), KIAA0614, HECTD3, HECTXEx1, HETCAm1, and HECTHe1 (figs. 2 and 3). The monophyly of class I is supported by a BPP of 1.0 and a BS value of 89% (fig. 2). Large HERCs were previously thought to be related to the family of small HERCs (class III in our tree), because they shared the RLD (Hadjebi et al. 2008), but our data corroborate that these families are paraphyletic and the domains have been independently acquired (Gong et al. 2003; Marín 2010).

HERC1 is an animal-specific family that has been lost in Arthropoda (*Daphnia pulex* and *Drosophila melanogaster*) and Hemichordata (*Saccoglossus kowalevskii*). HERC1 proteins have a specific domain architecture consisting of HECT, two RLDs, SPRY, and a variable number of WD40 repeats. In some cases, there is also a UBA domain. In humans, HERC1 binds to clathrin heavy chain and has GEF activity on ARF1, a GTPase involved in membrane trafficking in the Golgi apparatus (Rosa and Casaroli-Marano 1996). HERC1 also ubiquitinates the tumor suppressor TSC2 (involved in the tuberous sclerosis complex disease and perhaps in membrane trafficking [Chong-Kopera et al. 2006]).

The HERC2 family, which appears as a sister group to HERC1, is closely related to HERC1 and includes proteins from both Metazoa and Choanoflagellata. In mammals,

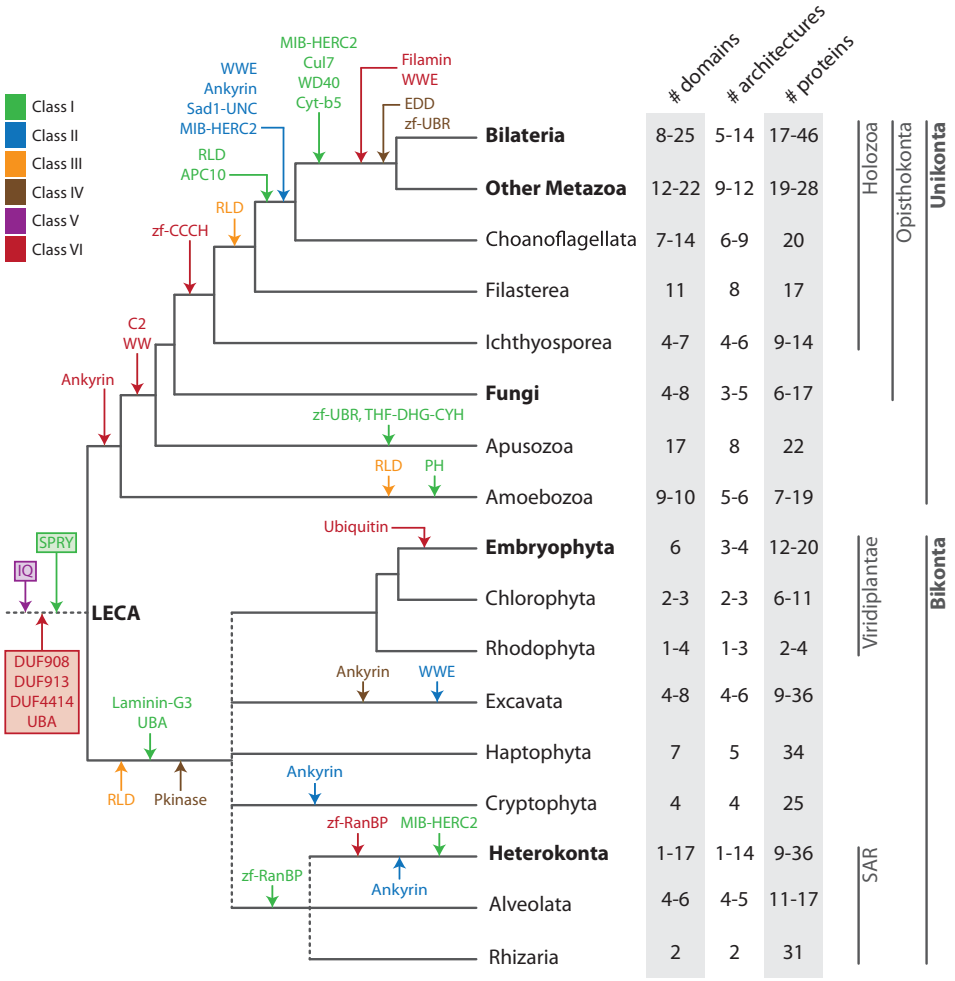


Fig. 5.—Pattern of appearance of new domains in HECT proteins within the tree of eukaryotes. Domains are color coded by class (see legend). Each label indicates the acquisition of a new domain, considering that the appearances of the same domain in different classes are independent acquisitions. Note that the domain list is nonextensive, because it only takes into account domains from proteins that have been phylogenetically classified. Domains in boxes at the LECA indicate the inferred domain content of the ancestral protein that gave rise to classes I (SPRY), V (IQ domain), and VI (DUF908, DUF913, DUF4414, and UBA). The ancestral architectures of the classes III, II, and IV LECA proteins were a single HECT domain. Note that RLD, WWE, Ankyrin, MIB-HERC2, and UBA have been convergently acquired several times in the evolution of different classes. The adjacent table contains information about the number of HECTs in each clade, the number of distinct domains in these proteins and the number of distinct domain architectures (shown as intervals; see supplementary table S2, Supplementary Material online, for more detail). Taxa containing multicellular organisms are shown in bold.

HERC2s ubiquitinate and target BRCA1 (breast cancer suppressor) for degradation (Wu et al. 2010). They have a complex domain architecture with two RLDs and several protein recognition motifs: Cyt-b5 (Ozols 1989), MIB-HERC2 (also present in RING E3 Mib2 [Itoh et al. 2003]), Cul7 (present in RING E3s Cul7 [Kaustov et al. 2007]), ZZ, and APC10. This architectural diversification occurred at the origin of the Metazoa, since the choanoflagellate homologs from both *Monosiga brevicollis* and *Salpingoeca rosetta* have simpler architectures (RLD repeats and RLD, APC10, and SPRY domains, respectively).

The KIAA0614 family is a pan-eukaryotic family with homologs in Metazoa, Choanoflagellata, Heterokonta, Alveolata, Rhizaria, and Haptophyta. Some proteins have a SPRY domain, while proteins from *Phytophthora infestans* and *Tetrahymena thermophila* have an extra zf-RanBP.

The HECTD3 family contains animal proteins (bearing an APC10 domain) and a homolog from *Acanthamoeba castellanii*. Human HECTD3 ubiquitinates some proteins involved in neural development and brain function, such as Syntaxin-8 (Zhang et al. 2009) and Tara—which is also a regulator of cell growth, cytoskeletal actin reorganization and cell motility (Yu et al. 2008).

HERC2 and HECTD3 are the only HECT families with APC10 domains, and they both are exclusive to animals and choanoflagellates. APC10 domain is also found in the RING E3 APC/C

complex, which takes part in cell cycle control by regulating mitosis (Jin et al. 2008). In this context, APC10 is responsible for the regulation of substrate binding (Peters 2002).

The other families within this class (i.e., HECTEx1, HECTAm1, and HECTHe1) are named after their taxonomic content (Excavata, Amoebozoa, and Heterokonta) and are defined by their distinctive domain arrangements. For instance, HECTAm1 contains PH and SPRY motifs, and HECTHe1 and HECTEx1 have Laminin-G3 (capable to reversibly bind to specific complex sugars, an exclusive feature of these two families) and SPRY domains. Also, class I contains a clade with *Thecanomas trahens* proteins bearing various protein recognition domains that seem to have been independently acquired (fig. 2).

The SPRY domain is exclusive to class I HECTs and is present in most of its families, which suggests that it could have existed in the ancestral LECA protein that gave rise to this class. It has been reported that SPRY plays a role in the recognition of ubiquitination substrates (Nishiya et al. 2011).

Class II

The well-supported class II (BPP = 1.0; BS = 89%) is composed of four protein families: HECTD1, HECTHe2, UPL3/4, and Trip12 (figs. 2 and 3).

The HECTD1 family contains sequences from Metazoa and Choanoflagellata. They have a distinctive protein domain arrangement containing Sad1-UNC, MIB-HERC2 domains and, in some cases, Ankyrin repeats. Human HECTD1 polyubiquitinates Hsp90, a chaperone that controls cell motility, which is essential in brain development (Sarkar and Zohn 2011). The HECTHe2 family also contains proteins with Ankyrin repeats and is specific to Heterokonta, Cryptophyta, and Haptophyta. Their functions are still unknown.

Trip12 (also known as ULF) includes proteins from animals, unicellular Holozoa and Fungi. Animal Trip12s are defined by two protein recognition domains: HEAT repeats, which are Armadillo-like motifs that recognize ubiquitin degradation signals in E3s substrates (Tewari et al. 2010); and WWE, which recognizes the Ankyrin motif of Notch and ligand-binding domains of other proteins (Aravind 2001). Fungal Trip12s also have HEAT/Armadillo repeats with a similar function, for example, the yeast Ufd4 HECT (Tewari et al. 2010).

Trip12 activity hampers tumor suppression in humans by preventing the p53 response to oncogenic events: it promotes the degradation of ARF, an inhibitor of the RING E3 Mdm2 (which in turn targets p53 for degradation [Brooks and Gu 2006]). Trip12 also targets p16 (a murine negative cell cycle regulator during embryogenesis) to degradation (Kajiro et al. 2011).

The UPL3/4 family includes homologs from several Bikonta clades (Viridiplantae, Excavata, Cryptophyta, Haptophyta, and Rhodophyta). Some Viridiplantae proteins also have Armadillo repeats, which have been predicted to recognize nuclear localization signals (Downes et al. 2003). *Arabidopsis* UPL3 polyubiquitinates some unknown regulator of trichome development (Downes et al. 2003); and both UPL3 and UPL4 collaborate in the regulation of Gibberellin cell signaling (Coates 2008). However, concrete substrates remain elusive.

Class III: Small HERCs, E6AP, and Other Families

Class III (BPP = 1.0; BS = 88%) includes small HERCs, HECTD2, E6AP (all of them named after the human proteins within them), and HECTX (Marín 2010) composed of Unikonta proteins. However, class III also includes proteins from Bikonta species (Viridiplantae, SAR, Cryptophyta, Haptophyta, and Excavata) that cannot confidently be assigned to any family, branching in an unclear position related to HECTD2, E6AP, and HECTX, but with low nodal supports.

The family of small HERCs includes proteins from animals, Choanoflagellata and Filasterea clades. It embodies human proteins HERC3, 4, 5, and 6, that is, the remaining HERC proteins that were formerly considered to be closely related to large HERCs 1 and 2 (see class I). So, any a priori functional or evolutionary similarities between these families need to be re-assessed. For instance, in contrast to large HERCs, the RLD motifs from small HERCs do not act as guanine nucleotide exchange factors (Rotin and Kumar 2009).

Indeed, convergent acquisition of RLD domains seems to be a common event in HECT evolutionary history: they are also present in several non-holozoan "HERC-like" proteins that cannot be assigned to any specific family (*A. castellanii*, *Toxoplasma gondii*, *Ectocarpus siliculosus*, *Cyanidioschyzon merolae,* and *Emiliania huxleyi* from class III; and *P. infestans* from class I). RLD domains intervene in a wide variety of cellular processes (RNA processing and transport, RNA mating, imitation of mitosis, chromatin condensation, guanine-nucleotide-exchange factor, protein recognition in DNA binding, and ubiquitination), which could explain their high "promiscuity."

Human small HERCs have important functions. For example, HERC3 binds Ub, PLIC1, or PLIC2 (Ub-like proteins) to endocytic proteins, thus regulating vesicular transport (Cruz et al. 2001). HERC4 is essential for spermatogenesis in mice (Cruz et al. 2001), and HERC5 is involved in the immune response related to interferon signaling pathways and polyubiquitinates IκB (inhibitor of the pro-inflammatory transcription factor NF-κB) (Kroismayr et al. 2004; Dastur et al. 2006).

The E6AP family (also known as E3A or UBE3A) includes the human protein E6AP (one of the first described HECTs), as well as proteins from animals, *Capsaspora owczarzaki, Sphaeroforma arctica,* and *Mortierella verticillata*, although the latter has poor nodal support. Human E6AP is known for its role in the inactivation of tumor suppressor p53 through proteasomal degradation (Scheffner 1998). E6AP is a good example of complex interplay between E3, in which different E3s have different antagonistic roles. For instance, human E6AP is polyubiquitinated by UBR5/EDD (another HECT E3,

discussed later) (Tomaic et al. 2011), as well as being enhanced (in an ubiquitin-independent manner) by HERC2 (Kühnle et al. 2011).

The HECTD2 family is an Opisthokonta-specific family that includes sequences from animals and Fungi, but not from unicellular Holozoa. HECTD2 proteins have a single HECT domain. Murine and human HECTD2 are known to intervene in protein degradation in neurodegeneration processes (Lloyd et al. 2009).

HECTX contains proteins from Cnidaria and Placozoa proteins, as well as from Filasterea, Fungi, and Amoebozoa. Thus, the lack of HECTX in bilaterians genomes is probably due to a secondary loss.

Class IV

Class IV includes four families: UBR5/EDD, G2E3, GL-Metazoa, and GL-Bikonta. The latter three are extremely divergent at the sequence level (figs. 2 and 3). The nodal support for this class is weak (fig. 2), but both Bayesian and ML analyses recovered the clade. In contrast, the nodal support for all of the families, except GL-Bikonta, is very good (BPP = 1.0 and BS = 99–100%).

The UBR5/EDD family includes proteins from animals (which have an EDD domain for binding ubiquitn, a zf-UBR protein recognition motif and a PABP domain) and architecturally simpler homologs from the choanoflagellate *Sal. rosetta* and the filasterean *Cap. owczarzaki*. Human EDD and *Dro. melanogaster* HYD act as general tumor suppressors by ubiquitinating E6AP (Tomaic et al. 2011), which increases p53 levels and induces cell senescence (Smits 2012). EDD and HYD also ubiquitinate TopBP1 (a topo-isomerase that intervenes in DNA damage response [Honda et al. 2002]) and negatively regulate Hh (hedgehog pathway) and Dpp (decapentaplegic pathway) expression, two crucial elements in the *Drosophila* eye disc development process (Lee 2002).

The G2E3, GL-Metazoa, and GL-Bikonta families are composed of proteins with a highly divergent HECT domain, with different domain arrangements that could confer them their own functional specificities. For instance, some proteins from *Naegleria gruberi* and *E. siliculosus* (GL-Bikonta) have unusual protein kinase domains of unknown function; and human and murine G2E3s have a non-functional HECT domain and three unconventional RING/PHD-like zinc fingers, two of which have been proved to have ubiquitin ligase activity (Brooks et al. 2008). None of these zinc fingers has been clearly classified as either PHD or RING motifs, although Pfam identifies the noncatalytically active one as a PHD-like zf-HC5HC2H domain (which is consistent with the fact that PHD domains are unable to act as ubiquitin ligases [Scheel and Hofmann 2003]). The lack of functional constraints on the HECT sequence would explain its divergence from other HECT proteins.

The most parsimonious explanation for the evolution of class IV is that an ancestral LECA gene underwent a duplication that gave rise to 1) the holozoan EDD family (secondarily lost in Bikonta species), and 2) a fast-evolving group, including the G2E3, GL-Metazoa, and GL-Bikonta families.

Class V

Class V (BPP = 1.0; BS = 96%) contains five families with proteins from Unikonta and Bikonta: UBE3B, UBE3C, HECTFu2, UPL6, and UPL7 (figs. 2 and 3). Except for HECTFu2, proteins belonging to this class have an exclusive IQ domain that could have been present in the ancestral protein that gave rise to class V. IQ typically binds to calmodulin and is also present in proteins that interact with GTP regulatory and cell cycle proteins, receptors, and channel proteins (Rhoads and Friedberg 1997).

UBE3B is an Opisthokonta-wide family in which an IQ domain is present in some proteins from animals, Filasterea (*Cap. owczarzaki*) and Fungi (*M. verticillata*). Proteins from the animal family UBE3C also have an IQ domain. UBE3B is thought to play a role in the oxidative stress response in humans and *Caenorhabditis elegans* (Oeda et al. 2001), and UBE3C plays an undetermined role in inflammatory responses in the human airways, probably related to IκB ubiquitination (Pasaje et al. 2011).

The HECTFu2 family, defined here for the first time, is specific to Fungi and their proteins do not bear any particular N-terminal protein domain architecture. It has no known substrates.

The UPL6 and UPL7 families conform to two independent clades, both consisting of Embryophyta and Chlorophyta proteins. UPL7 also contains proteins from Alveolata and Heterokonta. Again, IQ domains are found in Embryophyta and Chlorophyta sequences from UPL7 and Embryophyta sequences from UPL6. Contrary to previous studies (Gong et al. 2003), we did not recover a sister-group relationship between UPL6 and UPL7.

Class VI: Nedd4-Like, HUWE1, HACE1, and Other Families

Class VI is a wide group that includes 13 families plus three unclassified clades (figs. 2 and 3). The Bayesian analysis provides a good nodal support for this class (BPP = 0.99), but the clade is not statistically supported by ML.

The Nedd4-like group contains all families with C2 and WW domains: HECW/NEDL (with 1–2 WWs; specific to animals) Nedd4, WWP-Itchy and Smurf (with 2–4 WWs; specific to Holozoa). This group also contains two unclassified clades consisting of apusozoan and fungal proteins (with the same protein domain architecture) and a clade with proteins from unicellular Holozoa (with its own domain arrangement consisting of C2 and a CCCH zinc finger). The C2 domain targets the enzyme to membranes by binding to lipids (Ponting and Parker 1996), whereas WW is a recognition domain that selectively picks target proteins, typically through PY motifs (Chen and Sudol 1995; Macias et al. 2002).

A possible explanation for the evolution of this group of families involves the assumption that one ancestral homolog was present in the genome of the last Apusozoa–Opisthokonta common ancestor, which underwent independent diversifications in Apusozoa and Opisthokonta.

The Nedd4 family includes proteins from all holozoan lineages. In animals, Nedd4s are key downregulators of several receptors involved in cell signaling and membrane trafficking. For example, Nedd4s are responsible for the ubiquitination and stability of the insulin-like growth factor I receptor (Vecchione et al. 2003); *Dro. melanogaster* Nedd4 targets Notch receptor for proteasomal degradation (Sakata et al. 2004); and human Nedd4-1 ubiquitinates EGF (epidermal growth factor) receptor and ACK (a tyrosine kinase signaling factor) in response to EGF overexpression itself (Lin et al. 2010).

The WWP-Itchy family is also specific to Holozoa. It includes WWP1, WWP2, and Itchy, three human proteins that have been studied in depth, as well as Su(dx) from *Dro. melanogaster*. WWP-Itchy proteins regulate endosomal sorting and signaling by polyubiquitinating Notch in humans, mice, and *Cae. elegans* (Qiu et al. 2000; Wilkin et al. 2004; Shaye and Greenwald 2005). They also regulate the Hippo pathway: WWP1, WWP2, and Itchy polyubiquitinate AMOT (regulator of YAP/Yorkie, the central member of the Hippo pathway, which is essential for the constitution of a fully functional pathway [Sebé-Pedrós et al. 2012; Wang et al. 2012]). Itchy also polyubiquitinates Warts/Lats, another member of the Hippo pathway found in Opisthokonta (Ho et al. 2011). Moreover, human Itchy polyubiquitinates the transcription factors p63 and p73 (Rossi et al. 2005, 2006).

Within the Smurf family (present in all holozoan lineages except Ichthyosporea), DSmurf (*Dro. melanogaster* homolog) is known to regulate imaginal disc development (Liang et al. 2003) and embryonic dorsal-ventral patterning (Podos et al. 2001) by polyubiquitinating MAD (Dpp pathway); and human Smurfs (Smurf1 and 2) are known to antagonize TGFβ signaling, and therefore regulate cell growth and proliferation (Massagué and Gomis 2006).

The HECW family (or NEDL/Nedd4-like) contains animal HECTs, including human proteins NEDL1 (which stabilizes p53 in an ubiquitin-independent manner, thereby enhancing p53-mediated apoptosis [Li et al. 2008]) and NEDL2 (which stabilizes p73 [Miyazaki et al. 2003]).

The fungal, apusozoan, and unicellular-holozoan Nedd4-like clades are incertae sedis. As for the Nedd4-like fungal proteins (Fungi clade in fig. 2), only *Saccharomyces cerevisiae* Rsp5p has been characterized: It controls gene expression during nutrient limitation-driven stress (Cardona et al. 2009) and has various roles in intracellular trafficking (Belgareh-Touzé et al. 2008), and plasma membrane and cell wall organization (Kaminska et al. 2005). None of the Nedd4-like proteins from the apusozoan and unicellular-holozoan clade proteins has been characterized.

Class VI also includes several families characterized by a common domain architecture consisting of DUF908, DUF913, and DUF4414 (domains of unknown function). These three domains typically co-occur together in HECT proteins and are evolutionarily conserved in various Unikonta and Bikonta lineages, revealing an ancient origin for this group of proteins. These include HUWE1, HECTFu1 (HUWE1-like), UPL1/2, HECTAl1, and HECTHe3 families.

The HUWE1 family is named after the human protein within it (also known as UREB1, HectH9, KIAA0312, LASU1, ARF-BP1, or Mule). HUWE1 proteins have a complex domain architecture consisting of DUF908, DUF913, WWE, UBA, and DUF4414. It includes representatives from animals, *M. brevicollis* and Amoebozoa. The *M. brevicollis* has a single HECT domain, but proteins from Amoebozoa have the complete arrangement (except WWE). Human HUWE1 polyubiquitinates Myc (oncoprotein and transcription factor), which is essential for the transactivation of several Myc target genes, the recruitment of co-activator p300 and the induction of cell proliferation (Adhikary et al. 2005). It also enhances p53 stability by helping ARF inhibit p53 ubiquitination by Mdm2 (Brooks and Gu 2006), among other functions (Chen et al. 2005; Zhong et al. 2005; Hall et al. 2007).

The HECTFu1 family includes fungal proteins with a HUWE1-like N-terminal architecture (without WWE), and also some specific domains and simpler arrangements. There is indirect evidence that Tom1 (a yeast HUWE1-like protein) intervenes in Cdc6 posttranslational regulation (Hall et al. 2007).

UPL1/2 is a Viridiplantae-specific family that contains Embryophyta proteins with the characteristic DUF908-DUF913-UBA-DUF4414 N-terminal architecture and green algae proteins with a single HECT domain.

Both the HECTHe3 (present in Heterokonta) and HECTAl1 (present in Alveolata) families also contain the DUF4414-HECT arrangement.

Finally, there are four additional families with good nodal support and domain coherence within class VI: KIAA0317, HACE1, HECTHe4, and UPL5.

The HACE1 family contains proteins from all holozoan clades plus *A. castellanii*. HACE1 proteins have a variable number of Ankyrin repeats (typically two to three) and sometimes a PHD domain. The ubiquitinating activity of HACE1 is known to regulate Golgi complex disassembly and reassembly during mitosis (Tang et al. 2011), and also plays a role in various cancer processes (Zhang et al. 2007). The HACE1 and HUWE1 families were thought to be sister groups and, together, to be a sister group to the Nedd4-like group of proteins (Marín 2010); however, we did not recover such topology, but rather a polytomy of several families (fig. 2).

The KIAA0317 family is exclusive to Metazoa and Choanoflagellata (*Sal. rosetta*) clades. Most of them have Filamin repeats, which are only found in this family. They have no

known substrates, but Filamin is known to mediate protein recognition in other proteins and contexts (Ohta et al. 2006).

The HECTHe4 is specific to Heterokonta and includes *P. infestans* proteins with a distinctive zf-RanBP domain and other proteins with a HECT domain. Both ML and BI analyses have linked this family to the Nedd4-like group of proteins, but with low statistical support (fig. 2).

UPL5 is a Bikonta family that includes proteins from Viridiplantae (with a Ub domain), as well as from Rhizaria (*Bigelowiella natans*) and Cryptophyta (*Guillardia theta*) clades (with just a HECT domain). *Arabidopsis thaliana* UPL5 polyubiquitinates the WRKY53 transcription factor, which promotes leaf senescence (Miao and Zentgraf 2010). Ub-like domains within E3 enzymes probably allow for the interaction of these enzymes with other members of the pathway (Miao and Zentgraf 2010).

The Origins of Multicellularity and the Evolution of the HECT E3 System

As unicellular eukaryotes evolved into multicellular life forms, the need for more complex and finely tuned regulation mechanisms increased and met new regulatory requirements related to cell proliferation, adhesion, differentiation, ordered cell death, and extra/intracellular signaling. Therefore, and given that the ubiquitination pathway is an important regulatory layer responsible for key posttranslational modifications and protein turnover, one may expect expansions of the ubiquitination toolkit (including the HECT system) at the origin of multicellular clades. To ascertain whether this is the case, we analyzed the functional and molecular diversity of the HECT system in several eukaryote lineages.

Specifically, we used the relationship between the number of HECT proteins and the number of distinct N-terminal domain architectures of those proteins as an estimator of the diversity of the HECT system in every given genome. Our data show that the number of HECT proteins positively correlates with the number of distinct N-terminal domain architectures (fig. 6).

According to this, the HECT system is enriched in animals and unicellular Holozoa, the Heterokonta *P. infestans* and *E. siliculosus*, and the Apusozoa *T. trahens*. Conversely, Fungi, plant, Chlorophyta, Rhodophyta, and Excavata genomes are HECT-poor, with fewer proteins and little protein domain diversification. It is worth mentioning that some species such as the Rhizaria *B. natans* and the Haptophyta *Emi. huxleyi* have a high count of HECT proteins but a low degree of domain diversification.

The Apusozoa *T. trahens*, the sister group to Opisthokonta (Torruella et al. 2012), also shows a relatively rich HECT toolkit, much richer than plants and Fungi and similar in complexity to those of metazoans. Our data show that there are some HECT proteins that independently diversified within *T. trahens*. For instance, class I contains an unclassified *T. trahens* clade whose proteins have independently acquired different protein recognition domains (such as SPRY, ZZ, and zf-UBR). Also, the well-known Nedd4 group of HECTs dates back to the last common ancestor between Opisthokonta and Apusozoa. New apusozoan genomes will make it possible to gain further insights into the evolution of the HECT system in this lineage.

The diversity of HECTs in Heterokonta is highly variable. *Thalassiosira pseudonana* has a poor HECT system, whereas *E. siliculosus* (a multicellular brown alga) and especially *P. infestans* have a more diversified HECT system comparable with that of animals that most likely evolved from a small basal toolkit similar to that of *Tha. pseudonana*, according to the present phylogeny. Moreover, both *P. infestans* and *E. siliculosus* proteins have convergently acquired several architectures characteristic of Opisthokonta HECTs. For example, *P. infestans* proteins have recognition domains such as MIB-HERC2, UBA, SPRY, or RLD (typical of large HERC families), and *E. siliculosus* proteins have RLD and Kelch repeats.

Our analyses show that animals have the most expanded and diverse HECT system among eukaryotes, and their unicellular holozoan relatives (Choanoflagellata, Filasterea, and Ichthyosporea) have an intermediate diversity of the system (fig. 6). This suggests that there was a burst of HECT diversity at the onset of Metazoa, but that a relatively complex HECT system already existed in the animals' closest unicellular relatives. Indeed, the origin of most (17 out of 22) HECT families containing animal proteins (among those defined in this study) pre-dates the origin of animals (fig. 4). Rather, the higher degree of diversification of HECT in animals is explained by the acquisition of new domains in the N-terminal regions of HECTs. Leaving aside the hemichordate *S. kowalevskii* (a clear outlier to the general trend), animals have between 9 and 14 different HECT architectures, whereas their closest unicellular holozoan relatives have between four and nine arrangements.

The number of families present in each clade provides additional information on the degree of diversification of the HECT system in each taxon (fig. 4). For instance, 24 new families appear at some point during the evolution of the Opisthokonta lineage. The Holozoa are the most family-rich lineage, with 22 families, 5 of which are specific to Metazoa. Also, there are five families present in plants (all of which appear either at the origin of Bikonta or Viridiplantae). This reveals that in both animals and plants most HECT families pre-date the respective origins of multicellularity.

We also mapped the acquisition of N-terminal domains across the tree of eukaryotes (fig. 5). This is a common event within each class, and those architectures that appear at the base of multicellular clades and their closest unicellular relatives are of particular interest. Our data show that the acquisition of new domains is a common event in the holozoan clade, especially at the root of animals and Choanoflagellata (six domains) and at the node leading to Metazoa (eight domains). Indeed, there are five families (namely, EDD, HECTD3, HUWE1, UBE3B, and HERC2) in which animal

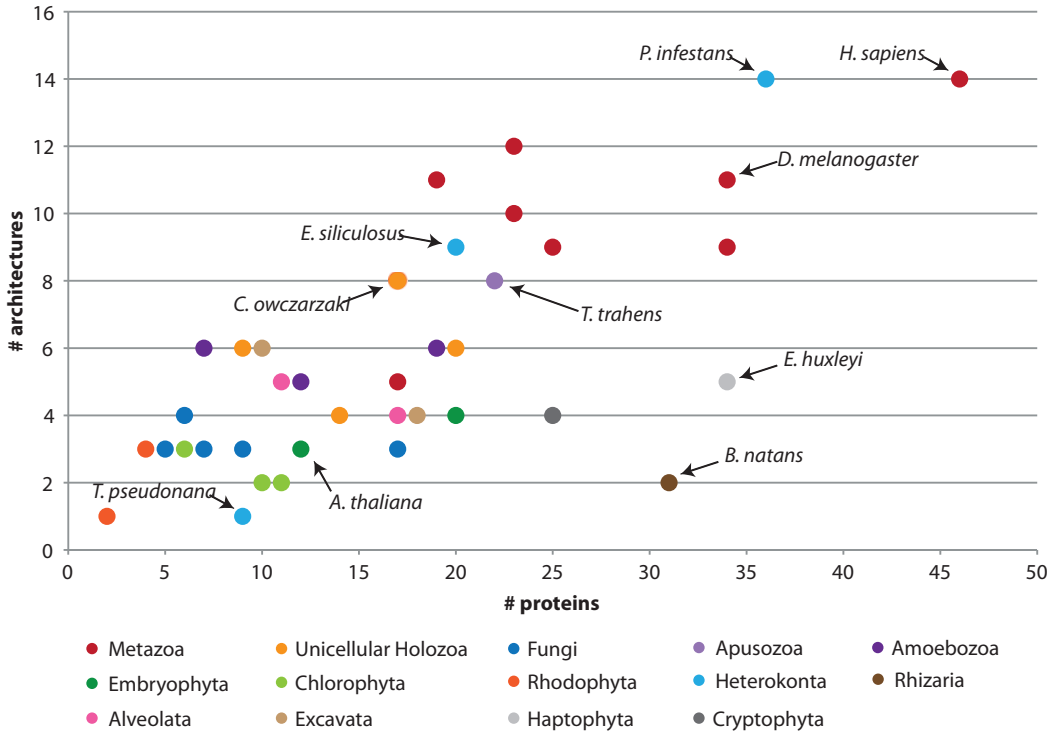


Fig. 6.—Number of HECT proteins against the number of distinct protein domain architectures found in all HECT-containing proteins for each species. Taxa are color-coded according to taxonomic assignment.

proteins have more complex architectures than those found in their unicellular relatives' homologs. Conversely, the acquisition of specific protein domains in other multicellular lineages such as Fungi and Embryophyta is minimal.

Overall, our data suggest that increases in both N-terminal architectural diversification and absolute number of proteins have shaped the evolutionary history of HECT ligases in eukaryotes. An increase in the protein number brings molecular duplicities that allow sub- or neofunctionalization of HECT proteins. N-terminal domain shuffling is a plastic and adaptable evolutionary mechanism that does not require a change of gene content. It can account for significant evolutionary changes in posttranslational regulation through the adjustment of substrate specificity and protein localization. Indeed, domain shuffling has been acknowledged as an important mechanism for explaining the evolution of multidomain proteins and the appearance of novel proteins, especially regarding the origin of new proteins in major transitions such as the acquisition of multicellularity in animals (Tordai et al. 2005; King et al. 2008; Suga et al. 2012).

It must be noted that HECTs are not the only set of E3 ligases of the ubiquitin system and they are not equally relevant in different eukaryotic lineages. This means that HECT-poor taxa such as plants or Fungi may not necessarily have a poor ubiquitination system. Indeed, *Ara. thaliana*, with just seven HECTs, has expanded their E3 proteins count in terms of F-box, RING and U-box ligases (Lespinet et al. 2002), compared to other eukaryotes. Conversely, E1 and E2 functions are each performed by a single type of enzymes. All E1 enzymes descend from a common ancestor that was co-opted into ubiquitin activating functions at the origin of eukaryotes, and, since then, has undergone duplications in Unikonta, Vertebrata, Heterokonta, and Kinetoplastida (Excavata) (Burroughs et al. 2009). Similarly, there is just one type of E2 enzyme for conjugating ubiquitin, and all (or most of) their known families were already present at the LECA (Burroughs et al. 2008; Michelle et al. 2009). Altogether, this shows that E1 and E2 enzymes radiated concomitantly prior to the LECA, when they were recruited for the ubiquitination pathway (Burroughs et al. 2008).

This pattern of evolution is markedly different from that showed by HECTs (in this study) and other E3 enzymes (Lespinet et al. 2002), which have undergone differential lineage-specific expansions—in the case of HECTs, those detected in Holozoa, Heterokonta, and maybe Apusozoa. This emphasizes the role of E3s as a specific and functionally specialized step of the ubiquitination pathway.

Conclusions

Our genomic survey and phylogenetic analysis classifies eukaryotic HECTs in six main classes, whose constituent proteins

probably descend from six ancestral proteins present in the LECA, assuming the "Unikont–Bikont" hypothesis for the rooting of the eukaryote phylogeny. These six classes include 35 identified protein families, as well as other proteins that cannot be classified with certainty.

We also show that, because the eukaryotic ancestor, the HECT system has increased its functional complexity and capacity to finely tune posttranslational protein regulation in several clades, especially—but not exclusively—in multicellular organisms. The system has also been simplified in other clades such as unicellular red algae.

The current diversity of the HECT system has been acquired through two parallel mechanisms: 1) the acquisition of new HECT families through protein duplication, and 2) the acquisition, by domain shuffling, of new protein domains that specifically recognize E3 substrates. We identified a positive correlation between the degree of domain diversification and the number of HECT proteins present in each genome.

Our analysis reveals that this domain syntax of HECT proteins is highly conserved across all eukaryotes: domain fusions always occur at the N-terminus of the proteins. This would be largely due to the physical constraints to catalytic activity imposed by the HECT proteins tertiary structure.

The HECT toolkit evolved in a largely independent manner in different eukaryote clades, often converging in similar domain architectures. Some taxa such as Holozoa are HECT-rich, with many HECT types and various domain arrangements, whereas other taxa such as fungi, plants, and green and red algae have HECT-poor genomes. Regarding the evolution of Holozoa, this study reveals that the onset of new families and new protein recognition motifs typically predate the emergence of animal multicellularity. However, animals further increased their HECT regulatory toolkit from their unicellular ancestor with six new HECT families.

Overall, we show a complex evolutionary scenario in which the HECT system has evolved toward different degrees of diversification in different clades, through family diversification and domain shuffling. Our genomic survey of HECT proteins clarifies the origin and evolution of different HECT families among eukaryotes and also represents a useful evolutionary framework for analyzing this important posttranslational regulatory mechanism.

Supplementary Material

Supplementary figures S1–S3 and tables S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank the Joint Genome Institute and Broad Institute for making data publicly available. This work was supported by an Institució Catalana de Recerca i Estudis Avançats contract, European Research Council starting grant ERC-2007-StG-206883, Ministerio de Economía y Competitividad (MINECO) grant BFU2011-23434 to I.R.-T., and a pregraduate Formación Profesorado Universitario grant from MINECO to A.S.-P.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

Adhikary S, et al. 2005. The ubiquitin ligase HectH9 regulates transcriptional activation by Myc and is essential for tumor cell proliferation. Cell 123:409–421.

Aravind L. 2001. The WWE domain: a common interaction module in protein ubiquitination and ADP ribosylation. Trends Biochem Sci. 26: 273–275.

Belgareh-Touzé N, et al. 2008. Versatile role of the yeast ubiquitin ligase Rsp5p in intracellular trafficking. Biochem Soc Trans. 36:791–796.

Brooks CL, Gu W. 2006. p53 ubiquitination: Mdm2 and beyond. Mol Cell. 21:307–315.

Brooks WS, et al. 2008. G2E3 is a dual function ubiquitin ligase required for early embryonic development. J Biol Chem. 283:22304–22315.

Burroughs AM, Balaji S, Iyer LM, Aravind L. 2007. Small but versatile: the extraordinary functional and structural diversity of the beta-grasp fold. Biol Direct. 2:18.

Burroughs AM, Jaffee M, Iyer LM, Aravind L. 2008. Anatomy of the E2 ligase fold: implications for enzymology and evolution of ubiquitin/Ub-like protein conjugation. J Struct Biol. 162:205–18.

Burroughs AM, Iyer LM, Aravind L. 2009. Natural history of the E1-like superfamily: implication for adenylation, sulfur transfer, and ubiquitin conjugation. Proteins 75:895–910.

Cardona F, Aranda A, Del Olmo M. 2009. Ubiquitin ligase Rsp5p is involved in the gene expression changes during nutrient limitation in *Saccharomyces cerevisiae*. Yeast 26:1–15.

Chen D, et al. 2005. ARF-BP1/Mule is a critical mediator of the ARF tumor suppressor. Cell 121:1071–1083.

Chen HI, Sudol M. 1995. The WW domain of Yes-associated protein binds a proline-rich ligand that differs from the consensus established for Src homology 3-binding modules. Proc Natl Acad Sci U S A. 92: 7819–7823.

Chong-Kopera H, et al. 2006. TSC1 stabilizes TSC2 by inhibiting the interaction between TSC2 and the HERC1 ubiquitin ligase. J Biol Chem. 281:8313–8316.

Coates J. 2008. Armadillo repeat proteins: versatile regulators of plant development and signalling. Plant Cell Monogr. 10:299–314.

Cruz C, Ventura F, Bartrons R, Rosa JL. 2001. HERC3 binding to and regulation by ubiquitin. FEBS Lett. 488:74–80.

Dastur A, Beaudenon S, Kelley M, Krug RM, Huibregtse JM. 2006. Herc5, an interferon-induced HECT E3 enzyme, is required for conjugation of ISG15 in human cells. J Biol Chem. 281:4334–4338.

Derelle R, Lang BF. 2012. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. Mol Biol Evol. 29:1277–1289.

Downes BP, Stupar RM, Gingerich DJ, Vierstra RD. 2003. The HECT ubiquitin-protein ligase (UPL) family in *Arabidopsis*: UPL3 has a specific role in trichome development. Plant J. 35:729–742.

Gong T-WL, Huang Li, Warner SJ, Lomax MI. 2003. Characterization of the human UBE3B gene: structure, expression, evolution, and alternative splicing. Genomics 82:143–152.

Hadjebi O, Casas-Terradellas E, Garcia-Gonzalo FR, Rosa JL. 2008. The RCC1 superfamily: from genes, to function, to disease. Biochim Biophys Acta. 1783:1467–1479.

Hall J, Kow E, Nevis K, Lu C. 2007. Cdc6 stability is regulated by the Huwe1 ubiquitin ligase after DNA damage. Mol Biol Cell. 18:3340–3350.

Hicke L. 2001. A new ticket for entry into budding vesicles-ubiquitin. Cell 106:527–530.
Ho KC, et al. 2011. Itch E3 ubiquitin ligase regulates large tumor suppressor 1 stability. Proc Natl Acad Sci U S A. 108:4870–4875.
Honda Y, et al. 2002. Cooperation of HECT-domain ubiquitin ligase hHYD and DNA topoisomerase II-binding protein for DNA damage response. J Biol Chem. 277:3599–3605.
Huang L, et al. 1999. Structure of an E6AP-UbcH7 complex: insights into ubiquitination by the E2-E3 enzyme cascade. Science 286:1321–1326.
Itoh M, et al. 2003. Mind bomb is a ubiquitin ligase that is essential for efficient activation of Notch signaling by delta. Dev Cell. 4:67–82.
Jin L, Williamson A, Banerjee S, Philipp I, Rape M. 2008. Mechanism of ubiquitin-chain formation by the human anaphase-promoting complex. Cell 133:653–665.
Kajiro M, et al. 2011. The E3 ubiquitin ligase activity of Trip12 is essential for mouse embryogenesis. PLoS One 6:e25871.
Kaminska J, et al. 2005. Rsp5 ubiquitin ligase affects isoprenoid pathway and cell wall organization in *S. cerevisiae*. Acta Biochim Pol. 52: 207–220.
Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33: 511–518.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Kaustov L, et al. 2007. The conserved CPH domains of Cul7 and PARC are protein-protein interaction modules that bind the tetramerization domain of p53. J Biol Chem. 282:11300–11307.
King N, et al. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature 451:783–788.
Kroismayr R, et al. 2004. HERC5, a HECT E3 ubiquitin ligase tightly regulated in LPS activated endothelial cells. J Cell Sci. 117:4749–4756.
Kühnle S, et al. 2011. Physical and functional interaction of the HECT ubiquitin-protein ligases E6AP and HERC2. J Biol Chem. 286: 19410–19416.
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.
Lee JD. 2002. The ubiquitin ligase hyperplastic discs negatively regulates hedgehog and decapentaplegic expression by independent mechanisms. Development 129:5697–5706.
Lespinet O, Wolf YI, Koonin EV, Aravind L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res. 12:1048–1059.
Liang Y-Y, et al. 2003. dSmurf selectively degrades decapentaplegic-activated MAD, and its overexpression disrupts imaginal disc development. J Biol Chem. 278:26307–26310.
Lin Q, et al. 2010. HECT E3 ubiquitin ligase Nedd4-1 ubiquitinates ACK and regulates epidermal growth factor (EGF)-induced degradation of EGF receptor and ACK. Mol Cell Biol. 30:1541–1554.
Li Y, et al. 2008. A novel HECT-type E3 ubiquitin protein ligase NEDL1 enhances the p53-mediated apoptotic cell death in its catalytic activity-independent manner. Oncogene 27:3700–3709.
Lloyd SE, et al. 2009. HECTD2 is associated with susceptibility to mouse and human prion disease. PLoS Genet. 5:e1000383.
Macias MJ, Wiesner S, Sudol M. 2002. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. FEBS Lett. 513: 30–37.
Marín I. 2010. Animal HECT ubiquitin ligases: evolution and functional implications. BMC Evol Biol. 10:56–68.
Maspero E, et al. 2011. Structure of the HECT:ubiquitin complex and its role in ubiquitin chain elongation. EMBO Rep. 12:342–349.
Massagué J, Gomis RR. 2006. The logic of TGFbeta signaling. FEBS Lett. 580:2811–2820.
Miao Y, Zentgraf U. 2010. A HECT E3 ubiquitin ligase negatively regulates *Arabidopsis* leaf senescence through degradation of the transcription factor WRKY53. Plant J. 63:179–188.
Michelle C, Vourc'h P, Mignon L, Andres CR. 2009. What was the set of ubiquitin and ubiquitin-like conjugating enzymes in the eukaryote common ancestor? J Mol Evol. 68:616–628.
Miyazaki K, et al. 2003. A novel HECT-type E3 ubiquitin ligase, NEDL2, stabilizes p73 and enhances its transcriptional activity. Biochem Biophys Res Commun. 308:106–113.
Mukhopadhyay D, Riezman H. 2007. Proteasome-independent functions of ubiquitin in endocytosis and signaling. Science 315: 201–205.
Nishiya T, et al. 2011. Regulation of inducible nitric-oxide synthase by the SPRY domain- and SOCS box-containing proteins. J Biol Chem. 286: 9009–9019.
Oeda T, et al. 2001. Oxidative stress causes abnormal accumulation of familial amyotrophic lateral sclerosis-related mutant SOD1 in transgenic *Caenorhabditis elegans*. Hum Mol Genet. 10:2013–2023.
Ohta Y, Hartwig JH, Stossel TP. 2006. FilGAP, a Rho- and ROCK-regulated GAP for Rac binds filamin A to control actin remodelling. Nat Cell Biol. 8:803–814.
Ozols J. 1989. Structure of cytochrome b5 and its topology in the microsomal membrane. Biochim Biophys Acta. 997:121–130.
Pasaje CF, et al. 2011. UBE3C genetic variations as potent markers of nasal polyps in Korean asthma patients. J Hum Genet. 56:797–800.
Peters J-M. 2002. The anaphase-promoting complex: proteolysis in mitosis and beyond. Mol Cell. 9:931–943.
Pickart CM, Fushman D. 2004. Polyubiquitin chains: polymeric protein signals. Curr Opin Chem Biol. 8:610–616.
Podos SD, Hanson KK, Wang YC, Ferguson EL. 2001. The DSmurf ubiquitin-protein ligase restricts BMP signaling spatially and temporally during *Drosophila* embryogenesis. Dev Cell. 1:567–578.
Ponting CP, Parker PJ. 1996. Extending the C2 domain family: C2s in PKCs delta, epsilon, eta, theta, phospholipases, GAPs, and perforin. Protein Sci. 5:162–166.
Punta M, et al. 2012. The Pfam protein families database. Nucleic Acids Res. 40:D290–D301.
Qiu L, et al. 2000. Recognition and ubiquitination of Notch by Itch, a Hect-type E3 ubiquitin ligase. J Biol Chem. 275:35734–35737.
Rhoads A, Friedberg F. 1997. Sequence motifs for calmodulin recognition. FASEB J. 11:331–340.
Rodríguez-Ezpeleta N, et al. 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Curr Biol. 17: 1420–1425.
Rosa J, Casaroli-Marano R. 1996. p619, a giant protein related to the chromosome condensation regulator RCC1, stimulates guanine nucleotide exchange on ARF1 and Rab proteins. EMBO J. 15: 4262–4273.
Rossi M, et al. 2005. The ubiquitin-protein ligase Itch regulates p73 stability. EMBO J. 24:836–848.
Rossi M, et al. 2006. The E3 ubiquitin ligase Itch controls the protein stability of p63. Proc Natl Acad Sci U S A. 103:12753–12758.
Rotin D, Kumar S. 2009. Physiological functions of the HECT family of ubiquitin ligases. Nat Rev Mol Cell Biol. 10:398–409.
Sakata T, Sakaguchi H, Tsuda L, Higashitani A. 2004. *Drosophila* Nedd4 regulates endocytosis of notch and suppresses its ligand-independent activation. Curr Biol. 14:2228–2236.
Sarkar A, Zohn I. 2011. Hectd1 regulates intracellular trafficking of Hsp90 to control its secretion and cell motility of the cranial mesenchyme. Dev Biol. 356:120.
Scheel H, Hofmann K. 2003. No evidence for PHD fingers as ubiquitin ligases. Trends Cell Biol. 13:285–7; author reply 287–288.
Scheffner M. 1998. Ubiquitin, E6-AP, and their role in p53 inactivation. Pharmacol Ther. 78:129–139.

Schwartz AL, Ciechanover A. 2009. Targeting proteins for destruction by the ubiquitin system: implications for human pathobiology. Annu Rev Pharmacol Toxicol. 49:73–96.
Sebé-Pedrós A, Zheng Y, Ruiz-Trillo I, Pan D. 2012. Premetazoan origin of the hippo signaling pathway. Cell Rep. 1:13–20.
Shaye DD, Greenwald I. 2005. LIN-12/Notch trafficking and regulation of DSL ligand activity during vulval induction in *Caenorhabditis elegans*. Development 132:5081–5092.
Smits V. 2012. EDD induces cell cycle arrest by increasing p53 levels. Cell Cycle 11:715–720.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science 297:89–91.
Suga H, et al. 2012. Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. Sci Signal. 5:ra35.
Tang D, et al. 2011. The ubiquitin ligase HACE1 regulates Golgi membrane dynamics during the cell cycle. Nat Commun. 2:501.
Tewari R, Bailes E, Bunting KA, Coates JC. 2010. Armadillo-repeat protein functions: questions for little creatures. Trends Cell Biol. 20:470–481.
Tomaic V, et al. 2011. Regulation of the human papillomavirus type 18 E6/E6AP ubiquitin ligase complex by the HECT domain-containing protein EDD. J Virol. 85:3120–3127.
Tordai H, Nagy A, Farkas K, Bányai L, Patthy L. 2005. Modules, multidomain proteins and organismic complexity. FEBS J. 272:5064–5078.
Torruella G, et al. 2012. Phylogenetic relationships within the opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. Mol Biol Evol. 29:531–544.
Vecchione A, Marchese A, Henry P. 2003. The Grb10/Nedd4 complex regulates ligand-induced ubiquitination and stability of the insulin-like growth factor I receptor. Mol Cell Biol. 23:3363–3372.
Verdecia MA, et al. 2003. Conformational flexibility underlies ubiquitin ligation mediated by the WWP1 HECT domain E3 ligase. Mol Cell. 11:249–259.
Wang C, et al. 2012. The Nedd4-like ubiquitin E3 ligases target angiomotin/p130 to ubiquitin-dependent degradation. Biochem J. 444:279–289.
Wilkin MMB, et al. 2004. Regulation of notch endosomal sorting and signaling by *Drosophila* Nedd4 family proteins. Curr Biol. 14:2237–2244.
Wu W, et al. 2010. HERC2 is an E3 ligase that targets BRCA1 for degradation. Cancer Res. 70:6384–6392.
Ying M, Zhan Z, Wang W, Chen D. 2009. Origin and evolution of ubiquitin-conjugating enzymes from *Guillardia theta* nucleomorph to hominoid. Gene 447:72–85.
Yu J, et al. 2008. The E3 ubiquitin ligase HECTD3 regulates ubiquitination and degradation of Tara. Biochem Biophys Res Commun. 367:805–812.
Zhang L, et al. 2007. The E3 ligase HACE1 is a critical chromosome 6q21 tumor suppressor involved in multiple cancers. Nat Med. 13:1060–1069.
Zhang L, Kang L, Bond W, Zhang N. 2009. Interaction between syntaxin 8 and HECTd3, a HECT domain ligase. Cell Mol Neurobiol. 29:115–121.
Zhong Q, Gao W, Du F, Wang X. 2005. Mule/ARF-BP1, a BH3-only E3 ubiquitin ligase, catalyzes the polyubiquitination of Mcl-1 and regulates apoptosis. Cell 121:1085–1095.

Associate editor: Purificación López-García

3.2. Evolution and classification of myosins, a paneukaryotic whole genome approach

Abstract – Myosins are key components of the eukaryotic cytoskeleton, providing motility for a broad diversity of cargoes. Therefore, understanding the origin and evolutionary history of myosin classes is crucial to address the evolution of eukaryote cell biology. Here, we revise the classification of myosins using an updated taxon sampling that includes newly or recently sequenced genomes and transcriptomes from key taxa. We performed a survey of eukaryotic genomes and phylogenetic analyses of the myosin gene family, reconstructing the myosin toolkit at different key nodes in the eukaryotic tree of life. We also identified the phylogenetic distribution of myosin diversity in terms of number of genes, associated protein domains and number of classes in each taxa. Our analyses show that new classes (*i.e.* paralogs) and domain architectures were continuously generated throughout eukaryote evolution, with a significant expansion of myosin abundance and domain architectural diversity at the stem of Holozoa, predating the origin of animal multicellularity. Indeed, single-celled holozoans have the most complex myosin complement among eukaryotes, with paralogs of most myosins previously considered animal-specific. We recover a dynamic evolutionary history, with several lineage-specific expansions (*e.g.* the 'myosin III-like' gene family diversification in choanoflagellates), convergence in protein domain architectures (*e.g.* fungal and animal chitin synthase myosins), and important secondary losses. Overall, our evolutionary scheme demonstrates that the ancestral eukaryote likely had a complex myosin repertoire that included six genes with different protein domain architectures. Finally, we provide an integrative and robust classification, useful for future genomic and functional studies on this crucial eukaryotic gene family.

Evolution and Classification of Myosins, a Paneukaryotic Whole-Genome Approach

Arnau Sebé-Pedrós[1,2,†], Xavier Grau-Bové[1,†], Thomas A. Richards[3,4], and Iñaki Ruiz-Trillo[1,2,5,*]

[1]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta, Barcelona, Catalonia, Spain

[2]Departament de Genètica, Universitat de Barcelona, Catalonia, Spain

[3]Life Sciences, The Natural History Museum, London, United Kingdom

[4]Biosciences, University of Exeter, United Kingdom

[5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys, Barcelona, Catalonia, Spain

*Corresponding author: E-mail: inaki.ruiz@multicellgenome.org, inaki.ruiz@ibe.upf-csic.es.

[†]These authors contributed equally to this work.

Accepted: January 14, 2014

Abstract

Myosins are key components of the eukaryotic cytoskeleton, providing motility for a broad diversity of cargoes. Therefore, understanding the origin and evolutionary history of myosin classes is crucial to address the evolution of eukaryote cell biology. Here, we revise the classification of myosins using an updated taxon sampling that includes newly or recently sequenced genomes and transcriptomes from key taxa. We performed a survey of eukaryotic genomes and phylogenetic analyses of the myosin gene family, reconstructing the myosin toolkit at different key nodes in the eukaryotic tree of life. We also identified the phylogenetic distribution of myosin diversity in terms of number of genes, associated protein domains and number of classes in each taxa. Our analyses show that new classes (i.e., paralogs) and domain architectures were continuously generated throughout eukaryote evolution, with a significant expansion of myosin abundance and domain architectural diversity at the stem of Holozoa, predating the origin of animal multicellularity. Indeed, single-celled holozoans have the most complex myosin complement among eukaryotes, with paralogs of most myosins previously considered animal specific. We recover a dynamic evolutionary history, with several lineage-specific expansions (e.g., the myosin III-like gene family diversification in choanoflagellates), convergence in protein domain architectures (e.g., fungal and animal chitin synthase myosins), and important secondary losses. Overall, our evolutionary scheme demonstrates that the ancestral eukaryote likely had a complex myosin repertoire that included six genes with different protein domain architectures. Finally, we provide an integrative and robust classification, useful for future genomic and functional studies on this crucial eukaryotic gene family.

Key words: origin of eukaryotes, LECA, Holozoa, eukaryote evolution, chitin synthase, Smad.

Introduction

The evolution of molecular motors was key to the origin and diversification of the eukaryotic cell. There are three major superfamilies of motor proteins: kinesins, dyneins, and myosins. The first two act as motors on microtubule filaments, while myosins function on actin (Vale 2003). Myosins participate in a variety of cellular processes, including cytokinesis, organellar transport, cell polarization, transcriptional regulation, intracellular transport, and signal transduction (Hofmann et al. 2009; Bloemink and Geeves 2011; Hartman et al. 2011). They bind to filamentous actin and produce physical forces by hydrolyzing ATP and converting chemical energy into mechanical force (Hartman and Spudich 2012). Both activities reside in the myosin head domain (PF00063). This head domain is accompanied by a broad diversity of N-terminal and/or C-terminal domains that bind to different molecular cargos, providing the functional specificity of the protein. Some myosins, such as myosins V and II, act as dimers that contact through their C-terminal coiled-coils, while others, such as myosins I, III, VI, VII, IX, X, XV, and XIX, act as monomers (Peckham 2011).

The identification of gene orthologs can be best accomplished by phylogenetic analyses, especially when complex architectures that are likely to undergo rearrangements are

involved (Koonin 2005; Sjölander et al. 2011; Leonard and Richards 2012; Gabaldón and Koonin 2013). Thus, myosin phylogenetic analysis is important to classify myosin paralog families and identify the ancestry of different gene architectures. Previous efforts have been made to classify the myosin family and to reconstruct its evolutionary diversification (Richards and Cavalier-Smith 2005; Foth et al. 2006; Odronitz and Kollmar 2007), although information from some key eukaryotic groups that have recently become available were missing from all of these studies. Therefore, there is a need to revise schemes of myosin evolution using improved taxon sampling and phylogenetic methods. This is important both to update the classification of myosins diversity and also understand the origin and evolutionary history of the wider gene family. Moreover, a precise reconstruction of the ancestral eukaryotic myosin toolkit (along with that of the other motor proteins [Wickstead and Gull 2007; Wickstead et al. 2010]) has important implications for understanding the phylogenetic patterns and functional attributes of early eukaryotes (Richards and Cavalier-Smith 2005).

Previous analyses, using different genome datasets and different phylogenetic methods provided conflicting hypotheses on myosin classification and the reconstruction of this ancestral toolkit. For example, Richards and Cavalier-Smith (2005) provided a classification of myosins based on two criteria: phylogenetic reconstruction and analysis of protein domain architecture. They inferred that the last eukaryotic common ancestor (LECA) had 3 of the 37 defined eukaryotic myosin types, including Myo_head-MYTH4/FERM, Myo_head-SMC-DIL, and Myo_head-TH1. In contrast, Foth et al. (2006), in a study focused on apicomplexan myosins, defined 29 classes and did not infer an ancestral complement. Also based on phylogeny, Odronitz and Kollmar (2007) defined 35 different myosin classes, most with an extremely restricted phylogenetic distribution. To make things more complex, different authors have used different criteria for classification, leading to inconsistencies in the classification and nomenclature between studies.

In this article, we present a new evolutionary history and classification of eukaryotic myosins. We use a significantly expanded taxon sampling than previous studies, in which, for the first time, all major eukaryotic lineages are represented. In particular, we include data from four previously unsampled eukaryotic lineages (Apusozoa, Rhizaria, Haptophyta, and Cryptophyta) so that all the major eukaryotic supergroups are represented (Roger and Simpson 2009). Evolutionary analyses have consistently demonstrated that the evolution of parasitic phenotypes is often accompanied by large-scale gene losses (Peyretaillade et al. 2011; Pomberta et al. 2012; Wolf et al. 2013). To overcome this problem, we here include free-living representatives of lineages that were previously represented only by parasitic taxa (such as *Ectocarpus siliculosus* and unicellular brown algae in Heterokonta/Stramenopiles and *Naegleria gruberi* in Excavata). Furthermore, we include data from taxa occupying phylogenetic positions that are key to understand major evolutionary transitions, including deep-branching fungi (the Chytridiomycota *Spizellomyces punctatus*), green algae, deeply derived plants, unicellular holozoan lineages (choanoflagellates, filastereans, and ichthyosporeans) and early-branching metazoans (ctenophores and sponges). We also use improved alignment and phylogenetic inference methods. We do not aim to infer a eukaryotic tree of life from the myosin genomic content (Richards and Cavalier-Smith 2005; Odronitz and Kollmar 2007). Convergence (Zmasek and Godzik 2012) (discussed later), gene fission (Leonard and Richards 2012), duplication, gene loss (Zmasek and Godzik 2011), and horizontal gene transfer (HGT) (Andersson et al. 2003; Andersson 2005; Marcet-Houben and Gabaldón 2010; Richards et al. 2011) are important phenomena in eukaryotes and, therefore, molecular markers such as the distribution pattern of gene orthologs need to be tested using gene phylogeny and updated as new genome sequences are released (Dutilh et al. 2007; House 2009; Shadwick and Ruiz-Trillo 2012). We based our myosin classification exclusively on phylogenetic affinity, which allowed us to identify: gene and domain loss, paralog groups, and convergent evolution of gene domain architecture. The use of updated phylogenetic methods and improved taxon representation allowed us to analyse the classification, evolutionary history, and functional diversification of myosins in new detail.

Materials and Methods

Taxon Sampling and Sequence Retrieval

Myosin sequences were queried in complete genome or transcriptome sequences of 62 taxa representing all known eukaryotic supergroups. Taxon sampling included 8 animals, 10 unicellular holozoans, 12 fungi, 1 apusozoan, 3 amoebozoans, 5 plants, 4 chlorophytes, 2 rhodophytes, 5 heterokonts, 5 alveolates, 1 rhizarian, 1 haptophyte, 1 cryptophyte, and 4 excavates (supplementary table S2, Supplementary Material online). The complete proteomes of all included species were analysed using Pfamscan (a HMMER search-based algorithm; Punta et al. 2012) with the default gathering threshold. Using custom Perl scripts, the resulting output files were parsed and all proteins containing a Myosin_head (PF00063) domain were extracted.

Phylogenetic Analyses

The sequences retrieved were aligned using the Mafft L-INS-i algorithm, optimized for local sequence homology (Katoh et al. 2002, 2005). The alignment was then manually inspected and edited in Geneious. This resulted in a matrix containing 353 amino acid residues, belonging to the Myosin_head domain (as this is the only conserved domain across all myosin classes). This way we avoid as well any effect

that convergently acquired protein domain architectures may have while inferring the phylogeny.

Maximum likelihood (ML) phylogenetic trees were estimated by RaxML (Stamatakis 2006) using the PROTGAMMALGI model, which uses the Le and Gascuel (LG) model of evolution (Le and Gascuel 2008) and accounts for between-site rate variation with a four category discrete gamma approximation and a proportion of invariable sites (LG + Γ + I). Statistical support for bipartitions was estimated by performing 1,000-bootstrap replicates using RaxML with the same model. Bayesian inference trees were estimated with Phylobayes 3.3 (Lartillot et al. 2009), using two parallel runs for 500,000 generations and sampling every 100 and with the LG + Γ + I model of evolution. Bayesian posterior probabilities (BPP) were used for assessing the statistical support of each bipartition.

Concurrent Domain Analysis

The domain architecture of all retrieved sequences was inferred with Pfamscan (Punta et al. 2012), using the gathering threshold as cutoff value. Then, the number of different concurrent domains (domains encoded within the same predicted open reading frame [ORF]) was calculated for each species using custom Perl scripts (excluding the myosin head domain itself). This information was further used to build Venn diagrams of shared concurrent domains between groups, using custom Bash scripts and the website: http://bioinformatics.psb.ugent.be/webtools/Venn/ (last accessed January 29, 2014).

Results and Discussion

Myosin Classification

Our genomic survey and phylogenetic analyses defined 31 myosin classes. Figure 1 displays their distribution across eukaryotic taxonomic groups and their canonical protein domain architecture for each class and subclass. Our data corroborated previous findings (Richards and Cavalier-Smith 2005; Foth et al. 2006; Odronitz and Kollmar 2007) and also identified a number of new families. This was somewhat expected, given that the number of myosin classes discovered has grown considerably since the pioneering studies of Cheney et al. (1993) and Goodson and Spudich (1993). For the sake of clarity, we incorporated the nomenclature used in previous studies (Cheney et al. 1993; Goodson and Spudich 1993; Hodge and Cope 2000; Berg et al. 2001; Thompson and Langford 2002; Richards and Cavalier-Smith 2005; Foth et al. 2006; Odronitz and Kollmar 2007; Syamaladevi et al. 2012), except for a number of classes in which there were conflicting names (see table S1, Supplementary Material online, for a comparison of nomenclature among studies). We dismissed and/or reused class names only on those cases in which we unambiguously inferred a different phylogenetic relationship, and therefore alternative classification, to that identified in previous analyses. Thus, our new updated and integrative classification provides a useful systematic framework for myosins.

Myosin I, the Largest Myosin Class, Has Five Subclasses

Myosin I (bootstrap support [BS] = 64%, BPP = 1.0; see fig. 2 and supplementary fig. S1, Supplementary Material online) comprises five subclasses including myosin Ik, newly identified here (BS = 79%, BPP = 0.99). Subclasses c/h, d/g, and a/b (named according to their vertebrate co-orthologs) have a tail composed of IQ domains (PF00612) and a myosin TH1 domain (PF06017). Co-orthologs of these four subclasses are present in several eukaryotic taxa (fig. 1). Interestingly, we find orthologs of each subclass in unicellular holozoans. Myosin Ik, which is found in choanoflagellates, filastereans, ichthyosporeans, and, with weaker support, in *Thecamonas trahens*, was lost in metazoans, and thus the diversification of these four subclasses (Ia/b, Ic/h, Id/g, and Ik) most likely occurred in the common ancestor of Holozoa prior to the radiation of Metazoa.

Myosin II Is Not a Valid Molecular Synapomorphy for Amorphea

Myosin II is the second largest class of myosins, and is characterized by a myosin N-terminal domain (PF02736) and a tail containing an IQ domain and a myosin tail domain (PF01576), consisting of several coiled-coil domains. Although myosin II was previously thought to be exclusive to amorpheans (also known as unikonts [Adl et al. 2012]) and was used as a phylogenetic marker (Richards and Cavalier-Smith 2005), a myosin II homolog was recently identified in the excavate *N. gruberi* (Odronitz and Kollmar 2007; Fritz-Laylin et al. 2010). Myosin II therefore probably had a deeper ancestry, although a HGT event from Amoebozoa to Excavata cannot be ruled out—especially considering the several cases of HGT that have recently been described between Heterolobosea and Amoebozoa (Andersson 2011). However, myosin proteins form numerous and specific interactions with actin filaments, plasma membrane, and numerous secondary protein complexes. Proteins with complex protein–protein interaction networks have been shown to be less likely to undergo HGT probably because integration into foreign protein interactions is limited (Jain et al. 1999; Cohen et al. 2011). Therefore, our favoured explanation for aberrant taxon distribution of myosin orthologs and domain architecture patterns identified in this study (as in the case of myosin VI discussed below) are patterns of multiple secondary loss or convergence, rather than HGT. Irrespective of whether the *N. gruberi* myosin II is a result of HGT or not, this shows that myosin II is no longer a valid molecular synapomorphy for amorpheans.

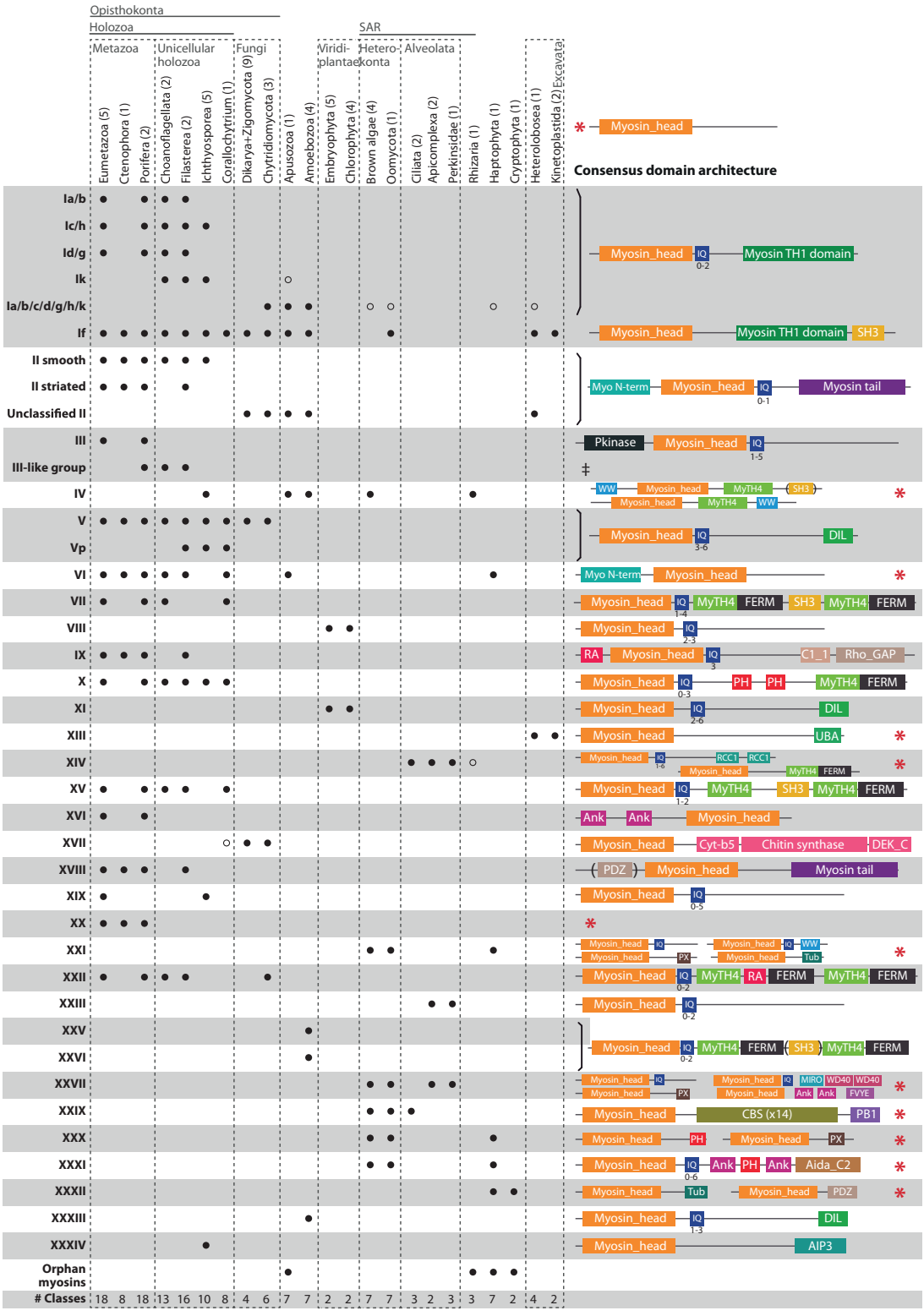


Fig. 1.—Phylogenetic distribution of myosin classes in eukaryotic genomes. The domain architectures for each class are shown, with a red asterisk on the right indicating that a single myosin head domain is also found in some sequences within that particular class. Filled circles indicate the presence of orthologs of a myosin class in a particular lineage. Unclear putative orthologs are shown with empty circles (see text). The presence of orphan myosins (i.e., species-restricted myosin classes) is also indicated. The total number of classes in each linage is indicated in the lower row. The number of species included in each taxonomic group is shown in parentheses. ‡See figure 3 for a detailed description of the various domain architectures.

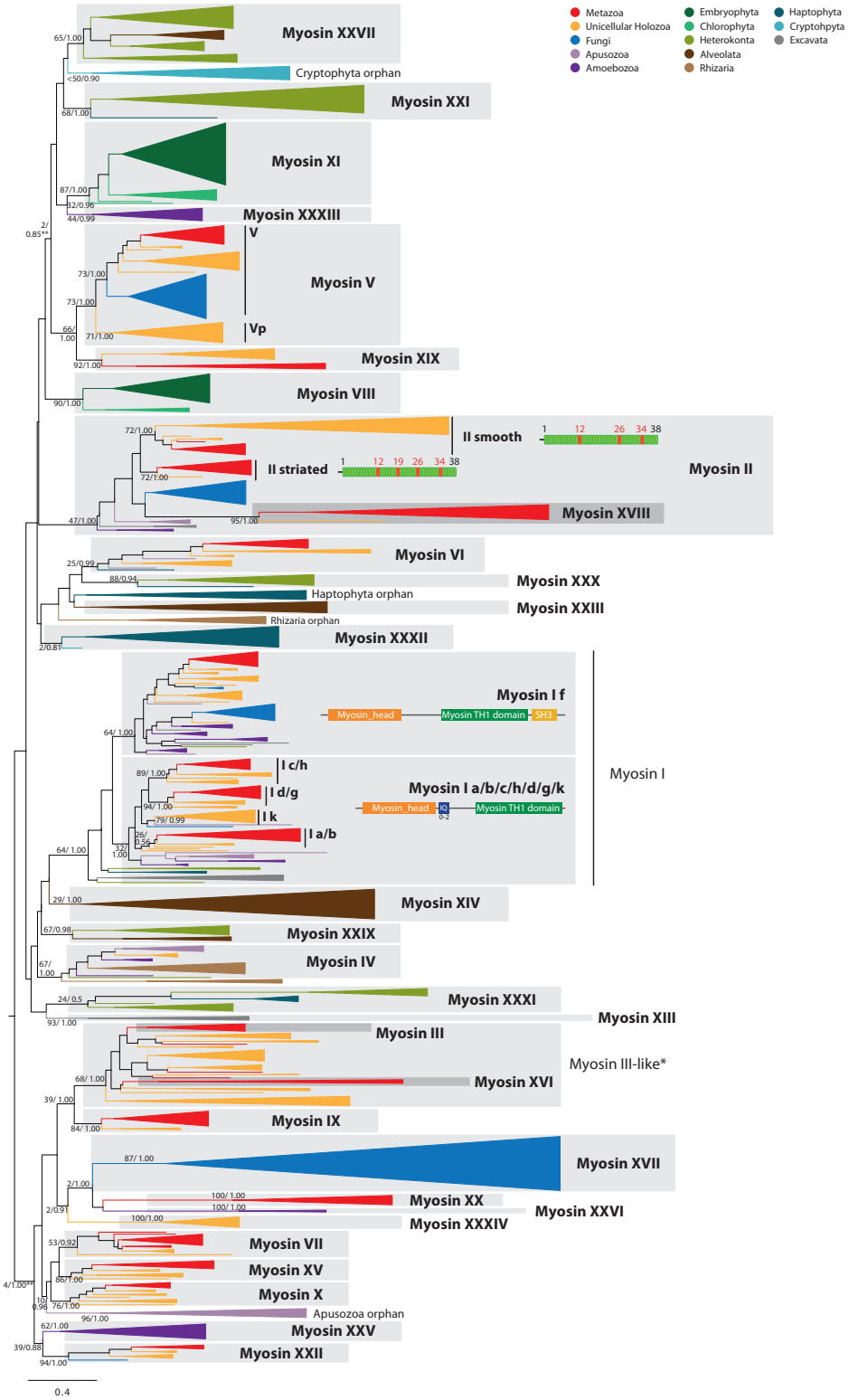


Fig. 2.—ML tree of myosin head domains. The tree is collapsed at key nodes and rooted using the midpoint-rooted tree option. Myosin classes are indicated. Nodal support was obtained using RAxML with 1,000 bootstrap replicates and BPP. Both values are shown on key branches. Taxa are color-coded according to taxonomic assignment (indicated in the upper right). The abbreviations are indicated in supplementary table S1, Supplementary Material online. Specific domain architectures are highlighted for myosin classes I and II (see text). **See figure 3 and supplementary figure S3 (Supplementary Material online) for more detail on the phylogeny of MyTH4-FERM and V-like myosins, respectively.

Striated Muscle Myosin II in Holozoa

Interestingly, myosin II is the major motor protein involved in actomyosin contraction in metazoan muscle and nonmuscle cells (Clark et al. 2007), providing contractile force during cytokinesis in the latter (Matsumura 2005), a function also performed by members of yeast myosin class II (East and Mulvihill 2011). Metazoans have two subclasses of myosin II, referred to here as smooth (Myo2) and striated (Myo11/zipper) muscle myosins (fig. 1), which have been shown to have architectural differences in the composition of their coiled-coil domains and to have originated most likely at the stem of Holozoa, although striated muscle myosin was later lost in unicellular holozoans (Steinmetz et al. 2012). We confirm this hypothesis by showing that an extant filasterean species, *Ministeria vibrans*, has a striated myosin homolog (BS = 72%, BPP = 1.0) with the extra 29 aa-based coiled-coil that is typical of striated muscle myosin II (fig. 2) (Steinmetz et al. 2012). We therefore infer that myosin II was derived early in the radiation of the eukaryotes and diverged into two classes in the holozoan lineage (smooth and striated), the latter being secondarily lost in ichthyosporeans and choanoflagellates.

Myosin III-Like: An Expanded Holozoan Clade

The myosin III class is characterized by an N-terminal Protein kinase domain (PF00069) and several IQ domains (fig. 1). It is strictly metazoan-specific, although a larger group of choanoflagellate, sponge, and filasterean sequences appear to be related to it (BS = 68%, BPP = 1.0) (figs. 1, 2, and 6). This group represents a choanoflagellate-specific expansion of myosin genes, with different domain arrangements, including some members with protein kinase domains, WW domains (PF00397), SH2 domains (PF00017), PH domains (PF00169), Y-phosphatase domains (PF00102), and others (discussed later; fig. 3). The metazoan-specific myosin XVI is also related to myosin III and myosin III-like sequences. Our data demonstrate that myosin III-like originated at the stem of the Filozoa clade (i.e., Filasterea, Choanoflagellata, and Metazoa), acquiring its definitive domain configuration (with an N-terminal protein kinase domain) and leading to the birth of an additional paralog class (myosin XVI) at the base of the Metazoa.

Myosin IV Is Not an Orphan Acanthamoeba castellanii Myosin

All myosin IV proteins have WW domains that can either be N-terminal or C-terminal to the Myosin_head domain, and a tail with a MyTH4 domain (PF00784), followed in some cases by a SH3 domain (in *T. trahens* and ichthyosporeans) (fig. 1). Previously considered an orphan myosin of the amoebozoan *Acanthamoeba castellanii* (Odronitz and Kollmar 2007), our results show that many other lineages have class IV myosins namely, ichthyosporeans, apusozoans, rhizarians, and heterokonts (BS = 67%, BPP = 1.0; figs. 1 and 2). Thus, despite its patchy distribution, it is likely that this myosin class was present in the LECA (fig. 4).

Myosin V and Related Myosins: A Large Assembly of Related Proteins

Class V myosins have an N-terminal Myosin_head domain and a C-terminal tail with IQ and a globular DIL domains (PF01843) (fig. 1). Myosin V and the structurally similar plant myosin XI carry a remarkable variety of cargo, including organelles, vesicles, and protein complexes (Li and Nebenführ 2008; Loubéry and Coudrier 2008). A relationship between myosin V and plant myosin XI has long been proposed due to their similar domain architectures (Richards and Cavalier-Smith 2005; Li and Nebenführ 2008). Moreover, the orthology between opisthokont myosin V and amoebozoan myosin V (renamed here as myosin XXXIII) was assumed but not well-supported phylogenetically (Foth et al. 2006; Odronitz and Kollmar 2007). Here, we show that all myosin V-like proteins cluster together phylogenetically with low ML nodal support in the global analysis (BS = 2%, BPP = 0.85), but maximum nodal support (BS = 100%, BPP = 1.00) if a closer outgroup is used (supplementary fig. S3, Supplementary Material online). This group includes other bikont myosins with different domain architectures. Therefore, we propose a unique ancestral origin in the LECA for the progenitor of this paralogous family (fig. 2; supplementary figs. S1–S3, Supplementary Material online). We group them in several classes, including plant myosin XI (BS = 87%, BPP = 1.0), opisthokont myosin V (BS = 73%, BPP = 1.0), amoebozoan myosin XXXIII (BS = 44%, BPP = 0.99) (formerly called myosin V, but phylogenetically not related to it), stramenopile + haptophyte myosin XXI (BS = 68%, BPP = 1.0), stramenopile + alveolate myosin XXVII (BS = 65%, BPP = 1.0), and a group of *Guillardia theta* orphan myosins (BS = 38%, BPP = 0.9) (these last three do not have the consensus myosin V architecture, presenting a wide variety of alternative domain architectures) (fig. 1). In the case of opisthokont myosin V, we confirm that myosin XIX is related to it (BS = 66%, BPP = 1.0), but we demonstrate that it is not a metazoan-specific class because it is also present in ichthyosporeans. Moreover, our phylogenetic trees strongly suggest that myosins V and Vp originated in the last common ancestor of opisthokonts (BS = 73%, BPP = 1.0) (supplementary fig. S3, Supplementary Material online). Myosin Vp was secondarily lost in fungi, metazoans, and choanoflagellates. Interestingly, the two filasterean species analysed have differentially lost one or the other, as *Capsaspora owczarzaki* has myosin Vp and *M. vibrans* has myosin V (fig. 1).

Myosin VI Is Mostly Specific to Opisthokonta and Apusozoa

The unique class VI myosins move toward the minus end of actin filaments, in contrast to all other known myosins.

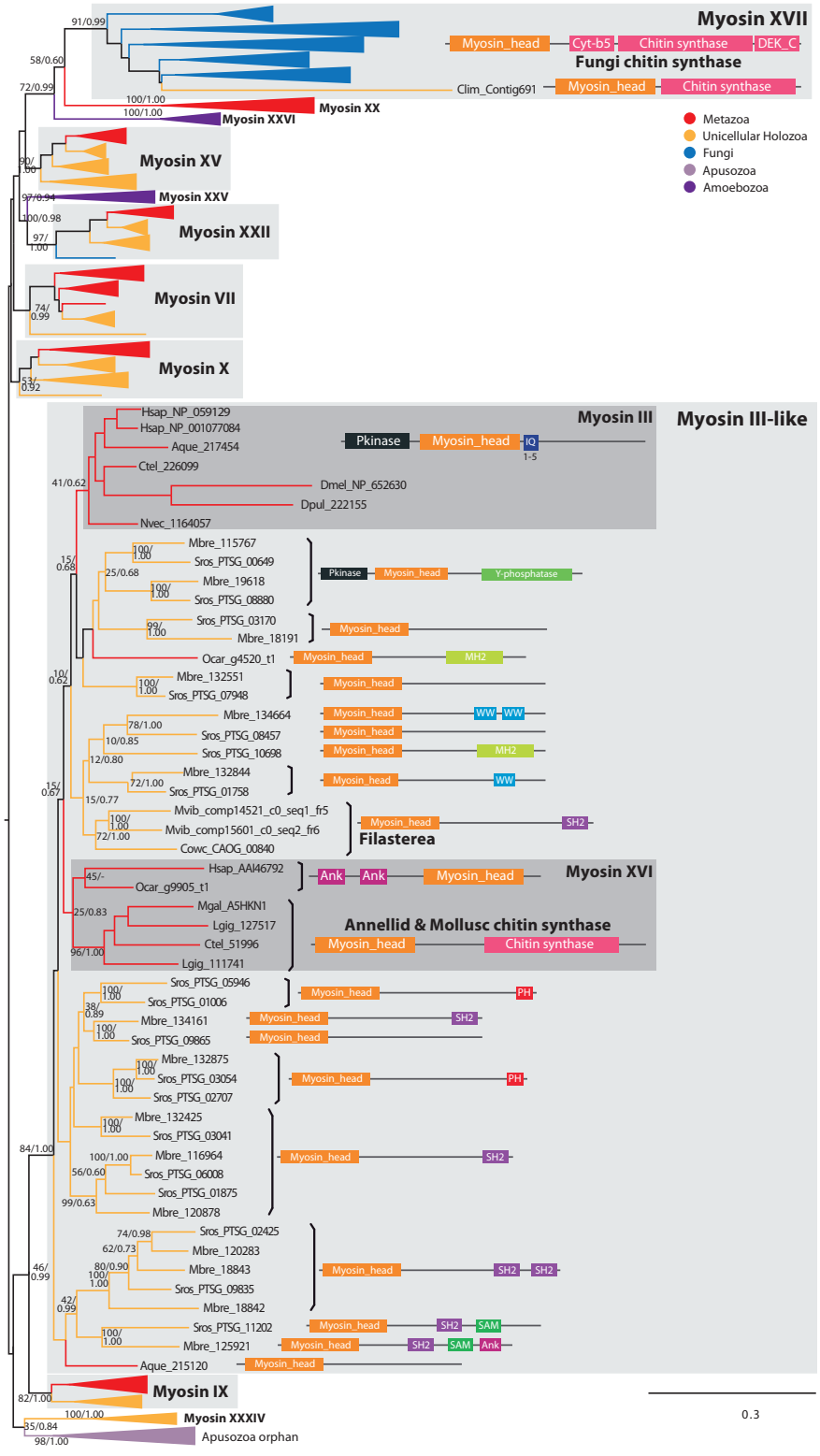


Fig. 3.—Convergent evolution of animal-fungal chitin synthases and myosin III lineage-specific expansion. ML tree of MyTH4-FERM myosin head domains. The tree is collapsed at key nodes and rooted using the midpoint-rooted tree option. Myosin classes are indicated. Statistical support was obtained by RAxML with 1,000 bootstrap replicates and BPP. Both values are shown on key branches. Taxa are color-coded according to taxonomic assignment as in figure 2. The protein domain architectures of key sequences are shown. The abbreviations are provided in supplementary table S1, Supplementary Material online.

Myosins from this class are involved in diverse processes such as cytokinesis, transcription regulation, and endocytosis (Roberts et al. 2004; Sweeney and Houdusse 2010). Our phylogeny shows that homologs of this class are present in metazoans, choanoflagellates, filastereans, *Corallochytrium limacisporum,* and apusozoans, but not in fungi or amoebozoans (fig. 1). Foth et al. (2006) found putative VI-like genes in alveolates, but our analysis places them within myosin XXIII (supplementary table S1, Supplementary Material online). Yet, we identified an ortholog in the haptophyte *Emiliania huxleyi* (BS = 25%, BPP = 0.99). It is not clear whether this non-amorphean myosin VI represents an ancestral member that was lost in all other bikonts, or whether it derives from a HGT event. The fact that this and a *T. trahens* homolog share a unique C-terminal RUN domain (PF02759) that is not found in any other myosin supports the latter possibility.

Myosins VII, IX, X, XV, XVIII, and XIX Are Holozoan Specific

Myosins VII, IX, X, XV, XVIII, and XIX were previously considered to be unique to animals (Odronitz and Kollmar 2007), but we demonstrate the presence of clear orthologs in unicellular holozoans as well. In mammals, myosin VII is a MyTH4-FERM myosin class found in structures based on highly ordered actin filaments, such as stereocilia and microvilli (Henn and De La Cruz 2005). Its members have a tail with two MyTH4 domains (PF00784), two FERM (PF00373) domains, likely the product of a partial gene tandem duplication, and addition of a SH3 domain. Myosin VII homologs are found only in metazoans, choanoflagellates and *Co. limacisporum* (fig. 1). Some authors described a group of amoebozoan proteins with a similar architecture, involved in chemotaxis and cell polarization (Breshears et al. 2010), and identified them as VII myosins. Yet, our phylogenetic analysis does not place them with the Holozoan VII class and, therefore, we reclassify them as myosin XXV (discussed later).

Myosin VII is phylogenetically related to myosins X and XV (the other MyTH4-FERM myosins found in metazoans, discussed later) and to a group of apusozoan orphan myosins, although with low nodal support in ML analysis (BS = 10%, BPP = 0.96) (fig. 2; supplementary figs. S1 and S2, Supplementary Material online). Our results therefore suggest that all three originated from a single ancestral protein in the last common ancestor of Holozoa (being differentially lost in some unicellular lineages; only the unicellular *Co. limacisporum* has orthologs of all three classes, XV, X, and VII). Interestingly, ctenophores have lost these three myosin classes. Myosin IX is composed of a N-terminal RA domain (PF00788) and a tail with IQ domains, a C1_1 domain (PF00130) and a RhoGAP domain (PF00620). Homologs of this class are found only in metazoans and filasterea (fig. 1).

Myosin X and XV are MyTH4-FERM classes of crucial importance for metazoan filopodia (Zhang et al. 2004; Bohil et al. 2006; Liu et al. 2008). The tail of myosins X is composed of a variable number of IQ motifs, two PH (PF00169), one MyTH4, and one FERM domain; while those of myosins XV are composed of two MyTH4, one FERM, and one SH3 domain. Myosin XVIII often has an N-terminal PDZ domain and has a C-terminal myosin tail domain. This family is present in the filasterean *C. owczarzaki* and all metazoans examined (BS = 95%, BPP = 1.0) (fig. 1). Although not statistically supported, myosin XVIII could be closely related to myosin II, as previously described (Foth et al. 2006). Finally, myosin XIX has a variable number of IQ domains and it is only found in eumetazoans and ichthyosporeans (BS = 92%, BPP = 1.0) (fig. 1). It is closely related to myosin V (BS = 66%, BPP = 1.0) (fig. 2; supplementary figs. S1–S3, Supplementary Material online).

