L’origen de la multicel·lularitat a metazous, una aproximació genòmica i funcional

The origin of metazoan multicellularity, a genomics and functional approach

Arnau Sebé Pedrós

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

L'origen de la multicel·lularitat a metazous, una aproximació genòmica i funcional

The origin of metazoan multicellularity, a genomics and functional approach

Memòria presentada per Arnau Sebé Pedrós per tal d'optar al títol de
Doctor per la Universitat de Barcelona

Dr. Iñaki Ruiz-Trillo	**Dr. Jaume Baguñà Monjo**	**Arnau Sebé Pedrós**
Director de la tesi	Director de la tesi i tutor	Autor

Barcelona, Març de 2013

Un altre cop vols agitar les aigües
del llac.
Està bé, però pensa
que no serveix de res tirar una sola pedra,
que has d'estar aquí des de la matinada
fins a la posta, des que neix la nit
fins al llevant
–tindràs la companyia
de les estrelles, podràs veure l'ocellassa
de la nit negra covant l'ou de la llum
del dia nou–,
assajant sempre cercles,
per si al cap de molts anys, tota una vida, et sembla
–i mai potser no n'estaràs segur–
que has assolit el cercle convincent.

Joan Vinyoli, *Cercles*
Antologia poètica
Edicions 62, 1999

Comprenc que, si per la ciència puc captar els fenòmens i enumerar-los, no puc aprehendre el món a través d'ella. Després d'haver seguit amb el tacte del dit el relleu d'aquest món tot sencer, no en sabria pas més que abans. I em doneu a triar entre una descripció que és certa, però que no m'ensenya res, i unes hipòtesis que pretenen ensenyar-me, però que no són certes.

Albert Camus, *El mite de Sísif*
Edicions 62, 2009

Table of contents

1. INTRODUCTION 3
1.1 Evolving multicellularity 3
1.2 Selective advantages and challenges 7
1.3 The multiple origins of multicellularity 9
1.4 The origin of Metazoa 13
1.5 The origin of metazoan genetic developmental programs 17
1.6 The unicellular relatives of Metazoa 18
1.7 Comparative genomics and the Urmetazoan genome 24
1.8 *Capsaspora owczarzaki* and its genome 26

2. OBJECTIVES 31

3. RESULTS 35
Informe dels directors sobre els articles publicats 35
3.1 Results R1: Ancient origin of the integrin-mediated adhesion and signalling machinery 39
3.2 Results R2: Integrin-mediated adhesion complex. Co-option of signalling systems at the dawn of Metazoa 55
3.3 Results R3: Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsapsora owczarzaki* 61
3.4 Results R4: Early evolution of the T-box transcription factor family 99
3.5 Results R5: Premetazoan Origin of the Hippo Signaling Pathway 117
3.6 Results R6: Insights into the origin of metazoan filopodia and microvilli 131
3.7 Results R7: Transcriptome remodelling during aggregative multicellularity in a close unicellular relative of Metazoa 155

4. DISCUSSION 175
4.1 A new view of *Capsaspora owczarzaki* life cycle 175
4.2 The origin of the metazoan multicellularity gene repertoire 177
4.2.1 Adhesion 177
4.2.2 Transcription factors 179
4.2.3 Signaling 181
4.3 Genetic sources of innovation in Metazoa 182
4.3.1 Molecular exaptation 182
4.3.2 *De novo* gene origin 183
4.3.3 Domain shuffling 184
4.3.4 Gene family expansion 185
4.3.5 Alternative splicing 187

4.3.6 New physical and regulatory interactions 189
4.4 The origin of Metazoa: from phyletic to spatial cell type distribution 190

5. CONCLUSIONS 197

6. REFERENCES 201

7. RESUM EN CATALÀ 217

FIGURES and TABLES:

Figure 1. Evolutionary origin of an altruistic gene 6
Figure 2. Predator-induced multicellularity in *Chlorella vulgaris* 7
Figure 3. Geochemistry and the evolution of eukaryotes 10
Figure 4. Timeline of the origins of multicellular eukaryotic clades 10
Figure 5. Phylogenetic distribution of multicellularity among eukaryotes 11
Figure 6. Origin of the major metazoan phyla 14
Figure 7. The geologic record of early animal evolution 15
Figure 8. Reconstruction of a putative Cryogenian demosponge 16
Figure 9. Phylogenetic relationships of Opisthokonta 20
Figure 10. The unicellular relatives of Metazoa 22
Figure 11. Reconstructing the ancestral Urmetazoan genome 25
Figure 12. Scientific literature about unicellular holozoans 26
Figure 13. *C.owczarzaki* actin and tubulin structures 175
Figure 14. *C.owczarzaki* ultrastructure 176
Figure 15. Metazoan cell junction types evolution 177
Figure 16. Genetic mechanisms of innovation in the transition to metazoan multicellularity 182
Figure 17. Domain shuffling and the origin of metazoan signaling receptors 185
Figure 18. Transcription factor evolution in eukaryotes 186
Figure 19. Intron length and intron density distribuition across eukaryotes 187
Figure 20. Alternative splicing modes across eukaryotes 188

Table 1. Genome statistic of *Capsaspora owczarzaki* and other eukaryotes 28
Table 2. Cell structures and behaviours in Opisthokonta 191

Un altre cop vols agitar les aigües
del llac.
Està bé, però pensa
que no serveix de res tirar una sola pedra,
que has d'estar aquí des de la matinada
fins a la posta, des que neix la nit
fins al llevant
–tindràs la companyia
de les estrelles, podràs veure l'ocellassa
de la nit negra covant l'ou de la llum
del dia nou–,
        assajant sempre cercles,
per si al cap de molts anys, tota una vida, et
sembla
–i mai potser no n'estaràs segur–
que has assolit el cercle convincent.

Joan Vinyoli, *Cercles*
Antologia poètica
Edicions 62, 1999

Comprenc que, si per la ciència puc captar els fenòmens i enumerar-los, no puc aprehendre el món a través d'ella. Després d'haver seguit amb el tacte del dit el relleu d'aquest món tot sencer, no en sabria pas més que abans. I em doneu a triar entre una descripció que és certa, però que no m'ensenya res, i unes hipòtesis que pretenen ensenyar-me, però que no són certes.

Albert Camus, *El mite de Sísif*
Edicions 62, 2009

Table of contents

1. INTRODUCTION 3

1.1 Evolving multicellularity 3

1.2 Selective advantages and challenges 7

1.3 The multiple origins of multicellularity 9

1.4 The origin of Metazoa 13

1.5 The origin of metazoan genetic developmental programs 17

1.6 The unicellular relatives of Metazoa 18

1.7 Comparative genomics and the Urmetazoan genome 24

1.8 *Capsaspora owczarzaki* and its genome 26

2. OBJECTIVES 31

3. RESULTS 35

Informe dels directors sobre els articles publicats 35

3.1 Results R1: Ancient origin of the integrin-mediated adhesion and signalling machinery 39

3.2 Results R2: Integrin-mediated adhesion complex. Co-option of signalling systems at the dawn of Metazoa 55

3.3 Results R3: Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsapsora owczarzaki* 61

3.4 Results R4: Early evolution of the T-box transcription factor family 99

3.5 Results R5: Premetazoan Origin of the Hippo Signaling Pathway 117

3.6 Results R6: Insights into the origin of metazoan filopodia and microvilli 131

3.7 Results R7: Transcriptome remodelling during aggregative multicellularity in a close unicellular relative of Metazoa 155

4. DISCUSSION 175

4.1 A new view of *Capsaspora owczarzaki* life cycle 175

4.2 The origin of the metazoan multicellularity gene repertoire 177

4.2.1 Adhesion 177

4.2.2 Transcription factors 179

4.2.3 Signaling 181

4.3 Genetic sources of innovation in Metazoa 182

4.3.1 Molecular exaptation 182

4.3.2 *De novo* gene origin 183

4.3.3 Domain shuffling 184

4.3.4 Gene family expansion 185
4.3.5 Alternative splicing 187
4.3.6 New physical and regulatory interactions 189
4.4 The origin of Metazoa: from phyletic to spatial cell type distribution 190

5. CONCLUSIONS 197

6. REFERENCES 201

7. RESUM EN CATALÀ 217

FIGURES and TABLES:

Figure 1. Evolutionary origin of an altruistic gene 6
Figure 2. Predator-induced multicellularity in *Chlorella vulgaris* 7
Figure 3. Geochemistry and the evolution of eukaryotes 10
Figure 4. Timeline of the origins of multicellular eukaryotic clades 10
Figure 5. Phylogenetic distribution of multicellularity among eukaryotes 11
Figure 6. Origin of the major metazoan phyla 14
Figure 7. The geologic record of early animal evolution 15
Figure 8. Reconstruction of a putative Cryogenian demosponge 16
Figure 9. Phylogenetic relationships of Opisthokonta 20
Figure 10. The unicellular relatives of Metazoa 22
Figure 11. Reconstructing the ancestral Urmetazoan genome 25
Figure 12. Scientific literature about unicellular holozoans 26
Figure 13. *C.owczarzaki* actin and tubulin structures 175
Figure 14. *C.owczarzaki* ultrastructure 176
Figure 15. Metazoan cell junction types evolution 177
Figure 16. Genetic mechanisms of innovation in the transition to metazoan multicellularity 182
Figure 17. Domain shuffling and the origin of metazoan signaling receptors 185
Figure 18. Transcription factor evolution in eukaryotes 186
Figure 19. Intron length and intron density distribuition across eukaryotes 187
Figure 20. Alternative splicing modes across eukaryotes 188

Table 1. Genome statistic of *Capsaspora owczarzaki* and other eukaryotes 28
Table 2. Cell structures and behaviours in Opisthokonta 191

Introduction

El nostre estudiant rumiava sobre el fenomen de les colònies de cèl·lules, va assabentar-se de semiorganismes, d'algues amb cèl·lules úniques, […] que tanmateix eren formacions pluricel·lulars; però si haguessin estat preguntades, no haurien sabut dir si volien ser considerades com a colònia d'individus unicel·lulars o com un ens unitari, i en a la seva autodefinició haurien dubtat, molt estranyades, entre el jo i el nosaltres.

Thomas Mann, *La muntanya màgica*
Edicions 62, 2007

1.1 Evolving multicellularity

The capacity of cells to act as integrated multicellular individuals is a major force that has structured the planet biosphere; especially during the Phanerozoic, since plants, animals and fungi appeared. The transition to multicellularity entails not only some degree of cell-cell adhesion and communication, but also a fundamental change in the nature of individuality. No longer each cell has an independent identity, but serves instead as part of a larger and more inclusive individual, defined by a suite of emergent traits.

But not all "multicellularities" are the same. First, we must differentiate between clonal/unitary and aggregative multicellularity. In the former, all cells in the group derive from an initial founder that undergoes successive rounds of cell division; resulting in a genetically identical population of cells. In the later, genetically distinct cells bind to each other (Grosberg and Strathmann 2007). In aggregative multicellularity, intraorganismal competition poses strong fitness challenges and, therefore, the aggregate is predicted to be evolutionary unstable (Aanen et al. 2008; Newman 2012). Indeed, in the eukaryotic lineages that have aggregative multicellularity, only transient cell aggregates are formed, never stable multicellular entities. This is the case of dictyostelids (Amoebozoa) (Schaap 2011), acrasid amoebas (Heterolobosea, Discicristata, Discoba) (Brown et al. 2011; Adl et al. 2012), *Guttulinopsis vulgaris* (Cercozoa, Rhizaria) (Brown et al. 2012), the genus *Sorogena* (Ciliata, Alveolata) (Lasek-Nesselquist and Katz 2001), the nucleariid *Fonticula alba* (Holomycota, Opisthokonta) (Brown et al. 2009) and the genus *Sorodiplophrys* (Labyrinthulomycetes, Heterokonta) (Dykstra and Olive 1975). In prokaryotes, aggregative multicellularity is found in Myxobacteria (Velicer and Vos 2009). From here on in this first section we will be referring to clonal multicellularity.

Defined only as clonally dividing cells that remain attached, multicellularity can be applied to filaments, clusters, balls or sheets of cells that arise via mitotic division from a single cell progenitor. Differentiation of somatic and reproductive cells is common (for example, in volvocine algae (Prochnik et al. 2010)), but more complex patterns of differentiation are not (Knoll 2011). This simple multicellularity is even found in Bacteria. For example, many cyanobacteria form filaments containing dozens, if not hundreds, of cells and some differentiate multiple cell types, including N-fixing

heterocysts and resistance akinetes (Bonner 1998; Tomitani et al. 2006). Other filament-forming bacterial lineages are the photosynthetic green non-sulfur bacteria (e.g., Chlorflexus), the large sulfur-oxidizing proteobacteria (e.g., Beggiatoa, Thioploca), some magnetotactic bacteria (Keim et al. 2004), and a remarkably diverse range of actinobacteria (e.g., Streptomyces). However, and despite their enormous ecological impact (e.g., in the form of biofilms and microbial mats), multicellular prokaryotes represent an evolutionary dead-end: large organisms and complex multicellularity are, with only trivial exceptions, the exclusive preserve of eukaryotes (Butterfield 2009).

There is, therefore, a significant "grey zone" in the definition of multicellularity, represented by more cryptic levels of intercellular integration. In contrast, complex multicellular organisms show, not only evidence of cell-cell adhesion, but also intercellular communication and, typically, tissue differentiation mediated by networks of regulatory genes (what is known as a stereotyped developmental program). This criterion applies strictly only to Embryophyta, Metazoa, Phaeophyta (brown algae, all classes), Bangiales and Florideophycidae red algae (many other rodophyte classes are unicellular), and some fungi (but, in this case, lacking individually partitioned cells and without a clear embryonic development; there are many secondarily unicellular species and simple filamentous species)(see below).

From a theoretical point of view, the transition to complex clonal multicellularity involved a hierarchical shift in Darwinian individuality (defined as when populations of these individuals display variation, heritability and reproduction). During this change, fitness is allocated from lower-level units, which relinquish their capacity to reproduce as independent units, into a higher level unit, that comes to reproduce as a larger whole (Michod and Herron 2006; Michod 2007; Rainey and Kerr 2010). Multi-level selection theory has provided the conceptual framework for understanding this transition by dissecting it into small analysable steps or stages. These stages include (Okasha 2006; Michod 2007; Rainey and Kerr 2010):

- Initial advantage of group formation. Undifferentiated groups of cells can arise under appropriate ecological conditions, cooperating to produce higher number of individual cells in the next generation. Here the group fitness is nothing more than the average or sum of the fitness of the individual cells that comprise the groups and, typically, selection at the group level affects only the few traits that allow group

formation (e.g., the secretion of an adhesive substance between the cells). This stage is readily explained by kin selection and traditional group selection theory.

- Origin of reproductive altruism within the group. At this step, fitness at the higher level is decoupled from fitness at the lower level. This is the big leap and the most problematic to understand. Basically two theories try to explain the evolution of collective reproduction (Rainey and Kerr 2010):
    - The first one proposes that the first propagules would come from cheats that appear within the group, gaining by mutation high reproductive rates and loosing cooperative behaviour. The group would collapse due to the action of these cheats and cells (basically cheats, since they grow faster) would disperse. Then, another mutation would revert the cheat phenotype to the collaborative phenotype, closing the cycle and producing again a cooperative group. This would be an incipient multicellular live cycle; but this model carries the burdensome of being interrupted by mutation at every stage of the cycle.
    - Alternatively, the uninterrupted model proposes that cell lineages would simply, by mutation, altruistically remove themselves from the reproductive line to generate some somatic benefit to the organism (Queller 2000). This model needs only a one-way mutation: to dead-end helper cells. One such cycle is easier to imagine if the altruistical mutation that removes a cell from the germ line is expressed conditionally, either in space and/or in time; this conditionality directly relates with our next point, the advent of a developmental program. An interesting paper by Nedelcu and Michod (2006) showed how reproductive altruism could arise from the co-option of an existing life-history gene (Figure 1): in the unicellular green algae *Chlamydomonas reindhartii*, an environmental cue (in this case light) activates a gene called regA (a SAND-domain transcription factor) that induces a change that maximises cell survival while detracting reproduction. The ortholog of this gene in the colonial green algae *Volvox carteri* works under developmental control to specify somatic cells, which contribute to colony survival and do not reproduce.

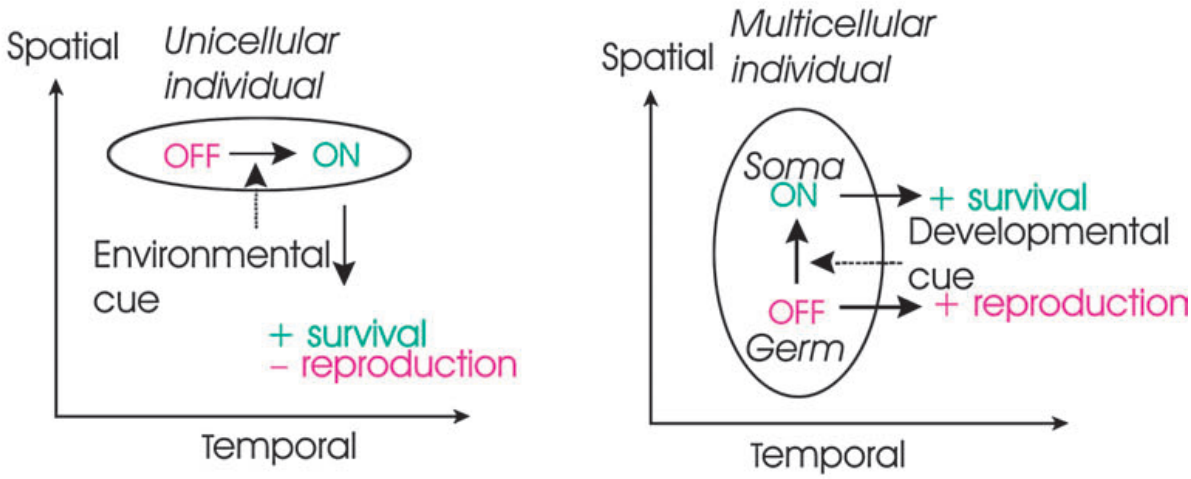


Figure 1. Evolutionary origin of an altruistic gene. Schematic representation of the change in expression pattern from a temporal context (environmentally induced) into a spatial context (developmentally induced) of a life-history gene co-opted as a developmental gene. From Nedelcu and Michod (2006).

- Cell differentiation. Conditional germ-line formation is the first of a succession of cell types and behaviours that can evolve once the multicellular individuality is established, especially as group size increases.

Ultimately, the process leads to the establishment of a clonal multicellular life cycle where the multicellular organism passes through a single-cell stage bottleneck from which, later on, an adult is formed through a process of embryonic development. This process ensures that each generation begins with a group of cells that shares all of their genes by descent. This is predicted to render them more evolvable (Wolpert and Szathmary 2002), as the variation among cells that arises within the parental life cycle is partitioned among offspring, rather than continuing within offspring. A prediction of this model is that, at least initially, these organisms would be small, as more cell divisions mean more occasions for mutations to occur and more division cycles in which such mutants can express their selfish replicative advantage (Queller 2000). Another reason to support this notion is that resources like oxygen and nutrients have to arrive to all cells in the three-dimensional group by simple diffusion. This fact establishes an upper size limit for a compact mass of cells (Knoll 2011).

1.2 Selective advantages and challenges

Several selective advantages of multicellularity have been hypothesised. First, multicellularity is an efficient way of incrementing size. This can also be achieved by hypertrophic cell growth, but there are physico-chemical limitations (surface-volume ratio, diffusion rates in the cytoplasm, etc.) that impose an upper limit to the size of a single cell (although some exceptionally large protists exist, like the 3-cm deep-sea testate amoeba *Gromia sphaerica* (Cercozoa, Rhizaria) (Matz et al. 2008)). Increased size enables escaping from heterotrophic predators. This was demonstrated in a classical experiment (Boraas et al. 1998) in which the unicellular green algae *Chlorella vulgaris* (Chlorophyta, Archaeplastida) was cultured with the predatory flagellate *Ochromonas vallescia* (Chrysopyceae, Heterokonta). In a few days, the algae started to form colonies by incomplete cell division (Figure 2). Initially, colonies were big, but this hindered cell nutrient uptake. So, finally, colony sized was stabilised in eight cells, enough to avoid being ingested by the flagellate while keeping an efficient nutrient absorption. After removing the predator, unicellular algae became predominant again. Therefore, the mere addition of a predator posed a selective pressure that efficiently selected colony-forming genotypes. The origin of heterotrophic eukaryotes (800 Ma) may have triggered the evolution of simple colonial forms in many eukaryotic lineages, some of which later evolved into more complex multicellularity.

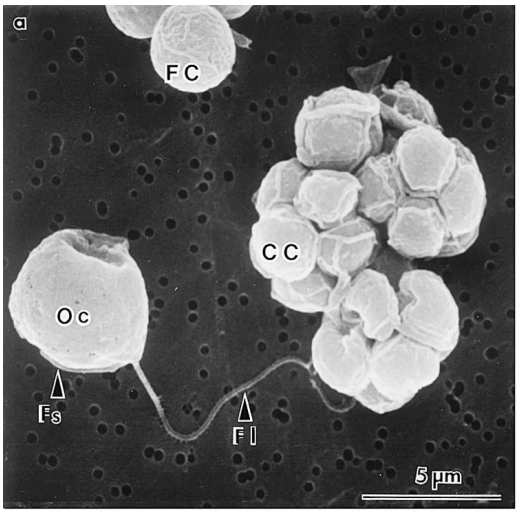


Figure 2. Predator-induced multicellularity in *Chlorella vulgaris*. The flagellate predator (Oc), with long (Fl) and short (Fs) flagella, a single Chlorella cell (FC) and a Chlorella colony (CC) sampled together from a culture 240 days after inoculation. From Boraas et al. 1998.

Second, multicellularity resolves metabolic and cellular trade-offs. For example, flagellar motility and mitosis compete for the microtubule-organizing centre (MTOC) (King 2004), which is used both to form the basal body that synthetises the flagella and to form the mitotic spindle used in chromosome segregation. This is why flagellated metazoan cells (including sperm, statocysts, etc.) never divide. By becoming multicellular both activities are made compatible, as some cells can divide while other

provide flagellar locomotion to the whole group. Another example is found in cyanobacteria, where photosynthesis and nitrogen fixation are biochemically incompatible and the two activities are split into different cell types.

In terrestrial habitats, spore dispersion is challenged by the absence of water currents. Some terrestrial eukaryotes, like dictyostelids or acrasids, aggregate to form dispersal fruiting bodies (Bonner 1998). Another advantage of multicellularity is a more efficient feeding, for example when secreting digesting enzymes (Grosberg and Strathmann 2007).

Finally, the neutralist view suggests that there is no need to invoke adaptive explanations because multicellularity, as any other form of complexity, evolves simply because "it can". It is just the result of irreversible accumulation of neutral or slightly deleterious mutations in small effective population sizes (Koonin 2011). The theory argues that these complex forms are not generally fitter than simpler forms (although it concedes that in some cases they can facilitate adaptation to new niches) and that these mutational neutral ratchets create an apparent directionality in evolution. Once multicellularity is established, the only selective pressure is to evolve mechanisms to avoid malfunction, such as programmed cell death or avoidance of uncontrolled cell proliferation (Koonin 2011).

Whatever evolutionary explanations for its origin, there are a several functions that complex multicellular organisms must perform:

- Cell adhesion: this can be achieved through extracellular cement substances, like pectins and hemicelluloses in plants or glycoprotein-based glues in fungi; or through specific transmembrane proteins that mediate cell-cell contacts, as is the case of cadherins and other adhesion proteins in animals.
- Cell-cell signaling: mechanisms that coordinate cells and transmit signals that, ultimately, modify specific cell behaviours by triggering differentiation programs. Plant hormones and animal developmental pathways like Notch, Hedgehog or Wnt are examples of such mechanisms. These systems are composed, alternatively, of a diffusible ligand and a transmembrane surface receptor (e.g., the Hedgehog and Wnt pathways in animals or the Brassinosteroid pathway in plants); or of two transmembrane receptors contacting each other (e.g., the Notch pathway in animals); or of a membrane-permeable ligand and an intracellular receptor (e.g., the

Auxin and Gibberellin pathways in plants and the steroid signaling in animals). In all cases, an intracellular signal transduction cascade follows.

- Cell differentiation: gene regulatory networks are established to control the proper proliferation of the cells (avoiding the appearance of non-cooperating cheaters) in the group. They also allow deploying specific cell behaviours in a spatial manner (in contrast with the temporal changes in cell type during a unicellular life cycle). This is achieved through transcription factors that control the expression of specific gene sets.

1.3 The multiple origins of multicellularity

Eukaryotes originated c. 2100 Ma (Anbar and Knoll 2002), being largely autotrophic at the beginning. Only later, around 800 Ma, we find evidence of heterotrophic eukaryotes: fossilised tests ascribed to lobose amoeba and other protists (Porter et al. 2003). In fact, around 800 Ma, there was marked increased in protist diversity (Knoll et al. 2006), a diversification that is hypothesised to be linked to geochemical changes (Figure 3). Although atmosphere was enriched in oxygen due to the great oxidation event linked with the origin of photosynthesis, during much of the Proterozoic Eon (2500-543 Ma) oceans were enriched in sulphide (Anbar and Knoll 2002), which is generally toxic to eukaryotes. It was approximately 800 Ma when sulfidic water masses began to recede, replaced by oxic waters (Figure 5)(Donoghue and Antcliffe 2010). This could have facilitated the diversification of eukaryotes. Moreover, increased oxygen levels could have facilitated the evolution of multicellularity in eukaryotes, as it would have increased the permissible size of a diffusion-limited multicellular organism (Knoll 2011).

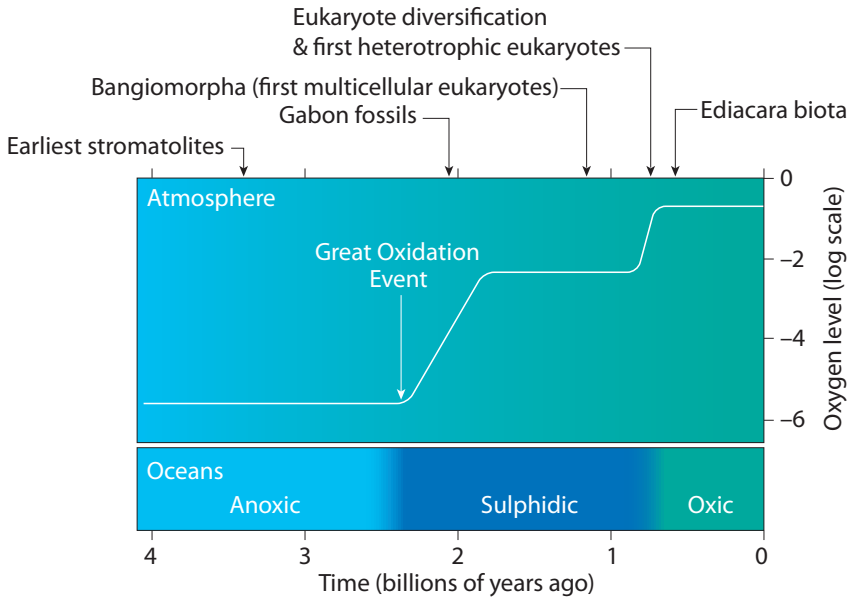


Figure 3. Geochemistry and the evolution of eukaryotes. Several events in the evolution of life are depicted, including the first evidence for stromatolites, the dating of Gabon fossils of putative multicellular organisms, the first multicellular eukaryotes and the big eukaryote diversification; together with major geochemical changes in atmosphere and oceans. Modified from Donoghue and Antcliffe 2010.

Although fossils of putative macroscopic multicellular organisms have also been described from 2100 Ma rocks in Gabon (Figure 3)(El Albani et al. 2010), it is not still clear whether these fossils record true multicellularity or colonies, eukaryotes or bacteria, or even if they are fossils or to abiotic structures (Knoll 2011). The oldest unequivocal multicellular eukaryotes appeared around 1200 Ma (Figure 3), these are the Bangiomorpha red algae, that have differentiated holdfasts and reproductive cells (Butterfield 2009; Knoll 2011). However, the vast majority of multicellular eukaryote clades appeared much later, after the oxygen increase 800Ma. This includes macrophyte green algae (Charales, Coeocharaetales, Zygnematales) 750 Ma (Becker 2012; Laurin-Lemay et al. 2012), metazoans c. 600 Ma (see below), embryophytes c. 450 Ma (Sanderson 2003), multicellular fungi c. 300 Ma (Stajich et al. 2009) and phaeophytes c. 130 Ma (Silberfeld et al. 2010)(Figure 4).

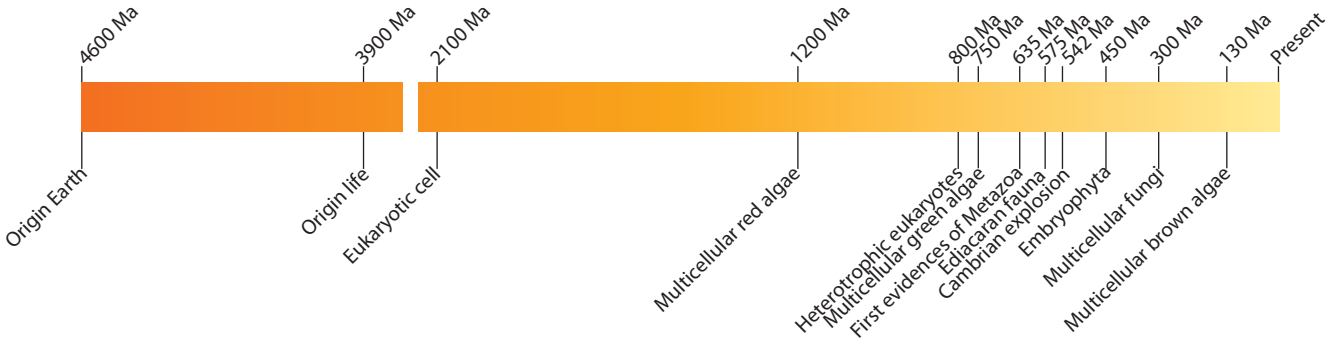


Figure 4. Timeline of the origins of multicellular eukaryotic clades.

Independently of when it happened, it is clear that multicellularity has repeatedly and independently evolved during eukaryote evolution. A conservative estimate is that it evolved sixteen times (Figure 5)(King 2004; Grosberg and Strathmann 2007), although more recent estimates elevates it to twenty-two (Adl et al. 2007; Knoll 2011). In any case, as previously mentioned, only in five cases we can speak of complex multicellularity: Embryophyta, Metazoa, Phaeophyta, Bangiales and Florideophycidae red algae, and some fungi.

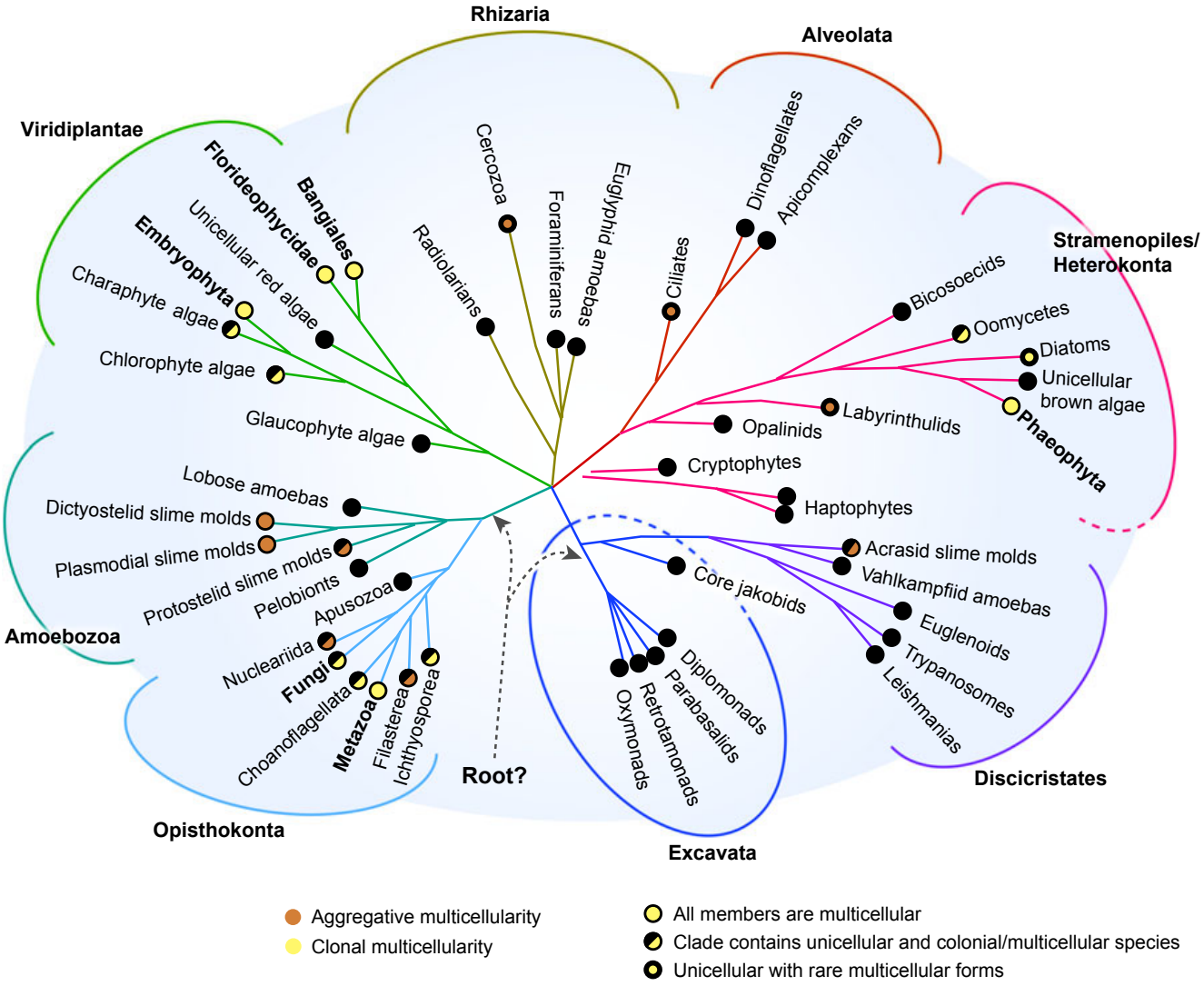


Figure 5. Phylogenetic distribution of multicellularity among eukaryotes. Taxa in bold include at least some complex multicellular representatives. Modified from Grosberg and Strathmann 2007.

In the case of embryophytes and metazoans, the embryonic development is well described and the molecular mechanisms associated to this development are known. In

contrast, little is known about red and brown algae. There are evidences of embryonic development in these cases too (Mshigeni and Lorri 1977; Bouget et al. 1998; Xie et al. 2010); but almost nothing is known about their developmental gene toolkit. Fungal multicellularity is, paradoxically, also poorly understood and quite problematic to categorise. Indeed, although complex multicellular fungi make up to 80-90% of described fungal diversity, almost nothing is known about fungal development and very few non-unicellular fungal genomes have been sequenced (Kües 2000; Stajich et al. 2009).

The modes of nutrition seem to be an important factor to explain the differences between complex multicellular lineages. For example, metazoans are the only phagotrophic organisms with complex multicellularity. It has been proposed that this is a crucial factor that explains the big gap between metazoans and non-metazoans, with no intermediate forms as thoe that we can see, for example, in plant sister lineages (Cavalier-Smith 2012). In autotrophs and osmotrophs, cells simply would need to stick together by producing an extracellular glue to become multicellular; but for phagotrophs cell aggregation severely interferes with feeding by intracellular ingestion. This would require more compelx tissue architecture, and cell behaviours. The origin of animals was therefore, inherently more difficult than the origin of other multicellularities due to the requirements of changing from single-cell phagotrophy to phagotrophy via a gut. The problem would have been initially solved with a sponge-like organization, that maintain a "protozoan" mode of feeding. In that sense, the choanoflagellate feeding mode (consisting of a flagellum surrounded with microvilli that trap particles) may have been an important step, as it could easily be accommodated in a pluricellular unit with autonomous feeding cells (much as what we can see in choanoflagellate colonies nowadays).

1.4 The origin of Metazoa

Dating the origin of metazoans has been for a long time, and still continues to be, a matter under intense dispute. The Burguess Shale-type Cambrian rocks have revealed an extraordinary explosion of metazoan forms that occurred around 542 Ma. Most of these fossils are indisputably identified as stem groups of extent metazoan phyla and classes (Morris 1989; Erwin et al. 2011)(Figure 6). In contrast, molecular clocks infer a much older origin of Metazoa, around 800 Ma (see, for example, Peterson et al. 2008)(Figure 6). Between these two extreme dates, we find the enigmatic Ediacaran fauna, which lived in the late Proterozoic (579-565 Ma), after the mid-Ediacaran Gaskiers Glaciation. These organisms have been presumptively classified into nine different clades (Erwin et al. 2011), including Arboreomorphs, Rangeomorphs (the oldest ones) (Figure 7), Triradialomorphs, Bilateralomorphs, Erniettomorphs, Kimberellomorphs (Figure 7), Pentaradialomorphs, Dickinsoniomorphs and Tetraradialomorphs. Some of these fossils have been ascribed to extant metazoan phyla. For example *Cloudina* is hypothesised to be a cnidarian (Figure 7), Kimberellomorpha is supposed to be the first known bilaterian (probably a mollusc), and Dickinsoniomorphs are hypothesised to be stem Placozoa or stem Eumetazoa.

The presence of bilateral and superficially segmented animal forms would mean that the toolkit for animal bilaterality was present, at least, around 575 Ma. But the assignment and classification of these fossils is still highly controversial. For example, recent work suggests that some of the Ediacaran fossils, including *Dickinsonia*, may have been instead lichen-like organisms or microbial colonies that lived on land (Retallack 2012). Another case are the polemic Doushantuo fossils (Figure 7), which date from late Ediacara and were initially described as developing embryos (Yin et al. 2007). However, they have recently been shown not to be animal embryos but probably encysting protists of unclear affinity (although they hold some surprising resemblance to, for example, colony-forming ichthyosporeans) (Huldtgren et al. 2011). Furthermore, synchrotron analysis has revealed that even those fossils that had been previously interpreted as gastrulating embryos (a quite convincing evidence of an animal embryo) are due to diagenetic artefacts (Cunningham et al. 2012). Therefore, Ediacaran embryology remains obscure.

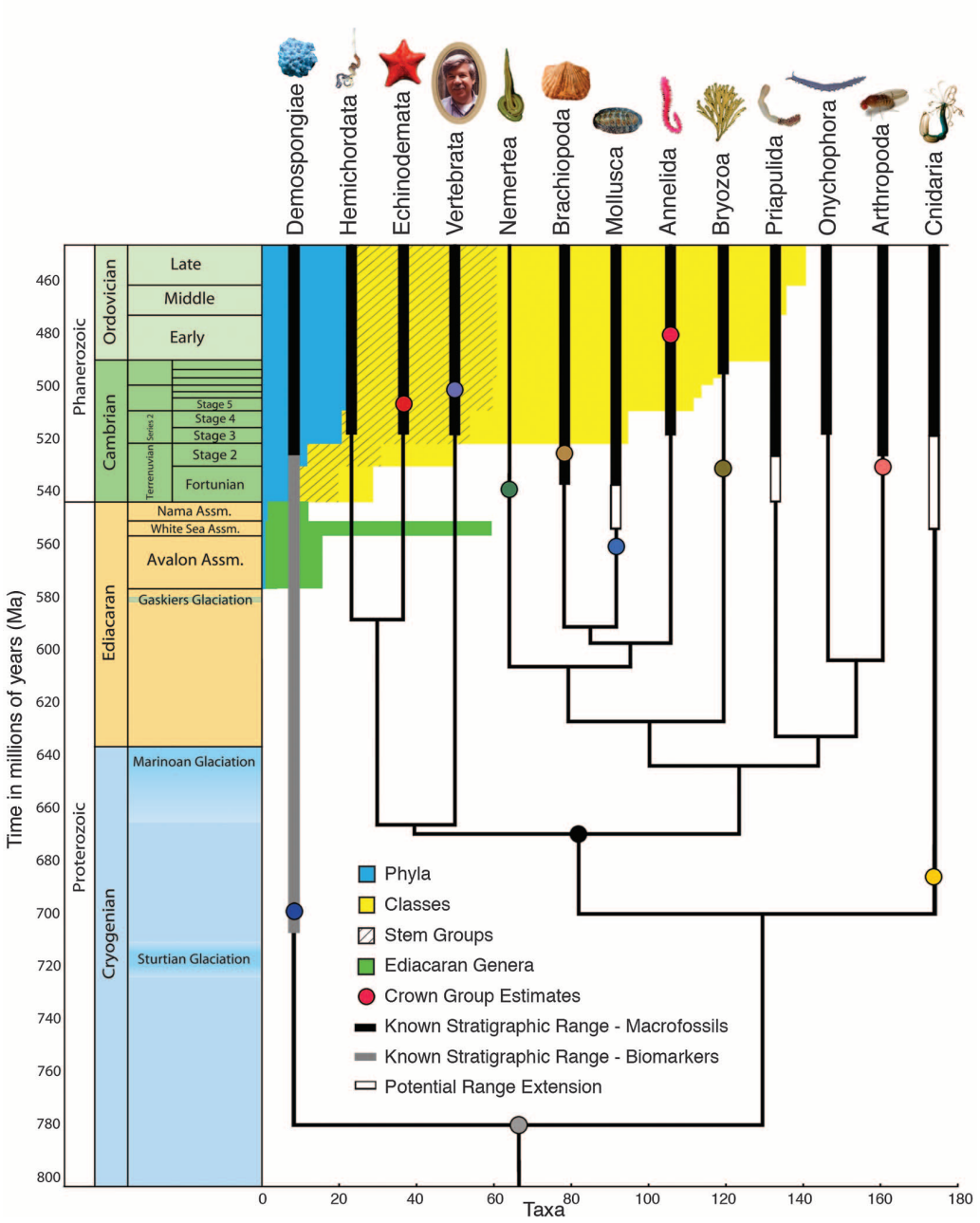


Figure 6. Origin of the major metazoan phyla. In green is the record of macroscopic Ediacaran fossils and in blue and yellow the known fossil record of animals at the phylum and class level, respectively (hatching indicates stem lineages, i.e., lineages that belong to a specific phylum but not to any of its living classes). Shown in thick black lines is the known fossil record of each of these 13 lineages through the Cryogenian-Ordovician. Coloured circles show molecular clock estimates of the respective crown groups. We can see that in many cases molecular clock estimates and fossil/biomarker register are quite in agreement, for example in the case of sponges or echinoderms; being Cnidaria the major exception. From Erwin 2011.

The late-Ediacaran origin of animals would fit well with the post-snowball earth theory of animal origins. This hypothesises that, during the repeated interspersed ice ages between 2400 Ma and 600 Ma that completely froze the Earth surface (Hoffman 1998), the demographic effects in the effective population size may have favoured the altruistic behaviour necessary for the origin of multicellularity (Boyle et al. 2007). Not only this theory is problematic, but also the snowball earth theory itself is at stake (Sansjofre et al. 2011). And, as we will see below, new evidence pushes further back in time the origin of animals.

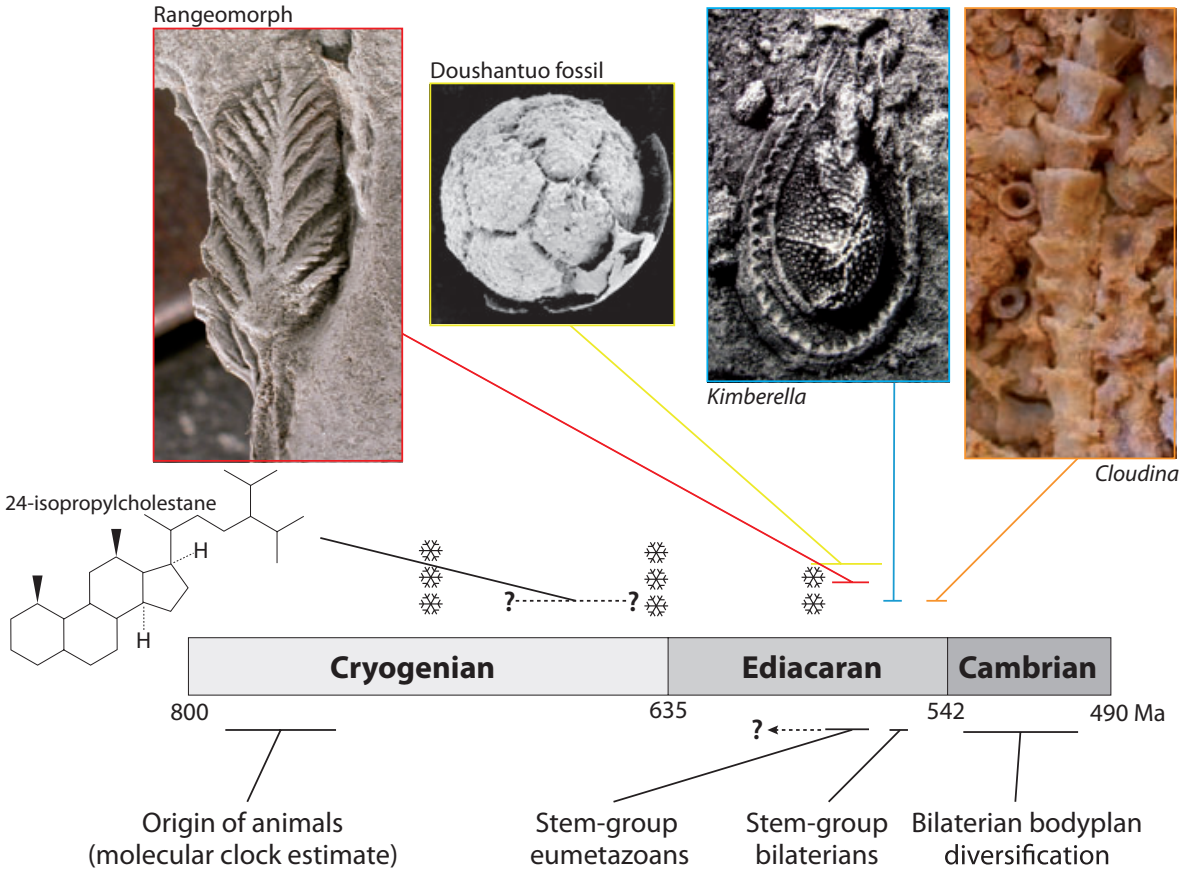


Figure 7. The geologic record of early animal evolution. Snowflake patterns indicate the Sturtian, Marionan and Gaskiers glaciations (from left to right). Modified from Knoll 2011.

In 2009, Love et al. reported the identification of abundant sedimentary 24-isopropylcholestanes, the hydrocarbon remains of C30 sterols produced by marine demosponges, dated before the end of the Marinoan glaciation (635 Ma) (Figure 11). An earlier study by Kodner et al. (2008) showed that this kind of sterols are not produced by choanoflagellates (the closest unicellular relatives of metazoans), supporting the notion that 24-isopropylcholestane is a good molecular proxy for demosponges. The earlier origin of non-spicule producing demosponges, compared with other sponges, would explain why the earliest mineralised sponge spicules occur only around 544 Ma (Brasier et al. 1997). Finally, a recent study (Maloof et al. 2010) describes possible sponge fossils of later Neoproteroic (around 660 Ma), between the Sturtian (710) and the Marionan (635) glaciations. The reconstruction of the fossils (Figure 8) shows an asymmetric body plan and the presence of an interconnected canal system, observations that suggest a sponge affiliation. Together, these results provide

compelling evidence that metazoans originated during the Cryogenian (between 850-635Ma). This inference is more close to the molecular clock estimates (see above).

Overall, the current picture (Figure 6) of the origin of Metazoa shows that rather than a sudden explosion of diversity, there was a long period of metazoan evolution without much diversification, followed by a first radiation that produced the Ediacaran fauna (with perhaps some stem metazoan phyla) and finally the burst of Cambrian diversity, that originated almost all extant metazoan phyla and classes.

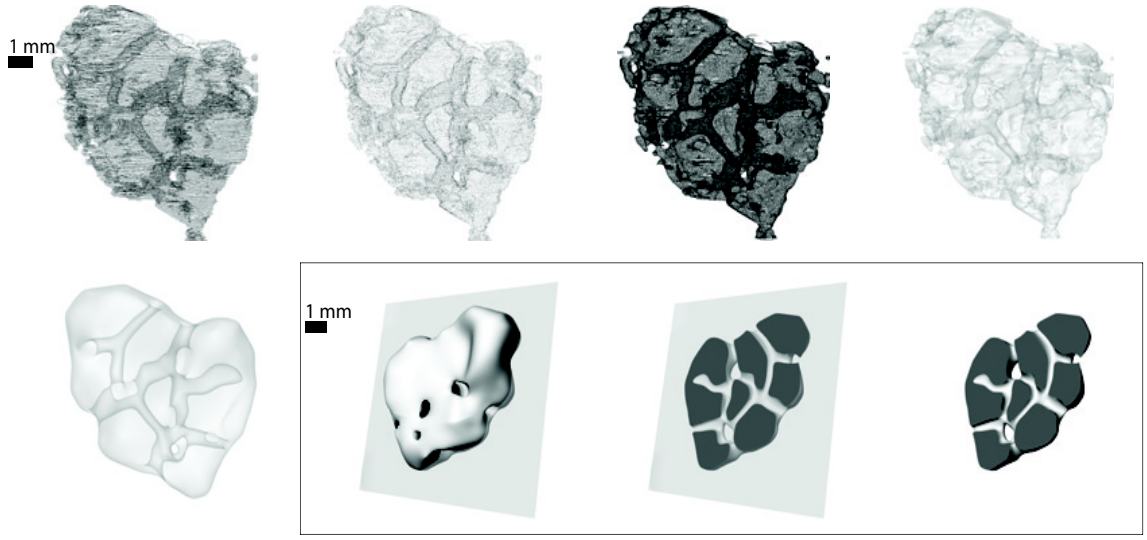


Figure 8. Digital reconstruction of a putative Cryogenian demosponge (c. 660 Ma). The fossil comes from the Trezona formation in southern Australia. Notice the presence of an interconnected canal system. From Maloof et al. 2010.