Myosin VIII and XI: The Green Lineage Myosins

Myosins VIII and XI are the only myosin classes present in plants and several chlorophytes (Peremyslov et al. 2011; fig. 1). Myosin VIII, whose monophyly is strongly supported (BS = 90%, BPP = 1.0), has a tail with IQ domains. As for myosin XI, several authors have pointed out its strong similarity to myosin class V in terms of domain architecture (Thompson and Langford 2002; Foth et al. 2006; Li and Nebenführ 2008). Here, we show that this class is found in embryophytes and chlorophytes and is well supported (BS = 87%, BPP = 1.0; fig. 1). This class is phylogenetically related to myosin V, and is included in a major myosin cluster that we name myosin V-like (fig. 2; supplementary figs. S1–S3, Supplementary Material online).

Myosin XIV: Myosins with a MyTH4-FERM Protein Domain Combination in a Ciliate

Myosin XIV has been shown to be involved in phagosome motility and nuclear elongation in the ciliate *Tetrahymena thermophila* (Williams and Gavin 2005; Foth et al. 2006). We find that this is an alveolate-specific class that has expanded in many species (specifically in ciliates) and that shows various domain architectures. Interestingly, the ciliate *Te. thermophila* has several myosin XIV homologs with MyTH4 and FERM domains, and is the only known bikont (non-amorphean) taxon with myosins that have a MyTH4-FERM protein domain combination. This configuration is very common in amorphean myosins, and was probably convergently acquired in the ciliates.

Myosin XVI and XVII: Convergence of Fungal and Animal Myosins with a C-terminal Chitin Synthase

Myosin XVII, also called chitin synthase, is a fungus transmembrane myosin with Cyt-b5 (PF00173), chitin synthase 2

(PF03142) and DEK_C (PF08766) domains in its tail, a domain combination unique to this class. Its monophyly is well supported (BS = 91%, BPP = 0.99), and it is phylogenetically related to amorphean FERM domain myosins. This chitin synthase class was thought to be specific to Fungi (James and Berbee 2012). Interestingly, the holozoan *Co. limacisporum* has a highly derived myosin that is associated with a chitin synthase domain and that is phylogenetically related to the fungal myosin XVII (fig. 3). This implies that class XVII chitin synthase precedes the appearance of the Opisthokonta and was lost in most holozoan lineages (except for *Co. limacisporum*) and so is not a valid synapomorphy for the fungi (James and Berbee 2012). Moreover, we also identified myosins with chitin synthases in annelids and molluscs (figs. 1 and 3), which are members of the XVI class. Thus, they are not orthologous to fungus chitin synthases, but rather appeared convergently in annelids and molluscs (fig. 3).

Myosin XXII: An Opisthokont-Specific Myosin with a Scattered Taxonomic Distribution

Myosin XXII is a MyTH4-FERM domain myosin found in some opisthokonts, including the chytrid fungus *S. punctatus*, filastereans, choanoflagellates, poriferans, and *Drosophila melanogaster*. Its tail is composed of an IQ, two MyTH4 and two FERM domains, with a RA domain (PF00788) between the first MyTH4 and the first FERM domain. It was secondarily lost in *Co. limacisporum*, ichthyosporeans, and many metazoans (fig. 1). Myosin XXII seems to be related to amoebozoan myosin XXV (fig. 2). They may comprise a single class, although there are some architectural differences between them (discussed later).

Myosin XXI, XXX, and XXXI: Heterokonta and Haptophyta Share Unique Myosins

These three myosin classes are found in heterokonts and haptophytes, which suggests that they were secondarily lost in rhizarians and alveolates (figs. 1 and 4) as these groups are thought to branch closer to heterokonts than haptophytes (Burki et al. 2012). Myosin XXI homologs present diverse myosin tail architectures, including IQ, WW (PF00397), PX (PF00787), and Tub (PF01167) domains. This class has become considerably expanded in the oomycete *Phytophthora infestans*. Myosin XXX homologs in *E. siliculosus* have a C-terminal PH domain and *P. infestans* homologs have a PX domain. Finally, the myosin XXXI class, in which we also include the old myosin XXXIII (Odronitz and Kollmar 2007), has a characteristic tail architecture in several heterokonts homologs, with a variable number of IQ domains, a PH domain flanked by two ankyrin domains, and a C-terminal Aida_C2 domain (PF14186).

Myosin XXV, XXVI, and XXXIII: Renamed Amoebozoa-Specific Myosins

The myosin XXV class (BS = 62%, BPP = 1.0) comprises amoebozoan sequences that were previously considered to be myosin VII homologs. They are MyTH4-FERM myosins known to have a role in cell adhesion and filopodia formation (Breshears et al. 2010). They show remarkable architectural similarities with both myosin XV and myosin VII (fig. 1), but seem to be phylogenetically related to myosin XXII (although they have different tail architectures and their sister-group relationship is low supported) (fig. 2; supplementary figs. S1 and S2, Supplementary Material online), and thus were classified as an independent class. Myosin XXVI (BS = 100%, BPP = 1.0) is another class of amoebozoan MyTH4-FERM myosins, which does not cluster with either myosin VII or myosin XXV. We suggest a common ancestry for a group of amorphean myosin classes that are generally characterised by the presence of MyTH4 domains. This group includes these two amoebozoan classes (XXV and XXVI; fig. 2; supplementary figs. S1, S2, and S4, Supplementary Material online), as well as myosins III, XVI, IX, XVII, XX, XXXIV, X, XV, VII, and XXII (fig. 3; supplementary fig. S4, Supplementary Material online).

Myosin XXXIII includes the amoebozoan sequences previously considered as class V myosin, and shares the same domain architecture as plant myosin XI. Our phylogenetic analysis does not support a close relationship between myosin XXXIII and myosin V; it rather demonstrates that they are related to the myosin V-like clade (fig. 2), leading us to rename the group as myosin XXXIII (fig. 2; supplementary figs. S1–S3, Supplementary Material online).

The Evolution of the Myosin Repertoire in Eukaryotic Genomes

Phylogenetic analysis allowed us to define broader groups of myosin classes and to reconstruct the evolution of the myosin toolkit across the eukaryotes. This reconstruction is based on the favored hypothesis for the root of eukaryotes, the unikont–bikont split (Stechmann and Cavalier-Smith 2002; Richards and Cavalier-Smith 2005), that has recently been recovered in a rooted multi gene concatenated phylogeny with a modification with regards to the placement of the apusozoan *T. trahens* within the unikonts (Derelle and Lang 2012). Based on this root, our data suggest that the LECA had at least six myosin types, with different protein domain architectures (fig. 4 for the reconstruction of LECA and other ancestral nodes). According to our reconstruction, LECA had the following: 1) an ancestral myosin I (progenitor paralog of the myosin I a/b/c/h/d/g/k ortholog subfamilies) with an architecture consisting of a myosin head domain followed by 0 to 2 IQ repeats and a C-terminal myosin TH1 domain; 2) a myosin If, with a myosin head domain followed by a myosin TH1 domain and a C-terminal SH3 domain; 3) a myosin II, with a myosin N-terminal domain, a myosin motor domain, 0 to 1 IQ

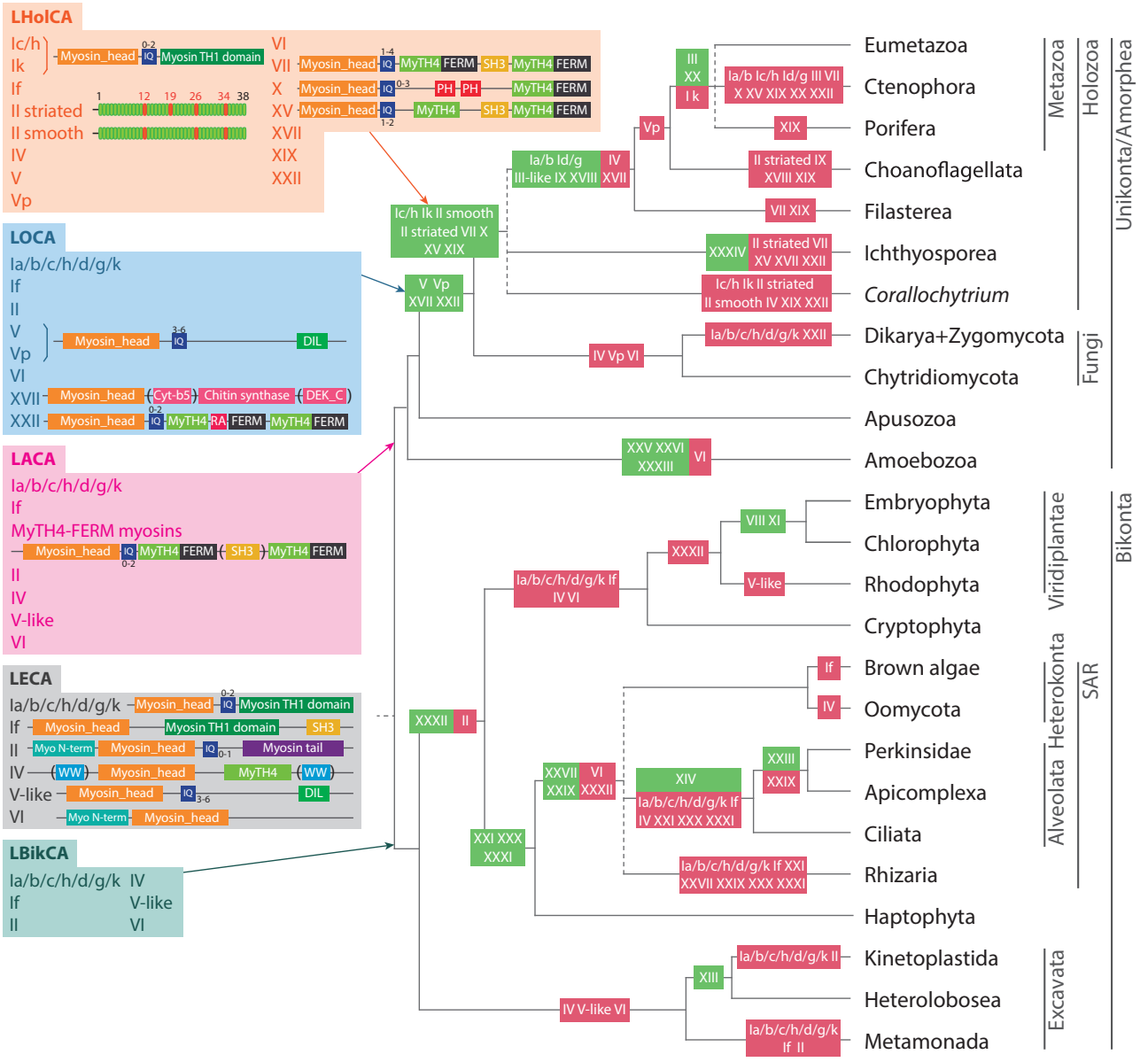


Fig. 4.—Reconstruction of myosin evolution in eukaryotes. Key ancestral nodes, including inferred domain architectures, are reconstructed, including LECA, LBikCA, LACA, modified after Derelle and Lang (2012), LOCA, and LHolCA. Domain architecture is only shown at the most ancient inferred presence of a particular myosin type (e.g., myosin if only at the LECA reconstruction). The appearance and loss of myosin classes are mapped in green and red, respectively. Dashed lines indicate unresolved phylogenetic relationships. Tree topology is based on different recent phylogenomic studies (Dunn et al. 2008; Hampl et al. 2009; Burki et al. 2012; Derelle and Lang 2012; Torruella et al. 2012; Laurin-Lemay et al. 2012; Sierra et al. 2013).

domains and a myosin tail domain; 4) a myosin IV with a myosin head domain followed by a MyTH4 domain and a characteristic WW domain (either C-terminal or N-terminal); 5) a myosin V-like myosin with a myosin head followed by variable number of IQ repeats and a C-terminal DIL domain; and 6) a myosin VI, with a myosin N-terminal domain followed by a myosin head domain.

In figure 4, we show the diversity of the myosin complement in the LECA genome under a modified version of the unikont–bikont root. Our reconstruction indicates that LECA possessed a minimum of six paralog families all encoding different protein domain architectures. Even if alternative rooting hypothesis are taken into account (Rodríguez-Ezpeleta et al. 2007; Wideman et al. 2013) the inferred number of myosin paralog families in the LECA is still high (supplementary figs. S5 and S6, Supplementary Material online). This result is consistent with the pattern observed in the kinesin gene family, which also demonstrated a diverse repertoire of

paralog families present in the LECA (Wickstead et al. 2010). Together these data suggest that the LECA possessed a complex and diversified actin and tubulin cytoskeleton and that this ancestral cell possessed a large number of complex eukaryotic cellular characteristics prior to the diversification of extant and sampled eukaryotic groups. Assuming this root, these results have two implications: 1) they strongly suggest that a large quantity of protein diversification and cellular complexity evolved between the point of eukaryogenesis (Martin et al. 2001) and LECA, and 2) indicate that gene loss and subsequent reduction in cytoskeletal systems played a significant role in the diversification of eukaryotes, a pattern that is increasingly apparent on other gene families and cellular systems (Wolf and Koonin 2013).

Our analysis reconstructed the LBikCA (Last Bikont Common Ancestor) with the same complement of myosins as the LECA (fig. 4). New classes appeared later in bikont evolution, such as myosin XIII at the stem of Kinetoplastida + Heterolobosea and myosin XXI, XXX, and XXXI at the stem of SAR + Haptophyta. Assuming the unikont–bikont root, our analyses demonstrate that many groups underwent secondary losses, with two extreme cases of complete loss of the myosin toolkit in the following: 1) metamonads (including *Trichomonas vaginalis* and *Giardia lamblia*) and 2) rhodophytes (including the unicellular *Cyanidioschyzon merolae* and the multicellular alga *Chondrus crispus*) (figs. 4 and 5).

The LACA (Last Amorphean Common Ancestor, modified by inclusion of Apusozoa [Derelle and Lang 2012]) added a new myosin type from LECA, a MyTH4-FERM myosin (Berg 2001; Richards and Cavalier-Smith 2005) that includes several phylogenetically related myosin classes (supplementary figs. S1, S2, and S4, Supplementary Material online). These myosins have a complex protein domain architecture including a myosin head domain followed by 0 to 2 IQ repeats, a MyTH4 domain, a FERM domain, in some cases a SH3 domain, and an additional MyTH4 and FERM domains (fig. 4). This ancestral protein domain architecture is found in diverse myosins from

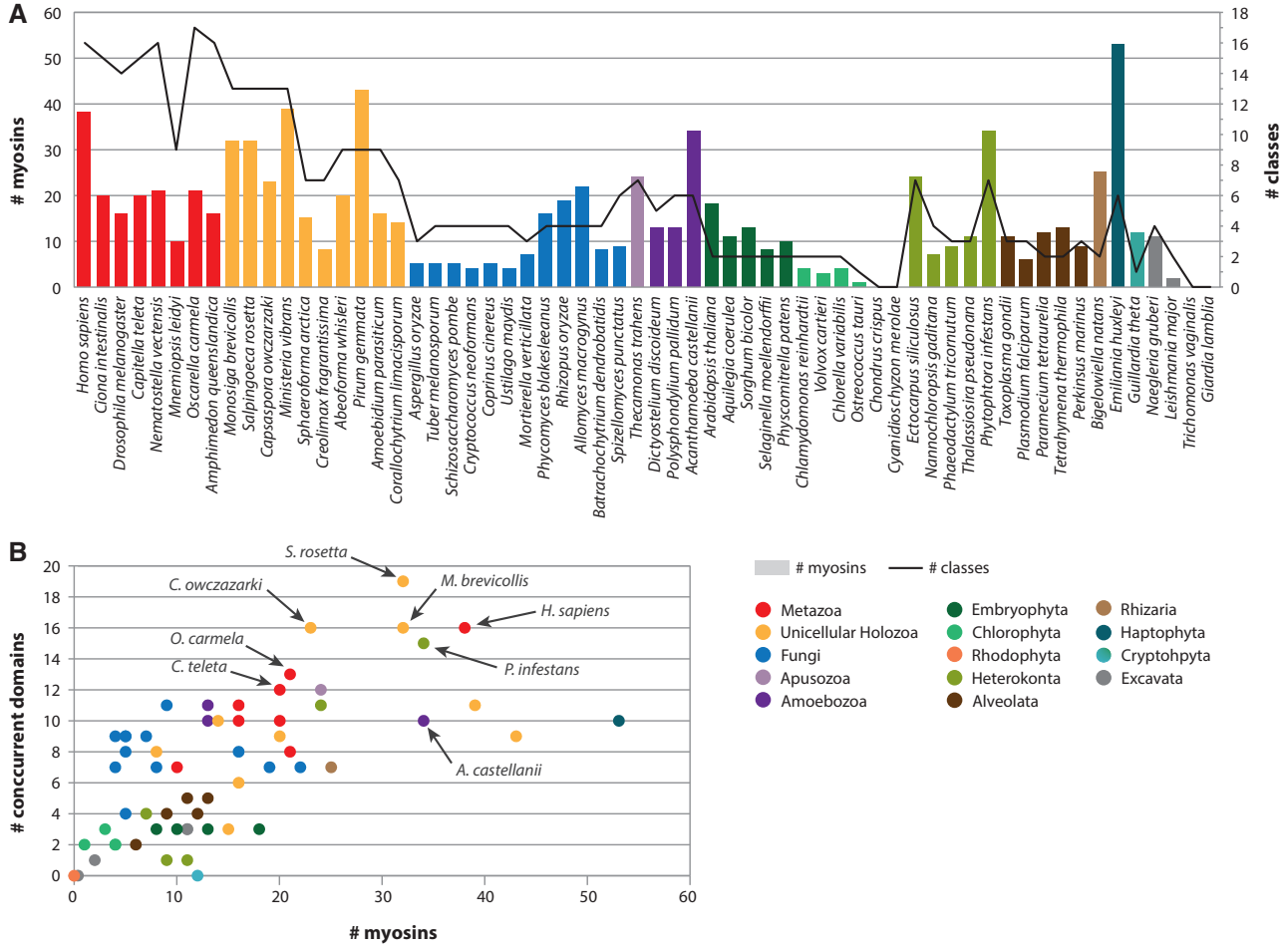


Fig. 5.—Phylogenetic patterns of myosin diversity. Taxa are color-coded according to taxonomic assignment. (*A*) Number of myosin genes (right *Y*-axis, columns) and number of myosin classes (left *Y*-axis, black line) in each species. (*B*) Number of concurrent protein domains (*Y*-axis) compared with the number of myosin genes (*X*-axis) in each species.

extant amoebozoans (classes XXV and XXVI) to holozoans (class VII and, with some variations, classes XV and X). In any case, the putative ancestral MyTH4-FERM myosin underwent major architectural rearrangements as the family expanded during diversification of the amorpheans (figs. 3 and 4).

The LOCA (Last Opisthokont Common Ancestor) had an even more complex myosin complement, with the addition of new myosin classes (as a consequence of the diversification of ancestral myosin types, such as myosin V-like or MyTH4-FERM myosins), including myosin V, myosin Vp, myosin XVII (a chitin synthase that is present in all fungi and in a single holozoan species, *Co. limacisporum*, discussed earlier), and myosin XXII. This complexity became even greater in the LHolCA (Last Holozoan Common Ancestor), which had the highest diversity of myosin types among all reconstructed ancestors (fig. 4). This diversity was further expanded during holozoan evolution, with little innovation at the stem of Metazoa.

Phylogenetic Patterns of Myosin Diversity and Protein Domain Combinations

Our data show that there are strong phylogenetic patterns across lineages, in terms of abundance and number of classes, and the diversity of concurrent domains (i.e., domains that appear together with the myosin head domain in a given protein or ORF).

The number of myosin genes varies markedly between lineages (fig. 5*A*). Holozoan genomes, as well as some amoebozoans and heterokonts, have the highest numbers of myosins of all eukaryotes. In particular, the haptophyte *Em. huxleyi* has the highest number of myosin genes (53), followed by the ichthyosporean *Pirum gemmata* (43), the filasterean *M. vibrans* (39), and the metazoan *Homo sapiens* (38). On the other hand, dikaryan fungi, plants, green algae, alveolates, and some excavates have few or no myosins.

A comparison of the abundance of myosin proteins with the diversity of myosin classes (fig. 5*A*), reveals that *Em. huxleyi*, which has a high number of myosins, has only six myosin classes. This implies that the high number of myosin homologs found in this species is due to class-specific expansions rather than possession of a wide diversity of ancestrally derived myosin types. In contrast, many unicellular holozoans, especially choanoflagellates and filastereans, and some metazoans (such as *H. sapiens* and the homoscleropmorph sponge *Oscarella carmela*) have a high diversity of myosin classes. In general, our data reveal a marked increase in the number of myosin classes at the origin of Holozoa, although some specific taxa, such as the ctenophore *Mnemiopsis leidyi* and the ichthyosporeans *Sphaeroforma arctica* and *Creolimax fragrantissima,* secondarily reduced their repertoire of myosins.

Myosin motor domains are found in a diverse collection of protein domain architectures, therefore another aspect that reflects differences in myosin diversity is the number of concurrent protein domains found associated with the motor domain (fig. 5*B*). The richest species in terms of protein domain diversity attached to the myosin motor domain within a putative ORF are the choanoflagellate *Salpingoeca rosetta*, the filastereans *M. vibrans* and *C. owczarzaki* and the metazoan *H. sapiens*. This implies that myosins were highly diversified prior to the origin and divergence of metazoans. Indeed, the sponge *O. carmela* also has a rich repertoire of concurrent domains, which corroborates (together with the fact that it has the richest range of myosin classes among analysed taxa) that the myosin repertoire was already rich and diverse in early metazoan evolution.

Interestingly, the oomycete plant pathogen *P. infestans*, which has a high number of myosin genes, also shows a remarkable diversity of concurrent protein domains (Richards and Cavalier-Smith 2005), a feature that has already been described for other gene families (Grau-Bové et al. 2013). In contrast, the myosin-rich taxon *Em. huxleyi* is relatively poor in both class diversity (fig. 5*A*) and protein domain diversity. The poorest taxa in protein domain diversity are plants, chlorophytes, excavates and alveolates. The cryptophyte *G. theta* represents an extreme case with no identified protein domains within the predicted ORF of any of its 11 myosins.

An examination at the concurrent protein domain composition of myosin in different taxa (fig. 6) reveals that 14 protein domains are conserved between amorpheans and bikonts (fig. 6*A*) with similar levels of innovation in both clades (20 and 21 new concurrent protein domains, respectively). A comparison of the most widely studied eukaryote clades (metazoans, embryophytes, and fungi [fig. 6*B*]) reveals that there are no specific concurrent domains in plants (only those present in myosin XI, which are shared by metazoan and fungus myosin class V) and in fungi there are only two specific domains (those associated with myosin XVII, i.e., DEK_C and Cyt-b5). In contrast, metazoans have many specific domains associated with myosins.

Within amorpheans (fig. 6*C*) there is a core of conserved domains (such as Myosin_tail_1 or Myosin_TH1) and a burst of innovation in the Holozoa. A closer look reveals that most of these domain combinations are present in unicellular holozoans, while little actual innovation occurred at the origin of metazoans (only the PDZ domain) (fig. 6*D*). In contrast, every single unicellular holozoan lineage has new specific associated domains: three in choanoflagellates (Mcp5_PH, SAM_2 and Y_phosphatase), two in filastereans (Rap_GAP and zf-MYND) and two in ichthyosporeans (AIP3 and LIM).

Within bikonts (fig. 6*E*) there are no protein domains shared by all major lineages and little innovation in protein domain combinations is observed, except in the case of haptophytes (five domains) and particularly in the SAR clade (Stramenopiles/Heterokonta, Alveolata, and Rhizaria). A closer look at the SAR clade (fig. 6*F*) reveals that this diversification of protein domains is largely lineage-specific, with five new domains in alveolates and thirteen new domains in heterokonts.

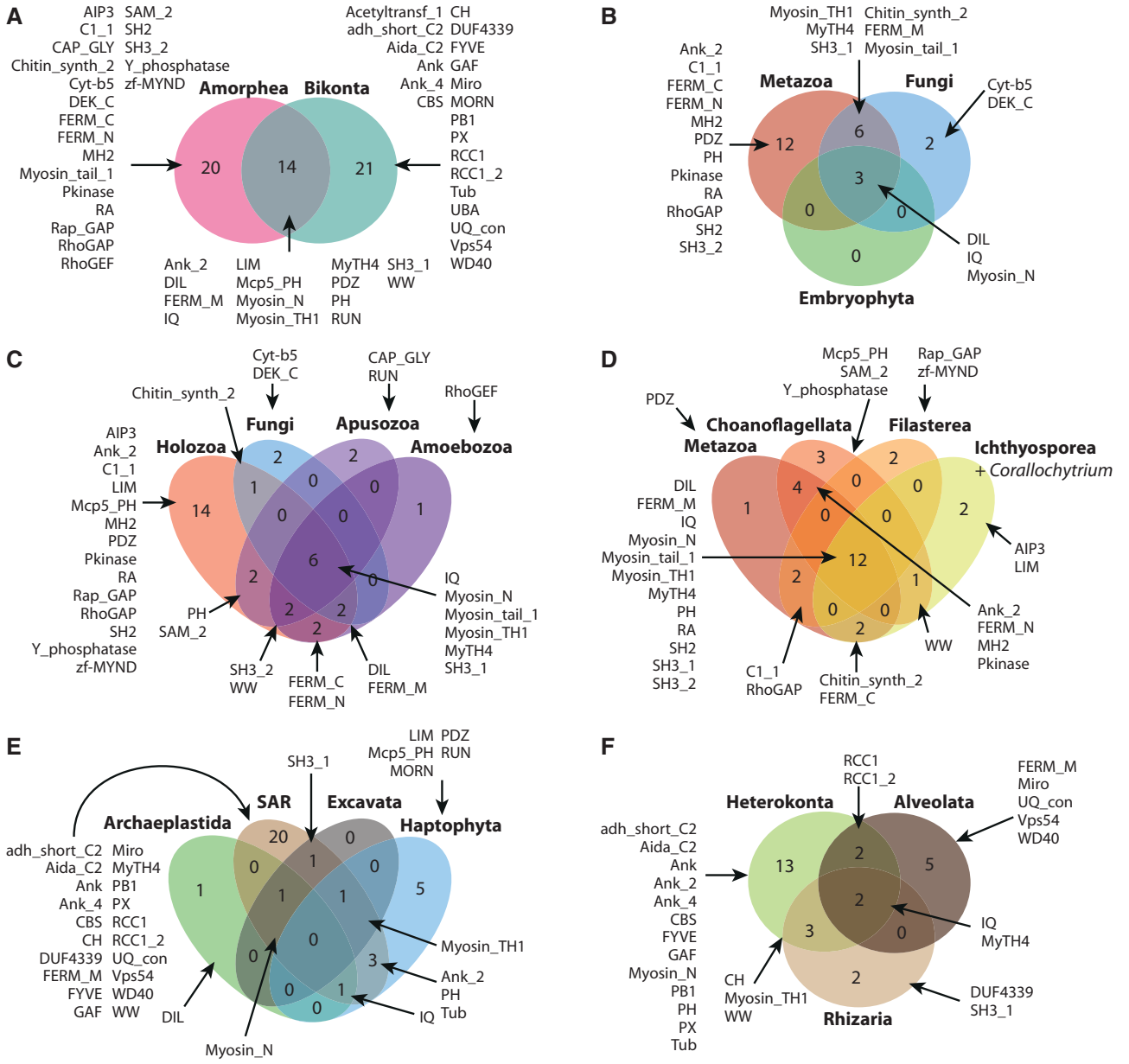


Fig. 6.—Myosin concurrent protein domains in different eukaryotic lineages. Venn diagrams show the number of shared and lineage-specific protein domains in different comparisons. The Pfam names of the various protein domains are indicated.

It is interesting to note that some of these shared protein domains were acquired convergently, for example the LIM domain in haptophytes and ichthyosporeans, the Mcp5_PH domain in haptophytes and choanoflagellates and the FERM_M domain in alveolates and amorpheans. This points to another source of homoplasy when considering protein domain architectures as evolutionary synapomorphies.

Lineage-Specific Myosin Diversifications

Our data show several lineage-specific expansions, often accompanied by major protein domain architecture rearrangements. This is the case, for example, of myosin class XXVII, which is expanded in both the oomycete *P. infestans* and the alveolate *Perkinsus marinus*, with unique protein domain architectures. Another example is the ciliate *Te. thermophila*, which has 12 myosin homologs of the alveolate-specific class XIV. In addition to the consensus architecture found in most alveolates, *Te. thermophila* myosin XIV is the only bikont myosin with the MyTH4-FERM domain combination, a domain architecture that was convergently acquired (compared with amorphean MyTH4-FERM myosins, discussed earlier).

The most spectacular lineage-specific expansion is that observed in choanoflagellate myosin III-like myosins (fig. 3). This phylogenetically defined group includes bona fide eumetazoan myosin III homologs, the related metazoan myosin XVI class (including annelid and mollusc chitin synthases), filasterean sequences (comprising a unique group), a single sequence of the sponge *Amphimedon queenslandica*, a single sequence of the sponge *O. carmela*, and several choanoflagellate myosins (15 from *Monosiga brevicollis* and 18 from *Sa. rosetta*). These choanoflagellate sequences have a wide diversity of protein domain rearrangements (fig. 3). Interestingly, many of these domains, like SH2 and Y-phosphatase domains, are related to tyrosine kinase signaling (Liu et al. 2011), a prominent feature of choanoflagellates (Manning et al. 2008). Sequences belonging to the myosin III-like group with a C-terminal SH2 domain were also identified in filastereans, which also have an extensive tyrosine kinase toolkit (Suga et al. 2012). Another interesting configuration found within this myosin III-like group is an *Sa. rosetta* and an *O. carmela* sequence with a C-terminal MH2 PF03166 domain. This domain is typically present in Smad transcription factors, where it is found at the C-terminal of the MH1 DNA-binding domain and acts as a protein binding motif that mediates cofactor interactions (Massagué et al. 2005). Interestingly, the MH2 domain is only found in choanoflagellates and metazoans, while Smad transcription factors are exclusive to animals (Sebé-Pedrós et al. 2011). The fact that the single MH2 domain found in choanoflagellates is associated with a myosin, together with that fact that the sponge *O. carmela* also has this configuration, suggests that MH2 initially appeared associated with myosins as a protein–protein interaction domain. Later on, early in metazoan evolution, MH2 was fused by domain shuffling to a MH1 DNA-binding domain to create the Smad transcription factors.

The Origin of the Metazoan Myosin Repertoire

Our results show that all metazoan myosin classes but one (Myosin XVI, also known as Dachs) have a premetazoan origin, many of them being holozoan innovations (fig. 6) (including myosin III-like, VII, IX, X, XV, XVIII, and XIX). Moreover, several subclass diversifications occurred in unicellular holozoans, for example in Myosin V (Myosin V and Myosin Vp), in Myosin I (Myosin I a/b, I/c/h, Id/g, and Ik) and in Myosin II (smooth and striated). In terms of number of myosins and diversity of concurrent domains (fig. 5), unicellular Holozoa have the highest counts among eukaryotes (even higher than most Metazoa). In fact, the choanoflagellate *Sa. rosetta* has the most diverse repertoire of myosin concurrent domains (fig. 5*B*), followed by another choanoflagellate (*Mo. brevicollis*), the filasterean *C. owczarzaki* and the metazoan *H. sapiens*. Overall, we can infer that the complexity of the myosin toolkit was extremely high before the advent of animal multicellularity and that this system is of paramount importance in extant unicellular holozoans.

Conclusions

We provide a robust updated myosin classification, based on ML and Bayesian phylogenetic methods and broad genomic taxon sampling that includes, for the first time, all major eukaryotic lineages. We provide a redefinition and/or confirmation of previously defined myosin classes (with an effort to reconcile myosin nomenclature between various previous classifications), and we assess the presence/absence of myosin classes in eukaryotes. Furthermore, we reconstruct a more complex myosin complement in the LECA genome than previously proposed, with six different myosin types and six different inferred domain architectures under the modified unikont–bikont root. Notably, we find strong phylogenetic patterns related to the complexity of the myosin system. Finally, we infer an intricate evolutionary history of the myosin gene family, including multiple lineage-specific expansions (such as the myosin III-like group in the choanoflagellate lineage), domain diversifications (specially in holozoans), secondary losses (in metamonads and rhodophytes), and convergences (e.g., in the fungal and metazoan myosin–chitin synthases). Taken together our results demonstrate that myosin gene family underwent multiple large-scale expansions and contractions in paralog families combined with extensive remodelling of domain architectures. As the diversity of this gene family directly relates to the function of the actin cytoskeleton, these results tell a story of extensive remodelling of this cytoskeleton system across the eukaryotes. These results also suggest that evolutionary inference of species relationships based on myosin distribution patterns is difficult without reliable phylogenetic analysis and comprehensive sampling. As such, the expansion of available genome data will provide a more accurate inference of the relative phylogenetic age of myosin classes and types—likely expanding the repertoire of myosins, and therefore the cellular complexity, of ancestral eukaryotic forms.

Supplementary Material

Supplementary data S1, figures S1–S6, and tables S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Patrick Steinmetz (Universität Wien) for his esteemed help and advice in the identification of structural motifs of myosins. We thank Andy Baxevanis (National Human Genome Research Institute), Scott A. Nichols (University of Denver) and Jonas Collén (Station Biologique Roscoff) for affably sharing unpublished protein sequences from *Mn. leidyi*, *O. carmela* and *Ch. crispus*, respectively.

I.R.-T. is an investigator from the Institució Catalana de Recerca i Estudis Avançats. T.A.R. is an EMBO Young Investigator, Leverhulme Early Career Fellow. T.A.R.'s work on motor protein evolution is supported by the Biotechnology and Biological Sciences Research Council (BBSRC - BB/G00885X/2). This work was supported by a European Research Council starting grant (ERC-2007-StG-206883) and a Ministerio de Economía y Competitividad (MINECO) grant (BFU2011-23434) to I.R.-T.; the Gordon and Betty Moore Foundation grant GBMF3307, the National Environment Research Council grant to T.A.R.; pregraduate Formación de Profesorado Universitario grant from MINECO to A.S.P and pregraduate Formación de Personal Investigador grant from MINECO to X.G.B.

Literature Cited

Adl SM, et al. 2012. The revised classification of eukaryotes. J Cell Biol. 59: 429–493.
Andersson JO. 2005. Lateral gene transfer in eukaryotes. Cell Mol Life Sci. 62:1182–1197.
Andersson JO. 2011. Evolution of patchily distributed proteins shared between eukaryotes and prokaryotes: Dictyostelium as a case study. J Mol Microbiol Biotechnol. 20:83–95.
Andersson JO, Sjögren AM, Davis LAM, Embley TM, Roger AJ. 2003. Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. Curr Biol. 13:94–104.
Berg JS, Powell BC, Cheney RE. 2001. A millennial myosin census. Mol Biol Cell. 12:780–794.
Bloemink MJ, Geeves MA. 2011. Shaking the myosin family tree: biochemical kinetics defines four types of myosin motor. Semin Cell Dev Biol. 22:961–967.
Bohil AB, Robertson BW, Cheney RE. 2006. Myosin-X is a molecular motor that functions in filopodia formation. Proc Natl Acad Sci U S A. 103: 12411.
Breshears LM, Wessels D, Soll DR, Titus MA. 2010. An unconventional myosin required for cell polarization and chemotaxis. Proc Natl Acad Sci U S A. 107:6918–6923.
Burki F, Okamoto N, Pombert J-F, Keeling PJ. 2012. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. Proc Biol Sci. 279:2246–2254.
Cheney RE, Riley MA, Mooseker MS. 1993. Phylogenetic analysis of the myosin superfamily. Cell Motil Cytoskeleton. 24:215–223.
Clark K, Langeslag M, Figdor CG, van Leeuwen FN. 2007. Myosin II and mechanotransduction: a balancing act. Trends Cell Biol. 17:178–186.
Cohen O, Gophna U, Pupko T. 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol. 28:1481–1489.
Derelle R, Lang BF. 2012. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. Mol Biol Evol. 29:1277–1289.
Dunn CW, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.
Dutilh BE, et al. 2007. Assessment of phylogenomic and orthology approaches for phylogenetic inference. Bioinformatics 23:815–824.
East DA, Mulvihill DP. 2011. Regulation and function of the fission yeast myosins. J Cell Sci. 124:1383–1390.
Foth BJ, Goedecke MC, Soldati D. 2006. New insights into myosin evolution and classification. Proc Natl Acad Sci U S A. 103:3681–3686.
Fritz-Laylin LK, et al. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell 140:631–642.
Gabaldón T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. Nat Rev Genet. 14:360–366.
Goodson HV, Spudich JA. 1993. Molecular evolution of the myosin family: relationships derived from comparisons of amino acid sequences. Proc Natl Acad Sci U S A. 90:659–663.
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 Biol Evol. 5:833–847.
Hampl V, et al. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc Natl Acad Sci U S A. 106:3859–3864.
Hartman MA, Finan D, Sivaramakrishnan S, Spudich JA. 2011. Principles of unconventional myosin function and targeting. Annu Rev Cell Dev Biol. 27:133–155.
Hartman MA, Spudich JA. 2012. The myosin superfamily at a glance. J Cell Sci. 125:1627–1632.
Henn A, De La Cruz EM. 2005. Vertebrate myosin VIIb is a high duty ratio motor adapted for generating and maintaining tension. J Biol Chem. 280:39665–39676.
Hodge T, Cope MJ. 2000. A myosin family tree. J Cell Sci. 113:3353–3354.
Hofmann WA, Richards Ta, de Lanerolle P. 2009. Ancient animal ancestry for nuclear myosin. J Cell Sci. 122:636–643.
House CH. 2009. The tree of life viewed through the contents of genomes. In: Gogarten MB, Gogarten JP, Olendzenski LC, editors. Methods in molecular biology, Vol. 532. Berlin: Springer. p. 141–61.
Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A. 96: 3801–3806.
James TY, Berbee ML. 2012. No jacket required—new fungal lineage defies dress code: recently described zoosporic fungi lack a cell wall during trophic phase. Bioessays 34:94–102.
Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33: 511–518.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 39:309–338.
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.
Laurin-Lemay S, Brinkmann H, Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. Curr Biol. 22:R593–R594.
Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol. 25:1307–1320.
Leonard G, Richards TA. 2012. Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. Proc Natl Acad Sci U S A. 109:21402–21407.
Li J-F, Nebenführ A. 2008. The tail that wags the dog: the globular tail domain defines the function of myosin V/XI. Traffic 116:290–298.
Liu BA, et al. 2011. The SH2 domain-containing proteins in 21 species establish the provenance and scope of phosphotyrosine signaling in eukaryotes. Sci Signal. 4:ra83.
Liu R, et al. 2008. Sisyphus, the *Drosophila* myosin XV homolog, traffics within filopodia transporting key sensory and adhesion cargos. Development 135:53–63.
Loubéry S, Coudrier E. 2008. Myosins in the secretory pathway: tethers or transporters? Cell Mol Life Sci. 65:2790–2800.
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. Proc Natl Acad Sci U S A. 105:9674–9679.
Marcet-Houben M, Gabaldón T. 2010. Acquisition of prokaryotic genes by fungal genomes. Trends Genet. 26:5–8.

Martin W, Hoffmeister M, Rotte C, Henze K. 2001. An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. Biol Chem. 382:1521–1539.
Massagué J, Seoane J, Wotton D. 2005. Smad transcription factors. Genes Dev. 19:2783–2810.
Matsumura F. 2005. Regulation of myosin II during cytokinesis in higher eukaryotes. Trends Cell Biol. 15:371–377.
Odronitz F, Kollmar M. 2007. Drawing the tree of eukaryotic life based on the analysis of 2,269 manually annotated myosins from 328 species. Genome Biol. 8:R196.
Peckham M. 2011. Coiled coils and SAH domains in cytoskeletal molecular motors. Biochem Soc Trans. 39:1142–1148.
Peremyslov VV, et al. 2011. Expression, splicing, and evolution of the myosin gene family in plants. Plant Physiol. 155:1191–1204.
Peyretaillade E, et al. 2011. Extreme reduction and compaction of microsporidian genomes. Res Microbiol. 162:598–606.
Pomberta J-F, et al. 2012. Gain and loss of multiple functionally related, horizontally transferred genes in the reduced genomes of two microsporidian parasites. Proc Natl Acad Sci U S A. 109:12638–12643.
Punta M, et al. 2012. The Pfam protein families database. Nucleic Acids Res. 40:D290–D301.
Richards TA, Cavalier-Smith T. 2005. Myosin domain evolution and the primary divergence of eukaryotes. Nature 436:1113–1118.
Richards TA, et al. 2011. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. Proc Natl Acad Sci U S A. 108:15258–15263.
Roberts R, et al. 2004. Myosin VI: cellular functions and motor properties. Philos Trans R Soc Lond B Biol Sci. 359:1931–1944.
Rodríguez-Ezpeleta N, et al. 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Curr Biol. 17: 1420–1425.
Roger AJ, Simpson AGB. 2009. Evolution: revisiting the root of the eukaryote tree. Curr Biol. 19:R165–R167.
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*. Mol Biol Evol. 28:1241–1254.
Shadwick JDL, Ruiz-Trillo I. 2012. A genomic survey shows that the haloarchaeal type tyrosyl tRNA synthetase is not a synapomorphy of opisthokonts. Eur J Protistol. 48:89–93.
Sierra R, et al. 2013. Deep relationships of Rhizaria revealed by phylogenomics: a farewell to Haeckel's Radiolaria. Mol Phylogenet Evol. 67: 53–59.
Sjölander K, Datta RS, Shen Y, Shoffner GM. 2011. Ortholog identification in the presence of domain architecture rearrangement. Brief Bioinform. 12:413–422.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science 297:89–91.
Steinmetz PRH, et al. 2012. Independent evolution of striated muscles in cnidarians and bilaterians. Nature 487:231–234.
Suga H, et al. 2012. Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. Sci Signal. 5:ra35–ra35.
Sweeney HL, Houdusse A. 2010. Myosin VI rewrites the rules for myosin motors. Cell 141:573–582.
Syamaladevi DP, Spudich JA, Sowdhamini R. 2012. Structural and functional insights on the Myosin superfamily. Bioinform Biol Insights. 6: 11–21.
Thompson RF, Langford GM. 2002. Myosin superfamily evolutionary history. Anat Rec. 268:276–289.
Torruella G, et al. 2012. Phylogenetic relationships within the opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. Mol Biol Evol. 29:531–544.
Vale RD. 2003. The molecular motor toolbox for intracellular transport. Cell 112:467–480.
Wickstead B, Gull K. 2007. Dyneins across eukaryotes: a comparative genomic analysis. Traffic 8:1708–1721.
Wickstead B, Gull K, Richards TA. 2010. Patterns of kinesin evolution reveal a complex ancestral eukaryote with a multifunctional cytoskeleton. BMC Evol Biol. 10:110.
Wideman AJG, Gawryluk RMR, Gray MW, Dacks JB. 2013. The ancient and widespread nature of the ER-mitochondria encounter structure. Mol Biol Evol. 30:2044–2049.
Williams SA, Gavin RH. 2005. Myosin genes in *Tetrahymena*. Cell Motil Cytoskeleton. 61:237–243.
Wolf YI, Koonin E V. 2013. Genome reduction as the dominant mode of evolution. Bioessays 35:829–837.
Wolf YI, Snir S, Koonin E V. 2013. Stability along with extreme variability in core genome evolution. Genome Biol Evol. 5:1393–1402.
Zhang H, et al. 2004. Myosin-X provides a motor-based link between integrins and the cytoskeleton. Nat Cell Biol. 6:523–531.
Zmasek CM, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. 12:R4.
Zmasek CM, Godzik A. 2012. This Déjà vu feeling—analysis of multidomain protein evolution in eukaryotic genomes. PLoS Comput Biol. 8: e1002701.

Associate editor: Geoff McFadden

3.3. Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi

Abstract – The Opisthokonta are a eukaryotic supergroup divided in two main lineages: animals and related protistan taxa, and fungi and their allies. There is a great diversity of lifestyles and morphologies among unicellular opisthokonts, from free-living phagotrophic flagellated bacterivores and filopodiated amoebas to cell-walled osmotrophic parasites and saprotrophs. However, these characteristics do not group into monophyletic assemblages, suggesting rampant convergent evolution within Opisthokonta. To test this hypothesis, we assembled a new phylogenomic dataset via sequencing 12 new strains of protists. Phylogenetic relationships among opisthokonts revealed independent origins of filopodiated amoebas in two lineages, one related to fungi and the other to animals. Moreover, we observed that specialized osmotrophic lifestyles evolved independently in fungi and protistan relatives of animals, indicating convergent evolution. We therefore analyzed the evolution of two key fungal characters in Opisthokonta, the flagellum and chitin synthases. Comparative analyses of the flagellar toolkit showed a previously unnoticed flagellar apparatus in two close relatives of animals, the filasterean *Ministeria vibrans* and *Corallochytrium limacisporum*. This implies that at least four different opisthokont lineages secondarily underwent flagellar simplification. Analysis of the evolutionary history of chitin synthases revealed significant expansions in both animals and fungi, and also in the Ichthyosporea and *C. limacisporum*, a group of cell-walled animal relatives. This indicates that the last opisthokont common ancestor had a complex toolkit of chitin synthases that was differentially retained in extant lineages. Thus, our data provide evidence for convergent evolution of specialized lifestyles in close relatives of animals and fungi from a generalist ancestor.

CellPress

Current Biology
Report

Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi

Guifré Torruella,[1,2,12] Alex de Mendoza,[1,2,12] Xavier Grau-Bové,[1,2] Meritxell Antó,[1] Mark A. Chaplin,[3] Javier del Campo,[1,4] Laura Eme,[5] Gregorio Pérez-Cordón,[6] Christopher M. Whipps,[7] Krista M. Nichols,[8,9] Richard Paley,[10] Andrew J. Roger,[5] Ariadna Sitjà-Bobadilla,[6] Stuart Donachie,[3] and Iñaki Ruiz-Trillo[1,2,11,*]

[1]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, Barcelona 08003, Catalonia, Spain
[2]Departament de Genètica, Universitat de Barcelona, Avinguda Diagonal 645, Barcelona 08028, Catalonia, Spain
[3]Department of Microbiology, University of Hawaii at Manoa, Snyder Hall, 2538 McCarthy Mall, Honolulu, HI 96822, USA
[4]Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[5]Department of Biochemistry and Molecular Biology, Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, NS B3H 4R2, Canada
[6]Institute of Aquaculture Torre de la Sal, IATS-CSIC, Ribera de Cabanes s/n, Castelló 12595, Spain
[7]Environmental and Forest Biology, State University of New York College of Environmental Science and Forestry (SUNY-ESF), Syracuse, NY 13210, USA
[8]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA
[9]Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Boulevard East, Seattle, WA 98112, USA
[10]Centre for Environment Fisheries and Aquaculture Science, Weymouth Laboratory, Barrack Road, The Nothe, Weymouth, Dorset DT4 8UB, UK
[11]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona 08010, Catalonia, Spain
[12]Co-first author
*Correspondence: inaki.ruiz@ibe.upf-csic.es
http://dx.doi.org/10.1016/j.cub.2015.07.053

SUMMARY

The Opisthokonta are a eukaryotic supergroup divided in two main lineages: animals and related protistan taxa, and fungi and their allies [1, 2]. There is a great diversity of lifestyles and morphologies among unicellular opisthokonts, from free-living phagotrophic flagellated bacterivores and filopodiated amoebas to cell-walled osmotrophic parasites and saprotrophs. However, these characteristics do not group into monophyletic assemblages, suggesting rampant convergent evolution within Opisthokonta. To test this hypothesis, we assembled a new phylogenomic dataset via sequencing 12 new strains of protists. Phylogenetic relationships among opisthokonts revealed independent origins of filopodiated amoebas in two lineages, one related to fungi and the other to animals. Moreover, we observed that specialized osmotrophic lifestyles evolved independently in fungi and protistan relatives of animals, indicating convergent evolution. We therefore analyzed the evolution of two key fungal characters in Opisthokonta, the flagellum and chitin synthases. Comparative analyses of the flagellar toolkit showed a previously unnoticed flagellar apparatus in two close relatives of animals, the filasterean *Ministeria vibrans* and *Corallochytrium limacisporum*. This implies that at least four different opisthokont lineages secondarily underwent flagellar simplification. Analysis of the evolutionary history of chitin synthases revealed significant expansions in both animals and fungi, and also in the Ichthyosporea and *C. limacisporum*, a group of cell-walled animal relatives. This indicates that the last opisthokont common ancestor had a complex toolkit of chitin synthases that was differentially retained in extant lineages. Thus, our data provide evidence for convergent evolution of specialized lifestyles in close relatives of animals and fungi from a generalist ancestor.

RESULTS AND DISCUSSION

Broad Taxonomic Sampling Provides New Phylogenetic Insights into the Evolution of the Opisthokonta

Previous attempts to solve opisthokont phylogeny swayed between species-rich datasets with poor deep-node resolution based on small ribosomal subunit [1–3] and multigene supermatrices that included few taxa [4–6]. To improve upon our previously published phylogenomic dataset [6], we therefore sampled representative species in all described opisthokont lineages (see Table S1 and Supplemental Experimental Procedures). This included representatives of nucleariids, choanoflagellates, filastereans, and the two main lineages of Ichthyosporea (Dermocystidia and Ichthyophonida). In addition, we included two different strains of the enigmatic *Corallochytrium limacisporum*, a spherical free-living walled saprotroph found in coral reefs [7]. Originally classified as a thraustochytrid based on its morphology,

C. limacisporum has been unstably placed within the Opisthokonta in all molecular phylogenies to date because of the scarce molecular data available [8–11]. In order to improve the opisthokont outgroup, we also sampled the ancyromonad *Nutomonas longa* CCAP 1958/5 [12], which is putatively related to Apusomonadida [11]. Overall, we generated new transcriptomic data for 10 protistan taxa (11 strains in total, highlighted in bold in Figure 1), plus new genomic data from another strain (*Ichthyophonus hoferi*). This represents the broadest taxon sampling to date to infer the opisthokont phylogeny.

To investigate the phylogenetic relationships, we assembled two datasets comprising a total of 93 single-copy protein domains: one with 83 taxa and 18,218 aligned amino acid positions (S83), and the other with 70 taxa and 22,313 amino acid positions (S70). The latter dataset was constructed to maximize alignment length and to minimize topological artifacts by excluding putative problematic taxa with long branches (e.g., Microsporidia, Excavata) and high percentages of missing data (e.g., taxa with only expressed sequence tag data) (see Table S1). Both datasets were consistent in recovering the backbone of the eukaryotic phylogeny using both Bayesian inference (BI) (Figures 1 and S1C) and maximum likelihood (ML) (Figures S1A and S1B; see Supplemental Experimental Procedures for details).

As sister groups to Opisthokonta, we recovered Apusomonadida and Breviatea as recently reported [13], branching as independent lineages and not forming a monophyletic group or clustering with amoebozoans. Interestingly, the topology of the S83 dataset placed *Nutomonas longa* (Ancyromonadida) branching closer to the Excavata and not closely related to the Apusomonadida and Opisthokonta. This contrasts with previous analyses [11, 12] but is consistent with recent results based on multiple markers [14]. Within the Holomycota (which includes fungi and their protistan relatives), we recovered a clade formed by *Nuclearia* sp. and *Fonticula alba* (Discicristoidea) as the earliest-branching lineage [15]. This was followed by *Rozella allomycis* and Microsporidia [16] and the paraphyletic assemblage of Chytridiomycota (including Neocallimastigomycota) and Blastocladiomycota [17]. Finally, within the Holozoa we recovered Filasterea as the sister group to the clade formed by the Metazoa and Choanoflagellatea, as previously reported [5, 6].

Interestingly, we recovered *C. limacisporum* as a sister group to Ichthyosporea (including the two major groups Ichthyophonida and Dermocystida) [18] with both ML and BI methods. The S83 dataset recovered this position for *C. limacisporum* with weak support (56% ML bootstrap support [bs] and 0.8 BI posterior probability [pp]). However, support for this branch increased significantly (bs = 80%, pp = 0.84) when the long-branch taxa were excluded (see Figure 1 and Table S2). The position of the dermocystid *Sphaerothecum destruens* as sister group to the rest of ichthyosporeans was only moderately supported (S83: bs = 60%, pp = 0.97; S70: bs = 61%, pp = 0.87) but was consistently recovered in all analyses. Thus, the monophyletic group comprising Ichthyosporea and *C. limacisporum* appears to be the earliest-branching lineage in the Holozoa. We tentatively name this novel group "Teretosporea," meaning "rounded spores," through this study.

C. limacisporum is the only known free-living osmotroph in the Holozoa, whereas the ichthyosporeans thus far described are known to be associated with animal hosts as parasites or commensals [18], despite being frequently found in environmental surveys [3]. The life cycles of *C. limacisporum* and Ichthyosporeans [7, 18] are strikingly similar: both start as a single cell that grows as a coenocyte until it reaches maturation, when it undergoes schizogony. The dispersive amoeboid or flagellated progeny (merozoites) settle and close the cycle [18]. Chytrid fungi show a similar developmental mode, with both coenocytic growth and amoeboid or flagellated stages [19]. Similarly, fungi also evolved from phagotrophic ancestors (Discicristoidea, *Rozella*, and Aphelida [20]) to become saprotrophs and parasites. Moreover, some Ichthyosporea species (*A. parasiticum* and *I. hoferi*) present a mode of polar growth that clearly resembles fungal hyphae [21]. Thus, teretosporeans and fungi present tantalizing similarities regarding life style adaptations and morphologies.

The resulting opisthokont tree also confirms the convergent evolution of filose amoebas, Filasterea within the Holozoa and Discicristoidea within the Holomycota. Both lineages have evolved a similar cell morphology comprising long, actin-based filopodia [22], with some taxa going through an aggregative multicellular cell stage in their life cycles [23].

Independent Loss of the Flagellum within the Opisthokonta

A single posterior motile flagellum is a defining character of opisthokonts [2]. Our observation that both filose amoebas and fungal-like lineages evolved in independent branches within opisthokonts therefore predicts independent loss of the flagellum. To address this hypothesis, we analyzed the evolution of the flagellar toolkit [24, 25]. The molecules that comprise the flagellum include specialized tubulins (*epsilon*, *delta*) [26], the intraflagellar transport system (i.e., the IFT-A, IFT-B, and BBSome complexes [27]), and some motor molecules, mainly specialized subfamilies of dyneins and kinesins [24, 28] (Figure 2B). Large-scale genomic analyses have shown that the presence of these genes in a given genome correlates with the presence of a flagellum—revealing, in some cases, a previously unseen flagellar stage [28].

To clarify the evolution of the flagellum, we sought orthologs of a set of over 60 flagellum-specific proteins [24, 27, 28] in our taxon sampling (see Supplemental Experimental Procedures and Table S3). As expected, non-flagellated lineages such as Dikarya fungi, Discicristoidea, Ichthyophonida, and the filasterean *Capsaspora owczarzaki* yielded no significant hits (Figure 2A). This confirmed the recurrent secondary loss of the flagellum in at least four opisthokont lineages. In contrast, we found several proteins corresponding to key flagellar molecular components in the transcriptome of two taxa assumed not to be flagellated, the filasterean *M. vibrans* and the teretosporean *C. limacisporum*.

M. vibrans was originally described as a filose amoeba suspended in the water column by a stalk attached to the substrate. The stalk resembled a modified flagellum based on transmission electron microscopy (TEM) observations, which included structures resembling, according to the authors, doublet microtubules [2]. Interestingly, we observed the presence of axonemal dyneins, *epsilon* tubulin, and IFT-A/B complexes, clearly suggesting the presence of a flagellum in this species. Therefore, we tested whether the stalk is a modified flagellum by tubulin immunostaining on the original ATCC 50519 strain (see Supplemental Experimental Procedures). Confocal microscopy

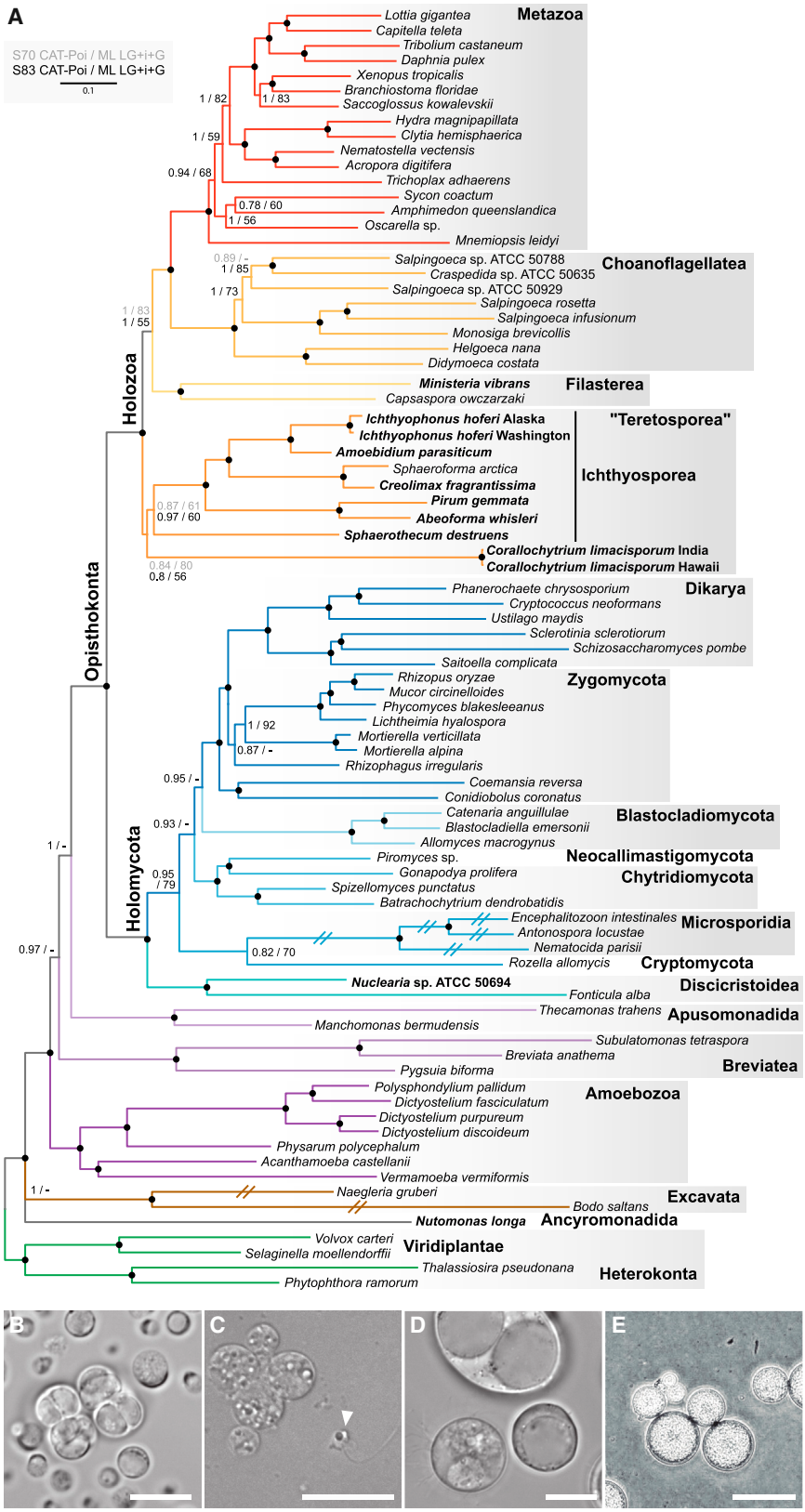


Figure 1. Phylogeny and Cell Biology of Opisthokonts

(A) Phylogenetic tree based on the 83-taxa matrix (see Tables S1 and S2 and Supplemental Experimental Procedures) and inferred by PhyloBayes under the CAT-Poisson model. Tree topology is the consensus of two Markov chain Monte Carlo chains run for 1,500 generations, saving every ten trees and after a burn-in of 25%. Split supports are posterior probabilities (pp) and nonparametric maximum likelihood (ML) bootstrap (bs) values obtained from 200 ML replicates using the LG+I+G model implemented in RAxML. Support values > 0.95 pp and > 95% bs are indicated by a bullet (•). The taxa sampled in this study are indicated in bold. For raw trees, see Figure S1.

(B–E) Light micrographs showing the coenocytic stage of representative species of the tentatively named "Terestosporea" (*Corallochytrium* + Ichthyosporea) sequenced in this study, including *Corallochytrium limacisporum* (B), *Sphaerothecum destruens* (C; arrowhead indicates flagellated zoospore), *Abeoforma whisleri* (D), and *Ichthyophonus hoferi* (E). Scale bar represents 10 μm in (B)–(D) and 100 μm in (E).

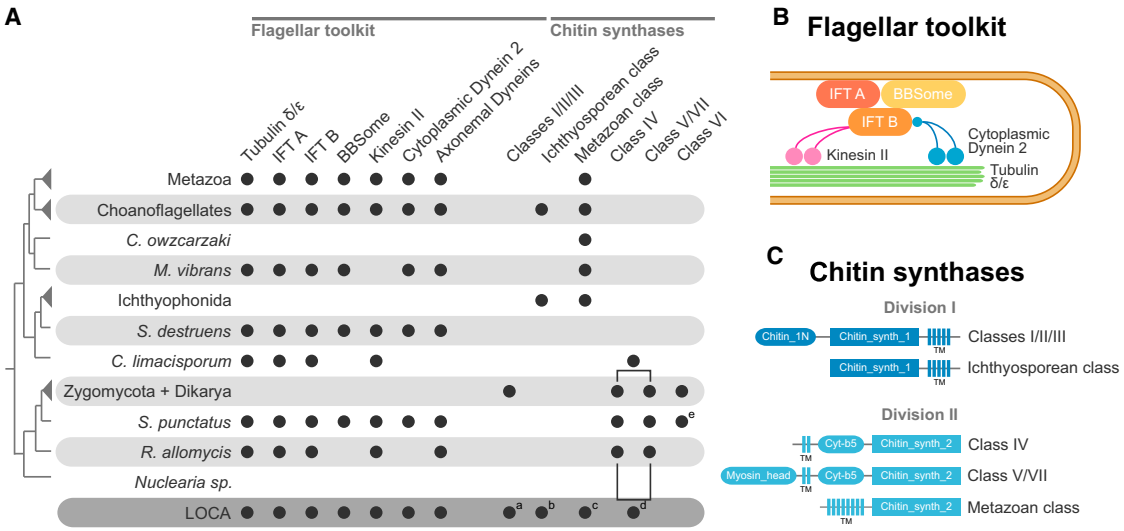


Figure 2. Multiple Independent Losses of the Flagellar Toolkit and CHS Genes in Opisthokonta

(A) Presence versus absence of key molecular components of the flagellar apparatus and chitin synthases (CHS) in distinct Opisthokonta lineages and taxa. [a]present in oomycetes, *Chlorella variabilis*, and *Paramecium tetraurelia*; [b]present in *Acanthamoeba castellanii*; [c]present in *Entamoeba histolytica* and *Thecamonas trahens*; [d]present in *Thalassiosira pseudonana*; [e]present in the chytrid *Batrachochytrium dendrobatidis*.

(B) Components of the flagellar apparatus and names of the molecular complexes. Adapted from [24]. See flagellar gene distribution in Table S3.

(C) Main chitin synthase classes and their canonical protein domain architectures (see CHS phylogeny in Figure S2).

revealed a tubulin protrusion branching from the cell body, which was specifically stained with α-tubulin (Figure 3A) and acetylated tubulin antibodies (Figures 3B and S3). Moreover, our own TEM observations revealed a putative dense basal body and a flagellar section with nine outer ring structures and central microtubules (Figure 3C). Our transcriptomic data and experimental analysis thus revealed a flagellar structure in *M. vibrans*. Consequently, the ancestral filasterean must have had a flagellum, which was secondarily lost from *C. owczarzaki*.

The transcriptome of *C. limacisporum* was found to contain *delta/epsilon* tubulins, IFT-A and IFT-B components, and the retrograde motor kinesin-II (Figure 2A). Although this organism does possess an ortholog of HEATR2 recently linked to motile cilia [29], we did not find evidence of flagellar motility components, such as cytoplasmic dynein 2 or any of the axonemal dyneins (heavy, light, and intermediate chains; Table S3). Consistent with the original description of *C. limacisporum* [7], we did not observe a flagellum using light and TEM microscopy, at least under the culturing conditions employed. Therefore, our data suggest that *C. limacisporum* has a cryptic flagellated stage in its life cycle, as has been inferred for other eukaryotes (i.e., *Aureococcus* and *Ostreococcus*) based on their genome sequences [28]. Consequently, within the Teretosporea, a flagellated stage would be a feature shared by *C. limacisporum* and Dermocystida that was secondarily lost from the Ichthyophonida (Figure 4). This confirms the recurrent loss of the flagellum in both filose amoeboid lineages (Discicristoidea and Filasterea) and specialized osmotrophic lineages (Fungi and Teretosporea).

At Least Four Chitin Synthases in the Last Opisthokonta Common Ancestor

Given the apparent similarities in the evolution of the Fungi and Teretosporea, we investigated the evolutionary history of another feature of fungal evolution, the cell wall. Chitin is a key biopolymer present in some fungal cell walls and animal cuticles [30], synthesized by chitin synthases (CHS), a large and complex multigene family. Several CHS classes have been described in fungi (classes I/II/III from division I and classes IV/V/VI/VII from division II) [31], with three ancestral classes known in animals [32]. Some fungal CHS classes are held as molecular synapomorphies of fungi (classes IV/V/VI/VII from division II), as they have been found exclusively in the genomes of fungi, including *R. allomycis* and microsporidian genomes [33]. Moreover, CHS homologs with uncertain classification have been found in other eukaryotes, including the oomycete *Saprolegnia monoica* [34], diatoms [35], and unicellular holozoans [18, 36].

To investigate which CHS classes are present in Teretosporea and to clarify their phylogenetic relationships with those in fungi and animals, we gathered CHS sequences from all eukaryotic supergroups and built a tree based on the chitin synthase domain (see Supplemental Experimental Procedures and Figure S2). This revealed three genes in *C. limacisporum* that belong to division II CHS and branch within the clade that comprises fungal classes IV/V/VII. These sequences consistently present the canonical functional motifs of fungal sequences (see Table S4). Interestingly, two of the genes encode an N-terminal myosin head domain, resembling genes from fungal classes V/VII [36] (Figure 2C). The myosin head of *C. limacisporum* CHS is sister group to fungal V/VII CHS, forming the myosin class XVII [37]. We thus propose that the CHS class IV/V/VII containing a myosin domain is an ancestral state in the Opisthokonta.

We also found that the Ichthyophonida contain CHS from both division I and division II clades. Ichthyophonida homologs from division I form a new clade with various eukaryotic sequences, including diatoms, choanoflagellates, and amoebozoans (Figures 2A and S2), revealing it also to be an ancestral class in

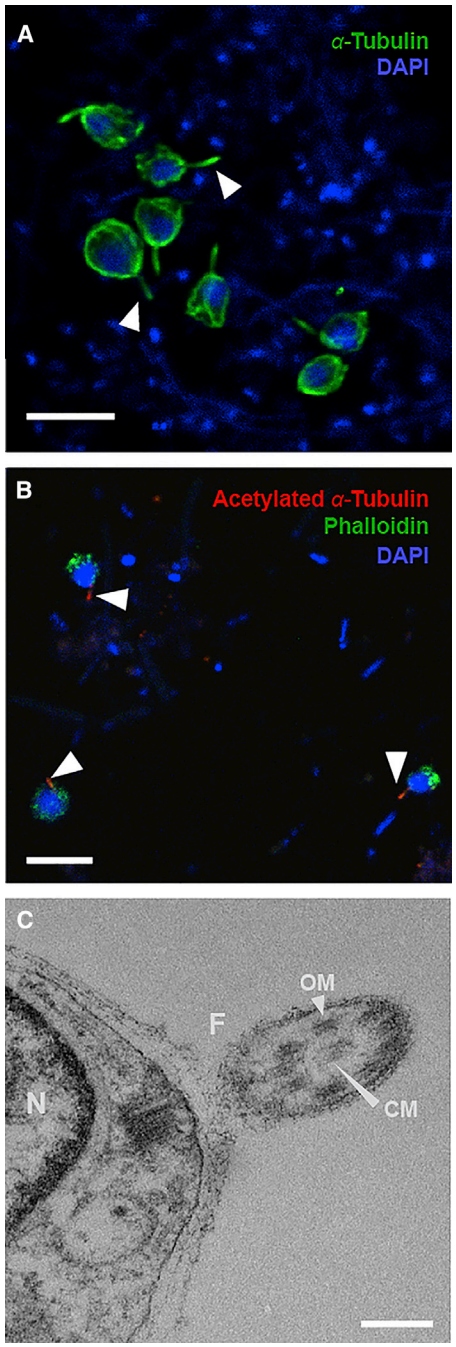


Figure 3. Confocal and Electron Microscopy of *Ministeria vibrans* Flagellum

(A and B) Confocal microscopy showing *Ministeria vibrans* ATCC 50519 stained with DAPI (blue) and anti-α-tubulin antibody 12G10 (Developmental Studies Hybridoma Bank) (green) (A) or with DAPI (blue), anti-acetylated-tubulin antibody T7451 (Sigma) (red), and phalloidin (green) (B). Arrowheads indicate the flagellar structure. Whereas the flagellar structure is specifically stained with cilia marker (acetylated tubulin) in (B), the cytoplasmic tubulin cytoskeleton is stained only with general anti-tubulin antibody in (A). *M. vibrans* feeds on bacteria, seen here as DAPI-stained bodies outside the cell. Scale bar represents 5 μm. See also Figure S3.

(C) TEM micrograph showing a transverse section of the flagellar structure of *M. vibrans*. N, nucleus; F, flagellar structure; OM, outer microtubules; CM, central microtubules. Scale bar represents 200 nm.

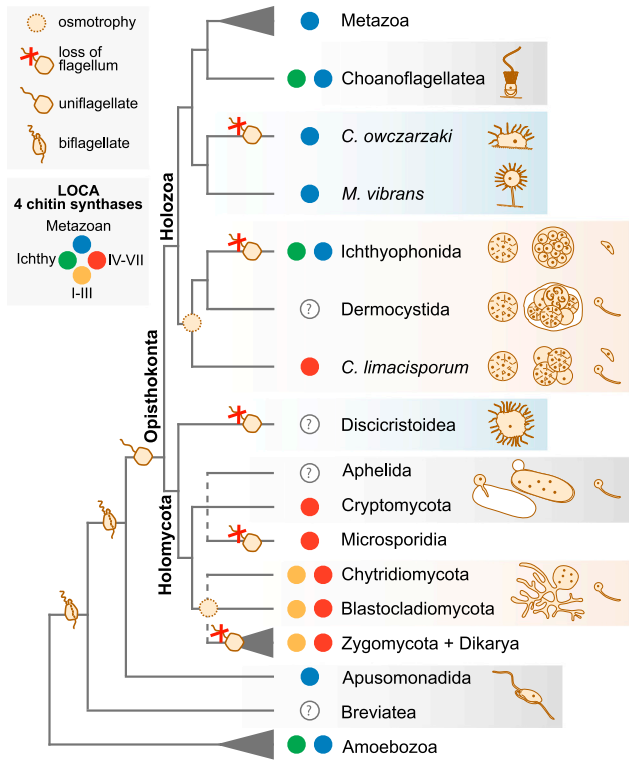


Figure 4. Evolution of Lifestyles and Some Cell Features of the Opisthokonts

Opisthokonta cladogram displaying lifestyle characteristics such as feeding mode, flagellated stage, CHS repertoire, and developmental mode (see Figure S4 for wheat germ agglutinin [WGA] staining) and ancestral state reconstruction of the last opisthokont common ancestor (LOCA). Choanoflagellate image is adapted from http://www.dayel.com/ (CC BY-SA 3.0).

the eukaryotes. Ichthyosporean division II CHS homologs belong to the Metazoan class, which is also present in other unicellular holozoans, apusomonads, and amoebozoans but is secondarily lost in fungi. Finally, fungal class I/II/III is found in several bikonts, including oomycetes and chlorophytes, suggesting an ancestral origin and secondary loss from the Holozoa. In summary, at least four ancestral paralogs of structurally different CHS (Figure 2C) were found in the last opisthokont common ancestor (LOCA), and secondary loss appears to have been common in descendant lineages (Figure 4). The presence of a complex CHS repertoire in the ancestor of all Opisthokonta, and the retention of rich CHS repertoires in the cell-walled lineages, suggests that the presence of chitin in the cell wall was an ancestral feature and not a fungal synapomorphy [33]. Consistent with this suggestion, Ichthyosporeans encoding a complex CHS repertoire showed chitin staining in the cell wall (Figure S4), and therefore only CHS VI class and the diversification of CHS IV/V/VII class into paralogous groups could be still considered fungal molecular synapomorphies.

A New Phylogenetic Framework for the Opisthokonts

By obtaining the transcriptomes of 10 new protist taxa (11 strains), plus the genome of an additional strain (12 strains in total), we have improved the previously biased representation

of genomic information for unicellular Opisthokonta. This allowed us to reassess the phylogenetic relationships among the opisthokonts through an unprecedented gene- and taxon-rich approach. Our dataset, with few missing data (Table S1), includes representatives from all opisthokont lineages, providing a stronger phylogenetic framework for internal relationships. Our phylogenetic analyses reveal a new clade: [Ichthyosporea + *C. limacisporum*], which we tentatively call Teretosporea, and which represents the earliest holozoan divergence (Figure 1).

Our data reveal that convergent evolution explains similarities in the lifestyles of the Fungi and Teretosporea as well as in Filasterea and Discicristoidea (Figure 4). The ancestral LOCA was most likely a filopodiated and flagellated generalist bacterivore [38]. Consequently, the specialized osmotrophic feeding mode, cell wall, and transition from saprotrophic to parasitic lifestyles in Fungi and Teretosporea occurred independently. This is not rare in eukaryotes, since similar adaptations are also found in stramenopiles such as the oomycetes and the thraustochytrids [39, 40]. However, our data provide the first example of such a process occurring in a close relative of animals. Through analysis of secondary loss of the flagellum and differential retention of ancestral CHS paralogs in opisthokonts, we have also provided molecular evidence to explain these lifestyle adaptations. Therefore, this study provides a striking example of convergent evolution through differential retention of ancestral genomic characters in the unicellular relatives of animals and fungi.

ACCESSION NUMBERS

The accession numbers for new data reported in this study are NCBI Sequence Read Archive: SRS502375, SRS502376, SRS721318, SRS725979, SRS725801, SRS726091, SRS724896, SRS725006, SRX179384; and NCBI BioProject: PRJNA290639.

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, four tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2015.07.053.

AUTHOR CONTRIBUTIONS

G.T., A.d.M., and I.R.-T. designed and coordinated the study. *C. limacisporum* isolation was performed by M.A.C. and S.D. *I. hoferi* genomic data was performed by K.M.N. and C.M.W. *S. destruens* RNA data was performed by R.P. Other strain cultivation and RNA extractions were performed by G.T., J.d.C., M.A., G.P.-C., and X.G.-B. Phylogenomics was performed by G.T., L.E., and A.J.R. Flagellum and CHS comparative genomics were performed by A.d.M. WGA staining was performed by A.d.M. and M.A. Immunostaining and TEM were performed by X.G.-B., M.A., A.S.-B., and G.T. Figures were assembled by X.G.-B., G.T., and A.d.M. G.T., A.d.M., and I.R.-T. wrote the manuscript. All authors commented on the manuscript.

ACKNOWLEDGMENTS

This work was supported by two grants (BFU2011-23434 and BFU2014-57779-P) from Ministerio de Economía y Competitividad (MINECO) and an ERC Starting Grant (ERC-2007-StG-206883) to I.R.-T. We also acknowledge support from Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (project 2014 SGR 619). G.T. was supported by a pre-graduate FI grant from the Catalan Government. A.d.M. and X.G.-B. were supported by a pre-graduate FPI grant from MINECO. L.E. was supported by a CGEB postdoctoral fellowship from the Tula Foundation, and A.J.R. was supported by a CIHR-NSHRF RPP grant (FRN# 62809). Assembly of the *I. hoferi* genome (USA) was supported by NSF grant number ACI-1053575. G.P.-C. was supported by a "Juan de la Cierva" grant, and A.S.-B. was supported by Generalitat Valenciana (PROMETEO FASE II-2014/085). We thank Dan Richter, Nicole King, Franz Lang, Matt Brown, Joseph Ryan, Andy Baxevanis, Ana Riesgo, Scott Nichols, Romain Derelle, Jordi Paps, Diego Mallo, Martin Kolisko, Txema Heredia, Philippe Lopez, Eric Bapteste, Arnau Sebé-Pedrós, Ana Franco-Sierra, and Jake L. Gregg for providing samples, data, technical assistance, and/or insightful comments.

Received: November 10, 2014
Revised: June 30, 2015
Accepted: July 22, 2015
Published: September 10, 2015

REFERENCES

1. Medina, M., Collins, A.G., Taylor, J.W., Valentine, J.W., Lipps, J.H., Amaral-Zettler, L., and Sogin, M.L. (2003). Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa. Int. J. Astrobiol. *2*, 203–211.
2. Cavalier-Smith, T., and Chao, E.E.-Y. (2003). Phylogeny of choanozoa, apusozoa, and other protozoa and early eukaryote megaevolution. J. Mol. Evol. *56*, 540–563.
3. del Campo, J., and Ruiz-Trillo, I. (2013). Environmental survey meta-analysis reveals hidden diversity among unicellular opisthokonts. Mol. Biol. Evol. *30*, 802–805.
4. Ruiz-Trillo, I., Roger, A.J., Burger, G., Gray, M.W., and Lang, B.F. (2008). A phylogenomic investigation into the origin of metazoa. Mol. Biol. Evol. *25*, 664–672.
5. Shalchian-Tabrizi, K., Minge, M.A., Espelund, M., Orr, R., Ruden, T., Jakobsen, K.S., and Cavalier-Smith, T. (2008). Multigene phylogeny of choanozoa and the origin of animals. PLoS ONE *3*, e2098.
6. Torruella, G., Derelle, R., Paps, J., Lang, B.F., Roger, A.J., Shalchian-Tabrizi, K., and Ruiz-Trillo, I. (2012). Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. Mol. Biol. Evol. *29*, 531–544.
7. 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. Bot. Mar. *30*, 83–89.
8. Ruiz-Trillo, I., Lane, C.E., Archibald, J.M., and Roger, A.J. (2006). Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. J. Eukaryot. Microbiol. *53*, 379–384.
9. Steenkamp, E.T., Wright, J., and Baldauf, S.L. (2006). The protistan origins of animals and fungi. Mol. Biol. Evol. *23*, 93–106.
10. Sumathi, J.C., Raghukumar, S., Kasbekar, D.P., and Raghukumar, C. (2006). Molecular evidence of fungal signatures in the marine protist Corallochytrium limacisporum and its implications in the evolution of animals and fungi. Protist *157*, 363–376.
11. Paps, J., Medina-Chacón, L.A., Marshall, W., Suga, H., and Ruiz-Trillo, I. (2013). Molecular phylogeny of unikonts: new insights into the position of apusomonads and ancyromonads and the internal relationships of opisthokonts. Protist *164*, 2–12.
12. Glücksman, E., Snell, E.A., and Cavalier-Smith, T. (2013). Phylogeny and evolution of Planomonadida (Sulcozoa): eight new species and new genera Fabomonas and Nutomonas. Eur. J. Protistol. *49*, 179–200.
13. Brown, M.W., Sharpe, S.C., Silberman, J.D., Heiss, A.A., Lang, B.F., Simpson, A.G., and Roger, A.J. (2013). Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. Proc. Biol. Sci. *280*, 20131755.
14. Cavalier-Smith, T., Chao, E.E., Snell, E.A., Berney, C., Fiore-Donno, A.M., and Lewis, R. (2014). Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and Amoebozoa. Mol. Phylogenet. Evol. *81*, 71–85.

15. Liu, Y., Steenkamp, E.T., Brinkmann, H., Forget, L., Philippe, H., and Lang, B.F. (2009). Phylogenomic analyses predict sistergroup relationship of nucleariids and fungi and paraphyly of zygomycetes with significant support. BMC Evol. Biol. *9*, 272.

16. James, T.Y., Letcher, P.M., Longcore, J.E., Mozley-Standridge, S.E., Porter, D., Powell, M.J., Griffith, G.W., and Vilgalys, R. (2006). A molecular phylogeny of the flagellated fungi (Chytridiomycota) and description of a new phylum (Blastocladiomycota). Mycologia *98*, 860–871.

17. James, T.Y., Kauff, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J., Celio, G., Gueidan, C., Fraker, E., Miądlikowska, J., et al. (2006). Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature *443*, 818–822.

18. Glockling, S.L., Marshall, W.L., and Gleason, F.H. (2013). Phylogenetic interpretations and ecological potentials of the Mesomycetozoea (Ichthyosporea). Fungal Ecol. *6*, 237–247.

19. Shelest, K., and Voigt, K. (2014). Genomics to study basal lineage fungal biology: Phylogenomics suggests a common origin. In Fungal Genomics, M. Nowrousian, ed. (Springer-Verlag), pp. 31–60.

20. Karpov, S.A., Mamkaeva, M.A., Aleoshin, V.V., Nassonova, E., Lilje, O., and Gleason, F.H. (2014). Morphology, phylogeny, and ecology of the aphelids (Aphelidea, Opisthokonta) and proposal for the new superphylum Opisthosporidia. Front. Microbiol. *5*, 112.

21. Gozlan, R.E., Marshall, W.L., Lilje, O., Jessop, C.N., Gleason, F.H., and Andreou, D. (2014). Current ecological understanding of fungal-like pathogens of fish: what lies beneath? Front. Microbiol. *5*, 62.

22. Sebé-Pedrós, A., Burkhardt, P., Sánchez-Pons, N., Fairclough, S.R., Lang, B.F., King, N., and Ruiz-Trillo, I. (2013). Insights into the origin of metazoan filopodia and microvilli. Mol. Biol. Evol. *30*, 2013–2023.

23. Sebé-Pedrós, A., Irimia, M., Del Campo, J., Parra-Acero, H., Russ, C., Nusbaum, C., Blencowe, B.J., and Ruiz-Trillo, I. (2013). Regulated aggregative multicellularity in a close unicellular relative of metazoa. eLife *2*, e01287.

24. Carvalho-Santos, Z., Azimzadeh, J., Pereira-Leal, J.B., and Bettencourt-Dias, M. (2011). Evolution: Tracing the origins of centrioles, cilia, and flagella. J. Cell Biol. *194*, 165–175.

25. Hodges, M.E., Scheumann, N., Wickstead, B., Langdale, J.A., and Gull, K. (2010). Reconstructing the evolutionary history of the centriole from protein components. J. Cell Sci. *123*, 1407–1413.

26. Findeisen, P., Mühlhausen, S., Dempewolf, S., Hertzog, J., Zietlow, A., Carlomagno, T., and Kollmar, M. (2014). Six subgroups and extensive recent duplications characterize the evolution of the eukaryotic tubulin protein family. Genome Biol. Evol. *6*, 2274–2288.

27. van Dam, T.J.P., Townsend, M.J., Turk, M., Schlessinger, A., Sali, A., Field, M.C., and Huynen, M.A. (2013). Evolution of modular intraflagellar transport from a coatomer-like progenitor. Proc. Natl. Acad. Sci. USA *110*, 6943–6948.

28. Wickstead, B., and Gull, K. (2012). Evolutionary biology of dyneins. In Dyneins, S. King, ed. (Elsevier), pp. 89–121.

29. Diggle, C.P., Moore, D.J., Mali, G., zur Lage, P., Ait-Lounis, A., Schmidts, M., Shoemark, A., Garcia Munoz, A., Halachev, M.R., Gautier, P., et al. (2014). *HEATR2* plays a conserved role in assembly of the ciliary motile apparatus. PLoS Genet. *10*, e1004577.

30. Merzendorfer, H. (2011). The cellular basis of chitin synthesis in fungi and insects: common principles and differences. Eur. J. Cell Biol. *90*, 759–769.

31. Ruiz-Herrera, J., and Ortiz-Castellanos, L. (2010). Analysis of the phylogenetic relationships and evolution of the cell walls from yeasts and fungi. FEMS Yeast Res. *10*, 225–243.

32. Zakrzewski, A.-C., Weigert, A., Helm, C., Adamski, M., Adamska, M., Bleidorn, C., Raible, F., and Hausen, H. (2014). Early divergence, broad distribution, and high diversity of animal chitin synthases. Genome Biol. Evol. *6*, 316–325.

33. James, T.Y., Pelin, A., Bonen, L., Ahrendt, S., Sain, D., Corradi, N., and Stajich, J.E. (2013). Shared signatures of parasitism and phylogenomics unite Cryptomycota and microsporidia. Curr. Biol. *23*, 1548–1553.

34. Leal-Morales, C.A., Gay, L., Fèvre, M., and Bartnicki-García, S. (1997). The properties and localization of Saprolegnia monoica chitin synthase differ from those of other fungi. Microbiology *143*, 2473–2483.

35. Durkin, C.A., Mock, T., and Armbrust, E.V. (2009). Chitin in diatoms and its association with the cell wall. Eukaryot. Cell *8*, 1038–1050.

36. James, T.Y., and Berbee, M.L. (2012). No jacket required–new fungal lineage defies dress code: recently described zoosporic fungi lack a cell wall during trophic phase. BioEssays *34*, 94–102.

37. Sebé-Pedrós, A., Grau-Bové, X., Richards, T.A., and Ruiz-Trillo, I. (2014). Evolution and classification of myosins, a paneukaryotic whole-genome approach. Genome Biol. Evol. *6*, 290–305.

38. Cavalier-Smith, T. (2013). Early evolution of eukaryote feeding modes, cell structural diversity, and classification of the protozoan phyla Loukozoa, Sulcozoa, and Choanozoa. Eur. J. Protistol. *49*, 115–178.

39. Spanu, P., and Kämper, J. (2010). Genomics of biotrophy in fungi and oomycetes–emerging patterns. Curr. Opin. Plant Biol. *13*, 409–414.

40. Richards, T.A., and Talbot, N.J. (2013). Horizontal gene transfer in osmotrophs: playing with public goods. Nat. Rev. Microbiol. *11*, 720–727.

3.4. The eukaryotic ancestor had a complex ubiquitin signaling system of archaeal origin

Abstract – The origin of the eukaryotic cell is one of the most important transitions in the history of life. However, the emergence and early evolution of eukaryotes remains poorly understood. Recent data have shown that the last eukaryotic common ancestor (LECA) was much more complex than previously thought. The LECA already had the genetic machinery encoding the endomembrane apparatus, spliceosome, nuclear pore, and myosin and kinesin cytoskeletal motors. It is unclear, however, when the functional regulation of these cellular components evolved. Here, we address this question by analysing the origin and evolution of the ubiquitin signalling system, one of the most important regulatory layers in eukaryotes. We delineated the evolution of the whole ubiquitin, SUMO and Ufm1 signalling networks by analysing representatives from all major eukaryotic, bacterial and archaeal lineages. We found that the ubiquitin toolkit had a pre-eukaryotic origin and is present in three extant archaeal groups. The pre-eukaryotic ubiquitin toolkit greatly expanded during eukaryogenesis, through massive gene innovation and diversification of protein domain architectures. This resulted in a LECA with essentially all of the ubiquitin-related genes, including the SUMO and Ufm1 ubiquitin-like systems. Ubiquitin and SUMO signalling further expanded during eukaryotic evolution, especially labelling and de-labelling enzymes responsible for substrate selection. Additionally, we analysed protein domain architecture evolution and found that multicellular lineages have the most complex ubiquitin systems in terms of domain architectures. Together, we demonstrate that the ubiquitin system predates the origin of eukaryotes and that a burst of innovation during eukaryogenesis led to a LECA with complex post-translational regulation.

The Eukaryotic Ancestor Had a Complex Ubiquitin Signaling System of Archaeal Origin

Xavier Grau-Bové,[†,1] Arnau Sebé-Pedrós,[†,1] and Iñaki Ruiz-Trillo[*,1,2,3]

[1]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Spain

[2]Departament de Genètica, Universitat de Barcelona, Barcelona, Spain

[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: inaki.ruiz@ibe.upf-csic.es.

Associate editor: Lars Jermiin

Abstract

The origin of the eukaryotic cell is one of the most important transitions in the history of life. However, the emergence and early evolution of eukaryotes remains poorly understood. Recent data have shown that the last eukaryotic common ancestor (LECA) was much more complex than previously thought. The LECA already had the genetic machinery encoding the endomembrane apparatus, spliceosome, nuclear pore, and myosin and kinesin cytoskeletal motors. It is unclear, however, when the functional regulation of these cellular components evolved. Here, we address this question by analyzing the origin and evolution of the ubiquitin (Ub) signaling system, one of the most important regulatory layers in eukaryotes. We delineated the evolution of the whole Ub, Small-Ub-related MOdifier (SUMO), and Ub-fold modifier 1 (Ufm1) signaling networks by analyzing representatives from all major eukaryotic, bacterial, and archaeal lineages. We found that the Ub toolkit had a pre-eukaryotic origin and is present in three extant archaeal groups. The pre-eukaryotic Ub toolkit greatly expanded during eukaryogenesis, through massive gene innovation and diversification of protein domain architectures. This resulted in a LECA with essentially all of the Ub-related genes, including the SUMO and Ufm1 Ub-like systems. Ub and SUMO signaling further expanded during eukaryotic evolution, especially labeling and delabeling enzymes responsible for substrate selection. Additionally, we analyzed protein domain architecture evolution and found that multicellular lineages have the most complex Ub systems in terms of domain architectures. Together, we demonstrate that the Ub system predates the origin of eukaryotes and that a burst of innovation during eukaryogenesis led to a LECA with complex posttranslational regulation.

Key words: **ubiquitin, SUMO, Ufm1, post-translational signaling, multicellularity, eukaryogenesis, LECA, FECA.**

Article

Introduction

Of the three domains of life, eukaryotes have the most complex forms of cell organization. Understanding the emergence and early evolution of the eukaryotic cell is a major challenge for evolutionary biology. Recent findings have profoundly changed our long-held view of a simple last eukaryotic common ancestor (LECA) (Cavalier-Smith 1987, 1991), pointing instead to an ancestor that was already equipped with the machinery required for many of the cellular processes occurring in extant eukaryotes. These include, for instance, the cell division machinery (Makarova et al. 2010), the endomembrane apparatus (Brighouse et al. 2010), the spliceosome (Collins and Penny 2005), nuclear pores (Mans et al. 2004), a wide repertoire of transcription factors (de Mendoza et al. 2013), the RNA interference machinery (Shabalina and Koonin 2008), and cytoskeletal motors (Wickstead and Gull 2011; Sebé-Pedrós et al. 2014). It is unclear, however, whether the LECA already used tightly regulated signaling pathways to control these cellular processes.

We know that signaling systems are crucial in complex cells, as they provide the basis for finely tuned regulation of processes such as transcription (Aravind et al. 2006; Turjanski et al. 2007; Whitmarsh 2007), the cell cycle (Harashima et al. 2013), interactions with the milieu (Seger and Krebs 1995; Deshmukh et al. 2010; Suga et al. 2012), and localization of components within the cell (Field and Dacks 2009; Brighouse et al. 2010). Many of these functions rely on kinase activity and posttranslational protein modification, two signaling strategies of prokaryotic origin that gained importance at the origin of eukaryotes (Aravind et al. 2006). In eukaryotes, posttranslational protein modification by ubiquitin (Ub) constitutes a major source of proteome regulation (Hochstrasser 2009). Thus, understanding the evolution of Ub signaling can provide clues not only into how the LECA regulated its cellular processes but also into the role of signaling systems during the origin and early evolution of eukaryotes. Despite some evolutionary studies devoted to specific gene families (Gagne et al. 2002; Marín 2009a, 2009b, 2010a, 2010b, 2010c, 2013; Eme et al. 2011; Grau-Bové et al. 2013), however, a global picture of the evolution of Ub posttranslational signaling in eukaryotes is still missing.

Ubiquitination consists of the posttranslational modification of proteins by the covalent attachment of Ub, a

76-residue peptide (Hochstrasser 2000). Ub can be linked to proteins in various ways: Monoubiquitination (tagging a single Lys residue of the substrate), multiubiquitination (tagging multiple Lys), and polyubiquitination (Ub chain linked by isopeptide bonds between specific Lys residues) (Hochstrasser 2009). The type of ubiquitination regulates the function of the substrate. For example, poly-ubiquitinated proteins are typically degraded at the 26S proteasomal complex, whereas mono/multiubiquitinated proteins are involved in endocytosis, membrane trafficking, regulation of kinase signaling, DNA repair, and chromatin regulation (Mukhopadhyay and Riezman 2007). Ubiquitination involves a sequential enzymatic cascade: An activating enzyme (E1) for the label, a conjugating enzyme (E2), and a ligating enzyme (E3) that covalently binds the label to the target protein. Moreover, there are specific peptidases (deubiquitinases [DUB]) that reverse the action of E3 ligases (Hochstrasser 2000).

Since the discovery of Ub, other posttranslational signaling pathways, collectively known as Ub-like systems, have been characterized. These systems use different labeling peptides, which often do not have significant sequence similarity with Ub but nonetheless have the same tertiary structure (a β-grasp fold [Hochstrasser 2000]). Ub-like systems share a common enzymatic cascade structure, although most of the specific proteins involved differ between systems (van der Veen and Ploegh 2012). Small-Ub-related MOdifier (SUMO) and Ub-fold modifier 1 (Ufm1) are two of the most relevant Ub-like systems. The SUMO peptide is 100 residues long and shares approximately 18% sequence identity with Ub (Bayer et al. 1998). SUMO acts on a wide range of proteins from various organisms and is involved in ribosomal biogenesis and nuclear functions such as transcription, chromosome organization, DNA repair, or nuclear transport (Johnson 2004; Kerscher et al. 2006; Gareau and Lima 2010). Ufm1 has no significant sequence identity with Ub (Komatsu et al. 2004). It has a narrower range of possible substrates (Hochstrasser 2009) and is involved in the regulation of the endoplasmic reticulum activity and membrane transport, as well as animal development (Komatsu et al. 2004; Tatsumi et al. 2011).

The three systems share the same E1 and E2 enzymes, both of which belong to ancient protein families present in Eukaryota, Bacteria, and Archaea. The prokaryotic E1s and E2s are involved in other signaling systems and were co-opted into new functions with the emergence of the early Ub system (Iyer et al. 2006; Burroughs et al. 2008, 2009; Michelle et al. 2009). Unlike E1s and E2s, there are numerous protein families acting as E3 ligases. A first division can be drawn between HECT and RING protein families, with different and independently evolved catalytic mechanisms (Deshaies and Joazeiro 2009; Rotin and Kumar 2009). RINGs can be further classified into two canonical protein families (C3H2C3, defined by the zf-RING_2 domain, and C3HC4 RINGs, represented by the zf-C3HC4, zf-C3HC4_2, and zf-C3HC4_3 domains) and many unconventional ones (U-box, zf-RING_LisH, RINGv, FANCL, IBR/RBR, and Sina). There are also multiprotein complexes with E3 activity, known as Cullin-RING ligases (CRLs). CRLs are composed of a specific RING type (zf-rbx1), a Cullin subunit (structural backbone of the complex), and different adaptor and target recognition subunits (Cardozo and Pagano 2004; Willems et al. 2004; Petroski and Deshaies 2005; Stone et al. 2005; Deshaies and Joazeiro 2009).

The ligase activity of E3s can be reversed by DUBs, isopeptidase enzymes that cleave Ub chains after the C-terminus of the peptide label (Amerik and Hochstrasser 2004). Some DUBs are specific to a particular kind of Ub linkage (usually Lys48 or Lys63) but most are unspecific and promiscuous (Komander et al. 2009). According to their catalytic mechanism, DUBs are divided into cysteine proteases (UCH, USP, OTU, and Josephin) and metalloproteases (JAB). Finally, the SUMO and Ufm1 systems employ specific E3 and peptidase protein families. There are two E3s (zf-MIZ, RINGs, and IR1-M) and three peptidases (ULP/SENP, WLM, and C97) in SUMO; and one E3 (DUF2042) and one peptidase (C78) in Ufm1.

In this work, we use comparative genomics to decipher the origin and evolution of three Ub-like systems: Ub itself, SUMO, and Ufm1. Our reconstruction shows that the ubiquitination toolkit of the LECA was as complex as that of most modern eukaryotes, in terms of diversity of gene families. Furthermore, various species of Archaea belonging to three different lineages (Euryarchaeota, Crenarchaeota, and Aigarchaeota) already had a minimal but complete ubiquitination toolkit. Thus, Ub signaling existed prior to the origin of eukaryotes and underwent a profound process of innovation during eukaryogenesis, resulting in a complex Ub system in the LECA. Analysis of the subsequent evolution of the Ub-like posttranslational systems in eukaryotes shows that E1 and E2 predate the LECA and underwent little innovation during early eukaryotic evolution, whereas most E3 families appeared concomitantly with eukaryotes and underwent multiple lineage-specific expansions and diversifications of protein domain architectures. We also describe two independent expansions of the Ub signaling system at the origins of multicellularity in animals and plants. Overall, we show that the complexity of the LECA involved the capacity to perform posttranslational regulation of different cell processes by Ub and Ub-like systems. This suggests that Ub signaling was key to the origin of eukaryotes and was later expanded in some specific, mostly multicellular, lineages.

Results

A Comparative Survey of the Ub System Reveals an Archaeal Origin and a Complex Toolkit in the LECA

To elucidate the origin and evolution of Ub-like systems, we first examined the presence and abundance of 40 protein families related to Ub, SUMO, and Ufm1 signaling in a broad range of eukaryotic genomes (see supplementary table S1, Supplementary Material online, and Materials and Methods). Specifically, we surveyed the generalist E1 and E2 enzyme protein families, 27 specific components of the Ub system (including the peptide label, E3s, and peptidases), 7 families related to SUMO, and 4 related to Ufm1 (see supplementary table S2, Supplementary Material online, and Materials and Methods). Our survey revealed that 38 of these 40 protein families are widespread among eukaryotic groups (fig. 1). We found that complete toolkits for Ub,

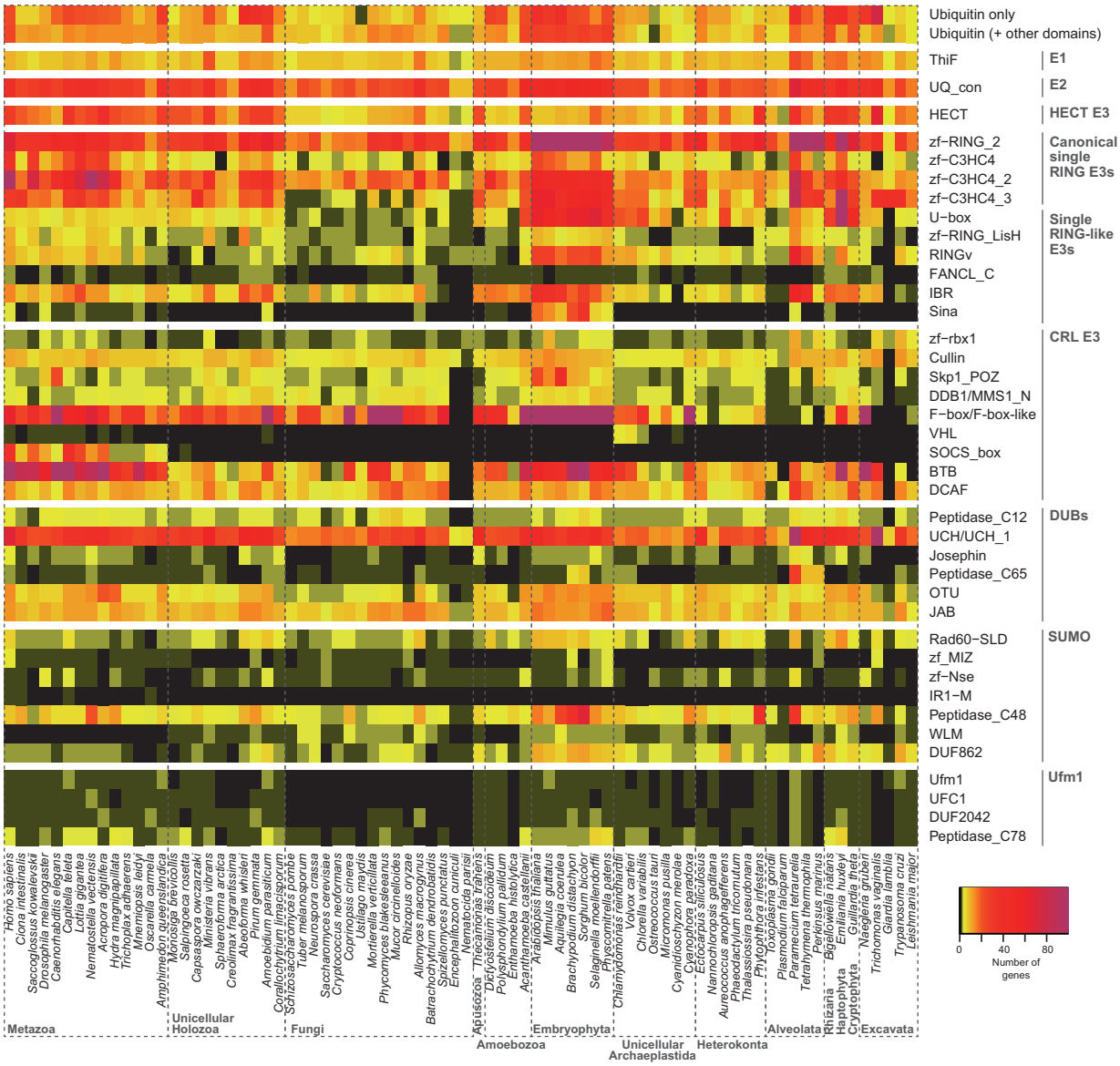


Fig. 1. Presence and abundance of the different components of the Ub, SUMO, and Ufm1 systems in eukaryotes. The heat map depicts absolute protein counts in each of the sampled genomes, according to the color scale. The Ub domain is divided into Ub-only (which includes Ub labels and poly-Ub peptides) and Ub + other domains (which includes proteins which make use of Ub domains for functions other than protein labeling).

SUMO, and Ufm1 systems exist in all the main groups of eukaryotes except for Fungi, in which Ufm1 is missing (see below). This phylogenetic distribution indicates that Ub, SUMO, and Ufm1 are ancient systems that were already present in the LECA (fig. 2).

To trace back the origin of the different signaling systems, we also examined a comprehensive database of prokaryotic genomes (see Materials and Methods). Although none of the analyzed bacterial genomes contained a complete Ub toolkit, many bacteria were found to possess signaling systems that employ JAB peptidases, and E1 and E2 enzymes akin to the ones acting in ubiquitination (Iyer et al. 2006; Hochstrasser 2009; Humbard et al. 2010). These bacterial homologs act in functional contexts unrelated to protein labeling, such as molybdopterin and thyamin biosynthesis (ThiF E1) and siderophora biosynthesis (JAB) (Iyer et al. 2006; Koonin 2006). We also found F-box, U-box, and DUB enzymes in a few genomes of obligate intracellular parasitic bacteria, such as *Agrobacterium tumefaciens*, *Legionella pneumophila*, *Candidatus* Amoebophilus asiaticus, or various Chlamydiae, probably as a result of independent horizontal gene transfer (HGT) events (Koonin et al. 2001; Spallek et al. 2009; Schmitz-Esser et al. 2010). Despite lacking Ub systems of their own, these pathogens exploit their hosts' by mimicking various signaling effectors (Spallek et al. 2009). Overall, the Ub-specific components analyzed clearly evolved after the origin of bacteria.

Unlike in bacteria, Ub-specific protein families were observed in many Archaea. Previous work by Nunoura et al. (2011) identified a bona fide eukaryotic-like Ub peptide and an E3 ligase in the Archaea *Caldiarchaeum subterraneum*. In our survey, we found evidence of eukaryotic-like Ub toolkits in three independent Archaea lineages: Crenarchaeota (including eight environmental genomes from the YNPFFA

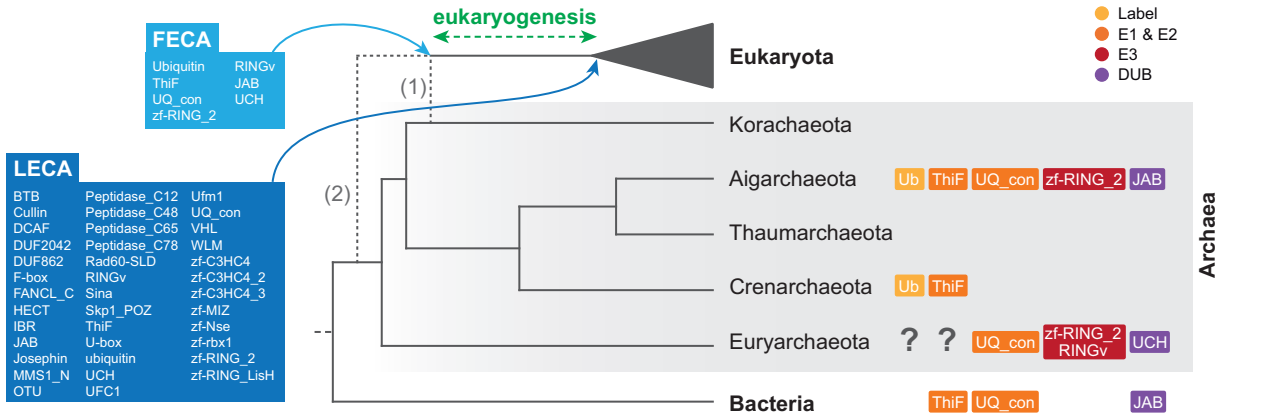


Fig. 2. Pre-eukaryotic evolution of Ub and Ub-like systems. The dashed lines indicate two possible phylogenetic scenarios: the Eocyte hypothesis for the origin of eukaryotes within Archaea (1) (Williams et al. 2012) and the "three domains" hypothesis for the relationships among Eukaryota, Bacteria and Archaea (2) (Woese et al. 1990). The reconstruction is the same with both hypotheses. The Ub, SUMO, and Ufm1 toolkits before and after the eukaryogenesis process (i.e., the FECA and the LECA) are shown. Boxes to the right of the cladogram represent the components of the Ub toolkit found in each archaeal group.

candidate group with Ub labels), Euryarchaeota (one environmental genome with a UCH DUB, C3H2C3s and a RINGv E3: marine group ii euryarchaeote SCGC AB-629-J06), and Aigarchaeota (11 environmental genomes from the pSL4 candidate group, seven of them with complete ubiquitination toolkits, and *C. subterraneum*, also with a complete toolkit) (fig. 2). Interestingly, the number of Ub-related genes in some of these genomes was found to be quite high, including nine C3H2C3 RING (zf_RING_2 domain) E3s in an aigarchaeote and up to six C3H2C3 RING plus a RINGv in the euryarchaeote. In addition, C3H2C3 RING genes have also been detected in two unclassified archaea (fig. 2 and supplementary file S1, Supplementary Material online).

To determine whether HGT of eukaryotic sequences into prokaryotic genomes could have occurred, we conducted Basic Local Alignment Search Tool (BLAST) similarity searches for all the protein families present in Archaea and phylogenetic analyses of Ub, UQ_con, and UCH (see Materials and Methods for details). None of the prokaryotic genes were found to be unexpectedly similar to eukaryotic sequences according to these methods. Thus, under the current taxon sampling, we can rule out a HGT origin for the archaeal toolkit (supplementary figs. S6 and S7 and file S3, Supplementary Material online).

In contrast, both SUMO and Ufm1 were found to be absent from Archaea and Bacteria. Thus, extant archaeal genomes contain a complete Ub toolkit that includes Ub label, E1 ThiF enzyme, E2 UQ_con enzyme, two different E3 ligases (C3H2C3 RING and RINGv), and two different DUBs (JAB and USP) (fig. 2), whereas SUMO and Ufm1 are specific to eukaryotes.

Evolution of Ub Signaling in Eukaryotes: Massive Secondary Losses, Few Gains, and Expansion of Gene Families

To better understand the evolution of the Ub system in eukaryotes, we examined the counts of two generalist gene families (E1 and E2 enzymes) and 38 protein families that are specific to a particular Ub-like system (peptide labels, E3 ligases, and peptidases) (fig. 1). We then reconstructed the patterns of gains and losses of each Ub-like signaling toolkit across eukaryotes using information of the phylogenetic distribution of each protein family (fig. 3). Finally, we also checked for statistically significant gene enrichments and depletions between eukaryotic groups (fig. 3), that is, significant quantitative changes in the number of proteins of a particular family. In contrast, gains and losses are defined as zero-to-one or one-to-zero state changes.

Our analysis indicates that the LECA already had most of the surveyed gene families, independently of whether we root eukaryotes between unikonts/amorpheans and bikonts (Derelle and Lang 2012) or between excavates and the rest (He et al. 2014). In particular, under the modified "unikont-bikont" hypothesis for the root of eukaryotes (fig. 3), we identified only two gains: SOCS-box and IR1-M gene families (part of the Ub and SUMO E3 toolkits, respectively). Under the assumption of the "Excavata-first" hypothesis, the sole difference was the appearance of Sina E3s after the divergence of excavates (supplementary fig. S1*A*, Supplementary Material online). Finally, using likelihood-based gain/loss reconstruction (supplementary fig. S1*B* and *C*, Supplementary Material online), we obtained a similar result compared with the parsimony-based analysis (33 and 36 gene families in the LECA, respectively, under the "unikont-bikont" hypothesis for the root of eukaryotes). This shows that the recruitment of novel machinery in Ub-like systems is a relatively exceptional event during eukaryotic evolution, especially when compared with the frequent losses of individual system-specific gene families.

Among Ub-like signaling systems, we found that ubiquitination is the most gene-rich pathway in most of the examined eukaryotes, followed by SUMO and Ufm1 (supplementary fig. S2, Supplementary Material online). Indeed, the proportion of Ub-related genes can add up to approximately 5% in some plant genomes (Smalle and Vierstra 2004; Stone et al. 2005), making it one of the most expanded gene toolkits in several eukaryotes. In the

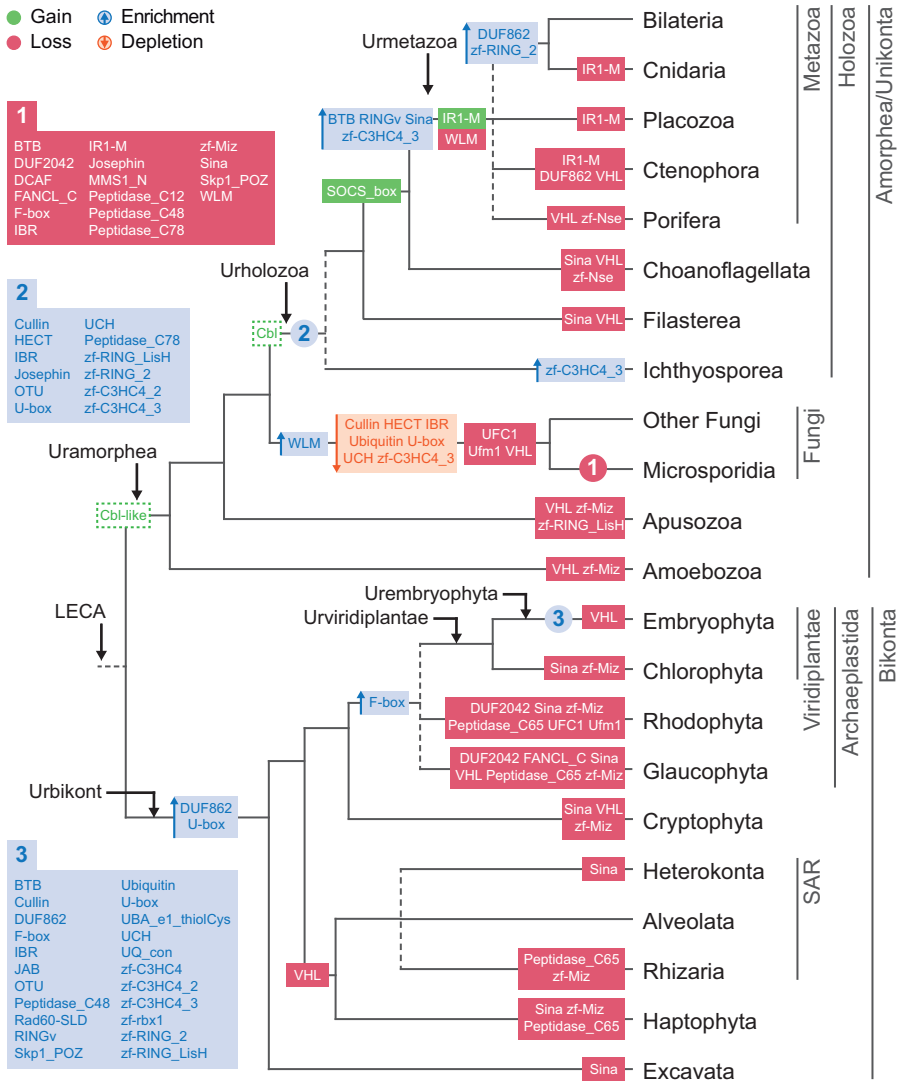


Fig. 3. Pattern of gains/losses and enrichments/depletions of the components of the Ub, SUMO, and Ufm1 systems in eukaryote evolution. The modified "unikont-bikont" hypothesis for the root of eukaryotes is assumed (Derelle and Lang 2012). See supplementary figure S1A, Supplementary Material online, for an alternative reconstruction with Excavata as the earliest branching lineage (He et al. 2014) and supplementary figure S1B and C, Supplementary Material online, for gain/loss reconstructions based on likelihood methods. Solid green and red boxes indicate gains and losses of gene families, respectively. Shaded blue and orange boxes indicate significant quantitative enrichments and depletions in the number of genes, respectively. Eukaryotic ancestors reconstructed in figure 5 are indicated by black arrows at the tree's nodes.

supplementary information, Supplementary Material online, we describe our findings for specific components of the system.

The Diversification of the Eukaryotic Ub System Is Driven by Architectural Rearrangements

To further analyze the diversification of Ub-like systems in eukaryotes, we used the array of domain architectures of each protein family as a proxy to assess the diversity and versatility of the Ub, SUMO, and Ufm1 toolkits. In particular, we compared the number of different protein domains that co-occur alongside the core protein domain of each protein family (see Materials and Methods). The most abundant families (e.g., canonical RINGs, F-box, BTB, DUBs, and deSUMOylases) are also the most diverse in terms of architectures (fig. 1 and supplementary fig. S4, Supplementary Material online), thereby implying a functionally diversifying gene expansion process.

To test whether there are phylogenetic patterns in the profiles of gene counts and architectural diversity of the Ub-like systems, we performed principal component analyses (PCA, see Materials and Methods for details) (fig. 4). The PCA based on gene counts revealed that embryophytes and metazoans have gene content profiles that differ from those of other eukaryotes (fig. 4A). In particular, we found that the principal component 1 identified a group of genomes rich in genes related to Ub-like signaling systems, including embryophytes, many animals (especially eumetazoans: *Homo sapiens*, *Capitella teleta*, or *Nematostella vectensis*) and ichthyosporeans (*Abeoforma whisleri*, *Pirum gemmata*, and *Amoebidium parasiticum*). Furthermore, PC2 differentiated most

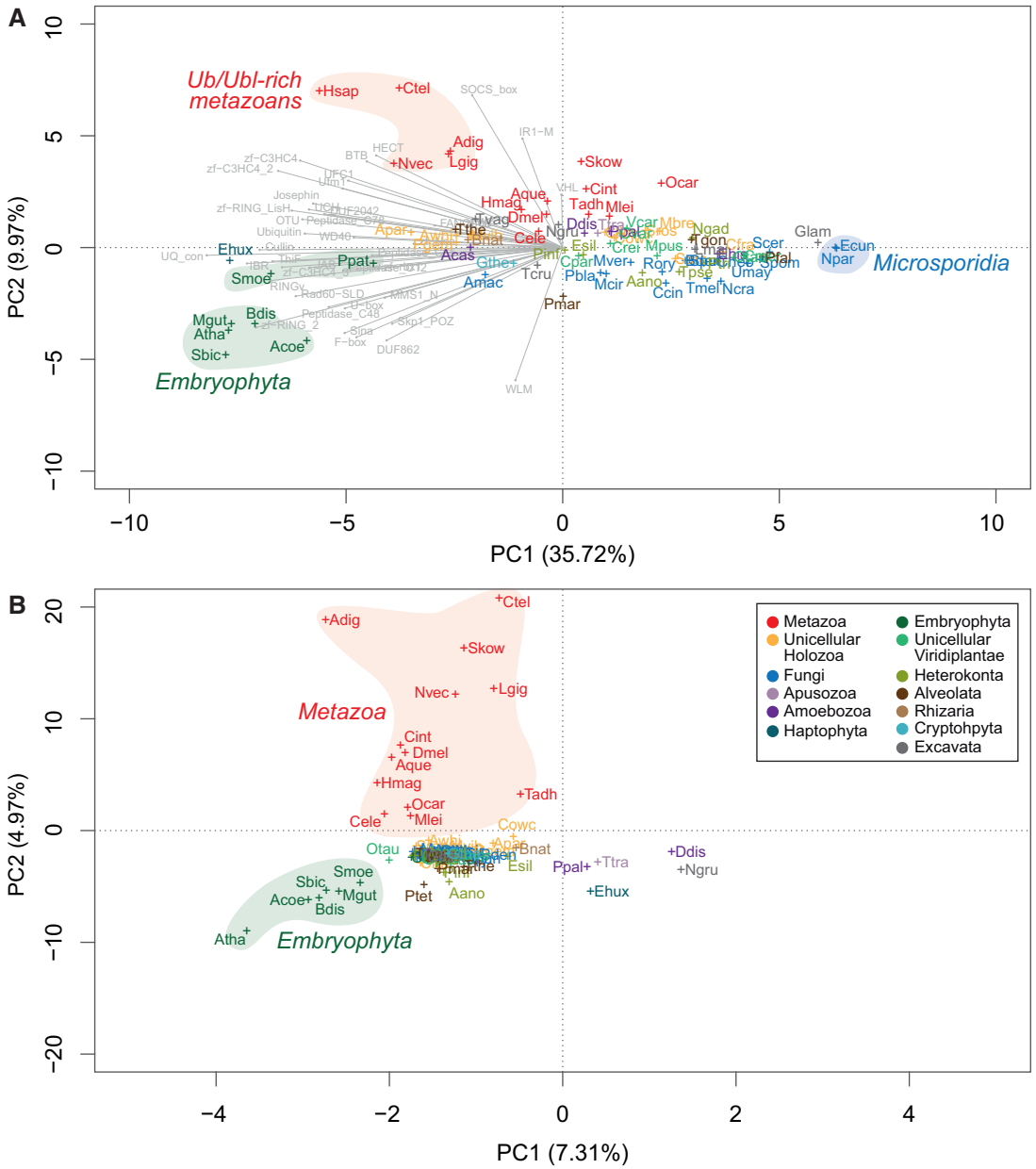


Fig. 4. Phylogenetic patterns in the composition of the Ub-like systems. (*A*) Clustering of 78 eukaryotic genomes in a multidimensional space for the gene count of the Ub, SUMO, and Ufm1 systems, using a PCA. The two first principal components are displayed, accounting for 33.97% and 10.09% of the variation, respectively. (*B*) Clustering of 78 eukaryotic genomes in a multidimensional space for the counts of the number of each different domain architectures in the Ub, SUMO, and Ufm1 systems, using a PCA. The two first principal components are displayed, accounting for 7.31% and 4.97% of the variation, respectively. Organisms names are color coded according to taxonomic assignment (see label). Shades indicate particular groups of genomes referred to in the main text. See supplementary table S1, Supplementary Material online, for a list of organism acronyms and Materials and Methods for details on the PCA analysis.

holozoans from embryophytes, which both clustered separately from the rest of the eukaryotes due to the loadings of many protein families that appeared or expanded in holozoans (e.g., HECT, BTB, SOCS-box, IR1-M, and C3HC4 RINGs) and plants (e.g., F-box, U-box, and C3H2C3 RINGs), respectively. The distinction between plants and holozoans (particularly animals) was also recovered by the PCA based on protein architectures (fig. 4*B*): Plants and animals, while sharing all the surveyed protein families, had specific sets of protein architectures that distinguished them from the rest of the eukaryotes.

The Ancestral Ub Toolkit Revealed by Domain Networks

To gain insight into the complexity of Ub-like signaling during eukaryotic evolution, we used the protein domain architectures of extant species to reconstruct ancestral domain

networks at various ancestral nodes of the eukaryotic tree (fig. 5) (see Materials and Methods). In particular, we inferred the network of accessory domains of genes related to Ub signaling in the urmetazoan, urholozoan, uramorphean, LECA, urembryophyte, urviridiplantae, and urbikont (fig. 5*A*–*G*, see fig. 3 for the phylogenetic positions of the reconstructed nodes).

We inferred that many Ub-related genes already employed multiple accessory protein domains (in black) in several Ub-related genes in the LECA (fig. 5*D*), although less than in most extant eukaryotes. For example, the LECA's Ub toolkit used highly promiscuous domains such as Ankyrin repeats (linked to C3HC4 RINGs), UBA (Ub-associated domain, linked to USPs and Ub), and LRR (linked to F-box). Architectural diversification during eukaryogenesis also led to specific domain combinations in E1 and E2 protein families, which use exclusive sets of accessory domains (e.g., E1s have UBA_e1_thiolCys, UBACT, and UBA_e1_C domains) and have little interconnection with other nodes. These E1 and E2 types are conserved in all the other ancestral nodes and characterize the eukaryotic Ub network. Also, the usage of multidomain proteins in the early eukaryote appeared as an important difference compared with archaeal systems, in which all genes encode single-domain proteins.

Since the origin of eukaryotes, the connectivity and network density of Ub and SUMO toolkits independently increased in Amorphea and Bikonta, although to a lesser extent in Bikonta. This led to rich signaling systems in multicellular animals and plants (fig. 5*A*–*C* and *E*–*G*), confirmed by the PCA based on domain architectures (fig. 4*B*). Nevertheless, we found that the network structure of the deep ancestors influenced later ancestors and extant organisms. For example, the urembryophyte's less extensively connected domain network could be traced back to the urbikont (fig. 5*E*–*G*). This phylogenetic inertia constrained the Ub and SUMO systems of plants, whose expansion was not accompanied by a significant increase of protein architectures. Conversely, the diversified toolkits of animals were recapitulated in the denser domain networks of the urmetazoan, the urholozoan, and the uramorphean (fig. 5*A*–*C*).

Despite these differences in network density, patterns common to all the ancestral networks emerged (fig. 5*A*–*C* and *E*–*G*). The most abundant catalytic machinery of Ub signaling employed a similar core of highly connected nodes in all the post-LECA ancestors. This included the C3HC4 variants (which shared most of their accessory domains and often co-occurred themselves), C3H2C3/zf-RING_2 (highly connected but not directly linked to other RINGs), IBR, or U-box. The CRL substrate recognition subunits BTB and F-box were both highly connected, particularly to protein-binding domains. In contrast, BTB and F-box shared few nodes, thus suggesting independent diversifications. For example, F-box often co-occurred with Kelch (in plants), LRR, and WD40, whereas BTB used Ankyrin, Kelch, BACK, and NPH3 (a signal-transducing motif that appears at the origin of plants).

Discussion

The Ancient Ub System and the Origin of Eukaryotes

Our data show that the core components of the eukaryote Ub system originated in Archaea and predate the process of eukaryogenesis that led to the LECA. In particular, the core Ub toolkit inferred from extant Archaea includes Ub, E1s, E2s, two different RING E3s, and two different DUBs (fig. 2). Interestingly, ubiquitination has been hypothesized to be a key mechanism for the symbiogenic origin of eukaryotes, during which it would be needed to act as a barrier against aberrant proteins resulting from the massive invasion of bacterial Group II introns into the host archaeal genome (Koonin 2006, 2011). Thus, our results are consistent with the presence of a complete Ub signaling toolkit in the theoretical proto-eukaryote, termed the first eukaryotic common ancestor (FECA) (Koonin 2011; Koumandou et al. 2013).

The initial toolkit was expanded during the stem phase of eukaryotic evolution with the addition of numerous new types of enzymes and an increase in the number of genes in some families (fig. 2). Similarly, the network of accessory domains of the LECA (fig. 5*D*) reveals that eukaryotic Ub-like systems switched to the use of multidomain protein families during their early evolution, whereas archaeal toolkits consist only of the catalytic protein domains. The presence of accessory domains within protein families reflect their ability to physically interact with other cellular components (Basu et al. 2008), which indicates that the rise of new protein families during eukaryogenesis was accompanied by an increasingly connected Ub domain architecture network. Interestingly, this increase in the LECA's regulatory potential was concomitant with the appearance of eukaryote-specific cellular functions regulated by ubiquitination, such as endocytosis, vesicle trafficking, and histone modification, as well as nuclei-specific DNA repair machinery. Altogether, we find that Ub signaling expanded in multiple ways as the first complex eukaryotes evolved.

Overall, our analyses indicate that the LECA had a rich and complex repertoire of Ub signaling genes, generating an extensive ancestral core machinery shared by most of the extant eukaryotic lineages. Given that some gene families were also secondarily, and recurrently, lost during eukaryotic evolution (fig. 3), our results suggest that there were two phases in the evolution of Ub signaling: 1) an initial period of rapid innovation during eukaryogenesis, in which the minimal FECA toolkit was enriched with new gene families exclusive to eukaryotes and 2) a long process of toolkit contraction (loss of gene families) in various eukaryotic lineages. These findings fit the biphasic model of reductive genome evolution proposed by Wolf and Koonin (2013) and strengthen the idea of eukaryogenesis as a burst of innovation in the history of life.

Diversification of Ub Signaling and the Origins of Multicellularity

Our data show that the core machineries of Ub, SUMO, and Ufm1 signaling were already present in the LECA (fig. 2). Subsequently, each eukaryotic group developed Ub-like

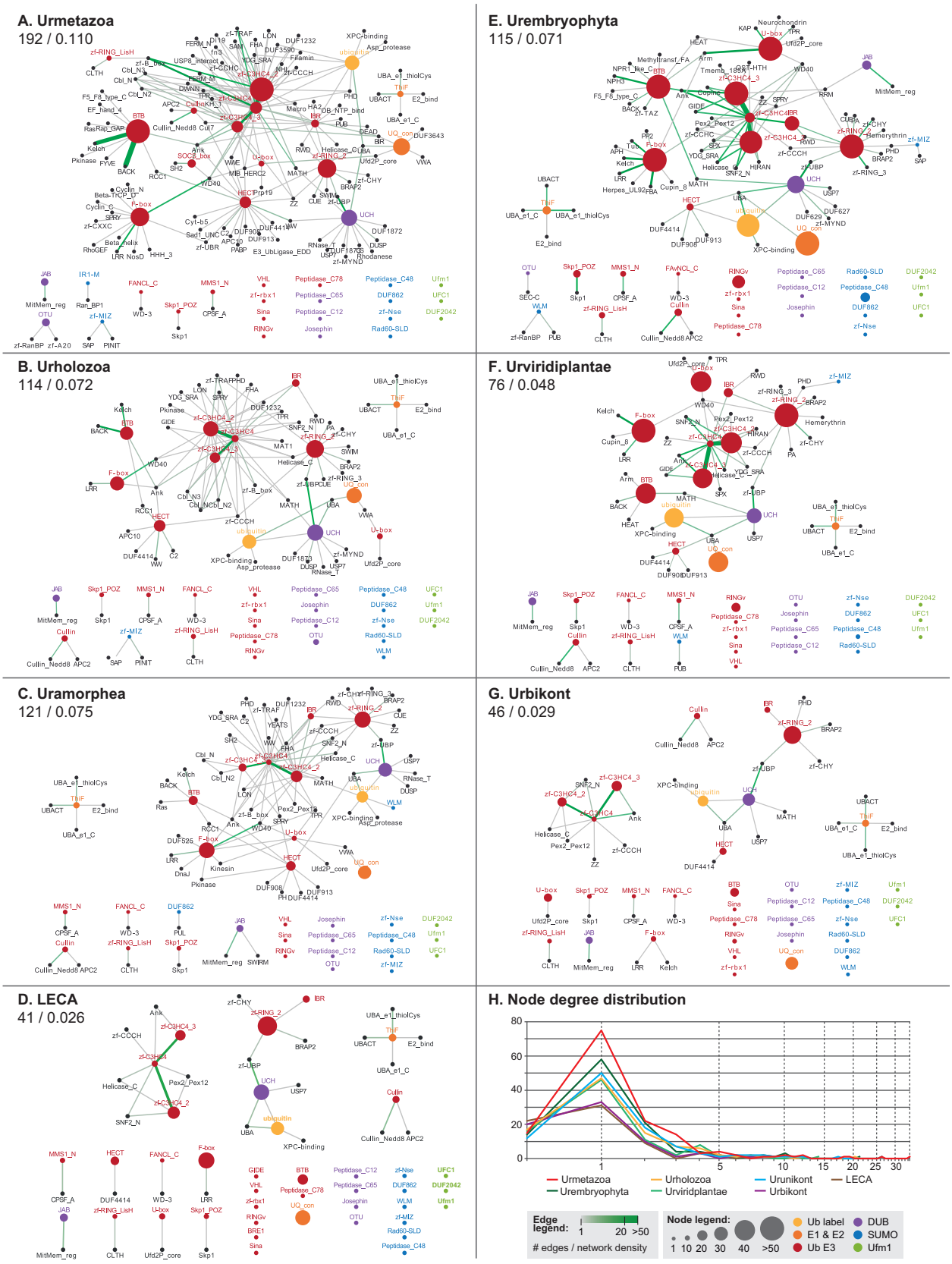


Fig. 5. Reconstruction of the ancestral networks of accessory domains of Ub, SUMO, and Ufm1 systems. The systems are reconstructed at the last common ancestors of (*A*) Metazoa, (*B*) Holozoa, (*C*) Amorphea, (*D*) Eukaryota, (*E*) Embryophyta, (*F*) Viridiplantae, and (*G*) Bikonta. Colored nodes represent core protein family domains, and black nodes represent their inferred accessory domains. The size of colored nodes is an estimation of the gene content of each ancestor. Edges link core with accessory domains and core domains between them and are color- and width coded according to the inferred number of such concurrences in each ancestor. For each network, the network density index, the number of edges, and the node degree distribution (*H*) are shown (see Materials and Methods). See figure 3 for the phylogenetic position of the reconstructed nodes.

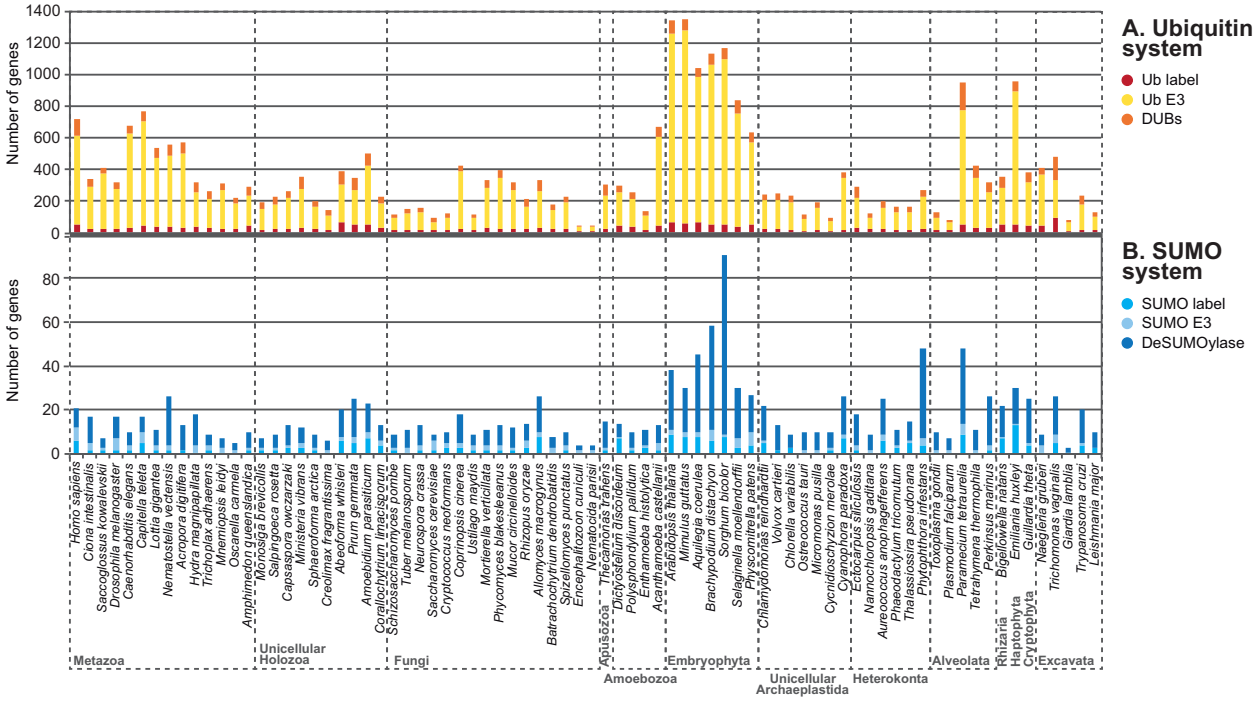


Fig. 6. Composition of Ub and SUMO toolkits. Number of Ub- and SUMO-related proteins (upper and lower charts, respectively), including the label itself, E3s, and peptidases. Note that specific deSUMOylases are consistently more abundant than SUMO E3s in most eukaryotes, whereas the opposite is true for Ub-related enzymes.

systems. This dynamic evolutionary history was mainly driven by lineage-specific gene expansions, architectural diversification of protein domains, occasional recruitment of new machinery, and abundant gene losses.

Gene expansions mostly affected E3 ligases and peptidases of Ub and SUMO toolkits, that is the effector enzymes responsible for substrate selection. Also, we found that the most enriched E3 and peptidase families often made use of promiscuous protein-binding domains, namely RINGs (canonical, IBR and U-box) and CRLs' substrate selector subunits (BTB and F-box), HECTs, and USPs. Likewise, HECTs are also rich in motifs that bind to lipids, complex sugars, and poly-A tails of RNA (Grau-Bové et al. 2013). The presence of such domains in the effector enzymes increases the substrate specificity and fine-tuned localization of Ub and SUMO (Tordai et al. 2005; Bhattacharyya et al. 2006; Di Roberto and Peisajovich 2013). Thus, the expansions of Ub and SUMO signaling brought an increased regulatory accuracy and functional diversification.

Our analysis also reveals that deSUMOylases are more abundant and diverse than SUMO E3s in most eukaryotes. The opposite pattern is found in ubiquitination, where Ub E3s outnumber DUBs (fig. 6). We therefore propose that two different strategies underlie the specificity of SUMO and Ub labeling in eukaryotes: SUMO relies on postlabeling regulation mediated by peptidases, whereas Ub depends on directed E3 activity. Consistent with this hypothesis, the expansion of SUMO peptidases in *Arabidopsis thaliana* entailed sub- and neofunctionalization events, whereas its E3s are often redundant (Chosed et al. 2006; Colby et al. 2006). In addition, humans, yeast, and *Ar. thaliana* can tune SUMOylation using a substrate-specific SUMO paralogs and paralog-specific peptidases (Saitoh and Hinchey 2000; Mukhopadhyay and Dasso 2007; Hickey et al. 2012). We also know that SUMO E2s can directly affect signaling in a nonspecific manner, without using E3s (Reverter and Lima 2005). We see how, from an identical pathway in the early eukaryote, different modes of posttranslational signaling regulation evolved for SUMO and Ub.

Comparing the two structural types of Ub E3s, we see that RING families are more abundant and architecturally diverse than HECTs in all eukaryotes (fig. 1 and supplementary fig. S4, Supplementary Material online). This might be explained by the fact that HECTs' tertiary structure is intrinsically constrained, as they require their catalytic site to be at the C-terminus to be active (Huang et al. 1999; Verdecia et al. 2003; Rotin and Kumar 2009). Consequently, they do not undergo C-terminal domain shuffling in any eukaryote (Grau-Bové et al. 2013). Also, the evolvability of RING-based catalysts was further increased by the emergence of CRLs, a combinatorial system of modular subunits with specific functions (e.g., interaction with E2s and substrates). Thus, historical and protein structural constraints explain the prevalence of RING-based catalysts in eukaryotes.

The greatest sophistication of Ub-like signaling systems is found in embryophytes and metazoans. These groups have the richest and most diverse Ub and SUMO systems among all eukaryotes (fig. 1). Moreover, the reconstruction of domain networks of ancestral Ub toolkits reveal that extensive innovation occurred at the origin of both animals and plants, probably through processes of domain shuffling that made use of already-in-place molecular machineries (fig. 5).

Although most of the surveyed protein families existed prior to the origins of animals and plants, we find that ubiquitination diversified extensively in these multicellular contexts through new domain combinations and gene number expansions (fig. 1 and supplementary fig. S4, Supplementary Material online). This may be due to the complex multicellularity of plants and animals, which requires fine-tuned regulation of cellular functions. Indeed, parallel to this complexification of posttranslational regulation, animals and plants are known to have a rich transcriptional regulation machinery, probably related to their complex development (de Mendoza et al. 2013).

Despite their similarities, the expansions of Ub-like signaling in multicellular animals and plants were independent: Each lineage expanded different protein families (fig. 4A) and diversified its toolkit with different accessory domains (fig. 5). This lack of protein architecture conservation among eukaryotes is common in other multidomain protein families (Basu et al. 2008, 2009). The rise and diversification of multidomain protein families by shuffling is also recurrent in animal genomes (Tordai et al. 2005) and is regarded as a key genomic event to explain the origin of multicellularity (King et al. 2008). Shuffling of ubiquitous and promiscuous domains is a major source of evolvability in eukaryotic signaling networks (Basu et al. 2008), as exemplified by tyrosine kinases (Deshmukh et al. 2010; Suga et al. 2012), Notch (King et al. 2008; Gazave et al. 2009), or Hedgehog toolkits (Snell et al. 2006; Adamska et al. 2007). Here, we identify independent bursts of innovation by domain shuffling underlying the complex Ub and SUMO systems of both animals and plants.

Conclusions

In summary, we found that Ub signaling predates the origin of eukaryotes, as core components of the pathway are present in three different archaeal groups: Aigarchaeota, Crenarchaeota, and Euryarchaeota. The Ub machinery of the earliest eukaryotes thus consisted of E1 and E2 enzymes (common to all three domains of life), two RING E3 types (canonical C3H2C3 and RINGv), and two peptidases (USP and JAB). This early Ub system underwent an important process of innovation during the eukaryogenic phase that led to the LECA.

We propose that three processes shaped Ub signaling during early eukaryotic evolution. First, almost all the Ub-related gene families seen in extant eukaryotes emerged at that time. This includes new catalytic mechanisms (e.g., HECTs and new peptidases) and, most importantly, two eukaryote-specific signaling systems (SUMO and Ufm1). Second, some gene families underwent massive expansions (e.g., RINGs and the highly versatile multisubunit CRLs). Finally, new and diverse protein domain architectures were acquired in both ancient and new enzyme families (e.g., E1s and CRLs' substrate selectors BTB and F-box). Altogether, these events identify the stem phase of eukaryotic evolution as a period of rapid and intense innovation in posttranslational signaling.

After the initial eukaryotic radiation, the Ub and Ub-like systems further evolved by protein family expansion and domain architectural diversification, in a largely lineage-specific manner. There was, however, little protein family innovation, with only IR1-M (animal SUMO E3s) and SOCS-box selectors (holozoan CRLs) evolving later on. These diversification processes particularly affected E3s ligases (in the case of the Ub system) and delabeling peptidases (in the case of the SUMO system) probably because they are in charge of the target selection specificity. In this sense, the diversification of domain architectures in these families is related to the substrate specificity, with new accompanying domains allowing selective interaction with other proteins, complex sugars, lipids or nucleic acids. This process of architectural innovation was especially intense at the origin of animals and plants, coinciding with their need for a precise regulation of multicellularity-related protein products and processes. Thus, alongside the eukaryogenic phase of Ub expansion, the origins of multicellular animals and plants represent the main bursts of innovation in Ub systems in eukaryotes.

Overall, our investigation into the diversity of early eukaryotic Ub signaling clearly points to an important burst of evolutionary innovation at the origin of eukaryotes. This suggests that the LECA was much more complex than previously thought, not only in terms of cellular machineries but also in terms of elaborate regulation systems such as Ub signaling.

Materials and Methods

We obtained all the proteins related to Ub, SUMO, and Ufm1 systems from a selection of 78 eukaryotic proteomes, the nonredundant Archaea and Bacteria protein database from National Center for Biotechnology Information (NCBI), and genomic data from the Microbial Dark Matter project (Rinke et al. 2013) (supplementary table S1, Supplementary Material online). The selection of eukaryotic taxa includes 14 animals, 10 unicellular holozoans, 16 fungi, 1 apusozoan, 4 amoebozoans, 7 embryophytes, 7 unicellular algae (chlorophytes, rhodophytes, and glaucophytes), 6 heterokonts/stramenopiles, 5 alveolates, 1 rhizarian, 1 haptophyte, 1 cryptophyte, and 5 excavates. We obtained the proteomes from publicly available databases, with the exception of *Oscarella carmela* and *Mnemiopsis leidyi*, kindly provided by Scott A. Nichols (University of Denver) and Andy Baxevanis (National Human Genome Research Institute), respectively. We also used RNA-Seq data generated in-house (*Ministeria vibrans*, *P. gemmata*, *Abeoforma whisleri*, *A. parasiticum*, and *Corallochytrium limacisporum*) (de Mendoza et al. 2013). We performed a Pfamscan on all eukaryotic proteomes and transcriptomes using Pfam A version 26 and selecting the gathering threshold as a conservative approach to minimize false positives (Punta et al. 2012). The identification of bacterial and archaeal sequences was done using HMMER (Eddy 1998), searching the hmm profiles of all the domains (supplementary table S2, Supplementary Material online) against the NCBI Bacteria and Archaea databases and the Microbial Dark Matter project database (Rinke et al. 2013).

We unambiguously assigned each protein of interest (including labeling peptides and E1, E2, E3, and delabeling enzymes) to a certain Pfam domain, referred to as the core defining domains of each protein family (see supplementary table S2, Supplementary Material online, for a complete list of

protein families, associated Pfam domains, and examples of specific genes in model organisms). The ThiF, zf-MIZ, and DCAF protein families were identified, refining the domain search with specific amino acid motifs. Specifically, proteins with ThiF and Moez/MoeB catalytic motifs do not have E1 activity and were discarded (Burroughs et al. 2009); zf-MIZ were selected by picking those architectures involving this domain combined with PINIT and/or SAP motifs; and DCAFs were identified by selecting proteins composed of WD40 domains and then retaining those that had a DWD motif (He et al. 2006; Hua and Vierstra 2011) with the following logo: [D | E] XXXX [I | L | V] [W | Y] [D] [I | L | V | M] [R | K].

Using R (R Development Core Team 2008), we built heat maps based on 1) the number of proteins involving a given core domain in each genome and 2) the number of accessory domains (i.e., total number of different domains that appear with a particular core domain in the same predicted ORF). Additional heat maps of the domain architectures in which each core domain is involved were built (supplementary fig. S5, Supplementary Material online). Statistical analyses were performed using R to detect enrichments or depletions in gene content in different lineages, using the Wilcoxon rank sum tests with a significance threshold of $P < 0.01$.

We used the BLAST (Camacho et al. 2009) to look for a potential HGT origin for the archaeal Ub, UQ_con, zf-RING_2, RINGv, and UCH proteins (supplementary fig. S7 and table S3, Supplementary Material online). We searched all the archaeal sequences (identified by HMMER searches, see above) with a cut-off value of 10^{-5} and against a combined database including the full NCBI nonredundant protein database, the Microbial Dark Matter database, and the full genomes and transcriptomes included in this study. We took the top 50 hits and searched them back to the same combined database, with a cut-off value of 10^{-10}. The network visualizations of this reciprocal BLAST analyses were generated using Cytoscape 3.1.1 (Smoot et al. 2011). We included the raw BLAST outputs in supplementary file S3, Supplementary Material online. Additionally, we performed phylogenetic analyses with UQ_con, UCH, and Ub families (zf-RING_2 and RINGv are not suitable for phylogenetic analysis because they are defined by short and poorly-conserved amino acid motifs). For these analyses, we used 1) all the Pfamscan-identified proteins from our selection of eukaryotes, 2) the identified archaeal sequences from NCBI and the Microbial Dark Matter databases, and 3) the top 100 hits from the BLAST searches in these databases. The alignments were performed using the Mafft L-INS-i algorithm, optimized for local sequence homology (Katoh and Standley 2013), and inspected and manually revised. We used the matched-pairs test of symmetry (Ababneh et al. 2006), implemented in Homo 1.2 for amino acids (http://www.csiro.au/Homo last accessed 1 October 2014), to determine whether the aligned sequences of amino acids are consistent with evolved under time-reversible conditions (assumed by most model-based phylogenetic programs). Based on the PP plots shown in supplementary figure S6A, Supplementary Material online, it was concluded that the data did not violate this assumption. The phylogenetic trees of UQ_con, UCH, and Ub were estimated using the Le and Gascuel (LG; 2008) evolutionary model with a discrete gamma (Γ) distribution of among-site variation rates (four categories), according to the respective analyses performed with ProtTest 3.4 (Darriba et al. 2011). The LG+Γ model with four categories was used in 1) maximum likelihood (ML) phylogenetic trees estimated with RaxML 7.2.8, using 100 bootstrap replicates as statistical support for the bipartitions (Stamatakis 2006) and 2) Bayesian inference trees calculated with PhyloBayes 3.3 (Lartillot et al. 2009), using two parallel runs for 500,000 generations and sampling every 100; and using Bayesian posterior probabilities as statistical support.

The reconstruction of ancestral states of each core element was inferred with Mesquite 2.75 using both a parsimony criterion and the AsymmMk likelihood model (http://mesquiteproject.org, last accessed 1 October 2014). We assumed two scenarios for the root of eukaryotes: 1) the modified "unikont-bikont" hypothesis (Derelle and Lang 2012) but renaming Unikonta as Amorphea (Adl et al. 2012) and 2) the "Discoba-first" hypothesis (He et al. 2014). For the relationships between Eukaryota, Bacteria, and Archaea, we contemplated both the "Eocyte" (eukaryotes root within Archaea) (Williams et al. 2012; Williams et al. 2013) and "three domains" hypotheses (Woese et al. 1990). The AsymmMk model was implemented with bias of 0.1 between gain and loss rates, with rates of change estimated by the model and taking into account branch lengths. To estimate the branch lengths, we built a multiprotein alignment with Hsp90, Hsp70, and actin homologs using Mafft L-INS-i (Katoh and Standley 2013), which was manually inspected. The matched-pairs test of symmetry performed using Homo showed that these sequences did not violate the time-reversibility assumption (supplementary fig. S1D, Supplementary Material online). In this case, ProtTest showed that the best evolutionary model for our data set was LG with a Γ distribution of four discrete categories and a proportion of invariable sites (LG+Γ+I). Using this model (PROTGMMAILG), we used RAxML with a fixed topology (consensus eukaryotic phylogeny, as in fig. 3 and supplementary fig. S1A, Supplementary Material online).

A PCA was performed using built-in R *prcomp* function, using scaling (so that all variables have unit variance before the analysis takes place) and a covariance matrix, and plotted using *bpca* R package. We used scaling because our data, although presenting the same units (counts of number of genes), show very different ranges of values (with some families having hundreds of genes and others just one or two). The PCA of the protein counts (fig. 4*A*) was based on the number of genes of each family in each species. In the PCA of protein domain architectures (fig. 4*B*), instead, the species were clustered based on the number of proteins with a particular domain architecture. To this end, we first created a list of all the existing protein domain architectures (for all protein families) and then counted how many proteins (with each particular architecture) each species has. These raw counts can be visualized in supplementary figure S5, Supplementary Material online.

Finally, we inferred the accessory protein domains of each protein family at ancestral nodes of the eukaryotic tree by comparing domain architectures (same raw data as for the PCA in fig. 4*B* and supplementary fig. S5, Supplementary Material online) within the corresponding clades. We represented these reconstructions as networks of co-occurring domains using Cytoscape 3.1.1 (Smoot et al. 2011). Our criterion linked core domains (central nodes, listed in supplementary table S2, Supplementary Material online) to accessory domains (other protein domains that co-occur with a core domain in the same protein) if such co-occurrence existed in at least the earliest-branching lineage of a clade and another internal taxon. We used a nested approach, first reconstructing the most external nodes and proceeding inward (e.g., first Bilateria, then Eumetazoa, followed by Metazoa, Holozoa, etc.). The abundance of each core domain (represented by the size of the node) at the reconstructed ancestors of particular clades was estimated with the median gene count of all the analyzed species in that clade (e.g., in the Urmetazoan in fig. 5*A*, the median of the counts of a particular core domain in all animals included in this study). The frequency of each domain co-occurrence (represented by the thickness of the edge between nodes) was estimated analogously. We calculated the network density index of each reconstructed ancestor using the Cytoscape NetworkAnalyzer module (Assenov et al. 2008).

Supplementary Material

Supplementary figures S1–S7, files S1–S3, and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a European Research Council Starting Grant (ERC-2007-StG-206883) and a grant (BFU2011-23434) from Ministerio de Economía y Competitividad (MINECO) to I.R.-T. X.G.-B. is supported by a pregraduate Formación del Personal Investigador grant from MINECO. The authors thank Andy Baxevanis (National Human Genome Research Institute) and Scott A. Nichols (University of Denver) for sharing unpublished protein sequences from *M. leidyi* and *O. carmela*, respectively. The authors also thank the reviewers for their thorough and much appreciated suggestions.

References

Ababneh F, Jermiin LS, Ma C, Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics* 22:1225–1231.

Adamska M, Matus DQ, Adamski M, Green K, Rokhsar DS, Martindale MQ, Degnan BM. 2007. The evolutionary origin of hedgehog proteins. *Curr Biol.* 17:R836–R837.

Adl SM, Simpson AG, Lane CE, Lukeš J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, et al. 2012. The revised classification of eukaryotes. *J Cell Biol.* 59:429–493.

Amerik AY, Hochstrasser M. 2004. Mechanism and function of deubiquitinating enzymes. *Biochim Biophys Acta.* 1695:189–207.

Aravind L, Iyer LM, Koonin EV. 2006. Comparative genomics and structural biology of the molecular innovations of eukaryotes. *Curr Opin Struct Biol.* 16:409–419.

Assenov Y, Ramírez F, Schelhorn S-E, Lengauer T, Albrecht M. 2008. Computing topological parameters of biological networks. *Bioinformatics* 24:282–284.

Basu MK, Carmel L, Rogozin IB, Koonin EV. 2008. Evolution of protein domain promiscuity in eukaryotes. *Genome Res.* 18:449–461.

Basu MK, Poliakov E, Rogozin IB. 2009. Domain mobility in proteins: functional and evolutionary implications. *Brief Bioinform.* 10:205–216.

Bayer P, Arndt A, Metzger S, Mahajan R, Melchior F, Jaenicke R, Becker J. 1998. Structure determination of the small ubiquitin-related modifier SUMO-1. *J Mol Biol.* 280:275–286.

Bhattacharyya RP, Reményi A, Yeh BJ, Lim WA. 2006. Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu Rev Biochem.* 75:655–680.

Brighouse A, Dacks JB, Field MC. 2010. Rab protein evolution and the history of the eukaryotic endomembrane system. *Cell Mol Life Sci.* 67:3449–3465.

Burroughs AM, Iyer LM, Aravind L. 2009. Natural history of the E1-like superfamily: implication for adenylation, sulfur transfer, and ubiquitin conjugation. *Proteins* 75:895–910.

Burroughs AM, Jaffee M, Iyer LM, Aravind L. 2008. Anatomy of the E2 ligase fold: implications for enzymology and evolution of ubiquitin/Ub-like protein conjugation. *J Struct Biol.* 162:205–218.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Cardozo T, Pagano M. 2004. The SCF ubiquitin ligase: insights into a molecular machine. *Nat Rev Mol Cell Biol.* 5:739–751.

Cavalier-Smith T. 1987. Eukaryotes with no mitochondria. *Nature* 326:332–333.

Cavalier-Smith T. 1991. Archamoebae: the ancestral eukaryotes? *Biosystems* 25:25–38.

Chosed R, Mukherjee S, Lois LM, Orth K. 2006. Evolution of a signalling system that incorporates both redundancy and diversity: *Arabidopsis* SUMOylation. *Biochem J.* 398:521–529.

Colby T, Matthäi A, Boeckelmann A, Stuible H-P. 2006. SUMO-conjugating and SUMO-deconjugating enzymes from *Arabidopsis*. *Plant Physiol.* 142:318–332.

Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol.* 22:1053–1066.

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

de Mendoza A, Sebé-Pedrós A, Sestak MS, Matejcic M, Torruella G, Domazet-Lošo T, Ruiz-Trillo I. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc Natl Acad Sci U S A.* 110:E4858–E4866.

Derelle R, Lang BF. 2012. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol Biol Evol.* 29:1277–1289.

Deshaies RJ, Joazeiro CA. 2009. RING domain E3 ubiquitin ligases. *Annu Rev Biochem.* 78:399–434.

Deshmukh K, Anamika K, Srinivasan N. 2010. Evolution of domain combinations in protein kinases and its implications for functional diversity. *Prog Biophys Mol Biol.* 102:1–15.

Di Roberto RB, Peisajovich SG. 2013. The role of domain shuffling in the evolution of signaling networks. *J Exp Zool B Mol Dev Evol.* 322:65–72.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* 14:755.

Eme L, Trilles A, Moreira D, Brochier-Armanet C. 2011. The phylogenomic analysis of the anaphase promoting complex and its targets points to complex and modern-like control of the cell cycle in the last common ancestor of eukaryotes. *BMC Evol Biol.* 11:265.

Field MC, Dacks JB. 2009. First and last ancestors: reconstructing evolution of the endomembrane system with ESCRTs, vesicle coat proteins, and nuclear pore complexes. *Curr Opin Cell Biol.* 21:4–13.

Gagne JM, Downes BP, Shiu S-H, Durski AM, Vierstra RD. 2002. The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in *Arabidopsis*. *Proc Natl Acad Sci U S A.* 99:11519–11524.

Gareau JR, Lima CD. 2010. The SUMO pathway: emerging mechanisms that shape specificity, conjugation and recognition. *Nat Rev Mol Cell Biol.* 11:861–871.

Gazave E, Lapébie P, Richards GS, Brunet F, Ereskovsky A V, Degnan BM, Borchiellini C, Vervoort M, Renard E. 2009. Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes. *BMC Evol Biol.* 9:249.

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 Biol Evol.* 5: 833–847.

Harashima H, Dissmeyer N, Schnittger A. 2013. Cell cycle control across the eukaryotic kingdom. *Trends Cell Biol.* 23:345–356.

He D, Fiz-Palacios O, Fu C-J, Fehling J, Tsai C-C, Baldauf SL. 2014. An alternative root for the eukaryote tree of life. *Curr Biol.* 24: 465–470.

He YJ, McCall CM, Hu J, Zeng Y, Xiong Y. 2006. DDB1 functions as a linker to recruit receptor WD40 proteins to CUL4-ROC1 ubiquitin ligases. *Genes Dev.* 20:2949–2954.

Hickey CM, Wilson NR, Hochstrasser M. 2012. Function and regulation of SUMO proteases. *Nat Rev Mol Cell Biol.* 13:755–766.

Hochstrasser M. 2000. Evolution and function of ubiquitin-like protein-conjugation systems. *Nat Cell Biol.* 2:E153–E157.

Hochstrasser M. 2009. Origin and function of ubiquitin-like proteins. *Nature* 458:422–429.

Hua Z, Vierstra RD. 2011. The cullin-RING ubiquitin-protein ligases. *Annu Rev Plant Biol.* 62:299–334.

Huang L, Kinnucan E, Wang G, Beaudenon S, Howley PM, Huibregtse JM, Pavletich NP. 1999. Structure of an E6AP-UbcH7 complex: insights into ubiquitination by the E2-E3 enzyme cascade. *Science* 286: 1321–1326.

Humbard MA, Miranda H V, Lim J-M, Krause DJ, Pritz JR, Zhou G, Chen S, Wells L, Maupin-Furlow JA. 2010. Ubiquitin-like small archaeal modifier proteins (SAMPs) in *Haloferax volcanii*. *Nature* 463:54–60.

Iyer LM, Burroughs AM, Aravind L. 2006. The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. *Genome Biol.* 7:R60.

Johnson ES. 2004. Protein modification by SUMO. *Annu Rev Biochem.* 73:355–382.

Katoh K., Standley D. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772–780.

Kerscher O, Felberbaum R, Hochstrasser M. 2006. Modification of proteins by ubiquitin and ubiquitin-like proteins. *Annu Rev Cell Dev Biol.* 22:159–180.

King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, et al. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451:783–788.

Komander D, Reyes-Turcu F, Licchesi JD, Odenwaelder P, Wilkinson KD, Barford D. 2009. Molecular discrimination of structurally equivalent Lys 63-linked and linear polyubiquitin chains. *EMBO Rep.* 10: 466–473.

Komatsu M, Chiba T, Tatsumi K, Iemura S, Tanida I, Okazaki N, Ueno T, Kominami E, Natsume T, Tanaka.. 2004. A novel protein-conjugating system for Ufm1, a ubiquitin-fold modifier. *EMBO J.* 23:1977–1986.

Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct.* 1:22.

Koonin EV. 2011. The logic of chance: the nature and origin of biological evolution. Upper Saddle River (NJ): Pearson Education.

Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol.* 55:709–742.

Koumandou VL, Wickstead B, Ginger ML, van der Giezen M, Dacks JB, Field MC. 2013. Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit Rev Biochem Mol Biol.* 48: 373–396.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol.* 25:1307–1320.

Makarova KS, Yutin N, Bell SD, Koonin E V. 2010. Evolution of diverse cell division and vesicle formation systems in Archaea. *Nat Rev Microbiol.* 8:731–741.

Mans BJ, Anantharaman V, Aravind L, Koonin EV. 2004. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. *Cell Cycle* 3:1625–1650.

Marín I. 2009a. RBR ubiquitin ligases: diversification and streamlining in animal lineages. *J Mol Evol.* 69:54–64.

Marín I. 2009b. Diversification of the cullin family. *BMC Evol Biol.* 9:267.

Marín I. 2010a. Animal HECT ubiquitin ligases: evolution and functional implications. *BMC Evol Biol.* 10:56–68.

Marín I. 2010b. Diversification and specialization of plant RBR ubiquitin ligases. *PLoS One* 5:e11579.

Marín I. 2010c. Ancient origin of animal U-box ubiquitin ligases. *BMC Evol Biol.* 10:331.

Marín I. 2013. Evolution of plant HECT ubiquitin ligases. *PLoS One* 8: e68536.

Michelle C, Vourc'h P, Mignon L, Andres CR. 2009. What was the set of ubiquitin and ubiquitin-like conjugating enzymes in the eukaryote common ancestor? *J Mol Evol.* 68:616–628.

Mukhopadhyay D, Dasso M. 2007. Modification in reverse: the SUMO proteases. *Trends Biochem Sci.* 32:286–295.

Mukhopadhyay D, Riezman H. 2007. Proteasome-independent functions of ubiquitin in endocytosis and signaling. *Science* 315:201–205.

Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H, Chee GJ, Hattori M, Kanai A, Atomi H, et al. 2011. Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res.* 39:3204–3223.

Petroski MD, Deshaies RJ. 2005. Function and regulation of cullin-RING ubiquitin ligases. *Nat Rev Mol Cell Biol.* 6:9–20.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. 2012. The Pfam protein families database. *Nucleic Acids Res.* 40:D290–D301.

R Development Core Team. 2008. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing [cited 2014 Oct 10]. Available from: http://www.R-project.org.

Reverter D, Lima CD. 2005. Insights into E3 ligase activity revealed by a SUMO-RanGAP1-Ubc9-Nup358 complex. *Nature* 435:687–692.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499:431–437.

Rotin D, Kumar S. 2009. Physiological functions of the HECT family of ubiquitin ligases. *Nat Rev Mol Cell Biol.* 10:398–409.

Saitoh H, Hinchey J. 2000. Functional heterogeneity of small ubiquitin-related protein modifiers SUMO-1 versus SUMO-2/3. *J Biol Chem.* 275:6252–6258.

Schmitz-Esser S, Tischler P, Arnold R, Montanaro J, Wagner M, Rattei T, Horn M. 2010. The genome of the amoeba symbiont "*Candidatus* Amoebophilus asiaticus" reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *J Bacteriol.* 192: 1045–1057.

Sebé-Pedrós A, Grau-Bové X, Richards TA, Ruiz-Trillo I. 2014. Evolution and classification of myosins, a paneukaryotic whole genome approach. *Genome Biol Evol.* 6:290–305.

Seger R, Krebs E. 1995. The MAPK signaling cascade. *FASEB J.* 9:726–735.

Shabalina SA, Koonin E V. 2008. Origins and evolution of eukaryotic RNA interference. *Trends Ecol Evol.* 23:578–587.

Smalle J, Vierstra RD. 2004. The ubiquitin 26S proteasome proteolytic pathway. *Annu Rev Plant Biol.* 55:555–590.

Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27:431–432.

Snell EA, Brooke NM, Taylor WR, Casane D, Philippe H, Holland PWH. 2006. An unusual choanoflagellate protein released by Hedgehog autocatalytic processing. *Proc R Soc B Biol Sci.* 273:401–407.

Spallek T, Robatzek S, Göhre V. 2009. How microbes utilize host ubiquitination. *Cell Microbiol.* 11:1425–1434.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Stone S, Hauksdóttir H, Troy A. 2005. Functional analysis of the RING-type ubiquitin ligase family of *Arabidopsis*. *Plant Physiol.* 137:13–30.

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. *Sci Signal.* 5:ra35–ra35.

Tatsumi K, Yamamoto-Mukai H, Shimizu R, Waguri S, Sou YS, Sakamoto A, Taya C, Shitara H, Hara T, Chung CH, et al. 2011. The Ufm1-activating enzyme Uba5 is indispensable for erythroid differentiation in mice. *Nat Commun.* 2:181.

Tordai H, Nagy A, Farkas K, Bányai L, Patthy L. 2005. Modules, multi-domain proteins and organismic complexity. *FEBS J.* 272:5064–5078.

Turjanski AG, Vaqué JP, Gutkind JS. 2007. MAP kinases and the control of nuclear events. *Oncogene* 26:3240–3253.

van der Veen AG, Ploegh HL. 2012. Ubiquitin-like proteins. *Annu Rev Biochem.* 81:323–357.

Verdecia MA, Joazeiro CA., Wells NJ, Ferrer J-L, Bowman ME, Hunter T, Noel JP. 2003. Conformational flexibility underlies ubiquitin ligation mediated by the WWP1 HECT domain E3 ligase. *Mol Cell.* 11: 249–259.

Whitmarsh AJ. 2007. Regulation of gene transcription by mitogen-activated protein kinase signaling pathways. *Biochim Biophys Acta.* 1773: 1285–1298.

Wickstead B, Gull K. 2011. The evolution of the cytoskeleton. *J Cell Biol.* 194:513–525.

Willems AR, Schwab M, Tyers M. 2004. A hitchhiker's guide to the cullin ubiquitin ligases: SCF and its kin. *Biochim Biophys Acta.* 1695: 133–170.

Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504: 231–236.

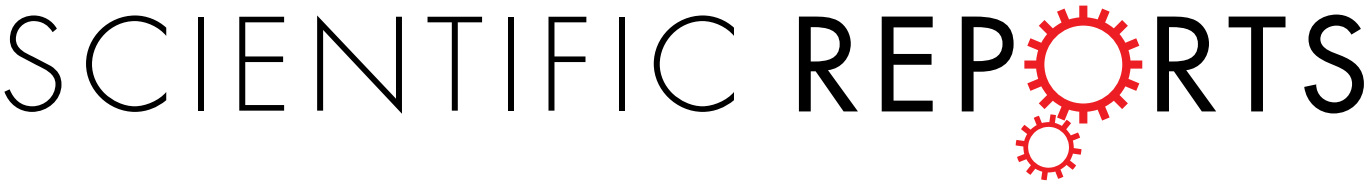
Williams TA, Foster PG, Nye TM, Cox CJ, Embley TM. 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc R Soc B Biol Sci.* 279:4870–4879.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A.* 87: 4576–4579.

Wolf YI, Koonin EV. 2013. Genome reduction as the dominant mode of evolution. *Bioessays* 35:1521–1878.

3.5. Origin and evolution of lysyl oxidases

Abstract – Lysyl oxidases (LOX) are copper-dependent enzymes that oxidize primary amine substrates to reactive aldehydes. The best-studied role of LOX enzymes is the remodeling of the extracellular matrix (ECM) in animals by cross-linking collagens and elastin, although intracellular functions have been reported as well. Five different LOX enzymes have been identified in mammals, LOX and LOX-like (LOXL) 1 to 4, showing a highly conserved catalytic carboxy terminal domain and more divergence in the rest of the sequence. Here we have surveyed a wide selection of genomes in order to infer the evolutionary history of LOX. We identified LOX proteins not only in animals, but also in many other eukaryotes, as well as in bacteria and archaea - which reveals a pre-metazoan origin for this gene family. LOX genes expanded during metazoan evolution resulting in two superfamilies, LOXL2/L3/L4 and LOX/L1/L5. Considering the current knowledge on the function of mammalian LOX isoforms in ECM remodeling, we propose that LOXL2/L3/L4 members might have preferentially been involved in making cross-linked collagen IV-based basement membrane, whereas the diversification of LOX/L1/L5 forms contributed to chordate/vertebrate-specific ECM innovations, such as elastin and fibronectin. Our work provides a novel view on the evolution of this family of enzymes.



OPEN

Origin and evolution of lysyl oxidases

Xavier Grau-Bové[1], Iñaki Ruiz-Trillo[1,3,4] & Fernando Rodriguez-Pascual[2]

Received: 20 October 2014
Accepted: 15 April 2015
Published: 29 May 2015

Lysyl oxidases (LOX) are copper-dependent enzymes that oxidize primary amine substrates to reactive aldehydes. The best-studied role of LOX enzymes is the remodeling of the extracellular matrix (ECM) in animals by cross-linking collagens and elastin, although intracellular functions have been reported as well. Five different LOX enzymes have been identified in mammals, LOX and LOX-like (LOXL) 1 to 4, showing a highly conserved catalytic carboxy terminal domain and more divergence in the rest of the sequence. Here we have surveyed a wide selection of genomes in order to infer the evolutionary history of LOX. We identified LOX proteins not only in animals, but also in many other eukaryotes, as well as in bacteria and archaea – which reveals a pre-metazoan origin for this gene family. LOX genes expanded during metazoan evolution resulting in two superfamilies, LOXL2/L3/L4 and LOX/L1/L5. Considering the current knowledge on the function of mammalian LOX isoforms in ECM remodeling, we propose that LOXL2/L3/L4 members might have preferentially been involved in making cross-linked collagen IV-based basement membrane, whereas the diversification of LOX/L1/L5 forms contributed to chordate/vertebrate-specific ECM innovations, such as elastin and fibronectin. Our work provides a novel view on the evolution of this family of enzymes.

Lysyl oxidases (LOX) are a family of copper-dependent amino oxidases for which important roles in cancer and vascular and fibrotic diseases have been proposed[1]. Five different LOX enzymes have been identified in mammals (LOX, and LOX-like 1 to 4), showing a high degree of homology in the catalytic carboxy terminal end and more divergence in the rest of the sequence[2]. While intracellular functions have been reported for LOX proteins, the primary role of this family of enzymes is the remodeling of the extracellular matrix (ECM), due to their capacity to convert lysine and hydroxylysine residues in collagens and elastin into highly reactive aldehydes, which eventually condense with other oxidized groups or intact lysines to form a variety of inter- and intrachain cross-linkages. The fundamental role of LOX proteins in ECM homeostasis has been demonstrated in experiments with mice lacking the LOX gene, which die just before or soon after birth by severe cardiovascular malformations, most likely involving defective elastogenesis[3]. Moreover, mice deficient in LOXL1, the closest mammal paralog of LOX, exhibit also cardiovascular defects, although they are perfectly viable and show a normal life span[4]. The remaining members (LOXL2-4) share the presence of four scavenger receptor cysteine-rich (SRCR) domains, a unique class of ancient, highly conserved polypeptide module present in a number of soluble and membrane-bound proteins for which no unifying function has been so far defined[5]. Recent work has described the capacity of LOXL2 and LOXL4 to enhance collagen IV deposition and assembly[6,7]. Nevertheless, it remains to be defined how this ECM remodeling capabilities fit together with the intracellular actions described for some of these SRCR-containing LOX members, such as the role of LOXL2 in the regulation of gene transcription[8,9].

It is beyond doubt that the numerous evolutionary transitions from unicellular to multicellular organisms that occurred within eukaryotes could have never happened without their organization into extracellular structures. In contrast to sessile algae, fungi, and plants, which acquired a comparatively uniform

[1]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Spain. [2]Centro de Biología Molecular "Severo Ochoa" Consejo Superior de Investigaciones Científicas (C.S.I.C.) / Universidad Autónoma de Madrid (Madrid), Madrid, Spain. [3]Departament de Genètica, Universitat de Barcelona, Barcelona, Spain. [4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. Correspondence and requests for materials should be addressed to F.R.-P. (email: frodriguez@cbm.csic.es)

composition in their cell walls, animals exhibit a complex and heterogeneous ECM, with multiple protein families involved in the construction of intricate structural networks, as well as many protein complexes devoted to intercellular adhesion and communication[10]. Recent genome data have revealed that some of the large, secreted, multidomain ECM components, including basement membrane-forming collagen IV and fibrillar collagens appear to be specific to the Metazoa[11]. Nevertheless, important domains from ECM proteins have a pre-metazoan origin. For instance, the filasterean *Capsaspora owczarzaki,* a close relative of Metazoa, has protein domains related to laminin and fibronectin, as well as a complete integrin adhesome[12–14]. Furthermore, choanoflagellates harbor many collagen motifs and domains otherwise specific to animals, such as the repeated GXY triple helical motif (even though these organisms lack fibrillar collagen)[15]. Domain shuffling of ancestral, premetazoan domains on the metazoan stem lineage have been proposed to give rise to the fibril-forming collagens, which are conserved throughout the metazoan evolutionary tree[16,17]. The same is true for collagen IV[18,19]. From these "founder genes", rounds of gene duplication and domain or exon shuffling have resulted in the formation of different classes, comprising currently 28 collagen genes in vertebrates, which play structural roles in soft tissues or act as templates for biomineralisation in bone or teeth[17,20]. However, this family expansion has not been universal for all metazoans. For example, *Drosophila* lacks any fibrillar collagens that were most likely secondarily lost[21]. Remarkably, chordates and, specifically, vertebrates have witnessed a significant number of ECM innovations, including not only the duplication of pre-existing deuterostome genes but also the generation of complex forms of collagen (transmembrane collagens, FACIT collagens, among others) or of specific protein innovations[22]. In particular, elastin is one of the vertebrate-specific ECM novelties, and has played a fundamental role in the evolution of a high-pressure, pulsatile blood circulation system[23].

Very limited information is available about the existence of LOX isoforms in non-bilaterian animals or other organisms. LOX-generated cross-links have been isolated from a sponge (*Haliclona oculata*), a sea urchin (*Strongylocentrotus droebachensis*), a sea cucumber (*Thyone briarius*), as well as from several annelids, echinodermates and molluscs[24,25]. Additionally, arthropodes like *Drosophila* have been reported to have two distinct LOX-like genes, whereas some chordates such as the cyprinidae *Danio rerio* (zebrafish) present up to 10 LOX genes[26–28]. A preliminary phylogenetic analysis of LOX genes revealed that human LOX and LOXL1 share a common ancestor and form an independent group from LOXL2, LOXL3 and LOXL4, being likely related to the *Ciona intestinalis* LOX1 and LOX2, respectively[22]. However, we lack an understanding of the evolutionary origin of the members of the LOX family, and how they relate to the evolution of the main ECM components such as collagens and elastin.

We here have surveyed a wide selection of genomes representing all the major eukaryotic and prokaryotic clades, aiming to reconstruct the evolutionary history of LOX enzymes. Our phylogenetic analyses, based on the conserved lysyl oxidase domain of LOX enzymes, show that LOX sequences are identifiable not only in animals, but also in many other eukaryotes, as well as in bacteria and archaea. This points at a much older origin than previously thought for LOX enzymes, preceding the origin of animals[21]. Our phylogenetic analyses show a significant expansion of LOX types during metazoan evolution, giving rise to three LOX families in Porifera (sponges) and two superfamilies in Eumetazoa (bilaterians and cnidarians). The LOXL2/L3/L4 superfamily is typically associated with SRCR domains, whereas LOX/L1/L5 display distinct N-terminal domains, and is related to the mammalian LOX and LOXL1. Based on the existing knowledge on the evolution of collagens and elastin, we propose here that LOXL2/L3/L4 members might contribute to the cross-linking of basement membrane collagen IV, whereas LOX/L1/L5 proteins may have evolved to cover the requirements of more sophisticated ECM in chordate/vertebrate phyla.

Results

The prokaryotic history of LOX enzymes. Figure 1 shows phylogenetic analysis of LOX enzymes in prokaryotes and eukaryotes (panel A, unrooted tree) and Holozoa only (panel B, using ichthyosporean LOX as tree root). The network of reciprocal blast hits with indication of their score is shown in Fig. 2. Complete phylogenies are shown in Supplementary Files S1 to S4, sequences in Files S5 and S6.

Besides the eukaryotic LOX enzymes, our survey identifies for the first time LOX in both Archaea and Bacteria. In particular, LOX-coding genes are widely distributed in Bacteria, being present in five major clades: Bacteroidetes, Actinobacteria, Proteobacteria, Gemmatimonadetes and Deinococcus-Thermus (Fig. 1A). In contrast, the archaeal LOX homologs cluster into two separate groups of thaumarchaeotes and euryarchaeotes (Fig. 1A). In fact, each of these archaeal groups are associated to bacterial LOX and appear to be composed of sequences from phylogenetically close organisms (Supplementary Fig. S1 and S2). This suggests that thaumarchaeotes and euryarchaeotes could have acquired LOX through two independent horizontal gene transfer (HGT) events from bacteria (Figs. 1A and 2), although identification of the bacterial donors is required to confirm this hypothesis.

Finally, it is interesting to note that, in contrast to eukaryotic LOX, most (except three) of the identified prokaryotic sequences exhibit simple protein domain architectures with just the LOX domain, with or without signal peptide and/or transmembrane region.