1.5 The origin of the metazoan genetic developmental programs

The evolutionary origin of metazoan development is another subject that has generated considerable speculation. The surface-to-interior gradients of oxygen and nutrients that are inherent to a three-dimensional multicellular structure may have been a first way to induce differentiation between outer and inner cells, especially considering the importance nutrient deprivation and hypoxia as signals for differentiation in many eukaryotes (Aguirre et al. 2005; Loenarz et al. 2011). These signals may have triggered differentiation programs to solve the problems of molecule diffusion in a colony, programs that ultimately led to overcome diffusion-imposed restrictions in size. A first way of doing this is by establishing cell-cell protein channels to facilitate diffusion between cells. But a much more efficient way is by creating a fluid space were superficial cells can pump nutrients, allowing bulk nutrient transfer (Beaumont 2009). Later, specialised tissues (the circulatory systems) evolved to perform similar functions. Therefore, simple morphogenesis may have been enough to solve the inherent metabolic limitations of the first multicellular animals. Control of morphogenesis through metabolism is arguably much less prominent in modern metazoans, but the regulative role of metabolic gradients is still detectable in processes such as vertebrate angiogenesis and, more extensively, in early-branching metazoans (Knoll 2011).

The question is how the genetic programs that control morphogenesis appeared, independently of the selective reasons for evolving them. For Newman and colleagues (Newman 2005; Newman et al. 2006; Newman and Bhat 2009; Newman 2012), at the very early stage of metazoan evolution, morphogenesis was highly plastic. It was based on self-organizing morphogenetic capabilities (e.g., differential cell adhesion, lateral inhibition or biochemical oscillation) that relied in a very basic genetic starting toolkit. Tightly genetically controlled development programs appeared later in evolution as "frozen accidents" of stabilizing selection, which turned organisms into less mutually interconvertible and plastic, and with a more stereotyped development (Newman 2012). Therefore, genetic control of morphogenesis became increasingly prominent during metazoan evolution, but the roots of animal multicellularity cannot be found in the complexity of patterning mechanisms displayed by extant metazoan phyla.

An alternative view held by Davidson and colleagues (Davidson and Erwin 2006) suggests that morphogenesis and spatial patterning of cells was based, from the very

beginning, on earlier-evolved cell type-specific genetic "kernels". These kernels are evolutionary inflexible gene regulatory networks (composed of transcription factors and cis-regulatory modules) that perform essential upstream functions in establishing metazoan body plans. Therefore, conservation of phyletic body plans is due to the retention, since pre-Cambrian times, of these kernels. The theory also predicts that only peripheral elements of the gene regulatory networks (such as effector genes) would change during later metazoan evolution.

However, both views are not so different and, indeed, even complementary. While the first stresses the role of genetic mechanisms that mobilise physical processes (the Dynamical Patterning Modules, *sensu* Newman) and gives a secondary (later) role to developmental transcription factors; the second emphasises the role of these transcription factors networks as essential to explain metazoan diversification. One could see Davidson's conserved kernels just as the very first outcome of Newman's "frozen accidents".

In any case, it is noteworthy that, once multicellular developmental programs were established, they became locked in by the accumulation of reinforcing genetic circuitry and fine-tuning. Embryonic multicellularity underpinned the establishment of irreducible complexity; an evolutionary ratchet (whether neutral or not) where uncontrolled cell proliferation, failures in pattern formation or in cell differentiation, and non-programmed cell death imposed severe penalties to fitness.

1.6 The unicellular relatives of Metazoa

Previous to perform any comparative genomic studies on the origin of metazoan multicellularity, we need a robust phylogenetic framework of the relationships between metazoans and close unicellular eukaryotes. But this is not a trivial question and it is a subject under continuous changes and updates. We will summarise here those clades of the eukaryotic tree of life that are more relevant to our question. Our description here is based on two of the most updated and accurate analyses on the subject to date, by Torruella et al. (2012) and by Paps et al. (2013).

Metazoans are embedded in a series of broader phylogenetic assemblies (for a review, see Paps and Ruiz-Trillo 2010)(Figure 9):

- Unikonts (Cavalier-Smith 2002): The name 'Unikonta' refers to the structure of the flagellar apparatus, composed of one or two basal bodies (centrioles plus their related microtubules) in conjunction with their associated cilia/flagella. According to Cavalier-Smith, the last common ancestor of Unikonta had only one centriole and one cilium per kinetid. In contrast, the last common ancestor of Bikonta (the group that includes plants, alveolates, rhizarians, stramenopiles, etc.) had two. This division is highly controversial, given that many extant unikonts have in fact two cilia. But recent phylogenomic analyses support it (Ruiz-Trillo et al. 2008; Derelle and Lang 2012). There are three major unikont lineages: Amoebozoa (that includes dictyotelids and other amoeboid protists), Apusozoa (the paraphyletic sister-group of opisthokonts, that includes apusomonads and ancyromonads (Paps et al. 2013)) and Opisthokonta (the group that includes animals, fungi and their respective unicellular relatives).
- Opisthokonta (Adl et al. 2005): opisthonkonts are unikonts with a single posterior flagellum (in at least one vital phase, although sometimes secondarily lost), one single pair of centrioles and mitochondria with flat cristae (compartments outlined by the mitochondrion inner membrane). The monophyly of Opisthokonta has been confirmed by multiple studies (Lang et al. 2002; Medina et al. 2003; Steenkamp et al. 2006; Ruiz-Trillo et al. 2008; Torruella et al. 2012) and by a molecular synapomorphy consisting of a 12-amino acid insertion in the elongation factor-1alpha gene (Baldauf and Palmer 1993). Opisthokonta includes two lineages: Holomycota (Liu et al. 2009) (fungi and their unicellular relatives, the nucleariids) and Holozoa (Lang et al. 2002) (animals and their unicellular relatives).
- Holozoa (Lang et al. 2002): Holozoa is the phylogenetic assemblage that includes Metazoa and their unicellular relatives. The most recent phylogenomic analyses (Torruella et al. 2012) show that there are, at least, three independent unicellular lineages close to Metazoa: Ichthyosporea, Filasterea and Choanoflagellata (see below).
- Filozoa (Shalchian-Tabrizi et al. 2008): is the group formed by Filasterea, Choanoflagellata and Metazoa.

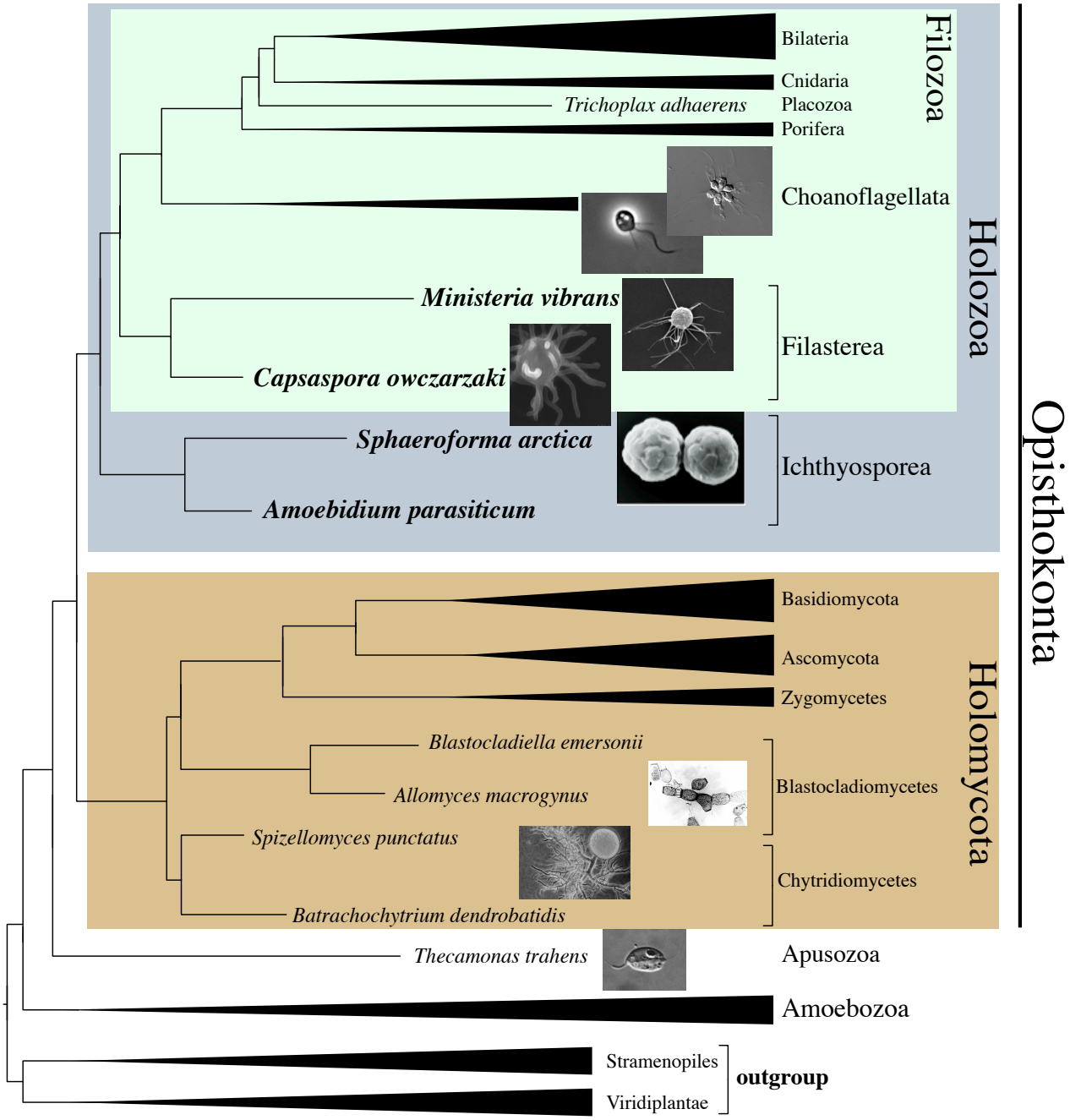


Figure 9. Phylogenetic relationships of Opisthokonta. The tree is based on the phylogenomic analysis from Torruella et al 2012. Figure courtesy of Guifré Torruella Cortés.

Metazoans themselves can be divided in diploblastics and Bilateria. The former is a paraphyletic arrangement of the earliest-branching metazoans, that is, Porifera (sponges), Ctenophora (comb jellies), Cnidaria (sea anemones and jellyfishes) and

Placozoa (*Trichoplax adhaerens*). The later is a monophyletic clade that includes all animals with bilateral symmetry and three germ layers (Telford 2006). Bilateria are further divided into Protostomia (Lophotrocozoa and Ecdysozoa) and Deuterostomia (Chordata and Ambulacraria) (Dunn et al. 2008; Hejnol et al. 2009). Despite its key position to understand the origin of Metazoa, the phylogenetic relationships of early-branching animals are far from understood. Some phylogenies support Porifera as the earliest-branching animals (Philippe et al. 2009; Sperling et al. 2009; Pick et al. 2010) (but they do not agree on whether sponges are monophyletic or paraphyletic) and others support ctenophores (Dunn et al. 2008; Hejnol et al. 2009). Neither are clear the relations between them, although in all these works the position of Cnidaria as sister group of Bilateria (forming the monophyletic Eumetazoa clade) seems to be settled.

Choanoflagellates are the closest unicellular relatives of Metazoa, a position that has been confirmed in recent years by several phylogenetic studies (Figure 10A,B) (Carr and Leadbeater 2008; Torruella et al. 2012). They comprise around 140 species of heterotrophic single-celled flagellates, mostly marine. Some of them are able to form colonies (Figure 10A) through clonal division and with intercellular bridges (Fairclough et al. 2010; Dayel et al. 2011). They bear a single flagellum surrounded by an actin-based microvilli collar. The flagellum is used for locomotion and to create a water flow that traps particles in the collar net. Choanoflagellates are strikingly similar to sponge choanocytes (feeding cells), a resemblance already noticed by James-Clark 150 years ago (James-Clark 1866). Choanoflagellates were the first unicellular holozoans to be studied in the context of the origin of animal multicellularity (King et al. 2003; King 2004; King 2005) and, recently, the genomes of two choanoflagellate species, *Monosiga brevicollis* and *Salpingoeca rosetta*, have been sequenced (King et al. 2008; Fairclough et al. 2013).

Filasterea is a group defined based on molecular data and includes two known species, *Capsaspora owczarzaki* (Figure 10D) and *Ministeria vibrans* (Figure 10C) (Ruiz-Trillo et al. 2004; Shalchian-Tabrizi et al. 2008). It is named after the thread-like tentacles that both genera share. *Capsaspora* amoeboid-like cells were isolated from the pulmonate snail *Biomphalaria glabrata* (Stibbs et al. 1979; Owczarzak et al. 1980), where they seem to feed from the larvae of the trematode *Schistosoma mansoni* that parasite the snail. *Capsaspora* is, thus, a symbiont of *B. glabrata* and was first named as *Nuclearia* sp. and considered a member of the nucleariids (the unicellular sister group of fungi)

(Zettler et al. 2001). *Ministeria vibrans* (Shalchian-Tabrizi et al. 2008) are free-living marine microorganisms with long un-branched filopodia.

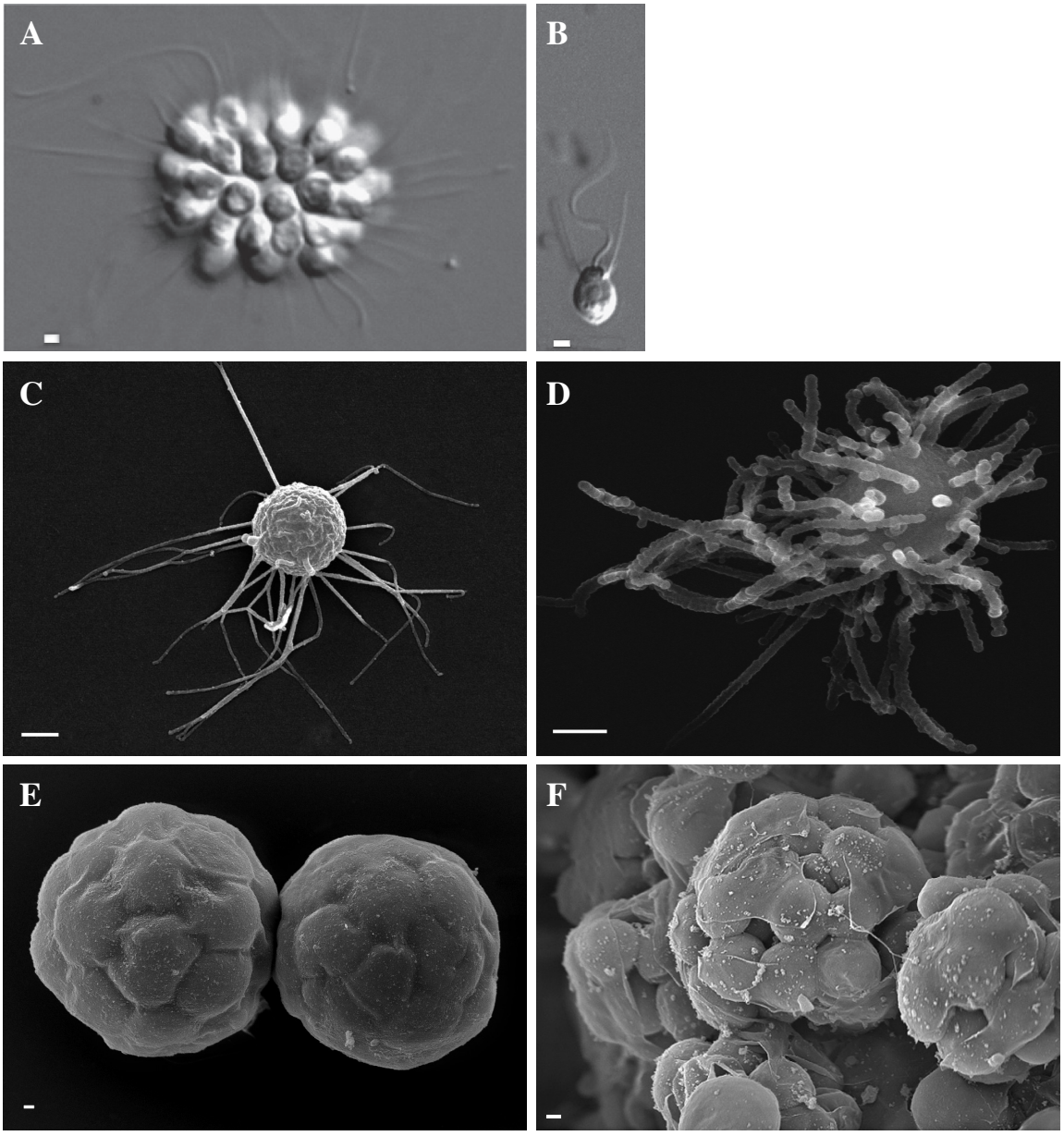


Figure 10. The unicellular relatives of Metazoa. A) *Salpingoeca rosetta* (Choanoflagellata) colony. B) *Salpingoeca rosetta* (Choanoflagellata) free swimmer. C) *Ministeria vibrans* (Filasterea) D) *Capsaspora owczarzaki* (Filasterea) E) *Sphaeroforma arctica* (Ichthyosporea) colony. F) *Creolimax fragrantissima* (Ichthyosporea) colony). A and B are DIC light microscopy pictures, modified from Dayel et al 2012. C, D, E and F are scanning electron microscopy pictures. All scale bars=1 μm**.**

Ichthyosporea are single-celled organisms, with flat mitochondrial cristae and some have syncytial colonial forms/sporangia (Mendoza et al. 2002; Marshall et al. 2008; Marshall and Berbee 2011), produced by hypertrophic growth and subsequent cellularization to form endospores. Most are animal parasites or endosymbionts,

although there are some free-living species. Ichthyosporeans are divided in two groups: Rhinosporidiacae/Dermocystida and Ichthyophonae. The former are well-studied human parasites, with a posterior flagellum and sporangia formation (Mendoza et al. 2002). The later include parasites of fish or aquatic arthropods and never have flagellum, but in several cases they have amoeboid cells and many, like *Sphaeroforma arctica* (Figure 10E) and *Creolimax fragrantissima* (Figure 10F) (Marshall et al. 2008), form colonies (Figure 10C). Initial analysis placed ichthyosporeans as the sister-group to Filasterea (Ruiz-Trillo et al. 2008), but recent studies have shown that Ichthyosporea is an independent lineage (Shalchian-Tabrizi et al. 2008; Torruella et al. 2012), the third closest unicellular relatives of metazoans.

The term "Choanozoa", which is used by some authors (Cavalier-Smith 2003; Shalchian-Tabrizi et al. 2008), refers to Choanoflagellata, Filasterea and Ichthyosporea; the three unicellular holozoan lineages. It is, of course, a paraphyletic assemblage.

Finally, there are two enigmatic unicellular clades that were considered to be within the Holozoa group. This is the case of the Aphelidea, intracellular parasites of algae with both an amoeboid and flagellated phases and a complex life cycle. They were placed together with Ichthyosporea (Gromov 2000), but have been recently demonstrated to branch with early-diverging fungi (Karpov et al. 2012). Another enigmatic lineage is formed by a single species, *Corallochytrium limacisporum* (Raghu-kumar 1987), a colony and amoeba forming protist that has been proposed to branch in different positions: sister-group of choanoflagellates (Zettler et al. 2001; Mendoza et al. 2002), related with fungi (Sumathi et al. 2006), related to ichthyosporeans (Ruiz-Trillo et al. 2004; Steenkamp et al. 2006) and branching between filastereans and choanoflagellates (Paps et al. 2013).

1.7 Comparative genomics and the Urmetazoan genome

The study of the origin of Metazoa experienced a revolution with the advent of the automatic genome sequencing techniques at the beginning of the XXIst century. The analyses of the genome sequences of early-branching metazoans such as the cnidarians *Hydra magnipapillata* and *Nematostella vectensis* (Putnam et al. 2007; Chapman et al. 2010), the placozoan *Trichoplax adhaerens* (Srivastava et al. 2008) and the demosponge *Amphimedon queenslandica* (Srivastava et al. 2010) provided the first comparative genomics approach to the precise definition of the common panmetazoan molecular toolkit (i.e. the genes that define the metazoan condition and provide essential functions for multicellularity). This metazoan genetic toolkit was composed of metazoan-specific genes such as cadherins and integrins for cell adhesion, diverse signalling receptors (Notch, Receptor Tyrosine kinases, Wnt receptors, etc.) and signal transduction cascades (Hippo pathway, etc.), as well as many transcription factors involved in metazoan development (T-box, Runt, NFkappaB, Myc/Max, etc.).

But this approach represents only half of the story, which could be completed by comparing this phylogenetically broad array of metazoan genomes with the genomes of diverse metazoan unicellular relatives (Ruiz-Trillo et al. 2007). The logic behind it is simple (Figure 11): by comparing metazoan genomes with those of their closest unicellular relatives we may define a minimal Urmetazoan genome content. By removing from this Urmetazoan genome those genes that are common to all eukaryotes,

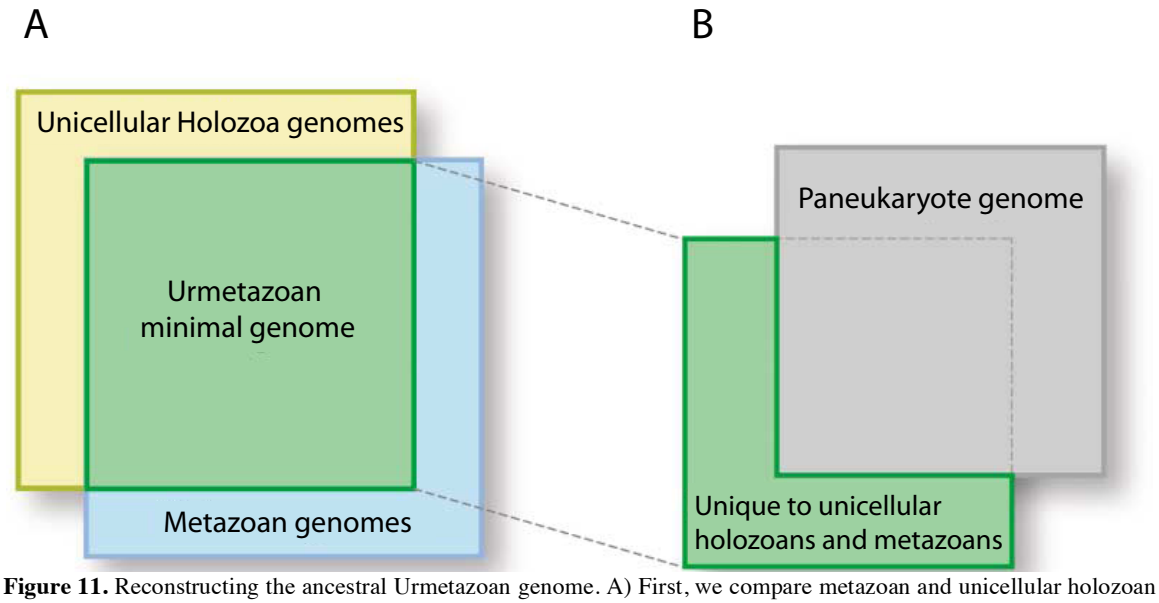


Figure 11. Reconstructing the ancestral Urmetazoan genome. A) First, we compare metazoan and unicellular holozoan genomes. B) Then, we extract those genes present in all eukaryotes and we end up with a set of genes that we infer were present in the unicellular ancestor of Metazoa. Modified from King 2004.

we end up with very good proxy of the genome content of the unicellular ancestor of metazoans.

As explained above, there are three unicellular lineages close to metazoans, the choanoflagellates, the filastereans and the ichthyosporeans. The sequencing of the genome of the choanoflagellate *Monosiga brevicollis* represented a hallmark in the study of the origin of animal multicellularity (King et al. 2008). It proved the power of this panholozoan comparative genomics approach, revealing that several genes previously considered metazoan-specific, such as tyrosine kinases or cadherins (Abedin and King 2008; Manning et al. 2008) were, in fact, present in the last common ancestor of metazoans and choanoflagellates. This, together with a growing body of knowledge about the basic choanoflagellate cell biology (Fairclough et al. 2010; Dayel et al. 2011) and the recent sequencing of the colony-forming choanoflagellate *Salpingoeca rosetta* (Fairclough et al. 2013), provides a completely new approach to study the origin of metazoans.

However, and given that gene loss is prominent among eukaryotes (Zmasek and Godzik 2011), the genomes of filastereans and ichthyosporeans are definitively needed in order to robustly infer the unicellular prehistory of metazoans. To this end, the UNICORN initiative, which comprises several labs around the world, aimed at sequencing the genomes of several key species to better understand the origin of multicellularity in both animals and fungi (Ruiz-Trillo et al. 2007). Among the genomes to be sequenced by the UNICORN project are the filasterean *Capsaspora owczarzaki* and the ichthyosporeans *Sphaeroforma arctica* and *Amoebodium parasiticum*.

The main handicap for the study of these species, in contrast with choanoflagellates (which have been studied for more than one century (James-Clark 1866)), is the lack of knowledge about their basic biology (life cycle, cell structure, natural habitat, etc.). Indeed, a quick check into the scientific literature shows that there are more and much older articles studying choanoflagellates than ichthyosporeans and, specially, than filastereans (Figure 16); the later being basically studied in the XXIst century.

It is in this context that this work has been developed, by studying *Capsaspora owczarzaki*, from both a genomic as well as a basic cell biology perspective.

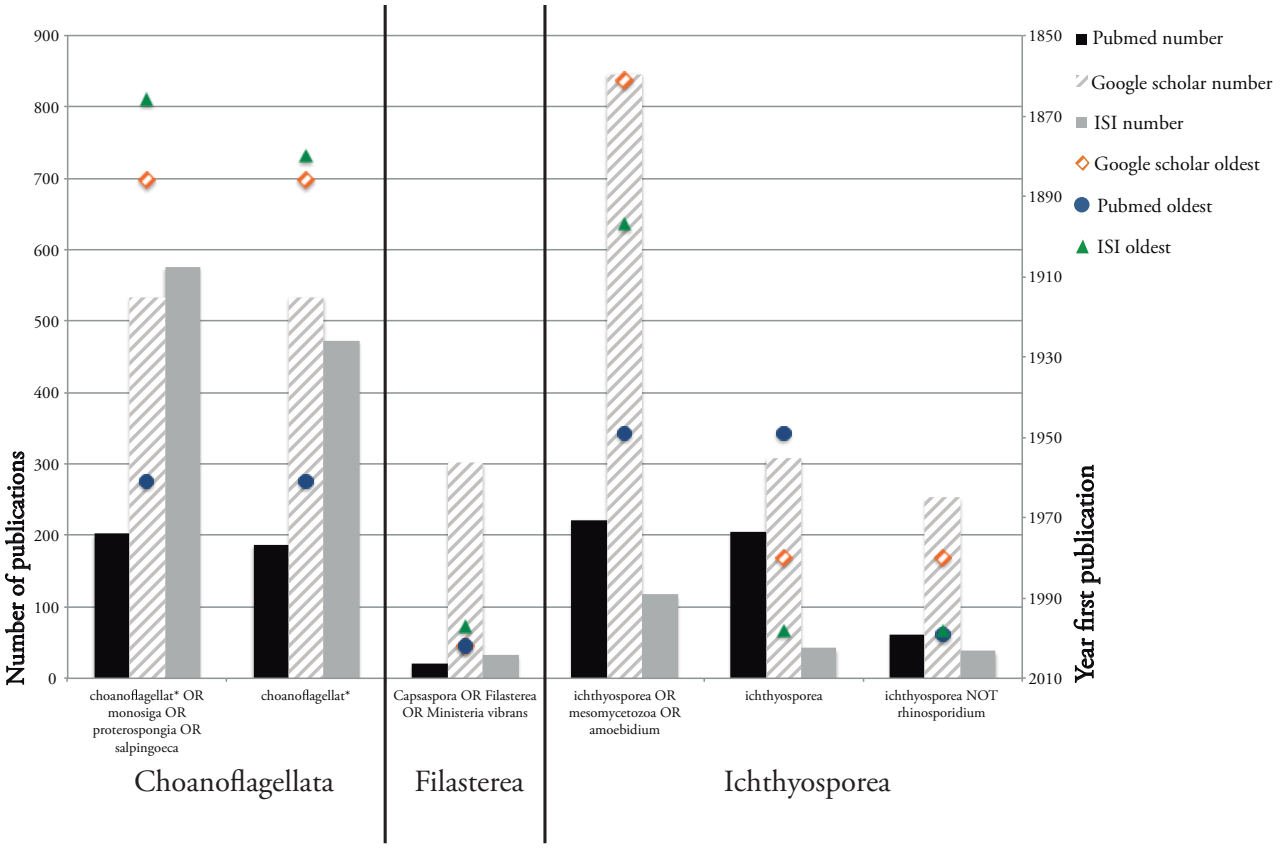


Figure 12. Scientific literature about unicellular holozoans. The graphic shows the results of the different searches (X-axis) in three different databases. Bars represent the number of publications (primary Y-axis) and dots show the year of the oldest publication (secondary Y-axis). Note: *Capsaspora* was described for the first time in 1979, but not classified until later.

1.8 *Capsaspora owczarzaki* and its genome

In 1979 and 1980, an amoeba-like symbiont was isolated from the mollusc *Biomphalaria glabrata* (Stibbs et al. 1979; Owczarzak et al. 1980) in the context of the study of this snail as an intermediate host of the widespread human trematode pathogen *Schistosoma mansoni*. The symbiont amoebas were obtained from the pericardial explants and mantle swabs of snails originally sampled in Puerto Rico. They were reported to be able to attack and kill sporocysts of the trematode. *C.owczarzaki* cells were described as 3-5 μm amoebas with a nucleus 1/3-1/2 of the diameter of the cell (containing a central nucleolus), long unbranched pseudopodia, mitochondria with flattened cristae, numerous phagosomes, lipid vacuoles, glycogen granules and a Golgi apparatus. Cells could encyst in response to crowding, generating 4-5 μm cysts with a double wall: the outer thin, irregular and loosely attached; and the inner thicker, smooth. These original works did not formally describe the species and, based on morphological

characters, they placed *C.owczarzaki* within in the genus *Nuclearia*.

Twenty years later, in 2001, Zettler et al. published a molecular phylogeny that attempted to resolve the phylogenetic relationships of the genus *Nuclearia*. They used three free-living *Nuclearia* species and the original *B.glabrata*-isolated amoeba. Surprisingly, the later did not cluster with the other three species. But, since there were no clear affinities with any other ophisthokonts, they refrained from reclassifying it.

In 2002, Hertel et. al used the original isolate from Puerto Rico and new isolates extracted from Brazilian snails, in a new attempt to classify the enigmatic amoeba. They amplified a sequence comprising 18S, ITS1, 5.8S, ITS2 and the beginning of the 28S rDNA gene regions. Their phylogenetic analysis showed the affinity of the amoebas with Mesomycetozoa (= Ichthyosporea) (Mendoza et al. 2002). They described the symbiont amoeba as a new genus and species: *Capsaspora owczarzaki*. The genus name is derived from the Greek *caps*="eat quickly" and *spora*="seed", for its ability to rapidly ingest digenetic trematode sporocysts in vitro. The species name was selected in honour of Dr Alfred Owczarzak, the author who isolated the organisms and provided a first ultrastructural description of the attack and consumption of sporocysts of the trematode *S. mansoni* by this organism.

In 2004, a new phylogenetic analysis based on three markers (small and large ribosomal subunits and actin) by Ruiz-Trillo et al., showed that, in fact, *C.owczarzaki* was an independent opisthokont lineage, branching between ichthyosporeans and choanoflagellates. The same authors published in 2008 a phylogenomic analysis confirming that *C.owczarzaki* is an independent lineage, either the sister-group to choanoflagellates and metazoans (as inferred by mitochondrial data) or the sister-group to Ichthyosporea (as inferred by ribosomal proteins). In that same publication, the authors showed, based on EST searches, that Capsaspora owczarzaki had some genes important for animal multicellularity, like MAGI protein or tetraspanin (Ruiz-Trillo et al. 2008).

More recently, a phylogenomic study by Torruella et al. (2012) resolved confidently that, similar to what was published in 2004, *C.owczarzaki* is an independent opisthokont lineage, the second closest unicellular relative of Metazoa.

Capsaspora owczarzaki can be easily cultured in the lab in sterile cell culture flasks at 23ºC. It grows in ATCC Medium 1034 (Modified PYNFH medium), composed of

bactopeptone, yeast extract, yeast nucleic acid, folic acid, hemin, fetal bovine serum and distilled water.

In the context of the UNICORN project, the *Capsaspora owczarzaki* ATCC30864 strain (the original one isolated in 1979 from Puerto Rico snails) genome was sequenced by Sanger chemistry. The raw reads were assembled into 84 scaffolds, spanning 28 Mb in total and with an 8x coverage. *Capsaspora owczarzaki* genome has 8567 predicted protein-coding genes, comprising 58.7% of the genome and with an average gene size of 3.2 Kb (Table 1). It is, therefore, a relatively compact genome, compared to the genomes of metazoans and choanoflagellate *Monosiga brevicollis* (Table 1).

Table 1. Genome statistics of *Capsaspora owczarzaki* and other eukaryotes. H.sa=Homo sapiens; N.vec=Nematostella vectensis; A.que=Amphimedon queenslandica; M.bre=Monosiga brevicollis; S.pombe=Schizosaccharomyces pombe; S.cer=Saccharomyces cerevisiae; D.dis=Dictyostelium discoideum.

	H. sa	N. vec	A. que	M. bre	**C. owc**	S. pombe	S. cer	D. dis
Genome size (Mb)	3101.8	357.0	167.1	41.6	**28.0**	12.6	12.1	34.1
% GC	40.9	40.6	31.1	54.9	**53.8**	36.1	38.3	22.4
Number of genes	22128	27273	30327	9171	**8657**	5155	5863	12474
Genome size/gene (Kb)	140.2	13.1	5.5	4.5	**3.2**	2.4	2.1	2.6
Transcribed % genome	1.2	7.6	21.4	39.7	**58.7**	57.2	72.4	61.8
Mean intron # per gene	8.8	4.3	4.7	6.6	**3.8**	1.0	0.1	1.5
Mean intron size (bp)	5645	799	251	171	**166**	82	203	139

Objectives

Whether mythic or scientific, the view of the world that man construct is always largely a product of imagination. For scientific process does not consist simply in observing, in collecting data, and in deducing from them a theory. One can watch an object for years and never produce any observation of scientific interest. To produce a valuable observation, one has first to have an idea of what to observe, a preconception of what is possible.

François Jacob, *Evolution and tinkering*.
Science, 1977. *196*(4295), 1161–116

The sequencing of the genome of *Capsaspora owczarzaki* and other species in key phylogenetic positions of the opisthokont tree of life (including early-branching fungi like *Spizellomyces punctatus* or the apusozoan *Thecamonas trahens*, sister group to all opisthokonts), in the context of the UNICORN sequencing project, opened a completely new window to approach the origin of metazoan multicellularity through comparative genomics. My Thesis has developed in this framework, with three main objectives:

1. Reconstruct the evolutionary history of different elements of the metazoan multicellularity molecular toolkit, using comparative genomics.

2. Evaluate the functional conservation of some of these elements between metazoans and their unicellular relatives.

3. Gain insights into the basic biology of *Capsaspora owczarzaki* in order to understand the phenotypic unicellular context in which these elements work.

Results

L'esperit […] torna a aparèixer al segon graó de l'escala de la vida, en el graó dels protozous, fills de les tenebres, que a les portes de la vida caminen a les palpentes com a les portes de la mort, i viuen donant al món els primers senyals de la separació entre l'esperit i la matèria.

Francesc Pujols, *Hiparxiologi*
Llibres de l'índex, 2003

INFORME DELS DIRECTORS SOBRE ELS ARTICLES PUBLICATS

Directors: Dr. Iñaki Ruiz-Trillo i Prof. Jaume Baguñà Monjo

Els set articles que conformen aquesta Tesi doctoral (tots ells com a primer co-autor) han estat publicats o estan en vies de ser-ho en revistes d'alt impacte en l'àmbit de l'Evolució, la Genètica o la Biologia; totes elles incloses en la base de dades PubMed. S'indiquen per a cadascuna d'elles, quan són disponibles, els índex d'impacte i la seva posició en els rànquings de diferents disciplines.
La gran majoria d'articles inclosos en aquesta tesi han estat realitzats en col·laboració amb altres grups d'investigació capdavanters en els seus respectius àmbits i també en col·laboració amb altres membres del grup.

Article R1

Sebé-Pedrós, A., Roger, A., Lang, F. B., King, N., & Ruiz-Trillo, I. (2010). **Ancient origin of the integrin-mediated adhesion and signaling machinery.** Proceedings of the National Academy of Sciences of the United States of America, 107(22), 10142–7.

Factor d'impacte (2010): 9.771
Posició dins l'àrea: Multidisciplinary Sciences 3/59 (Q1)

El doctorand va participar activament en la concepció, el disseny experimental i la redacció de l'article. Va realitzar la totalitat d'anàlisis inclosos en l'article (incloent cerca i anotació de gens, validació per RACE-PCR d'alguns dels nous gens descrits, alineaments i anàlisis filogenètics) i contribuí significativament a la seva discussió i interpretació.

Article R2

Sebé-Pedrós, A., & Ruiz-Trillo, I. (2010). **Integrin-mediated adhesion complex. Cooption of signaling systems at the dawn of Metazoa.** Communicative & integrative Biology, 3(5), 475–77.

Factor d'impacte (2010): pendent.
Posició dins l'àrea: pendent.

El doctorand va participar activament en la concepció, discussió i redacció de l'article.

Article R3

Sebé-Pedrós, A., De Mendoza, A., Lang, B. F., Degnan, B. M., & Ruiz-Trillo, I. (2011). **Unexpected repertoire of metazoan transcription factors in the unicellular holozoan Capsaspora owczarzaki.** Molecular Biology and Evolution, 28(3), 1241–54.

Factor d'impacte (2011): 5.510
Posició dins l'àrea: Biochemistry & Molecular Biology 49/286 (Q1)
Evolutionary Biolgy 7/45 (Q1)
Genetics & Heredity 20/156 (Q1)

El doctorand va participar activament en la concepció, el disseny experimental i la redacció de l'article, del qual n'és co-primer autor. Va realitzar, juntament amb l'altre co-primer autor, la totalitat d'anàlisis inclosos en l'article (incloent cerca i anotació de gens, validació per RACE-PCR d'alguns dels nous gens descrits, alineaments i anàlisis filogenètics)i contribuí significativament a la seva discussió i interpretació.

Article R4

Sebé-Pedrós, A., Ariza-Cosano, A., Weirauch, M. T., Leininger, S., Yang, A., Torruella, G., Adamski, M., Adamska, M., Hughes, T.R., Gómez-Skarmeta, J.L. & Ruiz-Trillo, I. **Early evolution of the T-box transcription factor family.**

Enviat i en revisió en una revista d'alt impacte.

El doctorand va participar activament en la concepció, el disseny experimental i la redacció de l'article, del qual n'és co-primer autor. Va realitzar la totalitat d'anàlisis *in silico* (cerca i anotació de gens, alineaments i anàlisis filogenètics) i va clonar els gens que posteriorment es testarien en el sistema heteròleg, el gripau *Xenopus laevis,* i també el que s'usaria per a realitzar el PBM. Així mateix, va participar en la realització dels experiments a *Xenopus*. Finalment, va contribuir significativament a la discussió i interpretació dels resultats generats.

Article R5

Sebé-Pedrós, A., Zheng, Y., Ruiz-Trillo, I., & Pan, D. (2012). **Premetazoan Origin of the Hippo Signaling Pathway.** Cell Reports, 1(1), 13–20.

Índex d'impacte (2012): pendent.
Posició dins l'àrea: pendent.

El doctorand va participar activament en la concepció, el disseny experimental i la redacció de l'article, del qual n'és co-primer autor. Va realitzar la totalitat d'anàlisis *in silico* (cerca i anotació de gens, alineaments i anàlisis filogenètics) i va clonar els gens que posteriorment es testarien en el sistema heteròleg de *Drosophila melanogaster*. Finalment, va contribuir significativament a la discussió i interpretació dels resultats generats.

Article R6

Sebé-Pedrós, A., Burkhardt, P., Sánchez-Pongs, N., Fairclough, S., Lang, B.F., King, N. & Ruiz-Trillo, I. **Insights into the origin of metazoan filopodia and microvilli.**

Enviat i en revisió en una revista d'alt impacte.

El doctorand va participar activament en la concepció, el disseny experimental i la redacció de l'article, del qual n'és co-primer autor. Va realitzar la totalitat d'anàlisis *in silico* (cerca i anotació de gens, alineaments i anàlisis filogenètics) i la microscòpia electrònica de rastreig, i va participar en la realització de les tècniques d'immunofluorescència així com en els anàlisis de nivells d'expressió gènica a partir de dades de transcriptòmica comparada. Finalment, va contribuir significativament a la discussió i interpretació dels resultats generats.

Article R7

Sebé-Pedrós, A., Irimia, M., del Campo, J., Parra-Acero, H., Russ, C., Haas, B.J., Blencowe, B.J, Nusbaum, C. & Ruiz-Trillo, I. **Transcriptome remodelling during aggregative multicellularity in a close unicellular relative of Metazoa.**

A punt per a enviar.

El doctorand va participar activament en la concepció, el disseny experimental i la redacció de l'article. Va realitzar l'aïllament i caracterització dels diferents estadis (incloent microscòpia electrònica de transmissió i rastreig), la generació de les línies clonals que constituirien les rèpliques biològiques, les extraccions d'RNA dels diferents estadis i rèpliques biològiques per a ésser seqüenciades, així com la totalitat d'anàlisis *in silico* posteriors (quantificació de nivells d'expressió, anàlisis estadístics, anàlisis d'ontologia gènica,...), a excepció de l'anàlisi d'*splicing* alternatiu. En aquest últim cas, va realitzar la totalitat de validacions per RT-PCR de les prediccions d'*splicing*. Finalment, va contribuir significativament a la discussió i interpretació dels resultats generats.

Iñaki Ruiz-Trillo Jaume Baguñà Monjo

Results R1

Ancient origin of the integrin-mediated adhesion and signaling machinery.

RESUM ARTICLE R1: Origen antic de la maquinària d'adhesió i senyalització per integrines.

L'evolució dels animals (metazous) des dels seus ancestres unicel·lulars requerí l'emergència de nous mecanismes d'adhesió i comunicació cel·lular. Un dels principals mecanismes d'adhesió en el desenvolupament animal és la maquinària d'adhesió i senyalització per integrines. El complex d'adhesió d'integrines és essencial per a la interacció entre les cèl·lules i la matriu extracel·lular, modulant diversos aspectes de la fisiologia cel·lular. Fins temps recents, aquesta maquinària s'ha considerat específica del món animal. En aquest article mostrem els resultat d'un anàlisi genòmic comparatiu de la maquinària d'adhesió per integrina, utilitzant noves dades genòmiques de diversos parents unicel·lulars dels animals i dels fongs. De forma inesperada, trobem que els components principals del complex d'adhesió per integrina es troben en el genoma del protozous *Amastigomonas* sp. (en l'actualitat anomenat *Thecamonas trahens*), membre del clade Apusozoa, i, per tant, l'origen d'aquest complex és anterior a la divergència dels Opisthokonta, el clade que inclou animals i fongs. A més, el nostre anàlisi suggereix que diversos elements clau del complex d'integrina es van perdre independentment en coanoflagel·lats i en fongs. Els nostres resultats emfatitzen el fet que molts gens que s'havien considerat claus per a l'origen dels animals tenen, de fet, un origen molt anterior. Això subratlla la importància de la co-opció de gens en la transició cap a la multicel·lularitat que va originar els animals.

Ancient origin of the integrin-mediated adhesion and signaling machinery

Arnau Sebé-Pedrós[a], Andrew J. Roger[b], Franz B. Lang[c], Nicole King[d], and Iñaki Ruiz-Trillo[a,e,1]

[a]Departament de Genètica and Institut de Recerca en Biodiversitat, Universitat de Barcelona, 08028 Barcelona, Spain; [b]Centre for Comparative Genomics and Evolutionary Bioinformatics, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada B3H 1X5; [c]Department of Biochemistry, Université de Montréal, Montréal, QC, Canada H3C 3J7; [d]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720; and [e]Institució Catalana per a la Recerca i Estudis Avançats at Parc Científic de Barcelona, 08028 Barcelona, Spain

Edited* by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, and approved April 28, 2010 (received for review February 24, 2010)

The evolution of animals (metazoans) from their unicellular ancestors required the emergence of novel mechanisms for cell adhesion and cell–cell communication. One of the most important cell adhesion mechanisms for metazoan development is integrin-mediated adhesion and signaling. The integrin adhesion complex mediates critical interactions between cells and the extracellular matrix, modulating several aspects of cell physiology. To date this machinery has been considered strictly metazoan specific. Here we report the results of a comparative genomic analysis of the integrin adhesion machinery, using genomic data from several unicellular relatives of Metazoa and Fungi. Unexpectedly, we found that core components of the integrin adhesion complex are encoded in the genome of the apusozoan protist *Amastigomonas* sp., and therefore their origins predate the divergence of Opisthokonta, the clade that includes metazoans and fungi. Furthermore, our analyses suggest that key components of this apparatus have been lost independently in fungi and choanoflagellates. Our data highlight the fact that many of the key genes that had formerly been cited as crucial for metazoan origins have a much earlier origin. This underscores the importance of gene cooption in the unicellular-to-multicellular transition that led to the emergence of the Metazoa.

cell adhesion | lateral gene transfer | metazoan origins | multicellularity

Little is known about how multicellular animals (metazoans) or fungi evolved from their single-celled or colonial ancestors. Cell adhesion and cell signaling are two important features of the multicellular metazoan lifestyle that were likely critical to the origin of Metazoa (1, 2). Recent data have shown that many of the major metazoan signaling pathways and cell adhesion systems are ubiquitous across the metazoan kingdom, including nonbilaterian lineages [sponges, placozoans, and cnidarians (3–6)]. These findings indicate that cell adhesion and cell signaling genes might have evolved before the origin of Metazoa. Consistent with this view, choanoflagellates, the unicellular putative sister group of Metazoa (7–11), have been shown to possess some genes involved in cell signaling and adhesion, such as tyrosine kinases and cadherins (1, 12–14). Expressed sequence tag surveys of other unicellular relatives of metazoans, such as *Capsaspora owczarzaki* and *Ministeria vibrans*, also yielded homologs of genes involved in metazoan cell adhesion and cell signaling (9, 15).

Here we report a comparative genomic survey of integrin-mediated adhesion machinery, a critical cell–matrix adhesion mechanism in metazoans that also plays a vital role in cell signaling (16–18). Integrin-mediated signaling occurs in two ways: as an "inside-out" signaling modulated through intracellular events, and as "outside-in" signaling that reacts via binding of a ligand to the receptor (17, 19, 20). Thus, integrins are involved in diverse cellular processes, including embryogenesis, cell spreading, cell migration, and proliferation (16–18). However, integrin adhesion and signaling seems to be absent from other multicellular organisms (e.g., plants and fungi) and is generally considered to be metazoan specific (2, 5, 21).

Integrins are heterodimeric transmembrane proteins composed of one α and one β subunit (17). The integrin-mediated process of linking the extracellular matrix to the intracellular actin cytoskeleton is made in concert with several cytoskeletal proteins that form adhesion-triggered signaling complexes (22): α-actinin and talin [both of which directly bind to the integrin β subunit (23–25)]; and paxillin and vinculin [both of which are scaffolding proteins that indirectly bind to integrin-β via talin and α-actinin (26, 27)]. An important element of the integrin adhesion machinery is the heterotrimer IPP complex, which is composed of ILK (integrin-linked kinase), PINCH (particularly interesting Cys-His–rich protein), and parvin (28, 29). This complex plays an important role in integrin-mediated signaling, regulating apoptosis, and cell dynamics (29). Finally, integrin-mediated signaling occurs mainly via two kinases known to be concentrated at the integrin adhesion machinery, namely c-Src tyrosine kinase and FAK (focal adhesion kinase) (22, 30, 31). Many other proteins are indirectly involved with the integrin adhesion complex (32), but here we focus on those most directly involved in the clustering of integrins into the adhesion complex (22).

The recent completion of genome sequences for five close relatives (some strictly unicellular, some colonial) of metazoans and fungi provides the opportunity to reconstruct the evolution of proteins required for integrin-mediated cell adhesion (ref. 33; see also http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html). By examining the genomes of the amoeba *C. owczarzaki*, two basal fungi, *Allomyces macrogynus* and *Spizellomyces punctatus*, the apusozoan *Amastigomonas* sp., and a choanoflagellate, *Proterospongia* sp., we find that the integrin adhesion and signaling machinery evolved in unicellular progenitors of apusozoan protists and opisthokonts (i.e., Fungi, choanoflagellates, and Metazoa). Integrin α and β and several other components of the integrin adhesion complex are absent from choanoflagellates and fungi and were presumably lost independently in these lineages. By comparing genome data from a broad sampling of unicellular taxa, we have been able to clarify the dynamic evolutionary history of the integrin adhesion complex.