LOX in unicellular eukaryotes. Our data also show the presence of LOX enzymes in different eukaryotic non-metazoan lineages (Fig. 1). Specifically, we identified LOX genes in the genomes of some

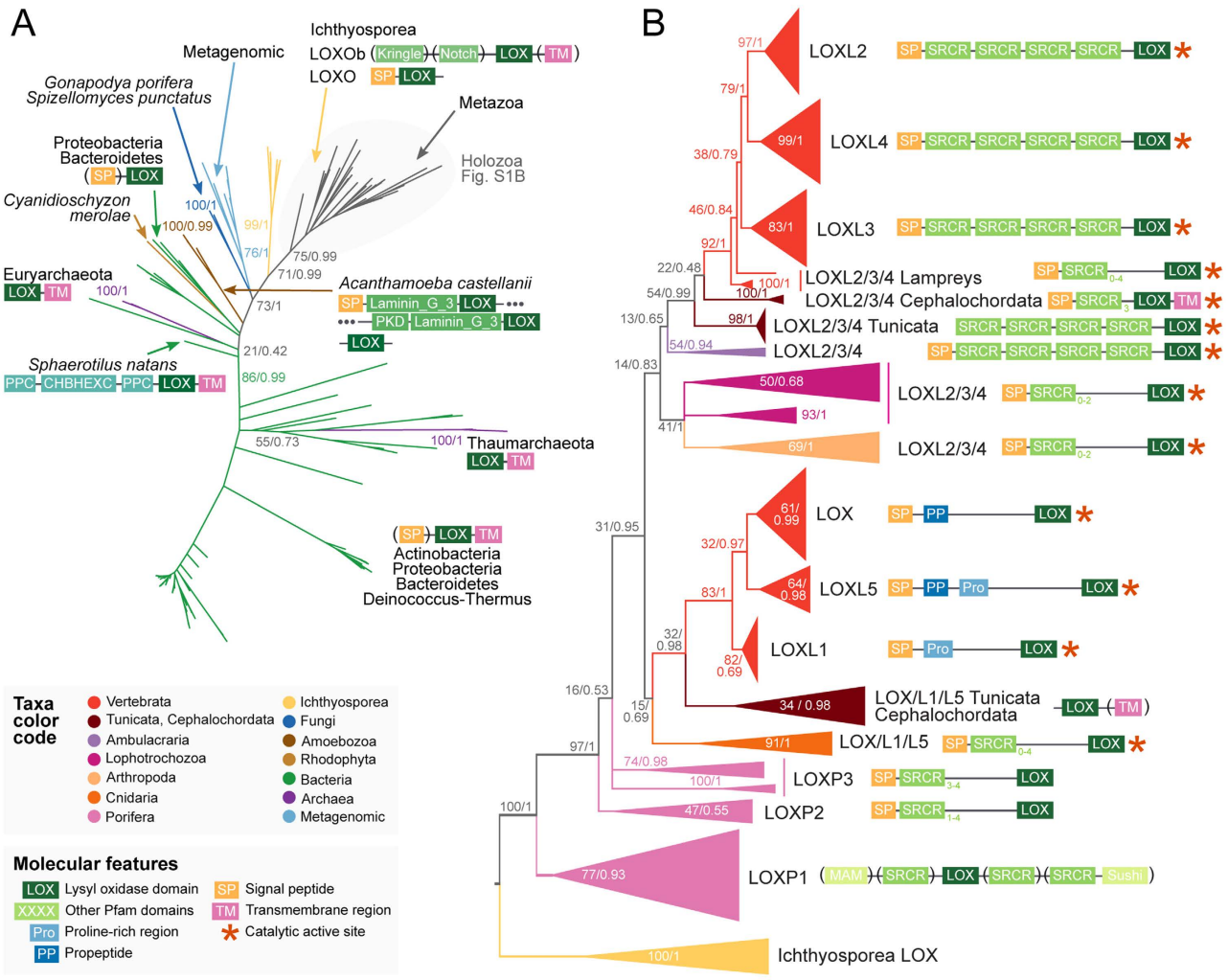


Figure 1. Phylogenetic trees of LOX enzymes in eukaryotes and prokaryotes. A) Unrooted tree of 154 LOX domains from eukaryotic and prokaryotic genomes as inferred by bayesian inference. **B**) Rooted tree of 129 LOX domains from an expanded selection of holozoans (animals and their unicellular relatives, see grey-shadowed area of part A), as inferred by bayesian inference. Nodal support values are shown at key branches (Maximum likelihood bootstrap support/Bayesian posterior probabilities). Sequences are color-coded according to their taxonomic assignment. The consensus protein domain architectures of each LOX family are shown adjacent to each phylogeny, including Pfam domains (green boxes), proline-rich and propeptide regions (blue), transmembrane regions (pink), signal peptide motifs (orange) and the Interpro 019828 motif (red asterisk). The trees are not to scale. See supplementary Figures S1, S2, S3 and S4 for detailed versions of these phylogenies, including scaled branches and complete nodal support.

Amorphea/Unikonta taxa (including animals, fungi and a number of unicellular clades), as well as from the Rhodophyta (red algae, from the Diaphoratickes supergroup).

The phylogenetic analysis of LOX recovers a major clade that includes opisthokont LOX homologs (all known animal enzymes, fungi and ichthyosporeans) together with a number of environmental metagenomic sequences (Fig. 1; BS 73%, BPP 0.99). Within fungi, we identify LOX homologs in the chytrid *Spizellomyces punctatus* and the monoblepharidomycete *Gonapodya prolifera*. Ichthyosporeans, which are a group of unicellular organisms closely related to animals[29], have also the most animal-like LOX genes according to our phylogeny (Fig. 1). They have two sets of LOX, one of which (LOXOb) has acquired C-terminal Kringle, PLAT and Notch protein domains (Fig. 1A). While the function of LOX in ichthyosporeans is at present unknown, the occurrence of the transmembrane region of Notch suggests some membrane-associated role akin to the SRCR-containing LOX of animals.

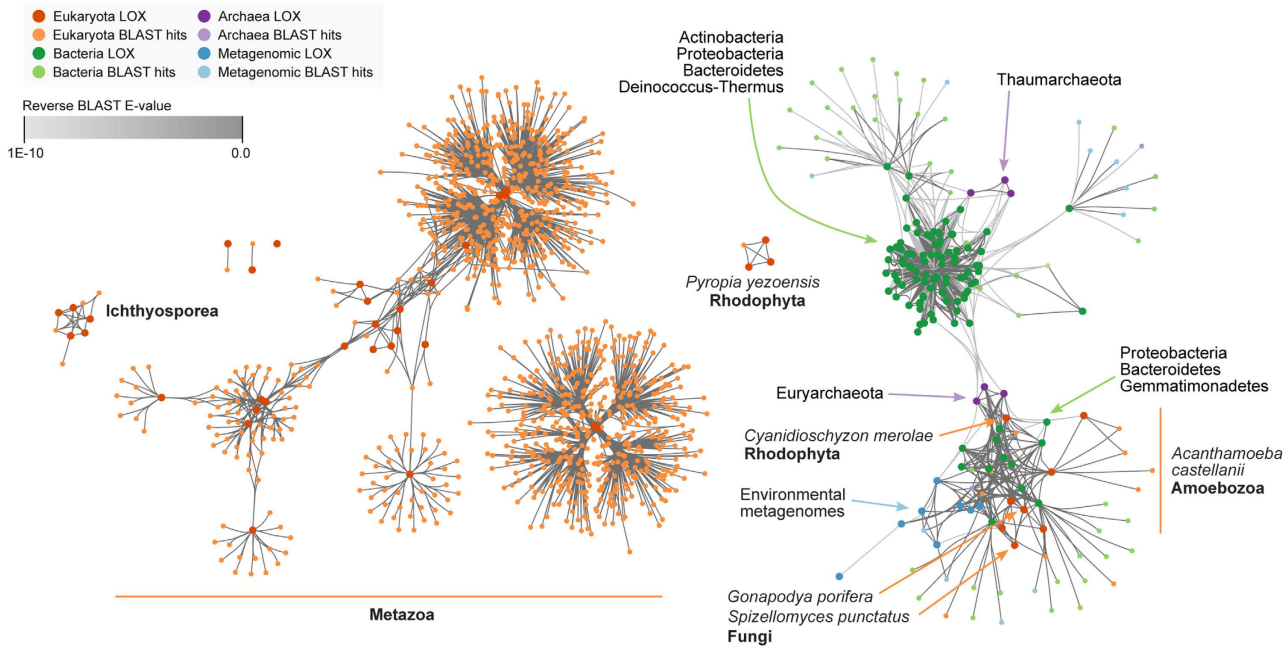


Figure 2. Network of reciprocal BLAST searches for LOX enzymes. Each node represents a LOX-containing protein. Nodes are connected by edges when they are reciprocal BLAST hits of each other (see Methods). Nodes are color-coded according to their taxonomic assignment (for some clusters of interest, further taxonomic details are also shown). Edges are color-coded according to the E-value of each BLAST hit.

We also identified LOX homologs in the unicellular amoebozoan *Acanthamoeba castellanii* and the rhodophytes *Cyanidioschyzon merolae* (unicellular algae) and *Pyropia yezoensis* (multicellular seaweed). However, they could not be unambiguously classified to any specific group, probably due to either low statistical support (*A. castellanii* and *C. merolae*) or insufficient data (*P. yezoensis*). According to the network of reciprocal BLAST (Fig. 2), the *C. merolae* LOX and the 4 copies of *A. castellanii* (BS 98%, BPP 0.99) seem to be related to prokaryotic, environmental or fungal sequences, whereas *P. yezoensis*' proteins cluster separately from the rest of the known LOX enzymes.

It is interesting to note that neither *A. castellanii* nor fungi have collagen-based ECM structures equivalent to those of animals. As for the multicellular seaweeds, they do have complex polysaccharide-based ECM, but do not possess collagen-based structures.

LOX diversification in animals. It is within animals where we found the greatest variety of LOX forms, with many duplications and frequent rearrangements of protein domain architectures (Fig. 1B).

We identified three groups of LOX enzymes specific to Porifera (sponges), termed LOXP1-3 (pink branches in Fig. 1B). Each of them has different protein domain architectures based on transmembrane SRCR domains, both N- and C-terminal. The LOXP1 family is only present in calcareous sponges (*Sycon ciliatum* and *Leucosolenia complicata*) and contains proteins with multiple domains, including not only SRCR but also MAM or Sushi. Given that LOXP1 is the earliest family present in animals, this means that the association between LOX and SRCR domains was already present at the origin of animals. LOXP2 and P3 families, both with the canonical N-terminal SRCR repeats, are present in demosponges (*Amphimedon queenslandica*), homoscleromorph (*Oscarella carmela*) and calcareous sponges.

A duplication event at the origin of eumetazoans gave birth to two animal LOX superfamilies that although not statistically supported, are recovered by both Maximum Likelihood and Bayesian inference analyses: LOX/L1/L5 (composed of homologs of human canonical LOX and LOXL1, plus the fish-specific LOXL5) and LOXL2/L3/L4 (homologs of human LOXL2, LOXL3 and LOXL4).

The LOX/L1/L5 superfamily (BS 15%, BPP 0.69) is present in cnidarians (dark orange branch in Fig. 1B), that have the ancestral SRCR-containing form, and chordates (red and dark red branches), that lack SRCR domains (Fig. 1B, see also a cladogram with domain gain/loss in Fig. 3). At the origin of vertebrates, this superfamily gives rise to the LOX, LOXL1 and LOXL5 (exclusive to various fish clades) gene families. LOXL1 enzymes have a N-terminal proline-rich region, also conserved in LOXL5 but lost in canonical LOX. Canonical LOX and LOXL5, in turn, share an exclusive propeptide region (Fig. 1B).

The LOXL2/L3/L4 superfamily (BS 14%, BPP 0.83) was lost in cnidarians and is only present in bilaterian genomes (Figs. 1B and 3). All the families retain the ancestral SRCR-containing form, with

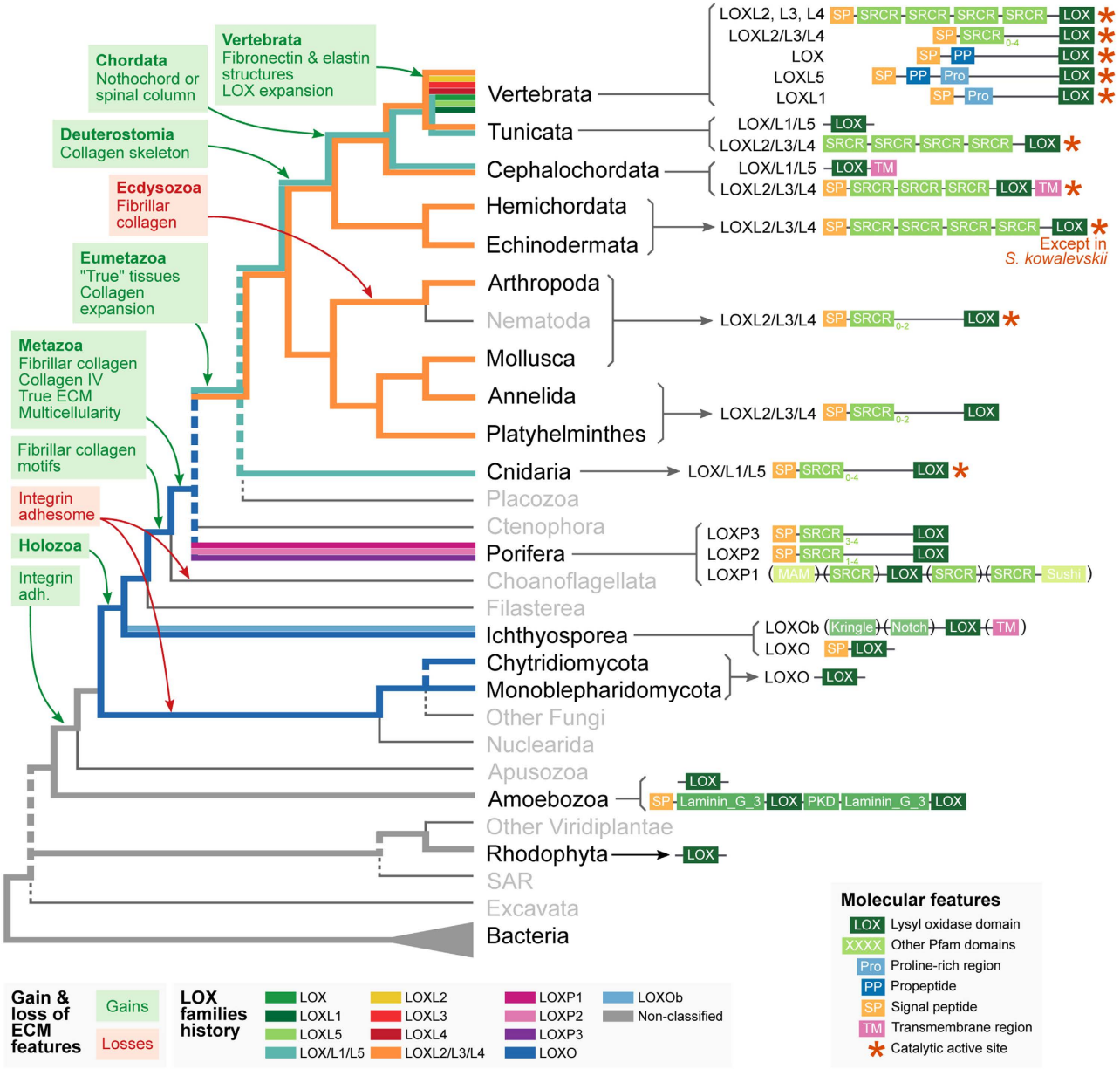


Figure 3. Reconstruction of the evolutionary history of LOX enzymes and ECM across the tree of life. The cladogram represents a consensus view of the eukaryotic tree of life (see Methods) with bacteria as outgroup. Each bold, colored line represents a LOX family (as indicated in the legend); its route along the tree represents their pattern of appearance and loss in each taxonomic group. Dashed lines represent unclear phylogenetic relationships. Green- and red-colored boxes represent gains and losses of ECM features, respectively. The consensus protein domain architectures of each LOX family are shown adjacent to each taxonomic group, including Pfam domains (green boxes), proline-rich and propeptide regions (blue), transmembrane regions (pink), signal peptide motifs (orange) and the Interpro 019828 motif (red asterisk).

variations in the number of repeats (Fig. 1B). This is the only LOX family present in protostomes (arthropods, molluscs, annelids and platyhelminths) and ambulacrarian deuterostomes (hemichordates and echinoderms). It is also present in tunicates and cephalochordates. The vertebrate-specific LOXL2, LOXL3 and LOXL4 families originated after the divergence of *Petromyzon marinus* (sea lamprey), which retains the ancestral type. All of them have four N-terminal SRCR repeats.

Overall, vertebrates have the highest count of LOX enzyme types among eukaryotes, with five widespread families (canonical LOX, LOXL1, LOXL2, LOXL3 and LOXL4), one family specific to fishes (LOXL5, found in actinopterygian, sarcopterygian and cartilaginous fishes) and one specific to lampreys (LOXL2/L3/L4). These LOX types display five different protein domain architectures (Fig. 1B).

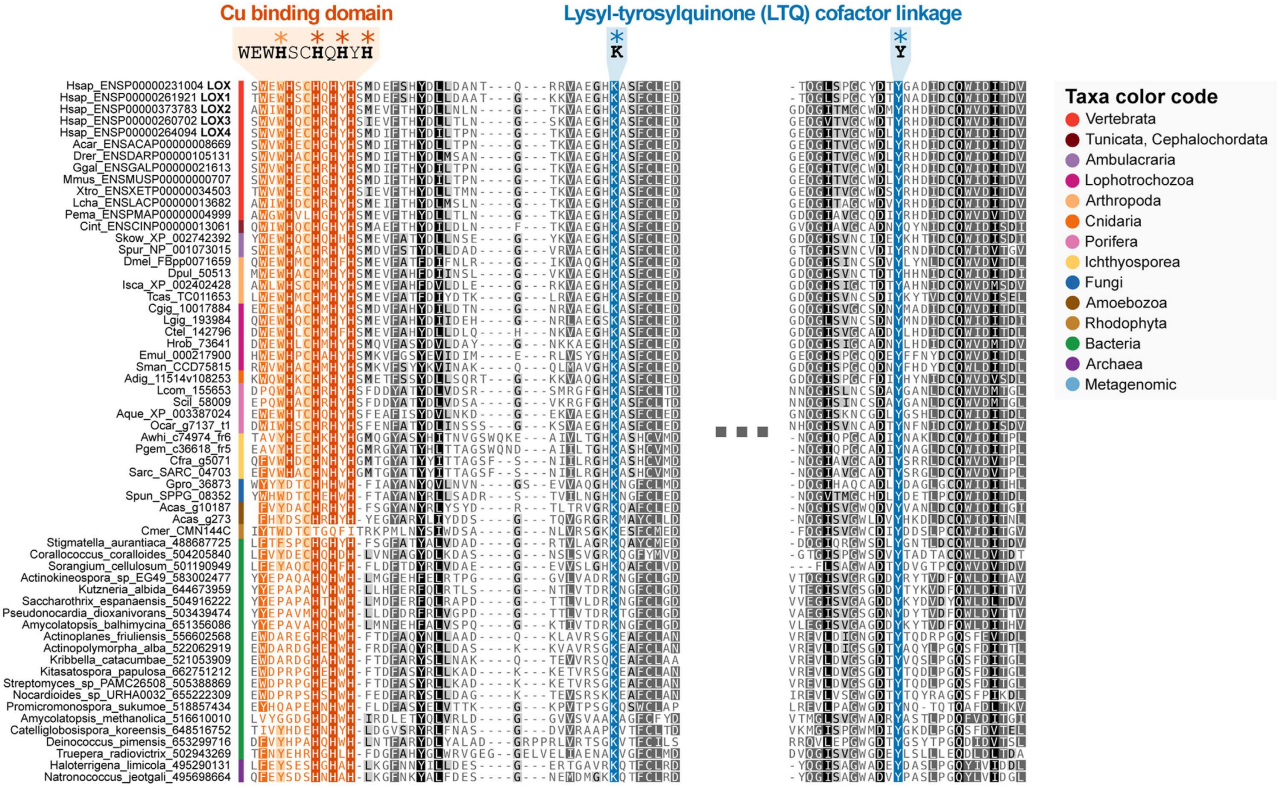


Figure 4. Multiple sequence alignment of catalytic LOX domains. 60 LOX proteins representing all of the groups analyzed in our study were aligned in order to inspect for conserved residues involved in the catalysis. Conserved residues highlighted in red constitute the cores of histidines forming the copper binding site within the InterPro 019828 motif (Lysyl oxidase). Note that the histidine depicted in orange within this motif is conserved in animals and ichthyosporeans, but not present in bacterial, fungal or amebozoan sequences. Strictly conserved lysine and tyrosine residues involved in LTQ cofactor linkage are highlighted in blue.

We could not identify any LOX gene in nematodes, nor in the placozoan *Trichoplax adhaerens* or the ctenophore *Mnemiopsis leidyi*.

Assessment of the catalytic activity of novel LOX homologs. The presence of LOX domains in previously unreported eukaryotes and prokaryotes raises the question of whether they are enzymatically active proteins or not. It has been demonstrated that LOX catalytic activity relies on the C-terminal domain of the protein, where two features are needed. First, the core of histidines forming the copper binding site, the so-called "copper-talon", which matches the conserved motif Interpro 019828 (WEW**H**SC**H**Q**H**Y**H**SMD in human LOX, Hsap_ENSP00000231004)[30]. Second, the lysine and tyrosine residues involved in the association with the lysyl tyrosyl quinone (LTQ) cofactor (K320 and Y355 in Hsap_ENSP00000231004)[31]. These key amino acids are widely conserved in all the groups analyzed in our study (Fig. 4, see also Supplementary Files S7 and S8) with the exception of the rhodophyte *C. merolae*, which lacks the histidine core. This observation predicts that these LOX homologs can be enzymatically competent to oxidize substrates. Interestingly, the first histidine residue within the copper binding site (H289 in Hsap_ENSP00000231004) is conserved in animals and ichthyosporeans, but is not present in bacterial, fungal or amebozoan sequences. Recent experimental evidence have provided useful information about whether the loss of this histidine residue can compromise the binding of copper, and therefore, the catalytic activity[32]. These authors sequentially mutated the histidine into alanines (being incapable to bind copper), and showed that the substitution of the first histidine did not significantly alter the ability of the enzyme to bind copper and oxidize substrates. Based on this report, it can be predicted that LOX domains identified in our work would display catalytic activity as they possess the core of the three essential histidines and the residues implicated in the LTQ linkage.

Discussion

Our results provide the most comprehensive up-to-date phylogenetic analysis of the family of LOX enzymes. A main conclusion is that the LOX domains are more widely distributed than previously thought, as we identify clear homologs in animals and other eukaryotes, as well as bacteria and archaea[22].

Based on our phylogenetic analyses with a wide taxon sampling, we can reconstruct the evolution and diversification of LOX enzyme families in eukaryotes and prokaryotes. With respect to the eukaryotic LOX enzymes, we identify a group of ichthyosporean and fungal LOX homologs as the closest relatives to the known animal enzymes (Fig. 1). This clearly indicates that this amino oxidase enzyme family was already present in the opisthokont ancestor, thus predating the origin of metazoans. Different scenarios could explain the origin of this opisthokont LOX according to our results. First, it could have been derived from an ancestral eukaryotic homolog from which the *A. castellanii* and *C. merolae* copies could have derived as well. Second, it could have been acquired by a horizontal gene transfer (HGT) event from bacteria to an ancestral opisthokont.

In order to understand the evolutionary history of LOX enzymes outside opisthokonts, we need to understand how LOX enzymes first appeared (in eukaryotes or prokaryotes) and whether HGT events took place (and when). However, the distribution of LOX cannot be conclusively explained by our phylogeny, as several non-exclusive scenarios would fit. For example, a potential explanation would be a bacterial origin of LOX, followed by a later transfer to eukaryotes (either by HGT or during the process of eukaryogenesis) and multiple secondary losses. Another possibility would be a later eukaryotic origin followed by a number of HGT events between eukaryotes and prokaryotes, and within prokaryotes as well.

In support of the HGT-driven scenarios, the genomes of *A. castellanii* and *C. merolae* are both known to have experienced multiple HGTs from bacteria, and the same is true for amoebozoan genes being transferred to prokaryotes[33–35] It is worth noting that HGT of metabolic genes from prokaryotes is an important factor underlying the diversification of eukaryotes, particularly in the case of amoebas such as *A. castellanii* or a hypothetical amorphean ancestor[12,33,36]. If this were the case, the acquisition of LOX by an ancestral microbial eukaryote would have had an important, delayed effect in the evolution of the ECM, as it eased the appearance of the current enzyme types essential for its formation.

The presence of LOX enzymes in bacteria raises the question of the function of LOX within these organisms. Several collagen-like proteins have been identified in bacteria, and for some of them, the formation of a stable triple helix has been demonstrated[21,37]. Some of the best characterized bacterial collagen-like proteins are the streptococcal Scl1 and Scl2, which are expressed on the cell surface of group A *Streptococcus* and contribute to bacterial pathogenicity through the binding to host ECM components including integrins and fibronectin[38,39]. Our analysis did not identify LOX isoforms in members of the *Streptococcus* genus, but, for example, in a number of *Streptomyces* species, for which collagen-like sequences have also been genome-annotated (see, for instance, Uniprot entries: D9WI30 or D6B4A5, www.uniprot.org). Nevertheless, a higher order structure reminiscent of intra- or interchain covalent association has not yet been described for bacterial collagen-like proteins, therefore making unlikely that LOX may cross-link bacterial collagenous material. While more studies are needed to elucidate the function of bacterial LOX enzymes, it can be hypothesized that LOX proteins may be a component of the enzymatic repertoire of bacterial metabolism transferred to eukaryotes and adapted to new functions, as suggested to have occurred, for instance, with the epigenetic machinery[40]. Interestingly, collagen-like proteins present in bacteria have also been proposed to originate from an HGT event from metazoans to bacteria[41].

Current views of the evolution of the animal ECM envision its constitution as the result of a gradual appearance of specific gene families and domains in pre-metazoan lineages, followed by remarkable expansions in animals. This is best exemplified by the presence of a fully functional integrin adhesome in *C. owczarzaki*, a unicellular filasterean with aggregative behavior that also has proteins with laminin and fibronectin motifs (although with different domain architectures than their animal counterparts)[12–14,42,43]. This is also the case of the choanoflagellates *Monosiga brevicollis* and *Salpingoeca rosetta*, that have proteins with collagen and laminin domains (also without a clear homologs in animals)[14,15,44]. Further refinement of these pre-existing protein families and the appearance of Metazoa-specific innovations provided the chordates and vertebrates with a wider repertoire of ECM proteins to fulfill novel functions in the vasculature or in the nervous system[18].

Our phylogenetic analysis of LOX revealed a relatively similar pattern of evolution: LOX domains were already present in unicellular eukaryotes (notably in the ichthyosporeans, that are closely related to Metazoa), and further expanded during metazoan evolution. Interestingly, unicellular organisms such as the ichthyosporeans *Sphaeroforma arctica*, *Creolimax fragrantissima*, *Pirum gemmata* and *Abeoforma whisleri* or the amoebozoan *Acanthamoeba castellanii*, display forms of LOX associated with domains thought to serve extracellular protein-protein interactions, for example PKD, Kringle or PLAT (with or without the presence of transmembrane regions), much in the same role that SRCR has been postulated to play in SRCR-containing LOX forms[2].

According to our study, SRCR domains first associated with LOX proteins in Metazoa, specifically in sponges (see Fig. 3). The SRCR domains in sponges are present both at N- and C-terminal, with and without association with other protein architectures, such as MAM or Sushi. Adult sponges consist of two layers of cells with epithelial features supported by a central cavity, the mesohyl, consisting of rigid material. Fibrillar and basement membrane collagens have been identified in the mesohyl and in the lamina were the two layers of cells attach, respectively[45,46]. Therefore, sponges constitute the first class of organisms where LOX enzymatic activities might have begun to sculpt the ECM. Whether LOX may have provided Porifera with novel capabilities such as spicule biomineralization or body stiffening required for

efficient water flow is at present unknown. It is worth mentioning that neither Ctenophora nor Placozoa have LOX genes. The origin of the eumetazoans witnessed the main branching of LOX isoforms, giving place to the LOXL2/L3/L4 and LOX/L1/L5 superfamilies (Fig. 1B and Fig. 3). The former kept the SRCR-LOX architecture invariably from arthropods to vertebrates, with minimal variations in the number of SRCR domains. The observation that this class of LOX is present in arthropods such as *Drosophila melanogaster*, which lacks fibrillar collagen, suggests that these LOX isoforms might preferentially (but not exclusively) cross-link basement membrane collagen IV, and thereby controlling ECM stiffness, as recently described[26,47]. In fact, collagen IV-cross linking activities for mammalian LOXL2 and LOXL4 have recently been reported[6,7]. Nevertheless, intracellular functions beyond matrix cross-linking have been also reported for LOXL enzymes, for instance transcriptional regulation or control of cell cycle and apoptosis for LOXL2[8,9].

In contrast to LOXL2/L3/L4, the LOX/L1/L5 superfamily experienced significant changes in domain architecture during evolution. While forms present in cnidarians retain SRCR domains, LOX/L1/L5 from tunicates and cepholochordates show no recognizable associated domains, and chordates and vertebrates display forms with propeptide and proline-rich regions typical of mammalian LOX and LOXL1 (Fig. 1B). As shown in Fig. 3, the appearance of LOX isoforms with these domain architectures is coincident with a significant expansion of vertebrate-specific ECM innovations, a circumstance reinforcing their widely accepted role as catalyzers of lysine-derived cross-links in fibrillar collagens and elastin. To this respect, LOX and LOXL1 have been reported to interact with tropoelastin through sequences in the N-terminal pro-regions[48]. Although the specific motifs within the pro-regions of LOX and LOXL1 that drive the association with elastin are not known, significant homology exists at the N-terminal sequence to support this interaction. Additionally, strong binding has been reported between LOX and fibulin-4 and LOXL1 and fibulin-5[4,49]. Fibulin-4 and -5 are essential proteins for the assembly of elastic fibers, and their interaction with LOX isoforms seems to facilitate the cross-linking of tropoelastin within elastic fibers[50]. Based on these observations, it can be inferred that LOX and LOXL1 forms evolved to contribute to elastogenesis, an assumption further reinforced by the result of the inactivation of these genes in mouse models, both giving rise to vascular phenotypes due to impaired elastic fiber formation[3,4].

It is interesting to mention that LOX and LOXL1 are proteolytically processed by bone morphogenic protein 1 (BMP1)/Tolloid-like metalloproteinases[51–54]. First identified as pro-collagen C-proteinases, this family of proteolytic enzymes has been described to cleave a wide repertoire of substrates[55]. It is worth mentioning that, with the exception of apolipoprotein 1 and gliomedin, which play unique roles in lipid metabolism and peripheral nervous system, respectively, BMP1 substrates belong to the category of ECM proteins or ECM-related factors, including fibrillar procollagens, small leucine-rich proteoglycans, basement membrane components, and mineralization factors, among many others[55]. The fact that LOX and LOXL1 forms are also cleaved by BMP1-related proteases suggests that the primary function of these LOX forms is matrix-oriented. LOX and LOXL1 needs to be processed to yield the catalytically active forms. Therefore, it is conceivable to propose that the proteolysis step serves as a quality control step to keep the LOX enzyme in a latent state until the proper substrate is encountered.

Another important vertebrate ECM innovation is fibronectin, an adhesive protein involved in many cellular responses with a significant role in wound healing[56]. In this context, the formation of a fibronectin matrix is critical for the subsequent assembly of types I and III collagen fibrils. The canonical LOX has been reported to interact with fibronectin through sequences both in the pro-region and in the C-terminal[57]. In fact, fibronectin may also contribute to the processing of the pro-enzyme, as fibronectin scaffolds support BMP1 binding through periostin[58,59]. Taken together, these evidences point out to a significant role for LOX and LOXL1, through their associated domains, in chordate/vertebrate-specific ECM building, particularly in the circulatory system and during tissue repair. Within these functions, it is interesting to note that LOXL5, present in early-branching vertebrate clades of fishes (Actinopterygii, Chondrichthyes and Sarcopterygii), contains both the proline-rich and propeptide regions. Thus, fishes retain both functionalities in the same enzyme, whereas its sister LOX family, present in the other vertebrates, has lost the proline-rich region. This probably reflects the specialization of the canonical LOX in particular functions in non-fish vertebrates.

In conclusion, our phylogenetic analysis of LOX proteins permits to trace the evolution of this family of enzymes, particularly in the context of the acquisition of the ECM components, collagen and elastin. Fig. 3 illustrates the appearance of LOX proteins within the elaboration of ECM components during eukaryotic evolution. Remarkable events include: 1) the presence of LOX forms in unicellular eukaryotes, associated to several domain architectures presumably serving extracellular protein-protein interactions; 2) the acquisition of SRCR domains as a specific feature of animals, presumably coincident with the appearance of true ECM in early metazoans; and 3) the generation of chordate/vertebrate LOX forms possibly supporting novel ECM innovations such as elastin and fibronectin.

Methods

Taxon sampling and sequence retrieval. LOX sequences were queried in complete genome or transcriptome sequences of 117 eukaryotic taxa representing all known eukaryotic supergroups, as well as all the major metazoan clades. Taxon sampling includes 37 metazoans, 10 unicellular holozoans, 24 fungi, 2 nucleariids, 1 apusozoan, 4 amoebozoans, 7 plants, 5 chlorophytes, 3 rhodophytes, 1 glaucophyte, 8 heterokonts, 6 alveolates, 1 rhizarian, 1 haptophyte, 1 cryptophyte and 6 excavates (Supplementary

Tables S1 and S2, list of sequences in Files S5 and S6). Prokaryotic sequences were queried in the NCBI non-redundant database and the Microbial Dark Matter Project database[60]. The proteins with LOX domains were retrieved from the complete proteomes with HMMER[61], using a Hidden Markov motif of the LOX domain as defined by Pfam (PF01186)[62]. These proteins were inspected using Pfamscan and manual alignments to assess the presence of protein domains including those found in mammalian LOX, such as the proline-rich and pro-peptide motifs, or scavenger receptor cysteine-rich domains[62].

Phylogenetic inference. The LOX domains (PF01186) of the retrieved sequences were aligned using the Mafft 7 L-INS-i algorithm, optimized for local sequence homology[63]. Two alignments were produced: 1) one containing eukaryotic, bacterial and archaeal proteins (154 sequences, 217 alignment positions; using eukaryotes from Supplementary Table S1); and 2) another one with just animal and ichthyosporean proteins (129 sequences, 283 aligned positions; using animals from Supplementary Table S2). According to ProtTest 3.4 analyses of each alignment[64], the most suitable evolutionary models were WAG+Γ+F and LG+Γ+I, respectively ("Γ" stands for a gamma distribution of among-site rate variation with 4 discrete categories; "I" means that a proportion of invariable sites is considered; and "F" means that empirical amino acid frequencies are inferred from the alignment). The phylogenetic trees of each of these alignments were inferred using the corresponding model of evolution, with two independent methods: Maximum Likelihood (ML) and Bayesian Inference (BI). ML trees were estimated with RAxML 8, starting from 100 random trees and selecting the best inference according to the Γ-based likelihood value[65]. Statistical support for bipartitions was estimated by performing 100 bootstrap replicates, using RaxML with the same evolutionary models. BI trees were estimated with Phylobayes 3.3[66] (which does not account for empirical amino acid frequencies nor invariable sites), running two parallel chains for each alignment. To decide when to stop the runs, we regularly performed a series of bpcomp tests on each pair of chains every 5,000 generations, consisting in burning-in the tree lists every 1% of the generations run so far. The final trees were built using the number of generations and burn-in values that yielded the lowest maxdiff statistics, sampling every 10 trees (provided it was under the 0.1 threshold recommended by Phylobayes). This resulted in 30,000 generations and 5% of burning for the animal and ichthyosporean alignment, and 60,000 and 7% for the eukaryotic and prokaryotic alignment. Bayesian posterior probabilities (BPP) were used for assessing the statistical support of each bipartition. Using these phylogenetic trees, the evolution of LOX enzymes across eukaryotes and prokaryotes was reconstructed, based on a consensus tree of life drawn from different studies[67–69].

Annotation of molecular features. The protein domain architectures of the retrieved sequences were analyzed using Pfamscan[70]. The full proteins were also analyzed with SignalIP 4.1[71] and TMHMM 2.0[72] to search for signal peptide cleavage sites and transmembrane helical domains, respectively (default parameters in both cases). To assess whether the identified LOX domains can have catalytic activity, the InterPro IPR019828 conserved site was searched[73]. Proline-rich and propeptide regions were manually checked in the alignments. Annotations of molecular features are provided in Supplementary Files S7 and S8.

Assessment of horizontal gene transfers. In addition to the information provided by phylogenetic inference, the possibility of horizontal gene transfer (HGT) events between taxa was tested using a reciprocal BLAST approach. Two sequences were considered to be connected if they were reciprocal BLAST hits of each other with an e-value $<10^{10}$, when queried against a combined database consisting of the full NCBI non-redundant protein database, the Microbial Dark Matter database and our selected eukaryotic taxon sampling (see above). The network visualizations of the reciprocal BLAST hits were generated using Cytoscape 3.1.1, clustering the nodes using the built-in force-directed algorithm[74].

References

1. Maki, J. M. Lysyl oxidases in mammalian development and certain pathological conditions. *Histol Histopathol* **24,** 651–660 (2009).
2. Csiszar, K. Lysyl oxidases: a novel multifunctional amine oxidase family. *Progress in nucleic acid research and molecular biology* **70,** 1–32 (2001).
3. Maki, J. M. *et al.* Inactivation of the lysyl oxidase gene Lox leads to aortic aneurysms, cardiovascular dysfunction, and perinatal death in mice. *Circulation* **106,** 2503–2509 (2002).
4. Liu, X. *et al.* Elastic fiber homeostasis requires lysyl oxidase-like 1 protein. *Nat Genet* **36,** 178–182 (2004).
5. Martinez, V. G., Moestrup, S. K., Holmskov, U., Mollenhauer, J. & Lozano, F. The conserved scavenger receptor cysteine-rich superfamily in therapy and diagnosis. *Pharmacol Rev* **63,** 967–1000 (2011).
6. Bignon, M. *et al.* Lysyl oxidase-like protein-2 regulates sprouting angiogenesis and type IV collagen assembly in the endothelial basement membrane. *Blood* **118,** 3979–3989 (2011).
7. Busnadiego, O. *et al.* LOXL4 is induced by TGF-beta1 through Smad and JunB/Fra2 and contributes to vascular matrix remodeling. *Mol Cell Biol* **33,** 2388–2401 (2013).
8. Herranz, N. *et al.* Lysyl oxidase-like 2 deaminates lysine 4 in histone H3. *Molecular cell* **46,** 369–376 (2012).
9. Moreno-Bueno, G. *et al.* Lysyl oxidase-like 2 (LOXL2), a new regulator of cell polarity required for metastatic dissemination of basal-like breast carcinomas. *EMBO molecular medicine* **3,** 528–544 (2011).
10. Engel, J. & Chiquet, M. in *The Extracellular Matrix: an Overview* (ed Mecham RP) 1–39 (Springer-Verlag, 2011).
11. Ozbek, S., Balasubramanian, P. G., Chiquet-Ehrismann, R., Tucker, R. P. & Adams, J. C. The evolution of extracellular matrix. *Mol Biol Cell* **21,** 4300–4305 (2010).

12. Sebe-Pedros, A., Roger, A. J., Lang, F. B., King, N. & Ruiz-Trillo, I. Ancient origin of the integrin-mediated adhesion and signaling machinery. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 10142–10147 (2010).
13. Suga, H. *et al.* The Capsaspora genome reveals a complex unicellular prehistory of animals. *Nat Commun* **4** (2013).
14. Williams, F., Tew, H. A., Paul, C. E. & Adams, J. C. The predicted secretomes of Monosiga brevicollis and Capsaspora owczarzaki, close unicellular relatives of metazoans, reveal new insights into the evolution of the metazoan extracellular matrix. *Matrix Biol* **37,** 60–68 (2014).
15. King, N. *et al.* The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* **451,** 783–788 (2008).
16. Exposito, J. Y. *et al.* 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. *J Biol Chem* **283,** 28226–28235 (2008).
17. Exposito, J.-Y., Valcourt, U., Cluzel, C. & Lethias, C. The Fibrillar Collagen Family. *Int J Mol Sci* **11,** 407–426 (2010).
18. Hynes, R. O. The evolution of metazoan extracellular matrix. *The Journal of Cell Biology* **196,** 671–679 (2012).
19. Srivastava, M. *et al.* The Trichoplax genome and the nature of placozoans. *Nature* **454,** 955–960 (2008).
20. Ivanova, V. P. & Krivchenko, A. I. A current viewpoint on structure and evolution of collagens. I. Fibrillar collagens. *J Evol Biochem Phys* **48,** 127–139 (2012).
21. Exposito, J.-Y. & Lethias, C. in *Evolution of Extracellular Matrix Biology of Extracellular Matrix* (eds Fred W. Keeley & Robert P. Mecham) Ch. **3,** 39–72 (Springer Berlin Heidelberg, 2013).
22. Huxley-Jones, J., Robertson, D. L. & Boot-Handford, R. P. On the origins of the extracellular matrix in vertebrates. *Matrix Biol* **26,** 2–11 (2007).
23. Wagenseil, J. E. & Mecham, R. P. Vascular extracellular matrix and arterial mechanics. *Physiological reviews* **89,** 957–989 (2009).
24. Eyre, D. R. & Glimcher, M. J. Comparative biochemistry of collagen crosslinks: Reducible bonds in invertebrate collagens. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **243,** 525–529 (1971).
25. Van Ness, K. P., Koob, T. J. & Eyre, D. R. Collagen cross-linking: distribution of hydroxypyridinium cross-links among invertebrate phyla and tissues. *Comparative biochemistry and physiology. B, Comparative biochemistry* **91,** 531–534 (1988).
26. Molnar, J. *et al.* Drosophila lysyl oxidases Dmloxl-1 and Dmloxl-2 are differentially expressed and the active DmLOXL-1 influences gene expression and development. *J Biol Chem* **280,** 22977–22985 (2005).
27. Gansner, J. M., Mendelsohn, B. A., Hultman, K. A., Johnson, S. L. & Gitlin, J. D. Essential role of lysyl oxidases in notochord development. *Developmental biology* **307,** 202–213 (2007).
28. van Boxtel, A. L. Lysyl oxidases in zebrafish development and teratogenesis, VU University Amsterdam, (2010).
29. Torruella, G. *et al.* Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Mol Biol Evol* **29,** 531–544 (2012).
30. Krebs, C. J. & Krawetz, S. A. Lysyl oxidase copper-talon complex: a model. *Biochim Biophys Acta* **1202,** 7–12 (1993).
31. Wang, S. X. *et al.* A crosslinked cofactor in lysyl oxidase: redox function for amino acid side chains. *Science* **273,** 1078–1084 (1996).
32. Lopez, K. M. & Greenaway, F. T. Identification of the copper-binding ligands of lysyl oxidase. *Journal of neural transmission* **118,** 1101–1109 (2011).
33. Clarke, M. *et al.* Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome biology* **14,** R11 (2013).
34. Ogata, H. *et al.* Genome sequence of Rickettsia bellii illuminates the role of amoebae in gene exchanges between intracellular pathogens. *PLoS genetics* **2,** e76 (2006).
35. Schmitz-Esser, S. *et al.* The genome of the amoeba symbiont "Candidatus Amoebophilus asiaticus" reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *Journal of bacteriology* **192,** 1045–1057 (2010).
36. Andersson, J. O. Gene transfer and diversification of microbial eukaryotes. *Annual review of microbiology* **63,** 177–193 (2009).
37. Xu, Y., Keene, D. R., Bujnicki, J. M., Höök, M. & Lukomski, S. Streptococcal Scl1 and Scl2 Proteins Form Collagen-like Triple Helices. *Journal of Biological Chemistry* **277,** 27312–27318 (2002).
38. Rasmussen, M., Edén, A. & Björck, L. SclA, a Novel Collagen-Like Surface Protein ofStreptococcus pyogenes. *Infection and Immunity* **68,** 6370–6377 (2000).
39. Lukomski, S. *et al.* Identification and Characterization of thescl Gene Encoding a Group A StreptococcusExtracellular Protein Virulence Factor with Similarity to Human Collagen. *Infection and Immunity* **68,** 6542–6553 (2000).
40. Aravind, L., Burroughs, A. M., Zhang, D. & Iyer, L. M. Protein and DNA modifications: evolutionary imprints of bacterial biochemical diversification and geochemistry on the provenance of eukaryotic epigenetics. *Cold Spring Harbor perspectives in biology* **6,** a016063 (2014).
41. Rasmussen, M., Jacobsson, M. & Björck, L. Genome-based Identification and Analysis of Collagen-related Structural Motifs in Bacterial and Viral Proteins. *Journal of Biological Chemistry* **278,** 32313–32316 (2003).
42. Sebe-Pedros, A. & Ruiz-Trillo, I. Integrin-mediated adhesion complex: Cooption of signaling systems at the dawn of Metazoa. *Communicative & integrative biology* **3,** 475–477 (2010).
43. Sebe-Pedros, A. *et al.* Regulated aggregative multicellularity in a close unicellular relative of metazoa. *eLife* **2,** e01287 (2013).
44. Fairclough, S. R. *et al.* Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. *Genome biology* **14,** R15 (2013).
45. Heinemann, S. *et al.* Ultrastructural studies on the collagen of the marine sponge Chondrosia reniformis Nardo. *Biomacromolecules* **8,** 3452–3457 (2007).
46. Boute, N. *et al.* Type IV collagen in sponges, the missing link in basement membrane ubiquity. Biology of t*he cell* **88,** 37–44 (1996).
47. Kim, S. N. *et al.* ECM stiffness regulates glial migration in Drosophila and mammalian glioma models. *Development* **141,** 3233–3242 (2014).
48. Thomassin, L. *et al.* The Pro-regions of lysyl oxidase and lysyl oxidase-like 1 are required for deposition onto elastic fibers. *J Biol Chem* **280,** 42848–42855 (2005).
49. Horiguchi, M. *et al.* Fibulin-4 conducts proper elastogenesis via interaction with cross-linking enzyme lysyl oxidase. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 19029–19034 (2009).
50. Papke, C. L. & Yanagisawa, H. Fibulin-4 and fibulin-5 in elastogenesis and beyond: Insights from mouse and human studies. *Matrix Biol* (2014).
51. Cronshaw, A. D., Fothergill-Gilmore, L. A. & Hulmes, D. J. The proteolytic processing site of the precursor of lysyl oxidase. *Biochem J* **306** (Pt 1), 279–284 (1995).
52. Trackman, P. C., Bedell-Hogan, D., Tang, J. & Kagan, H. M. Post-translational glycosylation and proteolytic processing of a lysyl oxidase precursor. *J Biol Chem* **267,** 8666–8671 (1992).
53. Uzel, M. I. *et al.* Multiple bone morphogenetic protein 1-related mammalian metalloproteinases process pro-lysyl oxidase at the correct physiological site and control lysyl oxidase activation in mouse embryo fibroblast cultures. *J Biol Chem* **276,** 22537–22543 (2001).
54. Borel, A. *et al.* Lysyl oxidase-like protein from bovine aorta. Isolation and maturation to an active form by bone morphogenetic protein-1. *J Biol Chem* **276,** 48944–48949 (2001).

55. Moali, C. & Hulmes, D. J. in *Extracellular Matrix:* Pathobiology *and Signaling*. (ed N. Karamanos) 539–561 (Walter de Gruyter, 2012).
56. To, W. & Midwood, K. Plasma and cellular fibronectin: distinct and independent functions during tissue repair. *Fibrogenesis & Tissue Repair* **4,** 21 (2011).
57. Fogelgren, B. *et al.* Cellular fibronectin binds to lysyl oxidase with high affinity and is critical for its proteolytic activation. *J Biol Chem* **280,** 24690–24697 (2005).
58. Maruhashi, T., Kii, I., Saito, M. & Kudo, A. Interaction between periostin and BMP-1 promotes proteolytic activation of lysyl oxidase. *J Biol Chem* **285,** 13294–13303 (2010).
59. Kudo, A. Periostin in fibrillogenesis for tissue regeneration: periostin actions inside and outside the cell. *Cellular and molecular life sciences* **68,** 3201–3207 (2011).
60. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499,** 431–437 (2013).
61. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic acids research* **39,** W29–37 (2011).
62. Punta, M. *et al.* The Pfam protein families database. *Nucleic acids research* **40,** D290–301 (2012).
63. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30,** 772–780 (2013).
64. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)* **27,** 1164–1165 (2011).
65. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)* **30,** 1312–1313 (2014).
66. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics (Oxford, England)* **25,** 2286–2288 (2009).
67. He, D. *et al.* An alternative root for the eukaryote tree of life. *Curr Biol* **24,** 465–470 (2014).
68. Derelle, R. & Lang, B. F. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol Biol Evol* **29,** 1277–1289 (2012).
69. Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452,** 745–749 (2008).
70. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic acids research* **42,** D222–230 (2014).
71. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Meth* **8,** 785–786 (2011).
72. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305,** 567–580 (2001).
73. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)* **30,** 1236–1240 (2014).
74. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)* **27,** 431–432 (2011).

Acknowledgements

This work was supported by grants from Ministerio de Economía y Competitividad (MINECO; Plan Nacional de I+D+I: SAF2012-34916 to F.R-P., BFU2011-23434 to I.R.-T.), Comunidad Autónoma de Madrid (2010-BMD2321, FIBROTEAM Consortium to F.R-P.), Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (2014 SGR 619 to I.R.-T.) and European Research Council Starting Grant (ERC-2012-Co-616960 to I.R.-T.). X.G.-B. is supported by a pregraduate Formación del Personal Investigador grant from MINECO. We thank Scott A. Nichols (University of Denver) for kindly sharing unpublished protein sequences from *Oscarella carmela*, Jonas Collén (Station Biologique Roscoff) for *Chondrus crispus*, and Maja Adamska (Sars International Centre for Marine Molecular Biology) for *Sycon ciliatum* and *Leucosolenia complicata*.

Author Contributions

X.G.-B. performed the analysis and prepared the figures. All authors (X.G.-B., I.R.-T., and F.R.-P.) designed the experiments, analyzed the data and contributed to manuscript text and reviewed its final version.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Grau-Bové, X. *et al.* Origin and evolution of lysyl oxidases. *Sci. Rep.* **5**, 10568; doi: 10.1038/srep10568 (2015).

3.6. Dynamics of genomic innovation in the unicellular ancestry of animals

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*. Previous analyses of animal unicellular relatives uncovered the premetazoan origin of many genes with multicellularity-related functions, *e.g.* developmental transcription factors or cell adhesion proteins. Here we show that genome architecture evolution was equally dynamic: an early burst of gene diversity in the holozoan ancestor was followed by independent episodes of synteny disruption, intron gain, and genome expansions in both unicellular and multicellular lineages. These punctuated innovations shaped the genomic prehistory of Metazoa, and offer a glimpse of the evolutionary trends shared by ancient and extant animal genomes.

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), Passeig Marítim de la Barceloneta, 37-49, 08003, Barcelona.

2. Departament de Genètica, Microbiologia i Estadística, Universitat de Barelona, Avinguda Diagonal 643, 08028, Barcelona.

3. Unité d'Ecologie, Systématique et Evolution, CNRS UMR 8079, Université Paris-Sud/Paris-Saclay, AgroParisTech, 91400 Orsay, France.

4. Department of Microbiology, University of Hawai'i at Mānoa, Snyder Hall, 2538 McCarthy Mall, Honolulu, HI 96822, USA.

5. Advanced Studies in Genomics, Proteomics and Bioinformatics, University of Hawai'i at Mānoa, Honolulu, Hawai'i, USA.

6. Faculty of Life and Environmental Sciences, Prefectural University of Hiroshima, Nanatsuka 562, Shobara, Hiroshima, 727-0023, Japan

7. Department of Biosciences, University of Exeter, Exeter, EX4 4QD, UK.

8. ICREA, Passeig Lluís Companys 23, 08010, Barcelona.

* To whom correspondence may be addressed: inaki.ruiz@ibe.upf-csic.es, xavier.graubove@gmail.com

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. *Previous analyses of animal unicellular relatives uncovered the premetazoan origin of many genes with multicellularity-related functions, e.g. developmental transcription factors or cell adhesion proteins. Here we show that genome architecture evolution was equally dynamic: an early burst of gene diversity in the holozoan ancestor was followed by independent episodes of synteny disruption, intron gain, and genome expansions in both unicellular and multicellular lineages. These punctuated innovations shaped the genomic prehistory of Metazoa, and offer a glimpse of the evolutionary trends shared by ancient and extant animal genomes.*

Keywords: holozoans, ichthyosporeans, comparative genomics, animal origins, gene family evolution, introns

I. Introduction

The transition from a unicellular organism to the first multicellular animal, more than 600 million years ago [1,2], 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 [3]. 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 [4–8], with their closest unicellular relatives in the Holozoa clade, such as the choanoflagellates *Monosiga brevicollis* and *Salpingoeca rosetta* [9,10], and the filasterean *Capsaspora owczarzaki* [11] (Fig. 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.

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 [10–15]. 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* [16,17], and clonal in *Salpingoeca* [10]. Thus, gene content has been extensively studied in the animal unicellular ancestry, but less attention has been devoted to the evolutionary dynamics of genome structure.

The evolution of genome size, non-coding regions, intron creation, and synteny conservation have been thoroughly explored within Metazoa but not in unicellular holozoans [9,18]. 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 analysed [19]; and the same happened with the ancestral animal diversity of cadherin and integrin adhesion systems before genomes from choanoflagellates and *Capsaspora* were analyzed [20,21]. 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* [22].

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 [23,24], and *Corallochytrium* develops colonies by binary, palintomic, cell division [25]. In both cases, completion of the cell cycle frequently involves release of propagules that restart the clonal proliferation [23,24]. 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 [26,27].

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* [22,27] (Ichthyosporea), and to the afore-mentioned *Salpingoeca rosetta*, *Monosiga brevicollis* (choanoflagellates) and *Capsaspora owczarzaki* (Filasterea), totaling 10 unicellular holozoan genomes (Fig. 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 structural 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.

II. 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 [28]. Gene annotation was performed using a

combination of *de novo* gene predictions and transcriptomic evidence derived from RNA sequencing experiments (see Methods). Together with *Capsaspora*, the two choanoflagellates and three already available ichthyosporeans, our dataset comprises 10 genomes from all unicellular Holozoa lineages – eight more than previous genome analyses [11,10]. Figure 1 summarizes the assembly and annotation statistics of all 10 unicellular holozoan genomes.

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 [22]. 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 (Fig. 2; BS=100%, BPP=1). Therefore, *Chromosphaera,* isolated from shallow marine sediments in Hawaii, is the first described putatively free-living dermocystid Ichthyosporea. Indeed, all described dermocystids are strict vertebrate parasites, 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) [29,30].

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* [25], are a sister-group to all the other holozoans (filastereans, choanoflagellates and animals) with improved statistical support (Fig. 2; BS=100%, BPP=0.85). Our analysis confirms the monophyly of Teretosporea, rejecting alternative scenarios such as the "Filasporea" hypothesis (a grouping of Filasterea + Ichthyosporea) [31,32], 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 [18,33] and most unicellular holozoans have relatively compact genomes, like *Corallochytrium*, *Capsaspora* or *Chromosphaera* (24.1, 27.9 and 34.6 Mb, respectively) (Fig. 1A). 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) (Fig. 1A) [7,18].

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) [18], 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 [34]. Indeed, 5-9% of the genome of *Salpingoeca*, *Sphaeroforma*, *Abeoforma* and *Pirum* are covered by TEs, whereas other holozoans are below 2.5% (Fig. 3A). A detailed examination of the TE complement of *Salpingoeca*, *Pirum* and *Abeoforma* revealed species-specific small sets of TE families, sharing high sequence identity, that accounted for the vast majority of copies (Fig. 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 (Fig. 3B). In *Abeoforma* and *Pirum*, TEs and other simple repeats comprised up to 17-34% of the genome, accompanied by unusually AT-biased nucleotide compositions (Fig. 1A). As a result of their highly repetitive genomes, partial gene models were frequent in *Pirum* and *Abeoforma* (Fig. 1A-Supplement 1). Consequently, they were excluded from comparative analyses with animals.

Finally, the smaller genomes of *Corallochytrium* and *Chromosphaera* were largely depleted of repetitive/satellite regions and TEs. This finding, together with their reduced intron content (see below, Fig. 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 [35,36]. Following this reasoning, we analyzed the microsyntenic gene pairs of unicellular holozoan genomes (Fig. 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 contrast, the analysis of microsynteny in *Capsaspora* revealed remarkable across-lineage conservation with the distant teretosporeans *Chromosphaera* and *Coral-*

lochytrium (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* (Fig. 3C-Supplement 1). 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 [21,27]. 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 [16], a typical cause of microsynteny conservation in animals [36]. Overall, gene linkage of most extant holozoans appears to be markedly different from their common ancestor, with specific gene pairings arising in Metazoa [35,36], 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) (Fig. 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.

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

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) [37,38]. The high intron density of multicellular animals has been linked to their higher organismal complexity, as it enables frequent alternative splicing and richer transcriptomes [39–41], provides physical space for transcription regulatory sites [42,43], and facilitates the diversification of gene families by exon shuffling [44]. 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 [37,45].

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 (Fig. 4A). Moreover, *Creolimax* and *Ichthyophonus* harbor longer introns than most protistan eukaryotes, similar in length to those of some animals (Fig. 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 *et al.* [46] (Fig. 4C). 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 (Fig. 4D-E). 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 [37,39]) (Fig. 4E). 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 (Fig. 4D-E). Prolonged periods of intron gain are uncommon in eukaryotes and, in the case of Metazoa, this phenomenon has been linked to inefficient purifying selection due to low effective population sizes [37,47,48]. Whether this is the case for the intron-rich *Creolimax*, *Sphaeroforma* and *Ichthyophonus*, yet, remains an open question. Estimates of population size for another symbiotic ichthyosporean, *Pseudoperkinsus tapetis*, are in the 10^6 to 10^7 range [49] – closer to most unicellular eukaryotes than to animals [50].

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 (Fig. 5; 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 intron sharing: 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 (Fig. 5). 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 (Fig. 4D), and 85.5% of its introns are of holozoan or earlier origin (Fig. 4F). Interestingly, introns with regulatory sites from *Capsaspora* (identified by [43]) have a similar, ancestral-biased, age distribution (Fisher's exact test, p-value=1; Fig. 4F). This hints at a decoupling between the evolutionary dynamics of introns and regulatory sites, despite sharing physical space in the genome.

Timing of gene family diversification in Holozoa

The *Monosiga* genome paper by King *et al.* [9] 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 [51–53], this process is specifically credited with the diversification of many gene families involved in complex signaling and/or multicellular integrated lifestyle in Metazoa [14,18,21,54–58].

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 [46] to reconstruct the content of specific protein domain architectures in ancestral genomes (available as Source Data SD7). 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) (Fig. 6A-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% (Fig. 6A). 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 (Fig. 6-Supplement 1) [59].

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 (Fig. 6A and 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 (Fig. 6A, 6-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 [11,15,12,56,60]. 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 (Fig. 6B, 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 ≥90% probability for the ancestral reconstructions; Methods and Fig. 6C). 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'.

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 [61,62]) but they were generally lower in animals than in their unicellular relatives and ancestors (Fig. 6D). Second, there was a strong negative relationship between modularity and the number of protein domains per gene family (Spearman's rank correlation coefficient, ρ_s=-0.96, p<0.001, Fig. 6E). Therefore, at the genome 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 [18,51]. This same effect was observed when we analyzed subsets of the proteome networks sharing a common function—for example, protein domains related to ECM, signaling, ubiquitination or protein-binding (with ρ_s in the range -0.32 to -0.84 and p<0.001; Fig. 6-Supplement 2).

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; Fig. 6D). Also, there was no negative relationship between the number of domains per community and the network modularity (ρ_s=0.12, p-value=0.32), meaning that the addition of new domains to TF genes occurred in a gene family-specific manner (Fig. 6E). This implies that the expanded TF repertoires of animal genomes [15] 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 recovered many examples of gene family-specific domain diversification in novel animal TFs (Table 1): 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 [15,19,63,64]. These TF families appeared as isolated clusters when we sorted protein domains by their pattern of co-occurrence in the reconstructed Metazoa LCA (Fig. 7A). Furthermore, we detected an unexpected premetazoan origin for some TF classes as per their domain combinations (Table 1). 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 [65]. However, we identified LIM-associated homeobox genes in multiple ichthyosporeans, *Corallochytrium* and *Capsaspora*. We classified these candidate genes according to HomeoDB [66] using [63] 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* (Fig. 7C). 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 animal from *Caenorhabditis elegans* to mammals or the sea walnut *Mnemiopsis* [67–69]. Given that transcriptionally regulated cell type specification has already been demonstrated in *Creolimax* [27], 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* [70]. Most eukaryotes have a consensus architecture composed of a central HAT/KAT11 domain (acetylase) flanked by 3 zinc fingers of TAZ (2) and ZZ (1) types (DNA-binding motifs) (Fig. 7D). 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 [19]. Intriguingly, *Capsaspora*'s epigenome contains p300/CBP-specific histone acetylation marks, but its relatively compact genome lacks distal enhancers [43].

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 (Fig. 7B). This pattern explains the lower network modularity of animal ECM genes (Fig. 6D-E). 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 [71]. 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, Fig. 7B).

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 [56]. 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 to Porifera); there are also non-structural genes like collectin receptors (Lectin-C) or the C1q complement subcomponent (C1q) [56,72–75]. Most collagen genes appeared and expanded in Metazoa, concomitantly with the ECM structures they associate with [56]. We found, however, a remarkable exception: a canonical type IV collagen gene in the filasterean amoeba *Ministeria vibrans*. 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) [56]. This *Ministeria* ortholog is composed of a pair of C4 domains at the C-terminus and multiple collagen 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* (Fig. 7E). Thus, a type IV collagen was already present in the common ancestor of Filasterea, Choanoflagellata and Metazoa.

III. 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 density 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 (Fig. 6B), 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 [14,58,76], cell adhesion systems [20,21,27] and transcription factors, often involved in developmental processes [15,19]. The expansion of protein diversity in early holozoans provided fertile ground for the frequent co-option of ancestral genes for multicellular functions in Metazoa [12]. Overall, our probabilistic reconstruction of the genome content of unicellular animal ancestors (available as Source Data SD7) 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) (Fig. 7C-E), and canonical Type IV collagens, a key element of the animal ECM [56] (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*). Interestingly, the independent origin of intron-dense genomes in animals and ichthyosporeans is mirrored by two different modes of alternative splicing of transcripts dominating in each clade. In animals, exon skipping is a common mechanism of transcriptome expansion by isoform creation [40,77]. In *Creolimax* and *Capsaspora*, however, exon skipping is rare: most of their alternatively spliced transcripts originate by dysfunctional intron retention [16,27]. The dominance of intron retention in the early Holozoa, therefore, makes their alternative splicing profiles more similar to the putative ancestral eukaryotic mechanisms than to Metazoa [77].

The emergence of larger genomes in Metazoa, however, cannot be explained solely by intron gain and gene family expansion [33]. Unfortunately, other factors such as the contribution of TE invasions (Fig. 3B) or the extension of intron sites, are not possible to date at the holozoan-wide evolutionary scale due to the lack of con-

served 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 C*apsaspora* (median intergenic regions: 373 bp) has intragenic *cis*-regulatory elements key to its temporal regulation of cell differentiation [43], but the putative regulatory functions in the larger intergenic regions of *Creolimax*, *Spaheroforma* and *Salpingoeca* (median intergenic 900-1200bp) remain uncharacterized. It is tantalizing to note that 1) *Creolimax* and *Salpingoeca* exhibit temporal differentiation of cell types [10,27] , and 2) their intergenic median sizes are in line with those of *Amphimedon* (885bp) (Source Data SD1), a demosponge with bilaterian-like promoters and enhancers, including distal regulation [70,78]. However, the ancestral gene linkages conserved across Metazoa, frequently due to common *cis*-regulation [36], appear to be animal innovations absent in unicellular holozoans (Fig. 3-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 [12]). 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 [18].

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 (Fig. 3D), yet choanoflagellates are more similar to the animal ancestor in terms of intron structure (Fig. 5) and have lower rates of protein domain diversity loss (Fig. 6D). 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 (Fig. 3C). However, *Capsaspora*'s lack of close sister groups hampers comparative studies of faster-evolving genomic features, be it the regulatory circuitry [43], or co-option of genetic tool-kits for its unique aggregative development [16].

The seven 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 (Fig. 5) and is relatively similar to animals in terms of coding sequence conservation (Fig. 3D), but it harbors a secondarily expanded genome with disrupted gene linkage (Fig. 3A, C). Another example is *Corallochytrium* and *Chromosphaera*, both with massive simplifications of intron content (Fig. 4D), but higher synteny conservation with the inferred ancestral Holozoa (Fig. 3C). Also, the diversity of protein domain combinations of *Chromosphaera* is the highest among ichthyosporeans (in line with values of animals and holozoan ancestors; Fig. 6A) and phylogenetic distances to animal orthologs are comparatively low (Fig. 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 [30], 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 [27]. 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 (Fig. 3A-C).

Genomic plasticity in the animal ancestry

The genomes of extant Metazoa are subject to overlapping evolutionary dynamics for different traits, from gene family expansions and depletions to conservation of gene structure and local order [18]. Overall, our analyses show how these processes extend back to the unicellular prehistory of Metazoa: we reconstruct conservation patterns between animals and their direct ancestors, and differential effects on their unicellular relatives, for instance, the shared and independent protein diversifications and intronization events. Such rich evolutionary dynamics in premetazoan genomes mirrors the premetazoan origin of various key multicellularity-related genes, which is accompanied by unicellular- and multicellular-specific expansions [11,9,10,15,13,12,14]; and by the plasticity of cell types proposed for ancestral holozoans [27,79]. Consequently, we see how the genomes of ancestral premetazoans were subject to the same processes observed in most animal phyla: a thorough exploration of the genomic space, and no trait left to tinker with.

IV. 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 25ml for 5 days. DNA was extracted using a standard protocol by which cells were lysed in the extraction buffer composed of Tris-HCL, 50mM EDTA, 500mM NaCl and 10mM ß-mercaptoethanol. DNA was purified with phenol:chlorophorm: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 125bp ($250{\cdot}10^6$ reads, 250bp insert size), MP 50bp ($108{\cdot}10^6$ reads, 6kb fragment size); *Abeoforma*, PE 100bp ($73{\cdot}10^6$ reads, 600bp insert size), MP 100bp ($41{\cdot}10^6$ reads, 6kb fragment size); *Chromosphaera*, PE 125bp ($143{\cdot}10^6$ reads; insert size 250bp), MP 50bp ($114{\cdot}10^6$ reads, 5kb fragment size); and *Corallochytrium*, PE 100bp ($150{\cdot}10^6$ reads, 420bp insert size), MP 100bp ($47{\cdot}10^6$ reads, 3kb 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 25ml 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 125bp with an insert size of 250bp, 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 [80] and trimmed accordingly with Trimmomatic v0.33 [81] 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 [28] with the BayesHammer error correction algorithm [82]. For each organism, PE data was analyzed using Kmergenie [83] 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 [84]. 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 [85]. Genome assembly statistics (genome size, N50, L75) were calculated using Quast v2.3 [86], and completeness was assessed using the BUSCO v1.1 [87] 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 [88] and PASA v2.0.2 [89,90]), plus complementary *ab initio* predictors (based on hidden Markov models for gene structure: GeneMark-ES v4.21 [91] and SNAP [92]). These results were combined to produce a consolidated gene annotation using Evidence Modeler v1.1.1 [90].

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 [22]) [93,94]. 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 [22] (*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 [95,96]. 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 [97], 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.* [22], adding nine single-copy protein domains to the previous version (which included 78 alignments) according to the methodology developed in [98].

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 [22].

We used the BVD57 dataset to build ML phylogenetic trees using IQ-TREE v1.5.1 [99], under the LG model with a 7-categories free-rate distribution, and an frequency mixture model with 60 frequency component profiles based on CAT (LG+R7+C60) [100]. 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 [100]. Statistical supports were drawn from 1,000 ultrafast bootstrap values with a 0.99 minimum correlation as convergence criterion [101] and 1,000 replicates of the SH-like approximate likelihood ratio test [102], 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). A near-perfect correlation has been found for ultrafast bootstraps and regular non-parametric bootstraps for the LG+R7+PMSF ML analysis, thus validating our approach (Fig. 2).

We then used the same alignment to build a Bayesian inference tree with Phylobayes MPI v1.5 [103], using the LG exchange rate matrix with a 7-categories gamma distribution and the non-parametric CAT model [104] (LG+Γ7+CAT). A Γ7 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 1,231 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 Source Data SD10.

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 [22,105,106] and our own results. We remained agnostic about the internal topology of SAR [107], Fungi [22], the contentious hypotheses for the root of eukaryotes (namely, "Opimoda-Diphoda" or "Excavata-first") [105,106] and the earliest-branching animal group (Porifera or Ctenophora) [108]. 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 [109]. 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.*, 6 ichthyosporeans, 2 choanoflagellates, *Corallochytrium* and *Capsaspora*) and yielded 58,516 orthoclusters, 11,925 of which contained >1 gene.

Phylogenetic analysis of gene families

Retrieval of homologous protein sequences was performed by querying protein domain HMM profiles (as defined in the 29th release of Pfam [110]) against a database of protein sequences from 69 selected eukaryotic genomes and transcriptomes (Source Data SD10). Each of the following gene families was defined by its catalytic/representative protein domain: type IV collagen (PF01413) and TAZ zinc finger TFs with HAT/KAT11 domains (PF08214). 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 [66].

Protein alignments were built with MAFFT v7.245 [111], using the G-INS-i algorithm optimized for global homology. 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 (IQ-TREE v1.5.1) [99] and Bayesian inference (MrBayes v3.2.6) [112]. The optimal evolutionary models for each alignment were selected using ProtTest v3.4 [113], yielding LG+Γ4+i as the best model for the Collagen IV, HAT/KAT11 and LIM Homeobox phylogenies. For IQ-TREE [99], 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 1,000 ultra-fast bootstrap [101] replicates with a 0.99 minimum correlation as convergence criterion, and 1,000 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 1,000 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.

Analysis of repetitive elements

Repetitive regions were annotated in the genomes of all unicellular holozoans using RepeatMasker open-4.0.5 [114] and annotations from the 20150807 release of GIRI RepBase database [115]. We used the slow, high-sensitivity search with the Eukaryota-specific database 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 [4]). To account for spurious conservation of gene order, we assigned random positions to each gene using the bedtools v2.24.0 *shuffle* utility [116] 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 [117]). 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 [118].

Analysis of coding sequence conservation

From our ortholog database using 40 eukaryotic genomes (excluding *Pirum* and *Abeoforma*), 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 [111], trimmed the alignments using trimAL automated algorithm [119], and inferred maximum likelihood trees for each ortholog group using RAxML v8.2.0 [120] 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 [121,122]. 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 [122].

Comparative analysis of intron content

Intron content of a subset of 40 eukaryotic genomes (excluding *Abeoforma* and *Pirum*) was analyzed using a set of single-copy orthologous genes, and used to reconstruct ancestral states as described by Csűrös *et al.* [37,123,124]. 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 [111] 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.* [46,125], starting from the standard null model, running 1,000 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 [46]). 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 1 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 [117]). 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 [122]. 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 [43], 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 (≥50% ATAC site overlap with the intron sequence, calculated using bedtools v2.24.0 *intersect* utility [116]), 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*) were computed using Pfamscan and the 29th release of the Pfam database [110]. 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 [46] as implemented in Count [126]. 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 [127]. A maximum value of 1 indicate a strong community structure, while a minimum value of 0 indicate 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 [122,128]. Function-oriented domain subnetworks were obtained by retrieving orthologous groups that contained relevant domains, which were obtained from previous studies (transcription factors from [15,129], signaling domains from [12], ECM-related domains from [12,21,56], ubiquitination from [58]) and pfam2go annotations (for the subsets mentioned above, and also for protein-binding domains) [130]. Monotonic statistical dependence between modularity and the number of domains per community was tested using Spearman's rank correlation coefficient (ρ_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 [110]. 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 [126].

V. Author contributions

XGB and IRT designed and coordinated the study. TAR, XGB, GL and HS were in charge of the assembly and annotation of the genomes. XGB performed the comparative analyses and ancestral reconstructions. SD provided data for the phylogenomic analysis of *Chromosphaera perkinsii*, which was performed by XGB and GT. XGB wrote the manuscript. All the authors critically reviewed and approve the final manuscript.

VI. Accession numbers

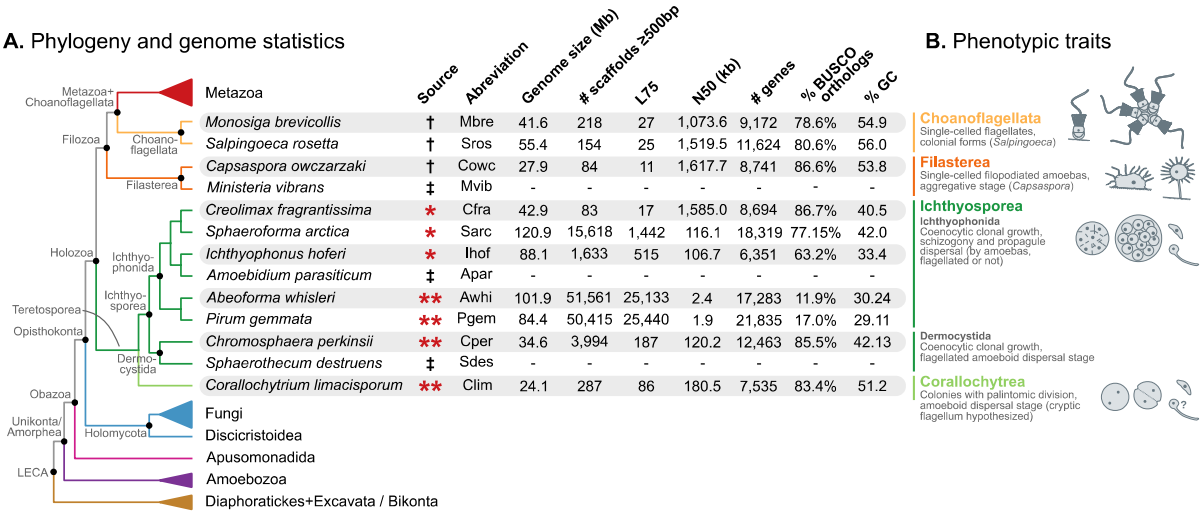
Genome sequencing and assembly data from *Corallochytrium*, *Abeoforma, Pirum* and *Chromosphaera* has been deposited in NCBI using the BioProject accession PRJNA360047. Transcriptome sequencing data from *Abeoforma* and *Chromosphaera* has been deposited in NCBI using the BioProject accession PRJNA360056.

VII. 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.

VIII. Figures

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 [11,9,10]; ‡ organisms for which transcriptomic data exists but no genome is available [22]. **B)** Overview of the phenotypic traits of each group of unicellular Holozoa, focusing on their multicellular-like characteristics. For further details, see [22–24,30]. Source Data SD1, SD8.



	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-Supplement 1. Comparisons of gene length of one-to-one orthologs from pair-wise comparisons of all 10 unicellular Holozoa. Dots around the diagonal lines indicate that orthologs from both organisms have identical lengths. Note that *Abeoforma* and *Pirum* have abundant incomplete orthologous sequences.

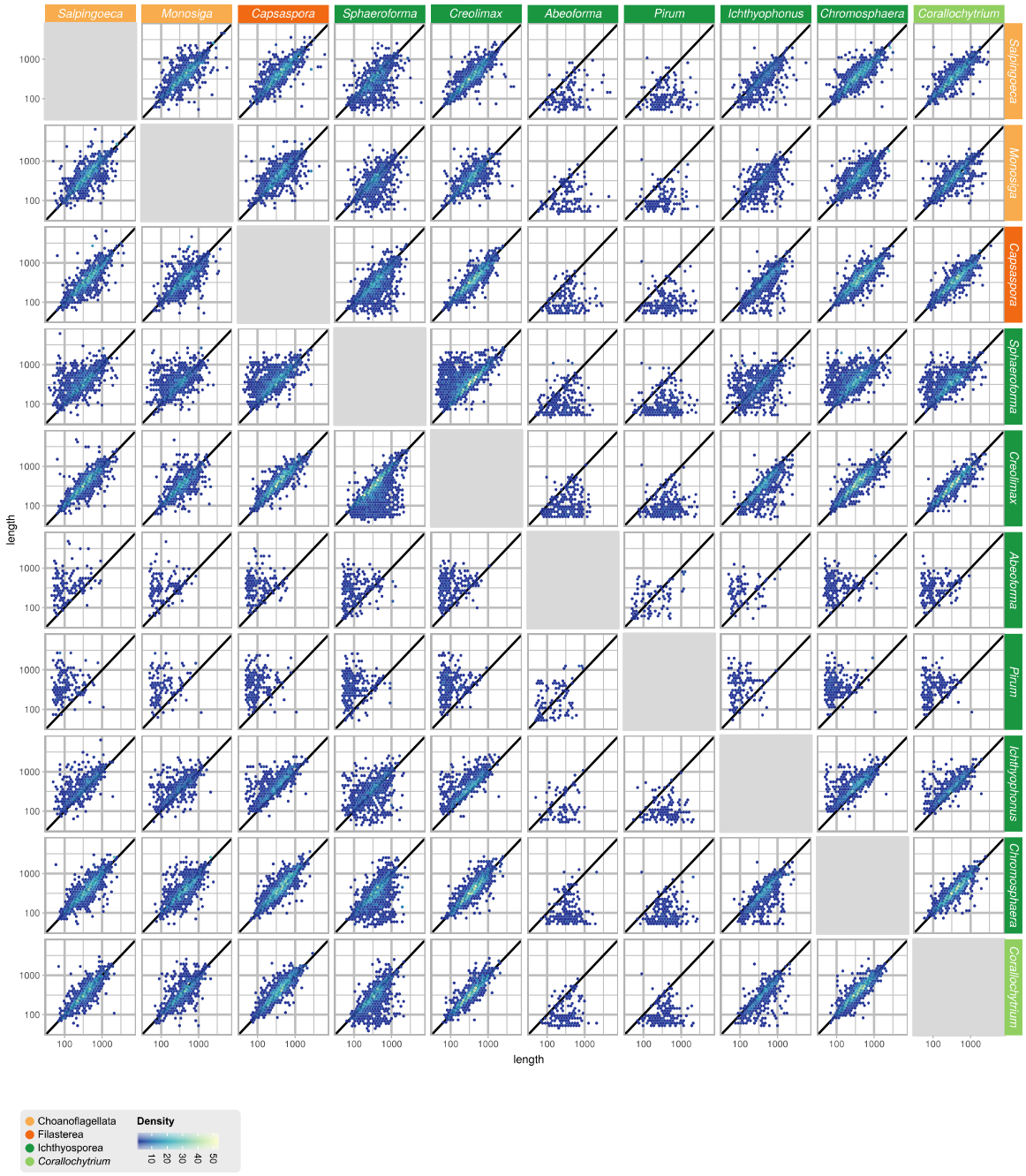


Figure 2. Phylogenomic tree of Unikonta/Amorphea. Phylogenomic analysis of the BVD57 taxa matrix. Tree topology is the consensus of two Markov chain Monte Carlo chains run for 1,231 generations, saving every 20 trees and after a burn-in of 32%. Statistical supports are indicated at each node: i) non-parametric maximum likelihood ultrafast-bootstrap (UFBS) values obtained from 1,000 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-Supplement 1 for raw trees with complete statistical supports. Source Data SD11.

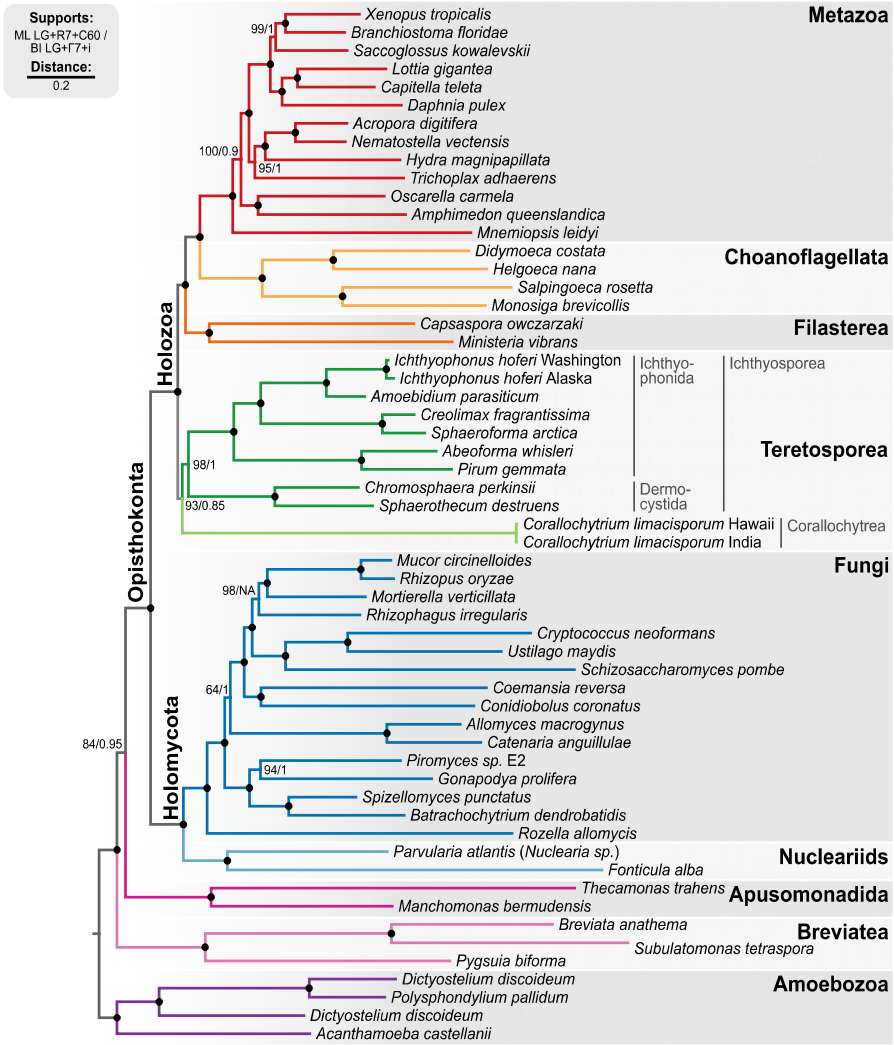
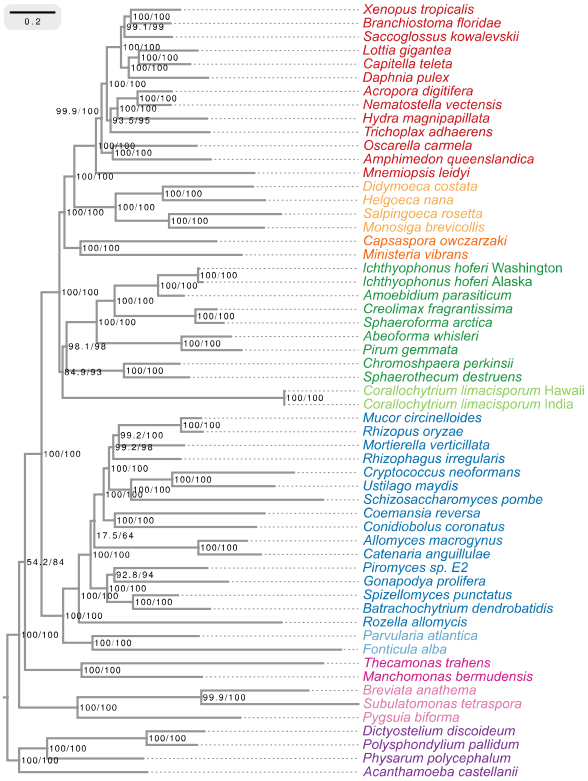
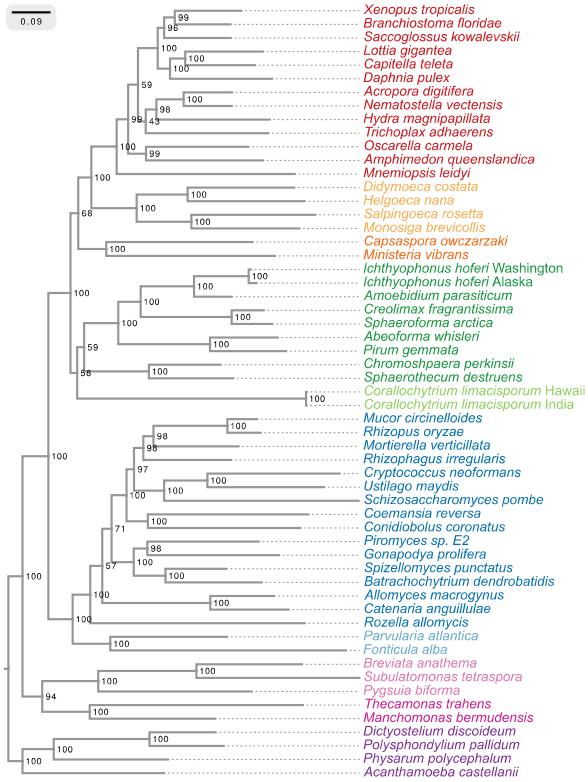


Figure 2-Supplement 1. Phylogenomic analysis of the BVD57 matrix using **A)** IQ-TREE maximum likelihood and the LG+R7+C60 model (supports are SH-like approximate likelihood ratio test / UFBS, respectively); **B)** IQ-TREE maximum likelihood and the LG+R7+PMSF model (fast CAT approximation; non-parametric bootstrap supports); and **C)** Phylobayes Bayesian inference under the LG+Γ7+CAT model (BPP supports).

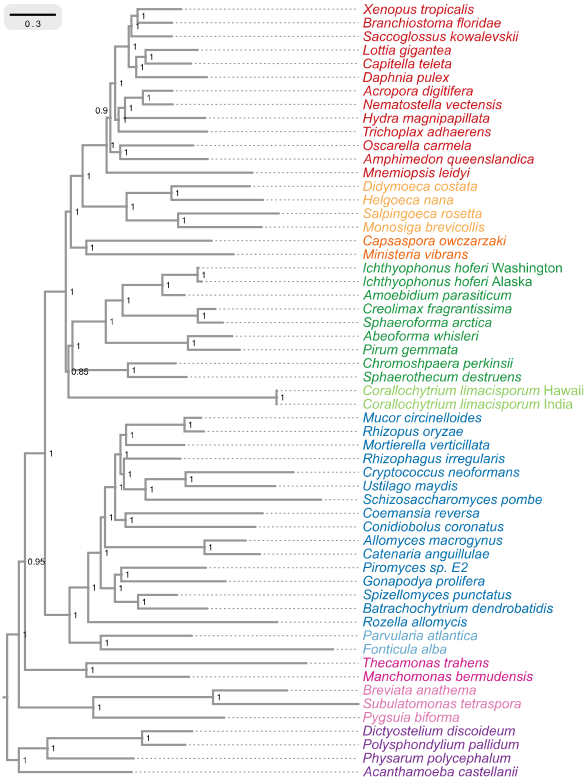
A. Phylogenomic analysis: BVD57, ML (LG+R7+C60)



B. Phylogenomic analysis: BVD57, ML (LG+R7+PMSF)



C. Phylogenomic analysis: BVD57, BI (LG+Γ7+CAT)



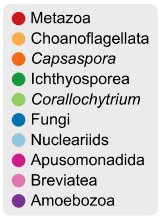


Figure 3. Patterns of genome evolution across unicellular Holozoa. A) Genome size and composition in terms of coding exonic, intronic and intergenic sequences of unicellular holozoan and selected metazoans. Percentage of repetitive sequences shown as black bars. Genome size of the Metazoa LCA from [18]. 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. Source Data SD1, SD2, SD3.

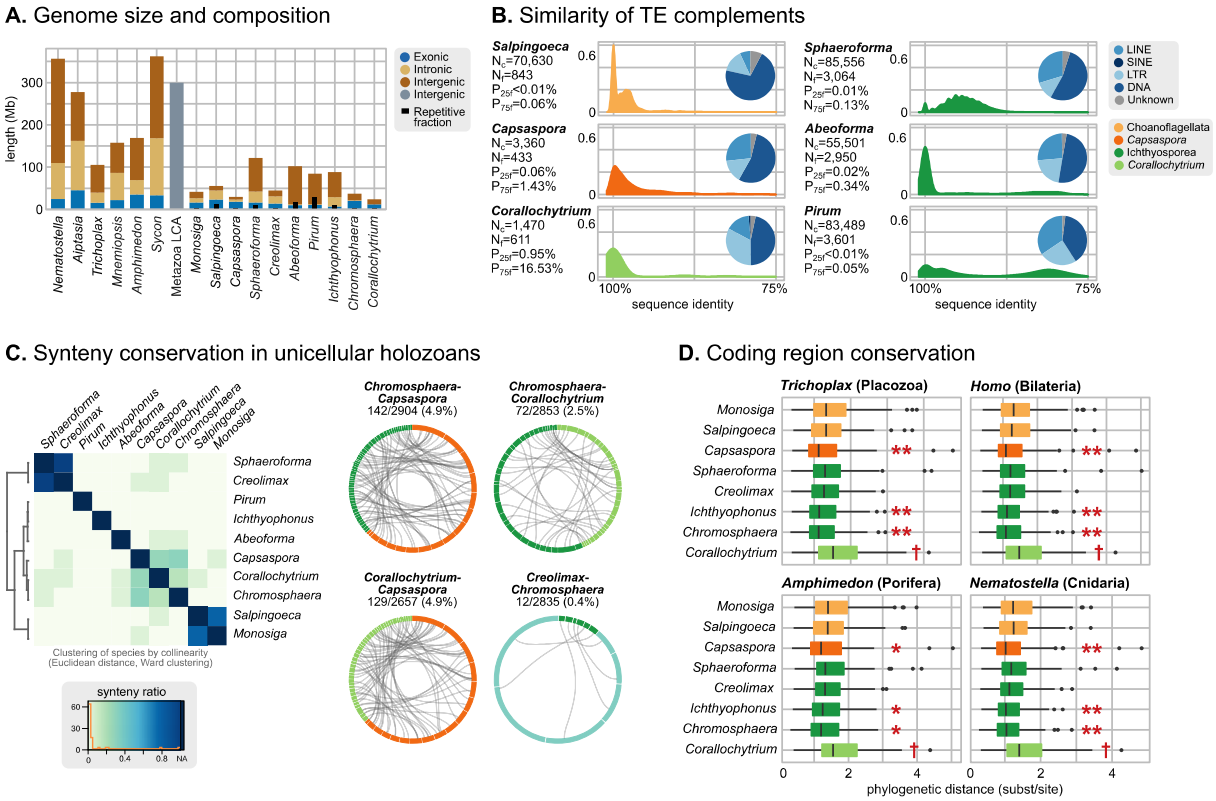


Figure 3-Supplement 1. A-J) Profile of transposable element (TE) composition of 10 unicellular Holozoa, including i) distribution of sequence similarity frequencies within the TE complement obtained from BLAST alignments (minimum 70% identity and 80-bp alignment length); ii) same data but using density-normalized plots; and iii) raw counts of hits for each TE family, indicating the number of families with hits (NFH) for each species. Each third panel illustrates how TE complements can be biased towards a handful of families with a high number of similarity hits (e.g. Monosiga or Pirum) or, conversely, exhibit even distributions (e.g. Corallochytrium). K. Heatmap of pairwise ratios of ortholog collinearity between 10 unicellular holozoan genomes. Species are manually ordered by taxonomic classification (no clustering).

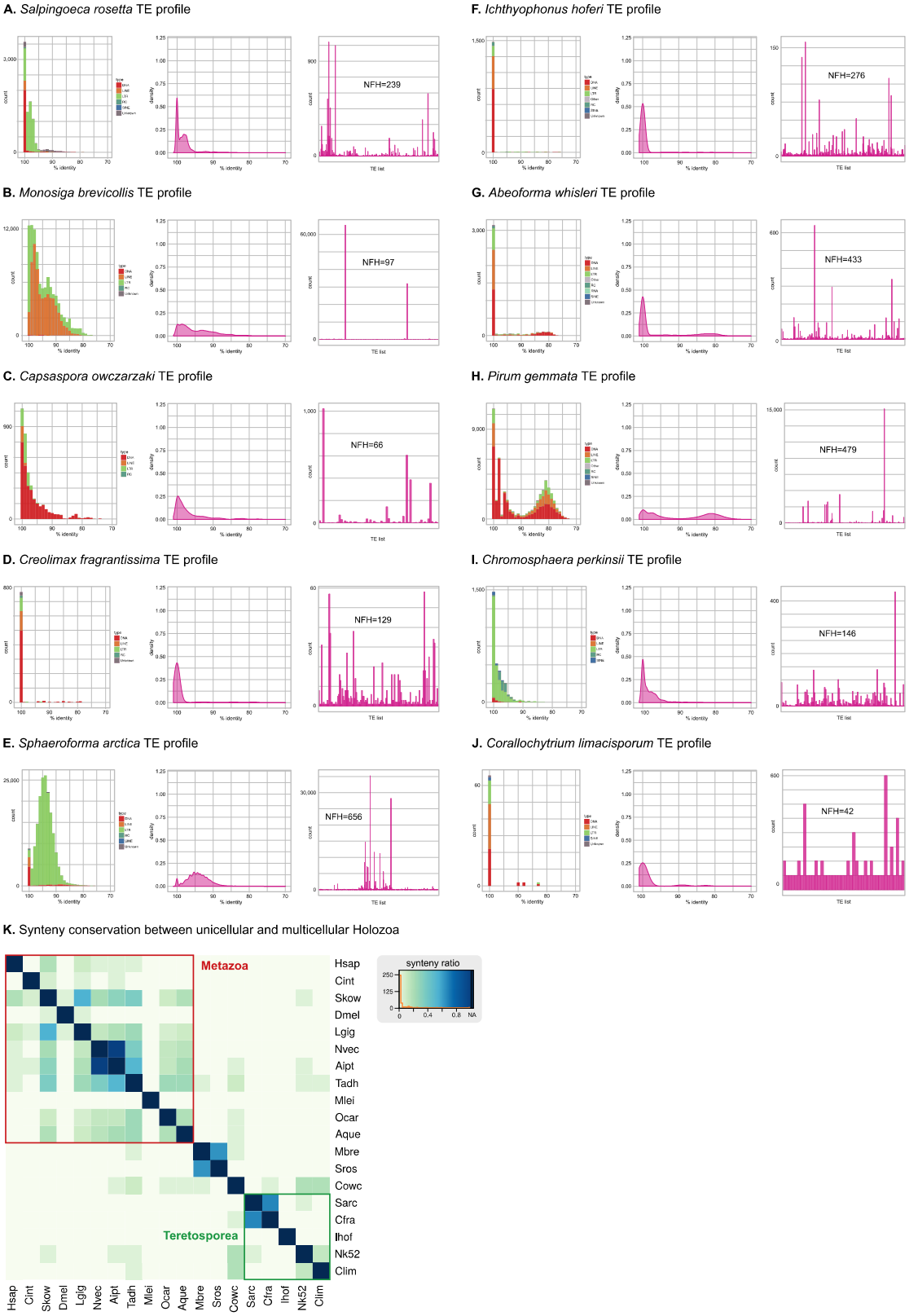


Figure 4. Intron evolution. A) Distribution of intron lengths and number of introns per gene in selected eukaryote genomes. Dots represent median intron lengths and vertical lines delimit the first and third quartiles. Color code denotes taxonomic assignment. Species abbreviation as in Figure 1 and SD10. **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. **C)** Classification of intron sites by conservation in protein alignments, as used in [46,125]. Grey boxes denote aligned amino acids with gaps (dashed lines). Intron sites (vertical lines) are conserved if they are present in various organisms at the same alignment position and codon phase. The method accounts for loss of intron sites (red crosses), independent gains at the same site (different codon phases), ambiguous sites (in poorly-aligned regions) and unclassifiable sites (non-homologous regions). **D)** 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 tree edges mark the continued process of dominant intron gains between the LECA and Metazoa/Ichthyophonida. Red and green bars represent the inferred number of intron gains (green) and losses (red) in ancestral nodes. E) 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). F) Phylostratigraphic analysis of the origin of *Capsaspora* introns, considering all sites (left) and those with putative regulatory sites (right; after [43]). Source Data SD4, SD5.

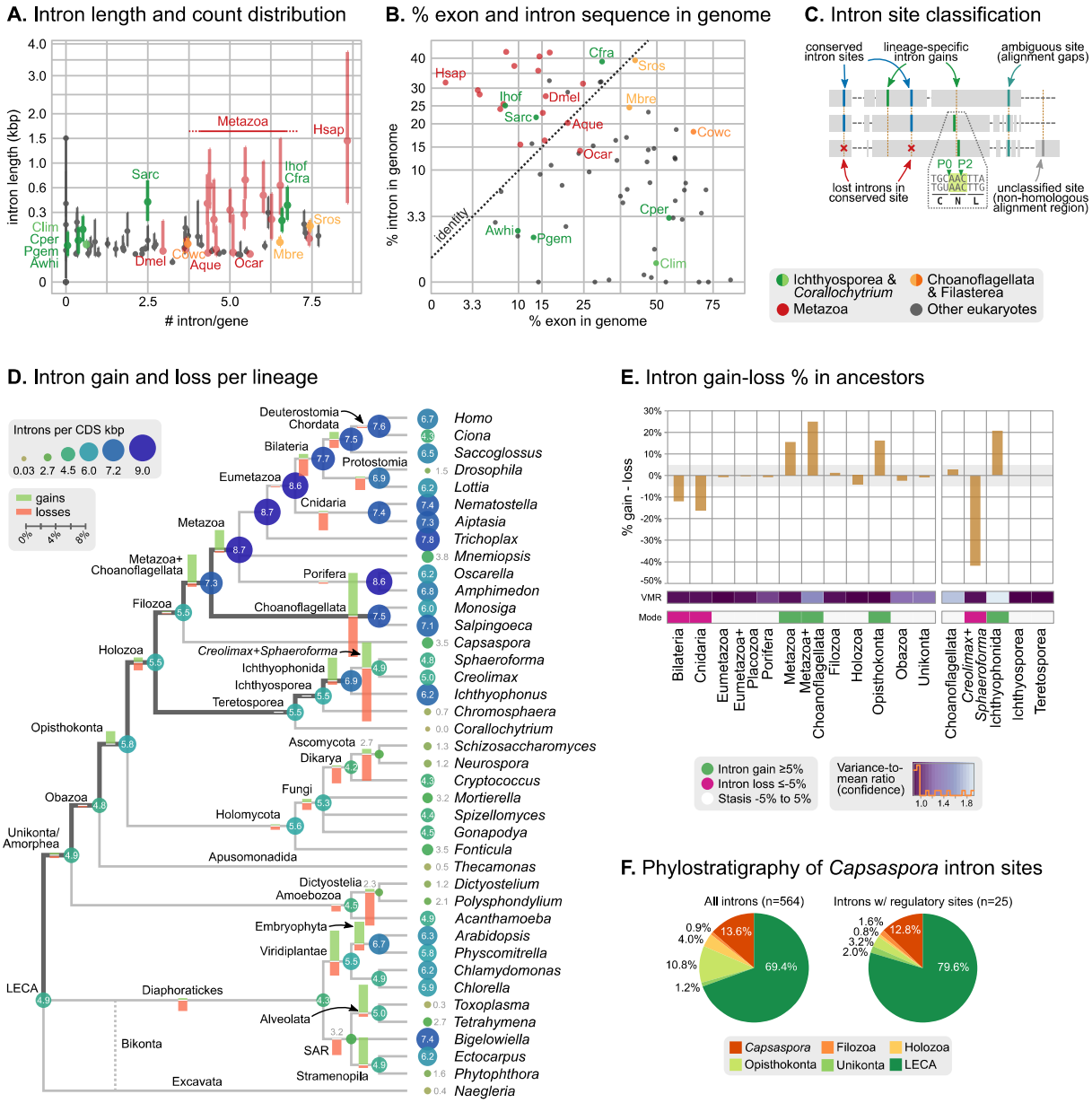


Figure 5. Profile of intron site presence across eukaryotes. Heatmap representing presence/absence of 4,312 intron sites (columns) from extant and ancestral holozoan genomes, plus the line of ascent to the LECA (rows). Intron sites and genomes have been grouped according to their respective patterns 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. Source Data SD5, SD6, SD7.

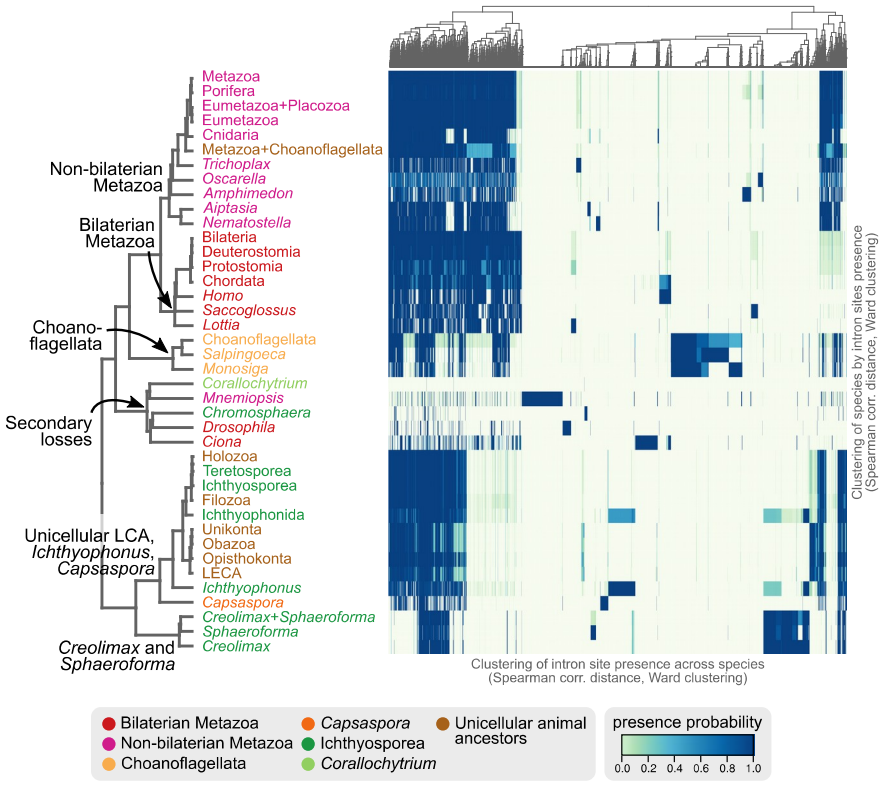


Figure 6. Evolution of protein domain architectures. A) Protein domain combination gain and loss per lineage, including extant genomes and ancestral reconstructed nodes. Diameter and color of circles denote the number of different domain combinations (in different gene families) in that node of the tree. Bolder tree edges mark the continued process of dominant intron gains between the Opisthokonta and Bilateria LCAs. 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). **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). **D and E)** Modularity and community size of the global network of domain pairs (upper panels) and the TF subnetwork (lower panels). 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 Methods and [127]). 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 rewiring 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 (ρ_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. Source Data SD5, SD6, SD7.

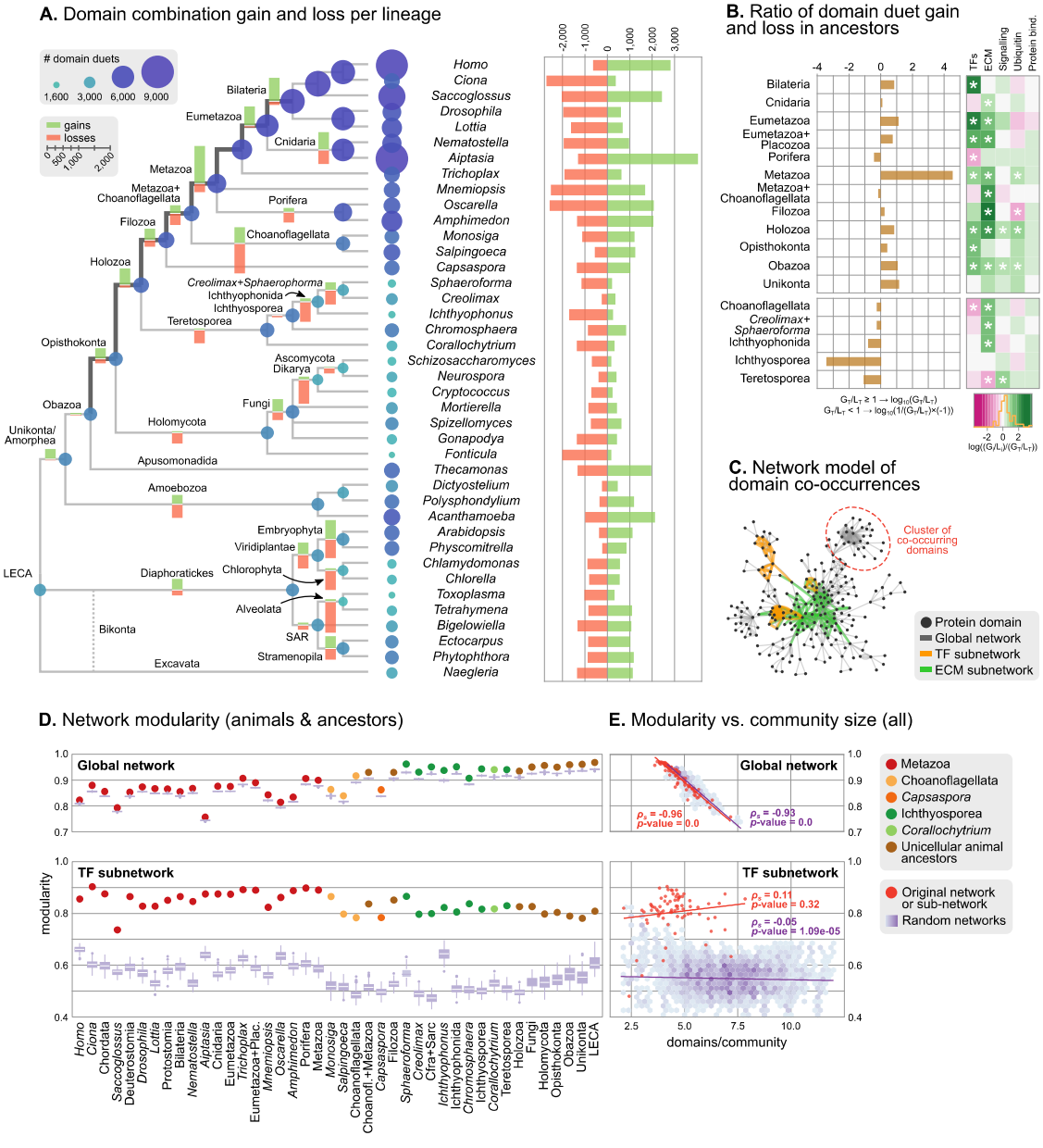


Figure 6-Supplement 1. A) Ancestral reconstruction of gains (green) and losses (red) of protein domains per lineage, based on Dollo parsimony. Note that, in contrast with the evolution of protein domain combinations here most nodes are dominated by losses. The most notable exceptions are the Metazoa and Opisthokonta LCAs (dominated by gains) and its unicellular ancestors from Holozoa to Choanoflagellata+Metazoa LCAs (in which gains and losses are in a relative stasis). Source data SD9. **B)** Log-ratio of gains-to-losses for single protein domains (brown bars) and protein domain pairs (blue bars), based on the respective ancestral reconstructions. Positive values thus mean that gains are greater than losses. Loss of single protein domains dominates in most nodes, but gains in protein domain combinations can nevertheless outweigh losses in many ancestors (e.g., Holozoa or Metazoa LCAs).

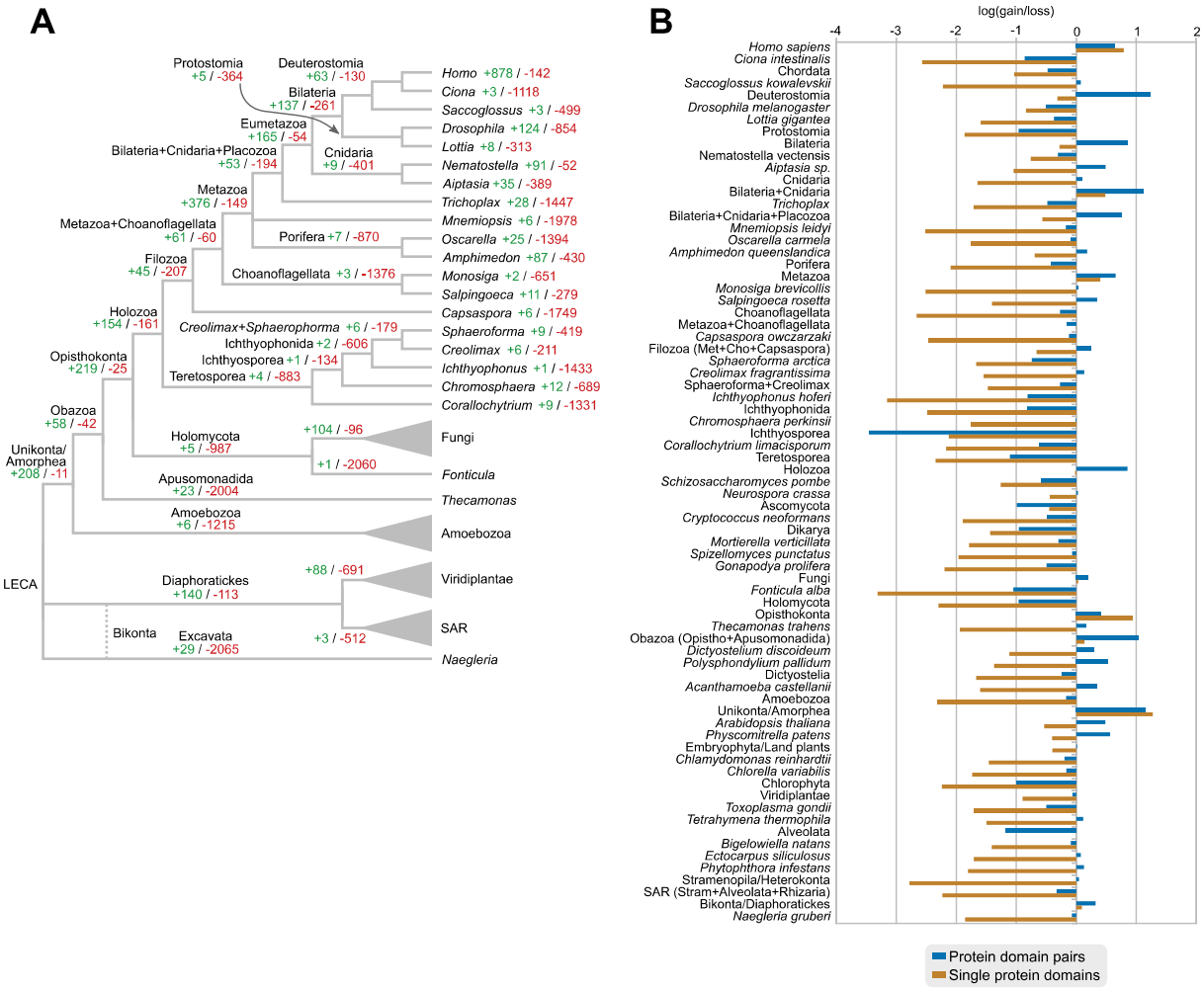


Figure 6-Supplement 2. Modularity and community size of the functional sub-networks based on domains related to **A)** signaling [12], **B)** ubiquitination [58], **C)** ECM [12,21,56] and **D)** protein binding [130] functions. Blue dots indicate real genomes, and the purple density plot indicates simulated rewired networks. Monotonic dependence between modularity and domains/community was tested for each set of data using Spearman's rank correlation coefficient (ρ_s); linear regression fits are included for clarity. Plots for sub-networks A-D exhibit the same decreasing trend as the global sub-network of Figure 6D-E, and contrast with the results for TFs.

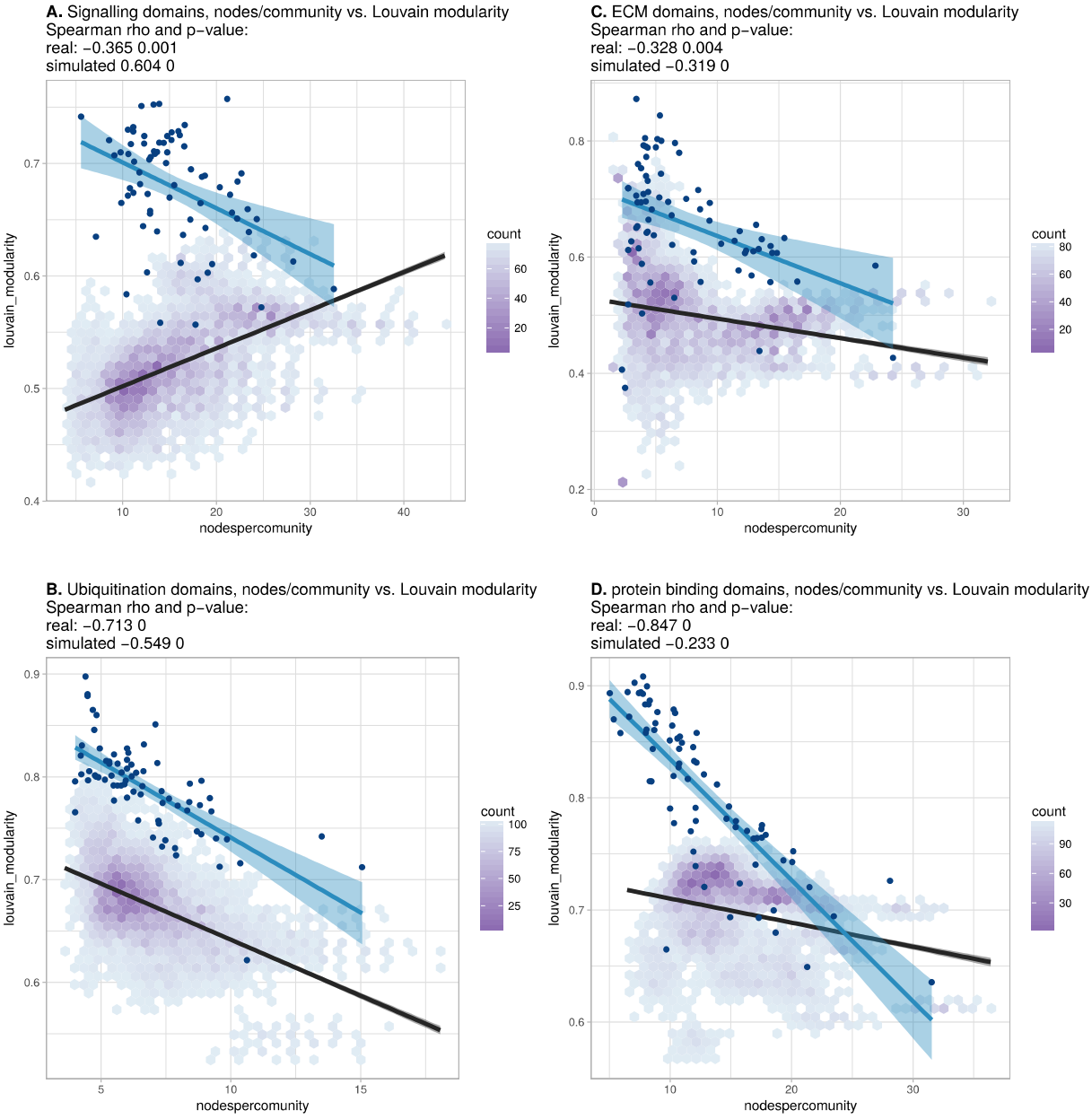


Figure 7. A and B) Protein domain co-occurrence matrices of transcription factor (TF) (**A**) or extracellular matrix (ECM)-related gene families (**B**), inferred at the LCA of Metazoa (≥90% probability). Horizontal and vertical axes of the heatmap represent individual protein domains and their mutual 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 Fig7-Supplement 1 for the complete BI and ML trees with statistical supports). Species abbreviation as in SD10. Source Data SD5, SD6, SD7, SD10.

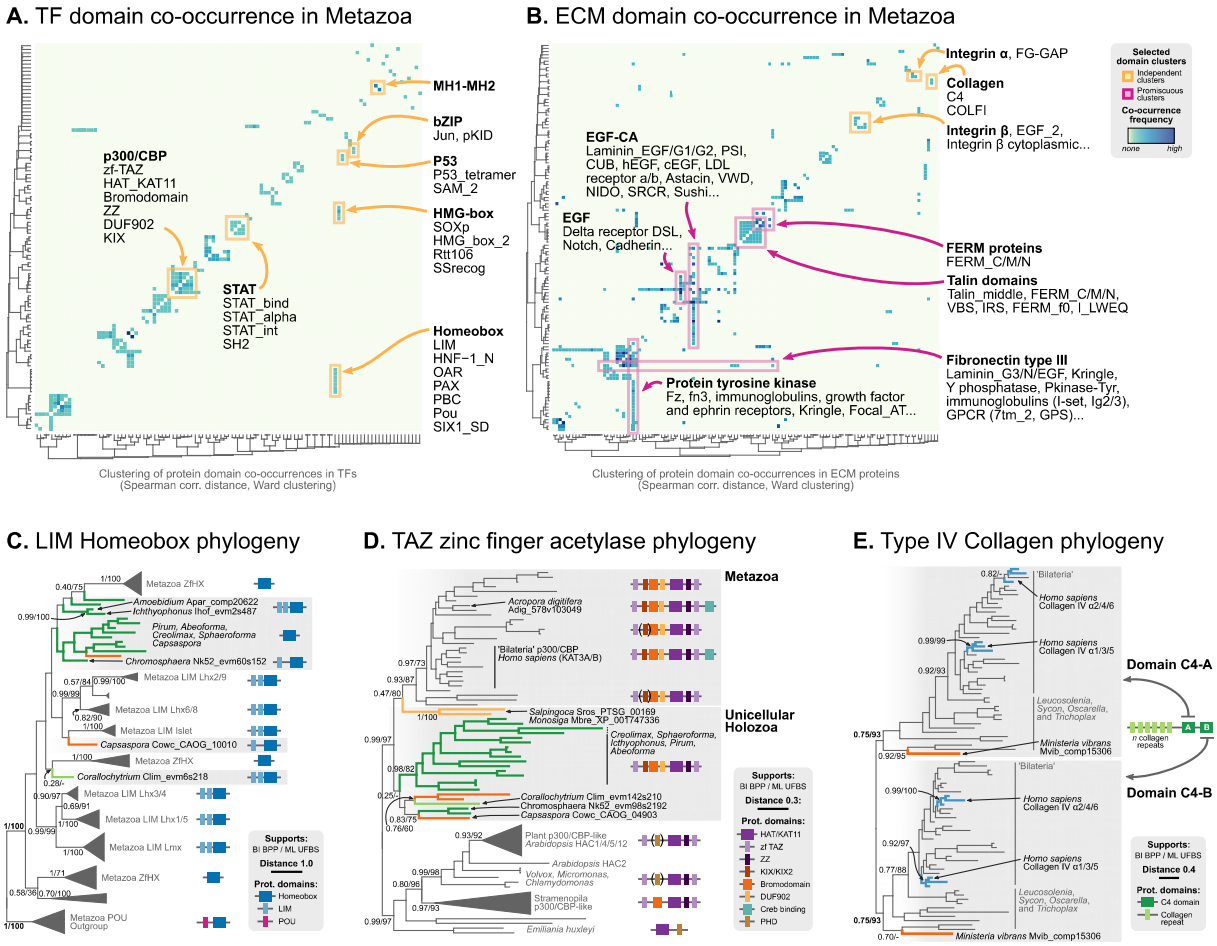
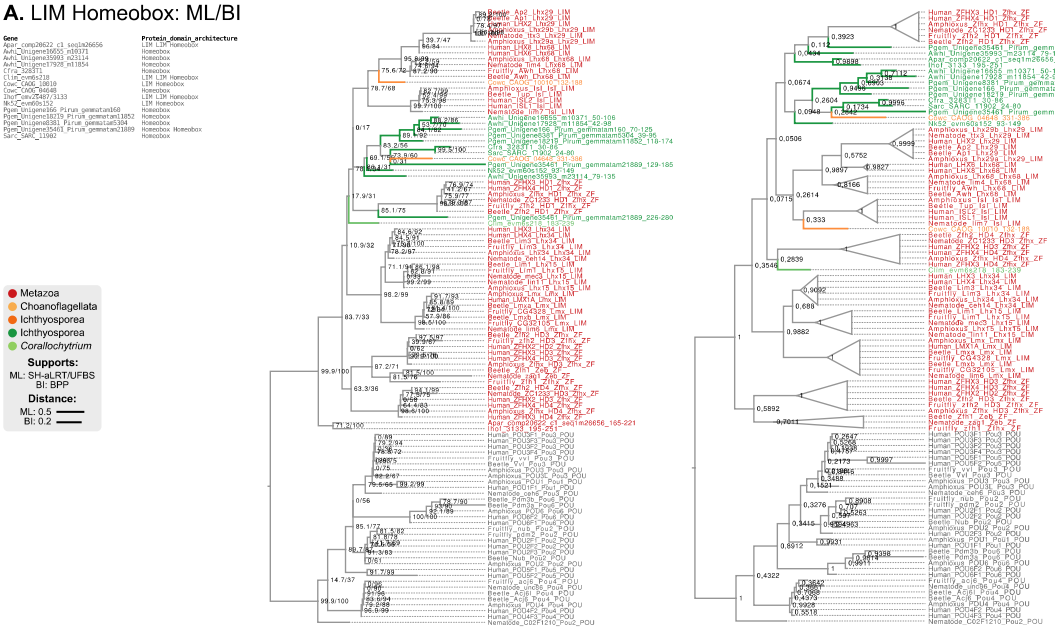
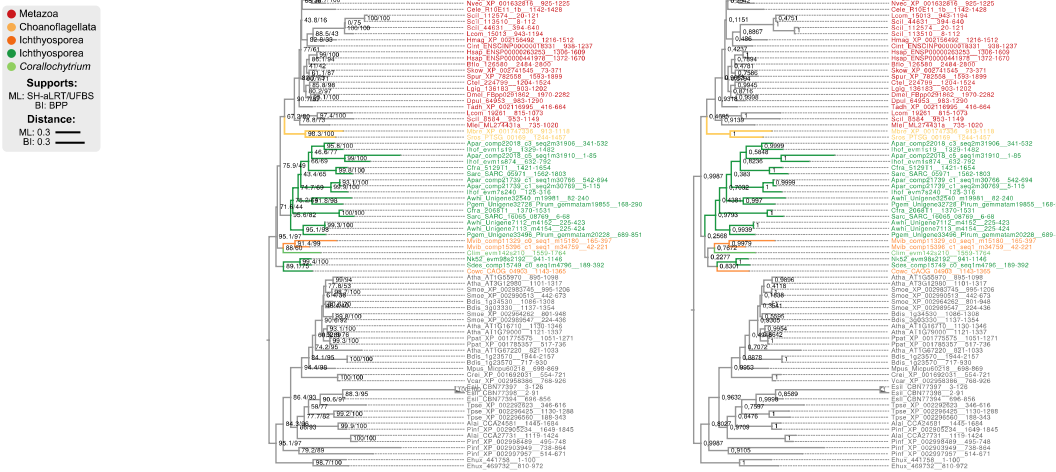


Figure 7-Supplement. Phylogenetic analysis of the **A)** LIM-Homeobox, **B)** p300/CBP, and **C)** Collagen Type IV, using Maximum likelihood in IQ-TREE (supports are SH-like approximate likelihood ratio test / UFBS, respectively) and Bayesian inference in Mr. Bayes (BPP statistical supports).

A. LIM Homeobox: ML/BI



B. p300/CBP: ML/BI



C. Type IV Collagen: ML/BI

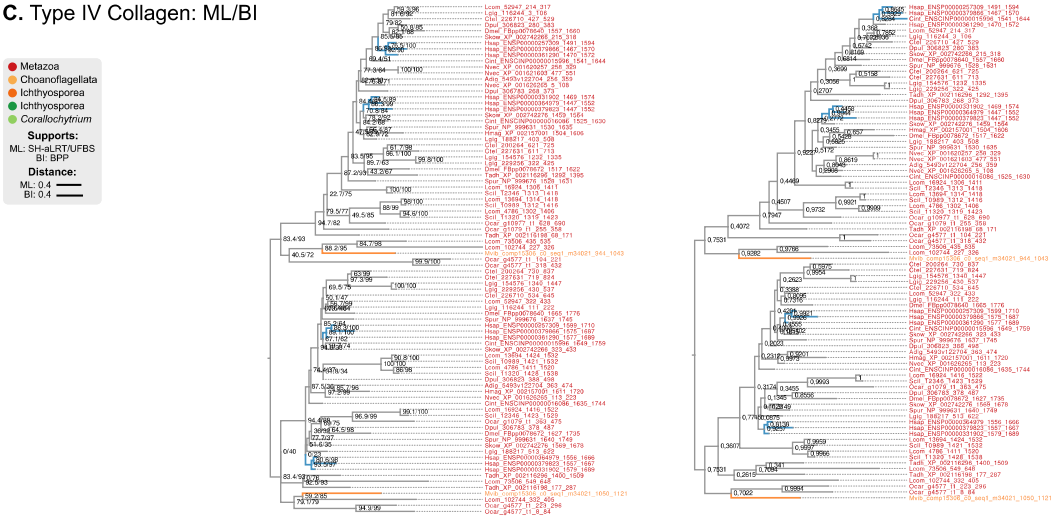


Table 1. Domain combinations that appear in transcription factor (TF) families in unicellular premetazoans, from the LCA of Unikonta/Amorphea to the LCA of Metazoa. First and second columns indicate the TF family and its inferred evolutionary origin, respectively (from [15]). Subsequent columns list 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. Source Data SD7.

Transcription factors	Metazoa	Metazoa +Choanof.	Filozoa	Holozoa	Opisthokonta	Unikonta
ARID	0.022 RFX_DNA_binding, RBB1NT, DUF3518			0.215 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, ELM2					
HLH	0.281 PAS_3, Hairy_orange		0.222 MITF_TFEB_C_3_N	0.001 Response_reg, CRAL_TRIO, PAS/9/11		
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-H2C2_2, zf-met				0.105 zf-C2H2_6	
zf-C2HC	0.136 MOZ_SAS					
zf-C4	0.600 Hormone_recep					
zf-GRF	0.537 Rnase_T					0.177 DUF2439, AAA_12
zf-MIZ	0.071 PINIT				0.030 SAP	0.026 PINIT
zf-TAZ				0.114 Bromodomain, DUF902, KIX		

IX. Source data

Source Data 1. Table of genome structure statistics, from the data-set of eukaryotic genomes used in the study. Includes genome size and portion of the genome covered by genes, exons, introns and intergenic regions. Used in Fig. 1 and 3.

Source Data 2. Annotated repetitive sequences from 10 unicellular Holozoa genomes. Includes transposable elements, simple repeats, low complexity regions and small RNAs. Used in Fig. 3.

Source Data 3. List of annotated transposable element families in 10 unicellular Holozoa genomes, with counts. Used in Fig. 3.

Source Data 4. 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 [125]. Used in Fig. 4.

Source Data 5. Reconstruction of intron site evolutionary histories, using a rates-across-sites Markov model for intron evolution, with branch-specific gain and loss rates [125]. Used in Fig. 4 and 5.

Source Data 6. 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 [126]. Used in Fig. 5, 6 and 7.

Source Data 7. 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 [126]. Used in Fig. 5, 6 and 7, and Table 1.

Source Data 8. 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 . Used in Fig. 1.

Source Data 9. Reconstruction of the evolutionary histories of individual protein domains, using Dollo parsimony and accounting for gains and losses [126]. Used in Fig. 5, 6 and 7, and Table 1.

Source Data 10. List of genome and transcriptome assemblies and annotations, including abbreviations, taxonomic classification and data sources.

Source Data 11. BVD57 phylogenomic dataset (see [22]), including 87 domains (with PFAM accession number) and unaligned sequences per species. Used in Fig. 2.

X. References

1. Budd GE, Jensen S. The origin of the animals and a "Savannah" hypothesis for early bilaterian evolution. Biol Rev. 2015;92: 446–473. doi:10.1111/brv.12239

2. dos Reis M, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PCJ, Yang Z. Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. Curr Biol. The Authors; 2015;25: 2939–2950. doi:10.1016/j.cub.2015.09.066

3. Grosberg RK, Strathmann RR. The Evolution of Multicellularity: A Minor Major Transition? Annu Rev Ecol Evol Syst. 2007;38: 621–654. doi:10.1146/annurev.ecolsys.36.102403.114735

4. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science (80-). 2007;317: 86–94. doi:10.1126/science.1139158

5. Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier ME a, Mitros T, et al. The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature. 2010;466: 720–726. doi:10.1038/nature09201

6. Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, et al. The ctenophore genome and the evolutionary origins of neural systems. Nature. 2014;510: 109–114. doi:10.1038/nature13400

7. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, et al. The *Trichoplax* genome and the nature of placozoans. Nature. 2008;454: 955–60. doi:10.1038/nature07191

8. Fortunato S a. V., Adamski M, Ramos OM, Leininger S, Liu J, Ferrier DEK, et al. Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. Nature. 2014;514: 620–623. doi:10.1038/nature13881

9. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, et al. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature. 2008;451: 783–8. doi:10.1038/nature06617

10. Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, et al. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. Genome Biol. 2013;14: R15. doi:10.1186/gb-2013-14-2-r15

11. Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E, et al. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. Nat Commun. 2013;4: 2325. doi:10.1038/ncomms3325

12. Richter DJ, King N. The Genomic and Cellular Foundations of Animal Origins. Annu Rev Genet. 2013; doi:10.1146/annurev-genet-111212-133456

13. Manning G, Young SL, Miller WT, Zhai Y. The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. Proc Natl Acad Sci. 2008;105: 9674–79. doi:10.1073/pnas.0801314105

14. Suga H, Dacre M, de Mendoza A, Shalchian-Tabrizi K, Manning G, Ruiz-Trillo I. Genomic Survey of Premetazoans Shows Deep Conservation of Cytoplasmic Tyrosine Kinases and Multiple Radiations of Receptor Tyrosine Kinases. Sci Signal. 2012;5: ra35-ra35. doi:10.1126/scisignal.2002733

15. de Mendoza A, Sebé-Pedrós A, Sestak MS, Matejcic M, Torruella G, Domazet-Lošo T, et al. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc Natl Acad Sci. 2013; doi:10.1073/pnas.1311818110

16. Sebé-Pedrós A, Irimia M, Del Campo J, Parra-Acero H, Russ C, Nusbaum C, et al. Regulated aggregative multicellularity in a close unicellular relative of metazoa. Elife. 2013;2: e01287. doi:10.7554/eLife.01287

17. Sebé-Pedrós A, Peña MI, Capella-Gutiérrez S, Gabaldon T, Ruiz-Trillo I, Sabidó E. High-throughput Proteomics Reveals the Unicellular Roots of Animal Phosphosignaling and Cell Differentiation. Dev Cell. 2016;In press: 1–12. doi:10.1016/j.devcel.2016.09.019

18. Simakov O, Kawashima T. Independent evolution of genomic characters during major metazoan transitions. Dev Biol. Elsevier; 2016; 0–1. doi:10.1016/j.ydbio.2016.11.012

19. Sebé-Pedrós A, de Mendoza A, Lang BF, Degnan BM, Ruiz-Trillo I. Unexpected Repertoire of Metazoan Transcription Factors in the Unicellular Holozoan *Capsaspora owczarzaki*. Mol Biol Evol. 2011;28: 1241–1254. doi:10.1093/molbev/msq309

20. Nichols SA, Roberts BW, Richter DJ, Fairclough SR, King N. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/ -catenin complex. Proc Natl Acad Sci. 2012;109: 13046–13051. doi:10.1073/pnas.1120685109

21. Sebé-Pedrós A, Roger AJ, Lang FB, King N, Ruiz-Trillo I. Ancient origin of the integrin-mediated adhesion and signaling machinery. Proc Natl Acad Sci. 2010;107: 10142–7. doi:10.1073/pnas.1002257107

22. Torruella G, De Mendoza A, Grau-Bové X, Antó M, Chaplin MA, Campo J Del, et al. Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. Curr Biol. 2015;25: 1–7. doi:10.1016/j.cub.2015.07.053

23. Mendoza L, Taylor JW, Ajello L. The class mesomycetozoea: a heterogeneous group of microorganisms at the animal-fungal boundary. Annu Rev Microbiol. 2002;56: 315–44. doi:10.1146/annurev.micro.56.012302.160950

24. Marshall WL, Celio G, McLaughlin DJ, Berbee ML. Multiple Isolations of a Culturable, Motile Ichthyosporean (Mesomycetozoa, Opisthokonta), *Creolimax fragrantissima* n. gen., n. sp., from Marine Invertebrate Digestive Tracts. Protist. 2008;159: 415–433. doi:10.1016/j.protis.2008.03.003

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

26. Suga H, Ruiz-Trillo I. Development of ichthyosporeans sheds light on the origin of metazoan multicellularity. Dev Biol. 2013; 1–9. doi:10.1016/j.ydbio.2013.01.009

27. de Mendoza A, Suga H, Permanyer J, Irimia M, Ruiz-Trillo I. Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. Elife. 2015;4: 7250–7. doi:10.7554/eLife.08904

28. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, et al. Assembling genomes and mini-metagenomes from highly chimeric reads. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2013;7821 LNBI: 158–170. doi:10.1007/978-3-642-37195-0_13

29. Del Campo J, Ruiz-Trillo I. Environmental survey meta-analysis reveals hidden diversity among unicellular opisthokonts. Mol Biol Evol. 2013;30: 802–805. doi:10.1093/molbev/mst006

30. Glockling SL, Marshall WL, Gleason FH. Phylogenetic interpretations and ecological potentials of the Mesomycetozoea (Ichthyosporea). Fungal Ecol. Elsevier Ltd; 2013; 1–11. doi:10.1016/j.funeco.2013.03.005

31. Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF. A phylogenomic investigation into the origin of Metazoa. Mol Biol Evol. 2008;25: 664–672. doi:10.1093/molbev/msn006

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

33. Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc B Biol Sci. 2015;370: 20140331. doi:10.1098/rstb.2014.0331

34. Elliott TA, Gregory TR. Do larger genomes contain more diverse transposable elements? BMC Evol Biol. 2015;15: 69. doi:10.1186/s12862-015-0339-8

35. Simakov O, Marletaz F, Cho S-J, Edsinger-Gonzales E, Havlak P, Hellsten U, et al. Insights into bilaterian evolution from three spiralian genomes. Nature. 2013;493: 526–31. doi:10.1038/nature11696

36. Irimia M, Tena JJ, Alexis MS, Fernandez-Minan A, Maeso I, Bogdanovic O, et al. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Res. 2012;22: 2356–2367. doi:10.1101/gr.139725.112

37. Csűrös M, Rogozin IB, Koonin E V. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. Ponting CP, editor. PLoS Comput Biol. 2011;7: e1002150. doi:10.1371/journal.pcbi.1002150

38. Carmel L, Wolf YI, Rogozin IB, Koonin E V. Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res. 2007;17: 1034–1044. doi:10.1101/gr.6438607

39. Rogozin IB, Carmel L, Csűrös M, Koonin E V. Origin and evolution of spliceosomal introns. Biol Direct. 2012;7: 28. doi:10.1186/1745-6150-7-11

40. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science (80-). 2012;338: 1587–93. doi:10.1126/science.1230612

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

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

43. Sebé-Pedrós A, Ballaré C, Parra-Acero H, Chiva C, Tena JJ, Sabidó E, et al. The Dynamic Regulatory Genome of *Capsaspora* and the Origin of Animal Multicellularity. Cell. 2016; 1–14. doi:10.1016/j.cell.2016.03.034

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

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

46. Csűrös M, Miklós I. A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. In: Apostolico A, Guerra C, Istrail S, Pevzner PA, Waterman M, editors. Research in Computational Molecular Biology. Venice; 2006. pp. 206–220. doi:10.1007/11732990_18

47. Lynch M. Intron evolution as a population-genetic process. Proc Natl Acad Sci. 2002;99: 6118–23. doi:10.1073/pnas.092595699

48. Lynch M, Conery JS. The origins of genome complexity. Science (80-). 2003;302: 1401–4. doi:10.1126/science.1089370

49. Marshall WL, Berbee ML. Population-level analyses indirectly reveal cryptic sex and life history traits of Pseudoperkinsus tapetis (Ichthyosporea, Opisthokonta): A unicellular relative of the animals. Mol Biol Evol. 2010;27: 2014–2026. doi:10.1093/molbev/msq078

50. Lynch M. The origins of eukaryotic gene structure. Mol Biol Evol. 2006;23: 450–468. doi:10.1093/molbev/msj050

51. Basu MK, Carmel L, Rogozin IB, Koonin E V. Evolution of protein domain promiscuity in eukaryotes. Genome Res. 2008;18: 449–61. doi:10.1101/gr.6943508

52. Basu MK, Poliakov E, Rogozin IB. Domain mobility in proteins: functional and evolutionary implications. Brief Bioinform. 2009;10: 205–16. doi:10.1093/bib/bbn057

53. Leonard G, Richards TA. Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. Proc Natl Acad Sci. 2012;109: 21402–21407. doi:10.1073/pnas.1210909110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1210909110

54. Tordai H, Nagy A, Farkas K, Bányai L, Patthy L. Modules, multidomain proteins and organismic complexity. FEBS J. 2005;272: 5064–78. doi:10.1111/j.1742-4658.2005.04917.x

55. Ekman D, Björklund AK, Elofsson A. Quantification of the elevated rate of domain rearrangements in metazoa. J Mol Biol. 2007;372: 1337–48. doi:10.1016/j.jmb.2007.06.022

56. Hynes RO. The evolution of metazoan extracellular matrix. J Cell Biol. 2012;196: 671–9. doi:10.1083/jcb.201109041

57. Deshmukh K, Anamika K, Srinivasan N. Evolution of domain combinations in protein kinases and its implications for functional diversity. Prog Biophys Mol Biol. 2010;102: 1–15.

58. Grau-Bové X, Sebé-Pedrós A, Ruiz-Trillo I. The Eukaryotic Ancestor Had a Complex Ubiquitin Signaling System of Archaeal Origin. Mol Biol Evol. 2015; msu334. doi:10.1093/molbev/msu334

59. Zmasek CM, Godzik A. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. BioMed Central Ltd; 2011;12: R4. doi:10.1186/gb-2011-12-1-r4

60. de Mendoza A, Sebé-Pedrós A, Ruiz-Trillo I. The Evolution of the GPCR Signaling System in Eukaryotes: Modularity, Conservation, and the Transition to Metazoan Multicellularity. Genome Biol Evol. 2014;6: 606–619. doi:10.1093/gbe/evu038

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

62. Xie X, Jin J, Mao Y. Evolutionary versatility of eukaryotic protein domains revealed by their bigram networks. BMC Evol Biol. BioMed Central Ltd; 2011;11: 242. doi:10.1186/1471-2148-11-242

63. Holland PWH, Booth HAF, Bruford EA. Classification and nomenclature of all human homeobox genes. BMC Biol. 2007;5: 47. doi:10.1186/1741-7007-5-47

64. Holland PWH. Evolution of homeobox genes. Wiley Interdiscip Rev Dev Biol. 2013;2: 31–45. doi:10.1002/wdev.78

65. Srivastava M, Larroux C, Lu DR, Mohanty K, Chapman J, Degnan BM, et al. Early evolution of the LIM homeobox gene family. BMC Biol. 2010;8: 4. doi:10.1186/1741-7007-8-4

66. Zhong Y, Holland PWH. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. Evol Dev. 2011;13: 567–568. doi:10.1111/j.1525-142X.2011.00513.x

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

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

69. Gadd MS, Bhati M, Jeffries CM, Langley DB, Trewhella J, Guss JM, et al. Structural Basis for Partial Redundancy in a Class of Transcription Factors, the LIM Homeodomain Proteins, in Neural Cell Type Specification. J Biol Chem. American Society for Biochemistry and Molecular Biology; 2011;286: 42971–42980. doi:10.1074/jbc.M111.248559

70. Gaiti F, Calcino AD, Tanurdžić M, Degnan BM. Origin and evolution of the metazoan non-coding regulatory genome. Dev Biol. 2016; doi:10.1016/j.ydbio.2016.11.013

71. Cromar G, Wong K-C, Loughran N, On T, Song H, Xiong X, et al. New Tricks for "Old" Domains: How Novel Architectures and Promiscuous Hubs Contributed to the Organization and Evolution of the ECM. Genome Biol Evol. 2014;6: 2897–2917. doi:10.1093/gbe/evu228

72. Aouacheria A, Geourjon C, Aghajari N, Navratil V, Deleage G, Lethias C, et al. Insights into Early Extracellular Matrix Evolution: Spongin Short Chain Collagen-Related Proteins Are Homologous to Basement Membrane Type IV Collagens and Form a Novel Family Widely Distributed in Invertebrates. Mol Biol Evol. 2006;23: 2288–2302. doi:10.1093/molbev/msl100

73. Heino J. The collagen family members as cell adhesion proteins. BioEssays. 2007;29: 1001–1010. doi:10.1002/bies.20636

74. Fahey B, Degnan BM. Origin and evolution of laminin gene family diversity. Mol Biol Evol. 2012;29: 1823–1836. doi:10.1093/molbev/mss060

75. Exposito JY, Larroux C, Cluzel C, Valcourt U, Lethias C, Degnan BM. 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. J Biol Chem. 2008;283: 28226–28235. doi:10.1074/jbc.M804573200

76. Grau-Bové X, Sebé-Pedrós A, Ruiz-Trillo I. A genomic survey of HECT ubiquitin ligases in eukaryotes reveals independent expansions of the HECT system in several lineages. Genome Biol Evol. 2013;5: 833–47. doi:10.1093/gbe/evt052

77. Irimia M, Roy SW. Origin of Spliceosomal Introns and Alternative Splicing. Cold Spring Harb Perspect Biol. 2014;6. doi:10.1101/cshperspect.a016071

78. Fernandez-Valverde SL, Degnan BM. Bilaterian-like promoters in the highly compact *Amphimedon queenslandica* genome. Sci Rep. 2016;6: 22496. doi:10.1038/srep22496

79. Mikhailov K V., Konstantinova A V., Nikitin MA, Troshin P V, Rusin LY, Lyubetsky V a, et al. The origin of Metazoa: a transition from temporal to spatial cell differentiation. BioEssays. 2009;31: 758–768. doi:10.1002/bies.200800214

80. Andrews S. FastQC [Internet]. 2014 [cited 3 May 2016]. Available: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc

81. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–20. doi:10.1093/bioinformatics/btu170

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

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

84. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014;24: 1384–1395. doi:10.1101/gr.170720.113

85. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1: 18. doi:10.1186/2047-217X-1-18

86. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29: 1072–5. doi:10.1093/bioinformatics/btt086

87. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31: 3210–2. doi:10.1093/bioinformatics/btv351

88. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 2011;27: 757–63. doi:10.1093/bioinformatics/btr010

89. Haas BJ, Delcher AL, Mount S.M. SM, Wortman JR, Smith RK, Hannick LI, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31: 5654–5666. doi:10.1093/nar/gkg770

90. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9: R7. doi:10.1186/gb-2008-9-1-r7

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

92. Korf I. Gene finding in novel genomes. BMC Bioinformatics. BioMed Central; 2004;5: 59. doi:10.1186/1471-2105-5-59

93. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29: 644–52. doi:10.1038/nbt.1883

94. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics. 2015;51: 11.14.1-19. doi:10.1002/0471250953.bi1114s51

95. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12: 656–64. doi:10.1101/gr.229202. Article published online before March 2002

96. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. Methods Mol Biol. 2016;1418: 283–334. doi:10.1007/978-1-4939-3578-9_15

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

98. Torruella G, Derelle R, Paps J, Lang BF, Roger AJ, Shalchian-Tabrizi K, et al. Phylogenetic Relationships within the Opisthokonta Based on Phylogenomic Analyses of Conserved Single-Copy Protein Domains. Mol Biol Evol. 2012;29: 531–544. doi:10.1093/molbev/msr185

99. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32: 268–274. doi:10.1093/molbev/msu300

100. Quang LS, Gascuel O, Lartillot N. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 2008;24: 2317–2323. doi:10.1093/bioinformatics/btn445

101. Minh BQ, Nguyen MAT, Von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 2013;30: 1188–1195. doi:10.1093/molbev/mst024

102. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol. 2010;59: 307–321. doi:10.1093/sysbio/syq010

103. Lartillot N, Rodrigue N, Stubbs D, Richer J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol. 2013;62: 611–5. doi:10.1093/sysbio/syt022

104. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004;21: 1095–109. doi:10.1093/molbev/msh112

105. Derelle R, Torruella G, Klimeš V, Brinkmann H, Kim E, Vlček Č, et al. Bacterial proteins pinpoint a single eukaryotic root. Proc Natl Acad Sci. 2015;112: E693–E699. doi:10.1073/pnas.1420657112

106. He D, Fiz-Palacios O, Fu C-J, Fehling J, Tsai C-C, Baldauf SLL. An Alternative Root for the Eukaryote Tree of Life. Curr Biol. Elsevier Ltd; 2014;24: 465–70. doi:10.1016/j.cub.2014.01.036

107. Burki F, Kaplan M, Tikhonenkov D V, Zlatogursky V, Minh BQ, Radaykina L V, et al. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc R Soc B Biol Sci. 2016;283: 20152802. doi:10.1098/rspb.2015.2802

108. Whelan N V, Kocot KM, Moroz LL, Halanych KM. Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci. 2015;112: 5773–8. doi:10.1073/pnas.1503453112

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

110. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40: D290-301. doi:10.1093/nar/gkr1065

111. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–80. doi:10.1093/molbev/mst010

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

113. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27: 1164–5. doi:10.1093/bioinformatics/btr088

114. Smit A, Hubley R, Green P. RepatMasker Open-4.0. Repeat Masker. 2015.

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

116. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. Oxford University Press; 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

117. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots: Various R Programming Tools for Plotting Data. 2016.

118. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 2009;19: 1639–45. doi:10.1101/gr.092759.109

119. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25: 1972–3. doi:10.1093/bioinformatics/btp348

120. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–3. doi:10.1093/bioinformatics/btu033

121. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20: 289–90.

122. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015.

123. Csűrös M, Holey JA, Rogozin IB. In search of lost introns. Bioinformatics. 2007;23: i87-96. doi:10.1093/bioinformatics/btm190

124. Csűrös M, Rogozin IB, Koonin E V. Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach. Mol Biol Evol. 2008;25: 903–911. doi:10.1093/molbev/msn039

125. Csűrös M. Malin: maximum likelihood analysis of intron evolution in eukaryotes. Bioinformatics. 2008;24: 1538–9. doi:10.1093/bioinformatics/btn226

126. Csűrös M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics. 2010;26: 1910–2. doi:10.1093/bioinformatics/btq315

127. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. United States; 2004;69: 26113. doi:10.1103/PhysRevE.69.026113

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

129. Weirauch MT, Hughes TR. A Handbook of Transcription Factors. Hughes TR, editor. Subcellular Biochemistry. Dordrecht: Springer Netherlands; 2011. doi:10.1007/978-90-481-9069-0

130. Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database: the classification re-

source after 15 years. Nucleic Acids Res. 2015;43: D213-21.
doi:10.1093/nar/gku1243

3.7. Correlated evolution of alternative splicing modes and gene architecture

Abstract – Alternative splicing (AS) is a major mechanism of transcriptome regulation in eukaryotes that can facilitate the diversification of the proteome and add an additional and responsive layer of gene expression control. We explore the landscape of AS across sixty species from all major eukaryotic lineages using high-coverage transcriptomic data, and uncover a consistent relationship between inter-specific shifts in the frequency of different modes of AS (exon skipping and intron retention) and the evolution of gene architecture. In particular, the advent of exon skipping-richer AS profiles, typical of animals and plants, is a readily evolvable feature present in both unicellular and multicellular species with conducive genome architectures – *e.g.*, intron-dense genomes with poorly defined splice sites. Using this approach, we uncover a pan-eukaryotic code of cis-features that determines the AS profile of extant eukaryotes. This code can be extended to infer the state of AS-dependent transcriptome regulation in ancestral eukaryotes for which genome architecture can be reconstructed – from the ancestors of multicellular animals or plants, to the earliest eukaryotes.

Correlated evolution of alternative splicing and gene architecture across eukaryotes

Xavier Grau-Bové[1,2], Iñaki Ruiz-Trillo[1,2,3], Manuel Irimia[4,5]

1. Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, 08003, Barcelona

2. Departament de Genètica, Microbiologia i Estadística, Universitat de Barelona, Avinguda Diagonal 643, 08028, Barcelona

3. ICREA, Passeig Lluís Companys 23, 08010, Barcelona

4. Centre de Regulació Genòmica, Barcelona Institute of Science and Technology, Dr. Aiguader 88, 08003, Barcelona

5. Universitat Pompeu Fabra (UPF), Plaça de la Mercè 10-12, 08002, Barcelona

Abstract

Alternative splicing (AS) is a major mechanism of transcriptome regulation in eukaryotes that can facilitate the diversification of the proteome and add an additional and responsive layer of gene expression control. We explore the landscape of AS across sixty species from all major eukaryotic lineages using high-coverage transcriptomic data, and uncover a consistent relationship between interspecific shifts in the frequency of different modes of AS (exon skipping and intron retention) and the evolution of gene architecture. In particular, the advent of exon skipping-richer AS profiles, typical of animals and plants, is a readily evolvable feature present in both unicellular and multicellular species with conducive genome architectures – e.g., intron-dense genomes with poorly defined splice sites. Using this approach, we uncover a pan-eukaryotic code of cis*-features that determines the AS profile of extant eukaryotes. This code can be extended to infer the state of AS-dependent transcriptome regulation in ancestral eukaryotes for which genome architecture can be reconstructed – from the ancestors of multicellular animals or plants, to the earliest eukaryotes.*

Keywords: Alternative splicing, exon skipping, intron retention, ancestral reconstruction, genome architecture, eukaryotic evolution.

I. Introduction

Alternative splicing (AS) is a regulatory process that allows the creation of multiple possible messenger RNA (mRNA) transcripts from a single gene, by differentially selecting splice sites in multi-exonic sequences (Breitbart et al. 1987). The possibility of creating multiple protein isoforms from a single gene renders them an effective mechanism to increase the proteomic repertoire of eukaryotic genomes (Nilsen and Graveley 2010; Graveley 2001), potentially leading to key evolutionary innovations (Bush et al. 2017; Gueroussov et al. 2015; Gracheva et al. 2011).

The main forms of AS among eukaryotes are the exclusion of particular exons (termed exon skipping or cassette exons) and retention of introns (Breitbart et al. 1987; Keren et al. 2010). These sources of transcript variation are widespread in eukaryotes, but pioneering studies revealed that the prevalence of each AS mode varied across eukaryotes: animals showed a bias towards exon skipping (ES) over intron retention (IR), which tended to be specifically favoured in fungi, plants and various protist lineages (McGuire et al. 2008; Kim et al. 2007). Thus, ES-dominated AS has been proposed to be a major contributor to the increased phenotypic complexity of animals (Irimia and Roy 2014): many isoforms have been found to be physiologically relevant (reviewed in (Kelemen et al. 2013; Nilsen and Graveley 2010)), for example by tuning the protein-protein interaction networks (Yang et al. 2016; Ellis et al. 2012; Buljan et al. 2012). Contrastingly, IR has been sometimes regarded as erroneous (Kim et al. 2008) or 'sloppy' splicing (Koonin et al. 2013), although it has a well-established functionality in down-regulation of gene expression via the nonsense-mediated decay (NMD) pathway (Wong et al. 2016; Brogna et al. 2016; Lykke-Andersen and Jensen 2015), nuclear retention (Le Hir et al. 2003) or intron detention (Boutz et al. 2015). Early dichotomic views regarding ES and IR have been further nuanced as IR was found to be much more frequent in animals than previously reported (Braunschweig et al. 2014). In parallel, high-coverage RNA-seq experiments uncovered high, animal-like ES frequencies in the chlorarachniophyte rhizarian *Bigelowiella natans* (Curtis et al. 2012); and, albeit with lower incidences, in eukaryotes as diverse as plants, holozoans, fungi and apicomplexans (Sebé-Pedrós et al. 2013; de Mendoza et al. 2015; Sorber et al. 2011; Bush et al. 2017; Kempken 2013).

The last eukaryotic common ancestor had an intron-dense genome (Csűrös et al. 2011) with heterogeneous splice sites (Irimia et al. 2007; Schwartz et al. 2008; Irimia and Roy 2008), and all the essential splicing machinery (the spliceosome, a complex of small nuclear RNAs, and assisting splicing factors) (Collins and Penny 2005). These observations have allowed to infer that the earliest eukaryotes already exhibited splicing-rich transcriptomes yielding multiple mRNA variants per gene, mostly by intron retention events (Roy and Irimia 2009; Irimia and Roy 2014; Koonin et al. 2013), and opened the path for the ancestral reconstruction of the AS profile of different eukaryotes by comparative genomic analyses of their intron/exon structure.

Here, we aim to examine the evolution of AS in eukaryotes under the light of genome architecture. In particular, we analyse the prevalence of both modes of alternative splicing—intron retention (IR) and exon skipping (ES)—across 60 species covering all major lineages (Fig. 1; Table S1) and uncover a set of genic features that influence the frequency of IR and ES in transcripts in multiple eukaryotes, thus suggesting the existence of a soft pan-eukaryotic *cis*-code for alternative splicing determination. Using this comprehensive data-set of joint transcriptomic and genomic data, we then investigate the IR-to-ES transition reported in complex multicellular lineages, animals and plants, by comparing their AS profiles and genome architectures with their closest unicellular relatives.

II. Results and discussion

Varying frequencies of intron retention and exon skipping across eukaryotes

We quantified the frequencies of ES and IR events at the whole-transcriptome level for each eukaryotic genome in our dataset. The analysis of ES events included a data-set of $1.84{\cdot}10^6$ exons from $3.61{\cdot}10^5$ genes of 60 eukaryotic genomes (Supplementary Table S1), with transcription evidence and varying frequencies of skipping. Each sufficiently expressed exon was thus classified into two categories: constitutive exons (0-2% skipping rate, or p_{ES}, representing 97.5% of the data-set) and positively skipped exons (p_{ES}=5-90%, 1.2% of the data). For the analysis of IR, an analogous data-set was built consisting of $1.04{\cdot}10^6$ introns from $2.33{\cdot}10^5$ genes, that were also classified into constitutively excluded and IR-positive introns (p_{IR} =0-2% and 56.5% of the data; and p_{IR} =5-90% and 23.1% of the data, respectively) (see Methods and Supplementary Tables S2 and S3).

We examined the global weight of each AS mode at each species, finding that IR events were more common than ES in all the surveyed species – including vertebrates and the chlorarachniophyte *B. natans* (Fig. 2A-B). Transcripts from all major eukaryotic lineages exhibit, at varying levels, evidence for both AS modes, concordantly with the proposed early emergence of dual AS in eukaryotes (Irimia and Roy 2014; Roy and Irimia 2009; Koonin et al. 2013). This result is in agreement with recent reports highlighting the pervasivity of intron retention across eukaryotes and challenges previous views of animal transcriptomes as being ES-dominated. A pos-

sible explanation for this disagreement is the association between high IR rates and low transcript expression levels, which hinders the detection of retained introns (particularly in earlier studies based on EST data (McGuire et al. 2008; Kim et al. 2007)) (see below).

Still, it remains the case that ES events are more frequent and affect more genes in animals than in any other eukaryotic group (Fig. 2C). We thus questioned when did ES-richer transcriptomic profiles appear in animal evolution: is it an animal innovation, or can it be traced back to their unicellular ancestors in the Holozoa clade? Previous examinations of ES in close unicellular relatives of animals, such as the filasterean amoeba *Capsaspora owczarzaki* (Sebé-Pedrós et al. 2013) and the ichthyosporean *Creolimax fragrantissima* (de Mendoza et al. 2015), reported low ES rates (~1-3% of their multiexonic genes), suggesting that ES-rich AS might be an animal innovation. Our analysis rather indicates a later origin for ES predominance at the root of Bilateria: whole-transcriptome ES frequencies were consistently higher in bilaterians than in cnidarians, the ctenophore *Mnemiopsis leidyi*, *Trichoplax adhaerens* and sponges (Fig. 2A). On the other hand, the relatively ES-rich transcriptome of the ichthyosporean *Sphaeroforma arctica* is a clear departure from the lower values exhibited by the other unicellular holozoans (6.5% of multiexonic genes).

In parallel, multicellular land plants also exhibit higher ES rates than most eukaryotes, including their colonial and unicellular relatives within the green lineage. There is, however, a notable exception: the colonial chlorophyte *Volvox carteri*, in which ES weight is higher than other algae (including its close unicellular relative *Chlamydomonas reinhardtii*) and some land plants (e.g. *Arabidopsis thaliana* or the lycophyte *Selaginella moellendorffii*) (Fig. 2A).

Tantalizingly, both *S. arctica* and *V. carteri* have larger genomes with higher intron densities than their closest relatives, be it other unicellular holozoans or chlorophytes – i.e., they have more available raw genomic material to produce transcript diversity. If the complex AS profiles of animals and plants were facilitated by high ancestral levels of splicing variability, as hypothesized in (Koonin et al. 2013; Irimia and Roy 2014), this phenomenon could therefore influence their unicellular relatives as well.

To examine the biological significance of each species' ES and IR profile, we examined how did the AS event affect the transcript reading frame. We found that alternatively skipped exons of ES-rich animal transcriptomes show a clear bias towards having 3-divisible lengths (henceforth, 3-n bias), i.e., their exclusion from the transcript does not cause frame-shift errors (Fig. 2D, blue dots, $p<0.01$ in Fisher's exact test). This is the case of ES-rich vertebrates and other bilaterians, but also the cnidarians *Aiptasia* sp., *Hydra magnipapillata* or the ctenophore *Mnemiopsis leidyi*. In animals like *H. sapiens* or *Drosophila melanogaster*, maintenance of the reading frame is associated with functional ES events (Sorek et al. 2004; Irimia et al. 2008). In contrast, ES events of most other eukaryotes show no 3-n bias; while plants, *S. arctica* ($p<0.01$) *and V. carteri* ($p<0.05$) exhibit an opposite bias towards non-3-divisible exon lengths. The lack of 3-n bias in ES has also been reported in *C. fragrantissima* (de Mendoza et al. 2015) and *B. natans* (Curtis et al. 2012), where reading frame disruptions have been proposed to be a consequence of noisy splicing rather than produce functional isoforms. Similarly, we did not detect significant biases in 3-divisibility in the intron lengths of IR events (Fig. 2D, grey dots, $p<0.01$ in Fisher's exact test), with the single exception of the alveolate *Tetrahymena thermophila*.

Thus, the widespread presence of ES appears to be associated with variable levels of spliceosomal noise, both in animals and other eukaryotes. In contrast, the 3-n bias of the ES events recorded in bilaterians, cnidarians and ctenophores implies that a shift towards a predominance of functional ES events (i.e., producing viable protein isoforms) is an exclusive animal innovation.

Influences of gene structure over IR and ES events

Previous studies have linked the varying modes of AS across genomes to differences in the gene architecture, such as the relative length of exons and introns, intron density, splicing site homogeneity and other *cis* signals (De Conti et al. 2013; Braunschweig et al. 2014; Barbosa-Morais et al. 2012). These features can explain why, despite the general dominance of IR, the fraction of genes affected by either AS mode varies widely across the analysed eukaryotes. Thus, we decided to explore the relationship between these features and AS modes at the eukaryote-wide level. In particular, we tested whether the ES-richer transcriptomes of the unicellular eukaryotes *S. arctica* and *V. carteri* could be linked to the emergence of animal-like genomic features.

We analysed the intron/exon structure and sequence composition of the genomic regions surrounding the alternatively spliced transcript segments, and correlated these parameters with the AS frequencies at the species level. In particular, we investigated the effect of global length of genes and transcripts; the length of the alternatively spliced exons (for ES) or introns (for IR) and its surrounding exons or introns (respectively); their % GC content and the differential GC content between exons and introns; the strength of the intron-exon junction definition at the 3' and 5' splice sites; the total number of spliceosomal introns in the given gene; the position of the AS event within the gene; and the transcript expression level (cRPKMs from pooled RNA-seq experiments). Our analysis identifies consistent relationships between ES and IR frequencies and gene architecture across the whole eukaryotic tree of life, conserved across genomes despite different dominant AS modes (Fig. 3 and 4).

In most analysed eukaryotes, alternatively skipped exons are shorter than constitutive exons (Fig. 3). In turn, ES events are associated with longer flanking introns, both upstream and downstream (at least in animals, unicellular holozoans and land plants). This result is consistent with the proposed mode of splicing by 'exon definition' typical of exons surrounded by long introns: the recognition of the 5' and 3' splice sites occurs across the exonic sequence (as intron ends are more distant); thus, interrupting this process can result in exon exclusion (De Conti et al. 2013). The positive relationship between ES and higher intron-to-exon length ratios also fits this principle.

Regarding IR, the 'intron definition' splicing model proposes the opposite scenario: impediments to the across-intron recognition of splice sites can lead to IR, and this would typically happen for short introns flanked by long exons (De Conti et al. 2013). As the median CDS length is relatively constant across eukaryotes (~1200-1400 bp (Elliott and Gregory 2015)), intron-poor and compact genomes would have longer exons and would in turn be dominated by IR. However, our analysis reveals that the influence of intron and flanking exons' length on IR is not homogeneous across eukaryotes: retained introns are indeed shorter than excluded ones in chordates (*Homo sapiens*, *Danio rerio*, *Xenopus tropicalis*, and *Ciona intestinalis*), some land plants (*Vitis vinifera*, *Mimulus guttatus*), and unicellular algae (*Emiliania huxleyi*, *B. natans*, *Naegleria gruberi* or *Guillardia theta*); but not in most other animals, unicellular holozoans, fungi or other protists (Fig. 4), which are IR-rich as well (Fig. 2B). The ratio of intron-to-exon has an equally uneven relationship with IR. However, higher intron densities (measured as introns per kbp of CDS) effectively negatively correlate with the level of IR, as expected. Overall, the dominance of IR in a given genome does not seem to be determined by a straight-forward relationship between intron length and density. Instead, positive or negative associations repeatedly emerge in a lineage- or species-specific manner.

Across all sampled eukaryotes, we identify a consistent relationship between positive cases of ES and IR and higher heterogeneity in the 5' and 3' splice sites. It has been frequently reported that heterogeneous splice sites favoured the emergence of AS-rich transcriptomes, functional and dysfunctional, in ancestral eukaryotes (Ast 2004; Schwartz et al. 2008; Irimia and Roy 2008; Irimia et al. 2007), a feature which was linked to the reported high intron densities in the line of descent from the LECA to animals and plants (Csűrös et al. 2011; Koonin et al. 2013). Here we show that heterogeneous splice sites also influence the IR and ES rates at the intra-specific level: poorly defined introns and exons are more subject to AS than those closer to the species consensus.

We also used the GC content of introns and exons to examine the effect of general sequence composition in AS. ES events are associated with high-GC exons in animals, but low-GC in most other eukaryotes (Fig. 3). In both groups of species, though, there is a positive relationship between ES and the differential of GC between flanking introns (GC-richer, compared to exons) and skipped exons (AT-richer). This association is maintained in IR events: retained introns have higher GC content than their flanking exons (Fig. 4).

Finally, we examined the effect of whole-transcript expression levels in AS: ES events are more frequent in lowly expressed genes across eukaryotes (Fig. 3). We expected a similar pattern in IR events, as they are associated with either down-regulation via NMD or random splicing errors (more prone to affect lowly expressed genes; see below and (Saudemont et al. 2017)). However, this is conspicuously not the case in many species (Fig. 4A). We did find, however, a clear negative correlation in IR-positive introns between the retention frequency (p_{IR}) and their expression level, thus extending the results from mammalian transcriptomes (Braunschweig et al. 2014) to all eukaryotes.

Overall, the ES events here detected across eukaryotes are globally associated with short exons flanked by longer introns, with weak 5' and 3' splice sites. Inasmuch these features are more common in the genomes of animals and plants than in most eukaryotes, similar genomic architectures in their unicellular relatives can be expected to produce relatively ES-richer transcriptomes. This is precisely the case for both *S. arctica* and *V. carteri* (Fig. 2A). First, their intron-dense genomes derive from lineage-specific intron gain processes not shared by animals and plants (respectively, at the root of ichthyophonid Ichthyosporea [Grau-Bové *et al.* 2017] and the root of Chlorophyceae + Trebouxiophyceae (Csűrös et al. 2011)). Second, their increased intron length distributions also point to independent origins of ES-conducive genomes (Fig. 1B). Overall, these results emphasize the possibility that ES-rich transcriptomes are a readily evolvable property, as long as underlying transcriptomic variability and a conducive genomic architecture co-exist.

A 'soft code' of *cis* features determines AS events across eukaryotes

The global consistency of these varied gene structural features in their association with both ES and IR raises the question as to whether they could be jointly used as predictors of the dominant AS mode in a given organism – in other words, a 'soft code' for alternative splicing determination. To this end, we developed species-specific logistic regression models for the sets of ES and IR events, taking into account the above-mentioned gene features as possible predictors. Similar predictive frameworks have already been put in place in taxonomically-restricted contexts such as mammals (Braunschweig et al. 2014) and vertebrates (Barbosa-Morais et al. 2012). The accuracy of each model was measured using

the area under the ROC curve for the corresponding regression (AUC ROC parameter; see Methods). This analysis yielded better-than-random predictions for both ES and IR in 95% of the surveyed species (AUC>0.5, p<0.01), and higher quality models in 41.5% (AUC>0.7, p<0.01) (Fig. 5 and Supplementary Table S4).

The combined predictive performance of the gene features was remarkably consistent for the IR models: 16 out of 22 surveyed variables appeared as significant (p<0.05, Wald test) and concordant (coefficient of the same sign) predictors in 75-100% of the surveyed species (Supplementary Table S6). This result is largely in agreement with the predictive model developed for mammal IR in (Braunschweig et al. 2014). The predictive performance of ES models was also concordant across species, although just 8 out of 21 variables showing showing a consistent predictive capacity. For example, the number of introns per gene was assigned a significant positive coefficient (p<0.05, Wald test) in 77.8% of the surveyed species; negative coefficients were assigned to the exon length, the ratio intron-to-exon lengths, exon GC%, 5' and 3' SS score and the expression level (70-100% across-species concordance) (Supplementary Table S5).

Overall, this approach allows us to define a set of tractable gene structural features with a similar influence on both ES and IR across multiple eukaryotic species. This issue is particularly interesting in an evolutionary context: we can indirectly infer the dominant AS mode in ancestral eukaryotes by reconstructing their intron densities (Csűrös et al. 2011), lengths (Elliott and Gregory 2015), the conservation of consensus splice sites (Schwartz et al. 2008; Irimia and Roy 2008), or any of the other relevant features.

Intron retention: functional regulation or dysfunctional noise?

The biological significance of the widespread and frequent IR events in eukaryotes remains poorly understood, with two main hypotheses proposed. First, that IR is a general way of fine-tuning eukaryotic transcriptomes by removing excess transcripts via NMD, that recognizes and degrades transcripts with premature stop codons (Lykke-Andersen and Jensen 2015; Braunschweig et al. 2014). Second, that IR is caused by random errors in the splicing process, and is a noisy, generally dysfunctional phenomenon (Melamud and Moult 2009; Pickrell et al. 2010). Both scenarios imply similar predictions and are thus difficult to test at the whole-transcriptome level (e.g., both the NMD and errorneous IR scenarios predict higher IR in lowly expressed transcripts, either due to their removal or to cell economy).

A recent study proposed an indirect method to estimate the fraction of erroneous transcripts at the whole-transcriptome level (Saudemont et al. 2017) by comparing the frequency of IR between 1) median genes and 2) a sub-set of highly constrained genes for which splicing errors are supposed to be negligible (due to higher costs). We defined species-specific sets of putatively IR-constrained transcripts using the gene features most consistently negatively-correlated with IR-positive events (Fig. 4): strongly defined 5' and 3' splice sites, long genes and a high number of introns per gene. We then estimated the fraction of erroneous IR for each species (Fig. 6, see Methods), finding that virtually all surveyed eukaryotes (96%) had fractions of erroneous IR F_{err}>20%, and a majority (68%) had F_{err}>50%. In the case of *H. sapiens*, we obtained a F_{err}=72.1%, very similar to the 72% value reported by (Saudemont et al. 2017) despite differences in methodology.

III. Conclusions

We identify a set of gene architectural features that influence the frequency of IR and ES events within a given species' transcriptome – the relative length of introns and exons, splice site definition strength and intron density. Since these *cis* gene features are globally coherent across species at the pan-eukaryotic level (Fig. 2 and 3), we conclude that they constitute a universal, albeit not deterministic 'soft code' that affects the dominant AS modes in different species. Our data-derived predictive model can be thus used to deduce the frequency of IR and ES events in reconstructed ancestral genomes.

Finally, our investigation into the IR-to-ES transitions in AS typical of complex multicellular organisms (plants and animals) reveals that independent evolution of conducive genome architectures within multicellular lineages (Bilateria, *V. vinifera*) and in sporadic unicellular allies (*V. carteri* or *S. arctica*) coincide with biases towards ES-richer transcriptomes – even when there is no evidence of such AS events leading to expanded proteomes via differential isoform translation (Fig. 2D). Indeed, whereas the IR-to-ES transitions appear to be attainable readily evolvable under the adequate genome architectural environment, the co-option of ES for regulated proteome expansion seems to be a largely animal-specific feature.

Overall, our results emphasize the effect of long-term genome evolutionary patterns in shaping AS, a fast-changing transcriptome regulatory layer. Thus, determining the circumstance behind the genome architecture evolution will be key to understand the emergence of functional, AS-rich transcriptomes in eukaryotes.

IV. Methods

Sources of genome and transcriptome data

We have assembled a data-set consisting of genome assemblies and annotations from 60 eukaryotic species for

which high-coverage Illumina RNA-seq libraries were already available (Supplementary Table S1). Specifically, we retrieved the genomic coordinates of genes, transcripts and exon sequences (GFF) in order to a set of canonical transcripts for each genome. If more than one isoform per gene was annotated, the longest possible CDS was considered to be the canonical transcript (a proxy with ~90% correspondence with proteomics-driven main isoform selection (Ezkurdia et al. 2015)).

In order to homogenize the experimental procedures used to built each RNA-seq library, we used only poly-A-selected libraries of single-end reads, trimmed down to a minimum of 50 base-pairs (bp) each, if appropriate (using FASTX Toolkit (Gordon 2017)). For those those species where RNA-seq experiments included more than one sample (biological and/or technical replicates, time series, growth conditions, or else), all reads were pooled into a single FASTQ file containing both reads and per-base quality information.

Detection and quantification of exon skipping and intron retention events

We followed the computational framework developed by Irimia *et al.* (Curtis et al. 2012; Barbosa-Morais et al. 2012; Sebé-Pedrós et al. 2013; de Mendoza et al. 2015) to detect and quantify alternative splicing events belonging to the intron retention (IR) and exon skipping (ES) categories.

Exon skipping detection

For each exon of the genome, we built a 'triplet' of composite exonic sequences consisting of 1) 42 bp from the 5' end of the first exon and 42 bp from the 3' end of the second exon (E1-E2 junction); 2) 42 bp fragments from the 5' end of the first exon and the 3' end of the third exon (E1-E3); and 3) 42 bp fragments from the 5' end of the second exon and the 3' end of the third exon (E1-E2). Hence, each triplet consisted of two canonical junctions (E1-E2 and E2-E3) and a non-canonical one that skipped the middle exon (E1-E3), all of them 84 bp-long (or less, if any exon was shorter than 42 bp).

Then, we computed the effective mappability of each junction in order to exclude exon-exon boundaries where RNA-seq mapping would be insufficient (Labbé et al. 2012). Specifically, we 1) built an artificial RNA-seq library consisting of all the possible reads derived from each junction in a 50 bp sliding window; 2) mapped these reads to the original junctions using *bowtie* v1.1.2, allowing a maximum of 3 mismatches (*-v 3*) and no multiple alignments (*-m 1*) (Langmead et al. 2009); and 3) removed all junction triplets for which at least one triplet had <20 effectively mappable positions (maximum is 35 for 50 bp reads, and ≥8 positions mapped from each exon).

Then, we aligned the pooled RNA-seq libraries to the remaining exon-triplets, using *bowtie* and the same parameters as above. We then corrected the number of mapped reads by the ratio obtained from dividing the mappable positions of that junction (between 20-35 bp) and the maximum theoretical mappability (35 bp).

The frequency of middle exon skipping of each exon-triplet (p_{ES}) was computed as follows:

$$p_{ES} = \frac{r_{E1E3}}{((r_{E1E2} + r_{E2E3})/2)}$$

where r denotes the mappability-corrected number of reads mapping in the E1-E2, E2-E3 and E1-E3 junctions (subindexes).

Finally, we classified exon junctions into three categories according to their mappability-corrected mapping values. A given exon-triplet was deemed ES-positive if the following conditions were fulfilled: $r_{E1E2}+r_{E2E3}+r_{E1E3} > 20$, $r_{E1E2}+r_{E2E3} > 2$, $r_{E1E3} > 1$, $p_{ES} > 2\%$ and $p_{ES} < 90\%$. If the total number of mapped reads was sufficient but $p_{ES} < 2\%$, the triplet was deemed ES-constitutive (i.e., negative). If any condition was not fulfilled, the triplet was deemed non-classifiable (i.e., NA).

Intron retention detection

For each intron of the genome, we built a triplet of composite exon-intron sequences (henceforth, 'intron-triplet') consisting of 1) 42 bp from the 5' end of the first exon and 42 bp from the 3' end of adjoining intron (E-I junction); 2) 42 bp fragments from the 5' end of the first exon and the 3' end of the second exon (E-E); and 3) 42 bp fragments from the 5' end of the intron and the 3' end of the second exon (I-E). Hence, each triplet consisted of one canonical junction (E-E) and two non-canonical ones that spanned the intron ends (E-I and I-E), all of them 84 bp-long (or less, if any exon or intron was shorter than 42 bp).

The mappability of each exon-intron junction was computed as specified above, discarding cases with <20 effectively mappable positions as well. We then aligned the same pooled RNA-seq libraries to the remaining exon-intron junctions using *bowtie*, and corrected the number of mapped reads.