Results

Integrins. Outside Metazoa, we found four integrin β and four integrin α genes in *C. owczarzaki*, and one β and one α in *Amastigomonas* sp. Interestingly, we found one of the integrin β domains (the extracellular domain) in the cyanobacterium *Trichodesmium*

Author contributions: A.S.-P. performed research; A.S.-P., F.B.L., and I.R.-T. analyzed data; A.J.R., F.B.L., and N.K. contributed new reagents/analytic tools; I.R.-T. designed research; and A.S.-P., A.J.R., F.B.L., N.K., and I.R.-T. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. GU320672–GU320675).

[1]To whom correspondence should be addressed. E-mail: inaki.ruiz@icrea.es.

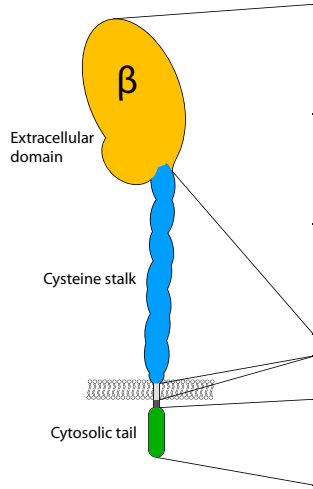
This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1002257107/-/DCSupplemental.

erythraeum (34), which lacks all of the other integrin β domains. Integrins were not detected in any other examined eukaryote. Interestingly, although an integrin α ortholog was thought to be present in the choanoflagellate *M. brevicollis* (12), we failed to detect a bona fide integrin α in either *Monosiga brevicollis* or *Proterospongia* sp. The putative integrin α from *M. brevicollis* (XP_001749484) did not pass any of our criteria (for example, reverse blast did not give integrin α hits; *Methods*). The *M. brevicollis* gene XP_001749484 shares with integrin α homologs the presence of some FG-GAP repeats domains, which are not specific to integrin α and are found in other nonintegrin proteins. A phylogeny made from FG-GAP repeats shows the *M. brevicollis* putative integrin α homolog clustering with nonintegrin bacterial proteins but not with integrin α (Fig. S1). On the other hand, our phylogenetic analysis of integrin β (Fig. S2) shows the four *C. owczarzaki* integrins clustering together with a bootstrap value (BV) of 60% (BV = 95% if the most divergent *C. owczarzaki* integrin β homolog is deleted from the analysis). The integrin β homologs of *Amastigomonas* sp. and *T. erythraeum* group together, with a BV of 85%. We were unable to recover any other integrin adhesion complex components in *Trichodesmium erythraeum*, and no other sequenced bacterial genome encodes the integrin β domain or any other component of the integrin adhesion complex.

We next analyzed whether β integrins from *C. owczarzaki* and *Amastigomonas* sp. and the integrin β extracellular domain from *T. erythraeum* have conserved the functional domains and motifs present in metazoan integrins (Fig. 1 and Fig. S3). The cation-binding motifs MIDAS, ADMIDAS, and LIMB, which are located in the extracellular domain (35, 36), are well conserved in the different nonmetazoan integrin β, except for *C. owczarzaki* integrin β4 (Fig. 1*A*). Moreover, *C. owczarzaki* integrin β1, β2, and β3 and *Amastigomonas* sp. integrin β have a clear expansion of the cysteine-rich stalk (Fig. 1*A* and Fig. S3), which accounts for their longer size relative to metazoan integrin β proteins. Other key motifs in metazoan integrin β proteins are the cytoplasmic integrin α-interacting motif and the NPXY motif, which plays a key role in protein interactions (17, 20, 37, 38). Both motifs are well conserved in *C. owczarzaki* integrin β1–β3 and *Amastigomonas* sp. integrin β (Fig. 1*A*). Finally, both *C. owczarzaki* and *Amastigomonas* sp. integrin β have predicted signal peptides and transmembrane domains.

We also examined the evolutionary conservation of nonmetazoan integrin α (Fig. 1*B*). Metazoan integrin α homologs typically have large extracellular regions with seven FG-GAP repeats that form a β propeller structure (36), with three DXD/NXD/NXXXD cation-binding motifs in the last three FG-GAP repeats (39). One specific and diagnostic feature of integrin α is

A Integrin β



Motif/Function	Metazoa	*Capsaspora β1*	*Capsaspora β2*	*Capsaspora β3*	*Capsaspora β4*	*Amastigomonas*	*Trichodesmium*
MIDAS Mg^{2+} and Mn^{2+} cations binding.	D	D^{119}	D^{123}	D^{238}	D^{128}	D^{115}	D^{59}
	S	S^{121}	S^{125}	S^{240}	S^{130}	S^{117}	S^{61}
	S	S^{123}	S^{127}	T^{242}	G^{132}	S^{119}	S^{63}
	E	E^{232}	E^{220}	N^{342}	N^{220}	E^{214}	E^{142}
	D	D^{268}	D^{248}	D^{370}	D^{252}	D^{245}	D^{173}
ADMIDAS Ca^{2+} cations binding.	S	S^{123}	S^{127}	T^{242}	G^{132}	S^{119}	S^{63}
	D	D^{126}	D^{130}	D^{245}	S^{135}	D^{122}	D^{66}
	D	D^{127}	D^{131}	D^{246}	D^{136}	D^{123}	D^{67}
	A	A^{357}	A^{338}	A^{454}	A^{338}	F^{330}	F^{246}
	D	D^{268}	D^{248}	D^{370}	D^{252}	D^{245}	D^{173}
LIMB Ca^{2+} cations binding.	D/E	E^{169}	D^{170}	E^{288}	E^{167}	D^{155}	D^{98}
	N	N^{227}	N^{216}	N^{338}	D^{215}	N^{209}	N^{138}
	D	D^{229}	D^{217}	D^{339}	D^{217}	D^{211}	D^{139}
	P	P^{231}	P^{219}	P^{341}	P^{219}	P^{213}	P^{141}
	E	E^{232}	E^{220}	N^{342}	N^{220}	E^{214}	E^{142}
CXCXXCXC	3	9	6	7	2	7	
KLLXXX**D** Interaction with integrin α.	1	1	1	1	0	1	
NPXY Interaction with intracellular scaffolding and signaling proteins.	1-2	1	1	1	0	2	

B Integrin α

Motif/Position/Function	Metazoa	*Capsaspora α1*	*Capsaspora α2*	*Capsaspora α3*	*Capsaspora α4*	*Amastigomonas*
DXD/NXD/NXXXD Extracellular. Cations binding.	3	3	3	3	1	3
KXG**FFXR** Cytosolic tail. Interaction with integrin β.	1	1	1	1	0	1

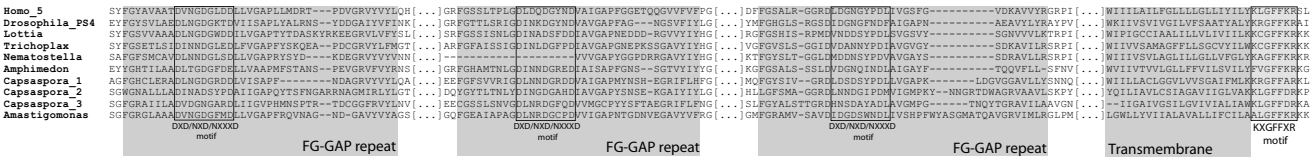


Fig. 1. Comparison of the functional domains and amino acid motifs between canonical metazoan and nonmetazoan integrins. (*A*) Integrin β amino acid motifs and (*B*) amino acid motifs and schematical alignment of integrin α (20, 35–39). Integrin β is divided into the integrin β extracellular domain (orange), integrin stalk (blue), interaction motif (gray), and cytoplasmic tail (green). The *T. erythraeum* homolog possesses only the integrin β domain. For the MIDAS, ADMIDAS, and LIMB motifs within the integrin β domain, we indicate the canonical metazoan amino acids and their positions within the whole protein. In red are shown nonconserved position and motifs. For the remaining motifs, the number of motifs present is indicated. Integrin α is divided into the three FG-GAP cation binding motifs and the integrin α–β interacting motif after the transmembrane domain. Key amino acids experimentally determined in the α–β interacting motifs are depicted in bold (17).

EVOLUTION

the short cytoplasmic tail that contains a KXGFFXR motif that interacts with integrin β. A signal peptide and a transmembrane domain are also typically found in integrin α homologs. These motifs are conserved, with some minor modifications, in the integrin α homologs of *C. owczarzaki* and *Amastigomonas* sp., but not in *C. owczarzaki* integrin α4, which has only two of the three cation-binding motifs and does not have a predicted signal peptide (Fig. 1*B*).

Scaffolding Proteins. Our investigations show that all scaffolding proteins involved in the integrin adhesion apparatus (that is, α-actinin, vinculin, paxillin, and talin) are common among unikonts (i.e., Opisthokonts+Amoebozoa; Figs. 2 and 3). Phylogenetic analyses of these proteins show, in general, topologies in agreement with organismal phylogeny (Figs. S4*A* and S5 *A* and *B*), except for paxillin, which did not have enough phylogenetic signal to recover a statistically significant topology.

IPP Complex. A complete IPP complex with all three components is only present in Metazoa, *C. owczarzaki*, the chytrid fungus *Batrachochytrium dendrobatidis*, and the apusozoan *Amastigomonas* sp. (Figs. 2 and 3). In the *Amastigomonas* sp. genome we found a partial gene encoding the N-terminal part of the ILK protein, which is composed of the three consecutive ankyrin repeats but failed to find a characteristic C terminus, which is a Ser/Thr kinase domain. Phylogenetic inference based on an alignment of the three ankyrin repeats shows that the putative ILK of *Amastigomonas* sp. branches within the canonical ILK homologs (Fig. S5*C*). Because the genome coverage of *Amastigomonas* sp. is at present still low, it is possible that the C-terminal part of *Amastigomonas* sp. ILK homolog is indeed present but not represented in the current assembly. In any case, the three components of the IPP complex are missing in both of the choanoflagellates, *M. brevicollis* and *Proterospongia* sp., and fungi other than *B. dendrobatidis*. Interestingly, *A. macrogynus* and *S. punctatus* possess just one component (PINCH and ILK, respectively) of the IPP complex (Figs. 2 and 3). A phylogenetic tree of ILK and several related kinases estimated from an alignment of the kinase domain alone shows that *C. owczarzaki*, *S. punctatus*, and *B. dendrobatidis* ILKs are related to metazoan ILKs (Fig. S5*D*). Similarly, phylogenetic trees of parvin and PINCH show a topology in agreement with organismal phylogeny (Fig. S6).

c-Src Tyrosine Kinase and FAK. Our searches show that c-Src is present in Metazoa, choanoflagellates, and *C. owczarzaki* (Fig. 2). The phylogenetic analysis of this protein family, which includes Abl kinases as an outgroup, shows that both choanoflagellates and *C. owczarzaki* c-Src tyrosine kinases group with metazoan ones (Fig. S4*B*). On the other hand, bona fide FAK are only present in Metazoa and *C. owczarzaki* (Fig. 2). *C. owczarzaki* FAK have all of the functional domains involved in its protein–protein interactions (31) (Fig. S7*A*). Interestingly, *M. brevicollis* has a gene encoding a tyrosine kinase domain that, by phylogenetic analysis, seems to be related to FAK (Fig. S7*B*), even though the predicted protein does not have the canonical domain structure of FAK.

Discussion

Our analyses show that the integrin-mediated cell adhesion machinery is not specific to metazoans, as previously thought (2, 5, 21). We found that the apusozoan *Amastigomonas* sp. has the integrin adhesion machinery, including all of the components of the canonical metazoan complex, except for the signaling molecules FAK and c-Src. Recent multigene analyses suggest that apusozoans are related to opisthokonts, most likely falling outside of this clade as their nearest sister group (8, 40, 41). However, they have also been proposed to be sister group to amoebozoans or represent a deeper-branching eukaryotic lineage, although these proposals derive from single-gene and statistically weakly supported phylogenies (see ref. 41 for a discussion). In any case, apusozoans clearly fall outside opisthokonts in all multigene phylogenetic analyses, and they do not share the characteristic translation elongation factor 1-α (EF1-α) insertion, a synapomorphy unique to the opisthokont lineages (see ref. 8 for *Ancyromonas* and *Apusomonas* and Fig. S8 for *Amastigomonas*). Regardless of whether apusozoans are (*i*) sister group to opisthokonts (40–42), (*ii*) sister group to amoebozans (41), or (*iii*) a deep eukaryotic lineage (41), our conclusion that many of the components of the integrin adhesome evolved well before the origin of Metazoa and Fungi is still valid. If apusozoans are sister group to amoebozoans or a deeper eukaryotic lineage, then these core integrin compo-

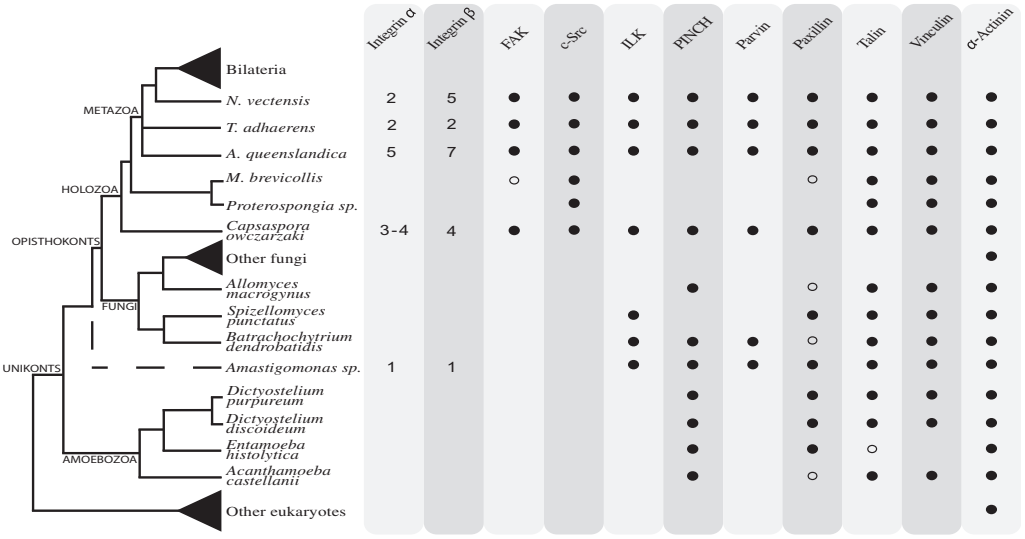


Fig. 2. Schematic representation of the eukaryotic tree of life showing the distribution of the different components of the integrin adhesion complex. The number of integrin homologs is shown. A black dot indicates the presence of clear homologs, whereas a hollow dot indicates the presence of putative or degenerate homologs. Absence of a dot indicates that a homolog is lacking in that taxon. The phylogenetic relationships are based in several recent phylogenetic studies (8, 9, 15, 40, 64, 65).

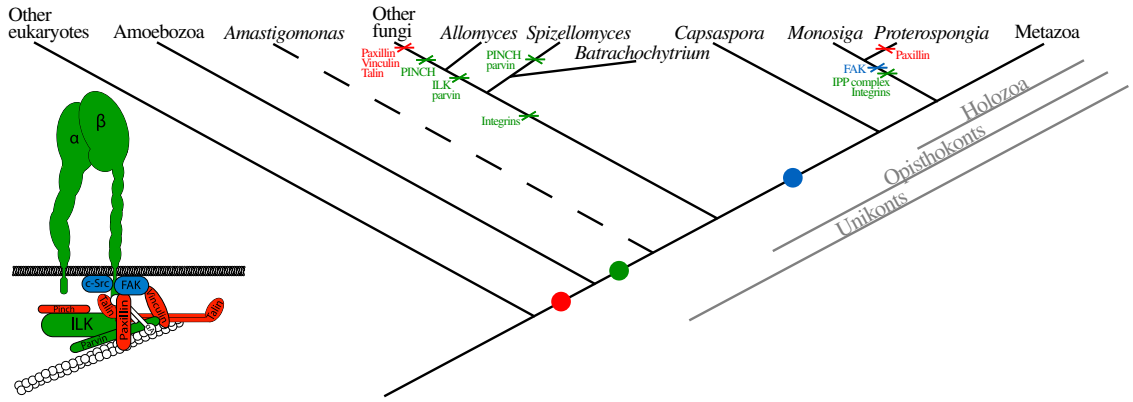


Fig. 3. Schematic representation of integrin-mediated cell-adhesion and cell-signaling evolution. *Left:* The canonical metazoan integrin adhesion complex. The colors correspond to the three main steps in the evolution of the integrin adhesion mechanism, as shown in the cladogram. Dots indicate origin, and crosses indicate losses. The branch leading to *Amastigomonas* sp. is shown dashed, because its phylogenetic position remains unresolved (see main text for discussion).

nents must also have been secondarily lost in the amoebozoan taxa whose genomes have been characterized to date.

We have also shown that *C. owczarzaki*, a specific sister group to choanoflagellates and Metazoa (9, 15, 43), has a canonical metazoan-type integrin adhesion and signaling machinery with a full repertoire of integrin adhesion complex components. Therefore, the canonical metazoan-type integrin adhesion machinery is probably specific to holozoans; that is, it originated before the divergence of *C. owczarzaki* from choanoflagellates+Metazoa, but likely after the Fungi+nucleariid+fonticulid clade had split from Holozoa (Fig. 3). Another possible scenario is that a canonical metazoan-type integrin adhesion machinery was present in the ancestor of apusozoans and opisthokonts, and both FAK and c-Src were subsequently lost within the apusozoan lineage.

A major implication of our taxon-rich comparative genomic survey is the emergence of a more complex scenario for the evolution of the integrin adhesion complex (Figs. 2 and 3). Under this scenario, which should be further tested with genome data from additional eukaryotic lineages, it is evident that several independent losses and diversifications of main components of the integrin adhesome have occurred over the course of evolution. For example, integrin α and β homologs seem to be absent from choanoflagellates and fungi. In fact, each choanoflagellate and fungal taxon we examined harbors a distinctive repertoire of integrin adhesome components resulting from different lineage-specific losses. This is most obvious in Fungi, where the loss of the IPP complex seems to be gradual, with the chytrid fungi taxa retaining all or some of the IPP components despite their lack of integrins. Specifically, *B. dendrobatidis* has the full IPP complex (ILK, PINCH, and parvin), whereas *S. punctatus* has just one of the components (ILK), and *A. macrogynus* has just PINCH. It is unclear what cellular functions the IPP complex components present in *B. dendrobatidis*, *S. punctatus*, and *A. macrogynus* might have in the absence of integrin subunits.

Of major interest for the origin of metazoans is the fact that choanoflagellates, which are the closest sister group of Metazoa (7, 8, 10, 11), have also lost many of the integrin components. Specifically, the two choanoflagellates analyzed here, *M. brevicollis* and *Proterospongia* sp., lack both integrin β and α, the full IPP complex, and one of the signaling molecules involved in the integrin adhesome, FAK (although *M. brevicollis* has a protein with a FAK-related tyrosine kinase domain; Fig. S7). Choanoflagellates do have c-Src, but they act in a different context than that of integrin adhesion, as recently demonstrated experimentally in *M. brevicollis* (44). Moreover, lineage-specific diversifications (independent of those occurring in metazoans) of both integrin α and β have occurred within the *C. owczarzaki* lineage. What roles these various homologs play in *C. owczarzaki* biology remains to be determined.

It is possible that functional differences between metazoan and nonmetazoan integrins exist that would explain their conservation in unicellular vs. multicellular contexts. Our analysis of the functional domains shows that both metazoan and nonmetazoan integrins are quite similar, the only difference being the longer size of the protein in nonmetazoan ones (Fig. 1 and Fig. S3). More importantly, both integrin α and β in *C. owczarzaki* and in *Amastigomonas* sp. possess all of the critical interacting amino acid motifs in their cytoplasmic tails (Fig. 1). Thus, we can assume that they too work as heterodimers and that they interact and function similarly to metazoan homologs. Functional analysis will be needed to test this hypothesis.

Many of the scaffolding proteins (talin, vinculin, paxillin) most likely evolved in the common ancestor of amoebozoans and opisthokonts, where they had ancestrally different functions (as in present day amoebozoans). They were coopted to further work as part of a metazoan-specific integrin adhesome (i.e., their presence in opisthokonts and in amoeboans should not be interpreted as a signature of an ancient integrin-mediated adhesion apparatus) (Fig. 3). For example, it has been shown that the talin homolog of the amoebozoan *D. discoideum* interacts with an NPXY motif (the same motif found in integrin β) of the cytoplasmic tail of an adhesion molecule called SibA (45). Thus, it is possible that an ancestral integrin β independently acquired an NPXY motif allowing it to recruit talin.

Our comparative genomic study not only deciphers the evolutionary history of the integrin adhesome, but it also highlights the importance of a broad taxonomic sampling in these kinds of studies. In particular, a broader taxonomic sampling within nonbilaterian metazoans was key for the realization that many key genes in bilaterian development are indeed present in triploblastic metazoans (3–6, 46–49). Similarly, genome data from unicellular metazoan-related lineages is pushing back the times of origin of many gene families formerly believed to be metazoan specific to well into the Proterozoic. Such is the case, for example, of tyrosine kinases (14, 50, 51), some transcription factors (12, 52), membrane-associated guanylate kinases (53), or cadherines (13). The integrin-mediated signaling and adhesion machinery here presented add another striking example to this pattern and suggest that some of these protein families may have emerged even earlier in eukaryote evolution before the divergence of opisthokonts. Investigation of a variety of additional genomes from unicellular opisthokonts and other more distantly related protistan

lineages will be required to more precisely pinpoint the origins of these systems in early eukaryote evolution.

Integrin β in the Cyanobacterium *T. erythraeum* Is Derived from a Lateral Gene Transfer Event. Our search revealed the presence of a gene encoding an incomplete integrin β in the *T. erythraeum*. We believe the most plausible scenario to explain this observation is an interdomain lateral gene transfer (LGT) event in the eukaryote-to-prokaryote direction, because integrin β is present in many eukaryote taxa but only in a single known prokaryotic genome. The lack of introns in the *Amastigomonas* integrin β (in contrast to the other integrins described herein) may have facilitated its integration into a cyanobacterial genome as would the property of natural competence (i.e., the ability to take up DNA) known in Cyanobacteria (54). Even though eukaryote-to-prokaryote LGT events are not as common as LGTs in the opposite direction, other cases have been described in *T. erythraeum* (55).

Conclusions

We have demonstrated that a near-complete integrin adhesion complex had evolved in a unicellular common ancestor of metazoans and fungi and still exists in the Apusozoa, the putative sister group to opisthokonts. Furthermore, we have shown that the origins of most of the scaffolding elements of current integrin adhesion complex predate the origins of integrin proteins themselves, suggesting that an ancient scaffolding machinery was coopted to the integrin adhesion system. Moreover, the origin of the IPP complex probably represented one of the first signaling modules, coupling the integrin adhesion machinery with cell signaling to control cell behavior. Novel signaling systems based on tyrosine kinases appeared at a later stage, most likely within holozoans. Another implication of our analyses is that lineage-specific diversifications and lineage-specific losses have played a major role in the evolution of the integrin adhesome in opisthokonts. For example, both fungi and choanoflagellates have lost several important components of the integrin adhesion complex from their ancestors. Finally, from our study and that done by Abedin and King (13), we can conclude that the major cell–cell and cell–matrix adhesion mechanisms in metazoans, those mediated by cadherins and integrins, respectively, have a deeper evolutionary origin than previously thought. This adds to the growing evidence that major cell signaling and cell adhesion pathways crucial to metazoan development were present in premetazoan lineages (12, 50, 51, 53). Thus, the answers to what triggered the unicellular-to-multicellular transition that gave rise to metazoans may lie not only in the acquisition of new genes but also in the cooption of ancestral proteins into new functions and the evolution of more complex interactions.

Methods

Gene Searches. We performed searches for the two integrin subunits (α and β) plus all of the other proteins that are directly involved in the integrin-mediated adhesion and signaling complex (see the Introduction). Those proteins include α-actinin, vinculin, talin, paxillin, ILK, PINCH, parvin, FAK, and c-Src. A primary search to collect putative initial candidates was performed using the basic local alignment sequence tool (BLAST: blastp and tblastn) using *Homo sapiens* integrin adhesion proteins as queries and an e-value threshold of 10^{-05}. We blasted against completed or ongoing genome project databases at the National Center for Biotechnology Information (NCBI), the Joint Genome Institute, and the Broad Institute (see Fig. 2 for a list of the taxa considered), as well as against the *Amphimedon queenslandica* protein and genome database (Dr. Bernard M. Degnan). *C. owczarzaki* and *S. punctatus* genome assemblies and annotations are available at the Broad Institute Web site (http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html). In the case of *Proterospongia* sp., *Amastigomonas* sp., *A. macrogynus*, and *Acanthamoeba castellanii*, we assembled the trace data using the WGS assembler. We then annotated the genes of interest using both Genomescan (56) and Augustus (57) and performed local BLAST searches against both annotations. Assemblies and annotations for these taxa are available upon request (*Appendix S1*).

When the BLAST searches of genome data described above returned significant "hits", the sequences obtained were then reciprocally searched against the NCBI protein database by BLAST to confirm the validity of the sequences retrieved with the initial search (58). To identify distant homologs that might have escaped these simple searches, two additional methods were used. The same BLAST search was repeated using homologs from nonmetazoan taxa, such as *Dictyostelium*, *Capsaspora*, or *Amastigomonas*, as queries instead of *Homo* sequences. Additionally, for integrin β, integrin α, vinculin, talin, and FAK, we performed protein domain searches using HMMER3.0b2 (59) against the same genome databases, plus six-frame translations of all studied genomes. Finally, we checked the protein domain structure of all putative positives by searching the Pfam (http://pfam.sanger.ac.uk/search) and SMART (http://smart.embl-heidelberg.de/) databases. Signal peptides were identified using the SignalP 3.0 Server (60).

Confirmation of *C. owczarzaki* Integrin β by PCR. We confirmed the presence of integrin β in *C. owczarzaki* by RT-PCR and 3′ RACE PCR. mRNA was extracted using a Dynabeads mRNA purification kit (Invitrogen), and subsequent RT-PCR was performed using SuperScript III First Strand Synthesis kit (Invitrogen). The full sequences of the 5′ and 3′ ends of the four distinct *C. owczarzaki* integrin β cDNAs were obtained by RACE, using nested PCR with primers designed from initial analyses of the genome data. Both coding and noncoding strands were sequenced using an ABI Prism BigDye Termination Cycle Sequencing Kit (Applied Biosystems). New sequences were deposited in GenBank under accession nos. GU320672–GU320675.

Phylogenetic Analyses. Alignments were constructed for all proteins using the Muscle (61) plug-in of Geneious software (Biomatters), which were then manually inspected and edited. Only those species and those positions that were unambiguously aligned were included in the final phylogenetic analyses. Maximum likelihood (ML) phylogenetic trees were estimated by RaxML (62) using the PROTGAMMAWAGI model, which uses the Whelan and Goldman amino acid exchangeabilities and accounts for among-site rate variation with a four-category discrete gamma approximation and a proportion of invariable sites (WAG+ Γ+I). Statistical support for bipartitions was estimated by performing 100-bootstrap replicates using RaxML and the same model.

Bayesian analyses were performed with MrBayes 3.1 (63), using the WAG+Γ+I model of evolution, with four chains, a subsampling frequency of 100, and two parallel runs. Runs were stopped when the average SD of split frequencies of the two parallel runs was <0.01, usually around 1,000,000 generations. The two LnL graphs were checked and an appropriate burn-in length established; stationarity of the chain typically occurred after ≈15% of the generations. Bayesian posterior probabilities were used for assessing the confidence values of each bipartition.

ACKNOWLEDGMENTS. We thank the Joint Genome Institute (JGI), the Broad Institute, and the Baylor College of Medicine (BCM) for making data publicly available; Bernard Degnan for access to the *A. queenslandica* genome data; Jason Stajich for sharing unpublished *B. dendrobatidis* genome data; Kim C. Worley and the team of the *A. castellanii* genome project for accession to the genome data; Manuel Palacín, Romain Derelle, Alex de Mendoza, and Hiroshi Suga for helpful insights; and other members of UNICORN, Gertraud Burger, Michael W. Gray, and Peter W. H. Holland. Preliminary sequence data were obtained from the JGI, the Broad Institute, BCM, and National Center for Biotechnology Information Web sites. The genome sequences of *C.owczarzaki*, *A. macrogynus*, *S. punctatus, Amastigomonas* sp., and *Proterospongia* sp. are being determined by the Broad Institute of Massachusetts Institute of Technology/Harvard University under the auspices of the National Human Genome Research Institute and within the UNICORN initiative. This work was supported by an Institució Catalana per a la Recerca i Estudis Avançats contract, European Research Council Starting Grant 206883, and Grant BFU2008-02839/BMC from Ministerio de Ciencia e Innovación (MICINN) (to I.R.-T.). A.S.-P.'s salary was supported by a pregraduate Formación de Personal Universitario grant from MICINN. A.J.R.'s contribution was supported by Grant MOP 62809 from the Canadian Institutes of Health Research, and B.F.L.'s contribution by the Canadian Research Chair Program.

1. King N (2004) The unicellular ancestry of animal development. *Dev Cell* 7:313–325.
2. Rokas A (2008) The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annu Rev Genet* 42:235–251.
3. Putnam NH, et al. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317:86–94.
4. Srivastava M, et al. (2008) The Trichoplax genome and the nature of placozoans. *Nature* 454:955–960.

5. Nichols SA, Dirks W, Pearse JS, King N (2006) Early evolution of animal cell signaling and adhesion genes. *Proc Natl Acad Sci USA* 103:12451–12456.
6. Larroux C, et al. (2008) Genesis and expansion of metazoan transcription factor gene classes. *Mol Biol Evol* 25:980–996.
7. Lang BF, O'Kelly C, Nerad T, Gray MW, Burger G (2002) The closest unicellular relatives of animals. *Curr Biol* 12:1773–1778.
8. Steenkamp ET, Wright J, Baldauf SL (2006) The protistan origins of animals and fungi. *Mol Biol Evol* 23:93–106.
9. Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF (2008) A phylogenomic investigation into the origin of metazoa. *Mol Biol Evol* 25:664–672.
10. Ruiz-Trillo I, Lane CE, Archibald JM, Roger AJ (2006) Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. *J Eukaryot Microbiol* 53:1–6.
11. Carr M, Leadbeater BS, Hassan R, Nelson M, Baldauf SL (2008) Molecular phylogeny of choanoflagellates, the sister group to Metazoa. *Proc Natl Acad Sci USA* 105:16641–16646.
12. King N, et al. (2008) The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451:783–788.
13. Abedin M, King N (2008) The premetazoan ancestry of cadherins. *Science* 319: 946–948.
14. Suga H, et al. (2008) Ancient divergence of animal protein tyrosine kinase genes demonstrated by a gene family tree including choanoflagellate genes. *FEBS Lett* 582: 815–818.
15. Shalchian-Tabrizi K, et al. (2008) Multigene phylogeny of choanozoa and the origin of animals. *PLoS ONE* 3:e2098.
16. Hynes RO (1992) Integrins: Versatility, modulation, and signaling in cell adhesion. *Cell* 69:11–25.
17. Hynes RO (2002) Integrins: Bidirectional, allosteric signaling machines. *Cell* 110: 673–687.
18. Harburger DS, Calderwood DA (2009) Integrin signalling at a glance. *J Cell Sci* 122: 159–163.
19. O'Toole TE, et al. (1994) Integrin cytoplasmic domains mediate inside-out signal transduction. *J Cell Biol* 124:1047–1059.
20. Dedhar S, Hannigan GE (1996) Integrin cytoplasmic interactions and bidirectional transmembrane signalling. *Curr Opin Cell Biol* 8:657–669.
21. Whittaker CA, Hynes RO (2002) Distribution and evolution of von Willebrand/integrin A domains: Widely dispersed domains with roles in cell adhesion and elsewhere. *Mol Biol Cell* 13:3369–3387.
22. LaFlamme SE, Auer KL (1996) Integrin signaling. *Semin Cancer Biol* 7:111–118.
23. Sjöblom B, Salmazo A, Djinović-Carugo K (2008) Alpha-actinin structure and regulation. *Cell Mol Life Sci* 65:2688–2701.
24. Wegener KL, et al. (2007) Structural basis of integrin activation by talin. *Cell* 128: 171–182.
25. Critchley DR (2009) Biochemical and structural properties of the integrin-associated cytoskeletal protein talin. *Annu Rev Biophys* 38:235–254.
26. Ziegler WH, Liddington RC, Critchley DR (2006) The structure and regulation of vinculin. *Trends Cell Biol* 16:453–460.
27. Deakin NO, Turner CE (2008) Paxillin comes of age. *J Cell Sci* 121:2435–2444.
28. Legate KR, Montañez E, Kudlacek O, Fässler R (2006) ILK, PINCH and parvin: The tIPP of integrin signalling. *Nat Rev Mol Cell Biol* 7:20–31.
29. Nikolopoulos SN, Turner CE (2001) Integrin-linked kinase (ILK) binding to paxillin LD1 motif regulates ILK localization to focal adhesions. *J Biol Chem* 276:23499–23505.
30. Arias-Salgado EG, et al. (2003) Src kinase activation by direct interaction with the integrin beta cytoplasmic domain. *Proc Natl Acad Sci USA* 100:13298–13302.
31. Parsons JT, Martin KH, Slack JK, Taylor JM, Weed SA (2000) Focal adhesion kinase: A regulator of focal adhesion dynamics and cell movement. *Oncogene* 19:5606–5613.
32. Zaidel-Bar R (2009) Evolution of complexity in the integrin adhesome. *J Cell Biol* 186: 317–321.
33. Ruiz-Trillo I, et al. (2007) The origins of multicellularity: A multi-taxon genome initiative. *Trends Genet* 23:113–118.
34. Johnson MS, Lu N, Denessiouk K, Heino J, Gullberg D (2009) Integrins during evolution: Evolutionary trees and model organisms. *Biochim Biophys Acta* 1788: 779–789.
35. Valdramidou D, Humphries MJ, Mould AP (2008) Distinct roles of beta1 metal ion-dependent adhesion site (MIDAS), adjacent to MIDAS (ADMIDAS), and ligand-associated metal-binding site (LIMBS) cation-binding sites in ligand recognition by integrin alpha2beta1. *J Biol Chem* 283:32704–32714.
36. Xiong JP, et al. (2001) Crystal structure of the extracellular segment of integrin alpha Vbeta3. *Science* 294:339–345.
37. Brower DL, Brower SM, Hayward DC, Ball EE (1997) Molecular evolution of integrins: Genes encoding integrin beta subunits from a coral and a sponge. *Proc Natl Acad Sci USA* 94:9182–9187.
38. Tahiliani PD, Singh L, Auer KL, LaFlamme SE (1997) The role of conserved amino acid motifs within the integrin beta3 cytoplasmic domain in triggering focal adhesion kinase phosphorylation. *J Biol Chem* 272:7892–7898.
39. Knack BA, et al. (2008) Unexpected diversity of cnidarian integrins: Expression during coral gastrulation. *BMC Evol Biol* 8:136.
40. Brown MW, Spiegel FW, Silberman JD (2009) Phylogeny of the "forgotten" cellular slime mold, *Fonticula alba*, reveals a key evolutionary branch within Opisthokonta. *Mol Biol Evol* 26:2699–2709.
41. Kim E, Simpson AG, Graham LE (2006) Evolutionary relationships of apusomonads inferred from taxon-rich analyses of 6 nuclear encoded genes. *Mol Biol Evol* 23: 2455–2466.
42. Cavalier-Smith T, Chao EE (2003) Phylogeny of choanozoa, apusozoa, and other protozoa and early eukaryote megaevolution. *J Mol Evol* 56:540–563.
43. Ruiz-Trillo I, Inagaki Y, Davis LA, Sperstad S, Landfald B, Roger AJ (2004) *Capsaspora owczarzaki* is an independent opisthokont lineage. *Curr Biol* 14(22):R946–947.
44. Li W, Young SL, King N, Miller WT (2008) Signaling properties of a non-metazoan Src kinase and the evolutionary history of Src negative regulation. *J Biol Chem* 283: 15491–15501.
45. Cornillon S, et al. (2006) An adhesion molecule in free-living *Dictyostelium amoebae* with integrin beta features. *EMBO Rep* 7:617–621.
46. Simionato E, et al. (2007) Origin and diversification of the basic helix-loop-helix gene family in metazoans: Insights from comparative genomics. *BMC Evol Biol* 7:33.
47. Gauthier M, Degnan BM (2008) The transcription factor NF-kappaB in the demosponge *Amphimedon queenslandica*: Insights on the evolutionary origin of the Rel homology domain. *Dev Genes Evol* 218:23–32.
48. Technau U, et al. (2005) Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. *Trends Genet* 21:633–639.
49. Yamada A, Pang K, Martindale MQ, Tochinai S (2007) Surprisingly complex T-box gene complement in diploblastic metazoans. *Evol Dev* 9:220–230.
50. Pincus D, Letunic I, Bork P, Lim WA (2008) Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. *Proc Natl Acad Sci USA* 105:9680–9684.
51. 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 USA* 105:9674–9679.
52. Degnan BM, Vervoort M, Larroux C, Richards GS (2009) Early evolution of metazoan transcription factors. *Curr Opin Genet Dev* 19:591–599.
53. de Mendoza A, Suga H, Ruiz-Trillo I (2010) Evolution of the MAGUK protein gene family in premetazoan lineages. *BMC Evol Biol* 10:93.
54. Johnsborg O, Eldholm V, Håvarstein LS (2007) Natural genetic transformation: Prevalence, mechanisms and function. *Res Microbiol* 158:767–778.
55. Layton BE, et al. (2008) Collagen's triglycine repeat number and phylogeny suggest an interdomain transfer event from a Devonian or Silurian organism into *Trichodesmium erythraeum*. *J Mol Evol* 66:539–554.
56. Yeh RF, Lim LP, Burge CB (2001) Computational inference of homologous gene structures in the human genome. *Genome Res* 11:803–816.
57. Stanke M, et al. (2006) AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res* 34(Web Server issue):W435–W439.
58. Moreno-Hagelsieb G, Latimer K (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* 24:319–324.
59. Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14:755–763.
60. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340:783–795.
61. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
62. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
63. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
64. Minge MA, et al. (2008) Evolutionary position of breviate amoebae and the primary eukaryote divergence. *Proc Biol Sci* 276:597–604.
65. Liu Y, et al. (2009) Phylogenomic analyses predict sistergroup relationship of nucleariids and fungi and paraphyly of zygomycetes with significant support. *BMC Evol Biol* 9:272.

Supporting Information

Sebé-Pedrós et al. 10.1073/pnas.1002257107

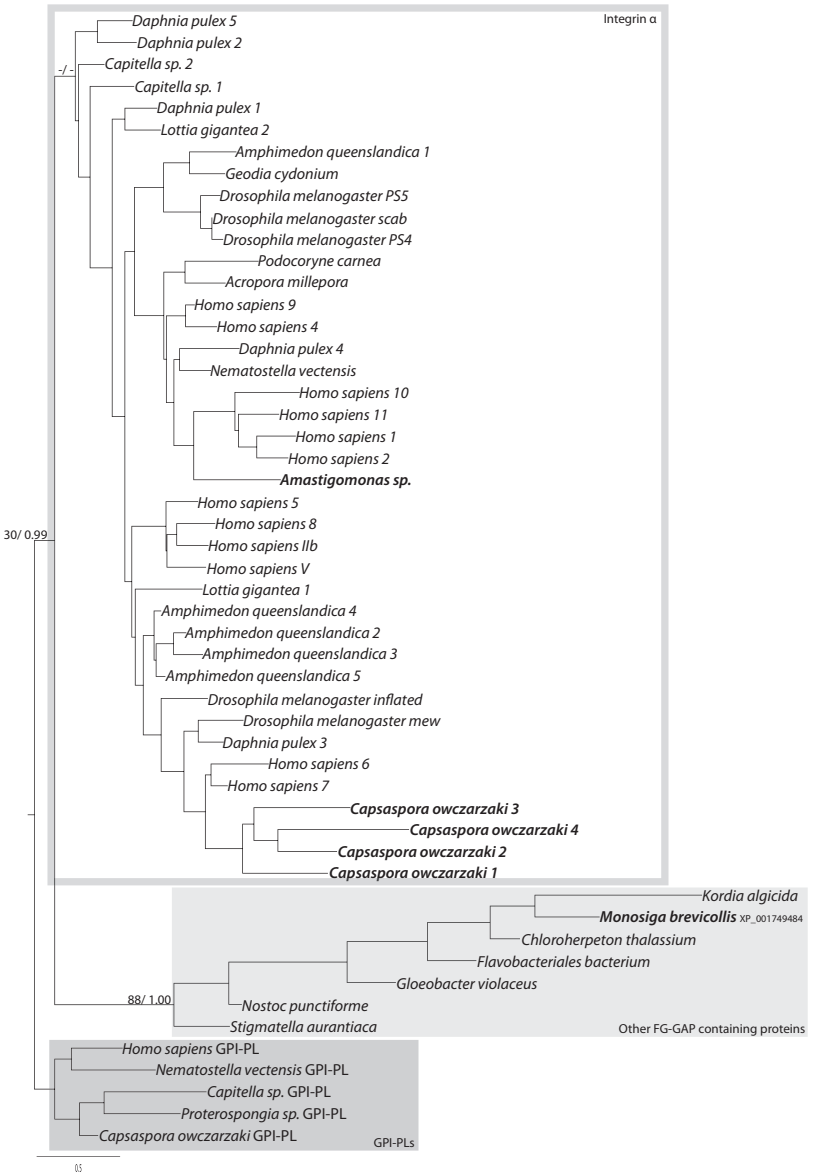


Fig. S1. Maximum likelihood tree of the integrin α homolog and other proteins containing FG-GAP repeats. Alignment has been done using the only common region between all these proteins, which are three consecutive FG-GAP repeats. The taxon sampling includes all of the integrin α homologs here described and a wide representation of metazoan homologs. The putative integrin α from *Monosiga brevicollis* and some FG-GAP repeat-containing bacteria proteins obtained when blasting the *M. brevicollis* sequence have also been included, together with several glycosylphosphatidylinositol phospholipases. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 1,000-bootstrap replicates (bootstrap value, BV) and Bayesian posterior probabilities. BV values are <50% for most branches. Both values are only shown for the some external key branches. The general topology is the same for Bayesian and maximum likelihood analyses.

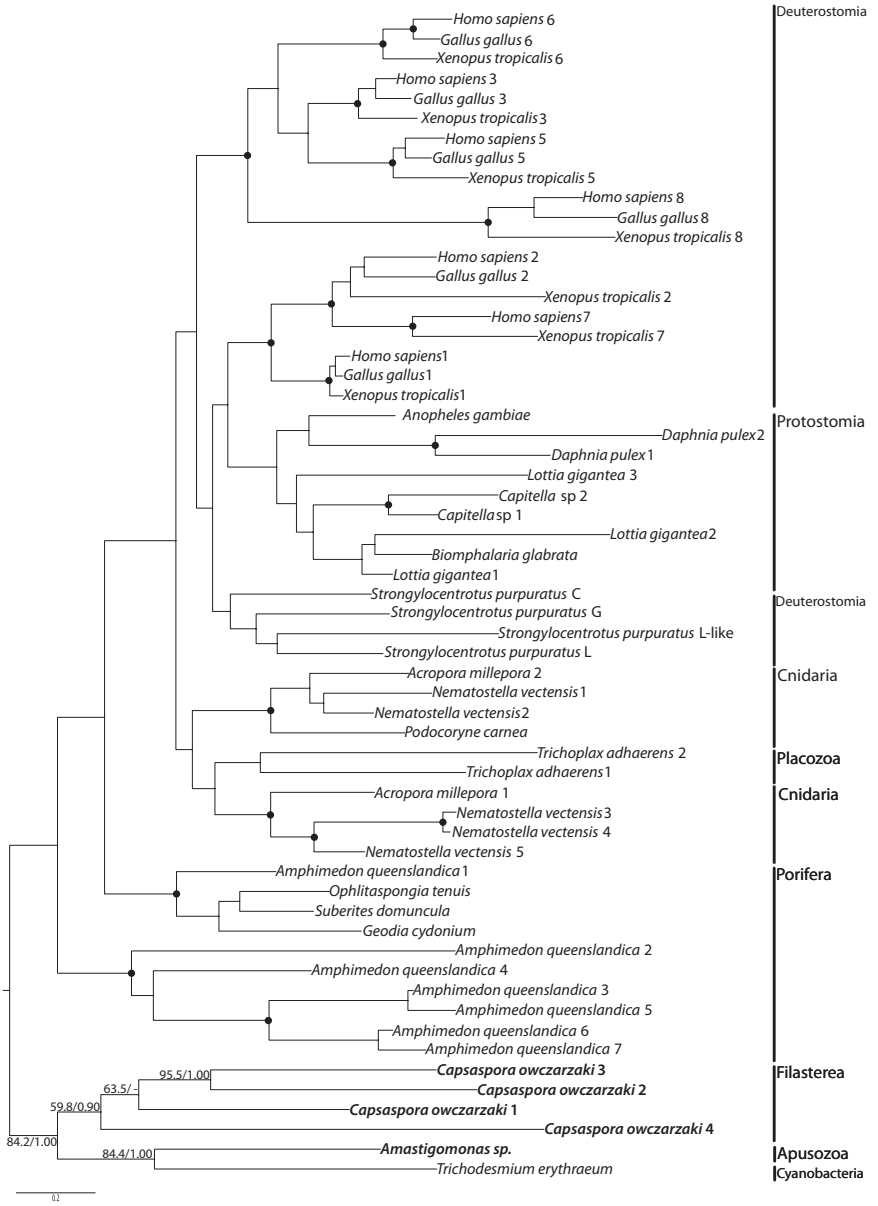
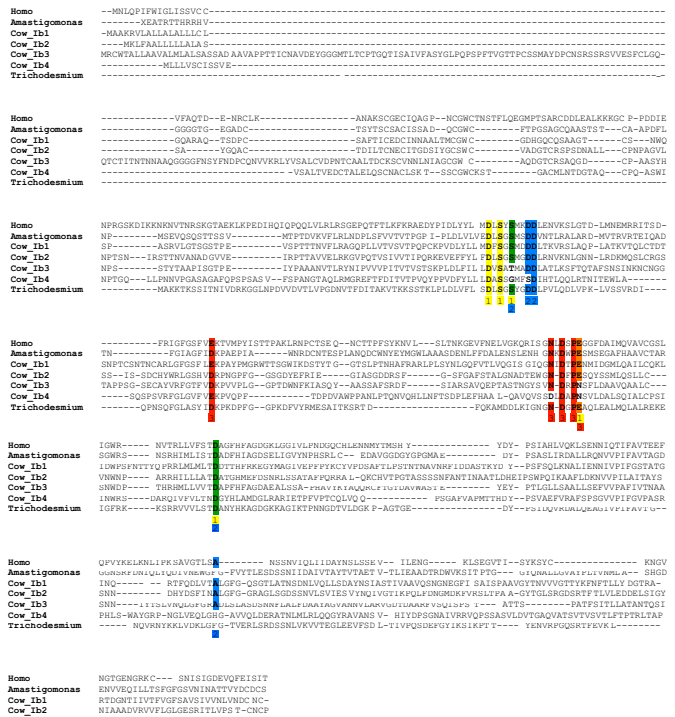


Fig. S2. Maximum likelihood tree of the integrin β protein. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 1,000-bootstrap replicates (bootstrap value, BV) and Bayesian posterior probabilities (BPP). Both values are shown on key branches. A black dot indicates BV >90% and BPP >0.95.

A. Whole Integrin β



B. Integrin β extracellular domain

Fig. S3. (*A*) Illustrative alignment of the whole integrin β, showing the integrin β domain, integrin stalk, transmembrane region, and cytoplasmic tail. Integrin homologs of *Capsaspora owczarzaki* (except integrin β-4, which is too derived), *Amastigomonas* sp., *Homo sapiens, Capitella* sp., and *Ophlitaspongia tenuis* are shown. The 56 conserved cysteines in metazoans are highlighted in green, whereas the 65 extra cysteines specific to protistan integrins are highlighted in red (see main text for more details). (*B*) Integrin β domain alignment for *H. sapiens*, *Amastigomonas* sp., *C. owczarzaki*, and *Trichodesmium erythraeum* to show in more detail the specific cation-binding motifs of Fig. 1 in main text, that is MIDAS (yellow, 1), ADMIDAS (blue, 2), and LIMB (red, 3). Orange and green means an amino acid shared by two motifs (indicated by the numbers below).

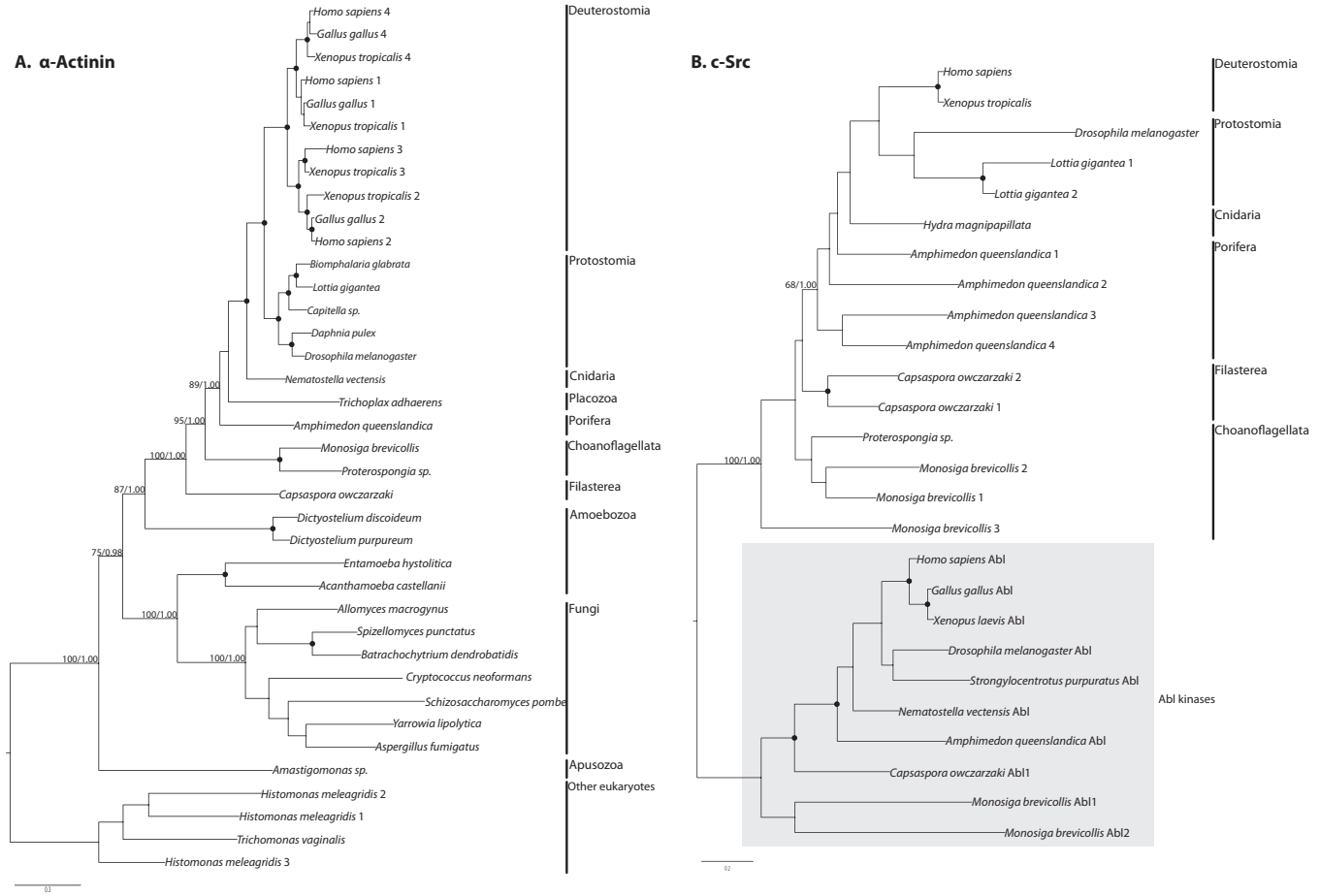


Fig. S4. Maximum likelihood tree of (*A*) α-actinin and (*B*) c-Src. The α-actinin tree (*A*) is rooted using bikont eukaryotes as an outgroup. The c-Src tree (*B*) is rooted using the closely related Abl kinase family as an outgroup. Statistical support was obtained by RAxML with 100-bootstrap replicates (bootstrap value, BV) and Bayesian posterior probabilities (BPP). Both values are shown in key branches. A black dot indicates BV >90% and BPP >0.95.

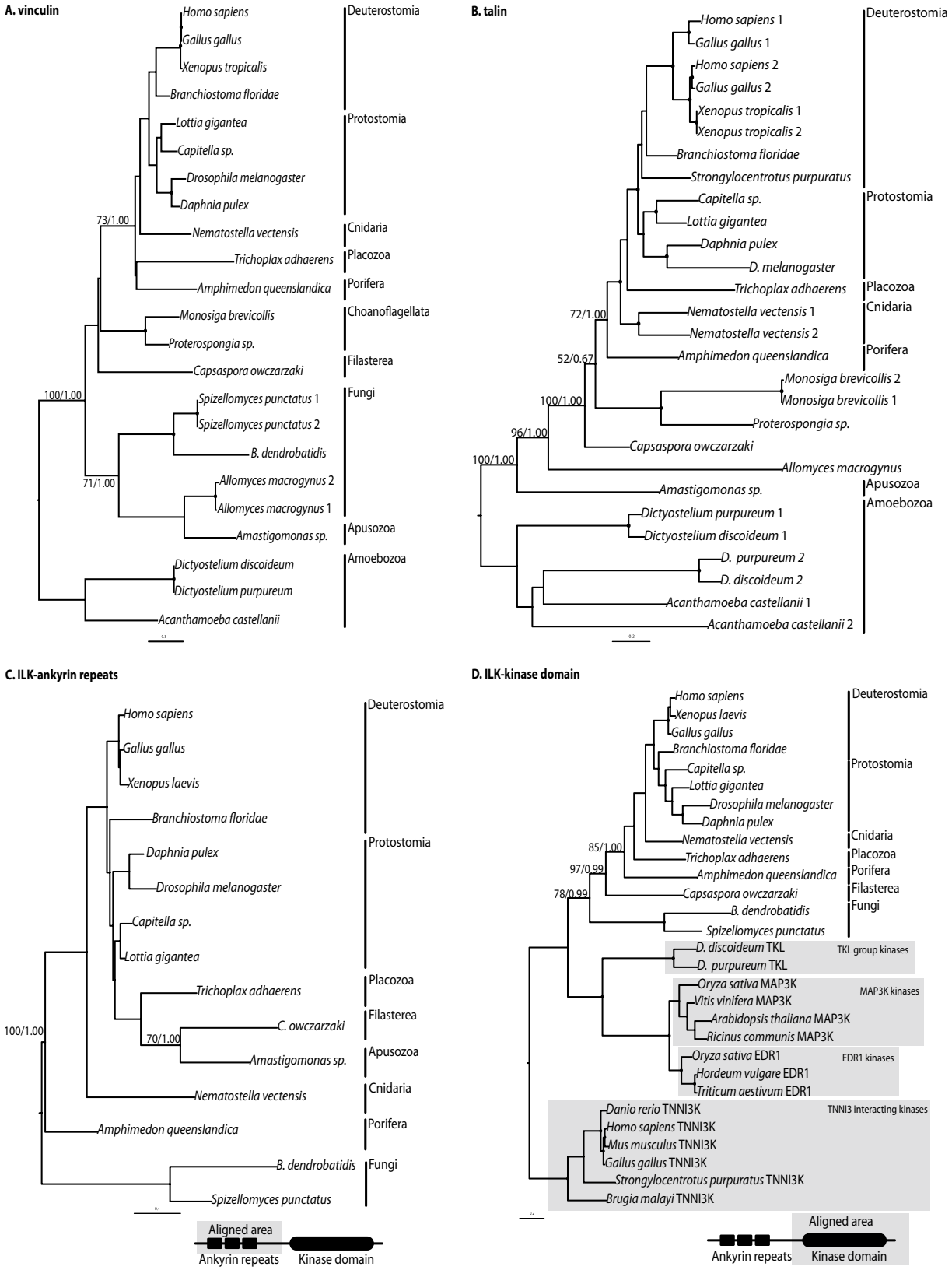


Fig. S5. Maximum likelihood tree of (*A*) vinculin protein, (*B*) talin protein, (*C*) integrin-linked kinase (ILK) protein based on the ankyrin repeats, and (*D*) ILK protein using the kinase domain. For each tree, statistical support was obtained by RAxML with 100-bootstrap replicates (bootstrap value, BV) and Bayesian posterior probability (BPP). Both values are shown in key branches. A black dot indicates BV >90% and BPP >0.95. Trees in *A* and *B* are rooted using the Amoebozoa as the outgroup. Tree in *C* is rooted using the midpoint-rooted tree option. Tree in *D* is rooted using several closely related kinase families as an outgroup.

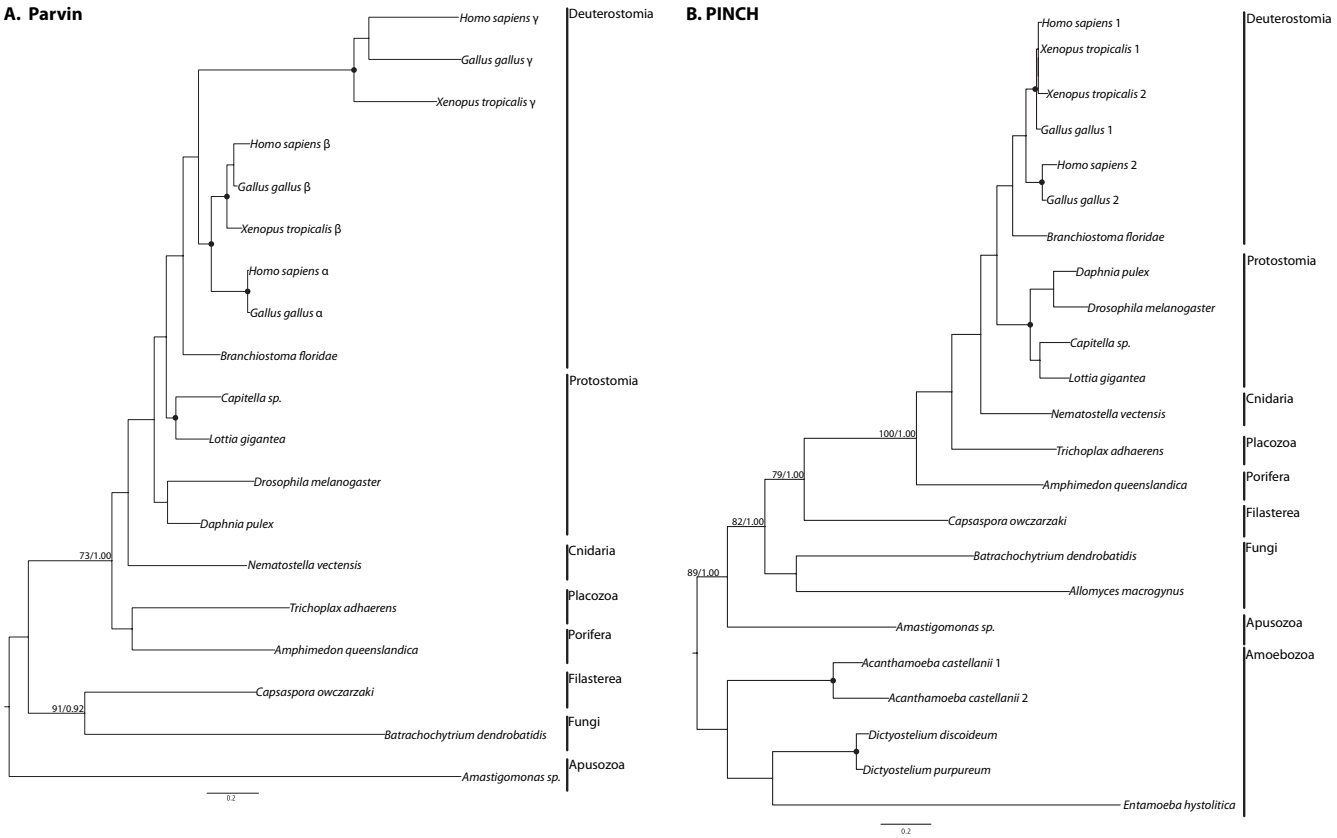


Fig. S6. Maximum likelihood tree of (*A*) the parvin protein and (*B*) the PINCH (particularly interesting Cys-His-rich) protein. The parvin tree is rooted using the midpoint-rooted tree option, whereas the PINCH tree is rooted using Amoebozoa as an outgroup. For each tree, statistical support was obtained by RAxML with 100-bootstrap replicates (bootstrap value, BV) and Bayesian posterior probability (BPP). Both values are shown in key branches. A black dot indicates BV >90% and BPP >0.95.

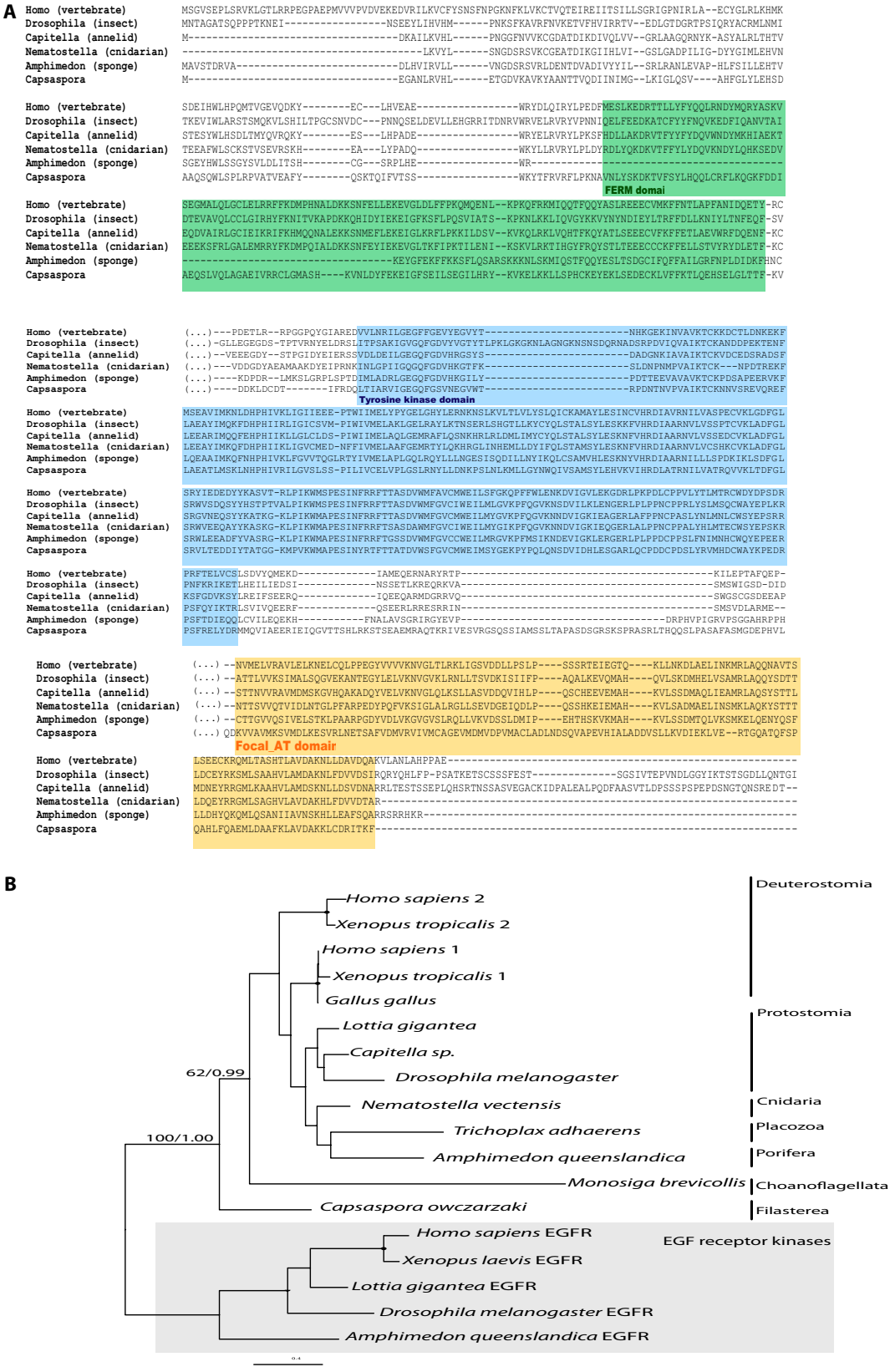


Fig. S7. (*A*) An illustrative focal adhesion kinase (FAK) alignment showing the different functional domains. (*B*) Maximum likelihood tree of the FAK protein with EGFR kinase family, based on the kinase domain. Tree is rooted using the EGFR kinase family as an outgroup. Statistical support was obtained by RAxML with 100-bootstrap replicates (bootstrap value, BV) and Bayesian posterior probability (BPP). Both values are shown in key branches. A black dot indicates BV >90% and BPP >0.95.

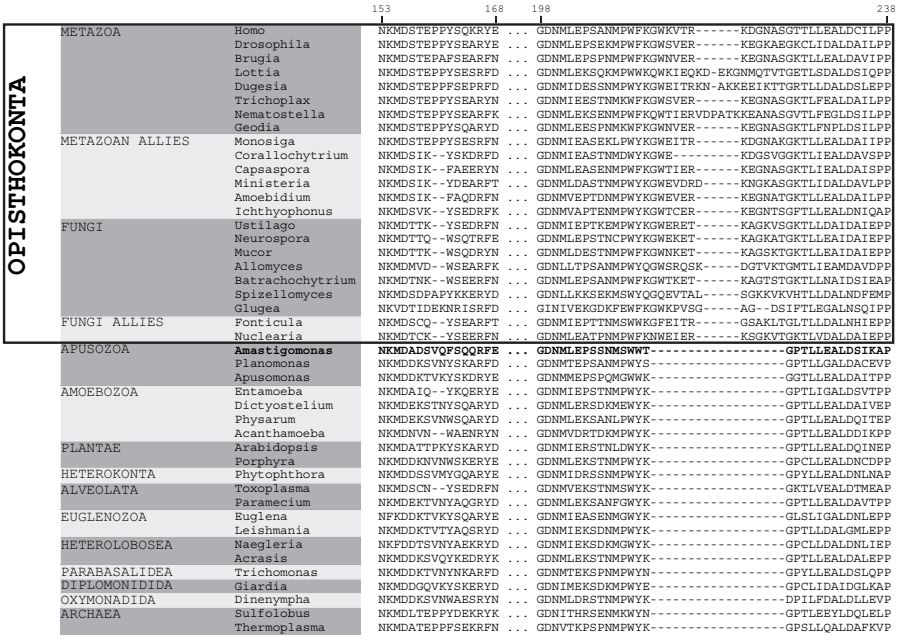


Fig. S8. Schematical alignment, based on the one shown by Steenkamp et al. (1), of a portion of the EF-1α gene showing the synapomorphic indel of opisthokonts. *Amastigomonas* sp. is shown in bold.

1. Steenkamp ET, Wright J, Baldauf SL (2006) The protistan origins of animals and fungi. *Mol Biol Evol* 23:93–106.

Other Supporting Information Files

Appendix S1 (DOC)

Results R2

Integrin-mediated adhesion complex. Cooption of signaling systems at the dawn of Metazoa.

RESUM ARTICLE R2: El complex d'adhesió d'integrina - co-opció de sistemes de senyalització a l'origen dels metazous

La maquinària d'adhesió per integrina és el principal sistema d'adhesió entre les cèl·lules i la matriu extracel·lular als animals. El complex d'adhesió d'integrina, el qual modula importants aspectes de la fisiologia cel·lular, està composat per integrines (subunitats beta i alpha) i diverses proteïnes de senyalització i de bastiment. Les integrines mai havien estat trobades en els eucariotes no-metazous analitzats fins fa poc, incloent fongs, plants i coanoflagel·lats (el llinatge unicel·lular germà dels animals). Arran d'això, les integrines i, com a tal, l'adhesió i senyalització que aquestes realitzen, eren considerades una innovació dels animals. Recentment, un estudi de genòmica comparada que inclou dades noves de diversos organismes unicel·lulars propers filogenèticament a animals i a fongs ha fet trontollar aquesta imatge. El complex d'adhesió i senyalització d'integrina no és específic dels animals, ans al contrari, és present en apusozous i en holozous unicel·lulars. Així, aquest important sistema d'adhesió i senyalització és anterior a l'origen dels fongs i els animals, i fou perdut secundàriament en fongs i coanoflagel·lats. Aquestes troballes suggereixen que la co-opció de gens jugà un important rol en l'origen dels animals, major del que fins ara es creia. En aquest article, hipotetitzem que la funció ancestral del complex d'adhesió d'integrina fou probablement la de senyalització.

Communicative & Integrative Biology 3:5, 475-477; September/October 2010; © 2010 Landes Bioscience

Integrin-mediated adhesion complex

Cooption of signaling systems at the dawn of Metazoa

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

[1]Departament de Genètica; Facultat de Biologia; Av. Diagonal 645 Universitat de Barcelona; [2]Institut de Recerca en Biodiversitat (Irbio); Parc Científic de Barcelona; [3]Institució Catalana per a la Recerca i Estudis Avançats (ICREA); Passeig Lluís Companys; Barcelona, Spain

The integrin-mediated adhesion machinery is the primary cell-matrix adhesion mechanism in Metazoa. The integrin adhesion complex, which modulates important aspects of the cell physiology, is composed of integrins (alpha and beta subunits) and several scaffolding and signaling proteins. Integrins appeared to be absent in all non-metazoan eukaryotes so-far analyzed, including fungi, plants and choanoflagellates, the sister-group to Metazoa. Thus, integrins and, therefore, the integrin-mediated adhesion and signaling mechanism was considered a metazoan innovation. Recently, a broad comparative genomic analysis including new genome data from several unicellular organisms closely related to fungi and metazoans shattered previous views. The integrin adhesion and signaling complex is not specific to Metazoa, but rather it is present in apusozoans and holozoan protists. Thus, this important signaling and adhesion system predated the origin of Fungi and Metazoa, and was subsequently lost in fungi and choanoflagellates. This finding suggests that cooption played a more important role in the origin of Metazoa than previously believed. Here, we hypothesize that the integrin adhesome was ancestrally involved in signaling.

Key words: multicellularity, integrins, cell-adhesion, Holozoa, cell-signaling, Capsaspora

Submitted: 06/04/10

Accepted: 06/04/10

Previously published online: www.landesbioscience.com/journals/cib/article/12603

DOI: 10.4161/cib.3.5.12603

*Correspondence to: Iñaki Ruiz-Trillo; Email: inaki.ruiz@icrea.es

Addendum to: 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 USA 2010; 107:10142–7; PMID: 20479219; DOI: 10.1073/pnas.1002257107.

Both cell adhesion and cell signaling, which are often correlated, are essential mechanisms for metazoan multicellularity.[1] Thus, elucidating the origin and evolution of those processes is key to further our understanding of metazoan origins. It has been inferred that the most ancestral cell junctions in metazoans are spot adherens junctions and focal adhesions.[2] The former are important for cell-cell adhesion and are molecularly based on cadherins, coupled with the actin cytoskeleton through catenins. On the other hand, focal adhesions are essential for cell-extracellular matrix (ECM) connection and are molecularly based on the integrin adhesion complex also linked to the actin cytoskeleton.

None of those two-cell adhesion systems are present in fungi or plants, but the analysis of the first genome sequence of a choanoflagellate, a group of single-celled and colony forming flagellates that are the closest relatives to metazoans, showed that cadherines appeared prior to the metazoan divergence. In fact, up to 23 different cadherins were described in the choanoflagellate *Monosiga brevicollis*, some of them structurally homologous to some metazoan cadherins (e.g., protocadherin or FAT) and others with completely new domain arrangements.[3] However, *M. brevicollis* does not have an integrin-mediated adhesion system. This leaves the integrin-mediated adhesion and signaling mechanism as one of the key inventions of the metazoan lineage.[4,5]

Molecular systematics has shown that choanoflagellates are not the only single-celled metazoan relatives. In fact, the opisthokonts are no longer a game of three (Metazoa, Fungi and choanoflagellates), but rather an eukaryotic clade teaming up with an increasing number of poorly known protist lineages, such as nucleariids, ichthyosporeans, filastereans (namely *Capsaspora owczarzaki* and *Ministeria*

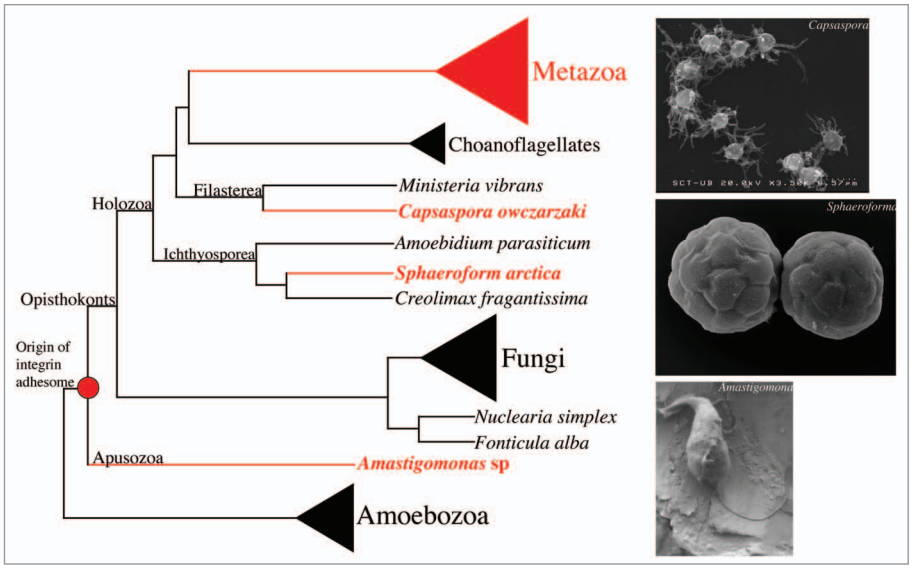


Figure 1. Schematic phylogenetic tree of the opisthokonts showing the putative origin of the integrin adhesion system with the current taxon sampling. Relationships are based on both published and unpublished molecular analyses.[6-10] Taxa in red are the ones encoding a complete integrin-mediate adhesion machinery. In bold taxa with complete genome sequences. Note that the genome of *S. arctica* is currently under way, but the trace data shows integrin hits.

vibrans), *Corallochytrium limacisporum* and, very recently, *Fonticula alba* (see **Fig. 1**).[6-10] That means those lineages should be taken into account when trying to infer the metazoan and fungi "genetic starter kit". In this regard, the UNICORN initiative aims to obtain the genome sequence of up to eleven protists lineages closely related to both fungi and metazoans.[11] Thanks to this project, the genome sequence of two cythrid fungi (*Allomyces macrogynus* and *Spizellomyces puctatus*), one filasterean (*C. owczarzaki*) and one choanoflagellate (Proterospongia sp.) have been obtained and are publicly available. Moreover, the complete genome of one apusozoan (Amastigomonas sp.), the putative sister-group of the opisthokonts[10,12] has also been sequenced.

This new information allows useful and taxonomically broader comparative genomic analyses, such as the one recently published by Arnau Sebé-Pedrós et al.[13] in which the repertoire of the different components of the integrin adhesome in opisthokonts and eukaryotes in general was investigated. The findings of Sebé-Pedrós are indeed unexpected, since a whole integrin-mediated machinery was found in two non-metazoan lineages, the filasterean *C. owczarzaki* and the apusozoan Amastigomonas sp (see **Fig. 1**). That the integrin-mediated complex is present in *C. owczarzaki* is by itself interesting enough, implying (1) that integrins are not exclusive to Metazoa and (2) that choanoflagellates lost such an important adhesion and signaling mechanisms. However, the most surprising finding is that the genome of the single-celled apusozoan Amastigomonas sp., which is not an opisthokont, also encodes the full repertoire (except for the two associated tyrosine kinases) of the integrin-mediated adhesion complex. These findings have large evolutionary implications. They not only take the integrin exclusivity out of metazoans, but also indicate that fungi (which comprise several species with complex multicellularity) secondarily lost this important cell adhesion and signaling machinery. Moreover, the finding that the origin of integrins is ancient means that the two most important metazoan cell adhesion mechanisms (cadherins and integrins) were already present in pre-metazoan lineages, and they were probably coopted for new functions in metazoans.

Undoubtedly, the presence of integrins in those unicellular organisms opens new and challenging questions. The first one, the actual role that this integrin machinery is playing in those single-celled organisms. Is the integrin machinery involved in sensing the extracellular environment? Such is the case, for example, of cadherins in choanoflagellates, which have been proposed be involved in the response to the extracellular environment, as bacterial prey capture.[3] Another question is whether integrins are present in other opisthokont lineages, such as ichthyosporean or nucleariids or *Fonticula alba* (which are the sister-group to Fungi). In this regard, the genome data of the remaining taxa to be sequenced by the UNICORN initiative (namely *F. alba*, the ichthyosporeans *Sphaeroforma arctica* and *Amoebidium parasiticum* and the free-living filasterean *M. vibrans*) will surely help unravel the evolutionary history of such important cell adhesion machinery.

Actually, a quick look at the current genome trace data of one ichthyosporean, *S. arctica*, whose genome is currently being sequenced, shows strong hits to integrins, suggesting that ichthyosporeans most probably also encode the integrin adhesion machinery. That means integrins are found in very different functional contexts, from single-celled amoeba endosymbiont crawlers (such as *C. owczarzaki*), to marine free-living flagellates (such as Amastigomonas sp.), to colony-forming fish and arthropod parasites (such as the ichthyosporean *S. arctica*), to fully multicellular eukaryotes (such as Metazoa) (see the **Fig. 1**). To us, this capacity to work in such different contexts suggests that integrins most probably had an ancestral role in signaling, for example in sensing the environment to modulate cell physiology and growth. The fact that the IPP signaling module is as ancient as integrins (it is also present in apusozoans, reviewed in ref. 13), and that the two tyrosine kinases associated with the integrin machinery (FAK and C-Src) were already present in the common ancestor of *C. owczarzaki*, choanoflagellates and Metazoa, seem to back up the signaling role of the ancestral integrins. The current cell-extracellular matrix role of the integrin complex in metazoans may have appeared by co-option, although we can not rule out the possibility that integrins in pre-metazoans played a cell adhesion role in, for example, the colony-forming ichthyosporeans. In

any case, functional analyses to elucidate the current role of the integrin adhesion system in *C. owczarzaki*, *S. arctica* or Amastigomonas sp., will surely be crucial to fully answer these new open questions.

Acknowledgements

We thank Lora L. Shadwick and John D.L. Shadwick for editing and critically reading the manuscript. This work was supported by an ICREA contract, an ERC Starting Grant (206883), and a grant (BFU2008-02839/BMC) from MICINN to IR.T. AS.P's salary was supported by a pre-graduate FPU grant from MICINN.

References

1. King N. The unicellular ancestry of animal development. Dev Cell 2004; 7:313-25.
2. Magie CR, Martindale MQ. Cell-cell adhesion in the cnidaria: insights into the evolution of tissue morphogenesis. Biol Bull 2008; 214:218-32.
3. Abedin M, King N. The premetazoan ancestry of cadherins. Science 2008; 319:946-8.
4. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman, et al. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature 2008; 451:783-8.
5. Rokas A. The origins of multicellularity and the early history of the genetic toolkit for animal development. Annu Rev Genet 2008; 42:235-51.
6. Ruiz-Trillo I, Inagaki Y, Davis LA, Sperstad S, Landfald B, Roger AJ. *Capsaspora owczarzaki* is an independent opisthokont lineage. Curr Biol 2004; 14:946-7.
7. Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF. A phylogenomic investigation into the origin of metazoa. Mol Biol Evol 2008; 25:664-72.
8. Steenkamp ET, Wright J, Baldauf SL. The protistan origins of animals and fungi. Mol Biol Evol 2006; 23:93-106.
9. 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.
10. Brown MW, Spiegel FW, Silberman JD. Phylogeny of the "forgotten" cellular slime mold, *Fonticula alba*, reveals a key evolutionary branch within Opisthokonta. Mol Biol Evol 2009; 26:2699-709.
11. Ruiz-Trillo I, Burger G, Holland PW, King N, Lang BF, Roger AJ, et al. The origins of multicellularity: a multi-taxon genome initiative. Trends Genet 2007; 23:113-8.
12. Kim E, Simpson AG, Graham LE. Evolutionary relationships of apusomonads inferred from taxon-rich analyses of 6 nuclear encoded genes. Mol Biol Evol 2006; 23:2455-66.
13. 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 USA 2010; 107:10142-7.

Results R3

Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki*.

RESUM ARTICLE R3: Inesperat repertori de factors de transcripció animals en l'holozou unicel·lular *Capsaspora owczarzaki*

Com els animals (metazous) s'originaren des dels seus ancestres unicel·lulars segueix sent una qüestió fonamental de la biologia. Elucidar l'evolució dels factors de transcripció crucials per a la regulació del desenvolupament animal és crític per a entendre l'origen dels animals. En aquest treball, analitzem el repertori de 17 famílies de factors de transcripció típics d'animals en l'holozou ameboide *Capsaspora owczarzaki*, un representant del llinatge unicel·lular més proper a animals i coanoflagel·lats. Anàlisis filogenètics i de genòmica comparada, amb el mostreig taxonòmic més ampli possible, ens permeten formular noves hipòtesis sobre l'origen i evolució dels factors de transcripció del desenvolupament animal. Mostrem que *Capsaspora owczarzaki* té un repertori de factors de transcripció sorprenentment elevat i determinem un origen molt més antic de factors de transcripció que es consideraven específics d'animals, com ara T-box o Runx. Malgrat tot, famílies de factors de transcripció que pre-daten l'origen dels animals, com ara homeodominis o bHLH, van expandir-se i diversificar-se significativament a l'arrel dels metazous i dels eumetazous.

Unexpected Repertoire of Metazoan Transcription Factors in the Unicellular Holozoan *Capsaspora owczarzaki*

Arnau Sebé-Pedrós,[†,1] Alex de Mendoza,[†,1] B. Franz Lang,[2] Bernard M. Degnan,[3] and Iñaki Ruiz-Trillo*[,1,4]

[1]Departament de Genètica & Institut de Recerca en Biodiversitat (Irbio), Universitat de Barcelona, Barcelona, Spain

[2]Department of Biochemistry, Université de Montréal, Montréal, Canada

[3]School of Biological Sciences, The University of Queensland, Brisbane, Queensland, Australia

[4]Institució Catalana per a la Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys, Barcelona, Spain

†These authors contributed equally to this work.

***Corresponding author:** E-mail: inaki.ruiz@icrea.es.

Associate editor: Billie Swalla

Abstract

How animals (metazoans) originated from their single-celled ancestors remains a major question in biology. As transcriptional regulation is crucial to animal development, deciphering the early evolution of associated transcription factors (TFs) is critical to understanding metazoan origins. In this study, we uncovered the repertoire of 17 metazoan TFs in the amoeboid holozoan *Capsaspora owczarzaki*, a representative of a unicellular lineage that is closely related to choanoflagellates and metazoans. Phylogenetic and comparative genomic analyses with the broadest possible taxonomic sampling allowed us to formulate new hypotheses regarding the origin and evolution of developmental metazoan TFs. We show that the complexity of the TF repertoire in *C. owczarzaki* is strikingly high, pushing back further the origin of some TFs formerly thought to be metazoan specific, such as T-box or Runx. Nonetheless, TF families whose beginnings antedate the origin of the animal kingdom, such as homeodomain or basic helix-loop-helix, underwent significant expansion and diversification along metazoan and eumetazoan stems.

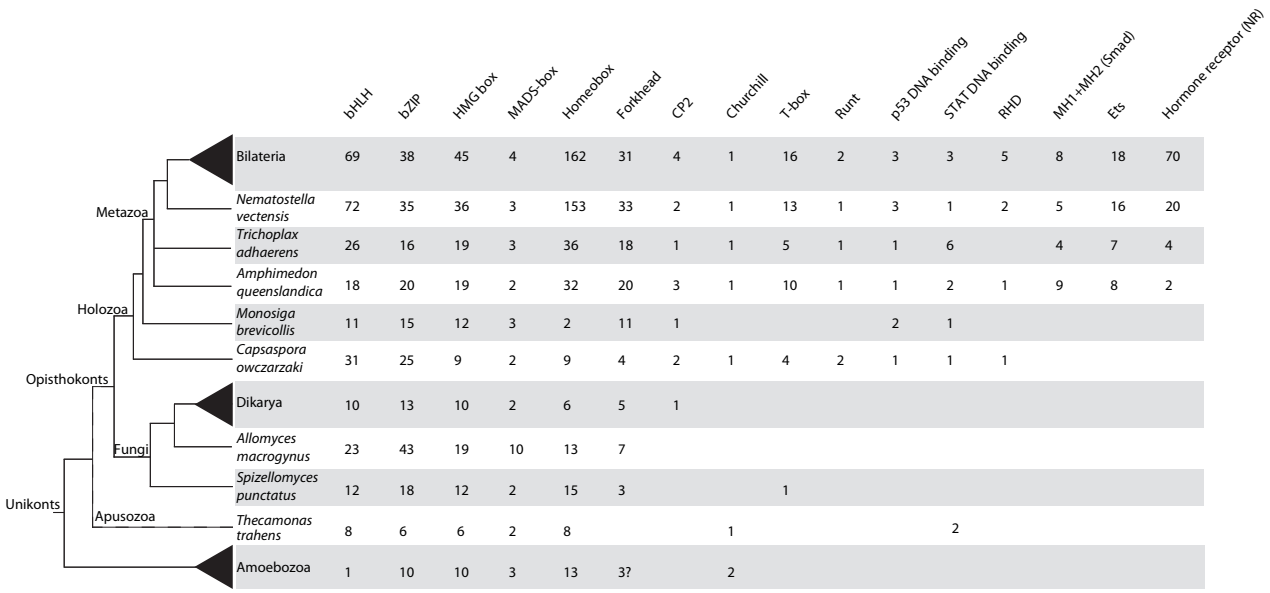
Key words: multicellularity, T-box, homeodomain, brachyury, origin Metazoa, choanoflagellates.

Introduction

What genomic changes took place at the dawn of the Metazoa remains a major biological question. Transcriptional regulation appears to be one of the most crucial aspects of animal development. Thus, understanding the early evolution of the transcriptional regulatory machinery is critical for drawing a complete picture of metazoan origins. Transcription factors (TFs) act as regulators of cell fate, cell cycle, patterning, proliferation, development, and differentiation in metazoans (Larroux et al. 2008). Previous studies have shown that most TFs that play important roles in bilaterian development originated before the divergence of extant animal phyla (Larroux et al. 2006, 2008; King et al. 2008; Degnan et al. 2009; Srivastava et al. 2010). However, the complexity of most TF families appears to have increased during early eumetazoan evolution, with cnidarians having a TF gene repertoire typically being two to three times larger than that of sponges and placozoans (Putnam et al. 2007; Degnan et al. 2009; Srivastava et al. 2010). Based on comparative analyses, it has been hypothesized that the metazoan TF "toolkit" included members of the basic helix-loop-helix (bHLH), myocite enhancer factors 2 (Mef2), Fox, Sox, T-box, Ets, nuclear receptor (NR), Rel/nuclear factor-kappaB (NF-kappaB), basic-region leucine zipper (bZIP), and Smad families and a range of homeobox-containing classes, including ANTP, Prd-like, Pax, POU, LIM-HD, Six, and three-amino acid-loop extension (TALE) (for a review, see Degnan et al. 2009).

Comparative analyses including the holozoan choanoflagellate *Monosiga brevicollis*, the putative sister-group to metazoans, are greatly improving our understanding of metazoan TF evolution. The genome of *M. brevicollis* contains the standard set of TFs observed across eukaryotes but lacks most of the well-known metazoan TFs, except p53, Myc, and a putative Sox (King et al. 2008; Degnan et al. 2009). Under this scenario, metazoan-specific TFs appear to include ANTP, Prd-like, POU, LIM-HD, and six homeobox genes, group I Fox, most bHLH groups (except B), some bZIP families, Ets, Runx, Mef2, and NR families (Degnan et al. 2009).

To gain further insight into the evolution of TFs leading to the metazoan lineage, we characterized and analyzed all the TFs that supposedly constitute the metazoan TF toolkit in another close unicellular relative of animals, the amoeboid holozoan *Capsaspora owczarzaki*, putatively the sister-group to metazoans and choanoflagellates (Ruiz-Trillo et al. 2004, 2008; Shalchian-Tabrizi et al. 2008; Brown et al. 2009; see fig. 1). The complete genome sequence of *C. owczarzaki* (hereafter "*Capsaspora*") has recently been obtained under the "UNICORN project" at the Broad Institute (Ruiz-Trillo et al. 2007). In addition to the TFs outlined above, our survey of the *Capsaspora* genome in this study also included other TFs known to be important to animal development,



	bHLH	bZIP	HMG box	MADS-box	Homeobox	Forkhead	CP2	Churchill	T-box	Runt	p53 DNA binding	STAT DNA binding	RHD	MH1+MH2 (Smad)	Ets	Hormone receptor (NR)
Bilateria	69	38	45	4	162	31	4	1	16	2	3	3	5	8	18	70
Nematostella vectensis	72	35	36	3	153	33	2	1	13	1	3	1	2	5	16	20
Trichoplax adhaerens	26	16	19	3	36	18	1	1	5	1	1	6		4	7	4
Amphimedon queenslandica	18	20	19	2	32	20	3	1	10	1	1	2	1	9	8	2
Monosiga brevicollis	11	15	12	3	2	11	1				2	1				
Capsaspora owczarzaki	31	25	9	2	9	4	2	1	4	2	1	1	1			
Dikarya	10	13	10	2	6	5	1									
Allomyces macrogynus	23	43	19	10	13	7										
Spizellomyces punctatus	12	18	12	2	15	3			1							
Thecamonas trahens	8	6	6	2	8			1				2				
Amoebozoa	1	10	10	3	13	3?		2								

FIG. 1 Table of domain presence and number across unikonts. Columns represent all the PFAM domains analyzed in this study. The number of genes in each TF family was inferred from each organisms's proteome by PfamScan using the PfamScan default parameters. For *Monosiga brevicollis* and *Capsaspora owczarzaki*, the analyses were performed by HMMER 3.0 searches. For Smad proteins, containing one MH1 and one MH2 domain, the number shown is the minimal number of either MH1 or MH2. For Bilateria, Dikarya, and Amoebozoa the average number is shown. Bilateria includes *Homo sapiens*, *Ciona intestinalis*, *Drosophila melanogaster*, *Anopheles gambiae*, *Caenorhabditis elegans*, *Helobdella robusta*, and *Lottia gigantea*. Dikarya includes *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Cryptococcus neoformans*, *Yarrowia lypolitica*, *Ustilago maydis*, *Aspergillus niger*, *Neurospora crassa*, and *Phanerochaete chrysosporium*. Amoebozoa includes *Dictyostellium discoideum*, *Dictyostellium purpureum*, *Entamoeba histolytica*, *Entamoeba dispar*, and *Acanthamoeba castellanii*. The phylogenetic relationships are based on several recent phylogenomic studies (Burki et al. 2008; Ruiz-Trillo et al. 2008; Brown et al. 2009; Liu et al. 2009; Minge et al. 2009).

including Churchill, p53, Stat, and LSF/Grainyhead (GRH) (fig. 1). Comparative genomic analyses were performed on holozoan genomes, and in some cases, other recently sequenced opisthokont and apusozoan genomes, namely *Allomyces macrogynus*, *Spizellomyces punctatus*, and *Thecamonas trahens* (see Materials and Methods, fig. 1). These results show that the complexity of TFs in *Capsaspora* is very high, indicating that some TFs thought to be metazoan specific evolved prior to the metazoan and choanoflagellate divergence and were subsequently lost in the choanoflagellate lineage.

Materials and Methods

Taxonomic Sampling

We surveyed, and characterized, a list of metazoan TFs in *Capsapora*. In some cases, we extended our searches to the widest possible set of eukaryotic taxa. This was the case for those TF families with specific and unique domains: T-box (T-box DNA-binding domain), Runx (Runt DNA-binding domain), NF-kappaB (Rel homology domain [RHD]), Mef2 (MADS box + Mef2 domain), p53 (p53 DNA-binding domain), Stat (Stat DNA-binding domain), Churchill (Churchill domain), Smad (MH1 + MH2 domains), Ets (Ets domain), and NR. Our extended searches included published and publicly available eukaryotic genomes, and other UNICORN taxa, such as the basal fungi *A. macrogynus*, *S. punctatus*, and the apusozoan *T. trahens* (see http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html). For the remaining TF families (i.e., bZIP [bZIP domain], Fox [forkhead domain], Sox [HMG box], homeobox [homeodomain], bHLH [bHLH domain], and LSF/GRH [CP2 domain]), we classified those *Capsaspora* genes with homology to metazoan genes. To this end, we used published fungal, metazoan, and choanoflagellate homologs. We also characterized bZIPs and Mef2 in *M. brevicollis*.

Gene Searches

A primary search was performed using the basic local alignment sequence tool (BLAST: BlastP and TBlastN) using *Homo sapiens* proteins as queries against Protein and Genome databases with the default BLAST parameters and an *e* value threshold of 10×10^{-5} at the National Center for Biotechnology Information (NCBI) and against completed or on-going genome project databases at the Joint Genome Institute (JGI), the Broad Institute, as well as the *A. queenslandica* genome database (www.metazome.net/amphimedon). In the case of *T. trahens*, *A. macrogynus*, and *Acanthamoeba castellanii*, we assembled the trace data using the WGS assembler ("http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page" http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page). We then annotated the genes of interest using both Genescan (Burge and Karlin 1997) and Augustus (Stanke and Morgenstern 2005) and performed local BLAST searches. When the BLAST searches of the genome data described above returned significant

"hits," the sequences obtained were then reciprocally searched against the NCBI protein database by BLAST in order to confirm the validity of the sequences retrieved with the initial search. Hmmer searches using HMMER3. 0b2 (Eddy 1998) were also performed, with standard PFAM profiles in the case of widespread domains or with home-made profiles in the case of specific domains.

Protein Domain Arrangements

For all proteins, the presence of specific protein domains was further checked by searching the Pfam ("http://pfam.sanger.ac.uk/search"http://pfam.sanger.ac.uk/search) and SMART ("http://smart.embl-heidelberg.de/"http://smart.embl-heidelberg.de/) databases.

Polymerase Chain Reaction confirmation of *C. owczarzaki* T-box, Runx, and NF-kappaB Genes

We confirmed the presence of the three *Capsaspora* TFs that were formerly considered to be metazoan-specific TFs now identified in *Capsaspora* (Runx, T-box, and NF-kappaB), using gene-specific oligonucleotide primers. The mRNA was extracted using a Dynabeads mRNA purification kit (Invitrogen, Carlsbad, CA) and subsequent reverse transcriptase-polymerase chain reaction (RT-PCR) was performed using a Superscript III First Strand Synthesis kit (Invitrogen). The full sequence of the 5' and 3' ends of the cited *Capsaspora* TF cDNAs were obtained by RACE, using a nested PCR and with specific oligonucleotide primers designed from the original genome data. Both coding and noncoding strands were sequenced using an ABI PRISM BigDye Termination Cycle Sequencing Kit (Applied Biosystems, Foster City, CA). New sequences were deposited in GenBank under the following accession numbers: GU985459 (*Capsaspora* Bra-like), GU985460 (*Capsaspora* double-tbox), GU985461 (*Capsaspora* Tbox3), GU985462 (*Capsaspora* Runx1), GU985463 (*Capsaspora* Runx2), and GU985464 (*Capsaspora* NF-kappaB).

Phylogenetic Analyses

Alignments were constructed for the following gene families and classes: T-box, homeobox, Fox, Sox, bHLH, bZIP, LAG, signal transducer and activator of transcription (STAT), Mef2, p53, NF-kappaB, Churchill, HMG box, GRH/LSF, and Runx. Alignments were obtained using the MAFFT v.6 online server (Katoh, Kuma, Miyata, and Toh 2005; Katoh, Kuma, Toh, and Miyata 2005) and then manually inspected and edited in Geneious. Only those species and those positions that were unambiguously aligned were included in the final analyses. Maximum likelihood (ML) phylogenetic trees were estimated by RaxML (Stamatakis 2006) using the PROTGAMMAWAGI model, which uses the Whelan and Goldman (WAG) amino acid exchangeabilities and accounts for among-site rate variation with a four category discrete gamma approximation and a proportion of invariable sites (WAG + Γ + I). Statistical support for bipartitions was estimated by performing 100-bootstrap replicates using RaxML with the same model. Bayesian analyses were performed with MrBayes 3.1 (Ronquist and Huelsenbeck 2003), using the WAG + Γ + I model of evolution, with four chains, a subsampling frequency of 100 and two parallel runs. Runs were stopped when the average standard deviation of split frequencies of the two parallel runs was <0.01, usually at around 1,000,000 generations. The two LnL graphs were checked and an appropriate burn-in length established; stationarity of the chain typically occurred after ~15% of the generations. Bayesian posterior probabilities (BPP) were used to assess the confidence values of each bipartition.

Homeodomain Gene Assignment

An alignment with members of the ANTP, Paired-like, POU, and LIM homeodomain classes was constructed using published data from *Amphimedon*, *Drosophila*, and *Nematostella* and other already classified sequences (Larroux et al. 2008). A RaxML best tree resulting from this phylogeny was produced to obtain the fixed topology, which recovered monophyly for all four classes. From this tree, we manually created constrained topologies that represented all the possible positions of *Capsaspora* non-TALE homeodomains. Site-wise log-likelihoods were calculated for all the generated topologies with RaxML. Best-scoring ML trees were chosen using the likelihood-based approximately unbiased (AU) test as implemented in CONSEL (Shimodaira and Hasegawa 2001). The positions of *Capsaspora* homeodomain genes that could not be statistically excluded ($P \geq 0.05$) were taken into account. Whenever the significant positions fell in the branches that connect the different classes of homeodomains (POU, LIM, . . .), the homeodomain was not classified. When *Capsaspora* hits fell inside just one cluster (e.g., *Capsaspora6* inside paired-like), they were classified accordingly.