The frequency of intron retention of each intron-triplet (p_{IR}) was computed as follows:

$$p_{IR} = \frac{(r_{IE} + r_{EI})/2}{r_{EE}}$$

where r denotes the mappability-corrected number of reads mapping in the I-E, E-I and E-E junctions (subindexes).

Finally, we classified intron junctions into three categories according to their mappability-corrected mapping values. A given intron-triplet was deemed IR-positive if the following conditions were fulfilled: $r_{IE}+r_{EI}+r_{EE} > 20$, $r_{IE}+r_{EI} > 2$, $r_{EE} > 1$, $p_{IR} > 2\%$ and $p_{IR} < 90\%$. If the total number of mapped reads was sufficient but $p_{IR} < 2\%$,

the triplet was deemed IR-constitutive (i.e., negative). If any condition was not fulfilled, the triplet was deemed non-classifiable (i.e., NA).

Transcriptome-wide quantification of AS levels

We measured the weight of ES and IR at the species level (w_{ES} and w_{IR}) using the following formulae:

$$w_{\mathrm{ES}} = \frac{\sum p_{\mathrm{ES,i}}}{\sum \left(p_{\mathrm{ES,i}} + p_{\mathrm{no\text{-}ES,i}} \right)} \cdot n_{\mathrm{genes}}$$

$$w_{\mathrm{IR}} = \frac{\sum p_{\mathrm{IR,i}}}{\sum \left(p_{\mathrm{IR,i}} + p_{\mathrm{no\text{-}IR,i}} \right)} \cdot n_{\mathrm{genes}}$$

Where n_{genes} represents the number of genes with at least one mappable exon or intron junction, respectively. This value therefore corrects the sum of ES or IR frequencies in the genome by the number of ES- or IR-visible genes.

Analysis of gene features: architecture, splice sites and expression levels

For each exon or intron junction analysed, we recorded the following parameters describing the architecture of their corresponding gene: gene length, transcript length (CDS only), exon length and GC content (for intron junctions, we analysed the values upstream and downstream exons), intron length and GC (for exon junctions, we analysed the values of upstream and downstream introns), total number of introns in the gene, position of the junction within the gene sequence (bps from first codon), and a categorical variable describing whether the length of the central junction element is divisible by 3 or not (i.e., if its exclusion or retention can alter the open reading frame). These features were derived from the GFF annotation.

In addition, we analysed the conservation degree of 3' and 5' splice sites when compared to species-specific consensus. For each species, we built position-weighted matrices (PWM) from the alignments of all 3' (23 bp, 20 from the intron and 3 from the exon) and 5' splice sites (9 bp, 3 from the exon and 6 from the intron) using the consensus matrix function in the *Biostrings* R library (Pages et al.). Then, for each individual splice site in the genome, the distance from the PWM consensus was calculated. Relevant lengths for each splice site were taken from (Liebert et al. 2004).

Finally, we evaluated transcript expression levels using the mappability-corrected RPKM value (cRPKM), aligning the pooled RNA-seq libraries of each species to the predicted transcriptome using *bowtie*, and calculating transcript-specific effective mappabilities as detailed above and in (Labbé et al. 2012).

Statistical analysis of AS frequency and gene features

For each species and for each of the quantitative gene features listed above, significant differences between the values taken by the IR-/ES-positive triplets and the IR-/ES- negative triplets were evaluated using two independent Kolmogorov-Smirnov two-sample tests with complementary alternative hypotheses: first, we tested whether the empirical cumulative distribution of the positive triplets lied above the constitutive values; second, we tested whether it lied below. We used the Kolmogorov-Smirnov distance, or D statistic, as a measure of the distance between each distribution, recording this value as positive if $p<0.01$ in the first test, and negative if $p<0.01$ in the second one.

Then, tested whether there were differences in the values taken by each gene feature within the IR-/ES-positive categories. First, we binned positive triplets into four categories (2-20%, 20-40%, 40-60%, 60-90%) and tested whether the gene feature values originate from the same distribution with the Kruskall-Wallis rank sum test, with significance for $p<0.05$. Then, we analyzed whether positive triplets exhibited monotonic correlations between their AS frequency and the values of each gene feature using the Spearman's rank correlation coefficient (ρ, significant for $p<0.05$).

Finally, we tested if the frequency of 3-divisible lengths in retained introns or skipped exons significantly differed from that of constitutive triplets using a Fisher's exact test (significant for $p<0.01$).

We used the R *stats* library to perform all statistical tests here mentioned (R Core Team 2015).

Prediction of AS modes using gene architecture features

Using our binary classification of positive/constitutive ES and IR events, we learned a binomial logistic regression models for each species and AS mode. We used 1) the positive and constitutive ES or IR events as the binary dependent variable, and 2) 20 quantitative gene features and the fraction of 3-divisible exons or introns (for ES and IR, respectively) as the putative independent predictors. The binomial logistic regression were built using the generalized linear model function from the R *stats* library (R Core Team 2015), using a K=10-fold cross-validation of the estimated coefficients. The predictive performance of each model was estimated with the area under its corresponding ROC curve (AUC ROC parameter), calculated using the original positive/constitutive data as the input, with the *pROC* R library (Robin et al. 2011). Significant differences between each model's predictions (AUC>0.5) and a null distribution of random predictions (AUC=0.5) were assessed using the Wilcoxon rank-sum test. The significance of each model's variable-specific coefficients was assessed using the Z-statistic significance according to

the Wald test (significant for $p<0.05$) (Supplementary Tables S4 and S5).

Estimation of the rate of error in IR

A recent study proposed a population-genetic framework that allows the estimation of the fraction of IR events due to dysfunctional splicing errors (in contrast to functional retentions) (Saudemont et al. 2017). They assume that 1) the fraction of functional IR is constant for different sets of genes (e.g. when classifying them by bins of expression); 2) the cost of erroneous IR is higher in highly expressed genes with many introns/longer CDS, as resource mis-allocation should be selected against. Therefore, the ration between the IR rates of highly-constrained and median genes can serve as a proxy to estimate the frequency of erroneous IR.

In particular, the fraction of erroneous IR (F_{err}) can be calculated from the ratio between p_{IR} of the median genes and IR-constrained transcript sets (r_{mc}), as long as selection for lack of errors is much higher in the former group (see (Saudemont et al. 2017) for details):

$$r_{mc} = \frac{p_{\text{IR,median}}}{p_{\text{IR,highly constrained}}}$$

$$F_{\text{err}} = 1 - \frac{1}{r_{mc}}$$

Having found an unequal relationship between IR-positive genes and lower expression levels across eukaryotes (Fig. 4), we deemed this parameter unsuitable for the definition of the IR-constrained transcript set. Instead, we used a combination of alternative features: for each species, we used the top 5-quantiles of gene length, number of introns/gene and either 3' or 5' splice site strength. The validity of this alternative approach is endorsed by the consistency in the reported values of erroneous IR of *H. sapiens* between this study (72.1%) and (Saudemont et al. 2017) (72.%). The median gene set is defined as genes in the 1-3 quantiles of cRPKM expression level.

V. Figures

Figure 1. Modes of AS and intron/exon structure across eukaryotes. A) Classification of AS events, after (Keren et al. 2010). Only exon skipping (ES) and intron retention (IR) are covered in this study. **B)** Intron density (introns/gene) and intron length distribution (in bp) across the 60 eukaryotic species here analysed. Dots represent median intron lengths and vertical lines delimit the first and third quartiles. Colour-coded according to taxonomy.

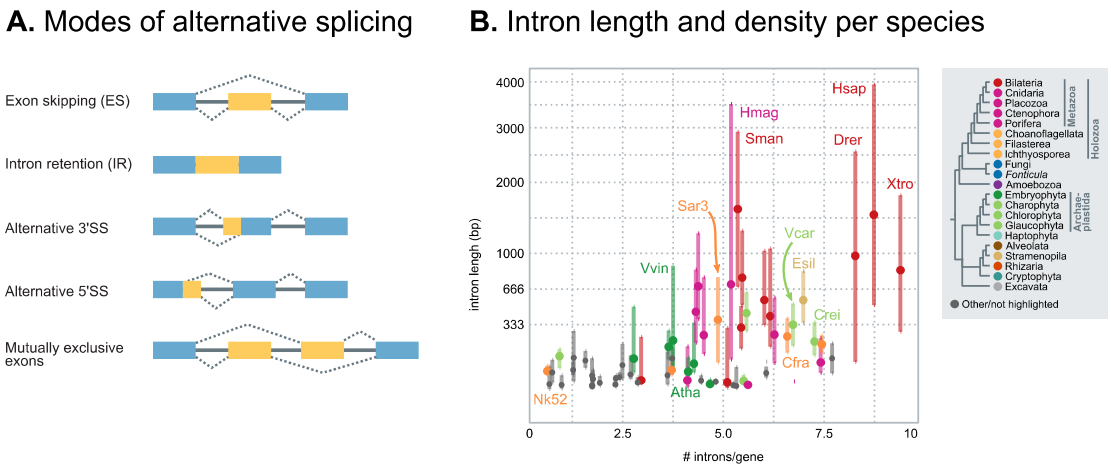


Figure 2. Weight of ES and IR at the whole-transcriptome level. A and B) Weight of exon skipping (A) and intron retention (B) in the transcriptome each species (arbitrary units of 'weight', see Methods), colour-coded according to taxonomy. **C)** Percentage of genes affected by ES and IR in each species, colour-coded according to taxonomy. **D)** Percentage of 3-divisible exons in ES analysis (blue dots) or introns in IR (grey dots). Highlighted dots mark species where there is an enrichment in either direction when comparing constitutive and positive AS events (Fisher's exact test, p<0.01). Highlight is colour-coded according to taxonomy.

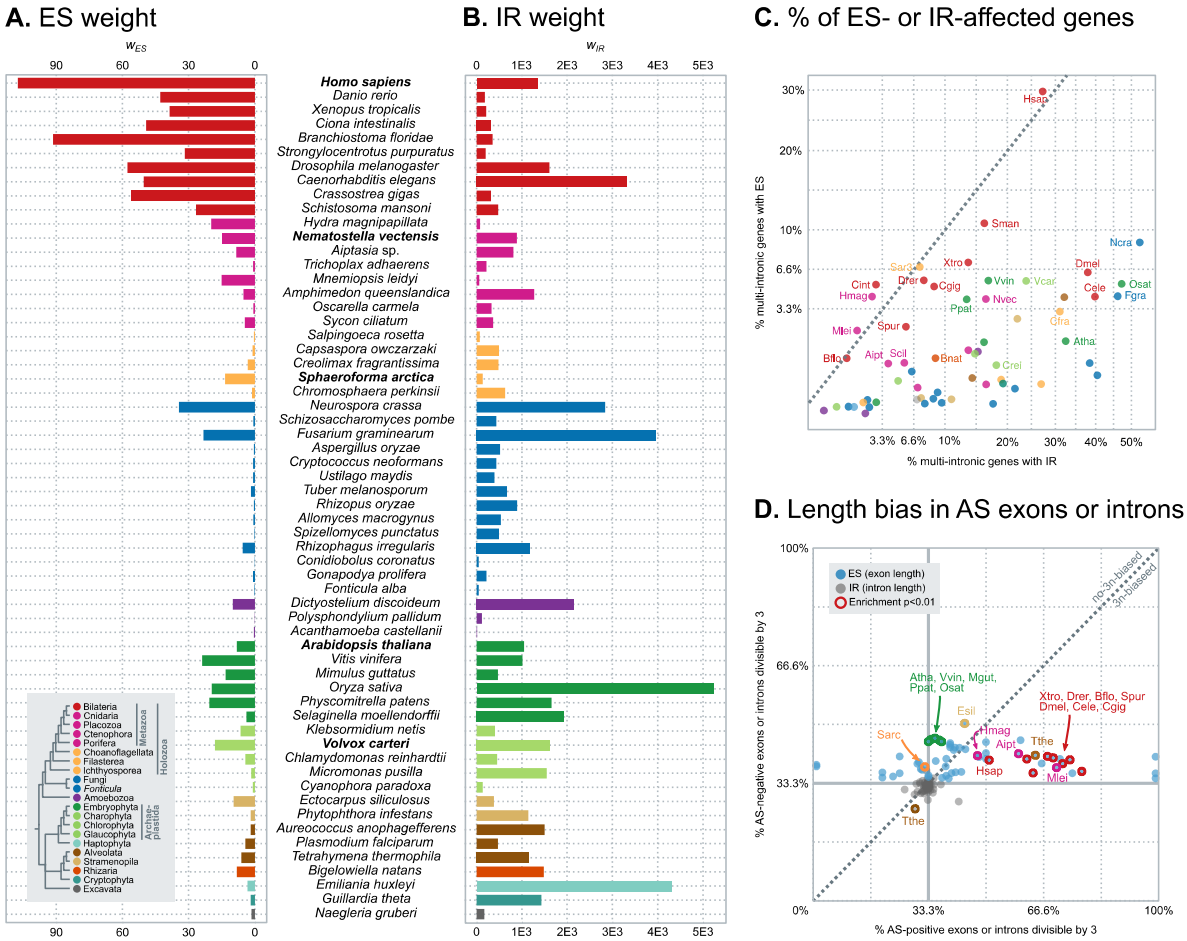


Figure 3. Gene structure and ES. A) Heatmap representing the distance between distributions of gene architecture values for ES-positive and ES-constitutive events, measured with the D statistic of the Kolmogorov-Smirnov two-sample test (significant if p<0.01; otherwise grey). D values are recorded as positive (green) or negative (red) according to two distinct Kolmogorov-Smirnov tests with complementary one-sided null hypotheses (i.e., positive green Ds reflect a positive correlation between ES and the given feature; negative red Ds the opposite). **B)** Selected examples of gene structure values for ES-constitutive (orange) and ES-positive (blue-green) events, from selected species (from A, in bold). For each pair of distributions, m_c and m_p represent the median value of the constitutive and positive distributions, respectively; significant differences tested using the Wilcoxon rank-sum test (p<0.05). Differently-scaled Y-axes marked in red.

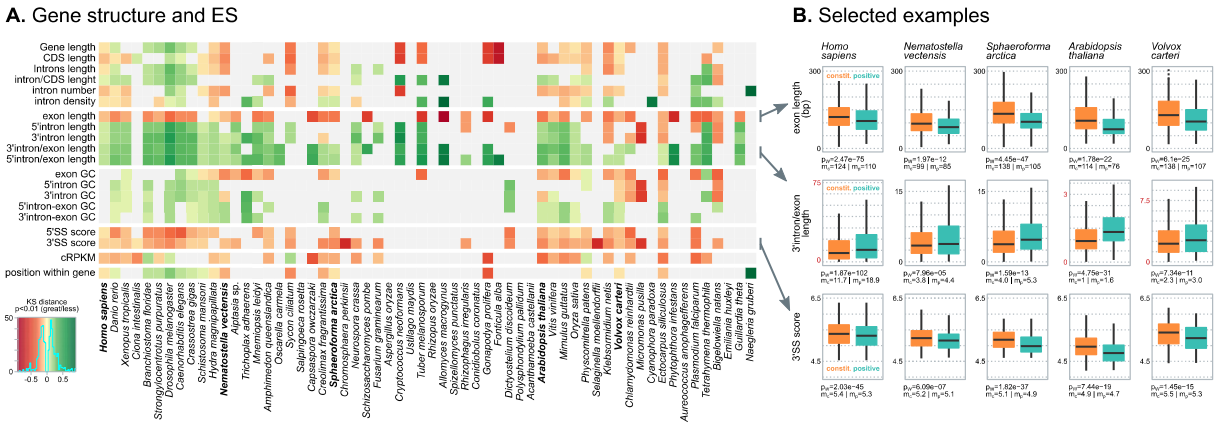


Figure 4. Gene structure and IR. A) Heatmap representing the distance between distributions of gene architecture values for IR-positive and IR-constitutive events, measured with the D statistic of the Kolmogorov-Smirnov two-sample test (significant if p<0.01; otherwise grey). D values are recorded as positive (green) or negative (red) according to two distinct Kolmogorov-Smirnov tests with complementary one-sided null hypotheses (i.e., positive green Ds reflect a positive correlation between IR and the given feature; negative red Ds the opposite). .**B)** Selected examples of gene structure values for IR-constitutive (orange) and IR-positive (blue-green) events, from selected species (from A, in bold). For each pair of distributions, m_c and m_p represent the median value of the constitutive and positive distributions, respectively; significant differences tested using the Wilcoxon rank-sum test (p<0.05). Differently-scaled Y-axes marked in red. In the case of expression levels (cRPKMs), IR-positive values have been binned in 5 categories (p_{IR}=2-10%, 10-20%, 20-30%, 30-50%, 50-90%). Monotonic dependence between expression and p_{IR} in IR-positive events was tested with Spearman's rank correlation coefficient (ρ, significant for p<0.05).

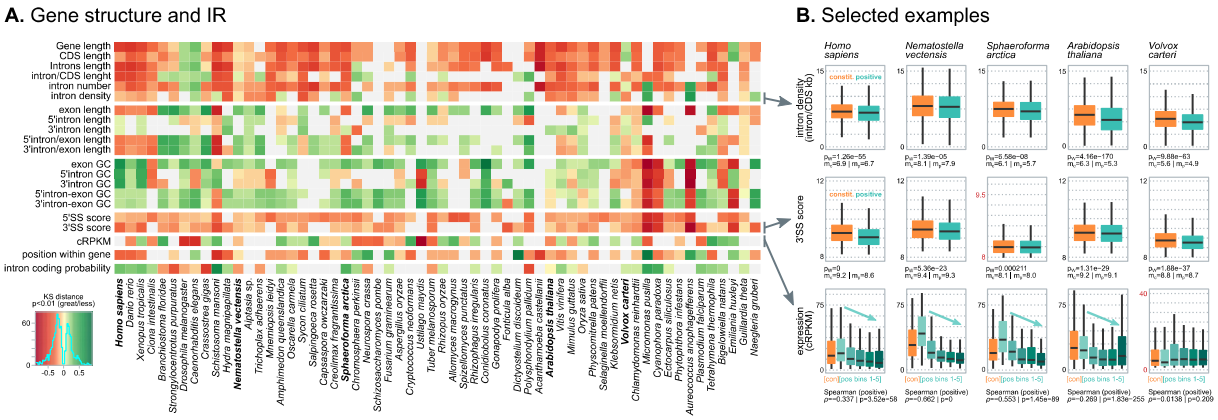


Figure 5. ROC curves of ES and IR. A) Species-specific ROC curves for ES logistic regression models. AUC (area under the ROC curve) values indicated for selected species. **B)** Species-specific ROC curves for IR logistic regression models. AUC values indicated for selected species. **C)** AUC of the ROC curve for the ES (horizontal axis) and IR (vertical) species-specific models. Highlighted quadrants delimit the species for which both ES and IR are >0.7 (green) or >0.5 (brown). Complete data as Supplementary Table S6.

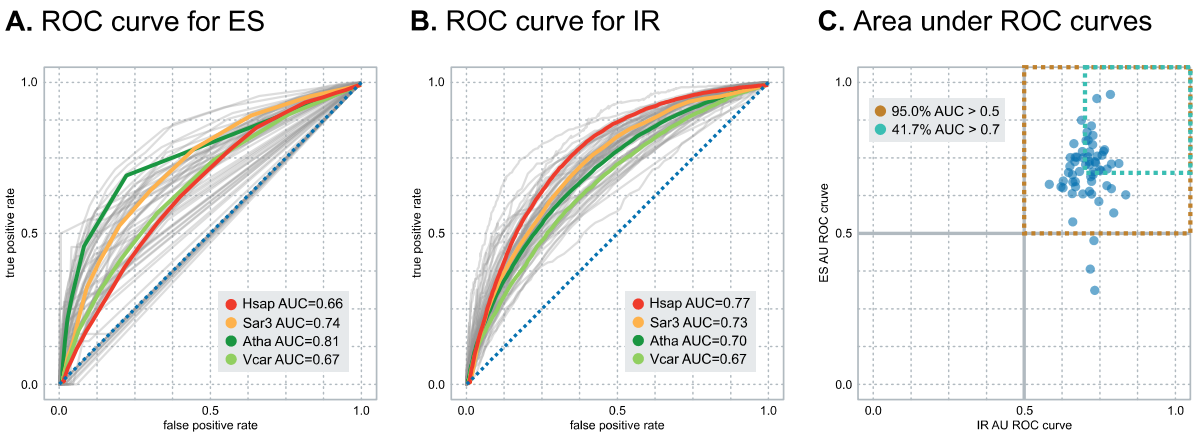
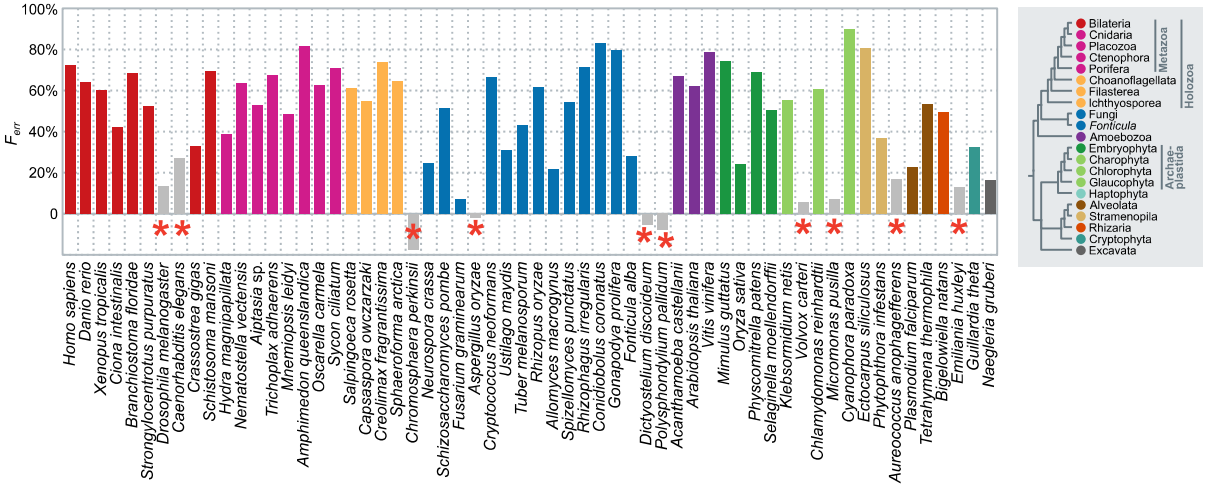


Figure 6. Fraction of erroneous IR per species. Colour-coded according to taxonomy. Species whose IR profile violates the assumptions of (Saudemont et al. 2017) are greyed-out and marked with a red asterisk. Note that violations can be *a priori* (i.e., when higher gene lengths, intron number and 5' or 3' SS scores do not correlate with IR) or *a posteriori* (cases in which $p_{\mathrm{IR,median}} < p_{\mathrm{IR,highly\ constrained}}$). Complete data as Supplementary Table S7.



VI. Supplementary material

Supplementary Table S1. Summary of the eukaryotic species used in this study, including the source of genome sequences, annotations, and RNA-seq experiments. Dataset accession numbers from SRA unless otherwise stated.

Species	Abbr.	Genome sequence and annotation	Transcriptome data	Taxonomy
Homo sapiens	Hsap	Ensembl 78	SRP007412	Metazoa
Danio rerio	Drer	Ensembl 78	ERP001280	Metazoa
Xenopus tropicalis	Xtro	Ensembl 78	SRP012375	Metazoa
Ciona intestinalis	Cint	Ensembl 78	SRP042651	Metazoa
Branchiostoma floridae	Bflo	Ensembl 78	SRR923751	Metazoa
Strongylocentrotus purpuratus	Spur	Ensembl Metazoa 27	SRP014690	Metazoa
Drosophila melanogaster	Dmel	Ensembl Metazoa 27	SRP001696	Metazoa
Caenorhabditis elegans	Cele	Ensembl 78	SRP000401	Metazoa
Crassostrea gigas	Cgig	Ensembl Metazoa 27	SRP014559	Metazoa
Schistosoma mansoni	Sman	Ensembl Metazoa 27	ERP000427	Metazoa
Trichoplax adhaerens	Tadh	Ensembl Metazoa 27	CRG	Metazoa
Hydra magnipapillata	Hmag	NCBI PRJNA12876	SRP051110	Metazoa
Nematostella vectensis	Nvec	Ensembl Metazoa 27	SRP018739	Metazoa
Aiptasia sp.	Aipt	(Baumgarten et al. 2015)	SRP047443	Metazoa
Mnemiopsis leidyi	Mlei	Ensembl Metazoa 27	SRP014828	Metazoa
Amphimedon queenslandica	Aque	Ensembl Metazoa 27	SRR1511618	Metazoa
Oscarella carmela	Ocar	(Nichols et al. 2012)	SRR1042012	Metazoa
Sycon ciliatum	Scil	(Fortunato et al. 2014)	ERP005418	Metazoa
Salpingoeca rosetta	Sros	Ensembl Protist 27	SRP005692	Choanoflagellata
Capsaspora owczarzaki	Cowc	(Suga et al. 2013)	(Sebé-Pedrós et al. 2013)	Filasterea
Creolimax fragrantissima	Cfra	(de Mendoza et al. 2015)	(de Mendoza et al. 2015)	Ichthyosporea
Sphaeroforma arctica	Sar3	Broad Institute Multicellularity Initiative	CRG	Ichthyosporea
Chromosphaera perkinsii	Nk52	Grau-Bové *et al.* 2017	Grau-Bové *et al.* 2017	Ichthyosporea
Neurospora crassa	Ncra	Ensembl Fungi 27	SRP016065	Fungi
Schizosaccharomyces pombe	Spom	Ensembl Fungi 27	ERP001483	Fungi
Fusarium graminearum	Fgra	Ensembl Fungi 27	SRP048401	Fungi
Aspergillus oryzae	Aory	Ensembl Fungi 27	SRP016952	Fungi
Cryptococcus neoformans	Cneo	Broad Institute Fungi Initiative	SRR847297	Fungi
Ustilago maydis	Umay	Ensembl Fungi 27	ERP001905	Fungi
Tuber melanosporum	Tmel	Ensembl Fungi 27	SRP028655	Fungi
Rhizopus oryzae	Rory	Broad Institute Fungi Initiative	SRP031602	Fungi
Allomyces macrogynus	Amac	Broad Institute Fungi Initiative	SRP022576	Fungi
Spizellomyces punctatus	Spun	Broad Institute Multicellularity Initiative	SRR343043	Fungi
Rhizophagus irregularis	Rirr	Ensembl Fungi 27	DRP002784	Fungi
Conidiobolus coronatus	Ccor	JGI v1	SRR427173	Fungi
Gonapodya prolifera	Gpro	JGI v3	SRR427152	Fungi
Fonticula alba	Falb	Broad Institute Multicellularity Initiative	SRP022580	Nuclearida
Dictyostelium discoideum	Ddis	Ensembl Protist 27 / Dictybase	SRP001567	Amoebozoa
Polysphondylium pallidum	Ppal	Ensembl Protist 27	SRP004023	Amoebozoa
Acanthamoeba castellanii	Acas	Ensembl Protist 27	SRP028620	Amoebozoa
Arabidopsis thaliana	Atha	Ensembl Plants 27	SRP052858	Embryophyta
Vitis vinifera	Vvin	Ensembl Plants 27	SRP065417	Embryophyta
Mimulus guttatus	Mgut	JGI GCF_000504015	SRP045683	Embryophyta
Oryza sativa	Osat	Ensembl Plants 27	SRP008821	Embryophyta
Physcomitrella patens	Ppat	Ensembl Plants 27	SRP011279	Embryophyta
Selaginella moellendorffii	Smoe	Ensembl Plants 27	SRP059539	Embryophyta
Klebsormidium netis (formerly *flaccidum*)	Kfla	(Hori et al. 2014)	SRP048567	Charophyta
Volvox carteri	Vcar	JGI 317_v2	SRP066714	Chlorophyta
Chlamydomonas reinhardtii	Crei	Ensembl Plants 27	ERP001997	Chlorophyta
Micromonas pusilla	Mpus	JGI 20110615	SRR847305	Chlorophyta
Cyanophora paradoxa	Cpar	Assembly (Price et al. 2012); annotation in-home	SRR363339	Glaucophyta
Ectocarpus siliculosus	Esil	(Cock et al. 2010)	SRP037532	Stramenopile
Phytophthora infestans	Pinf	Ensembl Protist 27	SRR1640225	Stramenopile
Aureococcus anophagefferens	Aano	Ensembl Protist 27	SRP045642	Stramenopile
Plasmodium falciparum	Pfal	Ensembl Protist 27	SRP003507	Alveolata
Tetrahymena thermophila	Tthe	Ensembl Protist 27	SRP016619	Alveolata
Bigelowiella natans	Bnat	Ensembl Protist 27	DRP003230	Rhizaria
Emiliania huxleyi	Ehux	Ensembl Protist 27	SRR847300	Haptophyta
Guillardia theta	Gthe	Ensembl Protist 27	SRR747855	Cryptophyta
Naegleria gruberi	Ngru	Ensembl Protist 27	CRG	Excavata

Supplementary Table S2. Data-sets of ES events per species, including including ES weight (w_{ES}) the total number of genes in the genome, total number of ES-visible genes and triplets, and ES-positive or constitutive genes and triplets.

Species	w_{ES}	Total genes	Visible genes	Positive genes	Visible triplets	Positive triplets
Hsap	107.39	20,346	14,891	4,426	144,856	6,310
Drer	42.87	26,459	19,585	1,059	169,001	1,187
Xtro	38.58	18,442	14,601	1,002	133,811	1,121
Cint	49.25	16,671	11,746	596	68,839	634
Bflo	91.50	50,817	24,075	257	93,922	295
Spur	31.78	28,842	16,447	398	84,347	429
Dmel	57.69	13,917	7,702	464	27,996	548
Cele	50.40	20,447	16,841	710	79,024	797
Cgig	56.07	26,089	16,900	834	124,040	1,002
Sman	26.71	10,772	7,236	773	46,086	872
Tadh	0.82	11,520	8,350	25	73,070	547
Hmag	19.80	20,047	12,270	518	78,597	504
Nvec	14.75	24,773	11,887	482	73,927	138
Aipt	8.39	29,271	15,427	138	87,201	25
Mlei	15.10	16,058	9,252	206	56,851	220
Aque	5.14	40,122	15,418	210	91,738	214
Ocar	0.71	11,152	6,983	25	52,493	26
Scil	4.57	26,105	11,783	108	87,809	109
Sros	0.39	11,624	8,896	8	71,161	8
Cowc	0.98	8,741	5,219	24	19,951	24
Cfra	3.19	8,694	6,629	216	47,796	223
Sar3	13.49	16,015	8,598	557	60,636	631
Nk52	1.30	12,463	1,903	7	3,376	7
Ncra	34.45	9,820	4,199	366	7,795	389
Spom	0.71	5,144	1,213	11	2,370	11
Fgra	23.25	14,164	6,148	261	12,983	273
Aory	0.38	12,074	5,795	5	12,977	5
Cneo	0.86	6,962	5,952	13	26,757	13
Umay	0.79	6,522	1,070	3	2,115	3
Tmel	1.78	7,496	4,591	26	14,029	26
Rory	0.44	17,459	7,717	6	22,950	6
Amac	0.73	18,773	8,079	4	18,378	4
Spun	0.17	8,952	6,332	5	33,335	5
Rirr	5.56	29,822	13,180	87	42,323	90
Ccor	0.12	10,635	3,734	2	11,740	2
Gpro	0.86	13,902	9,873	13	44,517	14
Falb	0.16	5,881	4,261	5	15,994	5
Ddis	10.03	13,212	3,679	48	6,369	51
Ppal	0.13	12,367	7,038	1	19,689	1
Acas	0.43	14,973	10,816	3	68,986	3
Atha	8.19	27,416	15,993	277	86,629	285
Vvin	23.93	29,927	14,867	800	75,983	870
Mgut	13.23	27,232	15,376	260	82,922	267
Osat	19.54	35,679	12,811	657	52,292	703
Ppat	20.67	32,273	15,732	635	89,317	676
Smoe	3.80	34,799	6,552	6	18,027	6
Kfla	6.53	16,544	12,568	154	75,255	164
Vcar	18.11	14,247	11,115	595	78,656	667
Crei	4.31	14,416	11,466	97	82,361	101
Mpus	1.64	10,672	2,307	10	4,181	10
Cpar	0.90	11,011	7,754	4	45,819	4
Esil	9.68	16,271	13,196	375	90,831	413
Pinf	1.84	17,785	6,369	8	15,729	8
Aano	1.89	11,520	2,842	4	7,806	4
Pfal	4.28	5,349	1,556	65	5,308	73
Tthe	6.12	24,725	13,698	68	66,726	70
Bnat	8.13	21,706	16,463	176	143,972	193
Ehux	3.29	38,544	11,604	6	34,446	6
Gthe	1.89	24,945	15,848	60	108,207	61
Ngru	1.54	15,709	2,470	3	4,333	3

Supplementary Table S3. Data-sets of IR events per species, including IR weight (w_{IR}) total number of genes in the genome, total number of IR-visible genes and triplets, and IR-positive or constitutive genes and triplets.

Species	w_{IR}	Total genes	Visible genes	Positive genes	Visible triplets	Positive triplets
Hsap	1,344.51	20,346	16,531	4,490	162,698	21,395
Drer	184.70	26,459	21,946	1,625	193,407	2,380
Xtro	211.03	18,442	16,031	2,136	151,731	3,748
Cint	313.43	16,671	14,078	415	85,365	503
Bflo	353.85	50,817	34,702	423	150,101	552
Spur	199.17	28,842	20,920	1,152	110,549	1,520
Dmel	1,611.73	13,917	10,549	3,998	39,813	13,582
Cele	3,323.27	20,447	19,178	7,626	100,093	33,735
Cgig	321.09	26,089	20,312	1,752	144,222	2,809
Sman	475.60	10,772	8,553	1,361	55,011	2,624
Tadh	222.07	11,520	9,465	636	82,586	425
Hmag	79.11	20,047	14,773	395	93,672	5,402
Nvec	888.60	24,773	14,971	2,421	86,799	1,311
Aipt	815.01	29,271	20,774	809	98,169	949
Mlei	59.20	16,058	10,870	190	70,536	222
Aque	1,271.00	40,122	21,454	2,866	118,865	6,720
Ocar	331.31	11,152	8,271	1,340	60,527	2,413
Scil	367.22	26,105	14,501	776	108,054	1,113
Sros	69.86	11,624	10,091	213	80,806	244
Cowc	496.20	8,741	7,052	1,330	28,985	2,442
Cfra	481.94	8,694	7,271	2,258	55,128	4,473
Sar3	137.69	16,015	10,566	734	69,650	1,106
Nk52	631.50	12,463	4,042	1,085	7,381	1,673
Ncra	2,839.07	9,820	7,734	4,059	15,648	6,406
Spom	438.14	5,144	2,279	875	4,664	1,468
Fgra	3,973.24	14,164	10,477	4,823	23,618	8,192
Aory	521.59	12,074	9,041	866	21,923	1,180
Cneo	436.94	6,962	6,604	598	33,903	771
Umay	397.86	6,522	2,276	488	4,340	679
Tmel	671.58	7,496	6,012	2,429	20,211	5,296
Rory	897.03	17,459	11,435	1,990	34,603	2,926
Amac	535.53	18,773	13,602	339	35,107	387
Spun	496.71	8,952	7,630	573	41,380	770
Rirr	1,180.46	29,822	21,017	1,282	64,074	1,982
Ccor	51.16	10,635	5,604	73	17,340	83
Gpro	219.90	13,902	11,881	1,013	57,292	1,345
Falb	49.56	5,881	5,151	122	20,846	143
Ddis	2,142.69	13,212	7,782	1,156	14,132	1,620
Ppal	120.31	12,367	9,933	222	29,457	252
Acas	13.02	14,973	12,767	50	81,438	55
Atha	1,048.28	27,416	20,143	6,530	108,390	15,505
Vvin	1,013.26	29,927	20,353	3,376	96,103	7,241
Mgut	472.05	27,232	20,055	3,178	103,976	5,399
Osat	5,232.34	35,679	17,133	8,077	55,033	18,535
Ppat	1,657.58	32,273	21,468	2,812	111,538	5,008
Smoe	1,930.30	34,799	11,739	348	26,740	426
Kfla	410.10	16,544	14,265	2,050	90,186	3,274
Vcar	1,619.91	14,247	12,260	2,900	90,756	8,291
Crei	452.33	14,416	12,445	2,237	95,443	4,137
Mpus	1,550.82	10,672	5,256	250	9,179	425
Cpar	137.96	11,011	9,167	73	56,605	119
Esil	381.87	16,271	14,642	3,207	105,794	5,928
Pinf	1,141.44	17,785	10,300	1,121	25,514	1,654
Aano	1,504.97	11,520	5,026	357	12,199	664
Pfal	470.07	5,349	2,656	852	7,964	1,708
Tthe	1,153.58	24,725	16,971	2,367	83,324	4,449
Bnat	1,487.28	21,706	18,183	1,597	162,525	3,429
Ehux	4,304.29	38,544	16,331	258	45,933	360
Gthe	1,431.51	24,945	18,641	3,577	126,930	10,694
Ngru	175.12	15,709	5,367	356	9,668	421

Supplementary Table S4. Area under (AU) the ROC curve for the ES and IR logistic regression models, per species.

Species	AU ROC ES	AU ROC IR
Hsap	0.6581	0.77
Drer	0.7717	0.7628
Xtro	0.7133	0.7748
Cint	0.6545	0.6699
Bflo	0.7528	0.6692
Spur	0.7381	0.6680
Dmel	0.8361	0.7084
Cele	0.8087	0.7116
Cgig	0.7559	0.7420
Sman	0.6481	0.7891
Tadh	0.6723	0.7231
Hmag	0.7015	0.6464
Nvec	0.6737	0.6294
Aipt	0.7526	0.6745
Mlei	0.7224	0.6513
Aque	0.7040	0.7287
Ocar	0.5380	0.6609
Scil	0.6945	0.7282
Sros	0.6055	0.7473
Cowc	0.6944	0.6596
Cfra	0.7772	0.7662
Sar3	0.7377	0.7349
Nk52	0.6393	0.7161
Ncra	0.7299	0.7891
Spom	0.7505	0.6949
Fgra	0.7079	0.7543
Aory	0.7975	0.6645
Cneo	0.7513	0.6525
Umay	0.4758	0.7310
Tmel	0.8563	0.7228
Rory	0.6544	0.6217
Amac	0.7280	0.7034
Spun	0.7275	0.7155
Rirr	0.6306	0.6941
Ccor	0.9607	0.7856
Gpro	0.8752	0.6895
Falb	0.6315	0.6589
Ddis	0.6277	0.8366
Ppal	0.9472	0.7406
Acas	0.7713	0.6926
Atha	0.8080	0.7047
Vvin	0.7109	0.7496
Mgut	0.7702	0.7167
Osat	0.6699	0.6610
Ppat	0.6879	0.7203
Smoe	0.6625	0.5830
Kfla	0.7343	0.7358
Vcar	0.6716	0.6691
Crei	0.7537	0.7108
Mpus	0.7317	0.8136
Cpar	0.5677	0.7966
Esil	0.7942	0.7267
Pinf	0.7051	0.6833
Aano	0.3106	0.7337
Pfal	0.8248	0.7277
Tthe	0.7390	0.7155
Bnat	0.7591	0.7578
Ehux	0.6301	0.7232
Gthe	0.6509	0.6285
Ngru	0.3813	0.7188

Supplementary Table S5. Consistency of significant coefficients of the species-specific ES logistic regression models ($p<0.05$ Wald tests), per feature. In green, manually selected most consistent predictors.

Gene feature	Coef. >0	Coef. <0	# significant species	Coef. >0 %	Coef. <0 %
Gene length	11	7	18	61.11%	38.89%
CDS length	5	7	12	41.67%	58.33%
Introns length	0	0	0	NA	NA
intron/CDS lenght	4	4	8	50.00%	50.00%
intron number	14	4	18	**77.78%**	22.22%
intron density	4	6	10	40.00%	60.00%
exon length	7	17	24	29.17%	**70.83%**
5'intron length	6	4	10	60.00%	40.00%
3'intron length	9	4	13	69.23%	30.77%
3'intron/exon length	2	9	11	18.18%	**81.82%**
5'intron/exon length	2	8	10	20.00%	**80.00%**
exon GC	3	22	25	12.00%	**88.00%**
3'intron GC	8	8	16	50.00%	50.00%
5'intron GC	13	6	19	68.42%	31.58%
3'intron-exon GC	0	0	0	NA	NA
5'intron-exon GC	0	0	0	NA	NA
5'SS score	1	32	33	3.03%	**96.97%**
3'SS score	1	40	41	2.44%	**97.56%**
cRPKM	0	24	24	0.00%	**100.00%**
position within gene	2	4	6	33.33%	66.67%
3n-divisible	8	13	21	38.10%	61.9

Supplementary Table S6. Consistency of significant coefficients ($p<0.05$) of the species-specific IR logistic regression models ($p<0.05$ Wald tests), per feature. In green, manually selected most consistent predictors.

Gene feature	Coef. >0	Coef. <0	# significant species	Coef. >0 %	Coef. <0 %
Gene length	6	24	30	20.00%	**80.00%**
CDS length	6	22	28	21.43%	**78.57%**
Introns length	0	0	0	NA	NA
intron/CDS lenght	6	20	26	23.08%	**76.92%**
intron number	4	24	28	14.29%	**85.71%**
intron density	6	20	26	23.08%	**76.92%**
intron length	27	5	32	**84.38%**	15.63%
5'exon length	8	9	17	47.06%	52.94%
3'exon length	13	1	14	**92.86%**	7.14%
5'intron/exon length	20	2	22	**90.91%**	9.09%
3'intron/exon length	21	2	23	**91.30%**	8.70%
intron GC	7	40	47	14.89%	**85.11%**
5'exon GC	6	22	28	21.43%	**78.57%**
3'exon GC	5	25	30	16.67%	**83.33%**
5'intron-exon GC	0	0	0	NA	NA
3'intron-exon GC	0	0	0	NA	NA
5'SS score	0	51	51	0.00%	**100.00%**
3'SS score	1	47	48	2.08%	**97.92%**
cRPKM	4	38	42	9.52%	**90.48%**
position within gene	0	0	0	NA	NA
3n-divisible	1	3	4	25.00%	75.00%
coding probability	12	3	15	**80.00%**	20.00%

Supplementary Table S7. Estimated fraction of erroneous IR events per species, and associated p_{IR} values. In red, species that violate assumptions of (Saudemont et al. 2017).

Species	F_{err}	$p_{IR,m}/p_{IR,hc}$	$p_{IR,m}$	$p_{IR,hc}$	# median triplets	# highly constrained triplets
Hsap	0.7211	3.59	9.89	2.76	23,329	1,446
Drer	0.6398	2.78	1.07	0.38	30,353	1,785
Xtro	0.5995	2.50	1.33	0.53	22,352	1,387
Cint	0.4213	1.73	2.84	1.64	3,492	335
Bflo	0.6826	3.15	0.79	0.25	8,329	633
Spur	0.5242	2.10	1.04	0.49	19,801	1,369
Dmel	*0.1332*	1.15	18.63	16.15	8,424	639
Cele	*0.2702*	1.37	19.57	14.28	20,909	1,647
Cgig	0.3300	1.49	1.54	1.03	20,680	1,638
Sman	0.6936	3.26	6.16	1.89	4,800	481
Tadh	0.3888	1.64	2.90	1.77	4,092	319
Hmag	0.6351	2.74	0.59	0.21	13,815	1,009
Nvec	0.5259	2.11	7.93	3.76	8,292	772
Aipt	0.6731	3.06	4.92	1.61	4,124	331
Mlei	0.4824	1.93	0.54	0.28	7,080	433
Aque	0.8173	5.47	6.52	1.19	12,583	1,151
Ocar	0.6272	2.68	4.49	1.67	7,103	747
Scil	0.7065	3.41	3.04	0.89	5,080	424
Sros	0.6118	2.58	0.65	0.25	4,084	413
Cowc	0.5461	2.20	8.03	3.65	3,484	185
Cfra	0.7375	3.81	8.72	2.29	8,440	933
Sar3	0.6427	2.80	1.63	0.58	9,903	858
Nk52	*-0.1750*	0.85	20.24	23.78	1,375	32
Ncra	0.2469	1.33	41.33	31.12	4,906	239
Spom	0.5126	2.05	24.70	12.04	1,009	43
Fgra	0.0728	1.08	41.04	38.05	5,742	274
Aory	*-0.0198*	0.98	7.62	7.77	2,336	104
Cneo	0.6628	2.97	6.72	2.27	2,021	177
Umay	0.3082	1.45	22.97	15.89	421	16
Tmel	0.4285	1.75	12.51	7.15	4,143	301
Rory	0.6137	2.59	9.27	3.58	3,729	313
Amac	0.2156	1.27	3.77	2.96	1,235	95
Spun	0.5408	2.18	8.59	3.94	1,533	107
Rirr	0.7148	3.51	7.09	2.02	4,077	335
Ccor	0.8320	5.95	1.09	0.18	1,541	78
Gpro	0.7955	4.89	2.45	0.50	7,869	547
Falb	0.2801	1.39	1.06	0.76	1,687	112
Ddis	*-0.0540*	0.95	26.78	28.23	1,768	72
Ppal	*-0.0789*	0.93	1.31	1.41	2,296	227
Acas	0.6687	3.02	0.12	0.04	11,330	1,218
Atha	0.6216	2.64	6.00	2.27	23,979	2,305
Vvin	0.7885	4.73	5.97	1.26	13,395	1,039
Mgut	0.7428	3.89	3.00	0.77	18,909	1,598
Osat	0.2410	1.32	31.30	23.75	16,885	1,450
Ppat	0.6891	3.22	9.11	2.83	7,941	856
Smoe	0.5039	2.02	17.73	8.80	525	55
Kfla	0.5547	2.25	3.31	1.47	12,270	1,203
Vcar	*0.0539*	1.06	13.24	12.53	7,392	925
Crei	0.6039	2.52	4.42	1.75	9,411	1,098
Mpus	*0.0724*	1.08	33.93	31.48	325	3
Cpar	0.8968	9.69	1.68	0.17	1,042	62
Esil	0.8056	5.14	3.32	0.65	19,438	2,115
Pinf	0.3679	1.58	13.92	8.80	1,885	110
Aano	*0.1698*	1.20	34.62	28.74	419	16
Pfal	0.2240	1.29	22.53	17.48	1,181	53
Tthe	0.5329	2.14	7.51	3.51	5,585	582
Bnat	0.4944	1.98	8.91	4.50	5,091	488
Ehux	*0.1272*	1.15	24.44	21.33	369	24
Gthe	0.3259	1.48	8.18	5.51	16,225	1,334
Ngru	0.1650	1.20	4.04	3.37	1,188	69

VII. References

Ast G. 2004. How did alternative splicing evolve? *Nat Rev Genet* **5**: 773–782.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science (80-)* **338**: 1587–93.

Baumgarten S, Simakov O, Esherick LY, Liew YJ, Lehnert EM, Michell CT, Li Y, Hambleton EA, Guse A, Oates ME, et al. 2015. The genome of Aiptasia, a sea anemone model for coral symbiosis. *Proc Natl Acad Sci* **112**: 201513318.

Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev* **29**: 63–80.

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 Res* **24**: 1774–1786.

Breitbart RE, Andreadis A, Nadal-Ginard B. 1987. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu Rev Biochem* **56**: 467–495.

Brogna S, McLeod T, Petric M. 2016. The Meaning of NMD: Translate or Perish. *Trends Genet* **xx**: 1–13.

Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM. 2012. Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Mol Cell* **46**: 871–883.

Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO. 2017. Alternative splicing and the evolution of phenotypic novelty. *Philos Trans R Soc B Biol Sci* **372**: 20150474.

Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Aury J-M, Badger JH, et al. 2010. The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* **465**: 617–21.

Collins L, Penny D. 2005. Complex Spliceosomal Organization Ancestral to Extant Eukaryotes. *Mol Biol Evol* **22**: 1053–1066.

Csűrös M, Rogozin IB, Koonin E V. 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes ed. C.P. Ponting. *PLoS Comput Biol* **7**: e1002150.

Curtis B a, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH, Hirakawa Y, et al. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492**: 59–65.

De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* **4**: 49–60.

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**: 7250–7.

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

Ellis JD, Barrios-Rodiles M, Çolak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM, et al. 2012. Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. *Mol Cell* **46**: 884–892.

Ezkurdia I, Rodriguez JM, Carrillo-De Santa Pau E, Vázquez J, Valencia A, Tress ML. 2015. Most highly expressed protein-coding genes have a single dominant isoform. *J Proteome Res* **14**: 1880–1887.

Fortunato S a. V., Adamski M, Ramos OM, Leininger S, Liu J, Ferrier DEK, Adamska M. 2014. Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. *Nature* **514**: 620–623.

Gordon A. 2017. FASTX Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/.

Gracheva EO, Cordero-Morales JF, González-Carcacía JA, Ingolia NT, Manno C, Aranguren CI, Weissman JS, Julius D. 2011. Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats. *Nature* **476**: 88–91.

Graveley BR. 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* **17**: 100–107.

Gueroussov S, Gonatopoulos-Pournatzis T, Irimia M, Raj B, Lin Z-Y, Gingras A-C, Blencowe BJ. 2015. An alternative splicing event amplifies evolutionary differences between vertebrates. *Science (80-)* **349**: 868–873.

Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, Sato S, Yamada T, Mori H, Tajima N, et al. 2014. Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nat Commun* **5**: 3978.

Irimia M, Penny D, Roy SW. 2007. Coevolution of genomic intron number and splice sites. *Trends Genet* **23**: 321–325.

Irimia M, Roy SW. 2008. Evolutionary Convergence on Highly-Conserved 3′ Intron Structures in Intron-Poor Eukaryotes and Insights into the Ancestral Eukaryotic Genome ed. B.J. Trask. *PLoS Genet* **4**: e1000148.

Irimia M, Roy SW. 2014. Origin of Spliceosomal Introns and Alternative Splicing. *Cold Spring Harb Perspect Biol* **6**.

Irimia M, Rukov JL, Penny D, Garcia-Fernandez J, Vinther J, Roy SW. 2008. Widespread evolutionary conservation of alternatively spliced exons in Caenorhabditis. *Mol Biol Evol* **25**: 375–382.

Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S. 2013. Function of alternative splicing. *Gene* **514**: 1–30.

Kempken F. 2013. Alternative splicing in ascomycetes. *Appl Microbiol Biotechnol* **97**: 4235–4241.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* **11**: 345–55.

Kim E, Goren A, Ast G. 2008. Alternative splicing: current perspectives. *BioEssays* **30**: 38–47.

Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res* **35**: 125–131.

Koonin E V, Csuros M, Rogozin IB. 2013. Whence genes in pieces: Reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. *Wiley Interdiscip Rev RNA* **4**: 93–105.

Labbé RM, Irimia M, Currie KW, Lin A, Zhu SJ, Brown DDR, Ross EJ, Voisin V, Bader GD, Blencowe BJ, et al. 2012. A Comparative Transcriptomic Analysis Reveals Conserved Features of Stem Cell Pluripotency in Planarians and Mammals. *Stem Cells* **30**: 1734–1745.

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Le Hir H, Nott A, Moore MJ. 2003. How introns influence and enhance eukaryotic gene expression. *Trends Biochem Sci* **28**: 215–220.

Liebert MA, Yeo G, Burge CB. 2004. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. **11**: 377–394.

Lykke-Andersen S, Jensen TH. 2015. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol* **16**: 665–677.

McGuire AM, Pearson MD, Neafsey DE, Galagan JE. 2008. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol* **9**: R50.

Melamud E, Moult J. 2009. Stochastic noise in splicing machinery. *Nucleic Acids Res* **37**: 4873–4886.

Nichols SA, Roberts BW, Richter DJ, Fairclough SR, King N. 2012. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/ -catenin complex. *Proc Natl Acad Sci* **109**: 13046–13051.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–63.

Pages H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: String objects representing biological sequences, and matching algorithms.

Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* **6**: 1–11.

Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber a. PM, Schwacke R, Gross J, Blouin N a., Lane C, et al. 2012. Cyanophora paradoxa Genome Elucidates Origin of Photosynthesis in Algae and Plants. *Science (80-)* **335**: 843–847.

R Core Team. 2015. R: A Language and Environment for Statistical Computing.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**: 77.

Roy SW, Irimia M. 2009. Splicing in the eukaryotic ancestor: form, function and dysfunction. *Trends Ecol Evol* **24**: 447–55.

Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necsulea A, Meyer E, Duret L. 2017. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *bioRxiv* 114215.

Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res* **18**: 88–103.

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.

Sorber K, Dimon MT, DeRisi JL. 2011. RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Res* **39**: 3820–3835.

Sorek R, Shamir R, Ast G. 2004. How prevalent is functional alternative splicing in the human genome? *Trends Genet* **20**: 68–71.

Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E, Carr M, Kerner P, Vervoort M, Sánchez-Pons N, et al. 2013. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. *Nat Commun* **4**: 2325.

Wong JJ-L, Au AYM, Ritchie W, Rasko JEJ. 2016. Intron retention in mRNA: No longer nonsense. *BioEssays* **38**: 41–49.

Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, Sun S, Yang F, Shen YA, Murray RR, et al. 2016. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164**: 805–817.

4. Discussion

Reconstruction of premetazoan genomes: when, whence.

God loves the noise as much as the signal.

Lewis M. Branscomb, 1980

4.1. Complementary views of ancestral Metazoa

The catalog of protistan relatives of Metazoa has expanded over the last decades thanks to the phylogenomic investigation of Opisthokonta (Lang *et al.* 2002; Cavalier-Smith and Chao 2003; Steenkamp *et al.* 2006; Ruiz-Trillo *et al.* 2008; Shalchian-Tabrizi *et al.* 2008; Torruella *et al.* 2012, 2015) and now includes four clades of unicellular holozoans branching after the divergence of Fungi: Choanoflagellata, Filasterea, Ichthyosporea and Corallochytrea. In the sections *R3 – Opisthokonta phylogenomics* and *R6 – Teretosporea genomes* we clarified the relationships between holozoan species by establishing the early-branching position of Teretosporea within Holozoa (the grouping of Ichthyosporea and *C. limacisporum*) and rejecting the 'Filasporea' hypothesis (the proposed grouping of Filasterea and Ichthyosporea). This stabilized opisthokont tree of life is thus essential to the interpretation of the diverse diverse array of cell morphologies and lifestyles exhibited by holozoan protists (Figure 4).

Key insights into the biology of the earliest animals have been obtained from the comparative analysis of animals and choanoflagellates (James-Clark 1866, 1871; King *et al.* 2008; Fairclough *et al.* 2013; Richter and King 2013), combined with the study of the palaeontological record and the Proterozoic-Phanerozoic ecology (Stanley 1973; Knoll 2011, 2014; Huldtgren *et al.* 2011; Budd and Jensen 2015; Cunningham *et al.* 2016). For example, we can infer that metazoans descend from an heterotrophic protist, that might as well have been eukaryvore or bacterivore (Stanley 1973; Knoll 2014); and comparisons of cell types and genetic tool-kits between choanoflagellates and animals suggest that it could have had a choanoflagellate-like microvilli collar around a single posterior flagellum (Maldonado 2004; Abedin and King 2008; Alegado and King 2014; Mah *et al.* 2014). Many illuminating analyses have revolved around the coloniality of choanoflagellates (Carr *et al.* 2017), including its transcriptomic characterization in *S. rosetta* (Fairclough *et al.* 2013) and the discovery of bacteria-produced environmental cues that trigger and enhance this multicellular behavior – thus pointing at shared environmental pressures with scattered metazoans that undergo similar processes during their development (Alegado *et al.* 2012; Woznica *et al.* 2016) and stressing a plausible close relationship between choanoflagellates and bacteria (as proposed in Yue et al. 2013). These similarities do not necessarily reflect the premetazoan state of key animal characters like feeding mode or coloniality, but are essential to understand the evolutionary pressures the early animals were subject to, and to develop better models to understand the transition to multicellularity (McFall-Ngai *et al.* 2013; Alegado and King 2014; Cavalier-Smith 2017).

In parallel, the inclusion of filastereans, ichthyosporeans and *C. limacisporum* in these comparative analyses with animals has allowed to expand the outlook by analyzing a wider diversity of unicellular lifestyles and, most interestingly, multicellular-like behaviours other than choanoflagellate coloniality: *C. limacisporum* produces tetrad cell assemblages that divide by palintomic cleavage (Raghukumar 1987), the ichthyosporean *C. fragrantissima* exhibits coordinated mitotic division in its multicnucleated stage (Suga and Ruiz-Trillo 2013), and *C. owczarzaki* can produce multicellular aggregates (Sebé-Pedrós *et al.* 2013a, 2016a). Crucially, there is evidence of

regulated temporal cell differentiation in all extant unicellular holozoan clades – at the transcriptomic (Sebé-Pedrós *et al.* 2013a; Fairclough *et al.* 2013; de Mendoza *et al.* 2015), proteomic (Sebé-Pedrós *et al.* 2016a) and epigenomic levels (Sebé-Pedrós *et al.* 2016b). The resulting perspective of premetazoan protists thus points at a relatively plastic ancestor with a complex life cycle and multiple cell types (Torruella 2014; Sebé-Pedrós *et al.* 2017). These observations have led to a renewed interest in views of animal multicellularity as a mosaic of premetazoan cell types that appeared after a temporal-to-spatial switch in cell differentiation programs (Mikhailov *et al.* 2009; Budd and Jensen 2015; Sebé-Pedrós *et al.* 2017) – originally formulated as the 'synzoospore' hypothesis (Zakhvatkin 1949). These views contrast with other proposals such as the 'placula' (Bütschli 1884) or the 'choanoblastea' hypotheses (a choanoflagellate-like, planctonic colony of undifferentiated protists later evolving spatial cell types, *cf.* Nielsen 2008; inspired by the Haeckelian 'gastrea', *cf.* Haeckel 1874).

This suggestive sketch of early animals can be greatly improved by combining comparative genomic analyses (King *et al.* 2008; Suga *et al.* 2013; Fairclough *et al.* 2013; Richter and King 2013) with the above-mentioned insights into the cell biology and development of unicellular holozoans. Specifically, under a clear phylogenetic framework of opisthokonts (*R3 – Opisthokonta phylogenomics* and *R6 – Teretosporea genomes*), we can confidently reconstruct key traits from the genomes of the metazoan ancestors before and after the transition to multicellularity, including gene content (as in sections *R1 – HECT*, *R2 – Myosin*, *R4 – Ubiquitin signaling*, *R5 – LOX*, *R6 – Teretosporea genomes*), genome size and architecture (section *R6 – Teretosporea genomes*), and even some aspects of the transcriptome regulation profile (section *R7 – Alternative splicing*). In the following sections, I will discuss how comparative genome analyses, including phylogenomics and ancestral reconstructions, can improve our understanding of animal origins.

4.2. Present eyes on past genomes: interpretations of ancestral reconstruction

One of the main possibilities opened up by comparative genomics is the reconstruction of ancestral genomes, and this particular type of evolutionary inference features prominently in the results presented in this dissertation (sections *R1 – HECT*, *R2 – Myosin*, *R4 – Ubiquitin signaling*, *R5 – LOX*, *R6 – Teretosporea genomes*). In most of these analyses, 'ancestral reconstruction' is a short-hand for inferring ancestral gene contents, mostly because genes (or their sub-parts, *i.e.* protein domains) are a tractable unit of homology that can be readily analyzed within a species phylogenetic framework. However, in the last section of my results I have also reconstructed ancestral states for other genomic characters, namely gene linkage and intron density (section *R6 – Teretosporea genomes*).

'Raw' ancestral reconstructions (*e.g.* a plain list of genes present in a specific ancestor) can be interpreted under different points of view. In the above-mentioned gene family analyses, I originally proposed different explanatory frameworks to construe the evolutionary processes. In particular, I have discussed the following propositions, in some form or another: 1) genome evolution is dominated by punctuated bursts of innovation followed by stasis and/or trait loss

(Wolf and Koonin 2013); 2) genomic innovations that play important roles in a given lineage can pre-date the divergence of that lineage, a circumstance referred to as pre-adaptation or exaptation (Gould and Vrba 1982); 3) phylogenetic inertia can shape lineage-specific innovations (Burt 2001). These points of view are not mutually exclusive and can sometimes build one on another.

In this section, I aim to re-examine my previous analyses of gene family evolution under the light of the above-mentioned explanatory frameworks. First, I will concisely present each framework, with its strengths and limitations, for the sake of clarity. Then, I will examine each ancestral reconstruction analysis separately (sections *R1 – HECT* and *R4 – Ubiquitin signaling*, *R2 – Myosin*, *R5 – LOX*, *R6 – Teretosporea genomes*) and discuss its interpretation under these complementary perspectives of genome evolution.

4.2.1. Three explanatory frameworks for genome evolution

A biphasic model: complexification and simplification

Wolf and Koonin (2013) argued that ancestral reconstructions of genome complexity exhibited a biphasic mode of evolution, in which gains are sudden and significant bursts of change coinciding with evolutionary radiations (and phenotypic/genomic innovation), whereas losses of complexity occur gradually at a quasi-regular pace. Key to this assessment is the fact that 'genomic complexity' can be associated with multiple traits for which ancestral reconstructions are possible: evolution of gene families, including paralogy and lateral gene transfers (Embley and Martin 2006; Lane and Martin 2010; Corradi and Slamovits 2011; Wolf *et al.* 2012; Schönknecht *et al.* 2014; Corradi 2015; Groussin *et al.* 2015; Pittis and Gabaldón 2016); origin of new protein domain folds (Zmasek and Godzik 2011, 2012); or analysis of intron content (Carmel *et al.* 2007b; Csűrös *et al.* 2011). The biphasic burst-and-loss model would act in specific gene families or genetic tool-kits and could be identified in all domains of life – from the last common ancestors of Archaea, Bacteria or Eukaryota, to more recent ancestors within these groups. Given that different gene sets can be independently affected by burst-and-loss processes even in the same line of descent, O'Malley *et al.* (2016) have recently emphasized the need to explicitly account for both complexification and simplification in comparative genomic studies.

Functional co-option of ancestral genes

The term 'exaptation' was proposed by Gould and Vrba (1982) to characterize evolutionary traits that appear as by-products of evolutionary processes other than their current adaptive function. 'Pre-adaptation', despite its arguably teleological connotations (Larson *et al.* 2013), is a similar term for this concept. Whatever the reasons for the emergence of a given trait (*e.g.* a specific gene), it can be later co-opted into new functions and gain new adaptive values. Many examples of this effect exist in the context of multicellularity, *e.g.* the Hippo/Warts signaling pathway (Sebé-Pedrós *et al.* 2012), the integrin adhesome (Sebé-Pedrós *et al.* 2010; de Mendoza *et al.* 2015), cadherins (Abedin and King 2008; Nichols *et al.* 2012) or developmental transcription factors such as *Brachyury* (Sebé-Pedrós *et al.* 2013b).

Path-dependence or inertia along the tree of life

Burt (2001) defined phylogenetic inertia as an evolutionary path during which characters with unchanged states will remain unchanged, while characters expressing directional change will maintain that trend (barring the effect of external forces) between different phylogenetic lineages. Therefore, it can help explain effects of path-dependence in genome evolution, *e.g.* the lineage-specific diversifications and enrichments of the transcription factor complement across eukaryotes, where major lineages make preferential use of different DNA-binding motifs (de Mendoza *et al.* 2013). Phylogenetic inertia has also been proposed to explain the conservation of *Hox* clusters in *Drosophila*: fly development does not require the precise temporal activation provided by shared regulatory elements in other animal *Hox* clusters, but it nevertheless maintains some remnants of gene collinearity (Duboule 2007; Negre and Ruiz 2007).

4.2.2. Phylogenetic inertia and adaptive potential shape the evolution of ubiquitin signaling

Two of the chapters of the present thesis examine the evolution of various genes involved in the ubiquitin signaling system across eukaryotes: I first presented a detailed examination of HECT origin and evolution using phylogenetic inference (section *R1 – HECT*), later complemented by expanding the comparative analysis to the complete set of enzymes involved in the ubiquitination pathway (section *R4 – Ubiquitin signaling*).

In this later survey of ubiquitin signaling genes across eukaryotes (section *R4 – Ubiquitin signaling*), we proposed the burst-and-loss archetype (Wolf and Koonin 2013) to explain the origin and diversification of ubiquitin ligases (also known as E3 enzymes). Indeed, the vast majority of enzyme classes involved in ubiquitin transfer in eukaryotes appeared in the lineage leading up to the last eukaryotic common ancestor, while later lineages were dominated by losses: 18 out of 20 protein families involved in ubiquitin ligation are exclusive eukaryotic innovations, including many variants of the RING E3s genes found in Archaea and a novel enzymatic fold (HECT) to perform the same function.

Our earlier examination of HECT evolution (section *R1 – HECT*), however, offered a more nuanced view of this early eukaryotic burst in ubiquitin ligase diversity. The first eukaryotic HECT gene had already duplicated into six classes (some of which with distinct multi-domain protein syntaxes) by the time the last eukaryotic common ancestor gave rise to extant lineages. Afterwards, further diversification occurred independently in the root of Bikonta/Diaphoratickes (five new families) and Unikonta/Amorphea (four), and this diversification process persisted in the opisthokont lineage (19 new families appearing at different nodes). Overall, we found that holozoans generally contained higher numbers of HECT genes than Fungi or bikont lineages (including the complex multicellular land plants), and that these were often more architecturally diverse as well. Given that ubiquitin signaling is ubiquitous across eukaryotes (Aravind *et al.* 2006; Hochstrasser 2009) and the necessity of regulated protein turnover is equally widespread, this result raised the hypothesis that some form of phylogenetic inertia (or path-dependence) in the early evolution of the HECT enzymes could explain the differences observed in extant genomes.

We followed the thread of this hypothesis in the above-mentioned analysis of the whole ubiquitin signaling pathway (section *R4 – Ubiquitin signaling*), where we examined the effects of phylogenetic inertia at different levels. This study consisted of a general reconstruction of ubiquitin signaling systems in eukaryotes, focusing not only on E3 ligases but also on E1 and E2 enzymes (upstream pathway elements responsible for ubiquitin activation and conjugation to the ligase, respectively), as well as deubiquitinase enzymes (antagonists of E3 ligases that remove the ubiquitin label from marked proteins).

Our first observation was that lineage-specific expansions do not equally affect all gene families involved in the pathway: paralogy and domain shuffling were common in E3 and deubiquitinases at the LCA of Holozoa, Metazoa and Embryophyta; but E1 and E2 enzyme families did not undergo lineage-specific expansions in any of the surveyed lineages. E1 and E2 genes code for non-ubiquitin-specific enzymes, co-opted from pre-existing prokaryotic pathways (Iyer *et al.* 2006; Burroughs *et al.* 2008, 2009; Michelle *et al.* 2009). This result fits the pattern proposed by Lespinet et al. (2002) regarding the role of lineage-specific gene expansions in genome evolution: the downstream effectors of signaling pathways are among the gene families most prone to undergo diversification processes, either via gene paralogy, protein domain shuffling (Basu *et al.* 2009), and/or sequence divergence (Kondrashov *et al.* 2002). The reasoning behind the prediction by Lespinet et al. (2002) is that the diversification of proximal signaling genes directly influences the interactions with their substrates, thus bearing a higher potential for evolutionary adaptations.

Second, as mentioned above, both HECTs and the diverse class of RING-like E3 enzymes are able to perform the ubiquitin ligation reaction (Deshaies and Joazeiro 2009; Rotin and Kumar 2009). This functional redundancy offered the possibility to test whether phylogenetic inertia could be affecting differently each eukaryotic lineage – favoring the expansion of different enzyme classes in different lineages. This was indeed the case: our quantitative analysis of gene content recovered the expansion of HECT in the LCA of Holozoa (among other enzymes; concordantly with section *R1 – HECT*); whereas the ubiquitin signaling expansions in the green lineage were mostly driven by variants of the RING zinc fingers (particularly in the LCA of Embryophyta). Our follow-up analysis of protein domain diversity within each enzyme family further stressed this point: the independent serial paralogy events occurred in both embryophytes and holozoans were accompanied by differential acquisitions of accessory protein domains, which contributed to different functional specificities.

Finally, we also observed how the bias towards diversifying the downstream effectors is not homogeneous across different signaling pathways. Indeed, in parallel to our analysis of ubiquitination, we surveyed the evolution of SUMO-mediated signaling as well. SUMO is a ubiquitin-like small signaling peptide whose transduction is mediated by E1 and E2 enzymes shared with the ubiquitination system, plus specific E3 ligases and de-SUMOylation enzymes (Hochstrasser 2000). Instead of having expanded its E3 complement, most eukaryotes have diversified its set of de-SUMOylation enzymes – the other downstream effector of the pathway. Evidence exists in *A. thaliana* that SUMO E3 paralogs are often functionally redundant, whereas de-SUMOylation enzymes paralogs can be substrate-specific (Chosed *et al.* 2006; Colby *et al.* 2006).

Thus, for both SUMOylation and ubiquitination, we identify an expansion of different types of downstream effectors, combined with relative stasis in the upstream generalist enzymes.

The evolvability of reversible signaling pathways

As mentioned above, the higher frequency of diversification in downstream E3 and de-labeling enzyme families (compared to the upstream and generalist E1 and E2) can be attributed to their substrate-specific role, and hypothetically linked to higher adaptive potential (Lespinet *et al.* 2002).

In order to gain further insights into the evolvability of signaling pathways, I will compare ubiquitin signaling with another widespread post-translational modification (PTM) system – protein phosphorylation by tyrosine kinases (Mulder 1839; Krebs and Fischer 1956; Suga *et al.* 2012, 2014). Coincidentally, both PTM systems underwent lineage-specific diversifications of effector enzymes in the LCA of Holozoa (Suga *et al.* 2012, 2014), and both are based on antagonistic writer/eraser enzymes with varying levels of substrate specificity (E3 ligases/deubiquitinases and kinases/phosphatases, respectivley). By analyzing protein phosphorylation dynamics, it has been demonstrated that writer/eraser systems permit a fine-tuned control of the signal deposition levels, which has been proposed to yield multiple benefits such as reduced noise, robustness to perturbation, accuracy (Lee and Yaffe 2016) and higher temporal and spatial specificity (Beltrao *et al.* 2013).

Another similarity between ubiquitination and phosphorylation is that lineage-specific diversifications do not evenly affect all the pathways' enzymes. Indeed, tyrosine kinases can be divided in two broad classes: cytoplasmic and membrane-bound receptors. The former are widely conserved across holozoans, whereas the latter are often subject to lineage-specific paralogy and domain-shuffling events (Suga *et al.* 2012, 2014; Fairclough *et al.* 2013) that tune many phosporylation-dependent processes in holozoans: environment sensing in choanoflagellates (King *et al.* 2003), cell adhesion and communication in *C. owczarzaki* (Sebé-Pedrós *et al.* 2010, 2013a, 2016a), or development in metazoans (Richards and Degnan 2009). These dual evolutionary dynamics are reminiscent of the alternate expansion of E3 and deubiquitinases across eukaryotes (HECTs in holozoans, various RINGs in embryophytes, etc.), and are also common in other signaling systems such as GPCRs (de Mendoza *et al.* 2014), Hippo/Warts (Sebé-Pedrós *et al.* 2012), or the animal- and choanoflagellate-specific repertoires of cadherins (Nichols *et al.* 2012).

A recent investigation of the phosphorylation dynamics in six different eukaryotes (animals, *S. cerevisiae* and *A. thaliana*) revealed conserved asymmetries between writer kinases and eraser phosphatases that could be extended to other PTM pathways such as ubiquitination (Smoly *et al.* 2017). Specifically, the authors argue that kinase writers are encoded by multiple genes with low translation levels, and that they act on a highly specific hierarchy of substrates. In contrast, phosphatases function in broader and unspecific contexts, and are encoded by few, highly-translated genes that are less essential and less responsive to environmental perturbations. Finally, Smoly et al. (2017) report that these asymmetries between writer and eraser enzymes are common to ubiquitination and histone acetylation. Thus, they propose that reversible writer/eraser PTMs are subject to common constraints favoring functionally (and phylogenetically) diverse writers,

and multi-purpose erasers. The advantages to such a scheme are noise reduction (as proposed for many other cell signals, *cf.* Alberts et al. 2014) and faster responses to changing cues.

Thus, a comparative analysis of ubiquitination and other PTM systems offers an sketch of possible alternatives to phylogenetic inertia to explain the uneven expansions of different ubiquitin-related enzymes. However, a large-scale characterization of ubiquitination dynamics in diverse eukaryotes is required in order to understand its role in unicellular and multicellular contexts, using proteome-wide assays of ubiquitin interaction networks (Zhang *et al.* 2017). With this kind of data, we could determine whether a steep increase in ubiquitinome complexity actually occurred at the origin of metazoans, as proposed for phosphorylation (Sebé-Pedrós *et al.* 2016a)); or whether ubiquitination is hierarchically organized in the same way as phosphorylation (Smoly *et al.* 2017). Finally, it will also allow to examine the differences between ubiquitin and SUMO dynamics, which seems to rely on de-labeling enzymes' specificity in apparent contradiction with the hypotheses of (Smoly *et al.* 2017).

4.2.3. Myosin exaptation in animals

Myosin molecular motors also appeared and diversified early in eukaryotic evolution (section *R2 – Myosin*), thus fitting the biphasic model of burst-and-loss proposed by Wolf and Koonin (2013). For example, we recovered six classes of ancestral eukaryotic myosins (none of which existed in prokaryotes, which lack myosin motors altogether), which were frequently lost in early branching eukaryotic groups: three in Excavata (IV, V-like and VI), one in Diaphoratickes (II) and two more that are repeatedly lost in Viridiplantae, Alveolata and Rhizaria (Ia/b/c/h/d/g/k and If). Losses of myosin families were even more frequent in more recent ancestors (*e.g.*, there were 22 family losses in various holozoan lineages).

However, regarding the evolution of animal myosin tool-kits, our data can also be interpreted in the light of exaptation: out of the 19 gene families present in Metazoa, just one is an animal innovation (XVI/Dachs), the rest of them having a premetazoan origin. Most innovations in the myosin complement were a consequence of gene duplications and protein domain shuffling at the root of Holozoa (with 15 new families, eight of which are serial duplications of ancestral classes I, II and V). Therefore, these results highlight both the richness of the myosin complements of unicellular holozoans (as originally discussed in *R2 – Myosin*) and the fact that key myosins involved in multicellular functions already existed before the origin of animals.

A suggestive example of possible exaptation is given by the Myosin II subfamilies *smooth* and *striated*. In animals, myosin II paralogs are involved contraction of both muscle and non-muscle cells (Clark *et al.* 2007). Our analysis confirmed the report by Steinmetz et al. (2012) that the *smooth* and *striated* myosin families, paralogs of the ancestral Myosin II class, appeared at the last common ancestor of Holozoa: *smooth* is present in all holozoan lineages, and *striated* in animals and *M. vibrans.* This paralogy event was initially thought to correlate with the homonymous types of muscular tissue (Goodson and Spudich 1993). However, it now seems clear that both cell types appeared in Bilateria, after a gradual process of gene specialization via sequential duplications of various gene tool-kits involved in building contractile cells, which notably included the

smooth/striated myosin II paralogs, as well as later duplications of the myosin essential light chains or myocardin (Steinmetz *et al.* 2012; Brunet *et al.* 2016).

The model of cell type evolution outlined in Brunet et al. (2016) offers interesting insights into how the ancestral myosin II paralogs could have tuned its function in a multicellular context. First, even though there is no cell type homology between striated and smooth muscles of bilaterians, cnidarians and ctenophores, the expression patterns of *striated* and *smooth* myosins exhibit cell type-specific differences across these animal lineages and also sponges (Dayraud *et al.* 2012; Steinmetz *et al.* 2012). Therefore, we can extrapolate the hypothesis of Brunet et al. (2016) in the following manner: if the division between contractile and non-contractile cell types is a common feature across all animals and frequently involves myosin II sub-functionalization (Brunet *et al.* 2016), a similar pattern could be possible in even earlier-branching unicellular holozoans which have, as we now know, an identical genetic tool-kit. In a premetazoan unicellular context, however, such sub-functionalization would have been temporal. This possibility would be in line with the view of animal origins as a temporal-to-spatial switch in the regulation of cell type specification programs (Sebé-Pedrós *et al.* 2013a, 2016b, 2017), and coincide as well with the view of holozoan cell types evolving as lineage-specific innovations (de Mendoza *et al.* 2015).

This extension of the hypothesis of Brunet et al. (2016) is, of course, purely speculative. However, it can illustrate how gene content reconstructions in unicellular animal ancestors can lead to testable predictions regarding the co-option of ancient genes for essential multicellular functions.

4.2.4. Lysyl oxidases pre-date the extracellular matrix

We recovered a premetazoan origin for the family of lysyl oxidase enzyme family (LOX), as they are present in the genomes of ichthyosporeans, early-branching fungi and prokaryotes (section *R5 – LOX*). In animals, LOX functions primarily as an organizer of collagen-based ECM structures, but its earlier functions remain unknown (Csiszar 2001). The ancestral holozoan had a single-copy LOX gene (LOXO) which most likely already functioned in an extracellular context (according to its domain content). In animals, LOXO acquired its characteristic SRCR transmembrane domain and then underwent parallel paralogy processes in eumetazoans (2 paralogs) and poriferans (3). These early animal LOX could have possessed the ability to cross-link ECM components, as attested from their widely conserved protein domain architecture (including transmembrane SRCR domains and signal peptides for secretion) and its activity in present-day sponges (Eyre and Glimcher 1971; Van Ness *et al.* 1988). We thus interpreted this result as an example of exaptation: extracellular oxidases, already-available in early holozoans, were recruited into the ECM after the origin of multicellularity. As the earliest eumetazoan LOX (LOXL2/3/4) is present in extant taxa that lack fibrillar collagens (*e.g.* arthropods), we thus suggested a preferential role in the cross-linking of the non-fibrillar collagen IV-based basement membrane (as reported in mammals and *D. melanogaster*).

In the light of this study, it is worth noting that we recently reported the existence of canonical type IV collagen in the filasterean *M. vibrans* (section *R6 – Teretosporea genomes*). This amoeba lacks animal-like ECM collagen structures or a basement membrane despite having non-orthologous collagen genes, as other unicellular holozoans (King *et al.* 2008; Richter and King 2013).

Interestingly, it also lacks LOX homologs. This suggests that the uniquely metazoan association between type IV collagens and lyslyl oxidases could have occurred by recruiting two pre-existing protein products, rather than being 'triggered' by the emergence of the substrate (type IV collagen) in animals. Albeit a relatively trivial inference, this conclusion can serve as a cautionary tale against inferring ancestral functions from apparent concordances in the evolutionary histories of specific genes.

4.2.5. Rates of gene family diversification in premetazoan genomes

Our analyses of the evolution of myosin genes, HECTs and the whole ubiquitin/ubiquitin-like signaling systems revealed high levels of variability in the architectures of accessory protein domains within each gene family. This phenomenon, also known as domain shuffling or rearrangement, is a frequent contributor to gene family diversification processes across eukaryotes, particularly in metazoan genomes (Tordai *et al.* 2005; Ekman *et al.* 2007; Basu *et al.* 2008, 2009; Zmasek and Godzik 2012). Accessory protein domains can fine-tune the functions of a protein by adjusting its substrate specificities, sub-cellular localization or affinities with other proteins within larger complexes. For example, the ubiquitination enzymes (HECTs and other E3 ligases) often diversified by acquiring accessory domains that contribute to the enzyme's specificity by directing the enzyme to particular substrates or cellular localizationstie (protein-, sugar-, lipid- or nucleic acid-binding domains). In the seminal analysis of the *M. brevicollis* genome (King *et al.* 2008), it became apparent that many animal-specific protein families had appeared by shuffling processes predating the emergence of multicellularity – *i.e.* in its immediate unicellular holozoans ancestors. As described above, this is precisely the case in the evolution of myosins and the ubiquitin/ubiquitin-like signaling pathways.

In our comparative genomic analysis of holozoan genomes (section *R6 – Teretosporea genomes*), we aimed to characterize the burst of gene family diversification in premetazoans, detected in our previous analyses (sections *R1 – HECT*, *R2 – Myosin*, *R4 – Ubiquitin signaling*, *R5 – LOX*) and in many genetic tool-kits relevant for multicellularity (Tordai *et al.* 2005; Ekman *et al.* 2007; Gazave *et al.* 2009; Deshmukh *et al.* 2010; Hynes 2012; Suga *et al.* 2012, 2013, 2014, de Mendoza *et al.* 2013, 2014). Specifically, the underlying hypothesis of the study was that an unusually high rate of protein innovation in the unicellular prehistory of animals, if conserved in extant metazoans, would imply a prominent role of exaptation at the origin of multicellularity. To examine this hypothesis, we 1) maintained our previous assumption that the syntax of protein domains within a gene family can be used as a proxy for its diversity, and 2) set about developing a genome-wide computational approach to reconstruct the ancestral rates of gene family diversification. Specifically, we first defined orthology relationships at the whole-genome level using sequence similarity clustering of 40 eukaryotic proteomes (a scalable alternative to phylogenetic inference, *cf.* Kristensen et al. 2011; Emms and Kelly 2015). Second, we recorded the pairs of protein domains ('bigrams', *cf.* Xie et al. 2011) within each gene family. And third, the species profile of presence/absence was analyzed with the maximum likelihood phylogenetic birth-and-death model developed by Csűrös and Miklós (2009) and Csűrös (2010), with an explicit species tree as a guide for the reconstruction.

This approach allowed to estimate the diversification rates in specific ancestral nodes, both at the whole-genome and gene family-specific levels. This analysis confirmed our earlier findings (sections *R1 – HECT*, *R4 – Ubiquitin signaling*) by revealing a continuous state of higher-than-average protein domain diversification in the premetazoan prehistory for genes related to the ECM, transcription factors, ubiquitination and other signaling systems (Figure 6; most relative diversification ratios>1). Notably, this historical profile is consistent with previous targeted analysis of gene family evolution: it recovers a peak in ECM innovation at the LCA of Filozoa, which coincides with the identification of a complete integrin adheshome in *C. owczarzaki* (Sebé-Pedrós *et al.* 2010); and decreased levels of TF and ECM diversification at the LCA of Apoikozoa, concordantly with their frequent secondary losses in these specific molecular tool-kits (Sebé-Pedrós *et al.* 2011; Richter and King 2013; de Mendoza *et al.* 2013).

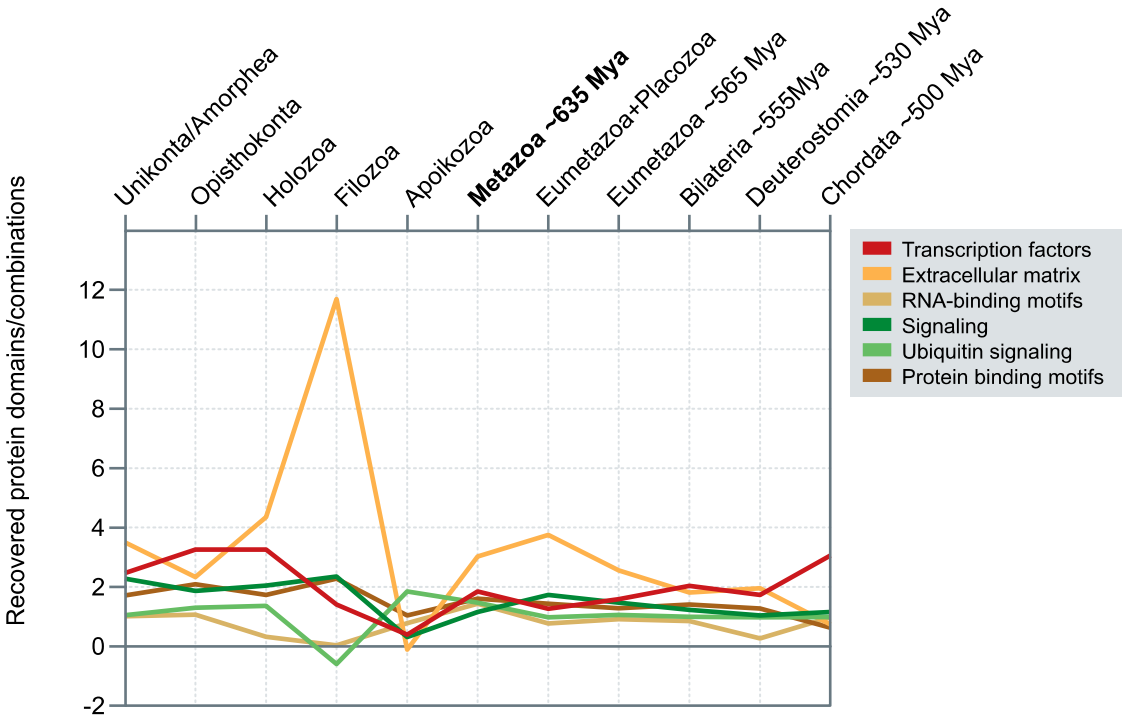


Figure 6. Ratio of protein domain gains-to-losses in different ancestral nodes, relative to gains in the the whole CA proteome. Metazoan LCA is highlighted in bold. Ratio>1: higher-than-average diversification; ratio<1: lower-than-average diversification. Data from *R6 – Teretosporea genomes*. Node ages for metazoan ancestors from Cunningham et al. (2016).

In turn, a closer examination of the orthologous groups that underwent diversification at the LCA of Holozoa (Figure 7) revealed a high frequency of PTM signaling-related functions, including protein kinases and phosphatases (Deshmukh *et al.* 2010; Suga *et al.* 2012, 2014), small GTPase and GPCR signaling (de Mendoza *et al.* 2014), ubiquitination (sections *R1 – HECT*, *R4 – Ubiquitin signaling*) and histone acetylation.

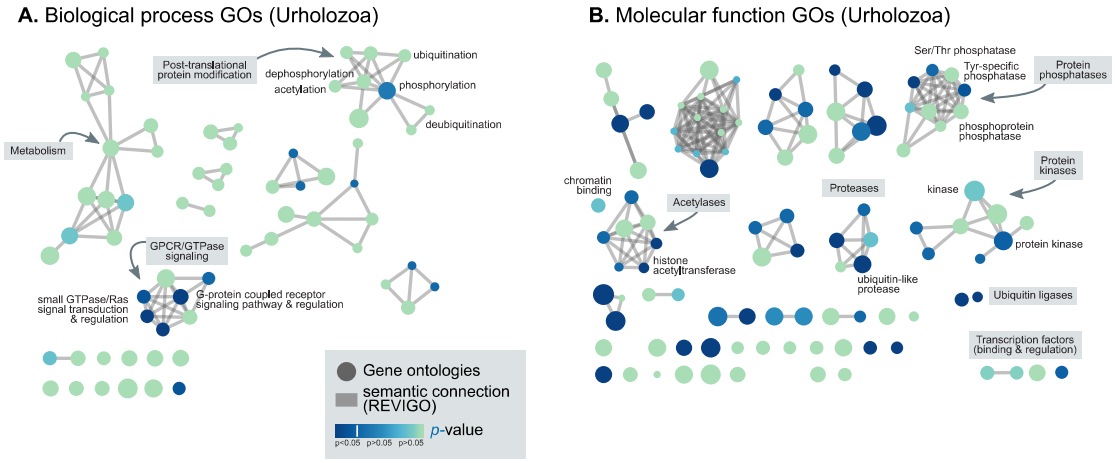


Figure 7. Networks of functional gene ontologies (GOs) of the clusters of orthologs that underwent protein domain diversification processes at the LCA of Holozoa, for **A)** biological processes and **B)** molecular functions. GO annotations obtained from a pooling of selected holozoans using EggNOG mapper with the eukaryotic database (Huerta-Cepas *et al.* 2016). Nodes represent GO terms, connected by edges according to the semantic network structure calculated in REVIGO (Supek *et al.* 2011). Nodes are size-coded according to relative weigth of each GO in the Uniprot database; and color-coded according to the *p*-value of a GO enrichment analysis performed with Ontologizer (Bauer *et al.* 2008), using a topology-weighted algorithm for related ontologies and pan-holozoan cluster of orthologs as a background. Protein domain diversification data from *R6 – Teretosporea genomes*.

We thus identify a progressive accretion of animal-like gene content in the unicellular ancestry of animals, comprehending genes with functions intimately related to the multicellular lifestyle – structural components of the extracellular matrix, regulators of transcription and chromatin states, intricate signaling pathways, etc. Overall, we identify 413 gene families undergoing protein diversification processes at the origin of Metazoa; 190 in the immediate unicellular holozoan ancestors (Apoikozoa, Filozoa and Holozoa LCAs); and 72 more genes families in the opisthokont LCA. In total, 614 clusters of orthologs diversified their domain architectures in the immediate unicellular ancestry of animals, 564 of which after the divergence of Fungi (data from *R6 – Teretosporea genomes*). These values, despite excluding the contribution of paralogy, are still higher than the estimations of 300-400 novel gene families at the root of Metazoa (Athanasopoulos *et al.* 2010; Domazet-Lošo and Tautz 2010; Simakov and Kawashima 2016). Therefore, from a purely quantitative point of view, the co-option of molecular 'pre-adaptations' appears to be a major source of proteomic innovation at the advent of animal multicellularity.

Hence, pinpointing the unicellular origin of specific molecular innovations is key to understand the evolution of the earliest animals. This holds true when examining the role of genes with conserved functions across the multicellular border, *e.g.* the partial redundancy of *C. owczarzaki*'s *Brachyury* transcription factor in *Xenopus* development (Sebé-Pedrós *et al.* 2013b) given the broad similarities in their downstream targets (Keller 2005; Sebé-Pedrós *et al.* 2016b). But this reasoning can also be applied to homologous genes or gene tool-kits that have different roles in animals and non-animal holozoans: *e.g.*, both integrins and cadherins function in multicellular adhesion and communication in animals, but these roles appear to be a consequence of the independent co-option of earlier extracellular sensing mechanisms (Abedin and King 2008; Sebé-Pedrós and Ruiz-Trillo 2010; de Mendoza *et al.* 2015). Thus, examining the circumstances of these parallel tool-kit co-options is essential to understand the role of cell adhesion in the first multicellular organisms.

Overall, these studies vindicate the utility of a protistan perspective of animals – a perspective that is all the most revealing when applied to the landscape of genome evolution.

4.3. Sampling new unicellular holozoan genomes

Currently, gene content analyses of unicellular holozoans can make use of the sequencing of ten genomes (2 choanoflagellates, King et al. 2008 and Fairclough et al. 2013; *C. owczarzaki*, Suga et al. 2013; and 7 teretosporean genomes presented in section *R6 – Teretosporea genomes*), and three transcriptomes (the filasterean *M. vibrans* and the ichthyosporeans *Amoebidium parasiticum* and *Sphaerothecum destruens*, Torruella et al. 2015). This trove of genomic data has allowed for an unprecedented level of detail in the reconstructions of the ancestral gene content of the metazoan LCA: an expanded taxon sampling gives more realistic estimates of gene age, and helps to assess the true frequency of lineage-specific gene losses that pervades most eukaryotic genomes (Lynch 2006b; Wolf and Koonin 2013).

As our view of premetazoan genome evolution has changed over time, it is worth asking which aspects of comparative genomics can benefit the most from the continued effort of sequencing new unicellular holozoan genomes. In order to provide some insights into this question, I will draw a metaphor from ecology: unicellular holozoan genomes are 'sampling sites' which can be probed in the search for homologs of animal genes, *i.e.* 'species' present in these 'sites'. We can explore the 'genomes-as-sites' metaphor with a specific example: assessing the richness of animal-like genes in unicellular holozoans, using accumulation curves and protein domain annotations from the available genome and transcriptome sequences, aiming to predict the total 'protein domain richness' of unicellular holozoans (Chiarucci *et al.* 2008).

The accumulation curve of animal gene discovery in unicellular holozoans (Figure 8) has a decreasing slope as more genomes are sampled: the number of typical animal gene families present in unicellular holozoa (real richness = 11,349) is not far from the maximum theoretical value as inferred by the Chao estimate (12,531.6±60.7; Chao 1987). A similar trend emerges for the analysis of animal-like protein domains (real richness = 5199-5663, depending on dataset; Chao estimate = 5560.8±37.9 or 6199.1±53.1). In contrast, when this analysis is applied to combinations of protein domains within specific gene families (from section *R6 – Teretosporea genomes*), we are still far from having probed all the genic diversity of unicellular holozoans (actual value = 10,540; Chao estimate = 20,924.5 ± 394.2).

The inaccuracy of the 'genomes-as-sites' metaphor[2] only allows for a limited interpretation of these results. First, as we sample more unicellular holozoans, the identification of new animal-like protein domain combinations appears to be more likely than spotting individual domains or gene

2. The 'genomes-as-sites' metaphor is not without problems: one has to assume that each unicellular holozoan genome is a completely independent sampling point (actually, they are semi-independent due to their shared ancestry); and divergence times between the metazoan LCA and the sampled unicellular holozoans vary depending on the lineage (shorter for choanoflagellates, longer for teretosporeans).

families. This result is in line with previous reports that highlighted that a greater deal of protein diversity can be discovered by analyzing changes in the domain arrangements than individual domains themselves (Levitt 2009; Moore *et al.* 2013). However, the Chao richness estimates refer to the expected number of animal-like traits (genes, domains or domain combinations) to be found if all unicellular holozoan genomes were sampled, a figure which will always be lower than the real 'richness' of the metazoan LCA (which is entitled to its share of exclusive innovations).

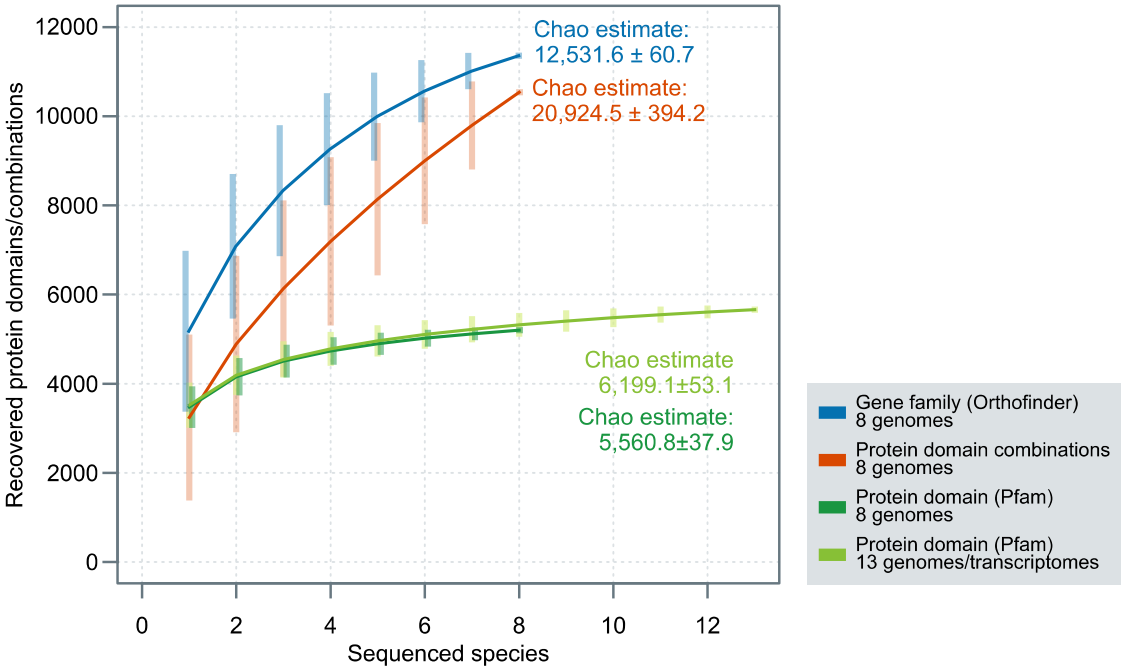


Figure 8. Accumulation curve of animal gene families (blue), protein domains (green) and protein domain combinations (orange) in unicellular Holozoa. The "8 genomes" dataset includes the complete genomes of *S. rosetta*, *M. brevicollis*, *C. owczarzaki*, *C. fragrantissima*, *S. arctica*, *I. hoferi*, *C. perkinsii* and *C. limacisporum*. The "13 genomes+transcriptomes" dataset expands the previous one with transcriptomes of *M. vibrans*, *S. destruens*, *P. gemmata*, *A. whisleri* and *A. parasiticum*. Accumulation curves and Chao estimates calculated using R vegan library (Oksanen *et al.* 2017).

Therefore, as new unicellular holozoan genomes become available, a reinforced interest in the detailed evolution of multi-gene families will be in order, so as to expand our knowledge of the premetazoan tool-kit for multicellularity by explicitly taking heed of the mosaic-like nature of multi-domain gene families (as proposed in Kristensen et al. 2011 and Haggerty et al. 2013). Such strategies of orthology inference need to account for two crucial aspects: a pluralistic view of protein evolution in which each domain can have its own homology relationships, and homoplasy of protein domain architectures (*i.e.* convergent evolution). Both homoplasy and mosaicism feature frequently in the results of my research (sections *R1 – HECT*, *R2 – Myosin*, *R4 – Ubiquitin signaling*, *R5 – LOX*, *R6 – Teretosporea genomes*), as I have examined in the previous section. The methodological framework of gene family evolution that I developed in section *R6 – Teretosporea genomes* is explicitly designed to analyze such complex homology relationships at the whole-genome level – without significantly increasing the computational costs of the inherent orthology inference step (Altenhoff *et al.* 2016).

Finally, the study of new genomes and transcriptomes of Holozoa remains essential to improve phylogenomic analyses and consolidate our views of eukaryotic tree of life. The robust phylogenetic framework presented in section *R3 – Opisthokonta phylogenomics* (Torruella *et al.* 2015), which rejected the 'Filasporea' hypothesis (*i.e.*, monophyly of Filasterea and Ichthyosporea), benefited from a broad taxon sampling that included all opisthokont protists (including unicellular holozoans and holomycetes such as nucleariids and *Fonticula alba*) and a deep sampling of choanoflagellates (8 species) and ichthyosporeans (7 species). Similarly, in *R6 – Teretosporea genomes*, I was able to improve the supports of both Ichthyosporea and Teretosporea monophyly due to the inclusion of an additional short-branch dermocystid Ichthyosporea, *C. perkinsii*. Thus, the taxon sampling richness appears to be of paramount importance in phylogenomic analyses, as recently highlighted in a comprehensive analysis of early animal diversification where the exclusion of slow-evolving poriferans was demonstrated to affect the branching order of ctenophores and poriferans themselves (Simion *et al.* 2017). As more holozoan genomes and transcriptomes become available, it will become necessary to 1) expand the current sets of phylogenomic markers and 2) update the phylogenetic framework with newer taxa and methods. This continued effort will have to take heed of the recent report of putative new filastereans and a relative of *C. limacisporum* (Tikhonenkov et al. 2016, and personal communication by Elisabeth Hehenberger).

4.4. Genomic architecture in the animal prehistory

Ever since the influential report of Lynch and Conery (2003), virtually every discussion regarding the evolution of genome architecture has contemplated their population-genetic framework – to put it concisely, complex genomes are a consequence of low purifying selection regimes imposed by low effective population sizes, which permit the accumulation of slightly deleterious genetic embellishments such as large C-values, frequent and long introns, or vast swathes of repetitive content. The debates spawning from this hypothesis, including critical (Cavalier-Smith 2010; Whitney and Garland 2010; Whitney *et al.* 2011; Roy 2016) and positive regards (Babenko *et al.* 2004; Koonin 2009, 2016; Koonin *et al.* 2013; Lartillot 2015) *inter alia* (Maeso *et al.* 2012; Lobkovsky *et al.* 2013; Elliott and Gregory 2015a), have since revolved around the relative influence of adaptation and neutrality in sculpting the evolution of genome organization.

In section *R6 – Teretosporea genomes* we shed light onto the evolution of genome architecture during the unicellular prehistory of animals. Despite the important disparities in extant genome sizes across holozoans (~24-120 Mb in unicellular lineages, ~20 Mb to >120 Gb in metazoans; Elliott and Gregory 2015a), we linked the larger unicellular holozoan genomes to secondary expansions, thus implying a lower C-value at the LCA of Holozoa (~25-35 Mb if it was to be in line with the relatively compact genomes of *C. owczarzaki*, *M. brevicollis*, *C. limacisporum* or *C. perkinsii*). We also reconstructed a relatively high intron density (5.5 introns/CDS kbp), a feature linked to low efficiencies in purifying selection (Csűrös *et al.* 2011) under the population-genetic framework of Lynch and Conery (2003). From these ancestral values, extant unicellular holozoans evolved

remarkably disparate genome architectures: *C. perkinsii* and *C. limacisporum* underwent massive intron loss (0.7 and 0.03 introns/CDS kbp), while choanoflagellates and *I. hoferi* followed an opposite, increasing trend (6-7.1 introns/CDS kbp) in parallel with animals (animal LCA: 8.7 introns/CDS kbp). Some ichthyosporeans seem to have punctually evolved larger genomes (88-120 Mb), as they are close relatives of species with more average genome sizes (*C. fragrantissima*, at 42.9 Mb); and larger genomes bear traces of recent and massive transposable element (TE) invasions. In addition, unicellular holozoans have starkly divergent lifestyles, including free-living saprotrophs (*C. limacisporum* and possibly *C. perkinsii*) and phagotrophs (choanoflagellates, *M. vibrans*) and a collection of organisms with symbiotic, parasitic, commensal or epibiontic relationships with animals (all the sequenced ichthyophonids and *C. owczarzaki*).

Given this variable landscape of genome architectures and life histories, I will now analyze the evolution of complex holozoan genomes under the light of the mutational-hazard hypothesis of Lynch and Conery (2003). As a proxy to genomic complexity (as per the definition of Wolf and Koonin 2013), I will mainly rely on reconstructions of intron density in holozoan ancestors, as introns are an analytically tractable 'genomic embellishment' (Carmel *et al.* 2007a; Csűrös *et al.* 2011) central to the population-genetic framework of proposed by Lynch (2002).

In particular, I envisage two different scenarios. First, whether complex genomic traits conserved between unicellular holozoan genomes and animals can be linked by a shared history of low purifying selection, *i.e.* a common population-genetic environment in their shared history. The results from section *R6 – Teretosporea genomes* do not to support this scenario: the ancestral reconstruction of intron density revealed two independent processes of intron gain in the ancestors of animals and ichthyophonid ichthyosporeans (8.7 and 6.9 introns/CDS kbp, respectively), from more moderate values in the ancestral holozoan (5.5 introns/CDS kbp). Therefore, it is implausible to assume that the effective populations had been permanently low between the LCAs of Holozoa and Metazoa (as implied by (Csűrös *et al.* 2011); animal range: 10^5-10^6). This result tentatively rejects the conjecture of a progressive accretion of architecturally complex genomes during the deepest unicellular prehistory of animals (*i.e.*, before the divergence of the extant unicellular lineages).

The second scenario contemplates the event of parallel but independent population-genetic pressures in unicellular and multicellular holozoans. In order to test this possibility, we can analyze whether the varying genome sizes and architectures of Teretosporea (the best-sampled unicellular holozoan clade) can be linked to differences in effective population in the same way this occurs in animals (Lynch and Conery 2003; Albertin *et al.* 2015; Simakov and Kawashima 2016). Specifically, according to the predictions of Lynch (2002, 2003, 2011), we can test whether ichthyosporean genomic complexities can be linked to ineffective purifying selection caused by small effective populations – for example, the large, intron-dense, repetitive and/or TE-rich genomes of *S. arctica* or *A. whisleri*.

In principle, it is possible to estimate the effective population of unicellular eukaryotes via its direct proportionality with the nucleotide-level heterozygosity (under certain assumptions[3]).

3. For low heterozygosity (H<1) values, $H=4N_e\mu$ in diploid genomes, or $H=2N_e\mu$ in haploid ones; N_e is effective population size and μ is the mutation rate (Lynch *et al.* 2011).

However, teretosporeans' heterozygosity levels can only be measured in cultured cell lines (prone to population-level artifacts such as culture-driven selection or bottlenecks), a problem further aggravated by the scarcity of teretosporeans in environmental surveys (Logares et al. 2014; de Vargas et al. 2015; del Campo et al. 2015; personal communication by David López-Escardó). Furthermore, a precise estimation of the effective population requires measuring the mutation rate, which is also unavailable for natural populations.

The scarce available data on cultured teretosporeans' heterozygosity levels (Table 2) has yielded a wide range of values (~0.05% to ~1.92% in *C. fragrantissima* and *A. whisleri* respectively). Thus, assuming that 1) measured heterozygosities homogeneously correlate with the natural population values and 2) the mutation rate is constant between species, these observations do not fulfill the predictions by Lynch and Conery (2003): higher heterozygosities (and effective populations) would be recovered in the largest, TE-richest, and intron-densest genomes (*P. gemmata*, *S. arctica*, *A. whisleri*). However, the above-mentioned pair of assumptions are weak and not supported by direct measurements of genetic diversity.

Table 2. Heterozygosity estimates in teretosporean genomes. Assembled genome sizes (from section *R6 – Teretosporea genomes*) can be compared to the haploid genome sizes as estimated in GenomeScope (Vurture *et al.* 2017), using a k-mer length of 21 bp, which also allows to estimate a range of per-site heterozygosity values in diploid genomes. Ploidy levels are unknown in Teretosporea, but a coincidence between assembly length and haploid genome size estimates can be an indirect hint of diploidy.

Sample/Isolate	Assembled genome size (Mb)	Haploid genome size (Mb)	Heterozygosity estimate	Model fit
C. fragrantissima (2011)	42.9	45.1-45.2	0.05-0.06 %	98.7-99.3 %
C. fragrantissima (2015)	45.8	44.4-44.4	0.06-0.07 %	97.5-97.8 %
S. arctica JP6010 (2011)	120.9	87.7-88.1	1.14-1.18 %	96.8-99.3 %
Sphaeroforma sirkka B5 (2016)	83.3	97.7-97.8	0.14-0.15 %	97.4-99.4 %
A. whisleri (2014)	101.9	100.2-101.2	1.87-1.92 %	94.4-97.2 %
P. gemmata (2015)	84.4	59.0-61-2	2.66-2.89 %	93.6-98-8 %
C. perkinsii (2015)	34.6	31.8-31.8	0.33-0.34 %	95.7-97.6 %
C. limacisporum Hawaii (2013)	24.1	23.1-23.2	0.21-0.22 %	98.6-99-4 %

In contrast, indirect indications of the population-genetic effect on genome architecture predicted by Lynch and Conery (2003) can be drawn from comparing free-living (with presumably larger populations) and symbiotic Teretosporea species. Under the light of this (equally speculative) assumption, the cosmopolitan and marine *C. limacisporum* has undergone an important process of genome streamlining: it harbours a relatively compact and TE-depleted genome (only maintaining a small set of highly similar, active TEs) in comparison with the larger repetitive regions of *P. gemmata*, *A. whisleri* or *S. arctica*; it has undergone intense intron loss and shortening compared to the LCA of Teretosporea (from 5.5 to 0.03 introns/CDS kbp; and shortest median intron length among unicellular holozoans); and it has an unusually high rate of microsynteny conservation (see below), as predicted in Lynch (2006a).

Regarding the ordering of genes within the genome, we did not identify any conservation of macrosyntenic arrangements between unicellular and multicellular holozoans, in line with previous results (Putnam *et al.* 2007; Suga *et al.* 2013; Simakov *et al.* 2013; Simakov and Kawashima 2016), thus implying that the conservation of chromosome-level gene linkages over long evolutionary timescales remains an exclusive animal feature (Simakov and Kawashima 2016). We did find, however, a small set of ~130 genes with conserved linkage between the filasterean *C. owczarzaki* and the early-branching teretosporeans *C. perkinsii* and *C. limacisporum*. To my knowledge, this represent a rare case of long-range microsynteny conservation across unicellular eukaryotic lineages—in the absence of persistent functional constraints, gene order conservation is expected to neutrally decay (Hurst *et al.* 2004; Koonin 2009)—and attests the slower pace of genome evolution of the filasterean *C. owczarzaki*. Its interpretation under the population-genetic conjecture of Lynch and Conery (2003) is, however, unclear: low effective populations have been linked to synteny disruption in animals (Albertin *et al.* 2015; Simakov and Kawashima 2016), and, more generally, to the loss of operon-like structures in eukaryotes as a passive consequence of a reduced efficiency in gene order maintenance (Lynch 2006b). Under the assumption of Lynch and Conery (2003), the evolutionary pathways linking *C. owczarzaki* with either *C. limacisporum* and *C. perkinsii* should have been dominated by relatively high effective populations – this can be argued in the case for the streamlined genomes of *C. perkinsii* and *C. limacisporum*, but the slight reduction in intron content of *C. owczarzaki* offers less clear evidence.

The specific processes behind the expansion in genome size and complexity at the origin of Metazoa (Simakov and Kawashima 2016) remain ultimately unknown. The bioenergetic argument —namely, that larger genomes are only sustainable in mitochondriate organisms with efficient energy outputs (Lane and Martin 2010; Lane 2011)—can explain the prokaryote-to-eukaryote transition and the higher frequency of larger genomes among the latter, but does not account for the mechanisms or advantages of such a phenomenon in multicellular organisms (although see Booth and Doolittle 2015; Lynch and Marinov 2017). Furthermore, the neutralist framework discussed above (Lynch and Conery 2003) has not yet been proven to be relevant in unicellular holozoans, and does not reject the possibility of a later, adaptive role for passively accumulating genomic embellishments. For example, high intron insertion rates can facilitate the emergence of exon skipping-based alternative splicing profiles (Lynch and Conery 2003; Lynch 2006a), which are known to expand the array of available transcript and protein products in Metazoa (Irimia and Blencowe 2012; Barbosa-Morais *et al.* 2012; Chen *et al.* 2014a) and lead to key evolutionary innovations (Gracheva *et al.* 2011; Gueroussov 2015; Bush *et al.* 2017). Introns can also harbour regulatory sites for proximal genes, which contributes to the maintenance of syntenic blocks (Le Hir *et al.* 2003; Irimia *et al.* 2012), and can foster gene family diversification by exon shuffling (Liu *et al.* 2005). In turn, mobilization of transposable elements has also been linked to the intron creation and lengthening processes that dominate metazoan ancient and recent evolution (Li *et al.* 2009; Csűrös *et al.* 2011), which enables exon skipping-based alternative splicing. TEs are also involved in complex adaptations such as conversions into coding genes (Hua-Van *et al.* 2011) and regulatory sequences (as 25% of the regulatory sites of some mammals; Jordan et al. 2003), as well as in the origin of the V(D)J recombination process used in immunoglobulin synthesis in vertebrates (Jones and Gellert 2004). Furthermore, the larger animal genomes can house non-coding elements that are key to maintain cell type-specific transcriptional profiles in multicellular organisms – namely,

the regulatory distal and developmental enhancers (Sebé-Pedrós *et al.* 2016b; Gaiti *et al.* 2016, 2017), richer complements of long non-coding RNA genes (Gaiti *et al.* 2015; de Mendoza *et al.* 2015; Sebé-Pedrós *et al.* 2016b), or topologically associated domains of transcriptional regulation (Lee and Iyer 2012; Seitan *et al.* 2013; Gaiti *et al.* 2016).

Overall, the expansion in genome size and complexity at the origin of Metazoa cannot be explained only by higher intron contents or gene family expansions alone (Elliott and Gregory 2015a). Similarly, the contribution of TE proliferation processes, while evidently essential to explain extant genome sizes (Elliott and Gregory 2015a; b), has only been successfully reconstructed at much shorter time-scales (Sessegolo *et al.* 2016; Hjelmen and Johnston 2017). Thus, in order to explain the complexification of genomic architecture that accompanied the transition to multicellularity, it will be essential to characterized the functions and types of elements that comprise the non-coding genome of unicellular holozoans.

4.5. Ancestral functions from ancestral architecture: evolution of alternative splicing

Alternative splicing (AS) is a mechanism of transcriptome regulation that has been described in a wide range of eukaryotes and is based on the differential selection of splice sites to produce multiple transcripts from a single gene (Gilbert 1978; Breitbart *et al.* 1987; Kim *et al.* 2007). It has been described as a powerful source of evolutionary innovations, *e.g.* enabling a diversification of the proteome or adding an additional layer of gene expression control (Graveley 2001; Nilsen and Graveley 2010; Gracheva *et al.* 2011; Gueroussov 2015). The early realization that different modes of AS—*e.g.* alternative exclusion of exons (exon skipping or ES) and intron retention (IR)—co-exist in most eukaryotes was followed by the discovery of lineage-specific biases in the frequency of each AS mode (Breitbart *et al.* 1987; Kim *et al.* 2007; McGuire *et al.* 2008). Some of these differences could be associated to specific gene structural features: higher splice site heterogeneity (Ast 2004; Roy and Irimia 2009) and intron density favor splicing variability (Koonin *et al.* 2013); and the relative length of exons and introns affects the definition of the excised portions of the pre-mRNA (De Conti *et al.* 2013). These and other *cis* signals have been found to influence the patterns of IR and ES across mammals and vertebrates (Barbosa-Morais *et al.* 2012; Braunschweig *et al.* 2014).

We expanded this framework (section *R7 – Alternative splicing*) to the whole eukaryotic kingdom and defined a set of gene structural features that concordantly influence the frequency and mode of AS in widely diverging lineages – 60 eukaryotes including land plants, multiple protists, animals and fungi. This 'soft AS code' can explain the observed differences in the frequency ES and IR across eukaryotic genomes (Kim *et al.* 2007; McGuire *et al.* 2008), as ES- or IR-conducive gene architectures can independently evolve multiple times. This result allows us to indirectly infer the dominant mode of alternative splicing in ancestral eukaryotic nodes, based on the reconstruction of the relevant genomic traits: intron densities, splice site heterogeneity, nucleotide composition

(GC content) and estimations of ancestral intron size (derived by correlation, for example, from genome size estimates; Elliott and Gregory 2015a).

According to our previous results from section *R6 – Teretosporea genomes*, the LCA of Holozoa is likely to have harboured a relatively compact genome (~30 Mb) that later underwent secondary expansions in some ichthyosporeans, choanoflagellates and the root of Metazoa. In parallel, we also inferred a moderate intron density (5.5 introns/CDS kbp), closer to the estimates for the ancestral eukaryote (4.9 or 4.3 introns/CDS kbp according to our analysis or Csűrös et al. 2011, respectively) than to the intron-richer ancestral metazoans (8.7 or 8.8 introns/CDS kbp depending on the analysis). Using this genome size (~30 Mb) and intron density (5.5 introns/CDS kbp), the average fraction of genic sequence in the highest-quality unicellular holozoan genomes (~66%), the relatively constant median CDS length (~1.3 kbp) and an estimation of ~10,000 genes in the LCA of Holozoa[4] , we can infer a typical intron length of ~93 bp in ancestral holozoans (ranging 48-148 bp under different assumptions; Table 3). The typical values are in line with the genomes of *C. owczarzaki*, *C. fragrantissima* or *S. rosetta*, which are relatively intron-dense (3.5-7.1 introns/CDS kbp) and produce IR-dominated transcriptomes (Sebé-Pedrós *et al.* 2013a; de Mendoza *et al.* 2015).

Table 3. Estimations of typical intron length at the LCA of Holozoa and Metazoa, under different scenarios. See text for details on the source of ancestral reconstruction and indirect estimations of each value. Ranges of Holozoa LCA genome size from *R6 – Teretosporea genomes*. Ranges of Holozoa LCA genic fraction from observation of extant unicellular holozoans' values. Ranges of Metazoa LCA genic fraction: higher bound from animals' average gene span in Elliott and Gregory (2015a); lower bound from animals' median intergenic span, from *R6 – Teretosporea genomes*.

LCA scenarios	Genome size	Gene number	Genic fraction	CDS length	Gene length	Intron density	Intron length
Holozoa (intermediate genome)	~ 30 Mb	~ 10,000	0.66	~ 1320 bp	~ 1999 bp	5.5 int/CDS kbp	~ 93.4 bp
Holozoa (large genome)	~ 35 Mb	~ 10,000	0.66	~ 1320 bp	~ 2333 bp	5.5 int/CDS kbp	~ 139.3 bp
Holozoa (small genome)	~ 25 Mb	~ 10,000	0.66	~ 1320 bp	~ 1666 bp	5.5 int/CDS kbp	~ 47.5 bp
Holozoa (small genic fraction)	~ 30 Mb	~ 10,000	0.50	~ 1320 bp	~ 1500 bp	5.5 int/CDS kbp	~ 24.7 bp
Holozoa (large genic fraction)	~ 30 Mb	~ 10,000	0.80	~ 1320 bp	~ 2400 bp	5.5 int/CDS kbp	~ 148.6 bp
Metazoa (small genic fraction)	~ 300 Mb	~ 20,000	0.38	~ 1489 bp	~ 5686 bp	8.7 int/CDS kbp	~ 380.0 bp
Metazoa (inter. genic fraction)	~ 300 Mb	~ 20,000	0.49	~ 1489 bp	~ 7609 bp	8.7 int/CDS kbp	~ 472.5 bp
Metazoa (large genic fraction)	~ 300 Mb	~ 20,000	0.64	~ 1489 bp	~ 9533 bp	8.7 int/CDS kbp	~ 620.9 bp

4. Estimation of the number of genes in the Holozoa LCA: the clusters of orthologous groups produced in section *R6 – Teretosporea genomes* were and analyzed using the birth-and-death model of Count, accounting for gains, losses and duplication; we then inferred the number of gene families present in that ancestor and the fraction that contained multiple paralogs (Csűrös 2010). The probabilistic model was trained using 2,000 randomly selected gene families. See Methods in section *R6 – Teretosporea genomes* for further details.

Applying the same logic to the LCA of Metazoa yields very different results (Table 3). First, we can infer that it had a higher intron density (8.7 introns/CDS kbp), a larger genome (~300 Mb; Simakov and Kawashima 2016) harbouring more genes (~20,000; Simakov and Kawashima 2016) that encoded for slightly longer CDS than unicellular holozoans (~1.5 kbp; Elliott and Gregory 2015a). Assuming different values for the genome genic fraction, we obtain typical intron sizes oscillating between 380-620 bp. Therefore, it seems safe to assume that the ancestral metazoan had longer introns than previous holozoans (~4-6-fold increase) – a feature which, coupled with higher intron densities and splice site heterogeneity, is a clear predictor of ES-richer transcriptomes. We can thus infer that the LCA of Metazoa probably exhibited a more diverse AS profile than its immediate ancestors: IR remained common and ES frequency increased.

Another finding from our pan-eukaryotic survey of AS is that, even if conducive genomes can promote splicing variability in the form of ES, these events do not necessarily lead to functional protein isoforms like those abundantly characterized in metazoans (Graveley 2001; Nilsen and Graveley 2010; Barbosa-Morais *et al.* 2012; Kelemen *et al.* 2013). In many bilaterian animals, ES events preferentially involve exons whose lengths are 3-divisible, and hence not prone to produce shifts in the reading frame when excluded from the transcript. This was conspicuously not the case in other ES-positive eukaryotes, such as land plants or *Volvox carteri* (which actually exhibited an opposite bias), non-bilaterian animals, *S. arctica*, *C. fragrantissima* or *Bigelowiella natans*. In similar cases, ES events have been attributed to spliceosomal noise akin to that underlying many events of IR (Curtis *et al.* 2012; de Mendoza *et al.* 2015). Thus, even if the LCA of Metazoa appears to have had an ES-conducive genome architecture, it is possible that it was dominated by noisy splicing as well. Interestingly, there are indications that isoform-permitting ES is not necessarily associated to high transcriptome-wide levels of ES: the IR-dominated *C. owczarzaki* appears to maintain a small network of dynamically regulated alternatively spliced exons (30 in total, 60% of which with 3-divisible lengths) that span known protein motifs of the final peptides (~66%) and are enriched in kinase signaling functions (Sebé-Pedrós *et al.* 2013a). This result hints at the possibility that, if functional AS is derived from the co-option of basal-rate splicing variability (Koonin *et al.* 2013), this phenomenon can still occur with testimonial levels of ES.

These results demonstrate that comparative genomics and ancestral reconstructions constitute a powerful tool for evolutionary analysis of ancestral eukaryotes: not only it allows to uncover the primary composition of ancestral genomes; it can also fuel inferences regarding their transcriptomic regulation and the role played by non-genomic sources of evolutionary innovation.

5. Conclusions

From protists to the prehistory of animals

Adventavit asinus,
pulcher et fortissimus
[The ass arrives,
beautiful and most brave]

Friedrich Nietzsche, Beyond Good and Evil:
Prelude to a Philosophy of the Future, 1866

The main conclusions of the present work are the following:

1. The evolution of HECT ubiquitin E3 ligases is marked by an early burst of gene diversification at coinciding with the emergence of eukaryotes, followed by differential lineage-specific expansions and contractions in later-branching lineages. The Holozoa, both multicellular and unicellular, bear the most diversified repertoire of HECT enzymes as measured by the syntax of co-occurring accessory protein domains.

2. The expanded analysis of the evolution of the ubiquitination pathway reveals that this signaling system already existed prior to the origin of eukaryotes, as the complete set of enzymes (including ubiquitin peptides, E1 activators, E2 conjugases, E3 ligases as well as label-removing enzymes) is found in 21 archaeal genomes from three different lineages (Crenarchaeota, Euryarchaeota and Aigarchaeota). However, most of the extant repertoire of ubiquitin-related enzymes originated in the lapse of time between the first eukaryotic ancestor (FECA) and the LECA, thus implying a burst of gene innovation during the eukaryogenesis process later followed (as in the case of HECTs) by differential gene loss and lineage-specific expansions.

3. The highest levels of protein diversity in the ubiquitination tool-kit are found in complex multicellular lineages (animals and plants) and their colonial/unicellular relatives (other holozoans, chlorophytes, glaucophytes) – and were reached by independent processes of paralogy/domain shuffling.

4. Myosin molecular motors are an eukaryotic innovation that underwent an early diversification process prior to the LECA, plus subsequent lineage-specific expansions peaking in the Holozoa. The core myosin complement of animals appeared well before the origin of multicellularity: out of 20 myosins present in the LCA of Metazoa, only one (XVI/Dachs) is an animal innovation; and ancestral holozoans already possessed paralogs of other animal myosin families (*e.g.*, the II-*striated* and II-*smooth* muscular myosins).

5. A phylogenomic investigation of holozoans revealed that the enigmatic *Corallochytrium limacisporum* is the sister-group to the Ichthyosporea, conforming a new clade tentatively named Teretosporea. The inclusion of the newly discovered *Chromosphaera perkinsii*, an early-branching dermocystid, improves the statistical support for Teretosporea and confirms the monophyly of Ichthyosporea.

6. A whole-genome phylogenetic comparative analysis revealed a continued process of gene family diversification in the unicellular ancestry of Metazoa, which was particularly intense for the complement of transcription factors, extracellular matrix genes and signaling genes (such as the ubiquitination repertoire). This allows us to quantify the strength of pre-metazoan innovation (614 gene families in LCAs from Opisthokonta to Apoikozoa) relative to animal-specific gene invention (300-400 genes).

7. The evolution of genome architecture in Holozoa was a dynamic process during which multiple traits independently changed in both unicellular and multicellular lineages. Since the LCA of Holozoa, a process of synteny disruption split up ancestral gene linkages

in most extant lineages (Ichthyosporea, choanoflagellates and animals). This event was subsequently followed by the establishment of new conserved syntenic arrangements in more recent ancestors within each individual lineage – chiefly in Metazoa.

8. The intron-dense genomes of animals and ichthyophonid Ichthyosporea appeared as a product of independent intron gain processes occurred after their early divergence within Holozoa. In contrast, whereas *C. limacisporum* and *C. perkinsii* were dominated by streamlining processes of intron loss.

9. *Capsaspora owczarzaki* harbours a slow-evolving genome in terms of synteny conservation, intron content and changes in the coding sequence – compared to the frequent lineage-specific genome architecture reconfigurations of choanoflagellates and teretosporeans.

10. The patterns of alternative splicing in extant eukaryotes are influenced by a soft code of gene architectural features involving intron density, relative length of exons and introns, splice site heterogeneity, or the gene expression levels, among others. These patterns differ between IR and ES, which allows us to identify the specific features that enable ES-rich transcriptomes (short exons in intron-dense genes, low expression, etc.). Thus, ES-rich AS profiles can evolve repeatedly in architecturally conducive genomes. This is notably the case of Metazoa (particularly Bilateria) and the intron-dense ichthyosporean *S. arctica*.

11. We can infer the relative strength of each AS mode in ancestral genomes by reconstructing their gene architecture. Under reasonable assumptions of typical gene features, we thus infer that the LCA of Metazoa had an AS-rich transcriptome dominated by IR—as it is still the case in early-branching extant animals—but with a higher frequency of ES events than previous holozoans.

12. The reconstruction of ancestral holozoan genomes by comparative methods is a powerful instrument to study the origin of animal multicellularity. The inference of gene contents in foregone premetazoan genomes can lead to testable predictions regarding the evolutionary paths mechanisms behind the origin of animal-specific features. In parallel, the analysis of premetazoan genome architecture reveals a dynamic process of reconfiguration affecting gene order, structure and composition – a thorough exploration of the genomic space that left no trait left to tinker with.

6. References

Publications and communications [1]

[1] *Ignore this* [2]
[2] *Increment by 2 before following*
[3] *Not true* [3] [4]
[4] *Ibid.*
[5] *True* [2] [6]
[6] *Actually a 1* [2] [2]

Randall Munroe, xkcd – Footnote labyrinths, 2013

Abedin M., King N., 2008 The premetazoan ancestry of cadherins. Science 319: 946–8.

Acemel R. D., Tena J. J., Irastorza-Azcarate I., Marlétaz F., Gómez-Marín C., *et al.*, 2016 A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. Nat. Genet. 48: 336–341.

Adl S. M., Simpson A. G. B., Lane C. E., Lukeš J., Bass D., *et al.*, 2012 The revised classification of eukaryotes. J. Cell Biol. 59: 429–93.

Albani A. El, Bengtson S., Canfield D. E., Bekker A., Macchiarelli R., *et al.*, 2010 Large colonial organisms with coordinated growth in oxygenated environments 2.1 Gyr ago. Nature 466: 100–104.

Albertin C. B., Simakov O., Mitros T., Wang Z. Y., Pungor J. R., *et al.*, 2015 The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature 524: 220–224.

Alberts B., Johnson A., Lewis J., Morgan D., Raff M., *et al.*, 2014 *Molecular Biology of the Cell*. Garland Science.

Alegado R. a, Brown L. W., Cao S., Dermenjian R. K., Zuzow R., *et al.*, 2012 A bacterial sulfonolipid triggers multicellular development in the closest living relatives of animals. Elife 1: 1–16.

Alegado R. A., King N., 2014 Bacterial Influences on Animal Origins. Cold Spring Harb. Perspect. Biol. 6.

Allwood A. C., Burch I. W., Kamber B. S., 2007 3.43 billion-year-old stromatolite reef from the Pilbara Craton of Western Australia: Ecosystem-scale insights to early life on Earth. Precambrian Res. 158: 198–227.

Altenhoff A. M., Boeckmann B., Capella-Gutierrez S., Dalquen D. A., DeLuca T., *et al.*, 2016 Standardized benchmarking in the quest for orthologs. Nat. Methods 13: 425–430.

Antcliffe J. B., Callow R. H. T., Brasier M. D., 2014 Giving the early fossil record of sponges a squeeze. Biol. Rev. 89: 972–1004.

Aouacheria A., Geourjon C., Aghajari N., Navratil V., Deleage G., *et al.*, 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. Mol. Biol. Evol. 23: 2288–2302.

Aravind L., Iyer L. M., Koonin E. V, 2006 Comparative genomics and structural biology of the molecular innovations of eukaryotes. Curr. Opin. Struct. Biol. 16: 409–19.

Ast G., 2004 How did alternative splicing evolve? Nat. Rev. Genet. 5: 773–782.

Athanasopoulos V., Barker A., Yu D., Tan A. H.-M., Srivastava M., *et al.*, 2010 The ROQUIN family of proteins localizes to stress granules via the ROQ domain and binds target mRNAs. FEBS J. 277: 2109–27.

Babenko V. N., Rogozin I. B., Mekhedov S. L., Koonin E. V., 2004 Prevalence of intron gain over intron los in the evolution of paralogous gene families. Nucleic Acids Res. 32: 3724–3733.

Barbosa-Morais N. L., Irimia M., Pan Q., Xiong H. Y., Gueroussov S., *et al.*, 2012 The evolutionary landscape of alternative splicing in vertebrate species. Science 338: 1587–93.

Basu M. K., Carmel L., Rogozin I. B., Koonin E. V, 2008 Evolution of protein domain promiscuity in eukaryotes. Genome Res. 18: 449–61.

Basu M. K., Poliakov E., Rogozin I. B., 2009 Domain mobility in proteins: functional and evolutionary implications. Brief. Bioinform. 10: 205–16.

Bauer S., Grossmann S., Vingron M., Robinson P. N., 2008 Ontologizer 2.0 - A multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24: 1650–1651.

Becker B., 2013 Snow ball earth and the split of Streptophyta and Chlorophyta. Trends Plant Sci. 18: 180–183.

Beltrao P., Bork P., Krogan N. J., Noort V. Van, 2013 Evolution and functional cross-talk of protein post-translational modifications. Mol. Syst. Biol. 9: 1–13.

Bengtson S., Sallstedt T., Belivanova V., Whitehouse M., 2017 Three-dimensional preservation of cellular and subcellular structures suggests 1.6 billion-year-old crown-group red algae (D Penny, Ed.). PLoS Biol. 15: e2000735.

Berney C. C., Pawlowski J., 2006 A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. Proc. R. Soc. B Biol. Sci. 273: 1867–72.

Bonner J. T. (Ed.), 2000a The Origin of Multicellularity. In: *First Signals*, The Evolution of Multicellular Development. Princeton University Press, pp. 19–48.

Bonner J. T. (Ed.), 2000b The Evolution of Signaling. In: *First Signals*, The Evolution of Multicellular Development. Princeton University Press, pp. 63–72.

Bonner J. T. (Ed.), 2000c The Basic Elements of Multicellular Development. In: *First Signals*, The Evolution of Multicellular Development. Princeton University Press, pp. 73–92.

Bonner J. T. (Ed.), 2000d Size and Evolution. In: *First Signals*, The Evolution of Multicellular Development. Princeton University Press, pp. 49–62.

Bonner J. T. (Ed.), 2000e Development in the Cellular Slime Molds. In: *First Signals*, The Evolution of Multicellular Development. Princeton University Press, pp. 93–130.

Booth A., Doolittle W. F., 2015 Eukaryogenesis, how special really? Proc Natl Acad Sci U S A 112: 10278–10285.

Boraas M. E., Seale D. B., Boxhorn J. E., 1998 Phagotrophy by flagellate selects for colonial prey: A possible origin of multicellularity. Evol. Ecol. 12: 153–164.

Borenstein E., Shlomi T., Ruppin E., Sharan R., 2006 Gene loss rate: a probabilistic measure for the conservation of eukaryotic genes. Nucleic Acids Res. 35: e7–e7.

Bouget F. Y., Berger F., Brownlee C., 1998 Position dependent control of cell fate in the Fucus embryo: role of intercellular communication. Development 125: 1999–2008.

Braunschweig U., Barbosa-Morais N. L., Pan Q., Nachman E. N., Alipanahi B., *et al.*, 2014 Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. 24: 1774–1786.

Breitbart R. E., Andreadis A., Nadal-Ginard B., 1987 Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. Annu. Rev. Biochem. 56: 467–495.

Brooke N., Garcia-Fernandez J., Holland P., 1998 The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. Nature 392: 920–922.

Brunet T., Fischer A. H., Steinmetz P. R., Lauri A., Bertucci P., *et al.*, 2016 The evolutionary origin of bilaterian smooth and striated myocytes. Elife 5: e19607.

Budd G. E., Jensen S., 2007 A critical reappraisal of the fossil record of the bilaterian phyla. Biol. Rev. 75: 253–295.

Budd G. E., Jensen S., 2015 The origin of the animals and a "Savannah" hypothesis for early bilaterian evolution. Biol. Rev. 92: 446–473.

Burroughs A. M., Jaffee M., Iyer L. M., Aravind L., 2008 Anatomy of the E2 ligase fold: implications for enzymology and evolution of ubiquitin/Ub-like protein conjugation. J. Struct. Biol. 162: 205–18.

Burroughs A. M., Iyer L. M., Aravind L., 2009 Natural history of the E1-like superfamily: implication for adenylation, sulfur transfer, and ubiquitin conjugation. Proteins 75: 895–910.

Burt D. B., 2001 Evolutionary stasis, constraint and other terminology describing evolutionary patterns. Biol. J. Linn. Soc. 72: 509–517.

Bush S. J., Chen L., Tovar-Corona J. M., Urrutia A. O., 2017 Alternative splicing and the evolution of phenotypic novelty. Philos. Trans. R. Soc. B Biol. Sci. 372: 20150474.

Buss L. W., 1987 *The Evolution of Individuality*. Princeton University Press.

Bütschli, 1884 Bemerkungen zur Gastraea-Theorie. Morphol. Jahrb.: 415–427.

Butterfield N. J., 2000 Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology 26: 386–404.

Butterfield N. J., 2007 Macroevolution and macroecology throught deep time. Palaeontology 50: 41–55.

Butterfield N. J., 2011 Animals and the invention of the Phanerozoic Earth system. Trends Ecol. Evol. 26: 81–87.

Campo J. del, Massana R., 2011 Emerging Diversity within Chrysophytes, Choanoflagellates and Bicosoecids Based on Molecular Surveys. Protist 162: 435–448.

Campo J. del, Mallo D., Massana R., Vargas C. de, Richards T. A., *et al.*, 2015 Diversity and distribution of unicellular opisthokonts along the European coast analysed using high-throughput sequencing. Environ. Microbiol. 17: 3195–3207.

Campo J. Del, Ruiz-Trillo I., 2013 Environmental survey meta-analysis reveals hidden diversity among unicellular opisthokonts. Mol. Biol. Evol. 30: 802–805.

Capra E. J., Laub M. T., 2012 Evolution of two-component signal transduction systems. Annu. Rev. Microbiol. 66: 325–47.

Carmel L., Rogozin I. B., Wolf Y. I., Koonin E. V, 2007a Evolutionarily conserved genes preferentially accumulate introns. Genome Res. 17: 1045–1050.

Carmel L., Wolf Y. I., Rogozin I. B., Koonin E. V, 2007b Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res. 17: 1034–1044.

Carr M., Leadbeater B. S. C., Hassan R., Nelson M., Baldauf S. L., 2008 Molecular phylogeny of choanoflagellates, the sister group to Metazoa. Proc. Natl. Acad. Sci. 105: 16641–6.

Carr M., Richter D. J., Fozouni P., Smith T. J., Jeuck A., *et al.*, 2017 A six-gene phylogeny provides new insights into choanoflagellate evolution. Mol. Phylogenet. Evol. 107: 166–178.

Cavalier-Smith T., 1986 The origin of fungi and pseudofungi. In: Rayner ADM, Brasier CM, Moore D (Eds.), *Evolutionary biology of the fungi*, Cambridge University Press, pp. 339–353.

Cavalier-Smith T., Paula Allsopp M. T. E., 1996 Corallochytrium, an enigmatic non-flagellate protozoan related to choanoflagellates. Eur. J. Protistol. 32: 306–310.

Cavalier-Smith T., 1998a Neomonada and the origin of animals and fungi. In: Coombs G, Vickerman K, Sleigh M, Warren A (Eds.), *Evolutionary relationships among protozoa*, Chapman & Hall, London, pp. 375–407.

Cavalier-Smith T., 1998b A revised six-kingdom system of life. Biol. Rev. 73: 203–66.

Cavalier-Smith T., Chao E., 2003 Phylogeny of Choanozoa, Apusozoa, and Other Protozoa and Early Eukaryote Megaevolution. J. Mol. Evol.: 540–563.

Cavalier-Smith T., 2010 Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution. Biol. Direct 5: 7.

Cavalier-Smith T., 2017 Origin of animal multicellularity: precursors, causes, consequences—the choanoflagellate/sponge transition, neurogenesis and the Cambrian explosion. Philos. Trans. R. Soc. B Biol. Sci. 372: 20150476.

Chao A., 1987 Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43: 783–791.

Chen E. H., Grote E., Mohler W., Vignery A., 2007 Cell-cell fusion. FEBS Lett. 581: 2181–2193.

Chen L., Bush S. J., Tovar-Corona J. M., Castillo-Morales A., Urrutia A. O., 2014a Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Mol. Biol. Evol. 31: 1402–1413.

Chen L., Xiao S., Pang K., Zhou C., Yuan X., 2014b Cell differentiation and germ-soma separation in Ediacaran animal embryo-like fossils. Nature: 0–5.

Chiarucci A., Bacaro G., Rocchini D., Fattorini L., 2008 Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. Community Ecol. 9: 121–123.

Choi H.-S., Kim J.-R., Lee S.-W., Cho K.-H., 2008 Why have serine/threonine/tyrosine kinases been evolutionarily selected in eukaryotic signaling cascades? Comput. Biol. Chem. 32: 218–21.

Chosed R., Mukherjee S., Lois L. M., Orth K., 2006 Evolution of a signalling system that incorporates both redundancy and diversity: Arabidopsis SUMOylation. Biochem. J. 398: 521–9.

Clark K., Langeslag M., Figdor C. G., Leeuwen F. N. van, 2007 Myosin II and mechanotransduction: a balancing act. Trends Cell Biol. 17: 178–186.

Colby T., Matthäi A., Boeckelmann A., Stuible H.-P., 2006 SUMO-conjugating and SUMO-deconjugating enzymes from Arabidopsis. Plant Physiol. 142: 318–32.

Conti L. De, Baralle M., Buratti E., 2013 Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip. Rev. RNA 4: 49–60.

Corradi N., Slamovits C. H., 2011 The intriguing nature of microsporidian genomes. Brief. Funct. Genomics 10: 115–124.

Corradi N., 2015 Microsporidia: Eukaryotic Intracellular Parasites Shaped by Gene Loss and Horizontal Gene Transfers. Annu. Rev. Microbiol. 69: 150720190645000.

Cromar G., Wong K.-C., Loughran N., On T., Song H., *et al.*, 2014 New Tricks for "Old" Domains: How Novel Architectures and Promiscuous Hubs Contributed to the Organization and Evolution of the ECM. Genome Biol. Evol. 6: 2897–2917.

Csiszar K., 2001 Lysyl oxidases: A novel multifunctional amine oxidase family. In: *Progress in nucleic acid research and molecular biology*,, pp. 1–32.

Csűrös M., Miklós I., 2009 Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. Mol. Biol. Evol. 26: 2087–2095.

Csűrös M., 2010 Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26: 1910–2.

Csűrös M., Rogozin I. B., Koonin E. V, 2011 A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes (CP Ponting, Ed.). PLoS Comput. Biol. 7: e1002150.

Cunningham J. A., Liu A. G., Bengtson S., Donoghue P. C. J., 2016 The origin of animals: Can molecular clocks and the fossil record be reconciled? BioEssays 1600120: 1–12.

Curtis B. a, Tanifuji G., Burki F., Gruber A., Irimia M., *et al.*, 2012 Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature 492: 59–65.

Damuth J., Heisler I. L., 1988 Alternative formulations of multilevel selection. Biol. Philos. 3: 407–430.

Dayel M. J., King N., 2014 Prey capture and phagocytosis in the choanoflagellate Salpingoeca rosetta. PLoS One 9: e95577.

Dayraud C., Alié A., Jager M., Chang P., Guyader H. Le, *et al.*, 2012 Independent specialisation of myosin II paralogues in muscle vs. non-muscle functions during early animal evolution: a ctenophore perspective. BMC Evol. Biol. 12: 107.

Delsuc F., Brinkmann H., Philippe H., 2005 Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6: 361–375.

Deshaies R. J., Joazeiro C. a P., 2009 RING domain E3 ubiquitin ligases. Annu. Rev. Biochem. 78: 399–434.

Deshmukh K., Anamika K., Srinivasan N., 2010 Evolution of domain combinations in protein kinases and its implications for functional diversity. Prog. Biophys. Mol. Biol. 102: 1–15.

Dixon J. R., Selvaraj S., Yue F., Kim A., Li Y., *et al.*, 2012 Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485: 376–380.

Domazet-Lošo T., Tautz D., 2010 A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature 468: 815–8.

Dornbos S., Bottjer D., Chen J.-Y., others, 2004 Evidence for seafloor microbial mats and associated metazoan lifestyles in Lower Cambrian phosphorites of Southwest China. Lethaia 37: 127–138.

Du Q., Kawabe Y., Schilde C., Chen Z., Schaap P., 2015 The Evolution of Aggregative Multicellularity and Cell–Cell Communication in the Dictyostelia. J. Mol. Biol. 427: 3722–3733.

Duboule D., 2007 The rise and fall of Hox gene clusters. Development 134: 2549–2560.

Eisen J. A., Fraser C. M., 2003 Phylogenomics: Intersection of Evolution and Genomics. Science 300.

Ekman D., Björklund A. K., Elofsson A., 2007 Quantification of the elevated rate of domain rearrangements in metazoa. J. Mol. Biol. 372: 1337–48.

Elliott T. A., Gregory T. R., 2015a What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos. Trans. R. Soc. B Biol. Sci. 370: 20140331.

Elliott T. A., Gregory T. R., 2015b Do larger genomes contain more diverse transposable elements? BMC Evol. Biol. 15: 69.

Embley T. M., Martin W., 2006 Eukaryotic evolution, changes and challenges. Nature 440: 623–630.

Eme L., Sharpe S. C., Brown M. W., Roger A. J., 2014 On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks. Cold Spring Harb. Perspect. Biol. 6.

Emms D. M., Kelly S., 2015 OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16: 157.

Erwin D. H., 1993 The origin of metazoan development: a palaeobiological perspective. Biol. J. Linn. Soc. 50: 255–274.

Erwin D. H., Laflamme M., Tweedt S. M., Sperling E. A., Pisani D., *et al.*, 2011 The Cambrian Conundrum: Early Divergence and Later Ecological Success in the Early History of Animals. Science.

Exposito J. Y., Larroux C., Cluzel C., Valcourt U., Lethias C., *et al.*, 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. J. Biol. Chem. 283: 28226–28235.

Eyre D. R., Glimcher M. J., 1971 Comparative biochemistry of collagen crosslinks: Reducible bonds in invertebrate collagens. Biochim. Biophys. Acta - Protein Struct. 243: 525–529.

Fahey B., Degnan B. M., 2012 Origin and evolution of laminin gene family diversity. Mol. Biol. Evol. 29: 1823–1836.

Fairclough S. R., Dayel M. J., King N., 2010 Multicellular development in a choanoflagellate. Curr. Biol. 20: R875–R876.

Fairclough S. R., Chen Z., Kramer E., Zeng Q., Young S., *et al.*, 2013 Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. Genome Biol. 14: R15.

Ferrier D. E. K., 2016 Evolution of Homeobox Gene Clusters in Animals: The Giga-Cluster and Primary vs. Secondary Clustering. Front. Ecol. Evol. 4: 1–13.

Fortunato S. a. V., Adamski M., Ramos O. M., Leininger S., Liu J., *et al.*, 2014 Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. Nature 514: 620–623.

Francis W. R., Eitel M., Vargas S., Adamski M., Haddock S. H. D., *et al.*, 2017 The genome of the contractile demosponge *Tethya wilhelma* and the evolution of metazoan neural signalling pathways. bioRxiv.

Gaiti F., Fernandez-valverde S. L., Nakanishi N., Calcino D., Yanai I., *et al.*, 2015 Dynamic and widespread lncRNA expression in the sponge and the origin of animal complexity. : 1–42.

Gaiti F., Calcino A. D., Tanurdžić M., Degnan B. M., 2016 Origin and evolution of the metazoan non-coding regulatory genome. Dev. Biol.

Gaiti F., Jindrich K., Fernandez-Valverde S. L., Roper K. E., Degnan B. M., *et al.*, 2017 Landscape of histone modifications in a sponge reveals the origin of animal cis-regulatory complexity. Elife 6.

Gazave E., Lapébie P., Richards G. S., Brunet F., Ereskovsky A. V, *et al.*, 2009 Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes. BMC Evol. Biol. 9: 249.

Gilbert W., 1978 Why genes in pieces? Nature 271: 501–501.

Glockling S. L., Marshall W. L., Gleason F. H., 2013 Phylogenetic interpretations and ecological potentials of the Mesomycetozoea (Ichthyosporea). Fungal Ecol.: 1–11.

Godfrey-Smith P., 2009 *Darwinian Populations and Natural Selection*. OUP Oxford.

Goodson H. V, Spudich J. a, 1993 Molecular evolution of the myosin family: relationships derived from comparisons of amino acid sequences. Proc. Natl. Acad. Sci. 90: 659–663.

Gould S. J., Vrba E. S., 1982 Exaptation—a Missing Term in the Science of Form. Paleobiology 8: 4–15.

Gracheva E. O., Cordero-Morales J. F., González-Carcacía J. A., Ingolia N. T., Manno C., *et al.*, 2011 Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats. Nature 476: 88–91.

Graveley B. R., 2001 Alternative splicing: increasing diversity in the proteomic world. Trends Genet. 17: 100–107.

Grosberg R. K., Strathmann R. R., 2007 The Evolution of Multicellularity: A Minor Major Transition? Annu. Rev. Ecol. Evol. Syst. 38: 621–654.

Groussin M., Boussau B., Szöllõsi G., Eme L., Gouy M., *et al.*, 2015 Gene acquisitions from bacteria at the origins of major archaeal clades are vastly overestimated . Mol. Biol. Evol. 33: msv249.

Gueroussov S., 2015 An alternative splicing event amplifies evolutionary differences between vertebrates. Science.

Haeckel E., 1874 Die Gastraea-Theorie, die phylogenetische Classification des Thierreichs und die Homologie der Keimblatter. Jenaische Zeitschrift für Naturwiss. 8: 1–55.

Haggerty L. S., Jachiet P.-A., Hanage W. P., Fitzpatrick D., Lopez P., *et al.*, 2013 A pluralistic account of homology: adapting the models to the data. Mol. Biol. Evol.

Hamdi B., Brasier M. D., Zhiwen J., Aharon P., Schidlowski M., *et al.*, 1989 Earliest skeletal fossils from Precambrian–Cambrian boundary strata, Elburz Mountains, Iran. Geol. Mag. 126: 283.

Hammerschmidt K., Rose C. J., Kerr B., Rainey P. B., 2014 Life cycles, fitness decoupling and the evolution of multicellularity. Nature 515: 75–79.

Hanschen E. R., Marriage T. N., Ferris P. J., Hamaji T., Toyoda A., *et al.*, 2016 The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. Nat. Commun. 7: 11370.

Hassett B. T., López J. A., Gradinger R., 2015 Two New Species of Marine Saprophytic Sphaeroformids in the Mesomycetozoea Isolated From the Sub-Arctic Bering Sea. Protist.

Heino J., 2007 The collagen family members as cell adhesion proteins. BioEssays 29: 1001–1010.

Heisler I. L., Damuth J., 1987 A Method for Analyzing Selection in Hierarchically Structured Populations. Am. Nat. 130: 582–602.

Herron M. D., Hackett J. D., Aylward F. O., Michod R. E., 2009 Triassic origin and early radiation of multicellular volvocine algae. 2009: 6–10.

Hertel L. A., Bayne C. J., Loker E. S., 2002 The symbiont Capsaspora owczarzaki, nov. gen. nov. sp., isolated from three strains of the pulmonate snail Biomphalaria glabrata is related to members of the Mesomycetozoea. Int. J. Parasitol. 32: 1183–1191.

Hir H. Le, Nott A., Moore M. J., 2003 How introns influence and enhance eukaryotic gene expression. Trends Biochem. Sci. 28: 215–220.

Hjelmen C. E., Johnston J. S., 2017 The mode and tempo of genome size evolution in the subgenus Sophophora (I V. Sharakhov, Ed.). PLoS One 12: e0173505.

Hochstrasser M., 2000 Evolution and function of ubiquitin-like protein-conjugation systems. Nat. Cell Biol. 2: E153-7.

Hochstrasser M., 2009 Origin and function of ubiquitin-like proteins. Nature 458: 422–9.

Holland P. W. H., Booth H. A. F., Bruford E. A., 2007 Classification and nomenclature of all human homeobox genes. BMC Biol. 5: 47.

Holland P. W. H., 2013 Evolution of homeobox genes. Wiley Interdiscip. Rev. Dev. Biol. 2: 31–45.

Holland P. W. H., 2015 Did homeobox gene duplications contribute to the Cambrian explosion? Zool. Lett. 1: 1.

Holstein T. W., 2012 The Evolution of the Wnt Pathway. Cold Spring Harb. Perspect. Biol. 4: a007922–a007922.

Hua-Van A., Rouzic A. Le, Boutin T. S., Filée J., Capy P., 2011 The struggle for life of the genome's selfish architects. Biol. Direct 6: 19.

Hudry B., Thomas-Chollier M., Volovik Y., Duffraisse M., Dard A., *et al.*, 2014 Molecular insights into the origin of the Hox-TALE patterning system. Elife 2014: 1–24.

Huerta-Cepas J., Forslund K., Szklarczyk D., Jensen L. J., Mering C. von, *et al.*, 2016 Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. bioRxiv.

Huldtgren T., Cunningham J. A., Yin C., Stampanoni M., Marone F., *et al.*, 2011 Fossilized nuclei and germination structures identify Ediacaran "animal embryos" as encysting protists. Science 334: 1696–9.

Huldtgren T., Cunningham J. A., Yin C., Stampanoni M., Marone F., *et al.*, 2012 Response to Comment on "Fossilized Nuclei and Germination Structures Identify Ediacaran 'Animal Embryos' as Encysting Protists." Science 335: 1169–1169.

Hunter T., 2009 Tyrosine phosphorylation: thirty years and counting. Curr. Opin. Cell Biol. 21: 140–146.

Hurst L. D., Pál C., Lercher M. J., 2004 The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. 5: 299–310.

Hynes R. O., 2012 The evolution of metazoan extracellular matrix. J. Cell Biol. 196: 671–9.

Irimia M., Blencowe B. J., 2012 Alternative splicing: decoding an expansive regulatory layer. Curr. Opin. Cell Biol. 24: 323–32.

Irimia M., Tena J. J., Alexis M. S., Fernandez-Minan A., Maeso I., *et al.*, 2012 Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Res. 22: 2356–2367.

Ispolatov I., Ackermann M., Doebeli M., 2012 Division of labour and the evolution of multicellularity. Proc. R. Soc. B Biol. Sci. 279: 1768–1776.

Iyer L. M., Burroughs A. M., Aravind L., 2006 The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. Genome Biol. 7: R60.

Jaillon O., Aury J., Noel B., Policriti A., Clepet C., *et al.*, 2007 The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463–467.

James-Clark H., 1866 Conclusive proofs of the animality of the ciliate sponges, and of their affinities with the Infusoria flagellata. Am. J. Sci. Series 2 V: 320–324.

James-Clark H., 1871 Note on the Infusoria flagellata and the Spongiae ciliatae. Am. J. Sci. s3-1: 113–114.

Jones J. M., Gellert M., 2004 The taming of a transposon: V(D)J recombination and the immune system. Immunol. Rev. 200: 233–248.

Jordan I. K., Rogozin I. B., Glazko G. V, Koonin E. V, 2003 Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 19: 68–72.

Kasting J. F., Siefert J. L., 2002 Life and the evolution of Earth's atmosphere. Science 296: 1066–8.

Keeling P. J., Inagaki Y., 2004 A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1 ␣. 101.

Kelemen O., Convertini P., Zhang Z., Wen Y., Shen M., *et al.*, 2013 Function of alternative splicing. Gene 514: 1–30.

Keller R., 2005 Cell migration during gastrulation. Curr. Opin. Cell Biol. 17: 533–541.

Kim E., Simpson A. G. B., Graham L. E., 2006 Evolutionary relationships of apusomonads inferred from taxon-rich analyses of 6 nuclear encoded genes. Mol. Biol. Evol. 23: 2455–2466.

Kim E., Magen A., Ast G., 2007 Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35: 125–131.

King N., Hittinger C. T., Carroll S. B., 2003 Evolution of key cell signaling and adhesion protein families predates animal origins. Science 301: 361–3.

King N., 2004 The unicellular ancestry of animal development. Dev. Cell 7: 313–25.

King N., Westbrook M. J., Young S. L., Kuo A., Abedin M., *et al.*, 2008 The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature 451: 783–8.

Kirk D. L., 2003 Seeking the Ultimate and Proximate Causes of Volvox Multicellularity and Cellular Differentiation. Integr. Comp. Biol. 43: 247–253.

Knoll A. H., Carroll S. B., 1999 Early animal evolution: emerging views from comparative biology and geology. Science 284: 2129–2137.

Knoll A. H., Hewitt D., 2011 Phylogenetic, Functional, and Geological Perspectives on Complex Multicellularity. In: *The Major Transitions in Evolution Revisited*,

Knoll A. H., 2011 The Multiple Origins of Complex Multicellularity. Annu. Rev. Earth Planet. Sci. 39: 217–239.

Knoll A. H., 2014 Paleobiological perspectives on early eukaryotic evolution. Cold Spring Harb. Perspect. Biol. 6: 1–14.

Kondrashov F. A., Rogozin I. B., Wolf Y. I., Koonin E. V, 2002 Selection in the evolution of gene duplications. Genome Biol. 3: research0008.

Koonin E. V, 2004 A Non-Adaptationist Perspective on Evolution of Genomic Complexity or the Continued Dethroning of Man. Cell Cycle 3: 278–283.

Koonin E. V, 2009 Evolution of genome architecture. Int. J. Biochem. Cell Biol. 41: 298–306.

Koonin E. V, Wolf Y. I., 2010 Constraints and plasticity in genome and molecular-phenome evolution. Nat. Rev. Genet. 11: 487–98.

Koonin E. V, 2011 *The Logic of Chance: The Nature and Origin of Biological Evolution* (FT Press, Ed.). Pearson Education, Upper Saddle River, New Jersey 07458.

Koonin E. V, Csuros M., Rogozin I. B., 2013 Whence genes in pieces: Reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. Wiley Interdiscip. Rev. RNA 4: 93–105.

Koonin E. V, 2016 Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. BMC Biol. 14: 114.

Krebs E. G., Fischer E. H., 1956 The phosphorylase b to a converting enzyme of rabbit skeletal muscle. Biochim. Biophys. Acta 20: 150–157.

Kristensen D. M., Wolf Y. I., Mushegian A. R., Koonin E. V, 2011 Computational methods for Gene Orthology inference. Brief. Bioinform. 12: 379–391.

Lane N., Martin W., 2010 The energetics of genome complexity. Nature 467: 929–34.

Lane N., 2011 Energetics and genetics across the prokaryote-eukaryote divide. Biol. Direct 6: 35.

Lang B. F., O'Kelly C., Nerad T., Gray M. W., Burger G., 2002 The closest unicellular relatives of animals. Curr. Biol. 12: 1773–1778.

Larson G., Stephens P. a., Tehrani J. J., Layton R. H., 2013 Exapting exaptation. Trends Ecol. Evol.: 1–2.

Lartillot N., 2015 Probabilistic models of eukaryotic evolution: time for integration. Phil. Trans. R. Soc. B 370: 20140338.

Lee B. K., Iyer V. R., 2012 Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation. J. Biol. Chem. 287: 30906–30913.

Lee M. J., Yaffe M. B., 2016 Protein Regulation in Signal Transduction. Cold Spring Harb. Perspect. Biol. 8: a005918.

Lespinet O., Wolf Y. I., Koonin E. V, Aravind L., 2002 The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res. 12: 1048–59.

Levin L., Gutiérrez D., Rathburn A., Neira C., Sellanes J., *et al.*, 2002 Benthic processes on the Peru margin: a transect across the oxygen minimum zone during the 1997–98 El Niño. Prog. Oceanogr. 53: 1–27.

Levin T. C., Greaney A. J., Wetzel L., King N., 2014 The rosetteless gene controls development in the choanoflagellate S. rosetta. Elife 3: e04070.

Levitt M., 2009 Nature of the protein universe. Proc. Natl. Acad. Sci. 106: 11079–84.

Li W., Tucker A. E., Sung W., Thomas W. K., Lynch M., 2009 Extensive, recent intron gains in *Daphnia* populations. Science 326: 1260–1262.

Liu M., Walch H., Wu S., Grigoriev A., 2005 Significant expansion of exon-bordering protein domains during animal proteome evolution. Nucleic Acids Res. 33: 95–105.

Liu Y., Steenkamp E. T., Brinkmann H., Forget L., Philippe H., *et al.*, 2009 Phylogenomic analyses predict sistergroup relationship of nucleariids and fungi and paraphyly of zygomycetes with significant support. BMC Evol. Biol. 9: 272.

Lobkovsky A. E., Wolf Y. I., Koonin E. V, 2013 Gene frequency distributions reject a neutral model of genome evolution. Genome Biol. Evol. 5: 233–242.

Logares R., Audic S., Bass D., Bittner L., Boutte C., *et al.*, 2014 Patterns of rare and abundant marine microbial eukaryotes. Curr. Biol. 24: 813–821.

Lonfat N., Duboule D., 2015 Structure, function and evolution of topologically associating domains (TADs) at HOX loci. FEBS Lett. 589: 2869–2876.

Lücking R., Huhndorf S., Pfister D. H., Plata E. R., Lumbsch H. T., 2009 Fungi evolved right on track. Mycologia 101: 810–822.

Lynch M., 2002 Intron evolution as a population-genetic process. Proc. Natl. Acad. Sci. 99: 6118–23.

Lynch M., 2003 *The Origins of Genome Architecture*. Sinauer Associates.

Lynch M., Conery J. S., 2003 The origins of genome complexity. Science 302: 1401–4.

Lynch M., 2006a The origins of eukaryotic gene structure. Mol. Biol. Evol. 23: 450–468.

Lynch M., 2006b Streamlining and Simplification of Microbial Genome Architecture. Annu. Rev. Microbiol. 60: 327–349.

Lynch M., 2007 The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. 104: 8597–8604.

Lynch M., Bobay L.-M., Catania F., Gout J.-F., Rho M., 2011 The repatterning of eukaryotic genomes by random genetic drift. Annu. Rev. Genomics Hum. Genet. 12: 347–366.

Lynch M., Marinov G. K., 2017 Membranes, energetics, and evolution across the prokaryote-eukaryote divide. Elife 6: 1–29.

Maeso I., Roy S. W., Irimia M., 2012 Widespread recurrent evolution of genomic features. Genome Biol. Evol. 4: 486–500.

Mah J. L., Christensen-Dalsgaard K. K., Leys S. P., 2014 Choanoflagellate and choanocyte collar-flagellar systems and the assumption of homology. Evol. Dev. 16: 25–37.

Maldonado M., 2004 Choanoflagellates, choanocytes, and animal multicellularity. Invertebr. Biol. 123: 1–22.

Maloof A. C., Rose C. V., Beach R., Samuels B. M., Calmet C. C., *et al.*, 2010 Possible animal-body fossils in pre-Marinoan limestones from South Australia. Nat. Geosci. 3: 653–659.

Manning G., Young S. L., Miller W. T., Zhai Y., 2008 The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. Proc. Natl. Acad. Sci. 105: 9674–79.

Marshall W. L., Celio G., McLaughlin D. J., Berbee M. L., 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.

Marshall W. L., Berbee M. L., 2011 Facing unknowns: living cultures (Pirum gemmata gen. nov., sp. nov., and Abeoforma whisleri, gen. nov., sp. nov.) from invertebrate digestive tracts represent an undescribed clade within the unicellular Opisthokont lineage ichthyosporea (Mesomycetozoea). Protist 162: 33–57.

McFall-Ngai M., Hadfield M. G., Bosch T. C. G., Carey H. V, Domazet-Lošo T., *et al.*, 2013 Animals in a bacterial world, a new imperative for the life sciences. Proc. Natl. Acad. Sci. 110: 3229–3236.

McGuire A. M., Pearson M. D., Neafsey D. E., Galagan J. E., 2008 Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biol. 9: R50.

Mendoza L., Ajello L., Taylor J. W., 2001 The taxonomic status of Lacazia loboi and Rhinosporidium seeberi has been finally resolved with the use of molecular tools. Rev. Iberoam. Micol. 18: 95–8.

Mendoza L., Taylor J. W., Ajello L., 2002 The class mesomycetozoea: a heterogeneous group of microorganisms at the animal-fungal boundary. Annu. Rev. Microbiol. 56: 315–44.

Mendoza A. de, Sebé-Pedrós A., Sestak M. S., Matejcic M., Torruella G., *et al.*, 2013 Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc. Natl. Acad. Sci.

Mendoza A. de, 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 Biol. Evol. 6: 606–619.

Mendoza A. de, 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: 7250–7.

Metchnikoff E., Metschnikoff E., 1886 *Embryologische Studien an Medusen : Ein Beitrag zur Genealogie der Primitiv-organe / von Elias Metschnikoff.* A. Hölder, Wien :

Michael T. P., 2014 Plant genome size variation: bloating and purging DNA. Brief. Funct. Genomics 13: 308–317.

Michelle C., Vourc'h P., Mignon L., Andres C. R., 2009 What was the set of ubiquitin and ubiquitin-like conjugating enzymes in the eukaryote common ancestor? J. Mol. Evol. 68: 616–28.

Michod R. E., Roze D., 2001 Cooperation and conflict in the evolution of multicellularity. Heredity (Edinb). 86: 1–7.

Michod R. E., 2003 On the Reorganization of Fitness During Evolutionary Transitions in Individuality. Integr. Comp. Biol. 43: 64–73.

Michod R. E., 2006 The group covariance effect and fitness trade-offs during evolutionary transitions in individuality. Proc. Natl. Acad. Sci. 103: 9113–9117.

Michod R. E., 2007 Evolution of individuality during the transition from unicellular to multicellular life. Proc. Natl. Acad. Sci. 104 Suppl: 8613–8.

Mikhailov K. V, Konstantinova A. V., Nikitin M. A., Troshin P. V, Rusin L. Y., *et al.*, 2009 The origin of Metazoa: a transition from temporal to spatial cell differentiation. BioEssays 31: 758–768.

Mills D. B., Canfield D. E., 2014 Oxygen and animal evolution: Did a rise of atmospheric oxygen "trigger" the origin of animals? BioEssays 36: 1145–1155.

Mills D. B., Ward L. M., Jones C., Sweeten B., Forth M., *et al.*, 2014 Oxygen requirements of the earliest animals. : 2–6.

Moore A. D., Grath S., Schüler A., Huylmans A. K., Bornberg-Bauer E., 2013 Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. Biochim. Biophys. Acta - Proteins Proteomics 1834: 898–907.

Moroz L. L., Kocot K. M., Citarella M. R., Dosung S., Norekian T. P., *et al.*, 2014 The ctenophore genome and the evolutionary origins of neural systems. Nature 510: 109–114.

Mshigeni K. E., Lorri W. S. M., 1977 Spore germination and early stages of development in Hypnea musciformis (Rhodophyta, Gigartinales). Mar. Biol. 42: 161–164.

Mulder G. J., 1839 Ueber die Zusammensetzung einiger thierischen Substanzen. J. für Prakt. Chemie 16: 129–152.

Nakatani Y., Takeda H., Kohara Y., Morishita S., 2007 Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res. 17: 1254–1265.

Narbonne G. M., Gehling J. G., 2003 Life after snowball: The oldest complex Ediacaran fossils. Geology 31: 27–30.

Narbonne G. M., 2005 The Ediacara Biota: Neoproterozoic Origin of Animals and Their Ecosystems. Annu. Rev. Earth Planet. Sci. 33: 421–442.

Negre B., Ruiz A., 2007 HOM-C evolution in Drosophila: is there a need for Hox gene clustering? Trends Genet. 23: 55–59.

Ness K. P. Van, Koob T. J., Eyre D. R., 1988 Collagen cross-linking: distribution of hydroxypyridinium cross-links among invertebrate phyla and tissues. Comp. Biochem. Physiol. Part B Comp. Biochem. 91: 531–534.

Newman S. a, 2012 Physico-genetic determinants in the evolution of development. Science 338: 217–9.

Nichols S. A., Roberts B. W., Richter D. J., Fairclough S. R., King N., 2012 Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/ -catenin complex. Proc. Natl. Acad. Sci. 109: 13046–13051.

Nielsen C., 2008 Six major steps in animal evolution: Are we derived sponge larvae? Evol. Dev. 10: 241–257.

Nilsen T. W., Graveley B. R., 2010 Expansion of the eukaryotic proteome by alternative splicing. Nature 463: 457–63.

Nitsche F., Weitere M., Scheckenbach F., Hausmann K., Wylezich C., *et al.*, 2007 Deep Sea Records of Choanoflagellates with a Description of Two New Species. Acta Protozool. 46: 99–106.

Nitsche F., Carr M., Arndt H., Leadbeater B. S. C., 2011 Higher level taxonomy and molecular phylogenetics of the Choanoflagellatea. J. Eukaryot. Microbiol. 58: 452–462.

Nosenko T., Schreiber F., Adamska M., Adamski M., Eitel M., *et al.*, 2013 Deep metazoan phylogeny: When different genes tell different stories. Mol. Phylogenet. Evol.

O'Malley M. A., Wideman J. G., Ruiz-Trillo I., 2016 Losing Complexity: The Role of Simplification in Macroevolution. Trends Ecol. Evol. 31: 608–621.

Oksanen J., Blanchet F. G., Friendly M., Kindt R., Legendre P., *et al.*, 2017 vegan: Community Ecology Package.

Olson B. J. S. C., Nedelcu A. M., 2016 Co-option during the evolution of multicellularity and developmental complexity in the volvocine green algae. Curr. Opin. Genet. Dev. 39: 107–115.

Ou Q., Xiao S., Han J., Sun G., Zhang F., *et al.*, 2015 A vanished history of skeletonization in Cambrian comb jellies. Sci. Adv.: 1–9.

Owczarzak A., Stibbs H. H., Bayne C. J., 1980 The destruction of Schistosoma mansoni mother sporocysts in vitro by amoebae isolated from Biomphalaria glabrata: an ultrastructural study. J. Invertebr. Pathol. 35: 26–33.

Patterson D. J., Nygaard K., Steinberg G., Turley C. M., 1993 Heterotrophic flagellates and other protists associated with oceanic detritus throughout the water column in the mid North Atlantic. J. Mar. Biol. Assoc. United Kingdom 73: 67–95.

Peterson K. J., Mcpeek M. A., Evans D. A. D., 2005 Tempo and mode of early animal evolution: inferences from rocks, Hox, and molecular clocks. 31: 36–55.

Pisani D., Pett W., Dohrmann M., Feuda R., Rota-stabelli O., *et al.*, 2015 Genomic data do not support comb jellies as the sister group to all other animals. : 1–6.

Pittis A. A., Gabaldón T., 2016 Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature 531: 101–4.

Pope B. D., Ryba T., Dileep V., Yue F., Wu W., *et al.*, 2014 Topologically associating domains are stable units of replication-timing regulation. Nature 515: 402–405.

Porter S., 2011 The rise of predators. Geology 39: 607–608.

Putnam N. H., Srivastava M., Hellsten U., Dirks B., Chapman J., *et al.*, 2007 Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317: 86–94.

Putnam N. H., Butts T., Ferrier D. E. K., Furlong R. F., Hellsten U., *et al.*, 2008 The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064–1071.

Ragan M. A., Goggin C. L., Cawthorn R. J., Cerenius L., Jamieson A. V, *et al.*, 1996 A novel clade of protistan parasites near the animal-fungal divergence. Proc. Natl. Acad. Sci. 93: 11907–11912.

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. Bot. Mar. 30: 83–89.

Rainey P. B., Monte S. De, 2014 Resolving Conflicts During the Evolutionary Transition to Multicellular Life. Annu. Rev. Ecol. Evol. Syst. 45: 599–620.

Ratcliff W. C., Herron M. D., Howell K., Pentz J. T., Rosenzweig F., *et al.*, 2013 Experimental evolution of an alternating uni- and multicellular life cycle in Chlamydomonas reinhardtii. Nat. Commun. 4: 2742.

Reis M. dos, Thawornwattana Y., Angelis K., Telford M. J., Donoghue P. C. J., *et al.*, 2015 Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. Curr. Biol. 25: 2939–2950.

Rho M., Zhou M., Gao X., Kim S., Tang H., *et al.*, 2010 Independent Mammalian Genome Contractions Following the KT Boundary. Genome Biol. Evol. 1: 2–12.

Richards G. S., Degnan B. M., 2009 The dawn of developmental signaling in the metazoa. Cold Spring Harb. Symp. Quant. Biol. 74: 81–90.

Richter D. J., King N., 2013 The Genomic and Cellular Foundations of Animal Origins. Annu. Rev. Genet.

Rossetti V., Schirrmeister B. E., Bernasconi M. V., Bagheri H. C., 2010 The evolutionary path to terminal differentiation and division of labor in cyanobacteria. J. Theor. Biol. 262: 23–34.

Rotin D., Kumar S., 2009 Physiological functions of the HECT family of ubiquitin ligases. Nat. Rev. Mol. Cell Biol. 10: 398–409.

Roy S. W., Irimia M., 2009 Splicing in the eukaryotic ancestor: form, function and dysfunction. Trends Ecol. Evol. 24: 447–55.

Roy S. W., 2016 Is Genome Complexity a Consequence of Inefficient Selection? Evidence from Intron Creation in Nonrecombining Regions. Mol. Biol. Evol. 33: 3088–3094.

Ruiz-Trillo I., Inagaki Y., Davis L. A., Sperstad S., Landfald B., *et al.*, 2004 Capsaspora owczarzaki is an independent opisthokont lineage. Curr. Biol. 14: R946-7.

Ruiz-Trillo I., Lane C. C. E., Archibald J. M., Roger A. J., 2006 Insights into the Evolutionary Origin and Genome Architecture of the Unicellular Opisthokonts Capsaspora owczarzaki and Sphaeroforma arctica. J. Eukaryot. Microbiol. 53: 379–384.

Ruiz-Trillo I., Burger G., Holland P. W. H., King N., Lang B. F., *et al.*, 2007 The origins of multicellularity: a multi-taxon genome initiative. Trends Genet. 23: 113–8.

Ruiz-Trillo I., Roger A. J., Burger G., Gray M. W., Lang B. F., 2008 A phylogenomic investigation into the origin of Metazoa. Mol. Biol. Evol. 25: 664–672.

Ryan J. F., Pang K., Schnitzler C. E., Nguyen A.-D., Moreland R. T., *et al.*, 2013 The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. Science 342: 1242592.

Sahoo S. K., Planavsky N. J., Kendall B., Wang X., Shi X., *et al.*, 2012 Ocean oxygenation in the wake of the Marinoan glaciation. Nature 488: 546–549.

Schönknecht G., Weber A. P. M., Lercher M. J., 2014 Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. Bioessays 36: 9–20.

Schopf J. W., 1993 Microfossils of the Early Archean Apex chert: new evidence of the antiquity of life. Science 260: 640–646.

Schultheiss K. P., Suga H., Ruiz-Trillo I., Miller W. T., 2012 Lack of Csk-Mediated Negative Regulation in a Unicellular Src Kinase. Biochemistry 51: 8267–8277.

Schulze F., 1885 Über das Verhältnis der Spongien zu den Choanoflagellaten. Sitzungsberichte der Königlich Preuss. Akad. der Wissenschaften 10: 1–13.

Sebé-Pedrós A., Roger A. J., Lang F. B., King N., Ruiz-Trillo I., 2010 Ancient origin of the integrin-mediated adhesion and signaling machinery. Proc. Natl. Acad. Sci. 107: 10142–7.

Sebé-Pedrós A., Ruiz-Trillo I., 2010 Integrin-mediated adhesion complex: Cooption of signaling systems at the dawn of Metazoa. Commun. Integr. Biol. 3: 475–7.

Sebé-Pedrós A., Mendoza A. de, Lang B. F., Degnan B. M., Ruiz-Trillo I., 2011 Unexpected Repertoire of Metazoan

Transcription Factors in the Unicellular Holozoan *Capsaspora owczarzaki*. Mol. Biol. Evol. 28: 1241–1254.

Sebé-Pedrós A., Zheng Y., Ruiz-Trillo I., Pan D., 2012 Premetazoan Origin of the Hippo Signaling Pathway. Cell Rep. 1: 13–20.

Sebé-Pedrós A., Irimia M., Campo J. Del, Parra-Acero H., Russ C., *et al.*, 2013a Regulated aggregative multicellularity in a close unicellular relative of metazoa. Elife 2: e01287.

Sebé-Pedrós A., Ariza-Cosano A., Weirauch M. T., Leininger S., Yang A., *et al.*, 2013b Early evolution of the T-box transcription factor family. Proc. Natl. Acad. Sci. 110: 16050–5.

Sebé-Pedrós A., Peña M. I., Capella-Gutiérrez S., Gabaldon T., Ruiz-Trillo I., *et al.*, 2016a High-throughput Proteomics Reveals the Unicellular Roots of Animal Phosphosignaling and Cell Differentiation. Dev. Cell In press: 1–12.

Sebé-Pedrós A., Ballaré C., Parra-Acero H., Chiva C., Tena J. J., *et al.*, 2016b The Dynamic Regulatory Genome of *Capsaspora* and the Origin of Animal Multicellularity. Cell: 1–14.

Sebé-Pedrós A., Degnan B. M., Ruiz-Trillo I., 2017 The origin of Metazoa, a unicellular perspective. Nat. Rev. Genet. in press.

Seitan V. C., Faure A. J., Zhan Y., Mccord R. P., Lajoie B. R., *et al.*, 2013 Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. : 2066–2077.

Sessegolo C., Burlet N., Haudry A., 2016 Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. Biol. Lett. 12: 20160407.

Shalchian-Tabrizi K., Minge M. a, Espelund M., Orr R., Ruden T., *et al.*, 2008 Multigene phylogeny of choanozoa and the origin of animals. PLoS One 3: e2098.

Shen X., Hittinger C. T., Rokas A., 2017 Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1: 126.

Shoguchi E., Shinzato C., Kawashima T., Gyoja F., Mungpakdee S., *et al.*, 2013 Draft Assembly of the Symbiodinium minutum Nuclear Genome Reveals Dinoflagellate Gene Structure. Curr. Biol. 23: 1399–1408.

Silberfeld T., Leigh J. W., Verbruggen H., Cruaud C., Reviers B. De, *et al.*, 2010 A multi-locus time-calibrated phylogeny of the brown algae (Heterokonta, Ochrophyta, Phaeophyceae): Investigating the evolutionary nature of the "brown algal crown radiation." Mol. Phylogenet. Evol. 56: 659–674.

Simakov O., Marletaz F., Cho S.-J., Edsinger-Gonzales E., Havlak P., *et al.*, 2013 Insights into bilaterian evolution from three spiralian genomes. Nature 493: 526–31.