Quantitative TF analyses in Unikont Taxa

To quantify the number of genes in each TF family, we used PfamScan using the PfamScan default parameters. The predicted proteomes used for PfamScan analysis were *Amphimedon queenslandica* (JGI), *Trichoplax adhaerens* (JGI), *Nematostella vectensis* (JGI), *H. sapiens* (NCBI), *Ciona intestinalis* (JGI), *Drosophila melanogaster* (NCBI), *Anopheles gambiae* (NCBI), *Caenorhabditis elegans* (NCBI), *Helobdella robusta* (JGI), *Lottia gigantea* (JGI), *Saccharomyces cerevisiae* (NCBI), *Schizosaccharomyces pombe* (NCBI), *Cryptococcus neoformans* (JGI), *Yarrowia lypolitica* (NCBI), *Ustilago maydis* (JGI), *Aspergillus niger* (JGI), *Neurospora crassa* (NCBI), *Phanerochaete chrysosporium* (JGI), *A. macrogynus* (Broad Institute), *S. punctatus* (Broad Institute), *Dictyostellium discoideum* (NCBI), *Dictyostellium purpureum* (JGI), *Entamoeba histolytica* (NCBI), *Entamoeba dispar* (NCBI), and *A. castellanii* (home-made prediction). For *M. brevicollis* and *Capsaspora*, the analyses were performed by HMMER 3.0 searches. For Smad proteins, containing one MH1 and one MH2 domain, the number was inferred by taking the minimal number of either MH1 or MH2 present in the proteomes.

Results and Discussion

Rel/NF-kappaB

The RHD is a conserved DNA binding and dimerization domain that is present in the N-terminal region of two protein families: nuclear factor activated T-cells (NFAT) and Rel/NF-kappaB. NFAT and Rel/NF-kappaB are involved in immune system processes in metazoans (Macian 2005). Rel/NF-kappaB also plays different roles in development and cell differentiation, receiving inputs from several signaling pathways (Hayden and Ghosh 2004). Until now, the RHD domain has not been identified outside metazoans and was thus considered a metazoan innovation (Gauthier and Degnan 2008).

However, we identified a single RHD domain in *Capsaspora* but failed to recover RHD from any other sequenced nonmetazoan taxa (fig. 1). Our phylogenetic analysis of the RHD domain shows the *Capsaspora* homolog branching off as sister-group of all metazoan Rel/NF-kappaB (supplementary fig. S1, Supplementary Material online). Furthermore, the *Capsaspora* RHD-domain-containing protein shares several key features with metazoan Rel and NF-kappaB homologs, such as 1) a highly conserved and specific recognition loop located within the RHD domain, which is involved in dimerization; 2) an IPTG or RHD2 domain, which confers binding specificity; 3) a basic nuclear-localization sequence; 4) a glycine−serine rich region; and 5) several ankyrin repeats, which are exclusive to metazoan NF-kappaB proteins (supplementary fig. S2, Supplementary Material online).

Thus, our data show that the RHD domain is not exclusive to metazoans as previously thought but rather it originated prior to the divergence of *Capsaspora* from choanoflagellates and metazoans. This implies that the RHD domain was subsequently lost in the choanoflagellate lineage.

Runx

The Runt DNA-binding domain defines a family of metazoan TFs (Runx) with essential roles in animal development (Coffman 2003; Robertson et al. 2009). They can act as transcriptional activators or repressors, in the latter case usually via corepressors of the Groucho/TLE family (Wheeler et al. 2000). Runx genes encode the Runt DNA-binding domain and heterodimerization domain and a C-terminal WRPY motif that interacts with the Groucho/TLE corepressor (Coffman 2003), except in the demosponge *A. queenslandica* and some bilaterian paralogs (specifically one of the two leech and planarian paralogs), which all lack the C-terminal WRPY motif (Robertson et al. 2009). A single Runx gene is present in *A. queenslandica*, *N. vectensis*, and *T. adhaerens*, although most bilaterians have several copies as a result of independent duplications (Rennert et al. 2003). Runx was previously considered to be metazoan specific (Robertson et al. 2009).

We failed to recover Runx genes from any other sequenced nonmetazoan genome except *Capsaspora*, which has two genes (fig. 1). Both *Capsaspora* Runxs possess key DNA-binding amino acids in the Runt motif (Wheeler et al. 2000; Sullivan et al. 2008), although only one of the paralogs (*Co_Runx1*) has the two Cys residues involved in redox regulation (Akamatsu et al. 1997) (supplementary fig. S3, Supplementary Material online). Interestingly, as in *A. queenslandica* and one of the two leech and planarian paralogs, both *Capsaspora* Runx lack the specific C-terminal WRPY Groucho-interacting motif. In contrast to *A. queenslandica*, however, *Capsaspora* does not encode Groucho in its genome. Neither does *Capsaspora* encode CBFβ, the heterodimeric-binding partner of the Runt domain that enhances its DNA affinity (Sullivan et al. 2008). This suggests that the Runt domain acts independently from CBFβ in *Capsaspora*. Our results show that Runx originated prior to the divergence of *Capsaspora* from choanoflagellates and metazoans, being secondarily lost in the choanoflagellate lineage. We hypothesize that Runx originally functioned independently of Groucho and CBFβ proteins and that the WRPY Groucho-interacting motif appeared in the eumetazoan lineage, as previously suggested (Robertson et al. 2009).

T-box

T-box TFs are characterized by an evolutionary conserved DNA-binding motif of 180−200 amino acids, the T-box domain (Smith 1999). They are key regulators of metazoan development (Muller and Herrmann 1997). The most well-known type of T-box is Brachyury, which has a key role in mesoderm specification (Marcellini et al. 2003), although its ancestral function may have been blastopore determination and gastrulation (Scholz and Technau 2003). T-box genes were previously generally considered to be metazoan specific (King et al. 2008; Larroux et al. 2008; Rokas 2008).

Here, we report the discovery of T-box genes in two nonmetazoan species. Three T-box genes are present in *Capsaspora* (one containing two consecutive T-box domains), and one gene exists in the basal chytrid fungus *S. punctatus* (fig. 1). Our searches, however, failed to recover T-box homologs from any other fungi (including the chytrids *A. macrogynus* and *Batrachochytrium dendrobatidis*) or other eukaryote (including the choanoflagellate *M. brevicollis*). Remarkably, all the T-box homologs from both *Capsaspora* and *S. punctatus* contain most of the key DNA-binding and dimerization amino acids of the metazoan T-box (Muller and Herrmann 1997; Bielen et al. 2007) (supplementary fig. S4, Supplementary Material online). The phylogenetic analysis of T-box domains (fig. 2) places one *Capsaspora* homolog (*Co*-Bra) inside the Brachyury family (bootstrap value [BV] = 50%). The two T-box domains in the *Capsaspora* "double-tbox" (Co-Dtbx1 and Co-Dtbx2) and the *S. punctatus* T-box clearly cluster together adjacent to the Brachyury family. The third *Capsaspora* homolog (Co-Tbx3) clusters within a group of unclassified T-box genes from the sponge *A. queenslandica* that may represent an independent and novel class of T-box genes. Our general topology supports the hypothesis that Brachyury is probably the ancestral class within the T-box family (Adell et al. 2003; Adell and Muller 2005; Larroux et al. 2008).

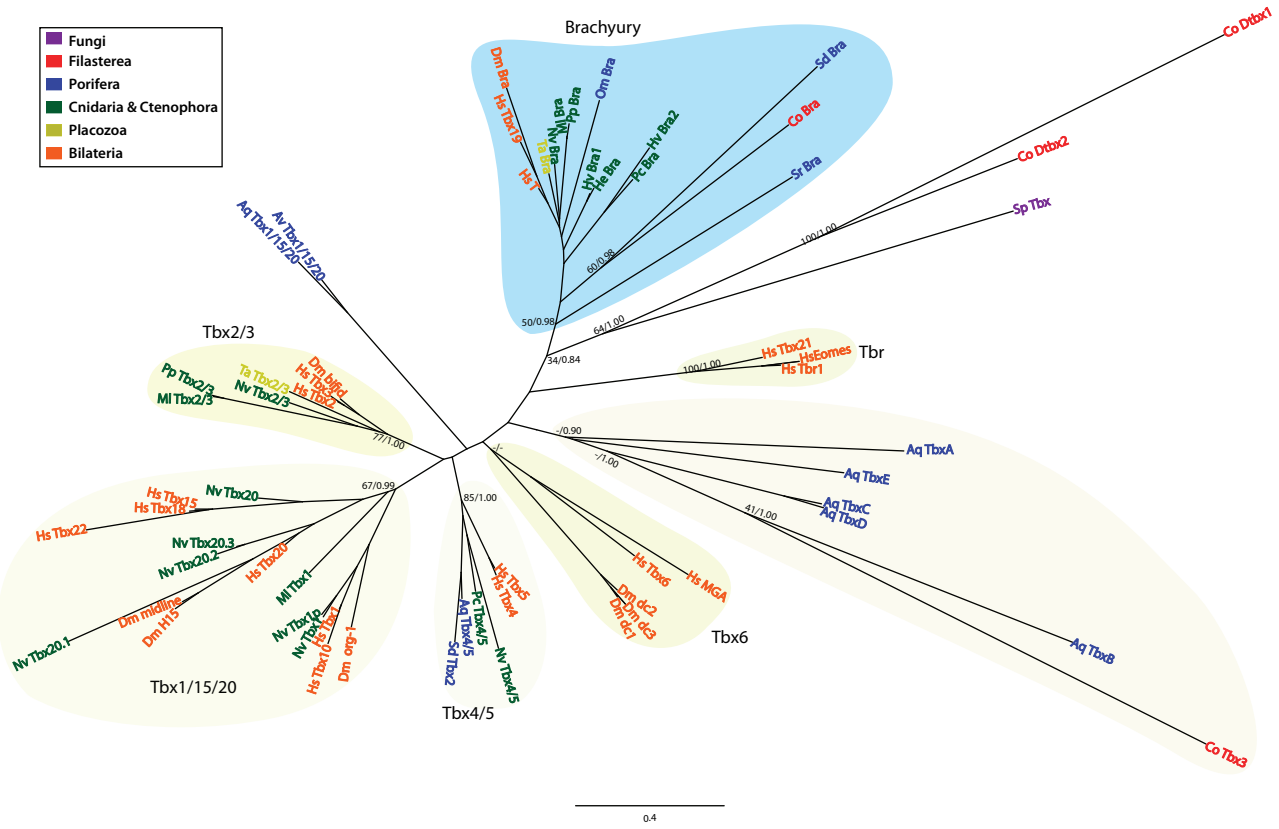


Fig. 2 ML tree of T-box domains showing the different T-box families. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 1,000 bootstrap replicates (BV) and BPP. Both values are shown on key branches. Colors show different taxonomic assignments. Aq (*Amphimedon queenslandica*), Av (*Axinella verrucosa*), Co (*Capsaspora owzarzaki*), Dm (*Drosophila melanogaster*), He (*Hydractinia echinata*), Hs (*Homo sapiens*), Hv (*Hydra vulgaris*), Ml (*Mnemiopsis leydi*), Nv (*Nematostella vectensis*), Om (*Oopsacas minuta*), Pc (*Podocoryne carnea*), Pp (*Pleurobrachia pileus*), Sd (*Suberites domuncula*), Sp (*Spizellomyces punctatus*), Sr (*Sycon raphanus*), Ta (*Trichoplax adhaerens*). Co-Dtbx1 and Co-Dtbx2 are the two T-box domains of the same T-box *Capsaspora* gene (for further details, see main text).

Moreover, our findings imply that T-box genes appeared not in metazoans but in the common ancestor of opisthokonts and were subsequently lost in most fungi and in choanoflagellates.

Churchill

Churchill is a zinc-finger TF that is involved in cell movement and cell fate determination (Londin et al. 2007). In *Xenopus* and chick, Churchill appears to regulate the T-box gene brachyury (Sheng et al. 2003). We have found orthologs of Churchill in *Capsaspora*, *T. trahens*, and, interestingly, also in the amoebozoan *A. castellanii* (fig. 1 and supplementary fig. S5, Supplementary Material online). This finding indicates a deeper origin of this gene than previously thought probably in the common ancestor of unikonts. This suggests that Churchill was secondarily lost in fungi, and choanoflagellates as well as in other amoebozoans. What role the Churchill orthologs play in *Capsaspora*, *T. trahens*, or *A. castellanii*, and whether, in *Capsaspora*, it is related at all to its T-box genes is unknown.

p53

The p53 tumor suppressor protein is a multifaceted TF that is involved in different cellular responses to DNA damage, such as DNA repair, cell cycle arrest, senescence, and apoptosis (Coutts and La Thangue 2005; Espinosa 2008). The p53 family includes p53, p63, and p73, the last two being more closely related to each other than to p53. The three p53 members have some differences in function and in the protein domain architecture. The p63 and p73 share an additional C-terminal sterile alpha motif (SAM) domain, whereas all three share a transcriptional activation domain, a DNA-binding domain, and C-terminal tetramerization domain (Nedelcu and Tan 2007). Choanoflagellates have both a p53 and a p63/73 classes (Nedelcu and Tan 2007).

Here, we characterize a unique member of the p53 gene family in *Capsaspora* (fig. 1 and supplementary fig. S6, Supplementary Material online), the gene encodes a SAM domain. The phylogenetic analysis places *Capsaspora*-p53/63/73 close to the choanoflagellate group (supplementary fig. S7, Supplementary Material online). The tree topology implies that the last common ancestor of holozoans had a single p53/63/73 gene, which followed independent divergences in vertebrates and choanoflagellates. In the absence of DNA damage, p53 appears to be downregulated by ubiquitination, which in vertebrates is carried out by the vertebrate-exclusive Mdm2 protein. However, other mechanisms of regulation have been proposed, such as ubiquitin

ligases or CREB-binding protein (CBP)/p300 (Shi et al. 2009). Interestingly, we identified CBP/p300 both in *Capsaspora* and *M. brevicollis* (see below), although whether CBP/p300 downregulates p53 in these holozoans remains unknown.

Stat

STAT proteins are TFs that, in response to a wide variety of extracellular signaling proteins, regulate the action of several genes that are involved in cell growth and homeostasis (Bromberg 2002; Levy and Darnell 2002). Structurally, STAT proteins have a N-terminal interacting domain, a STAT alpha domain with a coiled-coil structure involved in protein−protein interactions (e.g., it recruits HATs, specially CBP/p300), a STAT DNA-binding domain, a SH2 domain, and a C-terminal transactivation domain (Levy and Darnell 2002) (supplementary fig. S8, Supplementary Material online). The activation of STAT is mediated by the phosphorylation of a key tyrosine residue located after the SH2 domain (Levy and Darnell 2002). Our searches identified well-conserved STAT proteins in *Capsaspora*, *M. brevicollis*, and the apusozoan *T. trahens* (fig. 1). The STAT proteins from the latter two taxa appear, however, to be slightly truncated at the 5′ end (see supplementary fig. S8, Supplementary Material online). STAT proteins had previously been identified in amoebozoans (Kawata et al. 1997; Lee et al. 2008; Araki et al. 2010), but the protein domain analysis clearly showed that amoebozoan STAT are quite different from metazoan STAT proteins (supplementary fig. S8, Supplementary Material online). In contrast, the homologs from *M. brevicollis*, *Capsaspora*, and *T. trahens* are very similar to metazoan STATs. Moreover, a phylogenetic analysis of STATs using amoebozoan CudA proteins as outgroup (Yamada et al. 2008) showed amoebozoan-specific STATs as a sister-group to the holozoan + apusozoan clade (BV = 92%) (supplementary fig. S9, Supplementary Material online). As STAT proteins are present in extant apusozoans, these are likely to have been lost early in the fungal lineage.

Metazoan STAT proteins form part of the JAK signaling pathway, which is absent in nonmetazoan lineages (King et al. 2008). However, STAT proteins can interact with other receptor and nonreceptor tyrosine kinases (Kawata et al. 1997; Levy and Darnell 2002). Indeed, the distribution of STATs coincides with the distribution of tyrosine kinases among eukaryotes, being present in amoebozoans (Kawata et al. 1997; Goldberg et al. 2006), apusozoans and *Capsaspora* (Ruiz-Trillo I, unpublished data), choanoflagellates (King et al. 2008; Manning et al. 2008; Suga et al. 2008), and metazoans (Mayer 2008), all of which also have tyrosine kinases.

bZIP

bZIP TFs are named after the highly conserved structure containing a basic region and a leucine zipper (Hurst 1994). The bZIP proteins are ubiquitous among eukaryotes and are involved in several processes, such as environmental sensing and development (Deppmann et al. 2006). We have identified 25 and 15 bZIP proteins in *Capsaspora* and *M. brevicollis*, respectively. *Amphimedon queenslandica* has 20, and the average bilaterian, 38 (fig. 1). Interestingly, the chytrid fungus *A. macrogynus* has 43 bZIP genes, whereas most Dikarya have approximately 13 (fig. 1). We could only classify unambiguously seven and six of the bZIP proteins present in *Capsaspora* and *M. brevicollis*, respectively. A phylogenetic analysis including only the classified proteins showed that *Capsaspora* bZIPs correspond to PAR, C/EBP, Atf2, Oasis, Atf6, and CREB families, whereas the *Monosiga* homologs correspond to Atf4/5, Atf2, Oasis, and Atf6 families (fig. 3, see supplementary fig. S10, Supplementary Material online for a tree with all *Capsaspora* genes). Based on these analyses, we hypothesize that most, if not all, current metazoan bZIP families were present in the holozoan ancestor, with some of them subsequently being lost in choanoflagellates and *Capsaspora*. Some families, such as CREB, most likely underwent a protein domain rearrangement within metazoans, similarly to that described in other gene families (King et al. 2008; de Mendoza et al. 2010). Interestingly, all bZIP proteins that we identified in the unicellular relatives of metazoans belong to families that act strictly (Atf6, PAR, CREB, Oasis) or facultatively (Atf4/5, Atf2, C/EBP) as homodimers. This suggests that bZIP proteins in unicellular organisms may work mostly as homodimers, as already seen in yeast bZIP interactions (Deppmann et al. 2006). Our data suggest that although bZIP originated before the dawn of the Metazoa, their connectivity and combinatorial interactions may have increased in animals. For example, the *Capsaspora* homolog of CREB does not have the kinase-inducible activation domain that allows its interaction with p300/CBP (Giebler et al. 2000), even though p300/CBP is present in the *Capsaspora* genome.

bHLH

bHLH is a domain that is present in a large superfamily of TFs that are widespread among eukaryotes. In metazoans, they regulate critical developmental processes, such as neurogenesis, sex determination, myogenesis, and hematopoiesis (Jones 2004). This family of TFs has a DNA-binding basic region followed by two alpha helices separated by a variable loop region. Many bHLH proteins also include other domains that are involved in protein−protein interactions (Simionato et al. 2007). The bHLH proteins can act as homodimers or heterodimers to regulate gene expression. Metazoan bHLH have been grouped into six different higher order clades (A to F) (Simionato et al. 2007; Degnan et al. 2009) (for a general overview, see fig. 4). Group A, which includes genes such as MyoD and neurogenin, has only a bHLH domain and is exclusive to metazoans. Group B, which includes Myc or SREBP, has a leucine zipper 3′ to the bHLH domain and is found throughout the eukaryotes. Group C, which includes Clock and ARNT, has two PAS domains (PAS and PAS3) 3′ to the bHLH domain and is also thought to be exclusive to metazoans. Group D, which is metazoan specific, lacks the DNA-binding basic region, and hence, their members are unable to bind to DNA, acting as antagonists to Group A members. Most group

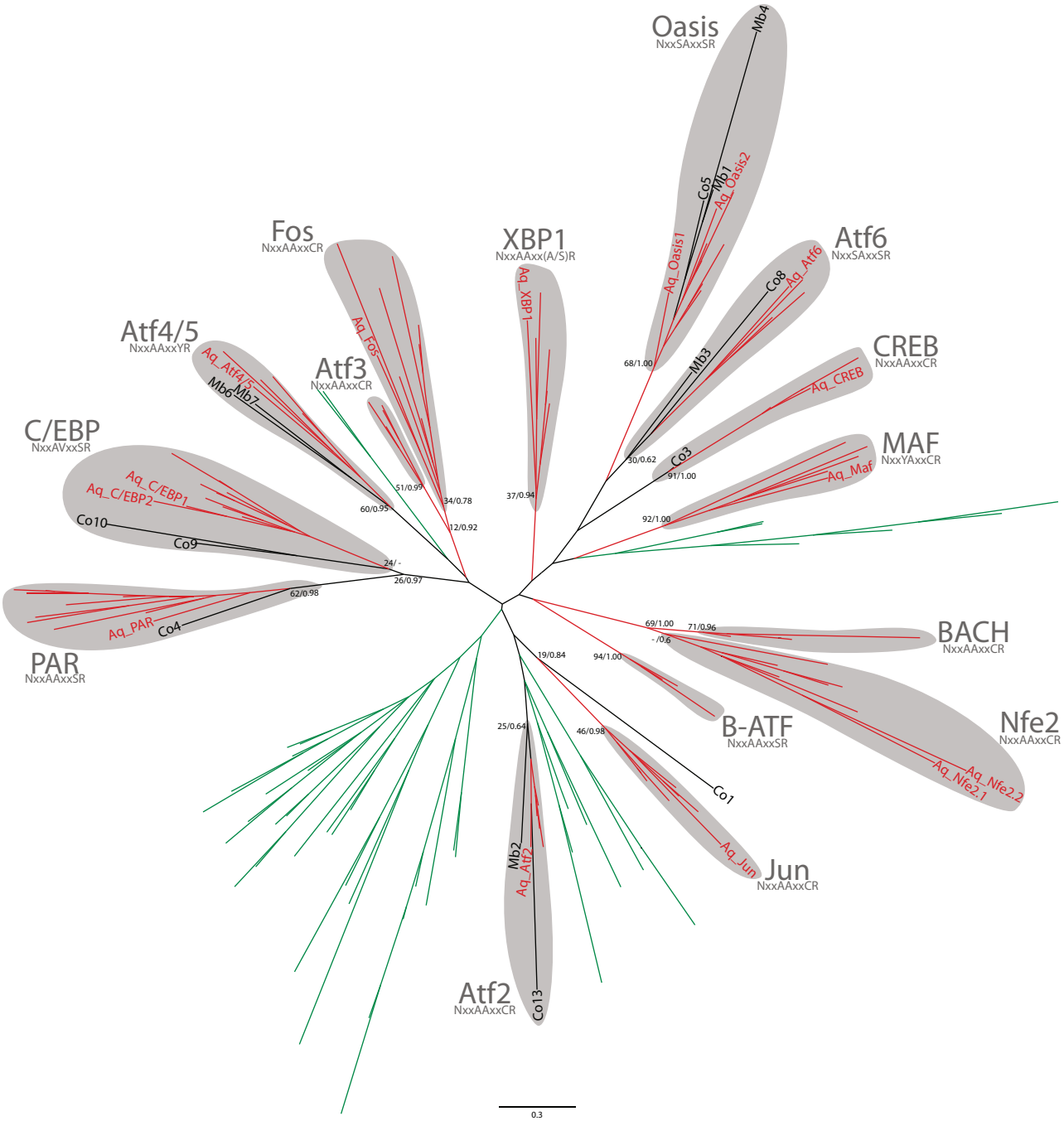
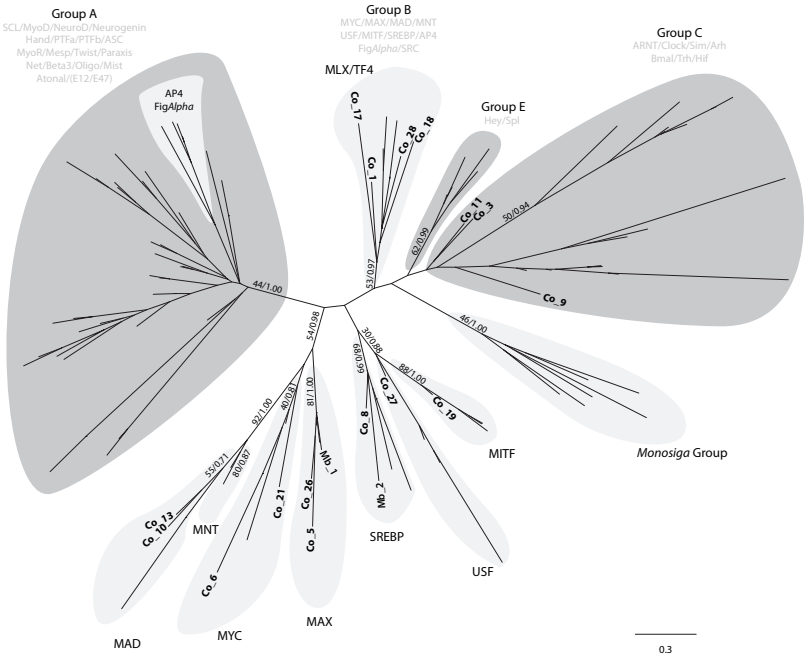


Fig. 3 ML tree of bZIP genes including the unambiguously assigned *Capsaspora* bZIP homologs. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 1,000 bootstrap replicates (BV) and BPP. Both values are shown on key branches. Aq (*Amphimedon queenslandica*), Co (*Capsaspora*), Mb (*Monosiga brevicollis*). Metazoan branches depicted in red and fungal branches in green. For each family, the signature sequence for DNA recognition is indicated and only proteins with this conserved motif are included in the family (Fujii et al. 2000). A tree including all *Capsaspora* bZIP genes is shown in supplementary figure S10 (Supplementary Material online).

E proteins include a metazoan-specific orange domain and a WRPW peptide in their carboxyl terminal part. Finally, Group F lacks the DNA-binding basic region but possesses a COE domain, which is involved both in dimerization and binding (Simionato et al. 2007).

We identified 31 bHLH proteins in *Capsaspora*, including orthologs of Myc, Mad, Max, SREBP, Mlx/TF4, MITF, and USF group B families, nonspecific homologs (ARNT-like) of group C, and some unclassifiable proteins (see figs. 1 and 4, supplementary figs. S11—S13, Supplementary Material online). This is more than double the bHLH genes found in *Monosiga* (11) and most fungi (around 10 in Dikarya). Compared with Metazoa, *Capsaspora* has a wider bHLH repertoire than the sponge *A. queenslandica* (18), but half what is present in cnidarians (72 in *N. vectensis*) or bilaterians (average of 69) (fig. 1). The choanoflagellate *M. brevicollis* contains Max, SREBP, Myc, and a lineage-specific group of bHLH genes. Thus, homologs of group C of

A



B

		Capsaspora owczarzaki	*Monosiga brevicollis*	*Amphimedon queenslandica*	*Nematostella vectensis*	*Bilateria*
Group A				•	•	•
Group B	MYC	•	•	•	•	•
	MAX	•	•	•	•	•
	MAD	•		•	•	•
	MNT				•	•
	MITF	•		•	•	•
	MLX/TF4	•		•	•	•
	SREBP	•	•	•	•	•
	USF	•		•	•	•
	AP4			•	•	•
	SRC				•	•
	Fig*Alpha*				•	•
Group C	ARNT/Bmal	•		•	•	•
	Ahr				•	•
	Clock			•	•	•
	Hif/Sim/Trh			•	•	•
Group D	Emc				•	•
Group E	Hey, H/ E(Spl)			•	•	•
Group F	Coe			•	•	•

Fig. 4 (*A*) ML tree of the bHLH domain including the unambiguously assigned *Capsaspora* bHLH homologs. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 100 bootstrap replicates (BV) and BPP. Both values are shown on key branches. Groups and families are defined by the classification of Simionato et al. (2007). The taxa used were *Homo sapiens*, *Nematostella vectensis*, *Amphimedon queenslandica*, *Monosiga brevicollis* (Mb) *and Capsaspora* (Co). For the sake of clarity, only the last two are specifically shown. (*B*) Table of the presence/absence of bHLH groups and families in some key taxa. *Amphimedon queenslandica*, *N. vectensis*, and *Bilateria* data obtained from Simionato et al. (2007). A tree including all *Capsaspora* bHLH genes is shown in supplementary figure S11 (Supplementary Material online).

bHLH were already present in the common ancestor of *Capsaspora*, choanoflagellates, and metazoans. Our data reveals that a basic Myc, MAX, Mxd/Mnt network of bHLH TFs was already present in the common ancestor of metazoans and *Capsaspora* and became more complex in multicellular lifestyles, incorporating, for example, Mnt and Mga. Interestingly, *Capsaspora* bHLH proteins are all homologs of those implicated in cell cycle and metabolism, and none of those are involved in differentiation. From our survey, we can also corroborate the two expansions periods (of bHLH groups and classes) in bHLH evolution previously inferred by Simionato et al. (2007) and later revised by Degnan et al. (2009), one before the divergence between *Capsaspora* and choanoflagellates + metazoans and another early in eumetazoan evolution (fig. 1). In contrast to bZIP TFs, there are some putative heterodimeric TF interactions among *Capsaspora* bHLH. For example, Myc, Mad, MAX, and Mlx can act as heterodimers in metazoans.

Mef2

Mef2 are the metazoan representatives of Type II MADS box genes. They are characterized by the presence of a Mef2 domain following the N-terminal MADS domain (Alvarez-Buylla et al. 2000). Mef2 genes play important roles in metazoan development, especially in the mesoderm (Potthoff and Olson 2007). Some authors considered Mef2 to be metazoan specific (Larroux et al. 2006; Degnan et al. 2009), although other authors had proposed *Saccharomyces* Smp1 and Rlm1 genes to be fungal homologs of metazoan Mef2 (Dodou and Treisman 1997). We identified canonical metazoan-type Mef2 in *Capsaspora* and in the cythrid fungi *S. punctatus* and *A. macrogynus* (fig. 1). The protein structure of *Capsaspora* and *S. punctatus* Mef2 closely resembles the canonical metazoan Mef2 (supplementary fig. S14, Supplementary Material online). We also identified a putative Mef2 homolog in *M. brevicollis* and in the amoebozoans *D. discoideum*, *D. purpureum*, *E. histolytica*, and *A. castellanii* as well as in the oomycetes *Phytophthora sojae*, *P. ramorum*, *P. infestans*, *and P. caspis*, although their sequences are divergent and have little similarity to the canonical metazoan Mef2 (supplementary fig. S14, Supplementary Material online). The fact that *Phytophthora* species encodes a Mef2 homolog may be explained by a lateral gene transfer (LGT) event because

they are the only analyzed eukaryotes outside opisthokonts and amoebozoans to have a mef2 gene. In fact, it has already been shown that some *Phytophthora* genes have a close relationship with amoebozoans genes (Tyler et al. 2006; Torruella et al. 2009). A phylogenetic analysis using fungal sequences as outgroup yields a clade that comprises all the taxa with a canonical metazoan-type Mef2 domain, that is all metazoans plus *Capsaspora*, *A. macrogynus*, and *S. punctatus* (supplementary fig. S15, Supplementary Material online). Our data show that the canonical metazoan Mef2 domain has a deeper origin than previously thought, with a conserved Mef2 domain present at least in the common ancestor of opisthokonts.

Fox

Fox genes TFs are characterized by the presence of a DNA-binding domain known as Forkhead box. Fox genes play important roles as regulators of both development and metabolism, and they seem to be specific to opisthokonts (Tuteja and Kaestner 2007a, 2007b; Shimeld et al. 2009). We identified four Fox genes in *Capsaspora*, none of them being part of the metazoan-specific class I but rather present in the supposedly opisthokont specific class II that also includes fungi (Larroux et al. 2008) (fig. 1 and supplementary fig. S16, Supplementary Material online). Interestingly, we identified three putative Fox genes in the amoebozoan *A. castellanii* (fig. 1), although their sequences are divergent compared with opisthokont ones. Thus, our results show that Fox genes are not specific to opisthokonts and were already present before the divergence of amoebozoans and opisthokonts.

HMG Box Genes

HMG box containing genes are TFs that are involved in genome stability, chromatin structure, and gene regulation (Stros et al. 2007). Metazoan-specific families are Sox and Tcf/Lef (Larroux et al. 2008). We characterized nine HMG box-containing proteins in *Capsaspora* (fig. 1 and supplementary fig. S17, Supplementary Material online), a similar number as those found in *Monosiga* (12) and Amoebozoa (average of 10) and significantly less than those found in Bilateria (average of 45). Two of *Capsaspora* HMG box genes have strong similarities to MATalpha box, typical sex-determinant genes that are present in Ascomycota (Fraser and Heitman 2003; Fraser et al. 2004). *Capsaspora* also encodes a HMG-B, a SSRP-1, and a SWI/SNF homolog, plus some HMG box containing genes that cannot confidently be assigned to any HMG box class.

Homeobox Genes

Homeobox genes encode an acid helix-turn-helix DNA-binding motif known as the homeodomain. Homeobox genes are known to have key roles in animal, plant, fungal, and amoebozoan development, such as regional patterning, regulation of cell proliferation, differentiation, adhesion, and migration (Gehring et al. 1994; Derelle et al. 2007). There are two large superfamilies, the canonical (non-TALE) class with a 60 amino acids homeodomain and the TALE superclass characterized by an insertion of three amino acids between helix 1 and 2 of the homeodomain (Mukherjee and Burglin 2007). Both TALE and non-TALE superclasses were already present in the ancestor of eukaryotes (Derelle et al. 2007). The two homeobox genes of the choanoflagellate *M. brevicollis* have already been characterized, both of them belonging to the TALE superclass, although they cannot confidently be assigned to any major metazoan homeobox family (King et al. 2008; Larroux et al. 2008). We identified nine homeodomain-containing genes in *Capsaspora*: three TALE and six non-TALE (fig. 1 and supplementary figs. S18—S22, Supplementary Material online). A phylogenetic analysis of these genes including members of all major families of homeodomains from metazoans, amoebozoans, and fungi failed to confidently assign *Capsaspora* homeobox genes to any of the major metazoan classes, except for one clear ortholog to the longevity assurance homolog (LAG-1) class. To further improve the resolution and classify the remaining *Capsaspora* homeobox genes, we performed phylogenetic analyses specific for TALE or non-TALE genes. This allowed us to assign one *Capsaspora* TALE homeobox gene to the PBC family, although it lacks the PBC N-terminal domain, and support is not very high. The remaining two *Capsaspora* TALE genes have an unclear phylogenetic relationship to other TALEs, although they appear to be closely related to the two *M. brevicollis* homeobox genes (supplementary fig. S19, Supplementary Material online). Interestingly, the sponge *A. queenslandica* appears not to have a homolog of PBC (supplementary fig. S19, Supplementary Material online) (Larroux et al. 2008). *Capsaspora* non-TALE genes appeared in unclear phylogenetic positions even with a restricted non-TALE only data set, although there is a potential homolog of LIM and two potential homologs of POU (supplementary fig. S20, Supplementary Material online). Thus, in order to classify them, we constructed different phylogenetic trees in which *Capsaspora* genes were forced to be members of a specific family and then we compared the likelihood values among all possible trees (for further details, see Material and Methods). Four *Capsaspora* non-TALE homologs appear to be at the root of the tree. Another one (*Capsaspora-6*) falls within the paired-like (Prd-like) clade with significant statistical support, this gene product possesses five of the six diagnostic amino acids of Prd-like genes (Galliot et al. 1999) (supplementary fig. S21, Supplementary Material online). However, it does not have the typical Q or K amino acid at position 50, and its intron is not located in the typical position (between codons 46 and 47), as consistently observed in metazoans. A specific phylogeny of ANTP, prd-like, LIM, and POU also supports this assignment but is not statistically significant (supplementary fig. S21, Supplementary Material online). The last *Capsaspora* non-TALE homeobox gene has a C-terminal TRAM LAG1 CLN8 (TLC) domain and a transmembrane domain, the characteristic domain architecture of the lass (longevity assurance homologs of yeast [Lag-1]) genes, which are considered to be homologs to fungal Lag genes. Interestingly,

phylogenetic analysis of the TLC domain showed that LAG genes with homeodomain are exclusive to metazoans and *Capsaspora*, whereas genes with the TLC domains and TRAM1 domain are found in amoebozoans, fungi, and metazoans (supplementary fig. S22, Supplementary Material online). Lass genes, however, are implicated in ceramide synthesis, the function of their homeodomain being unclear and their specific TF activity unknown (Teufel et al. 2009). Our data show that the repertoire of homeobox genes in metazoan unicellular relatives is larger than previously thought (see fig. 1), however, some specific homeobox gene classes, such as ANTP appear to be exclusive to the Metazoa. Genome data from additional unicellular relatives of metazoans will be needed to corroborate this.

CBP/p300

The CBP/p300 is a ubiquitous metazoan transcriptional coactivator that interacts with several TFs, acts as an acetyltransferase (Coutts and La Thangue 2005) and is involved in cell growth and development (Goodman and Smolik 2000). Specifically, CBP/p300 interacts with such TFs such as NF-kappaB (Perkins et al. 1997), Stat (Levy and Darnell 2002; Wojciak et al. 2009), Runx (Jin et al. 2004; Makita et al. 2008), p53 (Grossman 2001), CREB (Manna et al. 2009), and C/EBP (Manna et al. 2009). For example, CBP/p300 acetylates Runx genes (Jin et al. 2004) and ubiquitinates p53 (Shi et al. 2009). We have identified CBP/p300 homologs in both *Capsaspora* and *M. brevicollis*. This implies that CBP/p300 originated prior to the divergence of *Capsaspora* from choanoflagellates and metazoans. It is worth mentioning that this multifunctional cofactor seems to have evolved concomitant to the emergence of several holozoan TFs, such as Runx and NF-kappaB. This suggests that a relatively high level of regulatory complexity was already emerging on early in the holozoan lineage, well before the divergence of metazoan and choanoflagellate lineages.

LSF/GRH

The LSF/GRH family of TFs is characterized by the CP2 domain, and its members play important roles in bilaterians, being involved in vertebrate organogenesis, cell cycle progression, and cell survival and differentiation (Bray and Kafatos 1991; Uv et al. 1997; Veljkovic and Hansen 2004; Traylor-Knowles et al. 2010). LSF/GRH can be divided into two groups, the LSF/CP2 and the GRH subfamilies (Shirra and Hansen 1998; Traylor-Knowles et al. 2010). Members of the LSF/CP2 subfamily act as tetramers and possess an extra SAM domain C-terminal to the specific CP2 DNA-binding domain. Members of the GRH subfamily do not have the SAM domain and act as dimers.

The CP2 domain is present throughout the opisthokonts, including choanoflagellates (Traylor-Knowles et al. 2010) and seems to be a synapomorphy of this group of eukaryotes. Interestingly, it has been hypothesized that the GRH subfamily originated by duplication of an ancestral LSF/GRH-like gene at the origin of the Metazoa and was coopted to epidermal determination in metazoans (Traylor-Knowles et al. 2010). We identified two LSF/GRH genes in *Capsaspora*, a LSF-like and a GRH-like (fig. 1 and supplementary fig. S23, Supplementary Material online), although the *Capaspora* LSF-like gene lacks the characteristic C-terminal SAM domain found in metazoan LSF proteins. This domain may have been gained by domain shuffling before the split between metazoans and choanoflagellates, although the loss of this domain in *Capsaspora* cannot be ruled out. Our findings imply that the duplication of the LSF/GRH gene occurred before *Capsaspora* diversified from choanoflagellates and metazoans, and that GRH was lost in choanoflagellates. Thus, the presence of a GRH gene antedates the origin of the metazoan epithelium.

NRs, Smad, and Ets

Our data show that these three TFs families remain, at this time, metazoan specific because we did not identify any homologs in nonmetazoan taxa. The complete genome sequences of additional nonmetazoan taxa are needed to corroborate this hypothesis.

Origin and Early Evolution of Metazoan TFs

The repertoire of TFs in the holozoan *C. owczarzaki* reported here, and its comparison with metazoan, fungal, and choanoflagellate TFs provides important insights into the origin and evolution of TFs that are essential for metazoan multicellularity. This allows us to propose a new hypothesis regarding the origin of key metazoan TFs (see fig. 5). Some metazoan TF domains have deep origins being widespread in eukaryotes, such as HMG box, homeodomain (both TALE and non-TALE), bHLH, bZIP, or Mef2-like (see also Degnan et al. 2009). However, major diversifications of genes encoding some of these domains took place along metazoan and fungal stems (see fig. 1), generating lineage-specific classes and subfamilies. In regards to metazoan-specific TF gene families, there appears to have been two major expansions (fig. 5): one prior to the divergence of *Capsaspora*, choanoflagellates, and the Metazoa (e.g., in bZIP and bHLH) and another within the metazoan lineage (such as Sox, homeodomains, and further diversification of bZIP and bHLH). Several other TF domains, such as Churchill, STAT, and, most likely, Fox, were already present in the common ancestor of unikonts (i.e., amoebozoans, apusozoans, and opisthokonts). This finding changes previous views in which Churchill was considered exclusive to metazoans and Fox exclusive to opisthokonts (although the assignment of *A. castellanii* hits to Fox remain contentious). Although STAT domains were present in the common ancestor of unikonts, our data show that canonical metazoan-type STAT seem to be exclusive to apusozoans (*T. trahens*) and opisthokonts. A major challenge to previous proposals that T-box genes are metazoan innovations is the discovery of T-box genes in *S. punctatus* and *Capsaspora*. This means that T-box genes appeared before the divergence of fungi and holozoans. What role are these T-box genes playing in these nonmetazoan lineages remains to be studied.

Interestingly, some TFs appear to have evolved prior to the divergence of *Capsaspora* from choanoflagellates and metazoans, such as p53, Runx, and NF-kappaB; the latter

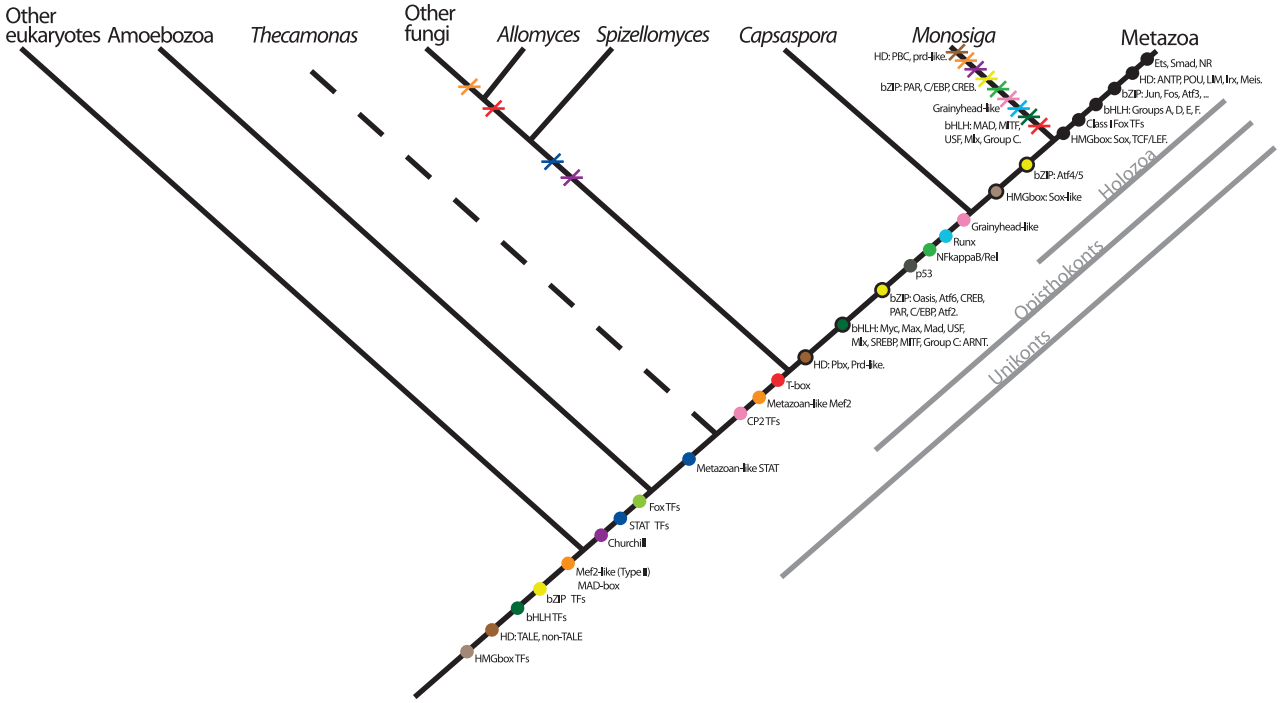


Fig. 5 Cladogram representing TF evolution among the analyzed taxa. Colors are unique for each domain class. A colored dot means the hypothetical origin of the domain. A black-circled dot indicates where a specific protein family appears in our taxon sampling. A cross means the loss of the domain or specific protein family in a lineage. Metazoan apomorphies are shown as black dots. The phylogenetic relationships are based on several recent studies (Burki et al. 2008; Ruiz-Trillo et al. 2008; Brown et al. 2009; Liu et al. 2009; Minge et al. 2009).

two previously being considered metazoan specific. This pattern of gene families that are relevant to metazoan multicellularity evolving prior to the emergence of the metazoan stem lineage is not new and has been observed in other cases, such as tyrosine kinases, cadherins, MAGUKs, and integrins (Abedin and King 2008; King et al. 2008; Manning et al. 2008; Suga et al. 2008; Degnan et al. 2009; de Mendoza et al. 2010; Sebe-Pedros et al. 2010). Finally, there are some TF domains that, under the current taxon sampling, appear to be metazoan innovations. These are ETS, Smad, and NRs. Moreover, some specific homeobox genes (ANTP, LIM, POU, Irx, Meis, Tgif, Six), bZIP classes (e.g., Jun, Fos), bHLH classes (A, D—F groups), and HMG box classes (Sox, TCL/lef) appear also to be metazoan specific (fig. 5), although we cannot rule out the possibility that some of these may have a more ancient origin and secondarily lost in nonmetazoan lineages. This new evolutionary scenario implies that significant lineage-specific TFs losses occurred within the choanoflagellate lineage. For example, Runx, T-box, RHD domain, GRH-like, and Churchill appear to have been lost in *M. brevicollis*. Whether this is specific to one choanoflagellate lineage (that of *M. brevicollis*) or to choanoflagellates in general remains unknown. Only genomic data from additional choanoflagellate taxa will resolve this issue. A similar pattern of lineage-specific loss in choanoflagellates has recently been shown for the integrin-mediated adhesion machinery (Sebe-Pedros et al. 2010).

A quantitative analysis (fig. 1) of TFs evolution suggests that several expansions occurred in Eumetazoa, such as bHLH and homeobox gene families and to a lesser degree HMG box and bZIP families. Specific domain expansions have already been reported in the Viridiplantae for bHLH and homeodomain proteins (Mukherjee et al. 2009; Pires and Dolan 2010). There are several theories about the correlation of these expansions with the transition to multicellularity (Derelle et al. 2007; Pires and Dolan 2010). On the other hand, some TF domains, such as CP2, Runt, MADS-box, Churchill, p53, and STAT, have similar number of members in unicellular and multicellular holozoans. *Capsaspora* TF complexity is quite high, with a wider range of bHLH and bZIP domain-containing proteins than in some early-branching metazoans such as *A. queenslandica* or *T. adhaerens*. Because *Capsaspora* has a complex (and not fully understood) life cycle, in which there is a symbiotic stage within the mollusc *Biomphalaria glabrata*, one may wonder whether the complexity of TFs identified in *Capsaspora* is due to LGT from the host or even from the trematode flatworm *S. mansoni*, a metazoan parasite of *B. glabrata*. Based on our phylogenetic analyses, we do not favor this hypothesis. None of the phylogenetic trees shown (all including bilaterians; some even *B. glabrata* homologs) show the *Capsaspora* homolog grouping closer to bilaterians than to other metazoans. Instead, we hypothesize that the common ancestor of *Capsaspora*, choanoflagellates, and metazoans had a richer TF repertoire than previously believed and that some TFs were subsequently lost in the choanoflagellate lineage (or at least in *M. brevicollis*).

In summary, our results show that the evolution of metazoan TFs includes the acquisition of new genes (some

of them via domain shuffling), gene cooption, and the diversification of ancestral domains increasing the combinatorial complexity. How these metazoan developmental TFs are functioning in unicellular organisms and how they were exapted into new functions in multicellular animals remains to be answered.

Supplementary Material

See supplementary material file 1 for figures S1—S7; file 2 for figures S8—S17; and file 3 for figures S18—S23. Supplementary material file 4 includes the annotation of the *Capsaspora* sequences included in this study. They are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The genome sequences of *C. owczarzaki*, *A. macrogynus*, *S. punctatus*, and *T. trahens* are being determined by the Broad Institute of MIT/Harvard University under the auspices of the National Human Genome Research Institute and within the UNICORN initiative. We thank JGI, BI, and BCM for making data publicly available. We thank Kim C. Worley and the team of the *A. castellanii* genome project for accession to the genome data. We thank Peter W. H. Holland for help in the assignment of homeodomain-containing proteins and Romain Derelle, Jordi Paps, and Hiroshi Suga for helpful insights. This work was supported by an ICREA contract, an European Research Council Starting Grant (ERC-2007-StG-206883), and a grant (BFU2008-02839/BMC) from Ministerio de Ciencia e Innovación (MICINN) to I.R.-T. A.S.'s salary was supported by a pregraduate FPU grant and A.d.M.'s salary from a FPI grant both from MICINN. B.F.L.'s and B.M.D.'s contributions were supported by the Canadian Research Chair Program and the Australian Research Council, respectively.

References

Abedin M, King N. 2008. The premetazoan ancestry of cadherins. *Science* 319:946—948.

Adell T, Grebenjuk VA, Wiens M, Muller WE. 2003. Isolation and characterization of two T-box genes from sponges, the phylogenetically oldest metazoan taxon. *Dev Genes Evol.* 213:421—434.

Adell T, Muller WE. 2005. Expression pattern of the Brachyury and Tbx2 homologues from the sponge Suberites domuncula. *Biol Cell.* 97:641—650.

Akamatsu Y, Ohno T, Hirota K, Kagoshima H, Yodoi J, Shigesada K. 1997. Redox regulation of the DNA binding activity in transcription factor PEBP2. The roles of two conserved cysteine residues. *J Biol Chem.* 272:14497—14500.

Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, Ribas de Pouplana L, Martinez-Castilla L, Yanofsky MF. 2000. An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proc Natl Acad Sci U S A.* 97:5328—5333.