Simakov O., Kawashima T., Marlétaz F., Jenkins J., Koyanagi R., *et al.*, 2015 Hemichordate genomes and deuterostome origins. Nature 527: 459–465.

Simakov O., Kawashima T., 2016 Independent evolution of genomic characters during major metazoan transitions. Dev. Biol.: 0–1.

Simion P., Philippe H., Baurain D., Jager M., Richter D. J., *et al.*, 2017 A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Curr. Biol.: 1–10.

Smith J. J., Keinath M. C., 2015 The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. : 1081–1090.

Smoly I., Shemesh N., Ziv-Ukelson M., Ben-Zvi A., Yeger-Lotem E., *et al.*, 2017 An Asymmetrically Balanced Organization of Kinases versus Phosphatases across Eukaryotes Determines Their Distinct Impacts (D Penny, Ed.). PLOS Comput. Biol. 13: e1005221.

Sperling E. A., Robinson J. M., Pisani D., Peterson K. J., 2010 Where's the glass? Biomarkers, molecular clocks, and microRNAs suggest a 200-Myr missing Precambrian fossil record of siliceous sponge spicules. Geobiology 8: 24–36.

Sperling E. A., Frieder C. A., Raman A. V., Girguis P. R., Levin L. A., *et al.*, 2013 Oxygen, ecology, and the Cambrian radiation of animals. Proc. Natl. Acad. Sci. 110: 13446–51.

Srivastava M., Begovic E., Chapman J., Putnam N. H., Hellsten U., *et al.*, 2008 The *Trichoplax* genome and the nature of placozoans. Nature 454: 955–60.

Srivastava M., Simakov O., Chapman J., Fahey B., Gauthier M. E. a, *et al.*, 2010 The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature 466: 720–726.

Stajich J. E., Berbee M. L., Blackwell M., Hibbett D. S., James T. Y., *et al.*, 2009 The Fungi. Curr. Biol. 19: R840-5.

Stanley S. M., 1973 An Ecological Theory for the Sudden Origin of Multicellular Life in the Late Precambrian. Proc. Natl. Acad. Sci. 70: 1486–1489.

Steenkamp E. T., Wright J., Baldauf S. L., 2006 The protistan origins of animals and fungi. Mol. Biol. Evol. 23: 93–106.

Steinmetz P. R. H., Kraus J. E. M., Larroux C., Hammel J. U., Amon-Hassenzahl A., *et al.*, 2012 Independent evolution of striated muscles in cnidarians and bilaterians. Nature 487: 231–4.

Suga H., Dacre M., Mendoza A. de, Shalchian-Tabrizi K., Manning G., *et al.*, 2012 Genomic Survey of Premetazoans Shows Deep Conservation of Cytoplasmic Tyrosine Kinases and Multiple Radiations of Receptor Tyrosine Kinases. Sci. Signal. 5: ra35-ra35.

Suga H., Chen Z., Mendoza A. de, Sebé-Pedrós A., Brown M. W., *et al.*, 2013 The *Capsaspora* genome reveals a complex unicellular prehistory of animals. Nat. Commun. 4: 2325.

Suga H., Ruiz-Trillo I., 2013 Development of ichthyosporeans sheds light on the origin of metazoan multicellularity. Dev. Biol.: 1–9.

Suga H., Torruella G., Burger G., Brown M. W., Ruiz-Trillo I., 2014 Earliest holozoan expansion of phosphotyrosine signaling. Mol. Biol. Evol. 31: 517–528.

Sumathi J. C., Raghukumar S., Kasbekar D. P., Raghukumar C., 2006 Molecular evidence of fungal signatures in the marine protist Corallochytrium limacisporum and its implications in the evolution of animals and fungi. Protist 157: 363–76.

Supek F., Bošnjak M., Škunca N., Šmuc T., 2011 REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6: e21800.

Szathmáry E., Smith J. M., 1995 The major evolutionary transitions. Nature 374: 227–232.

Szathmáry E., 2015 Toward major evolutionary transitions theory 2.0. Proc. Natl. Acad. Sci. 112: 10104–10111.

Tang F., Bengtson S., Wang Y., Wang X., Yin C., 2011 Eoandromeda and the origin of Ctenophora. 414: 408–414.

Tautz D., Domazet-Lošo T., 2011 The evolutionary origin of orphan genes. Nat. Rev. Genet. 12: 692–702.

Telford M. J., Budd G. E., Philippe H., 2015 Phylogenomic Insights into Animal Evolution. Curr. Biol. 25: R876–R887.

Tikhonenkov D., Janouškovec J., Hehenberger E., Burki F., Gawryluk R., *et al.*, 2016 The evolutionary importance of predatory flagellates: new deep branches on the eukaryotic tree of life. In: *International Scientific Forum on Protistology - Moscow*,

Tomitani A., Knoll A. H., Cavanaugh C. M., Ohno T., 2006 The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. Proc. Natl. Acad. Sci. 103: 5442–5447.

Tordai H., Nagy A., Farkas K., Bányai L., Patthy L., 2005 Modules, multidomain proteins and organismic complexity. FEBS J. 272: 5064–78.

Torruella G., Derelle R., Paps J., Lang B. F., Roger A. J., *et al.*, 2012 Phylogenetic Relationships within the Opisthokonta Based on Phylogenomic Analyses of Conserved Single-Copy Protein Domains. Mol. Biol. Evol. 29: 531–544.

Torruella G., 2014 Phylogeny and evolutionary perspective of Opisthokonta protists.

Torruella G., Mendoza A. de, Grau-Bové X., Antó M., Chaplin M. A., *et al.*, 2015 Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. Curr. Biol. 25: 1–7.

Towe K. M., 1970 Oxygen-collagen priority and the early metazoan fossil record. Proc. Natl. Acad. Sci. 65: 781–788.

Trommer G., Pondaven P., Siccha M., Stibor H., 2012 Zooplankton-mediated nutrient limitation patterns in marine phytoplankton: an experimental approach with natural communities. Mar. Ecol. Prog. Ser. 449: 83–94.

Umen J. G., 2014 Green Algae and the Origins of Multicellularity in the Plant Kingdom. Cold Spring Harb. Perspect. Biol. 6: 1–28.

Vargas C. de, Audic S., Henry N., Decelle J., Mahe F., *et al.*, 2015 Eukaryotic plankton diversity in the sunlit ocean. Science 348: 1261605–1261605.

Vurture G. W., Sedlazeck F. J., Nattestad M., Underwood C. J., Fang H., *et al.*, 2017 GenomeScope: Fast reference-free genome profiling from short reads. bioRxiv: 1–3.

Wellman C. H., Gray J., 2000 The microfossil record of early land plants. Philos. Trans. R. Soc. B Biol. Sci. 355: 712–717.

Whelan N. V, Kocot K. M., Moroz L. L., Halanych K. M., 2015 Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. 112: 5773–8.

Whitney K. D., Garland T., 2010 Did genetic drift drive increases in genome complexity? PLoS Genet. 6: 1–6.

Whitney K. D., Boussau B., Baack E. J., Garland T., 2011 Drift and genome complexity revisited. PLoS Genet. 7: 5–9.

Wolf Y. I., Makarova K. S., Yutin N., Koonin E. V, 2012 Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol. Direct 7: 46.

Wolf Y. I., Koonin E. V, 2013 Genome reduction as the dominant mode of evolution. Bioessays 35: 1521–878.

Woznica A., Cantley A. M., Beemelmanns C., Freinkman E., Clardy J., *et al.*, 2016 Bacterial lipids activate, synergize, and inhibit a developmental switch in choanoflagellates. Proc. Natl. Acad. Sci.: 201605015.

Wylezich C., Karpov S. A., Mylnikov A. P., Anderson R., Jurgens K., 2012 Ecologically relevant choanoflagellates collected from hypoxic water masses of the Baltic Sea have untypical mitochondrial cristae. BMC Microbiol. 12: 271.

Xiao S., Knoll A. H., Schiffbauer J. D., Zhou C., Yuan X., 2012 Comment on “Fossilized nuclei and germination structures identify Ediacaran ‘animal embryos’ as encysting protists”. Science 335: 1169; author reply 1169.

Xie X., Wang G., Pan G., Gao S., Xu P., *et al.*, 2010 Variations in morphology and PSII photosynthetic capabilities during the early development of tetraspores of Gracilaria vermiculophylla (Ohmi) Papenfuss (Gracilariales, Rhodophyta). BMC Dev. Biol. 10: 43.

Xie X., Jin J., Mao Y., 2011 Evolutionary versatility of eukaryotic protein domains revealed by their bigram networks. BMC Evol. Biol. 11: 242.

Yin Z., Zhu M., Davidson E. H., Bottjer D. J., Zhao F., *et al.*, 2015 Sponge grade body fossil with cellular resolution dating 60 Myr before the Cambrian. Proc. Natl. Acad. Sci. 112: E1453-60.

Yue J., Sun G., Hu X., Huang J., 2013 The scale and evolutionary significance of horizontal gene transfer in the choanoflagellate Monosiga brevicollis. BMC Genomics 14: 729.

Zakhvatkin A. A., 1949 The Comparative Embryology of the Low Invertebrates. Sources and Method of the Origin of Metazoan Development. Sov. Sci.: 395.

Zhang X., Smits A. H., Tilburg G. B. A. Van, Jansen P. W. T. C., Makowski M. M., *et al.*, 2017 An Interaction Landscape of Ubiquitin Signaling. Mol. Cell: 1–15.

Zmasek C. M., Godzik A., 2011 Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. 12: R4.

Zmasek C. M., Godzik A., 2012 This Déjà vu feeling--analysis of multidomain protein evolution in eukaryotic genomes. (CA Orengo, Ed.). PLoS Comput. Biol. 8: e1002701.

7. Annexes

Additional published results

7.1. Expression atlas of the deubiquitinating enzymes in the adult mouse retina, their evolutionary diversification and phenotypic roles

Additional publication

Esquerdo M, Grau-Bové X, Garanto A, Toulis V, Garcia-Monclús S, Millo E, López-Iniesta MJ, Abad-Morales V, Ruiz-Trillo I, Marfany G. 2016. Expression Atlas of the Deubiquitinating Enzymes in the Adult Mouse Retina, Their Evolutionary Diversification and Phenotypic Roles. PLoS One.

Impact Factor (2016): NA – previous available year: 3.057

Journal ranking: NA – previous available year: Multidisciplinary Sciences Q1 (11/63)

Authorship: XGB contributed to the experimental design of the phylogenetic analyses, and to their interpretation and discussion. Project conceptio and experimental design by AG and GM. Gene expression characterization by ME, AG, VT, SGM, EM, MJLI and VAM. Contributions of material and analytical tools by GM, XGB and IRT. Manuscript written by ME and GM.

Abstract – Ubiquitination is a relevant cell regulatory mechanism to determine protein fate and function. Most data has focused on the role of ubiquitin as a tag molecule to target substrates to proteasome degradation, and on its impact in the control of cell cycle, protein homeostasis and cancer. Only recently, systematic assays have pointed to the relevance of the ubiquitin pathway in the development and differentiation of tissues and organs, and its implication in hereditary diseases. Moreover, although the activity and composition of ubiquitin ligases has been largely addressed, the role of the deubiquitinating enzymes (DUBs) in specific tissues, such as the retina, remains mainly unknown. In this work, we undertook a systematic analysis of the transcriptional levels of DUB genes in the adult mouse retina by RT-qPCR and analyzed the expression pattern by in situ hybridization and fluorescent immunohistochemistry, thus providing a unique spatial reference map of retinal DUB expression. We also performed a systematic phylogenetic analysis to understand the origin and the presence/absence of DUB genes in the genomes of diverse animal taxa that represent most of the known animal diversity. The expression landscape obtained supports the potential sub-functionalization of paralogs in those families that expanded in vertebrates. Overall, our results constitute a reference framework for further characterization of the DUB roles in the retina and suggest new candidates for inherited retinal disorders.

PLOS ONE

RESEARCH ARTICLE

Expression Atlas of the Deubiquitinating Enzymes in the Adult Mouse Retina, Their Evolutionary Diversification and Phenotypic Roles

Mariona Esquerdo[1], Xavier Grau-Bové[1,2], Alejandro Garanto[1¤a], Vasileios Toulis[1], Sílvia Garcia-Monclús[1¤b], Erica Millo[1], Ma José López-Iniesta[1,3], Víctor Abad-Morales[1], Iñaki Ruiz-Trillo[1,2,4], Gemma Marfany[1,3,5]*

1 Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain, **2** Institut de Biologia Evolutiva (CSIC- Universitat Pompeu Fabra), Barcelona, Spain, **3** Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Instituto de Salud Carlos III, Barcelona, Spain, **4** Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, **5** Institut de Biomedicina de la Universitat de Barcelona (IBUB), Barcelona, Spain

¤a Current address: Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, Department of Human Genetics, Nijmegen, The Netherlands
¤b Current address: Sarcoma research group, Molecular Oncology Lab, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain
* gmarfany@ub.edu

Citation: Esquerdo M, Grau-Bové X, Garanto A, Toulis V, Garcia-Monclús S, Millo E, et al. (2016) Expression Atlas of the Deubiquitinating Enzymes in the Adult Mouse Retina, Their Evolutionary Diversification and Phenotypic Roles. PLoS ONE 11 (3): e0150364. doi:10.1371/journal.pone.0150364

Editor: Alfred S Lewin, University of Florida, UNITED STATES

Received: November 10, 2015

Accepted: February 12, 2016

Published: March 2, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was supported by grants BFU2010-15656 (MICINN) and SAF2013-49069-C2-1-R (MINECO) to G.M., and 2014SGR-0932 (Generalitat de Catalunya) grant (BFU-2011-23434) from Ministerio de Economía y Competitividad (MINECO) and co-funded by the Fondo Europeo de Desarrollo regional (FEDER) to I.R-T. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Ubiquitination is a relevant cell regulatory mechanism to determine protein fate and function. Most data has focused on the role of ubiquitin as a tag molecule to target substrates to proteasome degradation, and on its impact in the control of cell cycle, protein homeostasis and cancer. Only recently, systematic assays have pointed to the relevance of the ubiquitin pathway in the development and differentiation of tissues and organs, and its implication in hereditary diseases. Moreover, although the activity and composition of ubiquitin ligases has been largely addressed, the role of the deubiquitinating enzymes (DUBs) in specific tissues, such as the retina, remains mainly unknown. In this work, we undertook a systematic analysis of the transcriptional levels of DUB genes in the adult mouse retina by RT-qPCR and analyzed the expression pattern by *in situ* hybridization and fluorescent immunohistochemistry, thus providing a unique spatial reference map of retinal DUB expression. We also performed a systematic phylogenetic analysis to understand the origin and the presence/absence of DUB genes in the genomes of diverse animal taxa that represent most of the known animal diversity. The expression landscape obtained supports the potential subfunctionalization of paralogs in those families that expanded in vertebrates. Overall, our results constitute a reference framework for further characterization of the DUB roles in the retina and suggest new candidates for inherited retinal disorders.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Ubiquitination is a dynamic regulatory mechanism that controls cell processes such as protein quality control (via proteasome degradation), cellular signalling, transcriptional regulation or DNA repair [1–3]. As ubiquitination is reversible, cells deploy a large set of enzymes to conjugate (E1, E2 and E3 ligases) and deconjugate (deubiquitinating enzymes) ubiquitin moieties [4]. The human genome contains several hundreds of ubiquitin ligases, and close to 80 deubiquitinating enzymes (DUBs), indicating that: i) ubiquitination is a highly regulated process, and ii) substrate recognition specificity is inherent to the system.

Most data on the physiological relevance of ubiquitin has focused on its role as the tag molecule to target substrates to proteasome degradation, its role in cell cycle control and cancer, as well as its involvement in the molecular basis of neurodegenerative disorders [5,6]. Besides, a number of high-throughput approaches have focused on finding substrates for either ligases [7] or deubiquitinating enzymes (DUBs) [8]. Nonetheless, most high-throughput studies have been performed *in vitro* using mammalian cell cultures, and only recently, systematic assays in animal models have indicated the relevance of the ubiquitin pathway in the development, differentiation and maintenance of tissues and organs [9,10].

One of the tissues that requires a tight gene and protein regulation is the retina. The retina consists of structured layers of highly specialized neurons in the eye that capture and process light stimuli enabling vision [11]. Such a fine architecture turns retinal differentiation into an extremely complex mechanism that must be accurately regulated [12], and in which ubiquitin and ubiquitination play a relevant role. In fact, mutations in the genes encoding the E3 ligases TOPORS [13–15] and KLHL7 [16,17]; and in PRPF8, which belongs to the JAB1-MPN--MOV34 (JAMM) family of DUBs, are causative of the most prevalent retinal hereditary dystrophy, retinitis pigmentosa (RP). Moreover, protein homeostasis via the ubiquitin-proteasome system is also relevant to other retinal diseases and specific altered protein degradation has been associated to Stargardt's disease, age-related macular degeneration, glaucoma, diabetic retinopathy, and retinal inflammation (reviewed in [18]).

Lately, DUBs are becoming the focus of attention given that their specificity in substrate selection makes them key checkpoints of protein degradation and fate. Moreover, their fewer numbers (compared to E2 and E3 ligases) makes their functional analysis more feasible. An increasing number of reports propose DUBs as pharmacological targets in disease: cancer [19–21] and neurodegenerative diseases [6]. DUBs are classified into five different subfamilies depending on their catalytic domains [22]: Machado-Joseph Disease protein domain proteases (MJD), Ovarian Tumor proteases (OTU), Ubiquitin C-Terminal Hydrolases (UCH) and Ubiquitin-Specific Proteases (USP) are cysteine proteases, whereas JAB1/MPN/MOV34 family proteases (JAMM) are Zn^{2+} metalloproteases; overall adding up to 90 genes in the human genome, of which only 79 are predicted to be functional [1].

A recent review compiled the gathered knowledge of the functional roles of individual DUBs, focusing on their subcellular localization, levels of expression in human tissues, and gene mutation phenotype in human and model organisms [23], yet a comprehensive study on the expression pattern of DUBs in highly specialized tissues, such as the retina, has not been performed. Besides, previous comparisons of DUB mutant phenotypes in different model organisms attempt to directly assign, without a phylogenetic framework, orthology and function between invertebrate and vertebrate genes. Some of these assignments may need revision under robust phylogenetic data, since ubiquitin ligase and protease families have expanded in eukaryotes [24], and subfunctionalization and neofunctionalization are known to occur after gene expansion.

Thus, we here aimed to draw an expression pattern map for DUB genes in the mouse retina, by using RT-qPCR, *in situ* hybridization and immunohistochemistry. We have also applied comparative genomics to infer the basic protein domain architecture within the DUB subfamilies and illustrate their diversification within metazoans. These data combined with the reported phenotypes will help to identify relevant retinal genes and potential new candidates for retinal diseases. Overall, we provide a comprehensive reference framework on DUB function and their roles in neuronal tissues that will be useful for future functional and evolutionary studies.

Material and Methods

Ethics statement

All procedures in mice were performed according to the ARVO statement for the use of animals in ophthalmic and vision research, as well as the regulations of the Animal Care facilities at the Universitat de Barcelona. The protocols and detailed procedures were evaluated and approved by the Animal Research Ethics Committee (CEEA) of the Universitat de Barcelona (our institution), and were submitted and also approved by the Generalitat de Catalunya (local Government), with the official permit numbers DAAM 6562 and 7185.

Animal handling, tissue dissection and preparation of samples

Murine retina samples and eye slides were obtained from 2 month-old C57BL/6J (wild-type) and CD-1 (albino) animals. Animals were euthanized by cervical dislocation. Some retinas were dissected and immediately frozen in liquid nitrogen, while the rest were fixed in 4% paraformaldehyde (PFA) for 2 h at room temperature (RT), washed, cryoprotected overnight in acrylamide at 4°C, embedded in O.C.T. (Tissue-Tek, Sakura Finetech, Torrance, CA), frozen in liquid nitrogen and sectioned at -17°C.

RNA extraction and cDNA synthesis

For each sample, retinas from three different animals were pooled. Therefore, up to 9 animals in three independent replicates were analyzed. Retinas were homogenized using a Polytron PT 1200 E homogenizer (Kinematica, AG, Lucerne, Switzerland). Total RNA was extracted using the High Pure RNA Tissue Kit (Roche Diagnostics, Indianapolis, IN) following the manufacturer's instructions with minor modifications (increasing the DNAse I incubation step). Reverse transcription reactions were carried out using the qScript cDNA Synthesis Kit (Quanta Biosciences) following the manufacturer's protocol.

RT-qPCR

Quantitative reverse transcription PCR (RT-qPCR) was performed using the LightCycler® 480 SYBR Green I Master Mix (Roche Applied Science) and a LightCycler® 480 Multiwell Plate 384. The final reaction volume was 10 μl. Raw data was analyzed with the LightCycler® 480 software using the Advanced Relative Quantification method. *Gapdh* expression was used to normalize the levels of expression. *Rho* and *Cerkl* were considered as reference genes with high and low levels of expression, respectively, in the mouse retina. Three independent samples replicates were analyzed for each gene. Differences in gene expression levels within the same sample and between the samples were directly compared by their Z-score values. The mean and standard deviation of the Z-scores are plotted in Fig 1. The name and sequence of all the primers used for RT-qPCR and *in situ* hybridization are listed in S1 Table.

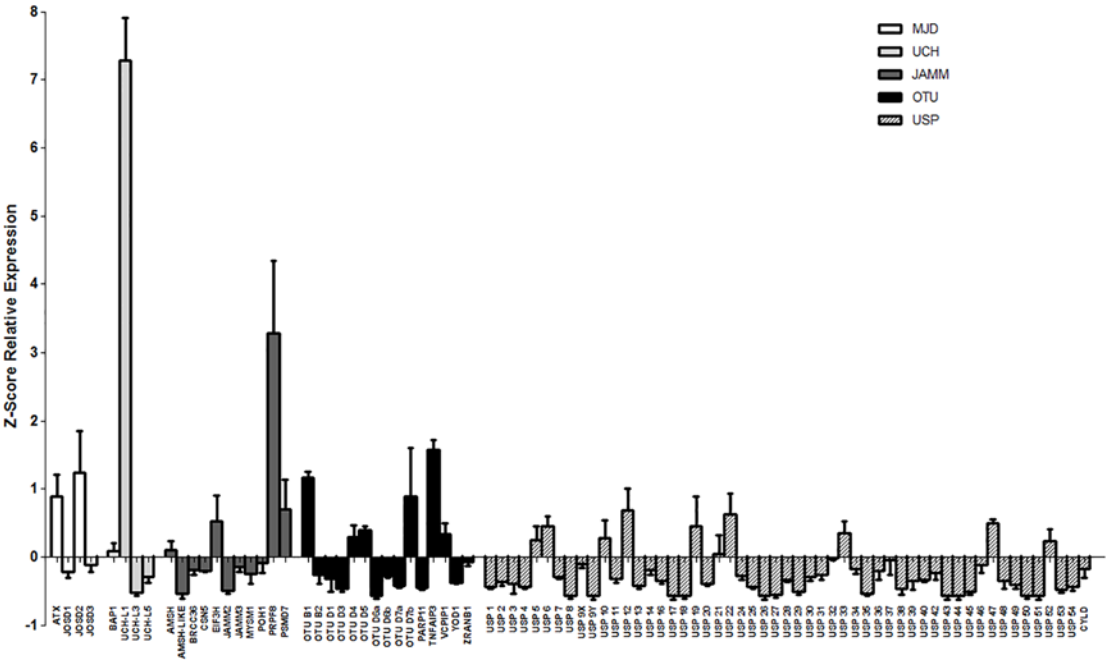


Fig 1. Relative expression levels of the five subfamilies of DUB enzymes. Gene expression values are the average of three independent samples (measured in three replicates), each sample contained retinas from three individuals. The expression levels are obtained as a ratio with *Gapdh* expression (used for normalization) per 10^4. The Z-score has been calculated for the whole set of genes per each sample, and mean and standard deviation has been obtained, so that the results can be directly compared among them. Negative values indicate when genes are expressed below the global mean of the gene expression obtained in the analysis, and positive values when genes are more highly expressed. To simplify the comparison, the graph starts at the negative values, being 0 the mean value of gene expression for the whole set of genes (87 in total) in each sample. **JAMM**- JAB1/MPN/MOV34 motif proteases; **MJD**- Machado-Joseph Disease protein domain proteases; **UCH**- Ubiquitin C-Terminal Hydrolases; **OTU**- Ovarian Tumor proteases; **USP**- Ubiquitin-Specific Proteases.

doi:10.1371/journal.pone.0150364.g001

In situ hybridization

For *in situ* hybridization (ISH), 16–18μm sections were recovered on commercial Superfrost Plus glass slides (Electron Microscopy Sciences, Hatfield, PA), dried 1 h at RT, rinsed three times for 10 min with phosphate-buffered saline (PBS), treated with 2 μg/ml proteinase K for 15 min at 37°C, washed twice for 5 min with PBS, and fixed with 4% PFA. Acetylation with 0.1 M triethanolamine-HCl (pH 8.0) containing first 0.25%, and then 0.5% acetic anhydride, was performed for 5 min each. Hybridization was carried out overnight at 55°C with digoxigenin-labelled riboprobes (2 μg/ml) in 50% formamide, 1 x Denhardt's solution, 10% dextran-sulfate, 0.9 M NaCl, 100 mMTris-HCl (pH 8.0), 5 mM EDTA (pH 8.0), 10 mM NaH_2PO_4, and 1 mg/ml yeast tRNA. For each gene, cDNA fragments generated by RT-PCR of approximately 400-700bp were subcloned into the pGEM-T® Easy Vector (Promega) and sense and antisense riboprobes were generated from the flanking T7 RNApol promoter. The name and sequence of all the primers used for RT-qPCR and *in situ* hybridization are listed in S1 Table.

After hybridization, the slides were washed in 2x SSC for 20 min at 55°C, equilibrated in NTE (0.5 M NaCl, 10 mM Tris-HCl pH 8.0, 5 mM EDTA) at 37°C, and then treated with 10 μg/ml RNase A in NTE at 37°C for 30 min. Subsequently, the sections were washed at 37°C in NTE for 15 min, twice in 2x SSC and 0.2x SSC for 15 min each, equilibrated in Buffer 1 (100 mM Tris-HCl pH 7.5, 150 mM NaCl), and blocked in Blocking Buffer (1% BSA and 0.1% Triton X-100 in buffer 1) for 1 h at RT. An anti-digoxigenin-AP conjugate antibody (1:1000; Roche Diagnostics, Indianapolis, IN) in Blocking Buffer was incubated overnight at 4°C. The sections were then washed twice in Buffer 1 for 15 min, once in Buffer 2 (100 mM Tris-HCl pH

9.5, 150 mM NaCl), and once in Buffer 2 supplemented with 50 mM $MgCl_2$ (5 min each) prior to adding the BM Purple AP Substrate (Roche Diagnostics, Indianapolis, IN). For each gene, antisense and sense ISH staining reactions were processed in parallel. The reaction was stopped in 1x PBS. Sections were cover-slipped with Fluoprep (Biomérieux, France) and photographed using a Leica DFC Camera connected to a Leica DM IL optic microscope (Leica Microsystems, Germany).

Fluorescent immunohistochemistry

For retina immunofluorescence, 16 μm sections were recovered on commercial Superfrost Plus glass slides (Electron Microscopy Sciences, Hatfield, PA), dried 30–45 min at RT, washed 10 min with PBS and blocked for 1 h with Blocking Buffer (2% Sheep Serum and 0.3% Triton X-100, in PBS 1x). Primary antibodies were incubated overnight at 4°C with Blocking Buffer. After incubation, slides were washed with PBS (3 x 10 min) and treated with DAPI (Roche Diagnostics, Indianapolis, IN) (1:300) and with secondary antibodies conjugated to either Alexa Fluor 488 or 561 (Life Technologies, Grand Island, NY) (1:300). After secondary antibody incubation slides were washed again in PBS (3 x 10min). Sections were mounted in Fluoprep and analyzed by confocal microscope (SP2, Leica Microsystems).

Primary antibodies and dilutions used were: 1:50 Rabbit anti-JOSD2 (Aviva Systems Biology); 1:50 Rabbit anti-JOSD3 (Aviva Systems Biology), 1: 50 Rabbit anti-ATXN3 (in house, a gift from Dr. S. Todi); 1:20 Rabbit anti-BAP1 (Abcam); 1:100 Rabbit anti-OTUD4 (Abcam ab106368), 1:100 Rabbit anti-PRPF8 (Abcam ab79237), 1:100 Rabbit anti-TNFAIP3 (Abcam ab74037), 1:100 Rabbit anti-UCHL3 (Abcam ab126703), 1:100 Rabbit anti-USP9X (Abcam ab19879), 1:100 Rabbit anti-USP13 (Abcam ab109264), 1:50 Rabbit anti-USP16 (Abcam ab135509), 1:100 Rabbit anti-USP22 (Abcam ab4812), 1:300 Rabbit anti-USP25 (in house), 1:250 Rabbit anti-USP28 (ABGEN AP2152b).1:500 for Mouse anti-Rhodopsin (Abcam, Cambridge, UK). Antibodies against AMSH (Biorbyt orb101007), JAB1 (Abcam ab12323), OTUB1 (Abcam ab76648), OTUD1 (Abcam ab122481), POH1 (Abcam ab8040), USP5 (Abcam ab154170) and USP45 (Novusbio H00085015) did not produce reproducible results.

Phylogenetic analyses

Protein sequences from each enzyme group were queried in complete genome sequences of 14 animal taxa (*Homo sapiens*, *Mus musculus*, *Danio rerio*, *Petromyzon marinus*, *Branchiostoma floridae*, *Saccoglossus kowalevskii*, *Strongylocentrotus purpuratus*, *Drosophila melanogaster*, *Daphnia pulex*, *Caenorhabditis elegans*, *Lottia gigantea*, *Capitella teleta*, *Nematostella vectensis* and *Acropora digitifera*) using the HMMER 3.1 algorithm. For each analyzed enzyme family (USP, UCH, OTU, MJD and JAMM) we searched all proteins containing the Hidden Markov motifs of their catalytic region as defined in Pfam (UCH/UCH_1, Peptidase_C12, OTU/Peptidase_C65, Josephin and JAB domains, respectively). Protein domain architectures of each retrieved protein were then computed using Pfamscan 1.5 and Pfam 27 database [25] of protein domains.

We aligned the catalytic region of each enzyme family using Mafft 7 L-INS-i [26](optimized for local sequence homology), and inspected each alignment matrix manually. The most suitable evolutionary model for the analyses, selected with ProtTest 3.4 [27], was LG+ Γ. We used RaxML 8.1.1 [28] to infer Maximum Likelihood trees of each family, with 100 bootstrap replicates as statistical supports. Complete sequences, alignments and phylogenies are provided in S1–S3 Files. Manual inspection of the trees allowed us to identify subfamilies, named after their human orthologs, based on their bootstrap support and conservation of protein domain architectures.

Results

Expression level of deubiquitinating enzymes in the mouse retina

A RT-qPCR was performed on mouse neuroretinas to assess the expression levels of the whole set of 87 mouse genes that encode the deubiquitinating enzymes belonging to the five aforementioned families (11 JAMM, 4 MJD, 15 OTU, 4 UCH, and 53 USP genes). Two reference genes, *Rhodopsin* and *Cerkl*, were included in the analysis due to their previously reported high and low levels of expression in the mouse retina, respectively [29]. The relative expression levels have been normalized to the expression of *Gapdh*, and the Z-score was calculated for the whole set of genes per each sample, so that they could be directly compared among them and between different samples. The results (mean and standard deviation of the Z-scores per each gene) are plotted in Fig 1, ordered by DUB family. A Z-score of zero indicates the mean value of expression for all the DUBs analyzed in the retina. Thus, genes with positive values have an expression above the mean, whereas genes with negative values show less expression than the mean (e.g. most USP genes).

The results showed that *Prpf8* was the highest expressed gene from the JAMM subfamily, followed by *Eif3h* and *Psmd7*. Both *Atxn3* and *Josd2* rendered the highest expression levels within the MJD subfamily. Concerning the OTU subfamily, *Otub1* and *Tnfaip3* produced the higher expression levels, followed by *Otud7b*, *Vcpip1*, *Otud4* and *Otud5*; whereas the levels of *Otud6a* were considered as negligible. *Uchl1* was the most highly expressed gene from the UCH family (and also with respect to all DUB genes), while *Uchl3* and *Uchl5* are lowly expressed in the retina. Finally, the genes from the large USP subfamily showed the lowest level of expression among all the DUB genes. Some USPs (20%) were highly expressed and showed positive Z-scores (*Usp5*, *Usp6*, *Usp10*, *Usp12*, *Usp19*, *Usp21*, *Usp22*, *Usp33*, *Usp47* and *Usp52)* whereas 25% of the USPs showed lower levels than the mean (*Usp8*, *Usp9Y*, *Usp17*, *Usp18*, *Usp26*, *Usp27*, *Usp29*, *Usp35*, *Usp43*, *Usp44*, *Usp45*,*Usp50*, and *Usp51*) (Fig 1).

Expression map of the DUBs in the mouse retina

Once the expression levels of all the DUB family members were assessed, we characterized and compared their expression pattern within the different layers of the mouse retina. We first decided to detect gene expression by mRNA localization using *in situ* hybridization (ISH) and then performed fluorescent immunohistochemistry of selected proteins.

For ISH, antisense (AS) riboprobes against a large group of DUBs were used on mouse retinal cryosections (Fig 2). As negative controls, the corresponding sense riboprobes (S) of each gene were generated and hybridized in parallel using the same conditions (see S1 Fig). The staining time was adjusted for each set of antisense/sense riboprobes so that a maximum signal was obtained in the antisense retinal sections with minimum background in the sense counterparts (for instance, *Prpf8* and *Tnfaip3* in situs stained in much less time than *Uchl5*, *Usp8* and *Usp18*, which required half a day). *Rhodopsin* was used as a positive control because of the reported high expression in the retina and its well-known localization in the inner segment of the photoreceptors. The large USP subfamily contains 57 members in the mouse genome but only a set of genes was considered for ISH. Representative ISH results are displayed in Fig 2. Our selection criteria included genes with relevant ocular phenotypes in systematic knockdown analyses of DUBs in *Drosophila* [9] and zebrafish [30].

Most DUBs are expressed ubiquitously throughout the layers of the murine retina, which would be compatible with a general role in the neuronal cell metabolism and regulation and thus, not restricted to particular retinal neurons. Nonetheless, specific patterns of expression were detected for particular DUBs. For instance, a strong hybridization signal in the plexiform

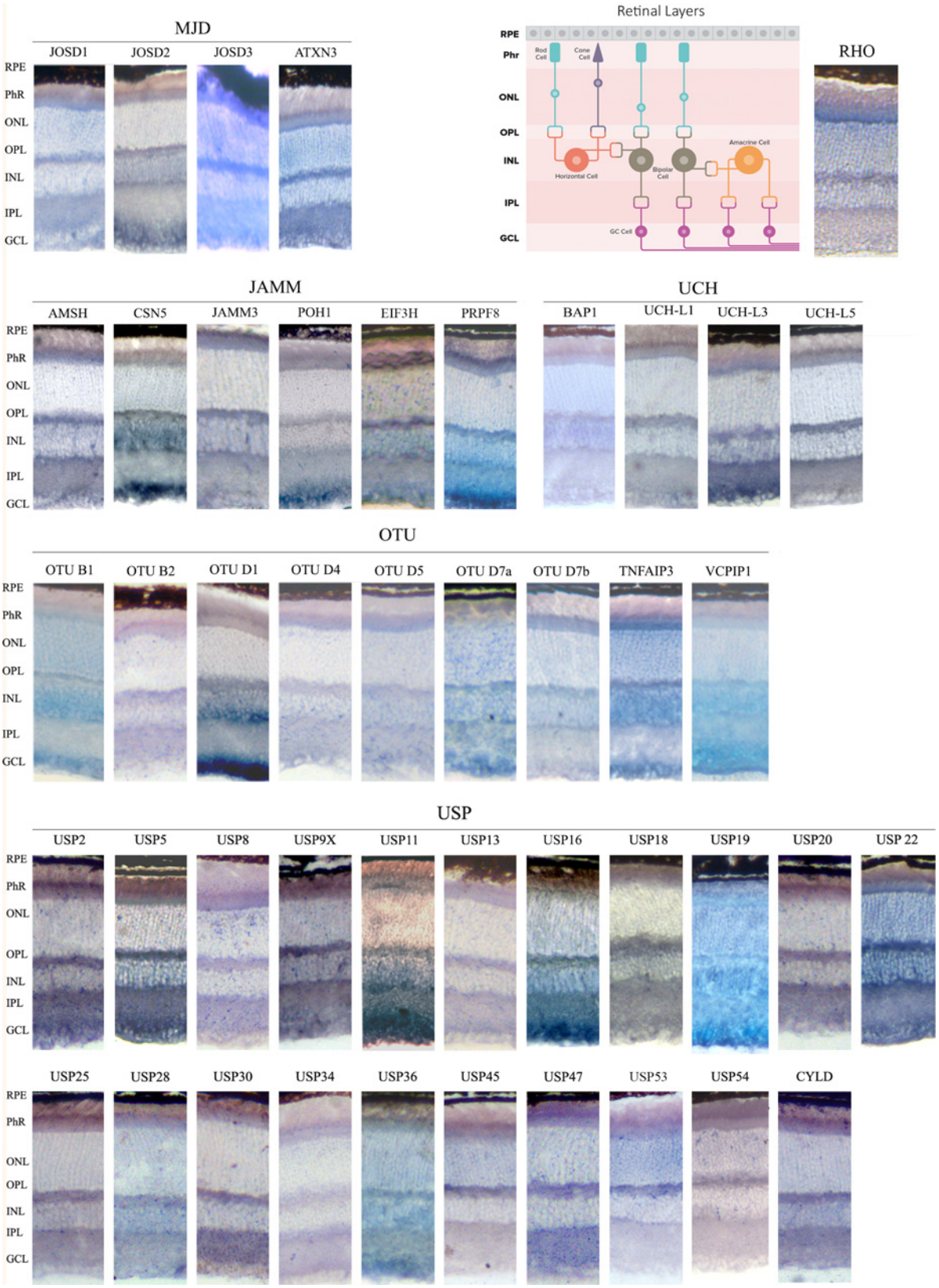

MJD
JOSD1
JOSD2
JOSD3
ATXN3
RPE
PhR
ONL
OPL
INL
IPL
GCL
Retinal Layers
Rod Cell
Cone Cell
Horizontal Cell
Bipolar Cell
Amacrine Cell
GC Cell
RHO
JAMM
AMSH
CSN5
JAMM3
POH1
EIF3H
PRPF8
UCH
BAP1
UCH-L1
UCH-L3
UCH-L5
OTU
OTU B1
OTU B2
OTU D1
OTU D4
OTU D5
OTU D7a
OTU D7b
TNFAIP3
VCPIP1
USP
USP2
USP5
USP8
USP9X
USP11
USP13
USP16
USP18
USP19
USP20
USP 22
USP25
USP28
USP30
USP34
USP36
USP45
USP47
USP53
USP54
CYLD

Fig 2. *In situ* hybridization of genes encoding DUB enzymes on retinal cryosections. Sections from wild-type C57BL/6J mouse retinas were hybridized using digoxigenin-labelled antisense riboprobes. Their corresponding sense riboprobes (negative controls) stained for the same length of time (lower panels in each row) are in the S1 Fig. The antisense *Rhodopsin* probe, which strongly labels the inner photoreceptor segment, was used as a positive control for the assay. **RPE**- Retinal pigmented epithelium; **Phr**- Photoreceptor cell layer; **ONL**- Outer nuclear layer; **OPL**. Outer plexiform layer; **INL**- Inner nuclear layer, **IPL**- Inner plexiform layer; **GCL**- Ganglion cell layer.

doi:10.1371/journal.pone.0150364.g002

layers was observed for *Uchl3*, *Uchl5*, *Usp2*, *Usp9X*, including in some cases the inner segment of the photoreceptor layer, as detected for *Amsh*, *Josd3*, *Atxn3* and *Usp47*. Some DUBs appear to be highly expressed in the GCL (*Csn5*, *Poh1*, *Prpf8*, *Josd2*, *Otud1*, *Vcpip1*, *Usp11*, *Usp5* and *Usp19*) in contrast to the pattern generated by *Usp8*, *Usp13*, *Usp30*, *Usp45* and *Usp54*, which yielded virtually no mRNA localization signal in the ganglion cells.

Several DUB genes of the USP family (*Usp5*, *Usp13*, *Usp19* and *Usp34*) were previously reported to be differentially expressed in the Retinal Pigmented Epithelium (RPE) by transcriptome analysis [31]. To assess their specific pattern of expression, and given that pigmented cells mask positive hybridization signals, we also performed ISH on albino retinas from CD-1 mice (S2 Fig). Although these four genes are expressed in this non-neuronal layer, their expression is not restricted to the RPE. In fact, *Usp5* and *Usp19* are very highly expressed throughout the retina (Fig 2). Comparison of the retinal expression pattern for these four genes did not show any detectable difference between C57BL/6J (wild-type black) and CD-1 (albino) mice strains.

Several genes, namely *Amsh-like*, *Brcc36*, *Jamm2*, *Mysm1* and *Psmd7* (JAMM group) and *Otud3*, *Yod1*, *Zranb1* (OTU group), did not render reproducible and reliable ISHs, even though several riboprobes spanning different gene regions were used. In most cases (e.g. *Amsh-like*, *Brcc36*, *Jamm2*, *Mysm1*, and *Otud3*) we obtained very low levels of expression and the signal was too faint to be distinguished from the negative control (sense riboprobe), or the sense and antisense riboprobes both produced signals of similar intensity. The ISH results of these genes are not included here.

Taking the ISH results together, we drew an atlas of expression for DUBs in the retina of adult mouse. In general, all analyzed genes except *Otud1* are expressed in the photoreceptors, and their mRNAs are localized in a wide range of intensities in the inner segment (perinuclearly) and the outer plexiform layer. Among layers, the GCL showed the most different pattern of gene expression. Notably, some DUBs, such as *Usp45*, *Usp53* and *Usp54*, are only detected in photoreceptors (PhR -inner segments, ONL (photoreceptor nuclei and perinuclei) and OPL (photoreceptor synapsis), whereas nearly no hybridization could be detected in the rest of retinal layers, which would suggest specific roles for these DUBs in this highly specialized photosensitive cells.

These ISH results prompted us to confirm and define more accurately protein localization within the retinal cell layers by fluorescent immunohistochemistry, since in cells with a highly specialized morphology, mRNA and protein localization might be different (e.g. the mRNA of rhodopsin is localized in the ribosome-rich photoreceptor inner segment whereas the protein is highly abundant in the membranous disks of the outer segment). We selected a group of DUBs for immunohistochemistry based on: i) particular ISH patterns, ii) relevance for eye phenotype in animal models, iii) putative functional diversification in phylogenetically closely related enzymes (see next section), and iv) antibody commercial availability and affinity. We selected 21 DUBs (the list of genes is detailed in the Material and Methods), of which 14 immunodetections rendered a reproducible and reliable signal (Fig 3 and S3 Fig).

Overall, the immunodetection confirms the ISH results since protein is detected in the same retinal cells than mRNA (Fig 3). Comparing RT-qPCR to ISH and immunohistochemistry results, high levels of retinal expression correlated with a ubiquitous expression pattern.

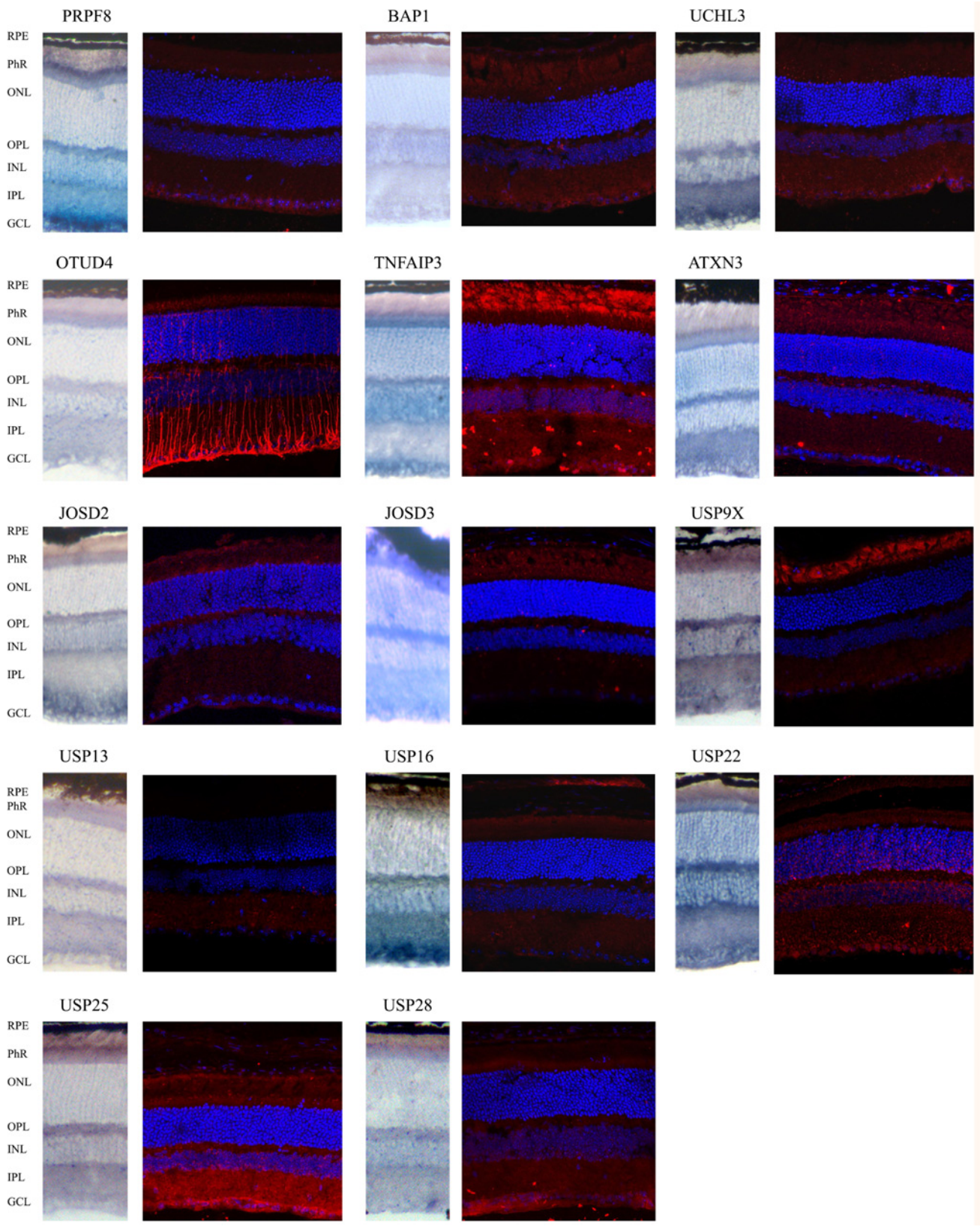

PRPF8
BAP1
UCHL3
OTUD4
TNFAIP3
ATXN3
JOSD2
JOSD3
USP9X
USP13
USP16
USP22
USP25
USP28
RPE
PhR
ONL
OPL
INL
IPL
GCL

Fig 3. Comparison of mRNA and protein immunodetection of selected DUBs in mouse retinal cryosections. Most analyzed genes render a consistent expression pattern when comparing mRNA and protein localization in the wild type mouse retina. The merge immunohistochemistry show DUBs immunodetected in red, and nuclei counter-staining with DAPI (in blue). Details in S3 Fig. **RPE**- Retinal pigmented epithelium; **Phr**- Photoreceptor cell layer; **ONL**- Outer nuclear layer; **OPL**. Outer plexiform layer; **INL**- Inner nuclear layer, **IPL**- Inner plexiform layer; **GCL**- Ganglion cell layer.

doi:10.1371/journal.pone.0150364.g003

Besides, some protein locations are worth mentioning as indicative of distinct functions in specific cellular compartments. For instance, OTUD4 is strongly detected in the axonal processes of bipolar and other retinal cells, supporting its involvement in neurodegeneration in human [32]. USP25 is mainly detected in the inner plexiform and ganglion cell layer; while USP9X and TNFAIP3 are particularly detected (but not exclusively) at the outer photoreceptor segment. Besides, USP22 is localized in the nucleus of ganglion cells, and perinuclearly in the rest of retinal neurons. For details, merge and separate immunodetection images, see S3 Fig.

DUB phylogenetic analysis, protein domain architecture and neuronal phenotypes

To provide a rational framework for gene expression patterns in extended families, it is crucial to have an understanding of the origin and phylogenetic closeness between the different DUB genes. Therefore, we performed a bioinformatic survey of DUB protein sequences across animal taxa. A recent phylogenetic analysis of the ubiquitin system across eukaryotes already showed that a massive expansion of ubiquitin ligases and proteases, which involves innovation and incorporation of new protein domains, occurred at the origin of animal multicellularity [24]. This was likely associated with the diversity of proteins and protein roles in different cell types. We here provide a comprehensive picture of the DUB families during the diversification of metazoans, related to previously described neuronal function, with an emphasis on eye and retinal phenotype.

Completely sequenced genomes from 14 species (from cnidarians to vertebrates) were queried with the catalytic region of each enzyme family (as defined in Pfam) in search of orthologs. Phylogenetic trees were generated using the retrieved sequences, and the statistical support for each node is also indicated (Fig 4A, 4B, 4C, 4D and 4E). For the sake of clarity, protein nomenclature is according to human DUBs. Highly similar sequences that expanded recently (during the pre-vertebrate/vertebrate expansion) and clustered together appear collapsed. The presence of an identified ortholog in each species/clade is represented with a black dot. Vertebrate species that present all the paralogs in a collapsed branch are circled in black. White dots mark the presence of homologs that could not be confidently assigned to a characterized DUB type, either because they are sister-group to various known DUB paralogs (and therefore represent the pre-duplication homolog), or because statistical support is too low to confidently cluster them with a specific ortholog. Question marks represent statistically supported clades that cannot be assigned to any known DUB (or group of paralogous DUBs). Protein motifs (as defined in Pfam) including the catalytic domain are drawn next to each branch to illustrate the diversity/conservation in protein architecture within each family. For detailed and complete phylogenetic trees, see S3 File.

Notably, the phylogenetic distribution of OTU DUBs reveals two different groups that appeared at the origin of eukaryotes OTUs with peptidase C65 domains (OTUB1 and OTUB2 in animals) and those with OTU domain [24] (Fig 4D). Given that i) these two catalytic domains diverged long before the origin of metazoans, ii) OTUB1/B2 protein domain architectures are clearly different from the other OTUs, iii) OTUB homologs are present in all metazoan clades, and iv) this split does not occur in any other family of DUBs, a new classification might be in order to acknowledge a new subfamily of DUBs.

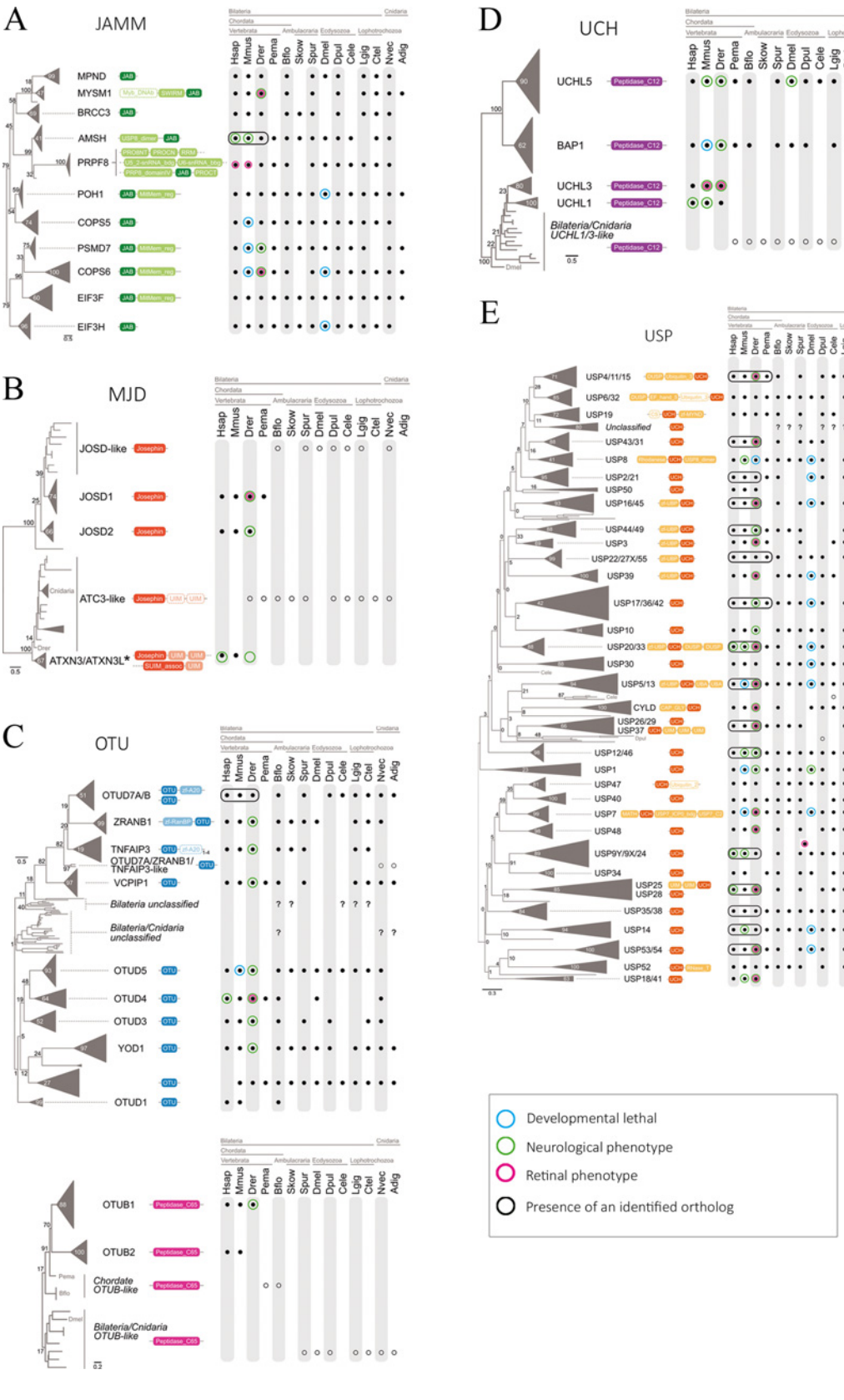

A
JAMM
MPND
MYSM1
BRCC3
AMSH
PRPF8
POH1
COPS5
PSMD7
COPS6
EIF3F
EIF3H
B
MJD
JOSD-like
JOSD1
JOSD2
ATC3-like
ATXN3/ATXN3L*
C
OTU
OTUD7A/B
ZRANB1
TNFAIP3
OTUD7A/ZRANB1/TNFAIP3-like
VCPIP1
Bilateria unclassified
Bilateria/Cnidaria unclassified
OTUD5
OTUD4
OTUD3
YOD1
OTUD1
OTUB1
OTUB2
Chordate OTUB-like
Bilateria/Cnidaria OTUB-like
D
UCH
UCHL5
BAP1
UCHL3
UCHL1
Bilateria/Cnidaria UCHL1/3-like
E
USP
USP4/11/15
USP6/32
USP19
Unclassified
USP43/31
USP8
USP2/21
USP50
USP16/45
USP44/49
USP3
USP22/27X/55
USP39
USP17/36/42
USP10
USP20/33
USP30
USP5/13
CYLD
USP26/29
USP37
USP12/46
USP1
USP47
USP40
USP7
USP48
USP9Y/9X/24
USP34
USP25
USP28
USP35/38
USP14
USP53/54
USP52
USP18/41
Hsap Mmus Drer Pema Bflo Skow Spur Dmel Dpul Cele Lgig Ctel Nvec Adig
Developmental lethal
Neurological phenotype
Retinal phenotype
Presence of an identified ortholog

Fig 4. Phylogenetic analysis of DUB genes and neuronal/retinal phenotype. Protein sequences from the catalytic region of each enzyme group were queried in complete genome sequences of 14 animal taxa and aligned. The protein domain architectures including the catalytic and accessory domain motifs are represented next to each DUB member (A, JAMM; B, MJD; C, OTU; D, UCH; and D, USP). Black dots indicate presence of the ortholog, whereas white dots indicate homologs that cannot be confidently assigned to a DUB type (see Results). Question marks represent statistically supported clades of uncharacterized DUBs. DUB sequences that are highly similar and cluster closely together appear collapsed under a common name. In general, invertebrates have a single representative member of the collapsed branch, whereas vertebrate genomes show one member of each paralog (species circled in black). *Acropora digitifera* USP homologs were excluded from the analysis as they impaired the resolution of the USP phylogeny. Genes reported to produce an abnormal neuronal phenotype when mutated are circled in magenta, whilst genes producing abnormal eye or retinal phenotype are circled in green. Genes whose mutation is lethal during developmental stages are circled in blue. An schematic summary of the DUB mRNA localization in the mouse retina (from ISH) is also presented next to the corresponding family. The intensity of the color indicates hybridization signal intensity. Retinal layers appear indicated as in Fig 2.

doi:10.1371/journal.pone.0150364.g004

The JAMM family has clear sequence assignment in all the analyzed animals, even though some species have secondarily lost some DUB members, e.g. *Acropora* (cnidarian), *C. elegans* (nematode), *Drosophila* (insect) *Saccoglossus* (hemichordate), and *Petromyzon* (sea lamprey, an early-branching vertebrate). These species also show specific gene loss for other DUB families, pointing to a divergent evolution in their lineages.

On the other hand, a clear expansion within each DUB family has occurred in the vertebrate lineage (Figs 4 and 5). When these duplicated members have rapidly diverged, the DUB protein sequences are in separate branches, but the common ancestry becomes evident since a single ancestral ortholog is present in the rest of clades (white dots in Fig 4). This is the case within the UCH (UCHL1 and UCHL3) and MJD families (JOSD1 and JOSD2). When the duplicated sequences have diverged but still branch closely together in the phylogenetic tree, the vertebrate paralogs have been collapsed into a single branch (black circles in Fig 4). This is particularly evident for USPs, where we can identify a single ancestral sequence in all invertebrate clades whereas several members are present in vertebrates (e.g. USP4/11/15. . .). Note that in the case of USP 18/41, a duplication event occurred only in the case of humans; as it is a single-species case, we have not included any black box on the figure. The *ATXN3* gene deserves specific mention, since its close paralog, *ATXN3L*, is a retrogene, that is, a gene generated by a very late retrotransposition event within the primate lineage.

The DUB gene expansion in animal phylogeny is visually summarized in the heat map of Fig 5. Color intensity reflects the number of genes per genome. It becomes evident that a burst of gene expansion within all DUB families was at the basis of the vertebrate lineage. Nonetheless, the innovation in the protein architectures with the acquisition of new domains accompanying the DUB catalytic signatures, pre-dates the origin of vertebrates in all the analyzed families, as vertebrate-like domain arrangements are often identified in other animal clades.

To complement our DUB expression study in the retina and in order to suggest relevant genes for hereditary visual disorders, we have compared the reported DUB mutant phenotypes of several animal models and human diseases, and viewed them under our new phylogenetic framework. We have specifically searched for early developmental lethality, neuronal phenotype and retinal alterations when available (Fig 4A, 4B, 4C, 4D and 4E). In the cases of neuronal phenotype, there is an accompanying alteration in the eye. However, most phenotypic assessment in the eye report only gross alterations, but a detailed retinal study has not yet been described for most animal models. For a detailed phenotypic trait list, see S2 Table and references therein.

In general, we observe that families with ancestral genes that have not been expanded in vertebrates (particularly the JAMMs) have a ubiquitous expression profile in the retina, suggesting a basic cell function. Moreover, mutations of these real orthologs produce consistent phenotypes through the analyzed taxa, arguing in favor of functional and evolutionary conservation. In contrast, for close paralog DUB genes arisen by duplication events in the vertebrate lineage,

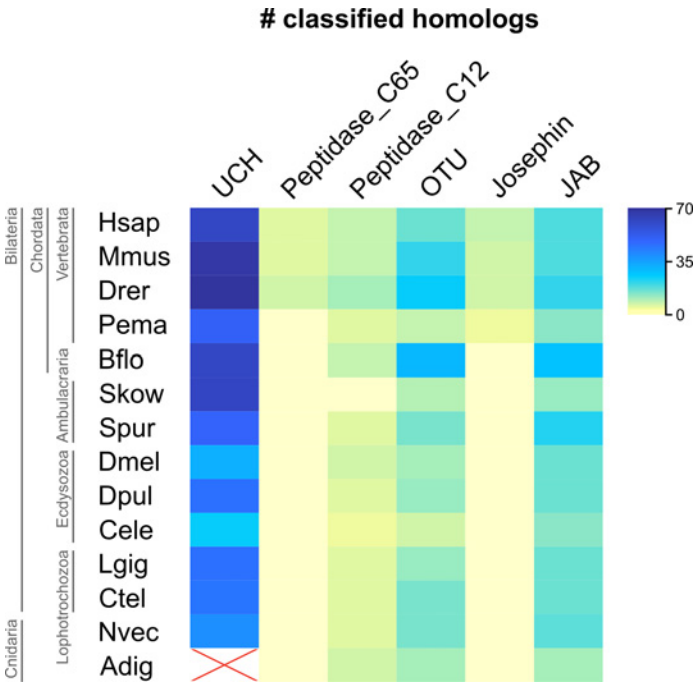


Fig 5. Counts of classified DUB homologs. Heatmap representing the number of classified genes in each analyzed genome. Increasing intensity reflects increasing number of genes. Only orthologs marked with black dots in Fig 4 are considered. *Acropora digitifera* USP homologs, excluded from the phylogenetic classification, are marked as not analyzed (NA).

doi:10.1371/journal.pone.0150364.g005

different patterns of retinal gene expression are often observed. A good example is OTUD7A/B (with one ancestral gene in most animals, and expanded in vertebrates), where OTUD7A is more highly expressed in the GCL and plexiform layers, whereas OTUD7B is more expressed in the photoreceptors. Similarly, UCHL3 and UCHL1 (both specific to vertebrates and associated to neuronal phenotypes) are expressed differently. Notably, UCHL3 (detected in the GCL and photoreceptors by ISH and immunodetection) produces eye specific retinal alterations, supporting subfunctionalization of these two paralogs. Other examples are included in the discussion.

Discussion

The ubiquitin-proteasome system (UPS) is currently viewed as one of the most dynamic and versatile cell regulators in eukaryotes. Perturbations of this system are known to be at the basis of many human disorders, particularly cancer and neurodegeneration [5,33]. Due to their ability to deconjugate ubiquitin, DUBs play a major regulatory role in the UPS. The disruption of DUB genes has dramatic consequences for the animal taxa analyzed, either during development or in adult stages, as shown by reports of the systematic DUB knockdown in zebrafish embryos and flies [9,30].

In mammals, several comprehensive surveys of DUBs have been reported resulting in: *in silico* inventories of the DUBs in the human genome [22,34]; identification of protein interactors by cell-based proteomics analysis [8]; studies of subcellular localization [1]; functional involvement in maintaining genome integrity in cells [35]. A recent review reported the expression levels of DUBs in human organs and the disease phenotypes associated to DUB mutations in humans and animal models [23]. Despite their importance, detailed expression and functional

analysis for most DUBs on particular tissues or organs, such as the retina, is still missing. We here aimed to fill this gap and produced a descriptive landscape of the expression of the complete set of DUBs in the mouse retina by combining mRNA and protein localization. We have also delineated a detailed evolutionary history of the different DUB families using phylogenetic analysis. We compared their protein domain architectures, and considered the neuronal and retinal phenotypes associated to each gene mutation/knockdown. We thus provide a reference framework for researchers interested in this visual tissue, either in physiological or in disease conditions, and suggest new avenues of research in DUBs as excellent candidates for retinal/visual hereditary disorders.

Differential levels of DUB gene expression in the adult mouse retina

Some genes that are barely expressed in the mouse retina (e.g. *Brcc36*, *Poh1*, *Bap1*, *Otub2*, and *Usp44*) are reported to be induced in replicative cells instead, being recruited to DNA damage sites where they regulate DNA repair and mitosis checkpoints [35]. These results are consistent with the fact that the adult retina is mostly formed by differentiated cells.

Among the genes highly expressed in the adult retina, *Uchl1*, *Atxn3*, *Otub1*, *Usp6*, *Usp22* and *Usp33* are also highly expressed in the brain [23]. In fact, *Uchl1*, *Otub1* and *Atxn3* are involved in neurodegenerative diseases in human, namely Parkinson's disease and cerebellar ataxia [6,36], thus indicating a relevant role in neurodegeneration. Our ISH results showed ubiquitous mRNA localization through all the retinal layers for these three genes, supporting a possible basal function in the retina. On the other hand, other DUB genes that are highly expressed in the brain [23], such as *Mysm1*, *Usp26*, *Usp29*, *Usp35* and *Usp51*, were barely expressed in the adult mouse retina; and genes that showed very low levels of expression when analyzed by qPCR within this work such as *Usp2*, *Usp25*, *Usp45*, *Usp53* and *Usp54* rendered eye phenotype when knocked-down in zebrafish [30]. Note that we performed RT-qPCR in whole adult neuroretinas at P60, and the role of these genes during development might be more relevant than in the adult stage. It is also worth noting that *Usp45*, *Usp53* and *Usp54* did show layer specificity, as they were mainly expressed in the photoreceptors (PhR inner segment, ONL and OPL), suggesting a specific role for these genes in photoreceptors and underscoring their role as potential candidates for visual disorders.

Immunohistochemical localizations also point to specific functions for some DUBs, e.g. OTUD4 is highly localized in axons; TNFAIP3 is highly expressed in the photoreceptor outer segment and GCL, and USP22 protein localization is mainly nuclear and perinuclear, thus suggesting that these genes may be good candidates for particular retinal phenotypes.

Phenotypic comparison of DUB mutants and gene expression profiles under the new evolutionary framework

Animal models have been generated by gene disruption (mouse) or knockdown (*Drosophila*, zebrafish) for some DUBs. When the DUB function is extremely relevant for cell cycle or cell differentiation, a lethal/early and extensive neuronal phenotype is consistently apparent in different organisms, as it is the case for most JAMMs and several USPs (see Fig 4 and S2 Table). In vertebrates, when some mutants show neuronal/brain affectation, a retinal/eye phenotype is also one of the accompanying phenotypic traits (examples are found in all the families). In fact, multiple vertebrate USP genes are present in paralogs (probably arising from the several rounds of genome duplication at the base of their linage), whereas their invertebrate relatives have a single homolog (black boxes in Fig 4). Therefore, it is not surprising that most USP knockdowns are lethal in *Drosophila* (where only a single member is present), whereas in vertebrates, the mutant phenotype mostly affect specific tissues, probably related to the larger

panoply of USP members and a higher functional diversification. For instance, in zebrafish the knockdown of *Usp33* (whose close relative homolog is *Usp20*) alters the nervous system development including the eye [9], which is consistent with a reported subcellular localization associated to microtubules and centrosomes; whereas the knockdown of the only member USP20/33 in *Drosophila* is lethal. Something very similar occurs with the knockdown of *Usp53* (whose close relative homolog is *Usp54*), which affects brain and eye development in zebrafish, whereas the knockdown of the single USP53/54 member is lethal in *Drosophila* (Fig 4B and S2 Table). For all the DUB families, orthologs share both high sequence similarities and consistent mutant phenotypes in vertebrates; overall, pointing to their functional conservation and supporting mouse and zebrafish models for assessing DUB roles in the human retina.

The knockdown phenotypes in different species are sometimes partially overlapping between neuronal and retinal alterations, probably due to subfunctionalization of different paralogs due to duplication events. For instance, *Usp5* and *Usp13* (encoding enzymes that expanded and diverged in the vertebrate lineage, and sharing 59.5% amino acid identities in human) showed a distinct pattern of expression in the mouse retina, with *Usp5* being highly expressed in the GCL in contrast to *Usp13*, which is barely expressed in this layer and the protein is mostly localized in the inner plexiform layer, thus indicating different roles despite sequence similarity. The knockdown of any of them severely alters zebrafish embryonic development and causes neurodegeneration (even though only the *Usp5* knockdown showed a clear eye phenotype), whereas in *Drosophila* the disruption of the single member *Usp5/13* alters eye development by increasing photoreceptor apoptosis, thus recapitulating neurodegeneration and retinal phenotype. Similarly, the close paralogs *Usp16* and *Usp45* have a contrasting expression pattern, with the former in GCL and plexiform layers, and the latter restricted to the photoreceptor cell layer, supporting again subfunctionalization or neofunctionalization of the vertebrate paralogs. Of note, the knockdown of Usp45 in zebrafish shows reduced eyes. Interestingly, *fat facets* (the ortholog of *Usp9X*, involved in endocytosis in the Notch pathway) limits the number of photoreceptors in *Drosophila* [37], while the human homolog *USP9X* has been involved in neurodegeneration, mental retardation, epilepsy and autism, as well as in cancer [38], but not yet in visual disorders. Nonetheless, the strong immunodetection in the outer segment of photoreceptors would indicate that it is also a good candidate for retinal dystrophies.

Finally, the only DUB-related gene that has been directly involved in human inherited retinal degeneration and causative of autosomal dominant Retinitis Pigmentosa is *PRPF8*, the JAMM-family member with the highest level of expression in the retina. Notably, PRPF8 (which is not properly a DUB since it is catalytically inactive) forms part of the splicing machinery [39]. Even though *PRPF8* is a housekeeping gene, its haploinsufficiency might cause a shift in the splicing patterns, which in turn alters the highly sensitive photoreceptors and triggers their apoptosis. Knock-in mice bearing human missense mutations also display retinal degeneration, thus strengthening the significance of this JAMM-gene in the retina [40].

Conclusions

In summary, our results show that data on the expression of the deubiquitinating enzyme gene family cannot be directly extrapolated between tissues or organs since cell requirements might be completely different, particularly in highly specialized and structured tissues, such as the retina. Therefore, in large families of seemingly redundant enzymes (such as DUBs) the integration of systematic expression maps together with a robust phylogenetic analysis and available phenotypic information provides an insightful reference framework for further functional characterization. This framework may be helpful for researchers working in the ubiquitin-

related field as well as for those working in the molecular bases of neurological and retinal disorders.

Supporting Information

S1 Fig. *In situ* hybridization of genes encoding DUB enzymes on mouse retina cryosections, with the comparison between antisense and sense riboprobes.
(PDF)

S2 Fig. *In situ* hybridization of genes encoding DUB enzymes on CD-1 (albino) mouse retina cryosections.
(PDF)

S3 Fig. Fluorescent immunohistochemistry of selected DUBs.
(PDF)

S1 File. Zip file containing the DUB catalytic domain sequences (per family) used for the phylogenetic analysis in FASTA format.
(ZIP)

S2 File. Zip file containing the sequence alignments obtained per each DUB family.
(ZIP)

S3 File. Zip file containing the complete phylogenetic trees with their corresponding bootstraps.
(ZIP)

S1 Table. Sequences of the primer pairs used in the reverse transcriptase Real Time qPCR and *in situ* hybridization.
(PDF)

S2 Table. Mutant neuronal and retinal phenotypes in different animal models and human listed per DUB family and gene.
(PDF)

Acknowledgments

We are grateful to D. Vystavělová and N. Peña-Auladell for technical support. We are also indebted to Dr. C. Arenas for advice on data statistical analysis. This study was supported by grants BFU2010-15656 (MICINN) and SAF2013-49069-C2-1-R (MINECO) to G.M., and 2014SGR-0932 (Generalitat de Catalunya) grant (BFU-2011-23434) from Ministerio de Economía y Competitividad (MINECO) and co-funded by the Fondo Europeo de Desarrollo regional (FEDER) to I.R-T. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceived and designed the experiments: AG GM. Performed the experiments: ME XG-B AG VT SG-M EM MJL-I VA-M. Analyzed the data: ME XG-B AG VT SG-M EM MJL-I VA-M IR-T GM. Contributed reagents/materials/analysis tools: GM XG-B IR-T. Wrote the paper: ME GM. Revised the final text: GM ME XG-B AG VT SG-M EM MJL-I VA-M IR-T.

References

1. Clague MJ, Coulson JM, Urbé S. Cellular functions of the DUBs. J Cell Sci. 2012; 125: 277–86. doi: 10.1242/jcs.090985 PMID: 22357969
2. Duncan LM, Piper S, Dodd RB, Saville MK, Sanderson CM, Luzio JP, et al. Lysine-63-linked ubiquitination is required for endolysosomal degradation of class I molecules. EMBO J. 2006; 25: 1635–1645. doi: 10.1038/sj.emboj.7601056 PMID: 16601694
3. Hunter T. The Age of Crosstalk: Phosphorylation, Ubiquitination, and Beyond. Mol Cell. 2007; 28: 730–738. doi: 10.1016/j.molcel.2007.11.019 PMID: 18082598
4. Clague MJ, Urbé S. Ubiquitin: same molecule, different degradation pathways. Cell. 2010; 143: 682–5. doi: 10.1016/j.cell.2010.11.012 PMID: 21111229
5. Dantuma NP, Bott LC. The ubiquitin-proteasome system in neurodegenerative diseases: precipitating factor, yet part of the solution. Front Mol Neurosci. 2014; 7: 70. doi: 10.3389/fnmol.2014.00070 PMID: 25132814
6. Ristic G, Tsou W-L, Todi S V. An optimal ubiquitin-proteasome pathway in the nervous system: the role of deubiquitinating enzymes. Front Mol Neurosci. 2014; 7: 72. doi: 10.3389/fnmol.2014.00072 PMID: 25191222
7. Kim TY, Siesser PF, Rossman KL, Goldfarb D, Mackinnon K, Yan F, et al. Substrate Trapping Proteomics Reveals Targets of the βTrCP2/FBXW11 Ubiquitin Ligase. Mol Cell Biol. 2014; doi: 10.1128/MCB.00857-14
8. Sowa ME, Bennett EJ, Gygi SP, Harper JW. Defining the human deubiquitinating enzyme interaction landscape. Cell. 2009; 138: 389–403. doi: 10.1016/j.cell.2009.04.042 PMID: 19615732
9. Tsou W-L, Sheedlo MJ, Morrow ME, Blount JR, McGregor KM, Das C, et al. Systematic analysis of the physiological importance of deubiquitinating enzymes. PLoS One. 2012; 7: e43112. doi: 10.1371/journal.pone.0043112 PMID: 22937016
10. Kang N, Won M, Rhee M, Ro H. Siah ubiquitin ligases modulate nodal signaling during zebrafish embryonic development. Mol Cells. 2014; 37: 389–98. doi: 10.14348/molcells.2014.0032 PMID: 24823357
11. Hoon M, Okawa H, Della Santina L, Wong ROL. Functional architecture of the retina: Development and disease. Prog Retin Eye Res. 2014; 42C: 44–84. doi: 10.1016/j.preteyeres.2014.06.003
12. Swaroop A, Kim D, Forrest D. Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. Nat Rev Neurosci. 2010; 11: 563–76. doi: 10.1038/nrn2880 PMID: 20648062
13. Schob C, Orth U, Gal A, Kindler S, Chakarova CF, Bhattacharya SS, et al. Mutations in TOPORS: a rare cause of autosomal dominant retinitis pigmentosa in continental Europe? Ophthalmic Genet. 2009; 30: 96–8. doi: 10.1080/13816810802695543 PMID: 19373681
14. Bowne SJ, Sullivan LS, Gire AI, Birch DG, Hughbanks-Wheaton D, Heckenlively JR, et al. Mutations in the TOPORS gene cause 1% of autosomal dominant retinitis pigmentosa. Mol Vis. 2008; 14: 922–7. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2391085&tool=pmcentrez&rendertype=abstract PMID: 18509552
15. De Sousa Dias M, Hernan I, Pascual B, Borràs E, Mañé B, Gamundi MJ, et al. Detection of novel mutations that cause autosomal dominant retinitis pigmentosa in candidate genes by long-range PCR amplification and next-generation sequencing. Mol Vis. 2013; 19: 654–64. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3611935&tool=pmcentrez&rendertype=abstract PMID: 23559859
16. Hugosson T, Friedman JS, Ponjavic V, Abrahamson M, Swaroop A, Andréasson S. Phenotype associated with mutation in the recently identified autosomal dominant retinitis pigmentosa KLHL7 gene. Arch Ophthalmol. 2010; 128: 772–8. doi: 10.1001/archophthalmol.2010.98 PMID: 20547956
17. Wen Y, Locke KG, Klein M, Bowne SJ, Sullivan LS, Ray JW, et al. Phenotypic characterization of 3 families with autosomal dominant retinitis pigmentosa due to mutations in KLHL7. Arch Ophthalmol. 2011; 129: 1475–82. doi: 10.1001/archophthalmol.2011.307 PMID: 22084217
18. Campello L, Esteve-Rudd J, Cuenca N, Martín-Nieto J. The ubiquitin-proteasome system in retinal health and disease. Mol Neurobiol. 2013; 47: 790–810. doi: 10.1007/s12035-012-8391-5 PMID: 23339020
19. Ramatenki V, Potlapally SR, Dumpati RK, Vadija R, Vuruputuri U. Homology modeling and virtual screening of ubiquitin conjugation enzyme E2A for designing a novel selective antagonist against cancer. J Recept Signal Transduct Res. 2014; 1–14. doi: 10.3109/10799893.2014.969375
20. Crosas B. Deubiquitinating enzyme inhibitors and their potential in cancer therapy. Curr Cancer Drug Targets. 2014; 14: 506–16. Available: http://www.ncbi.nlm.nih.gov/pubmed/25088039 PMID: 25088039

21. D'Arcy P, Brnjic S, Olofsson MH, Fryknäs M, Lindsten K, De Cesare M, et al. Inhibition of proteasome deubiquitinating activity as a new cancer therapy. Nat Med. 2011; 17: 1636–40. doi: 10.1038/nm.2536 PMID: 22057347

22. Nijman SMB, Luna-Vargas MP a, Velds A, Brummelkamp TR, Dirac AMG, Sixma TK, et al. A genomic and functional inventory of deubiquitinating enzymes. Cell. 2005; 123: 773–86. doi: 10.1016/j.cell.2005.11.007 PMID: 16325574

23. Clague MJ, Barsukov I, Coulson JM, Liu H, Rigden DJ, Urbé S. Deubiquitylases from genes to organism. Physiol Rev. 2013; 93: 1289–315. doi: 10.1152/physrev.00002.2013 PMID: 23899565

24. Grau-Bové X, Sebé-Pedrós A, Ruiz-Trillo I. The eukaryotic ancestor had a complex ubiquitin signaling system of archaeal origin. Mol Biol Evol. 2015; 32: 726–39. doi: 10.1093/molbev/msu334 PMID: 25525215

25. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42: D222–30. doi: 10.1093/nar/gkt1223 PMID: 24288371

26. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30: 772–80. doi: 10.1093/molbev/mst010 PMID: 23329690

27. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011; 27: 1164–5. doi: 10.1093/bioinformatics/btr088 PMID: 21335321

28. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–3. doi: 10.1093/bioinformatics/btu033 PMID: 24451623

29. Abad-Morales V, Domènech EB, Garanto A, Marfany G. mRNA expression analysis of the SUMO pathway genes in the adult mouse retina. Biol Open. 2015; doi: 10.1242/bio.201410645

30. Tse WKF, Eisenhaber B, Ho SHK, Ng Q, Eisenhaber F, Jiang Y-J. Genome-wide loss-of-function analysis of deubiquitylating enzymes for zebrafish development. BMC Genomics. 2009; 10: 637. doi: 10.1186/1471-2164-10-637 PMID: 20040115

31. Strunnikova N V, Maminishkis a, Barb JJ, Wang F, Zhi C, Sergeev Y, et al. Transcriptome analysis and molecular signature of human retinal pigment epithelium. Hum Mol Genet. 2010; 19: 2468–86. doi: 10.1093/hmg/ddq129 PMID: 20360305

32. Margolin DH, Kousi M, Chan Y-M, Lim ET, Schmahmann JD, Hadjivassiliou M, et al. Ataxia, dementia, and hypogonadotropism caused by disordered ubiquitination. N Engl J Med. 2013; 368: 1992–2003. doi: 10.1056/NEJMoa1215993 PMID: 23656588

33. Hussain S, Zhang Y, Galardy PJ. DUBs and cancer: the role of deubiquitinating enzymes as oncogenes, non-oncogenes and tumor suppressors. Cell Cycle. 2009; 8: 1688–97. Available: http://www.ncbi.nlm.nih.gov/pubmed/19448430 PMID: 19448430

34. Komander D, Clague MJ, Urbé S. Breaking the chains: structure and function of the deubiquitinases. Nat Rev Mol Cell Biol. 2009; 10: 550–63. doi: 10.1038/nrm2731 PMID: 19626045

35. Nishi R, Wijnhoven P, le Sage C, Tjeertes J, Galanty Y, Forment J V, et al. Systematic characterization of deubiquitylating enzymes for roles in maintaining genome integrity. Nat Cell Biol. Nature Publishing Group; 2014; 16: 1016–26, 1–8. doi: 10.1038/ncb3028 PMID: 25194926

36. Puschmann A. Monogenic Parkinson's disease and parkinsonism: clinical phenotypes and frequencies of known mutations. Parkinsonism Relat Disord. 2013; 19: 407–15. doi: 10.1016/j.parkreldis.2013.01.020 PMID: 23462481

37. Cadavid AL, Ginzel A, Fischer JA. The function of the Drosophila fat facets deubiquitinating enzyme in limiting photoreceptor cell number is intimately associated with endocytosis. Development. 2000; 127: 1727–36. Available: http://www.ncbi.nlm.nih.gov/pubmed/10725248 PMID: 10725248

38. Murtaza M, Jolly LA, Gecz J, Wood SA. La FAM fatale: USP9X in development and disease. Cell Mol Life Sci. 2015; doi: 10.1007/s00018-015-1851-0

39. Pena V, Liu S, Bujnicki JM, Lührmann R, Wahl MC. Structure of a multipartite protein-protein interaction domain in splicing factor prp8 and its link to retinitis pigmentosa. Mol Cell. 2007; 25: 615–24. doi: 10.1016/j.molcel.2007.01.023 PMID: 17317632

40. Graziotto JJ, Farkas MH, Bujakowska K, Deramaudt BM, Zhang Q, Nandrot EF, et al. Three gene-targeted mouse models of RNA splicing factor RP show late-onset RPE and retinal degeneration. Invest Ophthalmol Vis Sci. 2011; 52: 190–8. doi: 10.1167/iovs.10-5194 PMID: 20811066

8. Acknowledgements

Agraïments

If you don't have an extension cord I can get that too. Because we're friends! Right?

Randall Munroe, xkcd – Woodpecker, 2009

Per començar un doctorat cal tenir un pla, però és encara més divertit quan pots barrejar el pla amb la teva pròpia història. Això ho vull agrair a l'Iñaki, que em va oferir l'oportunitat de treballar amb ell i compartir la fascinació pel trenca-closques d'aquests anys: mirar fixament filogènies esmunyedisses i demanar-nos *què hi fa això aquí, si no serveix per res*? Però és més que això, perquè gràcies a ell he conegut també un grup humà fantàstic, una idea apassionada de la ciència, i tota la llibertat i els recursos que ens han permès els límits del temps i del possible. Gràcies: he crescut molt, i a sobre m'ho he passat bé.

Aquests límits els vaig aprendre precisament amb l'Arnau, l'Alex i el Guifré. Gràcies a la vostra consciència aguda del sentit de l'espectacle he gaudit molt de la tertúlia permanent sobre escatologia (em refereixo al *déluge universel*), narrativa creativa (el *happening* com a MacGuffin vital), ornitologia no sol·licitada (no he après res) i altra biologia en general (d'això sí que n'he après una mica). Sigui per l'espionatge, la cosificació o l'abús de la meva orella, tot plegat ha sigut molt divertit encara que sembli difícil de resumir sense caure, de bona fe, en la definició de la síndrome d'Estocolm. I també, encara que hi vaig poder coincidir menys temps, una abraçada a la Pons, el Javier i el Hiroshi, que representeu molt bé l'entorn científic totalment eclèctic on vaig acabar. Moltes gràcies per tot el que he après amb vosaltres.

Amb l'Helena, la Meri i el David hi he compartit una crisi generacional auto-imposada (que hem superat, oi?), i us agraeixo tot el que hem vist i acabat aprenent, a banda de l'humor terapèutic. I també la idea profundament marxista que «*mis manos son mi capital*», que, a diferència d'una muntanya, no es pot desmantellar. Però res s'atura i amb el temps he acabat vivint enmig d'un còctel d'alegria, entusiasme renovat i *pathos* en excés: Núria, Matija, Sebas, Gissela, Lule—heu sigut tants que ara temo deixar-me algú—i més tard Alicia, Javi «*El Breu*», Gema, Maria, Maria, Cristina—torna a agafar-me por de deixar-me gent—, Aleksandra, Alberto, les >1 estudiants d'estiu del David, Eduard, Andrej, Konstantina, Michelle i el darrer en arribar, Omaya. Després de la llarga llista, torno al començament: gràcies pel *pathos* renovat i l'entusiasme en excés, perquè així és una alegria compartir ciència.

També m'agradaria fer esment de tothom amb qui he compartit sostre aquests anys. Per descomptat, la primera casa i el Pau, la Berta i l'Alba a Sant Antoni. Per descomptat, les acollides de l'Arnau i la Cris al Clot i al desert. I per descomptat, l'Antònia i la segona casa a Sant Cugat (en més d'un sentit). I com que durant un temps vaig viure a una casa de Dickens a Exeter, això em fa pensar en el consol de no haver d'hagut de dormir mai al laboratori. Espero no ser l'únic. *Ecce homo!*

Exeter és una etapa breu però essencial d'aquests anys. Gràcies Tom pel refugi d'eremita on vaig començar a ordenar la meva vida intel·lectual, gràcies Guy per la paciència i les respostes, i gràcies també a la resta—Jonathan, Adam, Raquel, Jeremy, i tants altres—per la vostra acollida.

A banda del que és bo, també he col·leccionat enemics, alguns ficticis i alguns reals: *Il Padrino* dels oomicets, l'Institut d'Okinawa de Ciència i Tecnologia, els repartidors de llicències d'enzimologia que no saben que el sarcasme els apropa al diable, Parces i els talls d'electricitat. A vosaltres us dic: cap ressentiment.

***

Gràcies per tot a la meva família. Per començar, agrair el que ja no sóc a temps de dir al padrí Josep i l'àvia Carme. Lamento infinitament que no pugueu ser aquí ara, perquè no és possible arribar enlloc sense l'esforç i l'amor dels qui t'ensenyen el valor que tenen l'amor i l'esforç. Us trobo a faltar. I per continuar, als meus pares Xavi i Teresa, germans Josep i Carme, i a la tieta Carme. És difícil trobar un entorn allunyat de la meva feina mundana que a la vegada sigui tant i tant comprensiu i necessari per seguir-hi. Moltíssimes gràcies per tot: pel suport moral, físic, intel·lectual, per l'enllaç amb la terra, i per ser-hi sempre.

I a la Minchin, incondicional com sempre.

***

I per acabar, moltíssimes gràcies, Valèria, per tot. Per un cantó, pel teu suport i ajuda durant aquesta tesi i la teva precisió ètica i editorial. Però sobretot per les teves idees, per l'estímul constant, per l'alegria, per tot el que és important i que hem pogut fer junts durant aquests anys, i altre cop per tot. En vull més.