Araki T, van Egmond WN, van Haastert PJ, Williams JG. 2010. Dual regulation of a Dictyostelium STAT by cGMP and Ca2+ signalling. *J Cell Sci.* 123:837—841.

Bielen H, Oberleitner S, Marcellini S, Gee L, Lemaire P, Bode HR, Rupp R, Technau U. 2007. Divergent functions of two ancient Hydra Brachyury paralogues suggest specific roles for their C-terminal domains in tissue fate induction. *Development* 134:4187—4197.

Bray SJ, Kafatos FC. 1991. Developmental function of Elf-1: an essential transcription factor during embryogenesis in Drosophila. *Genes Dev.* 5:1672—1683.

Bromberg J. 2002. Stat proteins and oncogenesis. *J Clin Invest.* 109:1139—1142.

Brown MW, Spiegel FW, Silberman JD. 2009. Phylogeny of the "forgotten" cellular slime mold, Fonticula alba, reveals a key evolutionary branch within Opisthokonta. *Mol Biol Evol.* 26:2699—2709.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 268:78—94.

Burki F, Shalchian-Tabrizi K, Pawlowski J. 2008. Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. *Biol Lett.* 4:366—369.

Coffman JA. 2003. Runx transcription factors and the developmental balance between cell proliferation and differentiation. *Cell Biol Int.* 27:315—324.

Coutts AS, La Thangue NB. 2005. The p53 response: emerging levels of co-factor complexity. *Biochem Biophys Res Commun.* 331:778—785.

de Mendoza A, Suga H, Ruiz-Trillo I. 2010. Evolution of the MAGUK protein gene family in premetazoan lineages. *BMC Evol Biol.* 10:93.

Degnan BM, Vervoort M, Larroux C, Richards GS. 2009. Early evolution of metazoan transcription factors. *Curr Opin Genet Dev.* 19:591—599.

Deppmann CD, Alvania RS, Taparowsky EJ. 2006. Cross-species annotation of basic leucine zipper factor interactions: insight into the evolution of closed interaction networks. *Mol Biol Evol.* 23:1480—1492.

Derelle R, Lopez P, Guyader HL, Manuel M. 2007. Homeodomain proteins belong to the ancestral molecular toolkit of eukaryotes. *Evol Dev.* 9:212—219.

Dodou E, Treisman R. 1997. The Saccharomyces cerevisiae MADS-box transcription factor Rlm1 is a target for the Mpk1 mitogen-activated protein kinase pathway. *Mol Cell Biol.* 17:1848—1859.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* 14:755—763.

Espinosa JM. 2008. Mechanisms of regulatory diversity within the p53 transcriptional network. *Oncogene* 27:4013—4023.

Fraser JA, Diezmann S, Subaran RL, Allen A, Lengeler KB, Dietrich FS, Heitman J. 2004. Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms. *PLoS Biol.* 2:e384.

Fraser JA, Heitman J. 2003. Fungal mating-type loci. *Curr Biol.* 13:R792—R795.

Fujii Y, Shimizu T, Toda T, Yanagida M, Hakoshima T. 2000. Structural basis for the diversity of DNA recognition by bZIP transcription factors. *Nat Struct Biol.* 7:889—893.

Galliot B, de Vargas C, Miller D. 1999. Evolution of homeobox genes: Q50 paired-like genes founded the paired class. *Dev Genes Evol.* 209:186—197.

Gauthier M, Degnan BM. 2008. The transcription factor NF-kappaB in the demosponge Amphimedon queenslandica: insights on the evolutionary origin of the Rel homology domain. *Dev Genes Evol.* 218:23—32.

Gehring WJ, Affolter M, Burglin T. 1994. Homeodomain proteins. *Annu Rev Biochem.* 63:487—526.

Giebler HA, Lemasson I, Nyborg JK. 2000. p53 recruitment of CREB binding protein mediated through phosphorylated CREB: a novel

pathway of tumor suppressor regulation. *Mol Cell Biol.* 20:4849–4858.
Goldberg JM, Manning G, Liu A, Fey P, Pilcher KE, Xu Y, Smith JL. 2006. The dictyostelium kinome—analysis of the protein kinases from a simple model organism. *PLoS Genet.* 2:e38.
Goodman RH, Smolik S. 2000. CBP/p300 in cell growth, transformation, and development. *Genes Dev.* 14:1553–1577.
Grossman SR. 2001. p300/CBP/p53 interaction and regulation of the p53 response. *Eur J Biochem.* 268:2773–2778.
Hayden MS, Ghosh S. 2004. Signaling to NF-kappaB. *Genes Dev.* 18:2195–2224.
Hurst HC. 1994. Transcription factors. 1: bZIP proteins. *Protein Profile.* 1:123–168.
Jin YH, Jeon EJ, Li QL, Lee YH, Choi JK, Kim WJ, Lee KY, Bae SC. 2004. Transforming growth factor-beta stimulates p300-dependent RUNX3 acetylation, which inhibits ubiquitination-mediated degradation. *J Biol Chem.* 279:29409–29417.
Jones S. 2004. An overview of the basic helix-loop-helix proteins. *Genome Biol.* 5:226.
Katoh K, Kuma K, Miyata T, Toh H. 2005. Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Inform.* 16:22–33.
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.
Kawata T, Shevchenko A, Fukuzawa M, Jermyn KA, Totty NF, Zhukovskaya NV, Sterling AE, Mann M, Williams JG. 1997. SH2 signaling in a lower eukaryote: a STAT protein that regulates stalk cell differentiation in dictyostelium. *Cell* 89:909–916.
King N, Westbrook MJ, Young SL, et al. (37 co-authors). 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* 451:783–788.
Larroux C, Luke GN, Koopman P, Rokhsar DS, Shimeld SM, Degnan BM. 2008. Genesis and expansion of metazoan transcription factor gene classes. *Mol Biol Evol.* 25:980–996.
Larroux C, Fahey B, Liubicich D, Hinman V, Gauthier MF, Gongora M, Green K, WoÃàrheide G, Leys S, Degnan BP. 2006. Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. *Evol Dev.* 8:150–173.
Lee NSM, Rodriguez M, Kim B, Kim L. 2008. Dictyostelium kinase DPYK3 negatively regulates STATc signaling in cell fate decision. *Dev Growth Differ.* 50:607–613.
Levy DE, Darnell JEJ. 2002. Stats: transcriptional control and biological impact. *Nat Rev Mol Cell Biol.* 3:651–662.
Liu Y, Steenkamp ET, Brinkmann H, Forget L, Philippe H, Lang BF. 2009. Phylogenomic analyses predict sistergroup relationship of nucleariids and fungi and paraphyly of zygomycetes with significant support. *BMC Evol Biol.* 9:272.
Londin ER, Mentzer L, Sirotkin HI. 2007. Churchill regulates cell movement and mesoderm specification by repressing Nodal signaling. *BMC Dev Biol.* 7:120.
Macian F. 2005. NFAT proteins: key regulators of T-cell development and function. *Nat Rev Immunol.* 5:472–484.
Makita N, Suzuki M, Asami S, Takahata R, Kohzaki D, Kobayashi S, Hakamazuka T, Hozumi N. 2008. Two of four alternatively spliced isoforms of RUNX2 control osteocalcin gene expression in human osteoblast cells. *Gene* 413:8–17.
Manna PR, Dyson MT, Stocco DM. 2009. Role of basic leucine zipper proteins in transcriptional regulation of the steroidogenic acute regulatory protein gene. *Mol Cell Endocrinol.* 302:1–11.
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.
Marcellini S, Technau U, Smith JC, Lemaire P. 2003. Evolution of Brachyury proteins: identification of a novel regulatory domain conserved within Bilateria. *Dev Biol.* 260:352–361.
Mayer BJ. 2008. Clues to the evolution of complex signaling machinery. *PNAS* 105:9453–9454.
Minge MA, Silberman JD, Orr RJ, Cavalier-Smith T, Shalchian-Tabrizi K, Burki F, Skjaeveland A, Jakobsen KS. 2009. Evolutionary position of breviate amoebae and the primary eukaryote divergence. *Proc Biol Sci.* 276:597–604.
Mukherjee K, Brocchieri L, Burglin TR. 2009. A comprehensive classification and evolutionary analysis of plant homeobox genes. *Mol Biol Evol.* 26:2775–2794.
Mukherjee K, Burglin TR. 2007. Comprehensive analysis of animal TALE homeobox genes: new conserved motifs and cases of accelerated evolution. *J Mol Evol.* 65:137–153.
Muller CW, Herrmann BG. 1997. Crystallographic structure of the T domain-DNA complex of the Brachyury transcription factor. *Nature* 389:884–888.
Nedelcu AM, Tan C. 2007. Early diversification and complex evolutionary history of the p53 tumor suppressor gene family. *Dev Genes Evol.* 217:801–806.
Perkins ND, Felzien LK, Betts JC, Leung K, Beach DH, Nabel GJ. 1997. Regulation of NF-kappaB by cyclin-dependent kinases associated with the p300 coactivator. *Science* 275:523–527.
Pires N, Dolan L. 2010. Origin and diversification of basic-helix-loop-helix proteins in plants. *Mol Biol Evol.* 27:862–874.
Potthoff MJ, Olson EN. 2007. MEF2: a central regulator of diverse developmental programs. *Development* 134:4131–4140.
Putnam NH, Srivastava M, Hellsten U, et al. (20 co-authors). 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317:86–94.
Rennert J, Coffman JA, Mushegian AR, Robertson AJ. 2003. The evolution of Runx genes I. A comparative study of sequences from phylogenetically diverse model organisms. *BMC Evol Biol.* 3:4.
Robertson AJ, Larroux C, Degnan BM, Coffman JA. 2009. The evolution of Runx genes II. The C-terminal Groucho recruitment motif is present in both eumetazoans and homoscleromorphs but absent in a haplosclerid demosponge. *BMC Res Notes.* 2:59.
Rokas A. 2008. The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annu Rev Genet.* 42:235–251.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England).* 19:1572–1574.
Ruiz-Trillo I, Burger G, Holland PW, King N, Lang BF, Roger AJ, Gray MW. 2007. The origins of multicellularity: a multi-taxon genome initiative. *Trends Genet.* 23:113–118.
Ruiz-Trillo I, Inagaki Y, Davis LA, Sperstad S, Landfald B, Roger AJ. 2004. Capsaspora owczarzaki is an independent opisthokont lineage. *Curr Biol.* 14(22):R946–R947.
Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF. 2008. A phylogenomic investigation into the origin of metazoa. *Mol Biol Evol.* 25:664–672.
Scholz CB, Technau U. 2003. The ancestral role of Brachyury: expression of NemBra1 in the basal cnidarian Nematostella vectensis (Anthozoa). *Dev Genes Evol.* 212:563–570.
Sebe-Pedros A, Roger AJ, Lang FB, King N, Ruiz-Trillo I. 2010. Ancient origin of the integrin-mediated adhesion and signaling machinery. *Proc Natl Acad Sci U S A.* 107:10142–10147.
Shalchian-Tabrizi K, Minge MA, Espelund M, Orr R, Ruden T, Jakobsen KS, Cavalier-Smith T. 2008. Multigene phylogeny of choanozoa and the origin of animals. *PLoS One.* 3:e2098.
Sheng G, dos Reis M, Stern CD. 2003. Churchill, a zinc finger transcriptional activator, regulates the transition between gastrulation and neurulation. *Cell* 115:603–613.

Shi D, Pop MS, Kulikov R, Love IM, Kung AL, Grossman SR. 2009. CBP and p300 are cytoplasmic E4 polyubiquitin ligases for p53. *Proc Natl Acad Sci U S A*. 106:16275–16280.

Shimeld SM, Degnan B, Luke GN. 2010. Evolutionary genomics of the Fox genes: origin of gene families and the ancestry of gene clusters. *Genomics*. 95(5):256–260.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247.

Shirra MK, Hansen U. 1998. LSF and NTF-1 share a conserved DNA recognition motif yet require different oligomerization states to form a stable protein-DNA complex. *J Biol Chem*. 273: 19260–19268.

Simionato E, Ledent V, Richards G, Thomas-Chollier M, Kerner P, Coornaert D, Degnan BM, Vervoort M. 2007. Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics. *BMC Evol Biol*. 7:33.

Smith J. 1999. T-box genes: what they do and how they do it. *Trends Genet*. 15:154–158.

Srivastava M, Simakov O, Chapman J, et al. (33 co-authors). 2010. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* 466:720–726.

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

Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res*. 33:W465–W467.

Stros M, Launholt D, Grasser KD. 2007. The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. *Cell Mol Life Sci*. 64:2590–2606.

Suga H, Sasaki G, Kuma K, Nishiyori H, Hirose N, Su ZH, Iwabe N, Miyata T. 2008. Ancient divergence of animal protein tyrosine kinase genes demonstrated by a gene family tree including choanoflagellate genes. *FEBS Lett*. 582:815–818.

Sullivan JC, Sher D, Eisenstein M, Shigesada K, Reitzel AM, Marlow H, Levanon D, Groner Y, Finnerty JR, Gat U. 2008. The evolutionary origin of the Runx/CBFbeta transcription factors—studies of the most basal metazoans. *BMC Evol Biol*. 8:228.

Teufel A, Maass T, Galle PR, Malik N. 2009. The longevity assurance homologue of yeast lag1 (Lass) gene family (review). *Int J Mol Med*. 23:135–140.

Torruella G, Suga H, Riutort M, Pereto J, Ruiz-Trillo I. 2009. The evolutionary history of lysine biosynthesis pathways within eukaryotes. *J Mol Evol*. 69:240–248.

Traylor-Knowles N, Hansen U, Dubuc TQ, Martindale MQ, Kaufman L, Finnerty JR. 2010. The evolutionary diversification of LSF and Grainyhead transcription factors preceded the radiation of basal animal lineages. *BMC Evol Biol*. 10:101.

Tuteja G, Kaestner KH. 2007a. Forkhead transcription factors II. *Cell* 131:192.

Tuteja G, Kaestner KH. 2007b. SnapShot: forkhead transcription factors I. *Cell* 130:1160.

Tyler BM, Tripathy S, Zhang X, et al. (54 co-authors). 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313:1261–1266.

Uv AE, Harrison EJ, Bray SJ. 1997. Tissue-specific splicing and functions of the Drosophila transcription factor Grainyhead. *Mol Cell Biol*. 17:6727–6735.

Veljkovic J, Hansen U. 2004. Lineage-specific and ubiquitous biological roles of the mammalian transcription factor LSF. *Gene* 343:23–40.

Wheeler JC, Shigesada K, Gergen JP, Ito Y. 2000. Mechanisms of transcriptional regulation by Runt domain proteins. *Semin Cell Dev Biol*. 11:369–375.

Wojciak JM, Martinez-Yamout MA, Dyson HJ, Wright PE. 2009. Structural basis for recruitment of CBP/p300 coactivators by STAT1 and STAT2 transactivation domains. *EMBO J*. 28:948–958.

Yamada Y, Wang HY, Fukuzawa M, Barton GJ, Williams JG. 2008. A new family of transcription factors. *Development* 135: 3093–3101.

SUPPLEMENTARY MATERIAL 1

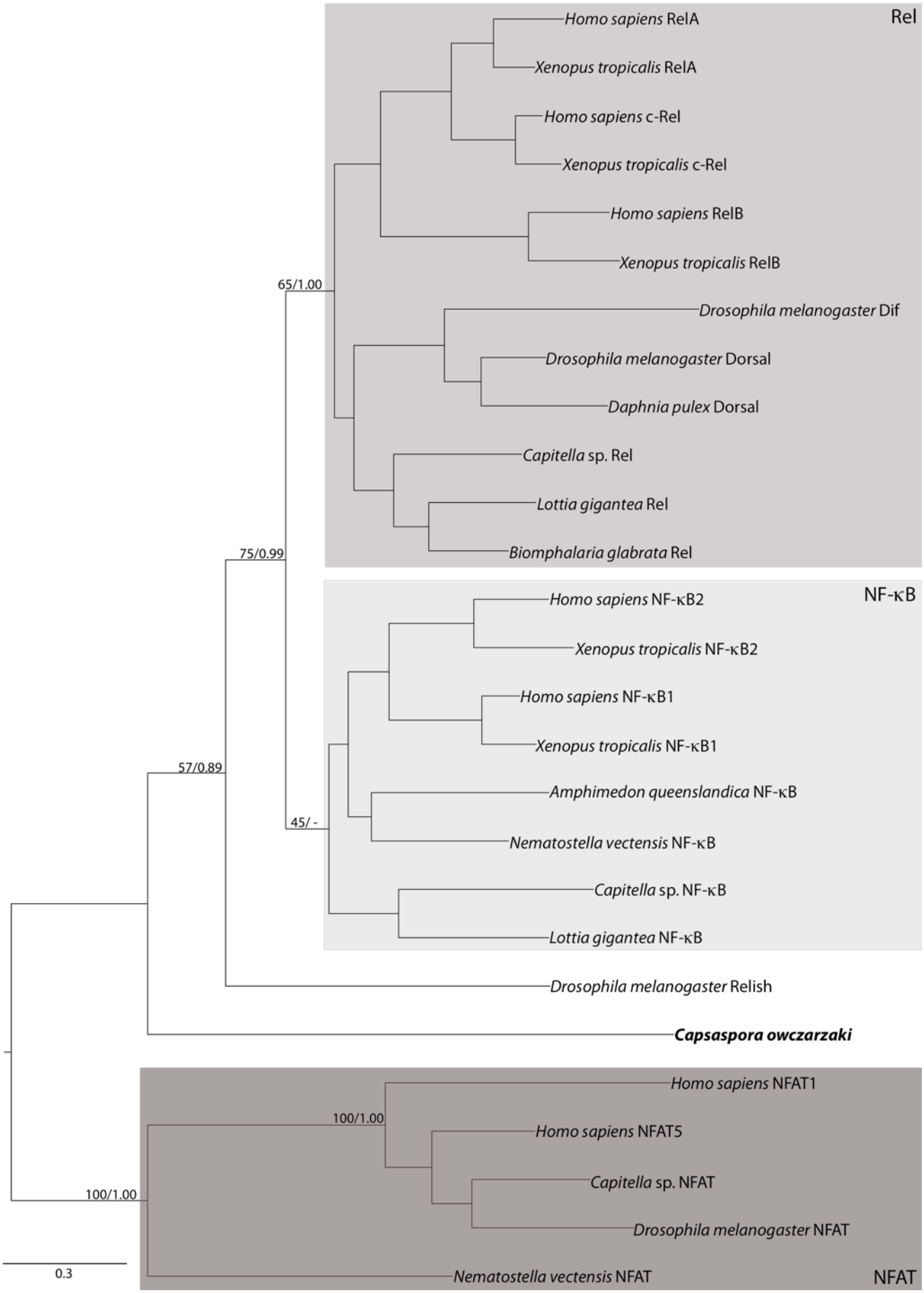
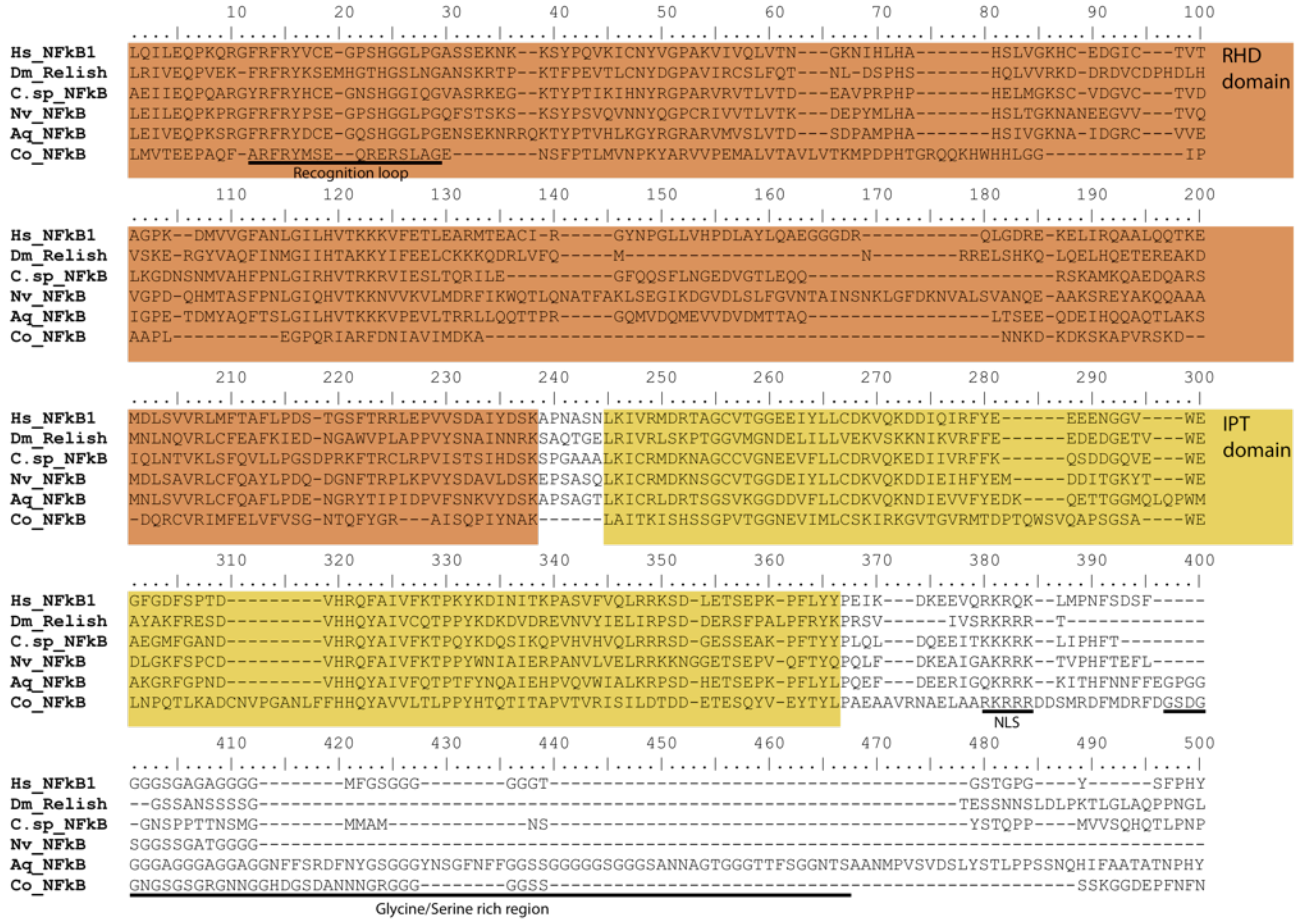
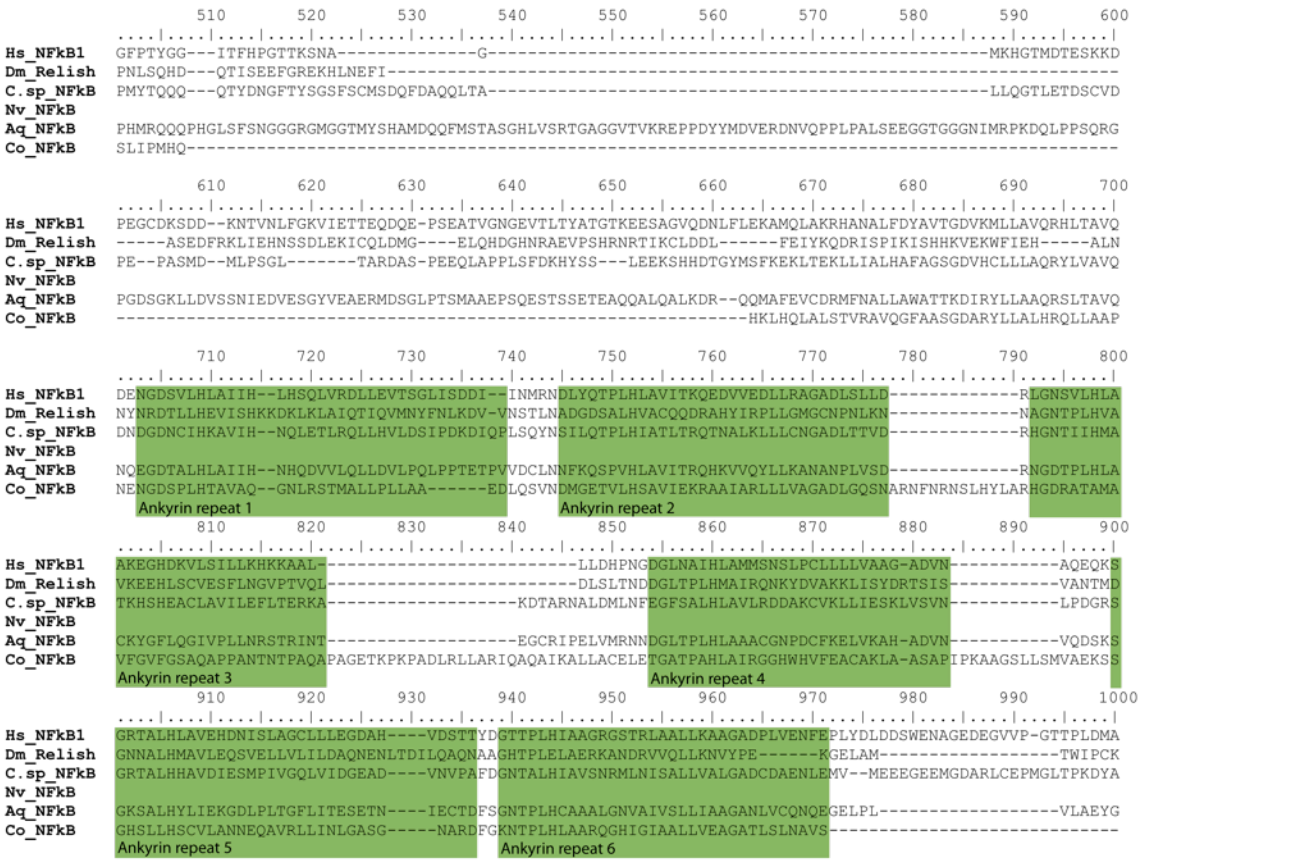


Figure S1. Maximum likelihood tree of RHD domain containing proteins NFAT, NF-kappaB and Rel using the common RHD domain. The tree is rooted using NFAT proteins. Statistical support was obtained by RaxML 100-bootstrap replicates (BV) and Bayes Posterior Probabilities (BPP). Both values are shown on key branches.


RHD domain
IPT domain
Recognition loop
NLS
Glycine/Serine rich region


Ankyrin repeat 1
Ankyrin repeat 2
Ankyrin repeat 3
Ankyrin repeat 4
Ankyrin repeat 5
Ankyrin repeat 6

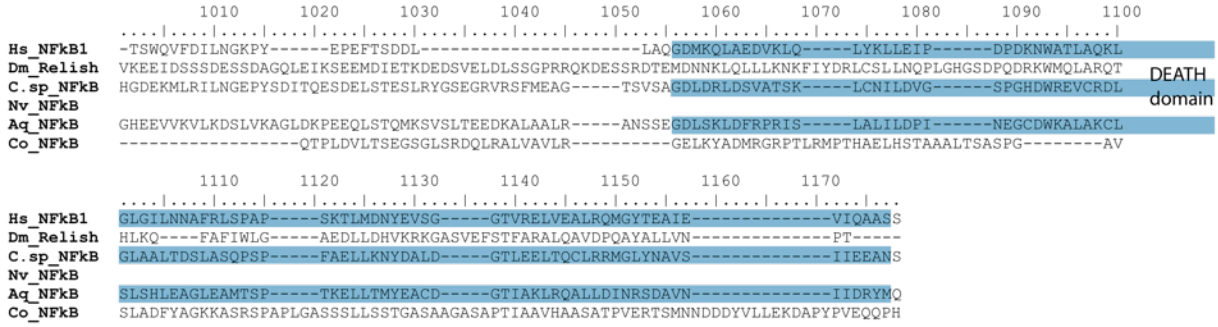


Figure S2. Alignment of NF-kappaB. Taxa include Hs (*Homo sapiens*), Dm (*Drosophila melanogaster*), C.sp (*Capitella sp.*), Nv (*Nematostella vectensis*), Aq (*Amphimedon queenslandica*) and Co (*Capsaspora owczarzaki*). Different Pfam domains (colored) and functional motifs (underlined) are shown. NLS: nuclear localization signal. See main text for further details.

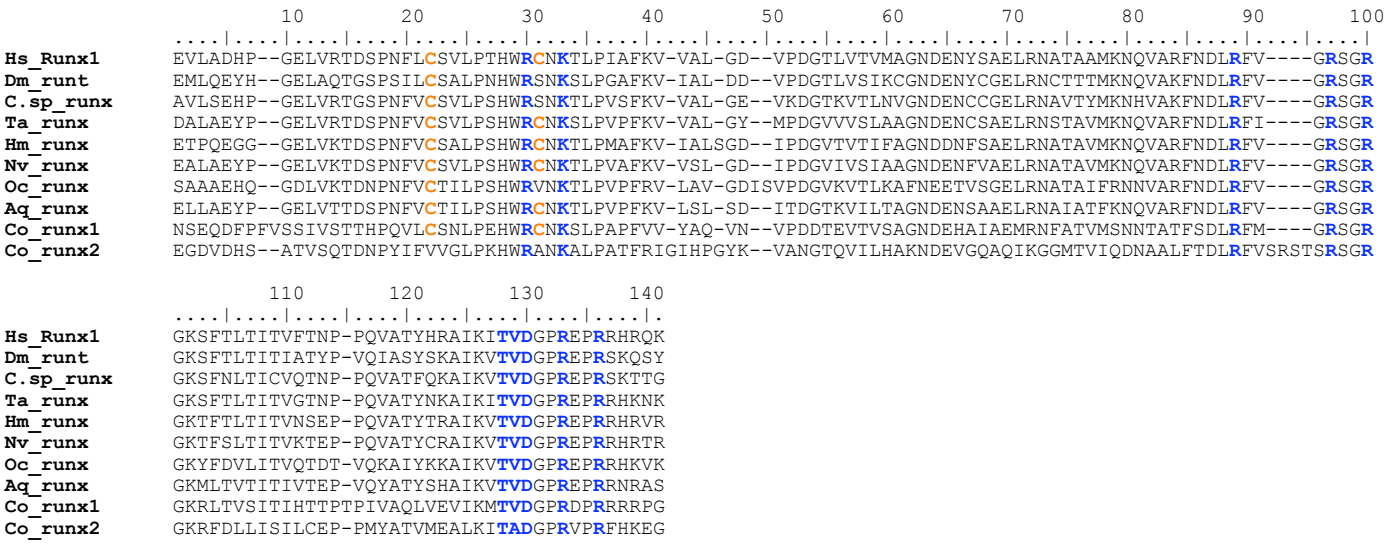


Figure S3. Alignment of the Runt domain. Taxa include Hs (*Homo sapiens*), Dm (*Drosophila melanogaster*), C.sp (*Capitella sp.*), Ta (*Trichoplax adhaerens*), Hm (*Hydra magnipapillata*), Nv (*Nematostella vectensis*), Oc (*Oscarella carmela*), Aq (*Amphimedon queenslandica*) and Co (*Capsaspora owczarzaki*). Key DNA binding aminoacids are highlighted in blue and the two Cys residues involved in redox binding affinity regulation are highlighted in orange.

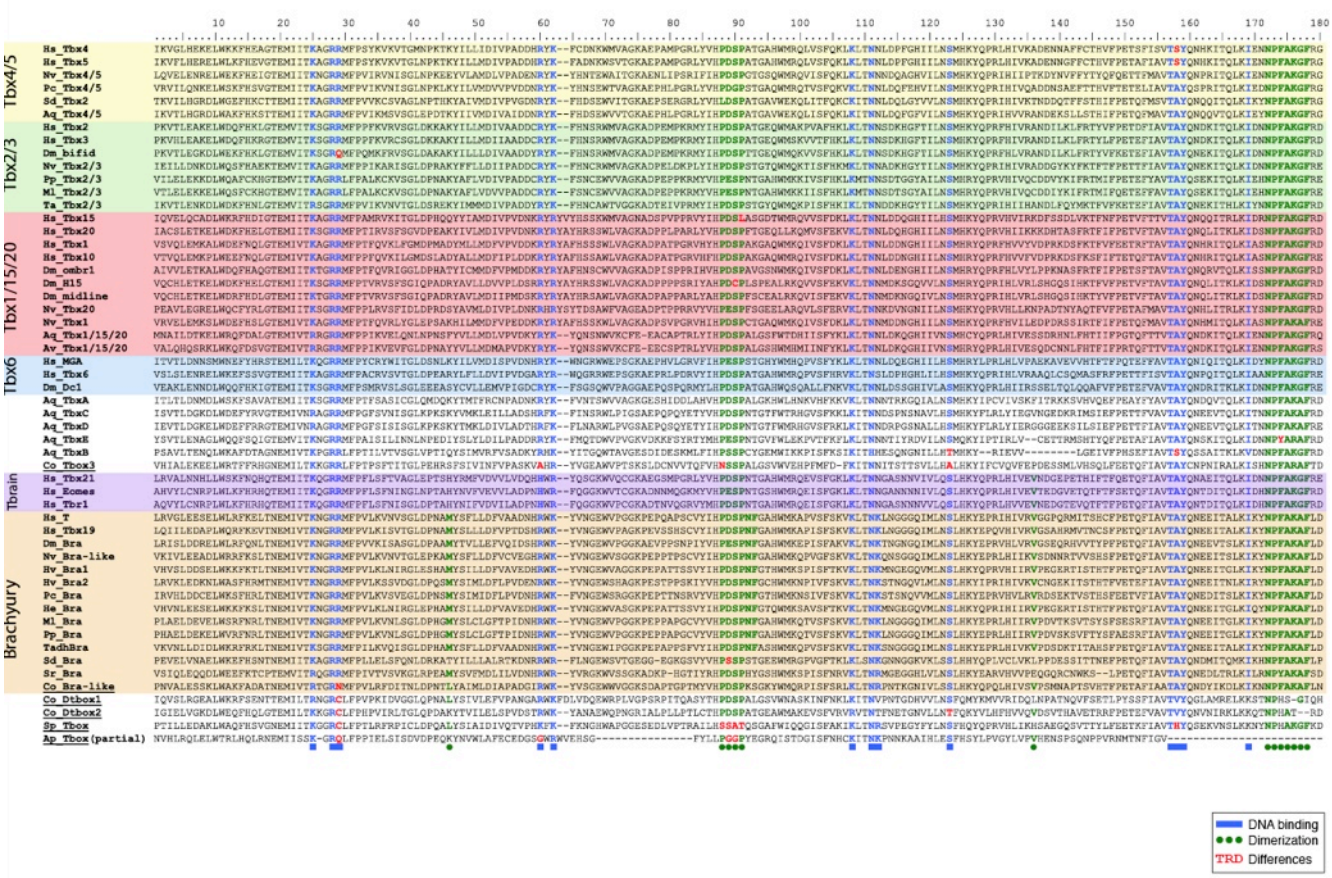


Figure S4. Alignment of the T-box domain with the different families shown in distinct colors. Taxa include Aq (*Amphimedon queenslandica*), Av (*Axinella verrucosa*), Co (*Capsaspora owczarzaki*), Dm (*Drosophila melanogaster*), He (*Hydractinia echinata*), Hs (*Homo sapiens*), Hv (*Hydra vulgaris*), Ml (*Mnemiopsis leydi*), Nv (*Nematostella vectensis*), Pc (*Podocoryne carnea*), Pp (*Pleurobrachia pileus*), Sd (*Suberites domuncula*), Sp (*Spizellomyces punctatus*), Sr (*Sycon raphanus*), and Ta (*Trichoplax adhaerens*). Key DNA binding amino acids are highlighted in blue and dimerization aminoacids in green. Non-conservative amino acidic changes are depicted in red.

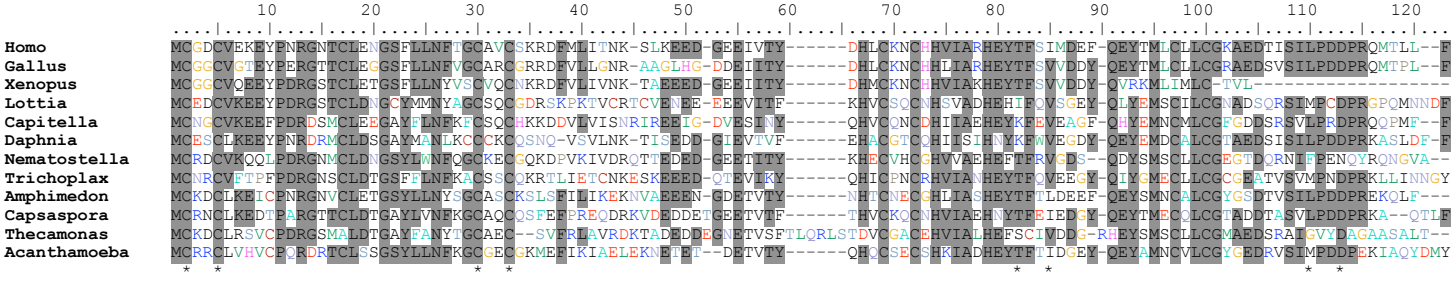


Figure S5. Alignment of Churchill domain. Asterisks indicate the CXXC motifs of the two C4-type zinc-fingers of the Churchill proteins (Sheng et al. 2003).

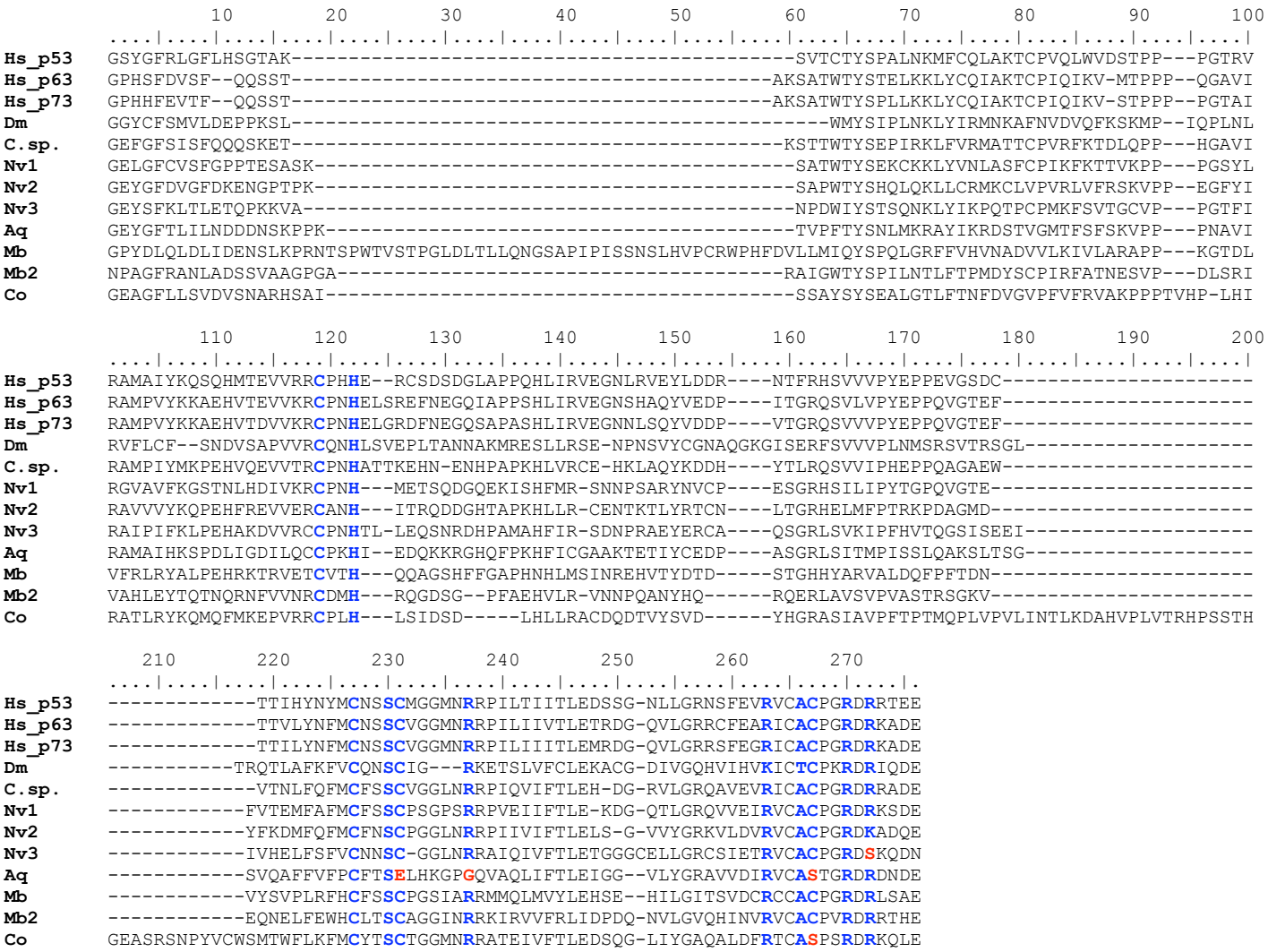


Figure S6. Alignment of the p53 DNA binding domain. Taxa include Hs (*Homo sapiens*), Dm (*Drosophila melanogaster*), C.sp (*Capitella sp.*), Nv (*Nematostella vectensis*), Aq (*Amphimedon queenslandica*), Mb (*Monosiga brevicollis*) and Co (*Capsaspora owczarzaki*). DNA binding aminoacidic motifs (Nedelcu and Tan 2007) are depicted in blue. Non-conservative aminoacidic changes are shown in red.

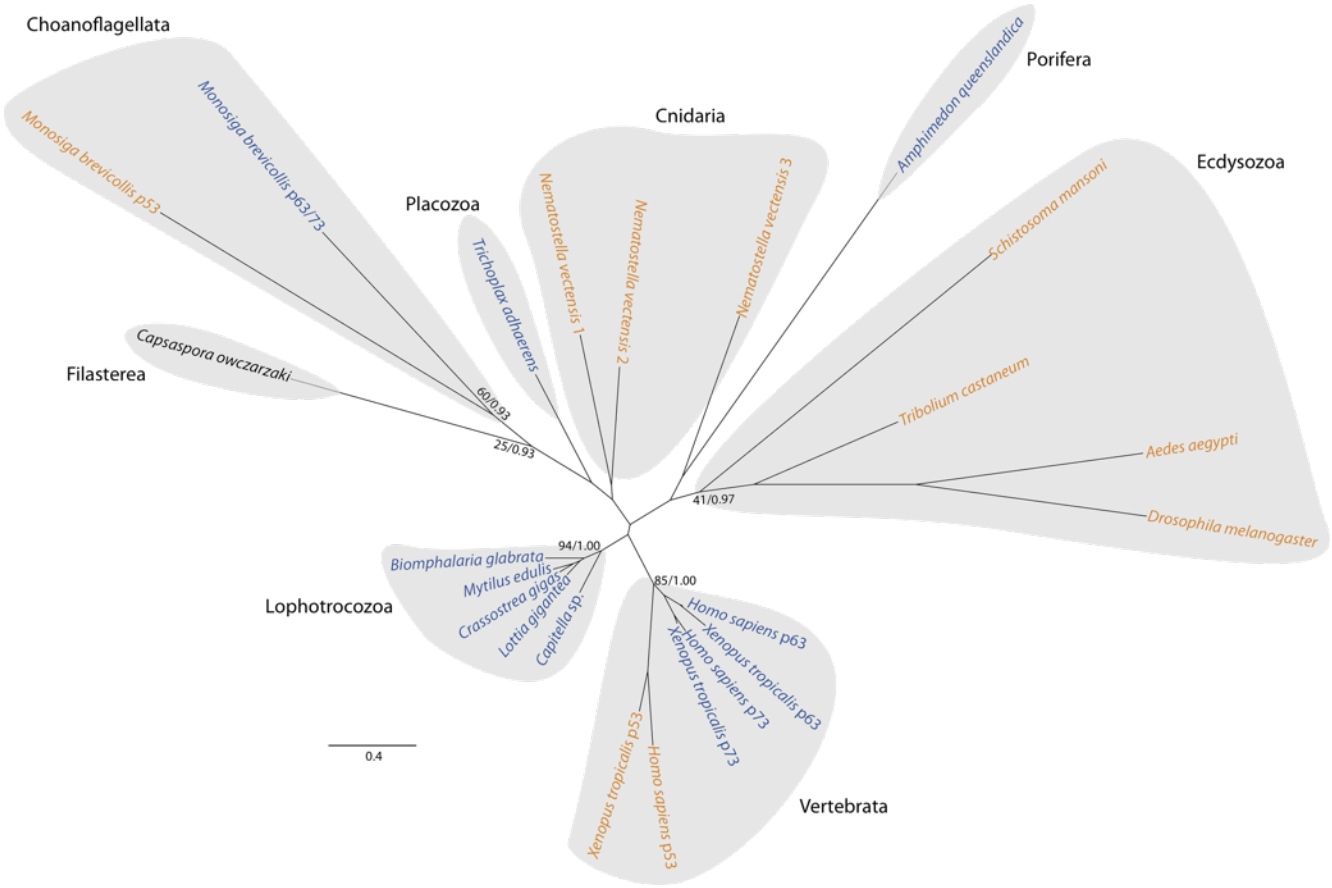


Figure S7. Maximum likelihood tree of p53 family. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 100-bootstrap replicates (bootstrap value, BV) and by Bayesian Posterior Probabilities (BPP). Both values values are shown on key branches. p63/73 subfamily members are shown in blue and p53 subfamily members in orange.

SUPPLEMENTARY MATERIAL 2

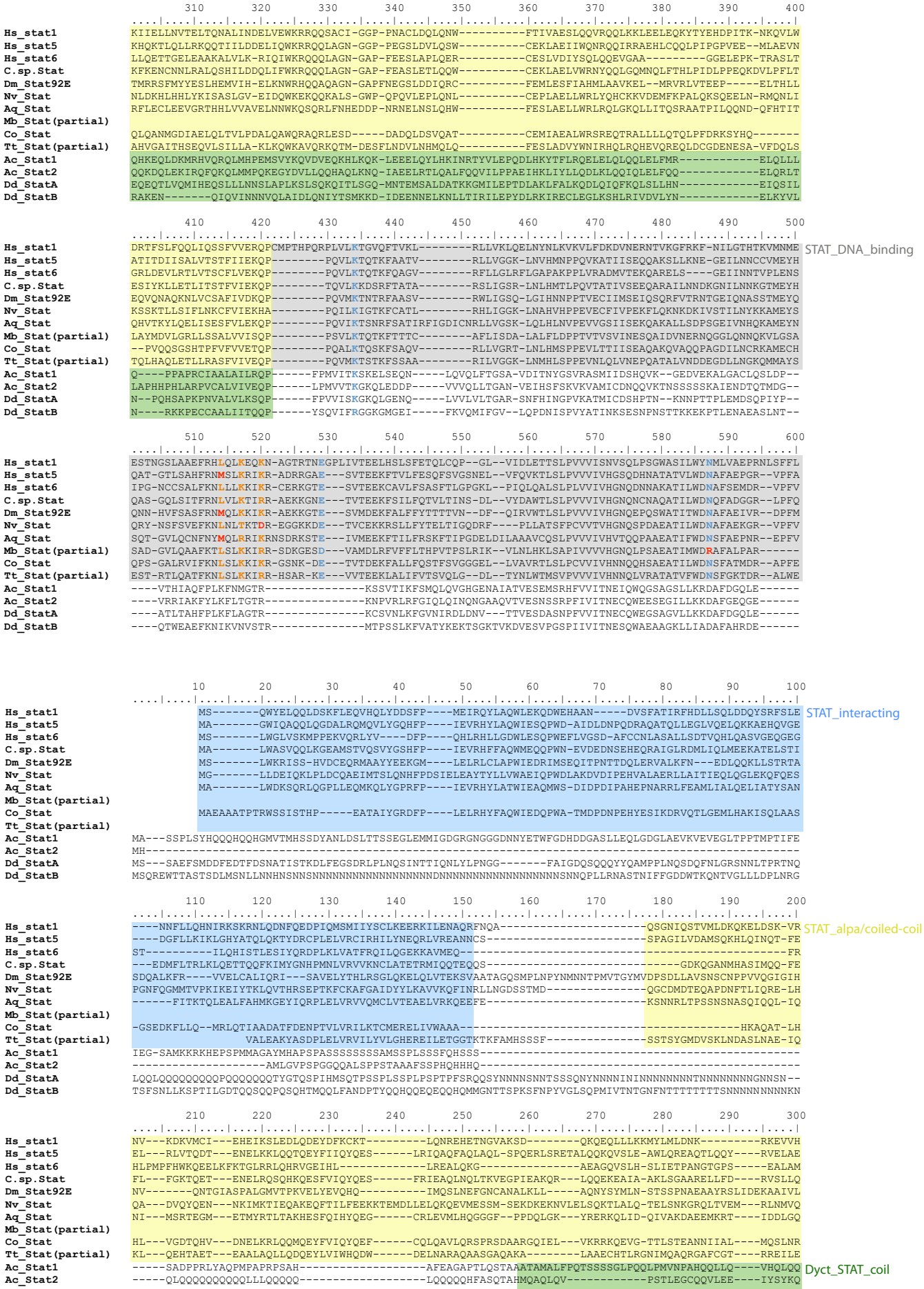


Figure S8. Alignment of Stat proteins. Taxa include Hs (*Homo sapiens*), Dm (*Drosophila melanogaster*), C.sp (*Capitella sp.*), Nv (*Nematostella vectensis*), Aq (*Amphimedon queenslandica*), Mb (*Monosiga brevicollis*), Co (*Capsaspora owczarzaki*), Tt (*Thecamonas trahens*), Ac (*Acanthamoeba castellanii*) and Dd (*Dictyostelium discoideum*). Different Pfam domains are shown with different colours. Key DNA binding aminoacids are highlighted in blue, nuclear importing aminoacid signals in orange, and phosphorylated regulative Tyr residue in purple. Non-conservative aminoacidic changes are depicted in red.

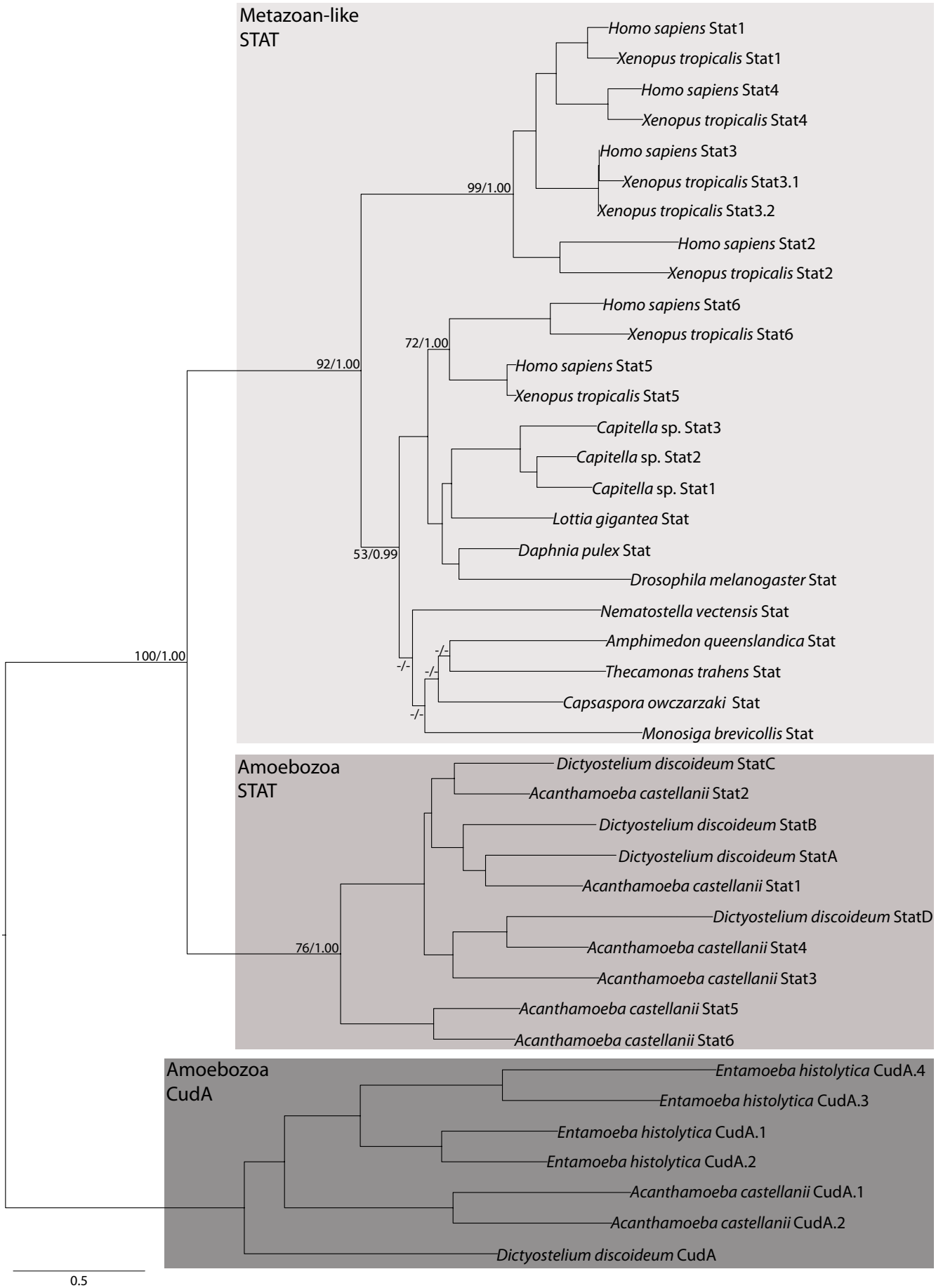


Figure S9. Maximum likelihood tree of Stat proteins. The tree is rooted using amoebozoan CudA proteins, which are distantly related to Stat proteins. Statistical support was obtained by RAxML with 100-bootstrap replicates (BV) and Bayesian Posterior Probabilities (BPP). Dashes (-) indicate no statistical support. Both values are shown on key branches.

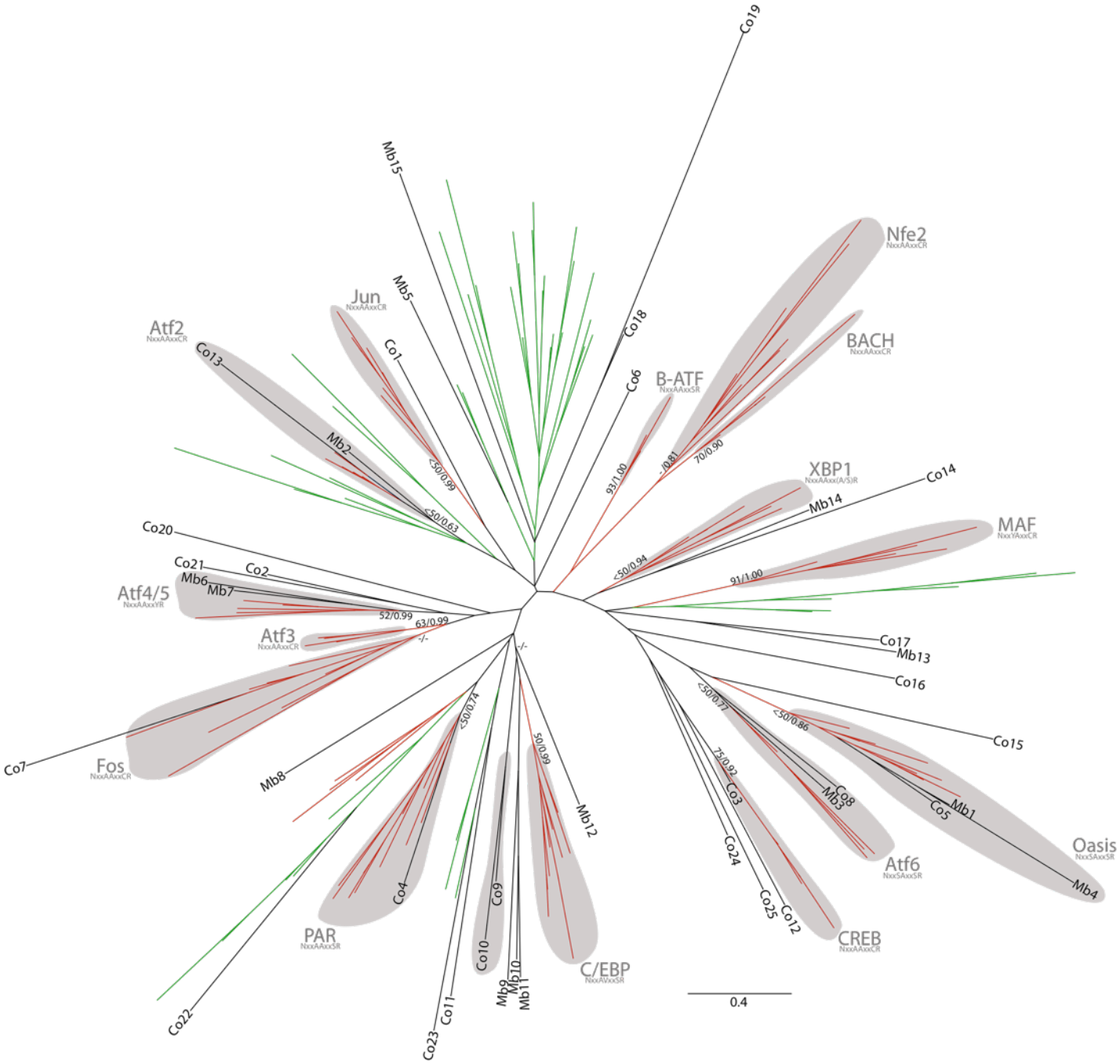


Figure S10. Maximum likelihood tree of bZIP genes including all *Capsaspora* homologs and some fungi. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 100-bootstrap replicates (BV) and Bayesian Posterior Probabilities (BPP). Both values are shown on key branches. Co (*Capsaspora owczarzaki*), Mb (*Monosiga brevicollis*). Metazoan branches depicted in red and fungal branches in green. For each family, the signature sequence for DNA recognition is indicated and only proteins with this conserved motif are included in the family (Fujii et al. 2000).

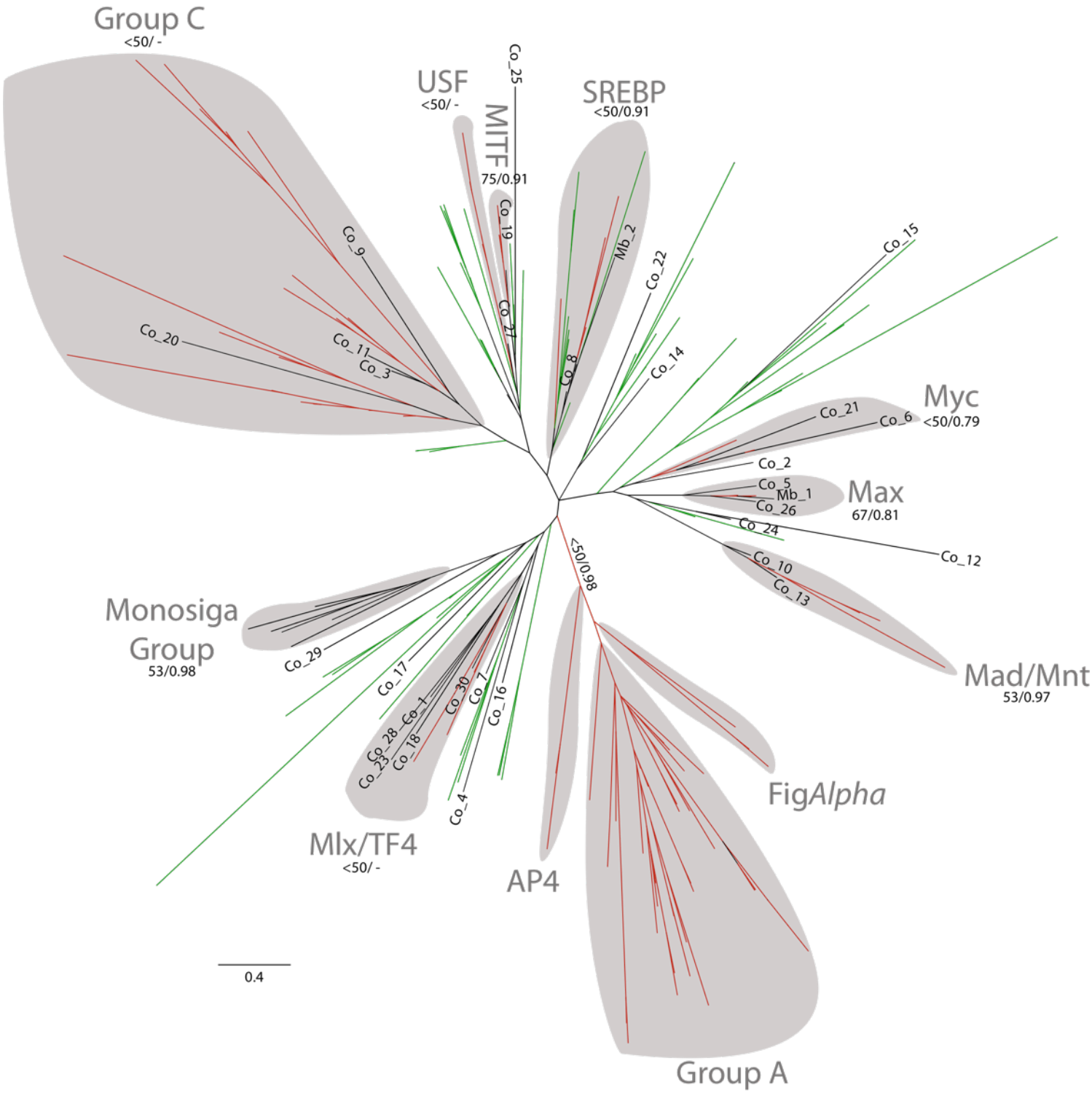


Figure S11. Maximum likelihood tree of bHLH domains including all *Capsaspora* homologs and some fungi. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 100-bootstrap replicates (BV) and Bayesian Posterior Probabilities (BPP). Both values are shown on key branches. Mb (*Monosiga brevicollis*), Co (*Capsaspora owczarzaki*). Metazoan branches depicted in red and fungal branches in green.

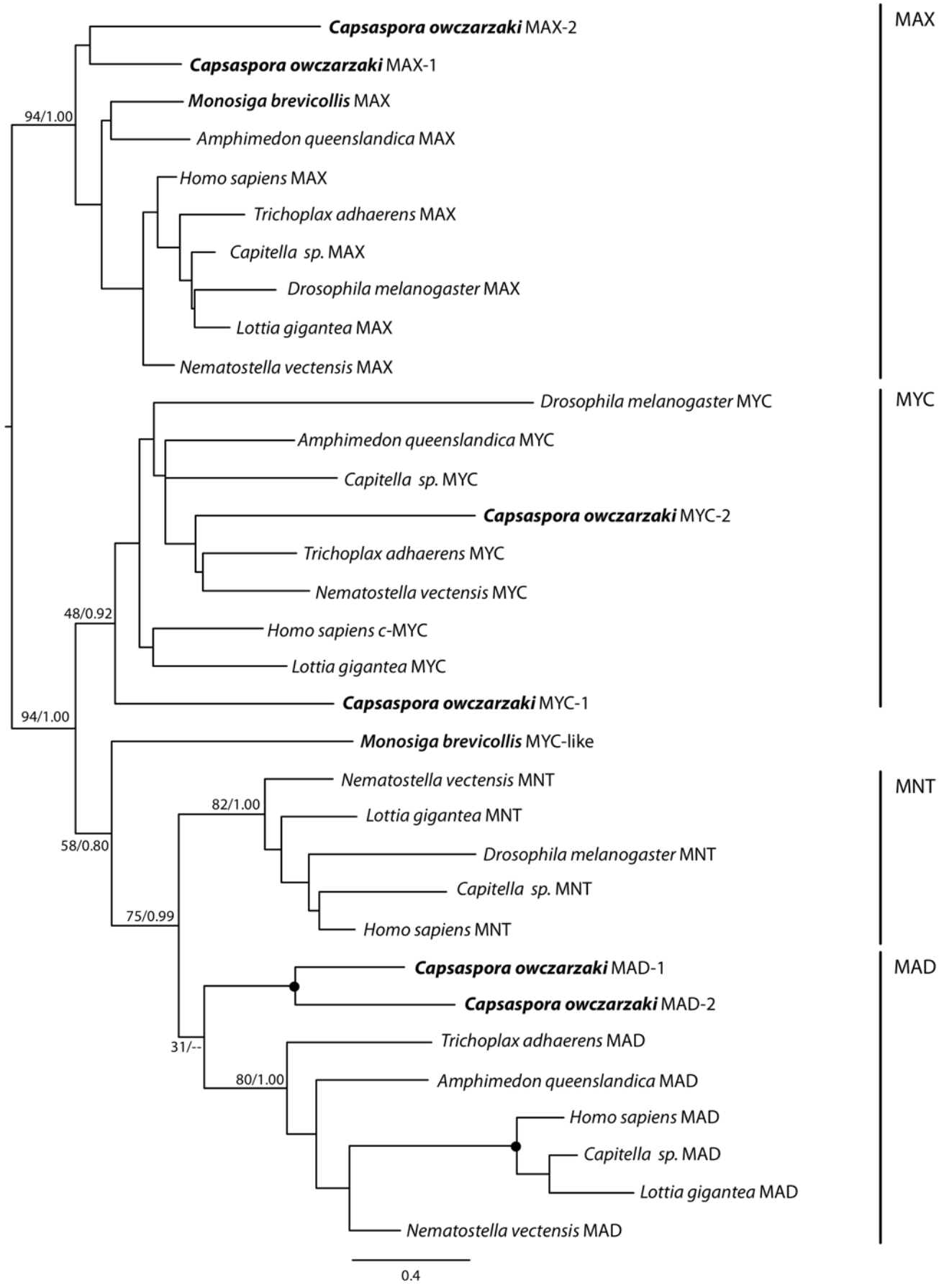


Figure S12. Maximum likelihood tree from the bHLH domain plus the LZ of Myc, Max, Mad and Mnt proteins. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 100-bootstrap replicates (BV) and Bayesian Posterior Probabilities (BPP). Both values are shown on key branches. A black dot indicates BV > 90% and BPP > 0.95.

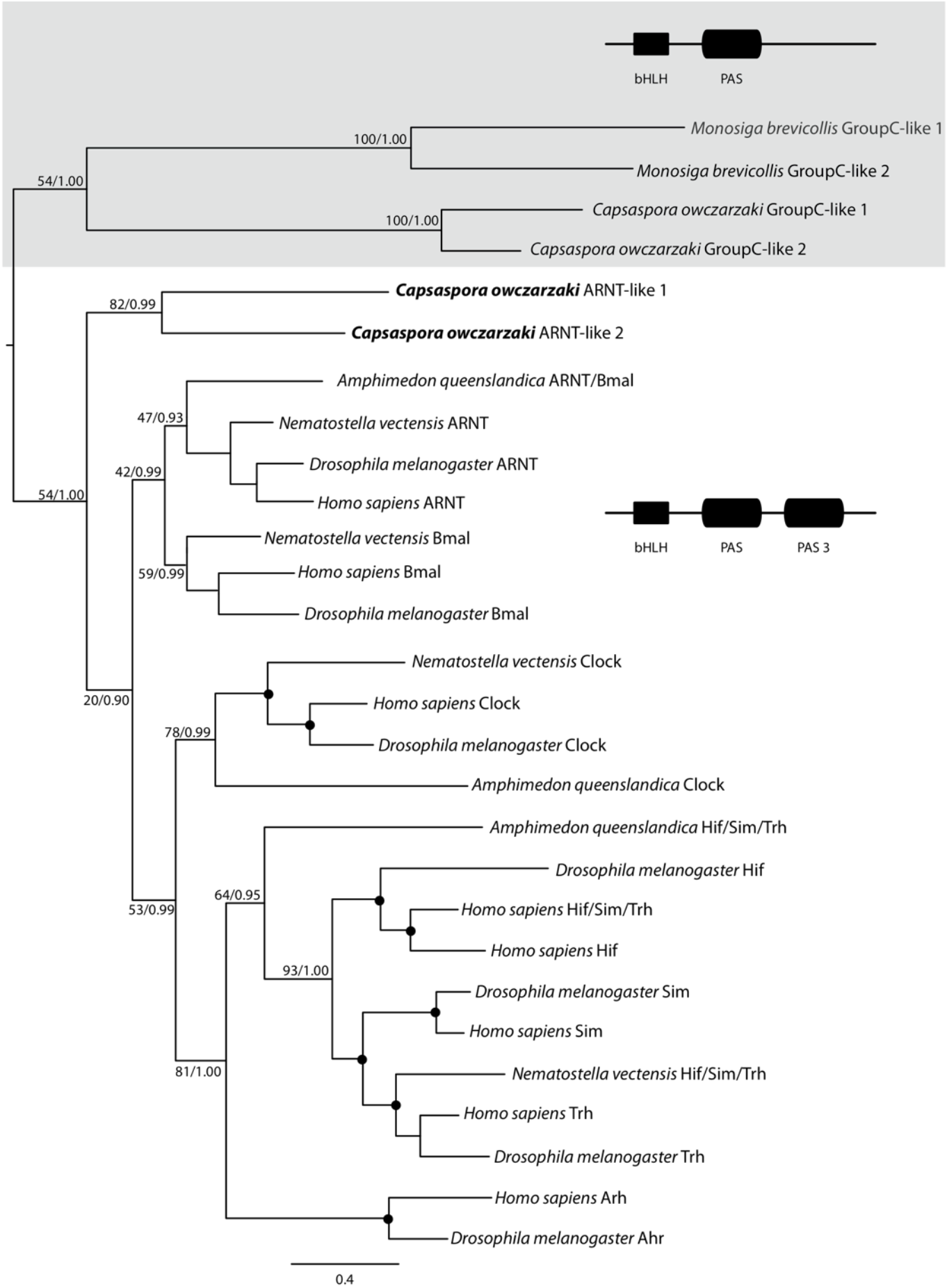


Figure S13. Maximum likelihood tree from the bHLH domain plus the PAS domain of Group C proteins. The tree is rooted using the Group-C-like proteins as the outgroup. Statistical support was obtained by RAxML with 100-bootstrap replicates (BV) and Bayesian Posterior Probabilities (BPP). Both values are shown on key branches. A black dot indicates BV > 90% and BPP > 0.95. Domain architectures from Pfam of the Group-C-like and Group C proteins are shown.

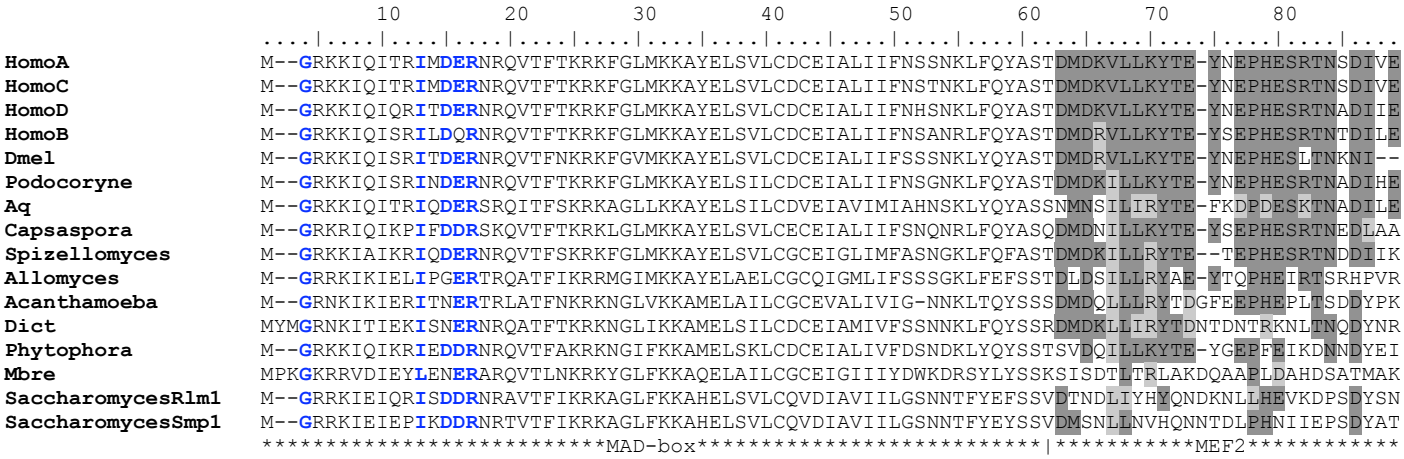


Figure S14. Alignment of MADS-box and mef2 domain. Conserved amino acids (dark grey), conservative changes (light grey) and key DNA binding aminoacids (blue) are depicted.

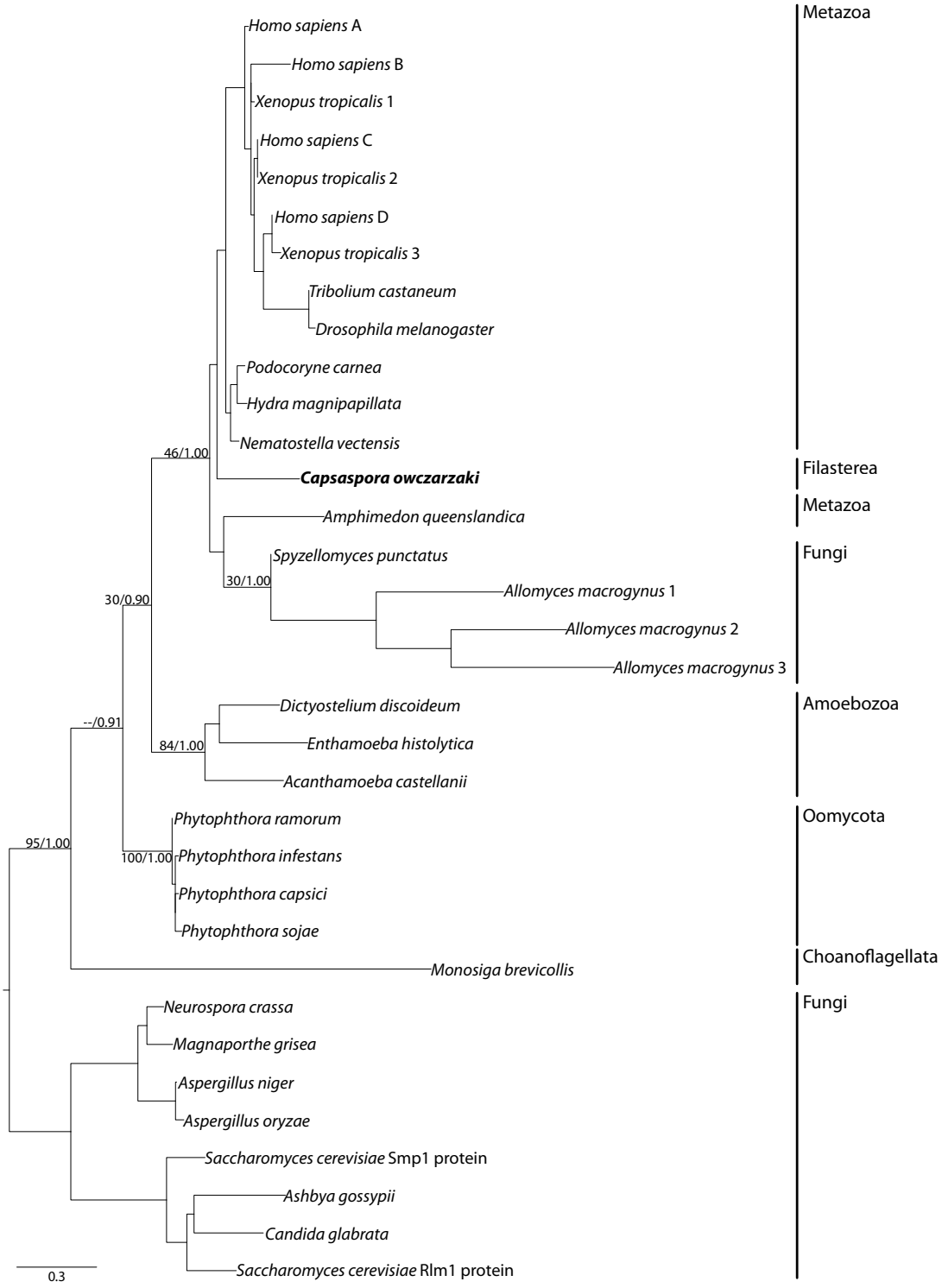


Figure S15. Maximum likelihood tree of Mef2-like genes. The tree is rooted using fungi. Statistical support was obtained by RAxML with 100-bootstrap replicates (BV) and Bayesian Posterior Probabilities (BPP). Both values are shown on key branches.

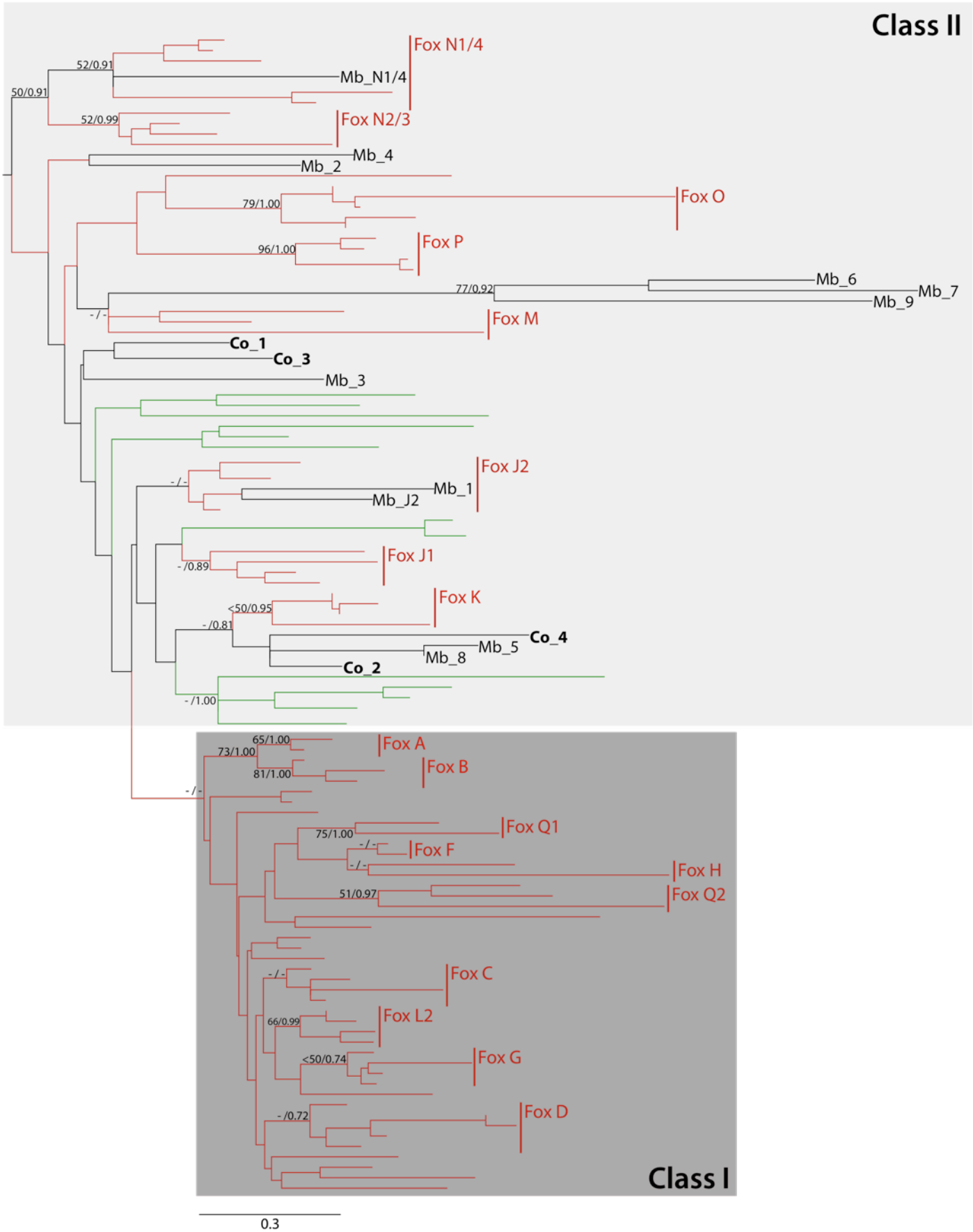


Figure S16. Maximum likehood tree of all *Capsaspora* and *Monosiga* homologs. The trees is rooted using the midpoint-rooted tree option. Statistical support was obtaind by RaxML with 100-bootstrap replicates (BV) and Bayes Posterior Probabilities (BPP). Both values are shown on key branches. Co (*Capsaspora owczarzaki*), Mb (*Monosiga brevicollis*). Metazoan branches depicted in red and fungal branches in green.

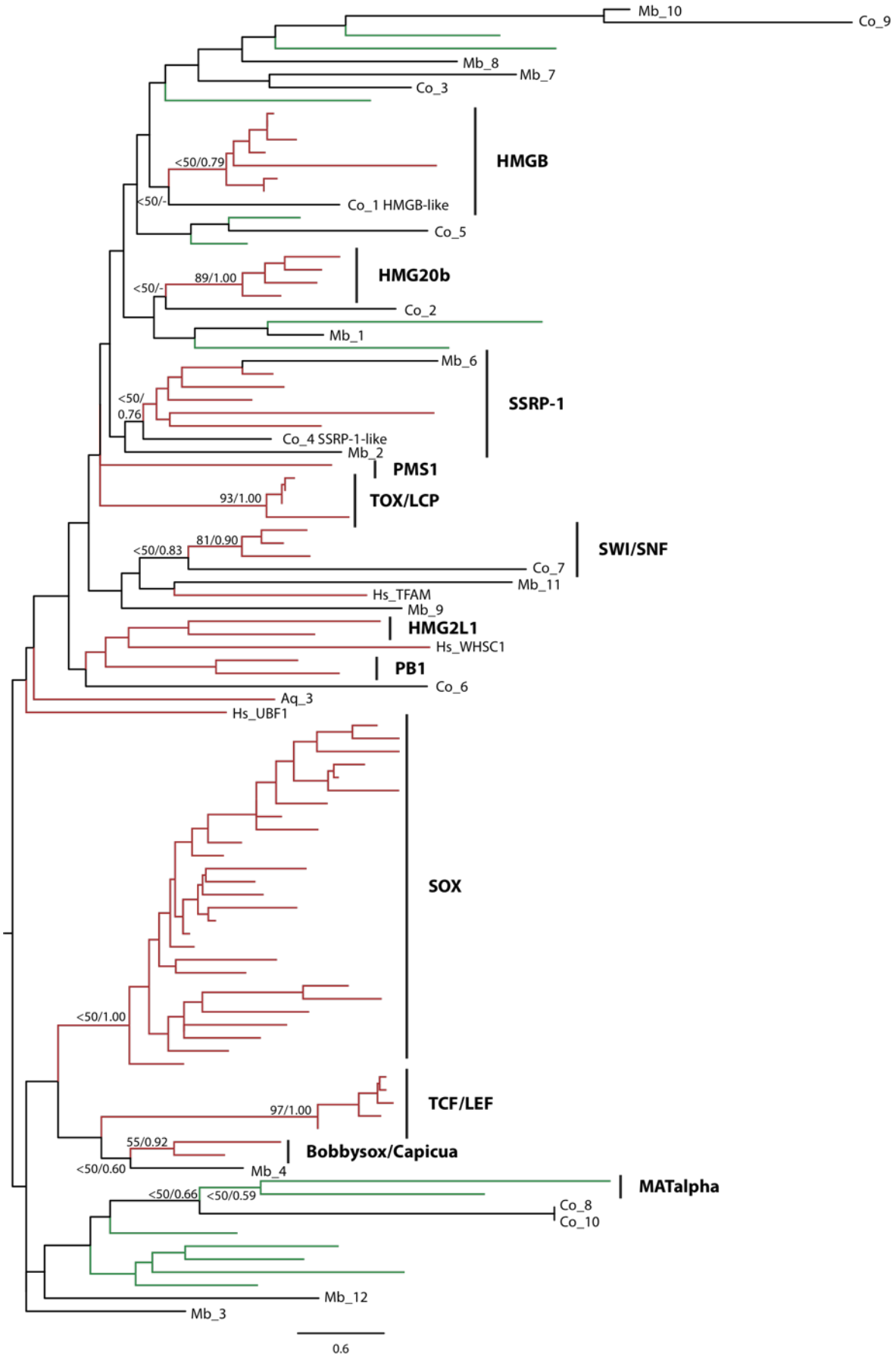


Figure S17. Maximum likelihood tree of HMGbox domains including all *Capsaspora* and *Monosiga* homologs. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 100-bootstrap replicates (BV) and Bayesian Posterior Probabilities (BPP). Both values are shown on key branches. Mb (*Monosiga brevicollis*), Co (*Capsaspora owczarzaki*), Hs (*Homo sapiens*). Metazoan branches depicted in red and fungal branches in green.

SUPPLEMENTARY MATERIAL 3

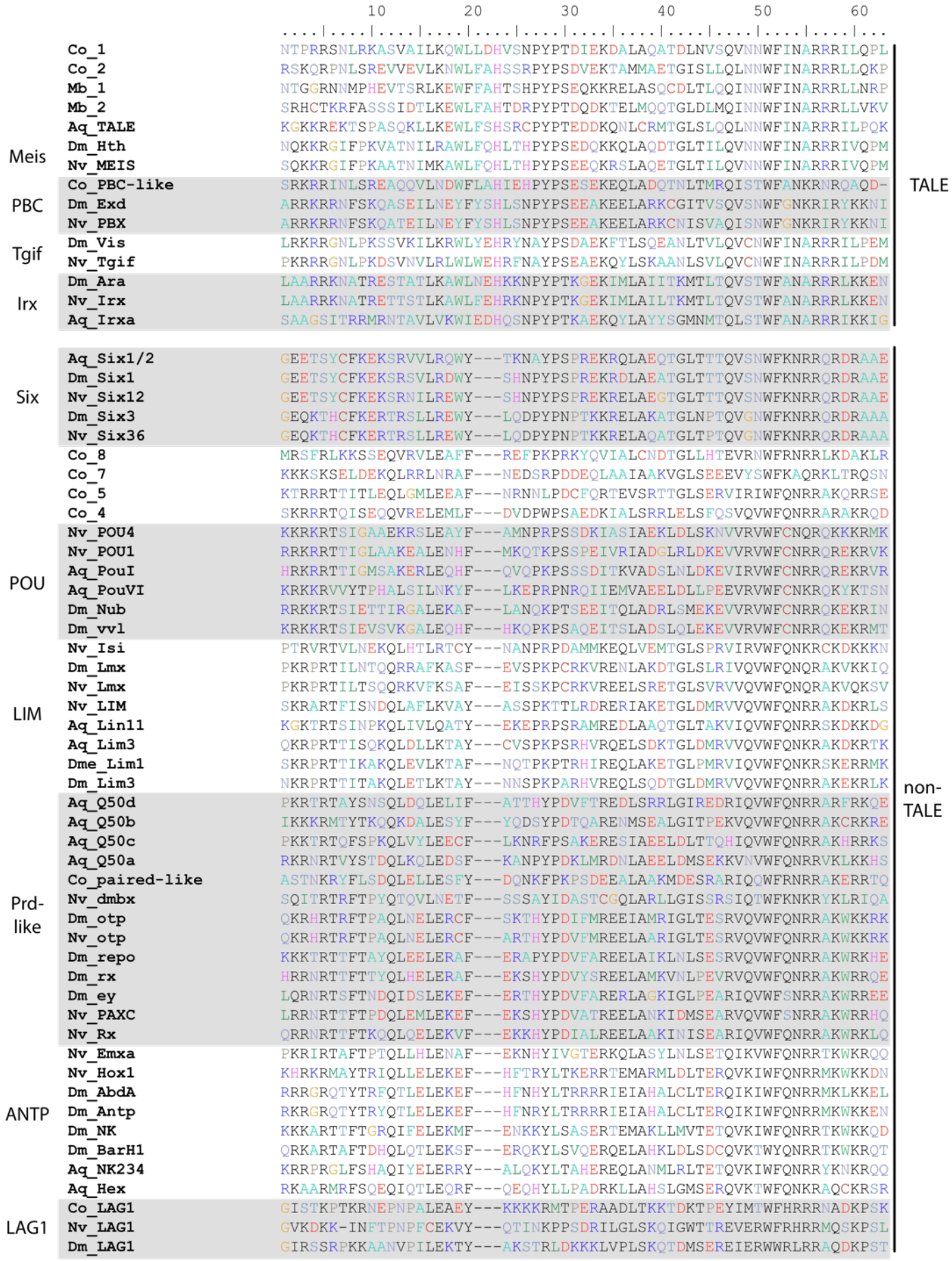


Figure S18. An illustrative homeobox alignment, showing the *Capsaspora* and *Monosiga* homeobox domains in contrast to other metazoan families. Taxa used includes Nv (*Nematostella vectensis*), Aq (*Amphimedon queenslandica*), Co (*Capsaspora owczarzaki*), Dm (*Drosophila melanogaster*) and Mb (*Monosiga brevicollis*).

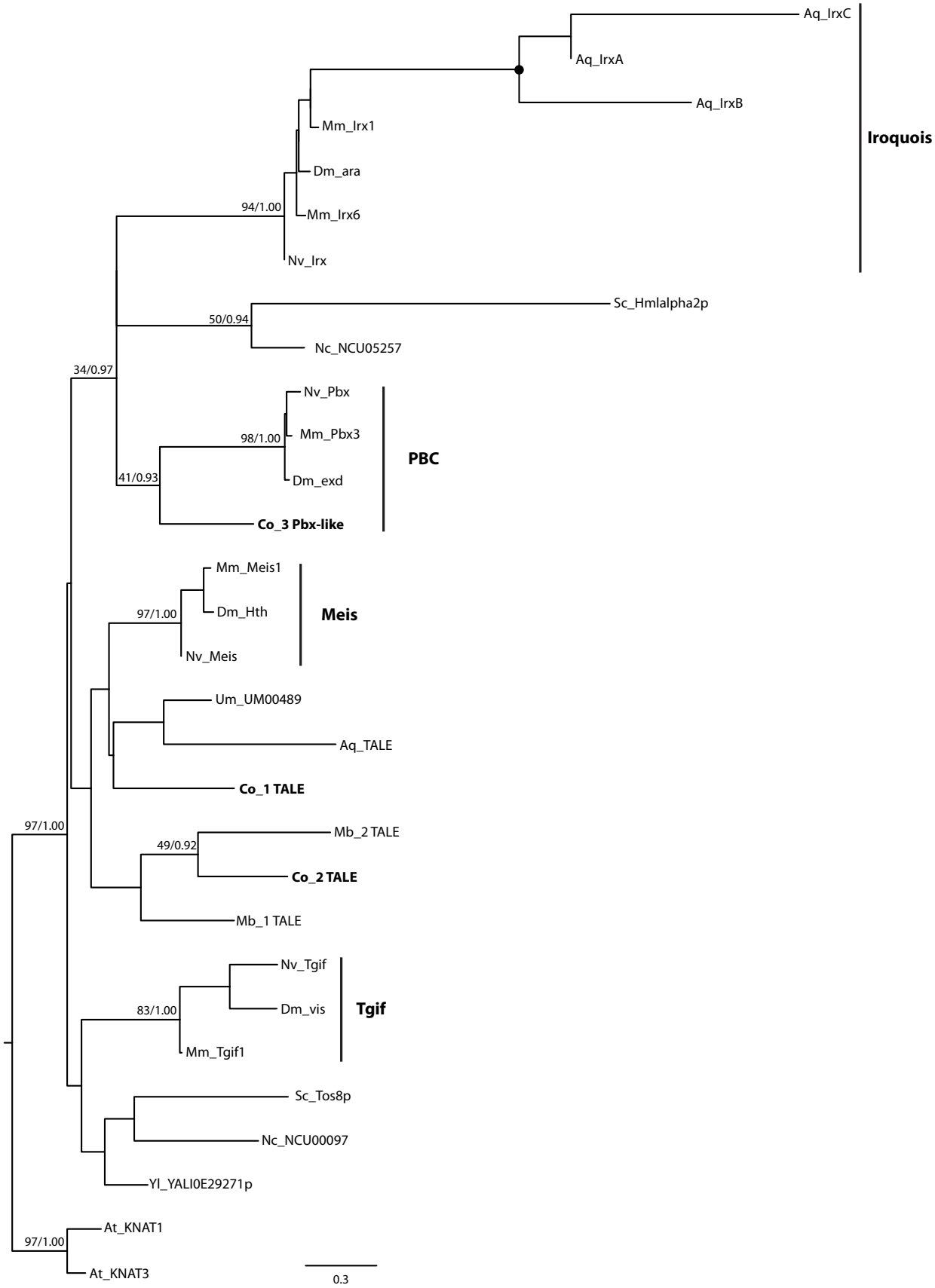


Figure S19. Maximum likelihood tree from the homeobox domain of TALE HDs. The tree is rooted using *Arabidopsis thaliana* sequences as outgroup. Statistical support was obtained by RAxML with 100-bootstrap replicates (bootstrap value, BV) and Bayesian Posterior Probabilities (BPP). Both values are shown on key branches. A black dot indicates BV > 90% and BPP > 0.95. Taxa used adapted from Larroux et al. Nv (*Nematostella vectensis*), Aq (*Amphimedon queenslandica*), Mb (*Monosiga brevicollis*), Co (*Capsaspora owczarzaki*), Dm (*Drosophila melanogaster*), Mm (*Mus musculus*), Sc (*Saccharomyces cerevisiae*), Um (*Ustilago maydis*), At (*Arabidopsis thaliana*), Nc (*Neurospora crassa*) and Yl (*Yarrowia lypolitica*).

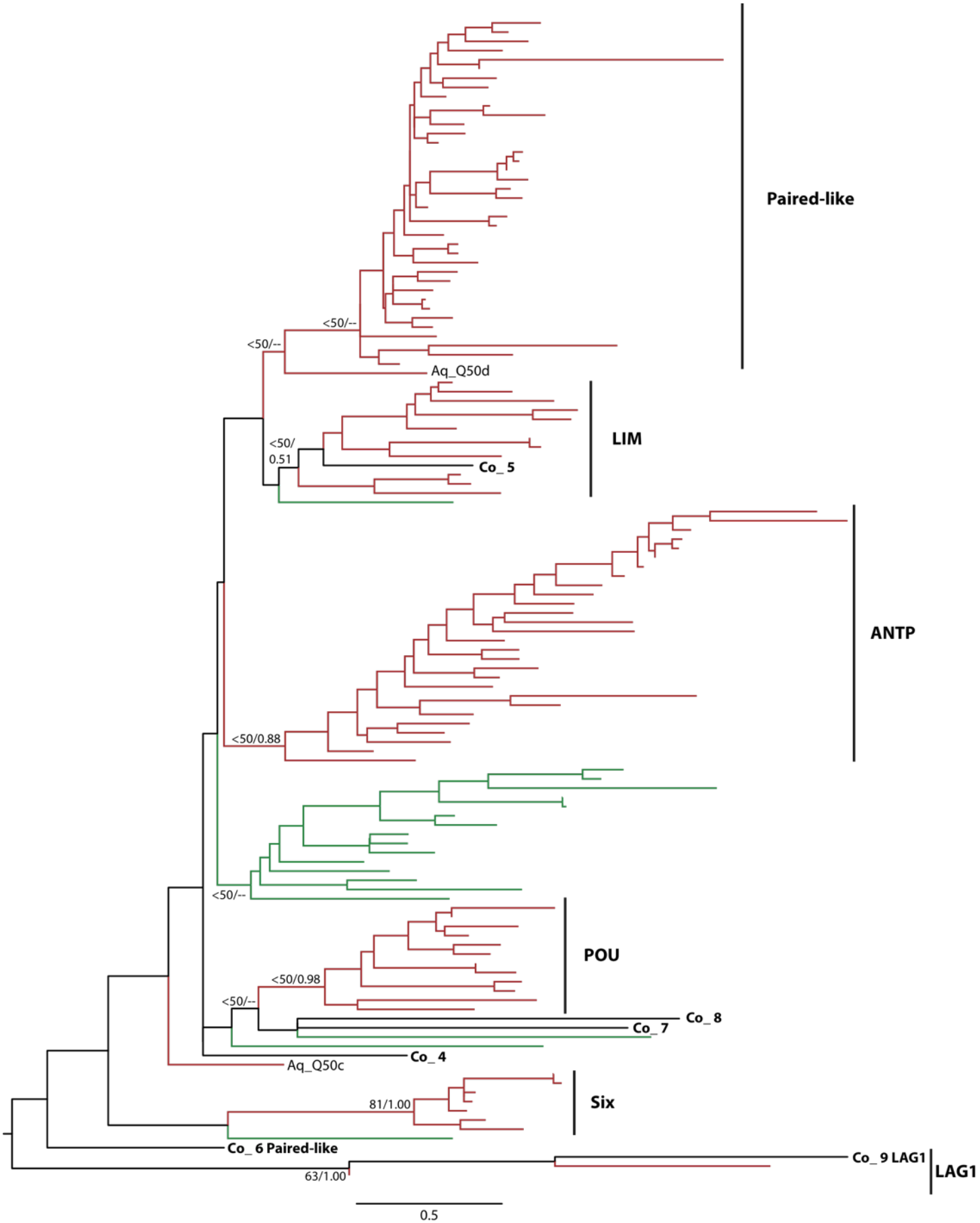


Figure S20. Maximum likelihood tree of non-TALE homeobox domains including all *Capsaspora* non-TALE homologs. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 100-bootstrap replicates (BV) and Bayesian Posterior Probabilities (BPP). Both values are shown on key branches. Mb (*Monosiga brevicollis*), Co (*Capsaspora owczarzaki*). Metazoan branches depicted in red and fungal branches in green.

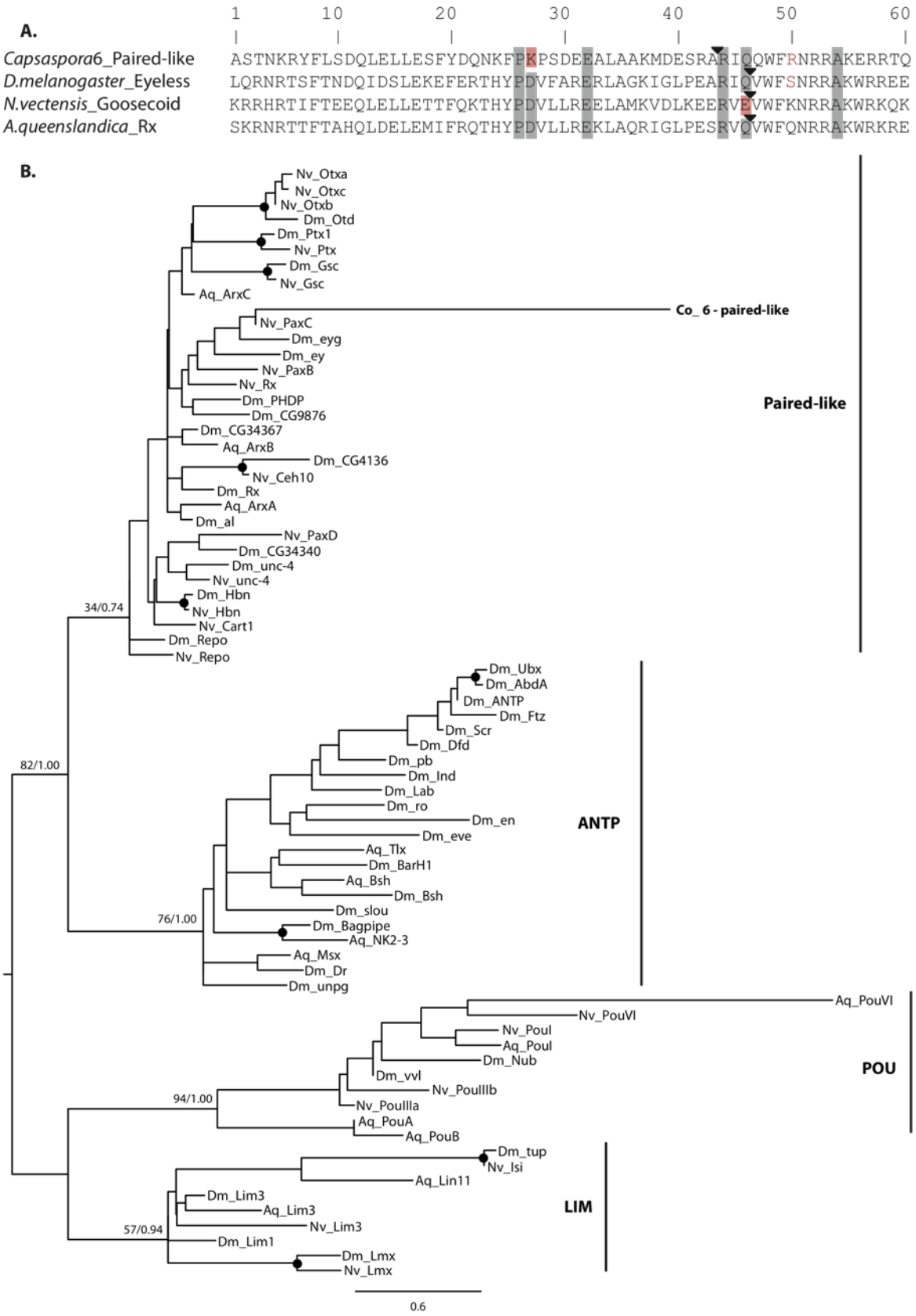


Figure S21. A) An illustrative homeobox alignment, showing the Capsaspora6 with some other archetypical Prd-like genes from different species. In grey the aminoacids that define Prd-like class (Galliot et al. 1999). Black triangle shows the intron position. **B)** Maximum likelihood tree from the homeobox domain of non-TALE HDs from ANTP, Prd-Like, POU and LIM-HD classes. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 100-bootstrap replicates (BV) and Bayesian Posterior Probabilities (BPP). Both values are shown on key branches. A black dot indicates BV > 90% and BPP > 0.95. Taxa used Nv (*Nematostella vectensis*), Aq (*Amphimedon queenslandica*), Co (*Capsaspora owczarzaki*) and Dm (*Drosophila melanogaster*).

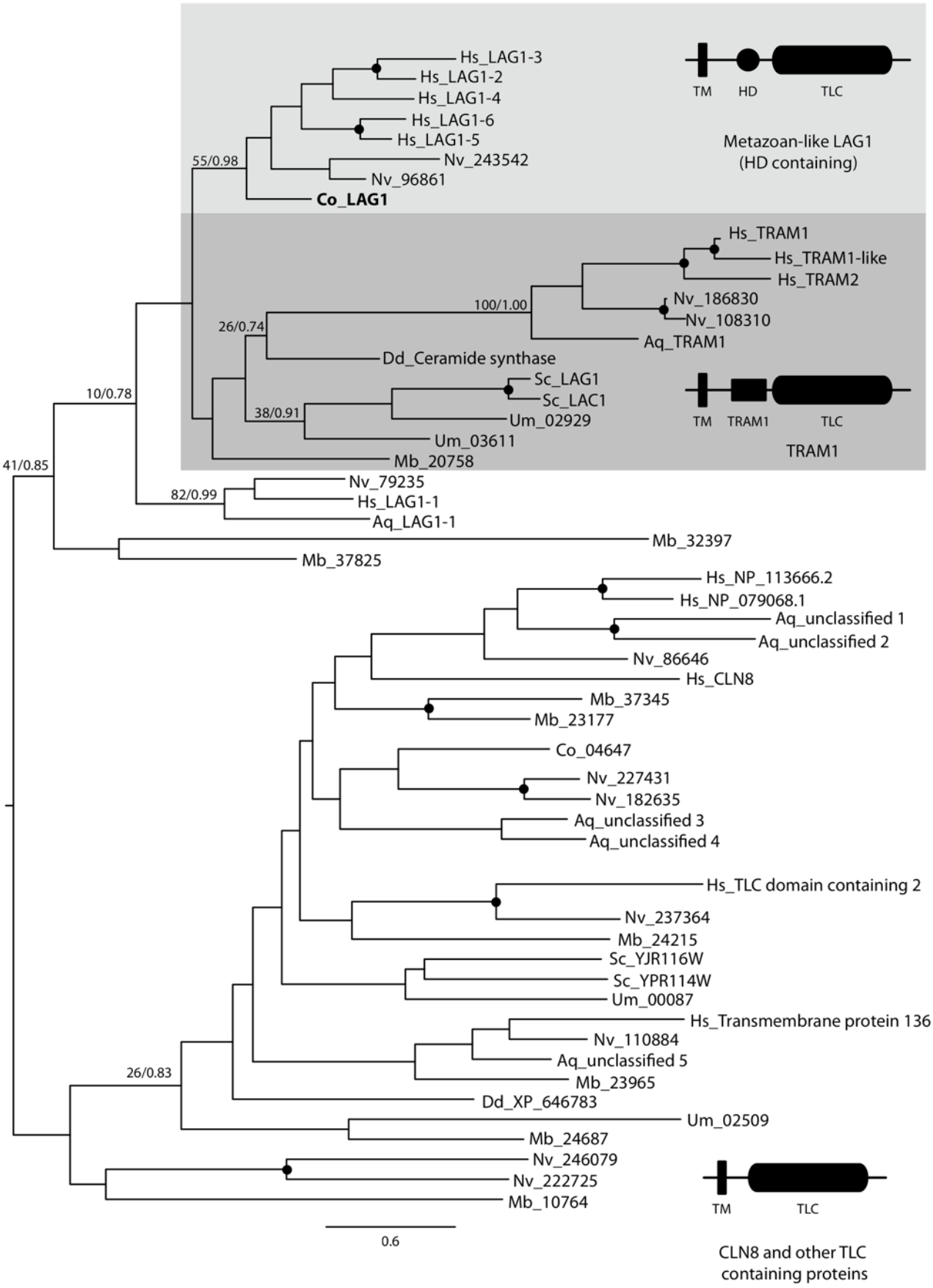


Figure S22. Maximum likelihood tree from the TLC domain. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 100-bootstrap replicates (BV) and Bayesian Posterior Probabilities (BPP). Both values are shown on key branches. A black dot indicates BV > 90% and BPP > 0.95. Taxa used Nv (*Nematostella vectensis*), Aq (*Amphimedon queenslandica*), Mb (*Monosiga brevicollis*), Co (*Capsaspora owczarzaki*), Hs (*Homo sapiens*), Sc (*Saccharomyces cerevisiae*), Um (*Ustilago maydis*) and Dd (*Dictyostelium discoideum*). PFAM domain architecture displayed in the different classes.

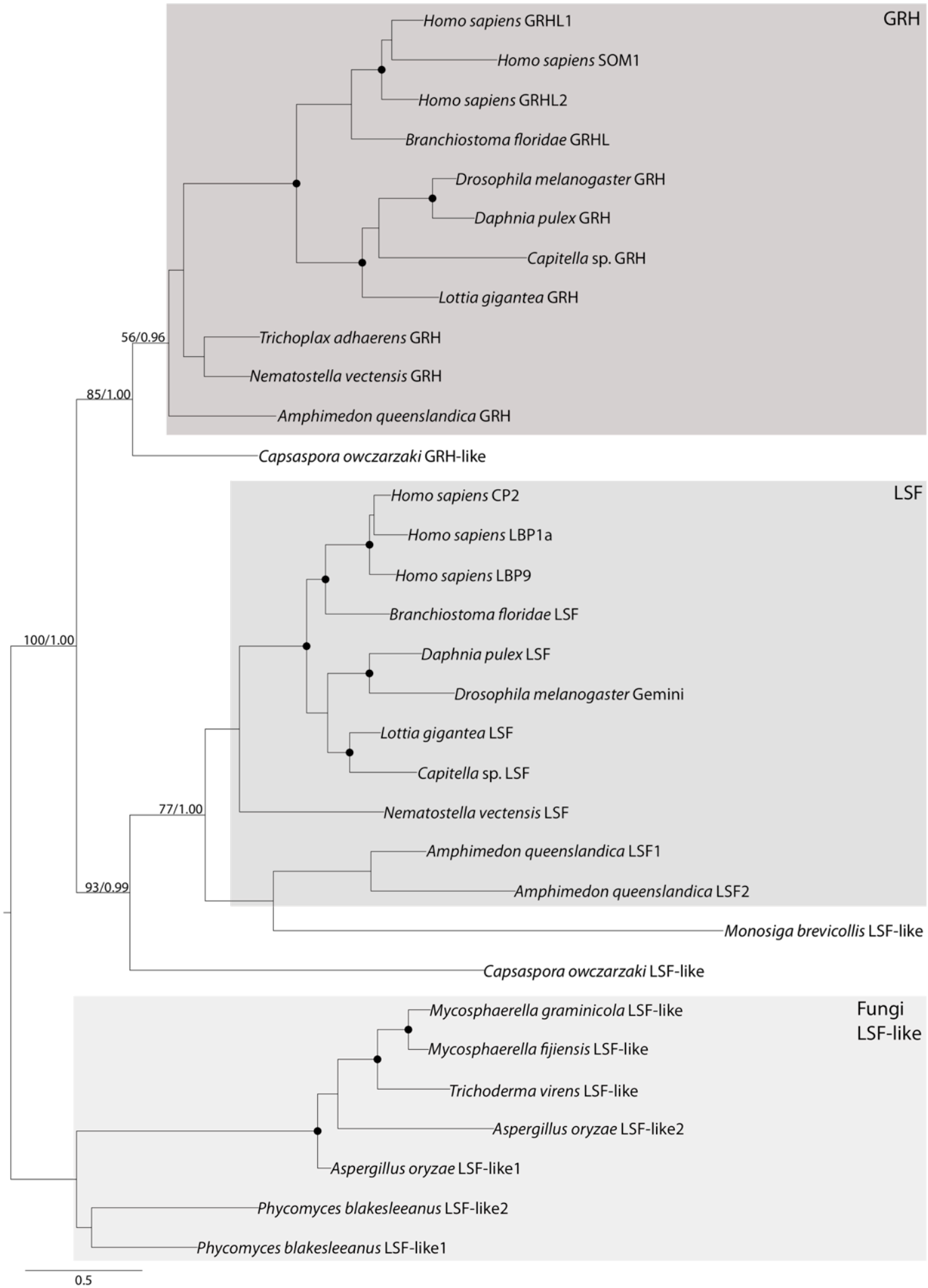


Figure S23. Maximum likelihood tree of CP2 domain containing genes. The tree is rooted using fungi. Statistical support was obtained by RAxML with 100-bootstrap replicates (BV) and Bayesian Posterior Probabilities (BPP). Both values are shown on key branches. A black dot indicates BV > 90% and BPP > 0.95.

Results R4

Early evolution of the T-box transcription factor family.

RESUM ARTICLE R4: Evolució primerenca de la família de factors de transcripció T-box

Els factors de transcripció del desenvolupament són elements crucials de la multicel·lularitat animal, amb els membres de la família de factors de transcripció T-box essent un dels més importants. Fins fa poc, els factors de transcripció T-box es consideraven exclusius d'animals. En aquest treball, mostrem la presència de gens T-box en diversos llinatges no-animals, incloent ictiosporis, filasteris i fongs. Els nostre resultats confirmen que Brachyury és el membre més antic de la família T-box, i establim que la família T-box es diversificà a l'origen dels animals. A més, demostrem conservació funcional de l'homòleg de Brachyury de *Capsaspora owczarzaki* en *Xenopus laevis*. Comparant els fenotips molecular de *C.owczarzaki* Brachyury amb els produïts per homòlegs de metazous basals, definim una diferència clara entre gens Brachyury de fongs, holozous unicel·lulars i animals; suggerint que l'especificitat de la classe Brachyury emergí a l'origen dels metazous. La determinació experimental de les preferències d'unió a DNA del gen Brachyury del protist resultà en un motiu d'unió molt similar a aquell dels gens Brachyury i altres classes T-box d'animals. Aquest resultat suggereix que l'especificitat funcional entre diferents classes de T-box és deguda no tant a canvis en l'especificitat d'unió com a interaccions amb cofactors alternatius; interaccions que s'establiren a l'origen dels metazous, coincidint amb l'expansió de la família T-box.

Early evolution of the T-box transcription factor family

Arnau Sebé-Pedrós[a,f,1], Ana Ariza-Cosano[b,1], Matthew T. Weirauch[c], Sven Leininger[d], Ally Yang[e], Guifré Torruella[a], Marcin Adamski[d], Maja Adamska[d], Timothy R. Hughes[e], José Luis Gómez-Skarmeta[b, 2], Iñaki Ruiz-Trillo[a,f,g 2]

[a]Institut de Biologia Evolutiva (UPF-CSIC), Passeig Maritim de la Barceloneta 37-49, 08003 Barcelona, Spain [b]Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Ctra. Utrera Km 1, Seville 41013, Spain [c]Center for Autoimmune Genomics and Etiology (CAGE) and Divisions of Rheumatology and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA [d]Sars International Centre for Marine Molecular Biology, Thormøhlensgt. 55, 5008 Bergen, Norway [e]Donnelly Centre and Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 3E1, CANADA [f]Departament de Genètica, Universitat de Barcelona, Av. Diagonal, 645, 08028 Barcelona, Spain [g]Institució Catalana per a la Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[1]These authors contributed equally to this work. [2]To whom correspondence should be adressed. E-mail: jlgomska@upo.es / inaki.ruiz@ibe.upf-csic.es

Developmental transcription factors are key players in animal multicellularity, with members of the T-box family of transcription factors being among the most important. Until recently, T-box transcription factors were thought to be exclusively present in metazoans. Here, we report the presence of T-box genes in several non-metazoan lineages, including ichthyosporeans, filastereans and fungi. Our data confirm that Brachyury is the most ancient member of the T-box family, and establish that the T-box family diversified around the origin of Metazoa. Moreover, we demonstrate functional conservation of a homolog of Brachyury of the protist *Capsaspora owczarzaki* in *Xenopus laevis*. By comparing the molecular phenotype of *C. owczarzaki* Brachyury with that of early-branching metazoans, we define a clear difference between fungal, unicellular holozoan, and metazoan Brachyury, suggesting that the specificity of Brachyury emerged at the origin of Metazoa. Experimental determination of the binding preferences of the protist Brachyury results in a similar motif to that of metazoan Brachyury and other T-box classes, suggesting that functional specificity between different T-box classes is likely achieved by interaction with alternative cofactors, as opposed to differences in binding specificity.

Multicellularity – Brachyury - Capsaspora owczarzaki - Porifera - Chytrid fungi - PBM

Introduction

Transcriptional regulation is a central aspect of animal development. Thus, an understanding of the early evolution of metazoan transcription factors is vital for achieving a better understanding of the origin of animals. The T-box family of genes is among the most important developmental transcription factors present in Metazoa. This family is characterised by an evolutionary conserved DNA-binding domain of 180-200 amino acids known as the T-box domain (1–3). Brachyury is the founding member, as well as the best-characterised member of the T-box family, with well-established roles in blastopore specification, mesoderm differentiation and, in chordates, notochord formation (4–6). It has been hypothesised that the ancestral role of Brachyury was primarily that of blastopore determination and gastrulation (5, 7).

Aside from Brachyury, other T-box classes include Tbx4/5, Tbx6, Tbx2/3, Eomes and Tbx1/15/20. With only a few exceptions (8), all classes of T-box genes are widespread among bilaterian animals, with a handful being identified and studied in non-bilaterian metazoans, such as cnidarians (5, 9), ctenophores (7, 10) and sponges (11–14). T-box genes were initially thought to be specific to metazoans (13, 15), but two recent studies revealed the presence of T-box genes in non-metazoan lineages (14, 16), including the unicellular filose amoeba *Capsaspora owczarzaki*, a close relative of the animals, and the chytrid fungus *Spizellomyces punctatus*. Both analyses did not identify T-box genes in any other sequenced eukaryote, suggesting that T-box genes were secondarily lost in choanoflagellates and most fungi. Interestingly, one of the T-box genes identified in *C. owczarzaki* is an homolog of Brachyury, making it the

only Brachyury gene identified outside of metazoans to date (16). However, the degree of conservation between *C. owczarzaki* and metazoan Brachyury genes and the presence of T-box genes in other unicellular opisthokonts remained unclear.

Here, we report a taxon-wide survey of T-box genes in several eukaryotic genomes and transcriptomes, including new data from the other known filasterean taxa (*Ministeria vibrans*), as well as several ichthyosporeans, which represent the earliest-branching holozoan group (17), calcarean sponges, and other recently sequenced genomes. We identify novel T-box genes in *M. vibrans*, in all of the ichthyosporeans, and in several early-branching Fungi. Our data pinpoints with unprecedented detail the evolutionary history of T-box transcription factors. We also confirm that Brachyury is the founding member of the T-box family and define new classes of T-box genes.

To obtain a glimpse into the possible function of the earliest Brachyury genes, we perform heterologous expression experiments of the Brachyury homologs from *C. owczarzaki, Sycon ciliatum* (Calcarea, Porifera) and *Nematostella vectensis* (Anthozoa, Cnidaria) in *Xenopus laevis*, a well-established model system for studying Brachyury functions (4, 7, 18). Our data show that *C. owczarzaki* Brachyury (CoBra) can partially rescue Xenopus embryos injected with a dominant negative XBra construct. However, CoBra, contrary to *S. ciliatum* Bra (SciBra) and *N. vectensis* Bra (NvBra), activates target genes known to be regulated by other T-box gene classes, but not by Brachyury. We also use protein binding microarrays to demonstrate that the binding specificity of CoBra is indistinguishable from that of metazoan Brachyury and other T-box genes. Together, our data suggest that the subfunctionalization of Brachyury and other T-box classes is due to changes in interactions with cofactors, as opposed to changes in the DNA binding recognition motif, and that this subfunctionalization occurred at the origin of the Metazoa, concomitant with the diversification of the T-box family.

Results & Discussion

Genomic survey of T-box genes in non-metazoan species

We have searched for T-box genes in recently sequenced eukaryotic genomes and transcriptomes. This genomic survey has greatly extended the number of non-metazoan taxa known to have T-box genes. Our analyses reveal that T-box genes are present in at least four fungi taxa, belonging to three different early-branching fungal lineages (19): *Spizellomyces punctatus* and *Gonapodya prolifera* (Chytridiomycota), *Pyromices* sp. (Neocallimastigomycota) and *Mortierella verticillata* (Mucoromycotina), all of which have a single T-box gene (Fig. 1). No T-box genes were found in higher-fungi (Dikarya), in agreement with previous surveys (13, 14, 16). This confirms that T-box transcription factors were lost during fungal evolution (16). We also identified two T-box genes in the filasterean *M. vibrans*, as well as in each of the ichthyosporeans analysed: seven in *Sphaeroforma artica*, six in *Creolimax fragrantissima*, five in *Abeoforma whisleri*, two in *Amoebidium parasiticum* and four in *Pirum gemmata*. We did not identify T-box genes in either of the two sequenced choanoflagellates (*Salpingoeca rosetta* and *Monosiga brevicollis*), confirming that T-box genes were also lost in choanoflagellates (16).

To classify the newly reported T-box genes, we performed a full phylogenetic analysis incorporating metazoan, protist and fungal T-box genes. The resulting tree demonstrates that all fungal T-box homologs, as well as one (*C. owczarzaki, M. vibrans, A. whisleri.,* and *P. gemmata*) or several (*C. fragrantissima, S. arctica,* and *A. parasiticum*) homologs from both filastereans and ichthyosporeans cluster at the base of the Brachyury class (Fig. S1). These results support the notion that Brachyury is the most ancient member of the T-box family (11, 13). Moreover, fungal, and especially filasterean, Brachyury genes have most of the T-box key DNA-binding and dimerization amino acids and even have conserved exclusive amino acid motifs of the Brachyury class (see Fig. S2). In contrast, this is not the case for the highly divergent T-box genes from ichthyosporeans, which lack most known functional T-box domain amino acids (Fig. S2).

The other filasterean and ichthyosporean T-box genes represent a new class of T-boxes that we call Tbx7, including homologs from the two filastereans (*C. owczarzaki* and *M. vibrans*), two ichthyosporeans (*A. whisleri* and *P. gemmata*), and sponges (*S.ciliatum, L. complicata* and *A.queenslandica*)(see Fig. 1 and Fig. S1). Among the Tbx7 group there is a *C. owczarzaki* T-box gene with two T domains, a configuration not present in any reported T-box gene. This is, however, not uncommon among other eukaryotic transcription factor families. It has been hypothesised that multiple DNA binding domains can increase the length and diversity of DNA recognition motifs recognisable by the limited number of DNA binding domain families

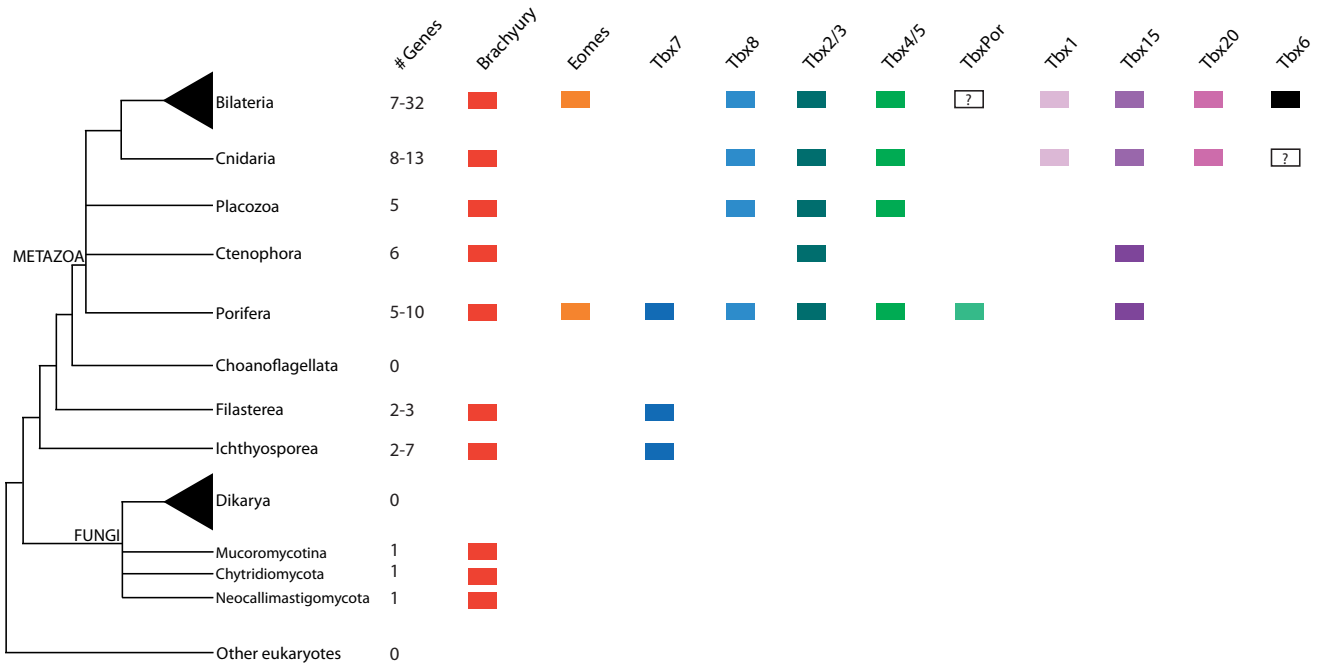


Fig. 1. Phylogenetic distribution of different T-box classes among opisthokonts. The first column indicates the minimum and maximum number of T-box genes found in each lineage. Consensus phylogenetic relationship are shown (17, 19, 22, 23, 48).

(20, 21). Whether this or other explanations account for the presence of this T-box gene in *C. owczarzaki* remains to be elucidated.

A revised evolutionary history of metazoan T-box classes

Previous studies have identified T-box genes in early-branching metazoans, resulting in a firm understanding of the repertoire of T-box genes in non-bilaterian animals (11–14). This knowledge has enabled a reconstruction of the evolutionary history of the T-box family. For example, it has been inferred that the Urmetazoan T-box complement included 3 classes (Tbx4/5, Brachyury and a putative Tbx1/15/20), with other classes being added in a step-wise manner through the evolution of metazoans. Thus, Tbx2/3, Tbx1, Tbx15 and Tbx20 originated within eumetazoans (Cnidaria+Bilateria), while Tbx6 and Eomes classes originated within bilaterians.

Our phylogenetic analysis, which includes new data not only from several fungi and unicellular relatives of Metazoa, but also from two calcarean sponges, allows us to re-evaluate the evolutionary history of the T-box family. Our data show that sponges, potentially the earliest-branching Metazoa (22, 23), have a much more complex complement of T-box genes than previously thought (Fig. 1).

Both the homoscleromorph sponge *Oscarella carmela* and the ctenophore *Mnemiopsis leidyi* have a Tbx1/15/20 homolog, which shares with all other Tbx1/15/20 members an exclusive amino acid insertion (Fig. S2). The presumed Tbx1/15/20 identified in the demosponges *Amphimedon queenslandica* and *Axinella verrucosa* (13) were previously thought to comprise a new demosponge-specific T-box class (14). We also recover this group, but surprisingly it also includes a sequence from the deuterostome *Saccoglossus kowalevski*. We have preserved the nomenclature TbxPor, following Holstien et al. (2010).

We identified a Tbx2/3 class member in the sponge O. carmella, as well as in the ctenophores *M. leidyi* and *Pleurobrachia pileus*. This suggests that the Tbx2/3 class was already present at the origin of animals. In agreement with previous results, we identified Tbx4/5 in most early-branching metazoans, except in ctenophores.

Our analyses recovered a clade of exclusively non-bilaterian representatives (the calcarean sponges *L. complicata* and *S. ciliatum*, the demosponge *A. queenslandica*, the filastereans *C. owczarzaki* and *M. vibrans*, and the ichthyosporeans *P. gemmata* and *A. whisleri*), all of which have highly divergent sequences (Fig. S2); we designate this clade Tbx7. We also define the group Tbx8, which to date includes only sponges (demosponges and *O. carmella*), *T. adharens*, *N. vectensis* and two bilaterians (*S. kowalevskii* and *L. gigantea*). Both groups appear to have been lost in some lineages during metazoan evolution.

As in previous studies (7, 14), our data do not support the monophyly of Tbx6 class, but no putative orthologs were identified in basal metazoans. Thus, this class likely evolved later during metazoan evolution. Further, in contrast to previous reports that considered the Eomes class as a bilaterian innovation (13), we could identify homologs in the calcarean sponges *L. complicata* and *S. ciliatum*.

Finally, our results suggest that Brachyury is the most widely distributed class of T-box genes, with members present in all major clades: sponges (Calcarea, Demospongia, Homoscleromorpha and Hexactinellida), ctenophores, placozoans, cnidarians, all analyzed bilaterians, and all non-metazoan taxa with T-box family members. Under this new scenario (Fig. 1), Brachyury was the ancestral T-box gene from which all other classes evolved. Further, two classes of T-boxes were already present at the origin of the Holozoa (Bra and Tbx7), and the Urmetazoan T-box complement was much larger than previously thought (Bra, Eomes, Tbx2/3, TbxPor, Tbx4/5, Tbx1/15/20, Tbx7 and Tbx8), suggesting that the diversification of T-box classes at the onset of Metazoa was therefore explosive, rather than step-wise.

Overall, our data show that T-box is an ancient transcription factor, with members present in several species belonging to five different non-metazoan lineages (Filasterea, Ichthyosporea and the basal fungi Neocallimastigomycota, Chytridiomycota and Mucoromycotina). Evolutionarily, the T-box family is highly dynamic, with multiple secondary losses (with the exception of Brachyury, which is conserved in many lineages, but lost for example in *C. elegans* (8) and *A. queenslandica*), some fast-evolving members (for example in sponges and ichthyosporeans), expansions (such as three paralogous Eumetazoan classes related to the ancestral Tbx1/15/20), and major rearrangements, such as the double T-box domain found in C. owczarzaki (16).

Functional conservation of *C. owczarzaki* and *S. ciliatum* Brachyury

Given its univocal phylogenetic position and the high degree of conservation at the amino acid level of *C. owczarzaki* Brachyury (Fig. S1, Fig. S2), we decided to test its functional conservation within a metazoan context. We also included another *C. owczarzaki* T-box in our analyses (CoTbox3, a member of the Tbx7 class). We used Xenopus as a model system, as it has previously been used to characterize T-box genes from early-branching metazoans (4, 7, 18).

Xenopus embryos injected with an mRNA encoding a dominant negative form of Brachyury (XBra_En) show defective gastrulation and impairment of muscle development (38). This phenotype is partially rescued by co-injection of XBra mRNA. We used embryos injected with XBra_En mRNA to compare the rescue capacity of XBra, *C. owczarzaki* Brachyury (CoBra) and *C. owczarzaki Tbox3* (CoTbx3) mRNAs (Fig. 2). Surprisingly, both *C. owczarzaki* genes rendered a proportion of rescued embryos similar to those observed in embryos injected with the endogenous XBra (Fig. 2), as determined by the general shape of the injected embryos and by in-situ hybridization for the muscle gene MyoD. This suggests that both CoBra and CoTbx3 can largely mimic endogenous XBra function.

We next evaluated if this similar rescue potential is the consequence of the capacity of these genes to activate similar downstream target genes. It has been show that not all T-box genes activate the same target genes. For example, Tbx6 (VegT), Eomes and Brachyury can all activate the mesendodermal genes Wnt11 and Sox17 while chordin is only activated by the first two but not by Brachyury (7, 24). This difference seems to be due to the ability of Brachyury to interact with the cofactor Smad1. This interaction, which takes place through an N-terminal domain of the Brachyury protein, allows the activation of Xom, a repressor of dorsal mesendodermal genes (25, 26). We therefore compare the ability of different T-box genes to activate these three target genes in Xenopus overexpression assays. For comparison, we also included the T-box gene of another non-metazoan, the fungus *S. punctatus* (SpBra), the Bra gene of the cnidarian *N. vectensis* (NvBra) and the two Bra paralogs of the calcarean sponge *S. ciliatum* (SciBra1 and SciBra2).

A summary of the molecular phenotypes is shown in Figure 3. Our data define three clear groups of molecular phenotypes. One is defined by metazoan homologs (NvBra, SciBra1 and SciBra2), which have a molecular phenotype similar to that obtained by the endogenous XBra Thus, both cnidarian and sponge homologs largely behave like XBra and strongly activate Sox17 and Wnt11, but are either unable to activate chordin or do it in a very low proportion of embryos and at low levels. Nevertheless, their potential to activate Sox17 and Wnt11 is somehow different. Thus, in embryos injected with NvBra, Sox17 activation is stronger than that of Wnt11, a result similar to what was found for the Brachyury of the ctenophore *Mnemiopsis leidyi* (7). On the other hand, the two homologs of S. ciliatum (SciBra1 and SciBra2), like XBra, activate Wnt11 more strongly

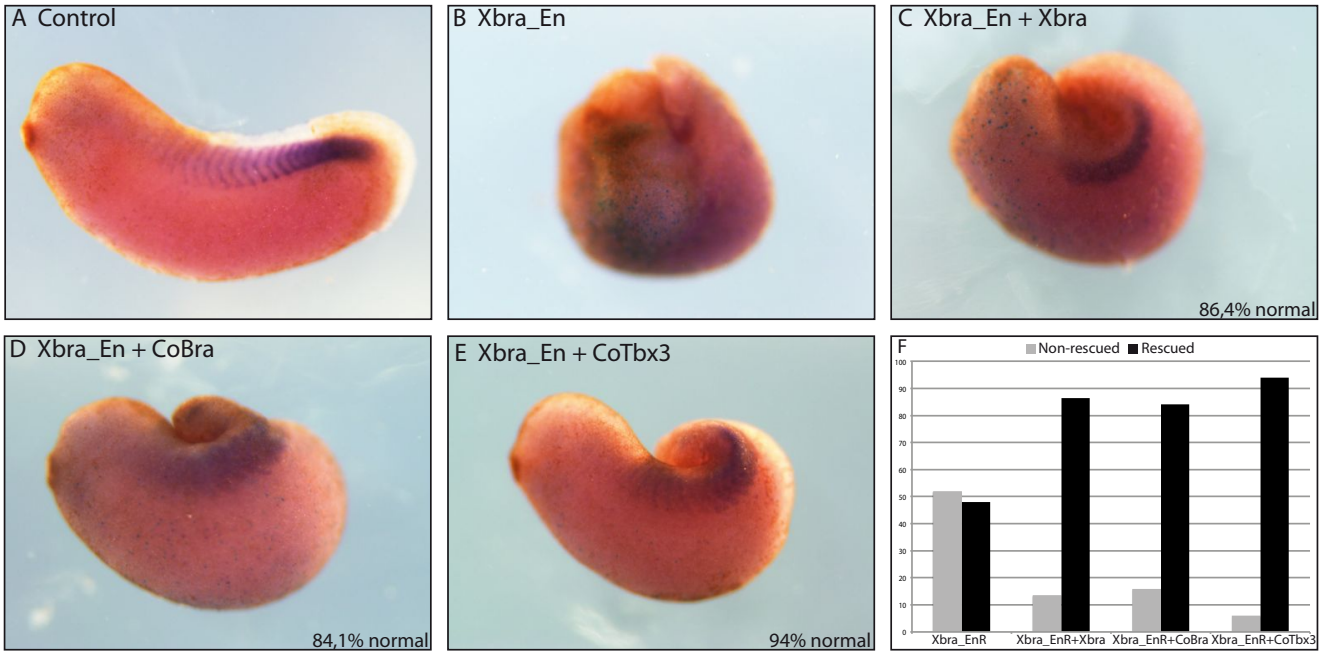


Fig. 2. Xenopus rescue experiments of *C. owczarzaki* Bra (CoBra) and Tbox-3 (CoTbx3). In situ hybridization of MyoD showing mesoderm formation. Embryos were injected with 500 pg of the corresponding mRNA. a) Control b) Embryo injected with a dominant negative Xbra mRNA (Xbra_En) c) Xbra_En injected embryo rescued injecting Xbra mRNA. d) Embryo rescued injecting CoBra mRNA. e) Embryo rescued injecting CoTbx3. d) The proportion of "rescued"/"non-rescued" phenotype embryos under the different conditions (b-e).

than Sox17. Despite their ability to partially rescue the loss of Xbra function, a different molecular phenotype was observed in embryos injected with *C. owczarzaki* homologs (CoBra and CoTbx3) mRNAs, which were able to strongly activate all three mesendoderm genes. This suggest a clear boundary between metazoan and non-metazoan Brachyury homologs, which may be explained by the ability of the metazoan Brachyury orthologs to interact with cofactors that restrict their function, such as Smad1. Interestingly, this factor is present in the genome of *S. ciliatum*, but not in the genome of C. owczarzaki (16). Finally, we found a different molecular phenotype in embryos injected with the fungus homolog (SpBra), which strongly activates Sox17, but not chordin or Wnt11.

Several amino acid motifs have been suggested to be key determinants of the specificity of Brachyury, compared to other T-box family members. We therefore asked whether any of these motifs could account for the differences we have observed between metazoan and non-metazoan homologs. Conlon et al. (2001) proposed that the presence of a Lysine in position 149 of Xbra accounts for its differential behaviour, compared to other T-box classes such as VegT (Tbox6) and Eomes, which instead have an Asparagine at this position. Our alignments (Fig. S2) indicate that this position is indeed conserved in the *N. vectensis* and *S. ciliatum* Brachyury proteins. However, despite the presence of an Arginine (R) instead of a Lysine (K) in the CoBra protein, we do not believe that this difference alone can explain the drastic phenotypic differences we observed between metazoan Brachyury and CoBra, especially when considering that R is a hydrophilic basic amino acid, being extremely similar to Lysine, and is very different from the neutral Asn (N) found in all other T-box classes. In fact, SciBra2 also has an R in this position and, nonetheless, it does not activate chordin (Fig. 3). Messenger et al. (2005) proposed that an N-terminal domain is responsible for the interaction between Brachyury and Smad1, which would restrict its function spatially. Indeed organisms whose Brachyury lack this domain, such as Drosophila Bra or ascidian Bra, are unable to behave as endogenous XBra (26). However, the ctenophore *M. leidyi*, the sponge *S. ciliatum*, and the cnidarian *N. vectensis* Bra homologs lack the conserved N-terminal region and, nonetheless, behave similarly to XBra. Finally, Bielen et al. (2007) proposed that a conserved motif in the C-terminal activation domain (called the R1 domain), which is present in Bilateria and Cnidaria, is responsible for Brachyury specificity. However, this domain is again not present in *M. leidyi* or *S. ciliatum* Brachyury homologs. Thus, our data, together with the results from Yamada et al (2010), suggest that the difference between metazoan and non-metazoan Brachyury homologs in their ability to mimic endogenous Xbra functions, cannot be explained by the presence of specific domains outside of the T domain, but instead are most likely caused by specific amino acids located inside the T domain. It is this conservation within the T domain that could account

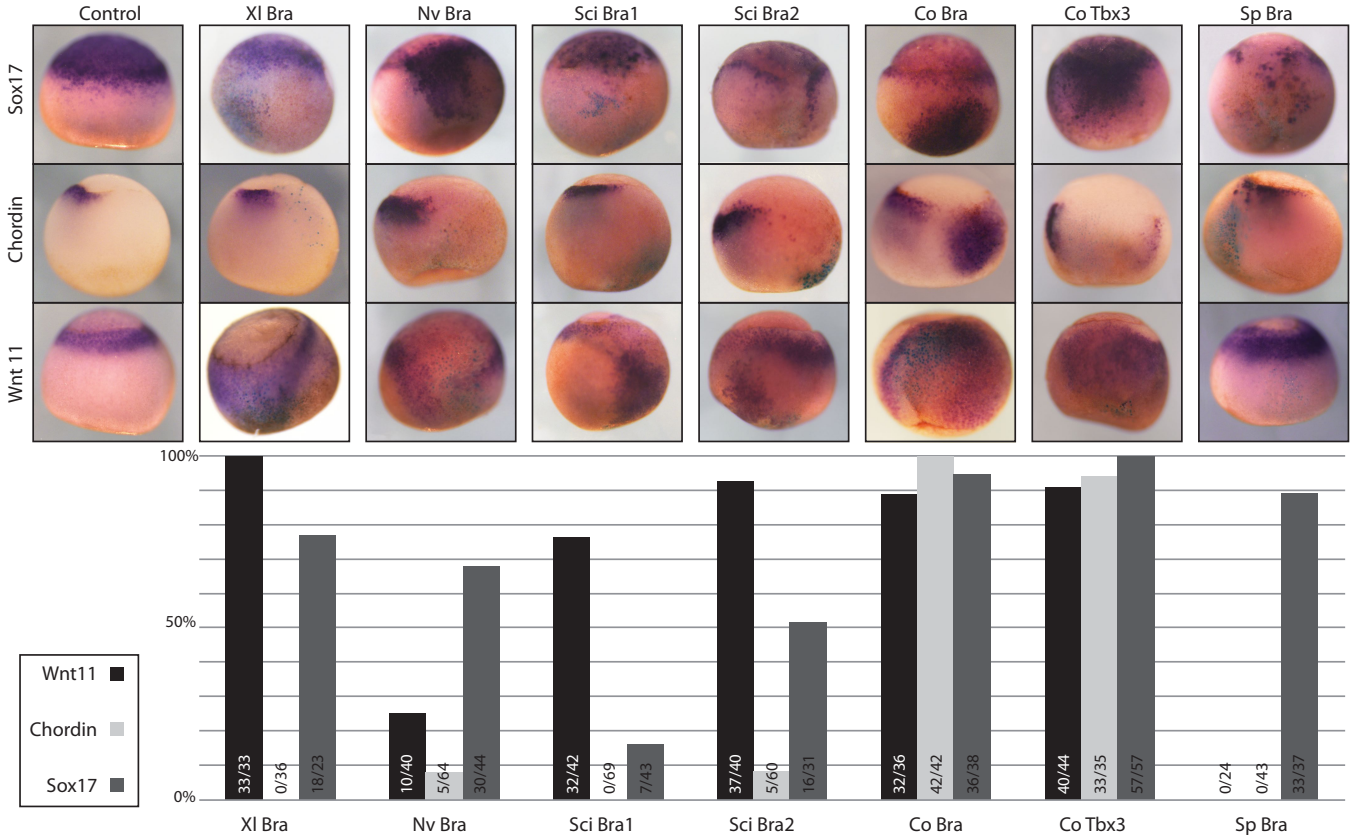


Fig. 3. Molecular phenotype of injected Xenopus embryos based on in-situ hybridization for the markers Wnt11, Chordin and Sox17. The graph depicts the percentage of injected embryos showing activation for each of the three markers. Total numbers (number embryos with activation/total injected embryos) are shown at the base. All embryos were injected with 1000 pg of the corresponding mRNA.

for the ability of sponge and ctenophore Bra to interact with the same common molecular components as that of Xbra and, therefore, the functional interchangeability between these genes.

C. owczarzaki Brachyury has a conserved T-box DNA binding motif

To further investigate the function of *C. owczarzaki* Brachyury, we determined its binding preferences using universal protein binding microarrays (PBM) (27, 28). Our results indicate that CoBra has a highly similar motif to that determined in the mouse Bra-homolog, called T (Fig. 4) (24, 29–31). Moreover, our results indicate that the T-box DNA recognition sequence is strongly conserved, both across a wide range of T-box classes (including Eomes (32), Tbx1, Tbx4 and Tbx2) and also across different organisms (Fig. 4, Fig. S1).

Thus, our data from the protist *C. owczarzaki* suggests that T-box genes have preserved a DNA recognition motif that has undergone very little change during evolutionary time, even with the diversification of the family at the origin of Metazoa. These results suggest that cooperative interactions of T-box genes with different cofactors, as opposed to differences in DNA binding sequence recognition, are the key means through which members of this family have diverged in function. Similar findings have been reported, for example, for Hox family transcription factors (33). Moreover, it is likely that regulation of temporal expression could contribute to differences in function. The conserved binding motif also helps to explain the ability of CoBra to rescue endogenous Xbra and to activate several downstream T-box targets in Xenopus, but without the specificity of XBra, probably due to the inability of CoBra to interact with cofactors. In sharp contrast, the Brachyury orthologs of the basal metazoan *S. ciliatum* can perfectly mimic the behaviour of endogenous XBra.

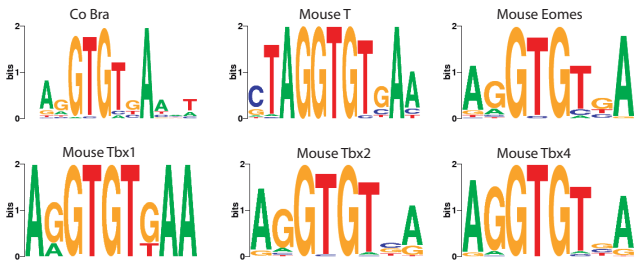


Fig. 4. CoBra binding motifs derived from Protein Binding Microarray data (see Methods). For comparison, different mouse T-box classes binding motifs also derived from PBM data (except mouse T, based on SELEX (29)).

Conclusion

Our data, which includes several previously unreported T-box genes from sponges, fungi, ichthyosporeans and filastereans, allows us to reconstruct T-box transcription factor family evolution with unprecedented detail. We have also analyzed the most conserved non-metazoan Brachyury homolog known to date, that of the filose amoeba *C. owczarzaki*, a close relative of Metazoa (17, 34).

Our results demonstrate that the repertoire of T-box transcription factors in pre-metazoans is much richer than previously thought, with members of this family present in several fungi, ichthyosporeans and filastereans. T-box genes evolved in the last common ancestor of all opisthokonts (Fig. 1), and were secondarily lost in higher fungi (Dikarya) and choanoflagellates. Phylogenetic analyses and molecular signatures confirm that Brachyury is the most ancient member of the T-box genes, being present in both fungi and unicellular holozoans. A new member of the family (Tbox7) evolved later, within the Holozoan clade, with members present in some ichthyosporeans, filastereans and sponges, but was secondarily lost in other metazoans. The T-box family radiated at the origin of Metazoa, in a highly dynamic scenario with some fast evolving classes (such as Tbx7 and Tbx8) and some classes that have been secondarily lost (such as Tbx7 and TbxPor, which are only present in sponges). After this initial period, the number of classes stabilised until the emergence of Tbx1, Tbx15 and Tbx20 from a common Tbx1/15/20 ancestor at the stem of Cnidaria+Bilateria, and the origin of Tbx6 at the stem of Bilateria.

The subfunctionalization of Brachyury seems to have been well established at the very origin of the Metazoa. However, the high number of T-box classes (including some, like Tbx7 or TbxPor, that were subsequently lost), the uneven distribution of T-box classes in sponges, and the presence of extremely fast evolving T-box genes in sponges suggest an early scenario of fast evolution of new T-box classes. These fast evolving T-box genes, present in the genomes of sponges and some protists, might simply be the remnants of this initial explosive evolution.

Results of our analyses indicate that the binding specificity of Brachyury is highly conserved among metazoan and non-metazoans, as well as between Brachyury and other T-box classes. This reinforces the idea that cofactor interactions may be responsible for the functional differences observed between different T-box classes, and may also explain why the Brachyury of *C. owczarzaki*, although clearly a Bra ortholog, does not have the ability to interact with cofactors in a Xenopus heterologous context, in sharp contrast to the Brachyury of sponges, ctenophores or cnidarians. Most likely, these restrictions were set at the origin of Metazoa, with the radiation of T-box classes, as evidenced by the perfect functional mimic of SciBra and M. leidyi Bra (7) with XBra. In that sense, both CoBra and CoTbx3 (a member of the Tbx7 class) behave as what we call “pan-Tbox” genes, activating all potential targets (like chordin) that will later in evolution be controlled by specific T-box classes (in the case of chordin, Tbx6 and Eomes). Through time, novel T-box specificities were established through evolution of functional interactions with cofactors.

Materials and methods

Microinjection of Brachyury genes into Xenopus embryos. The entire coding regions of Brachyury genes from different species (CoBra, CoTbox3, SciBra1, SciBra2, NvBra, SpBra) were inserted into the multicloning site of pCS2+ (35). mRNAs, prepared as previously described (36), were injected in Xenopus embryos at two-four-cell stage in a single blastomere at 1000pg per embryo. X-Gal staining was performed as described elsewhere (37).

Rescue experiments. A Xbra dominant negative construct (XBra-EnR) (38), was co-injected with Xbra, CoBra and CoTbox3 mRNAs into Xenopus embryos at 500pg per embryos each one.

Histochemistry. Xenopus embryos were fixed in MEMFA (0,1M MOPS, 2mM EGTA, 1mM $MgSO_4$, 3,7% formaldehyde, pH7,4) for 1hour at room temperature and them kept in methanol at -20ºC. Antisense RNA probes were prepared from cordin, wnt11 and sox17β cDNAs using digoxigenin (Roche). Xenopus embryos were hybrydized as described (39). After inmunostaining embryos were bleached by the treatment with 10% H_2O_2 in PBS under the light for 2-3 hours.

Gene searches & Phylogenetic analysis. A primary search was performed using the basic local alignment sequence tool (BLAST: BlastP and TBlastN) using Homo sapiens and C. owczarzaki proteins as queries against Protein, Genome and Transcriptome databases with the default BLAST parameters and an e-value threshold of e-5 at the National Center for Biotechnology Information (NCBI) and against completed or on-going genome project databases at the Joint Genome Institute (JGI) (for Piromyces sp., Gonapodya prolifera and other basal fungi available), the Broad Institute (for Mortierella verticillata, Salpingoeca rosetta, Sphaeroforma arctica and Spizellomyces punctatus), as well as the A. queenslandica genome database (www.metazome.net/amphimedon). In the case of Abeoforma whisleri, Pirum gemmata, Amoebidium parasiticum and Ministeria vibrans we assembled the trace RNAseq data using the Trinity assembler and in the case of C. fragrantissima we assembled trace genomic sequencing data using the WGS assembler ("http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title5Main_Page"). In both cases, we performed local BLAST searches and we annotated the sequences manually. We also performed Hmmer searches using HMMER3.0b2 (40) to confirm that we were retrieving all T-box orthologs.

Alignments were constructed using the MAFFT v.6 online server (41) and then manually inspected and edited in Geneious. Only those species and those positions that were unambiguously aligned were included in the final analyses. The best-fit model for our set of proteins was chosen using ProtTest server (42). Maximum likelihood (ML) phylogenetic trees were estimated by RaxML (43) using the PROTGAMMALG+Γ+I model, which uses the LG amino acid exchangeabilities and accounts for among-site rate variation with a four category discrete gamma approximation and a proportion of invariable sites. Statistical support for bipartitions was estimated by performing 100-bootstrap replicates using RaxML with the same model. Bayesian analyses were performed with MrBayes3.2 (44), using the LG+Γ+I model of evolution, with four chains, a subsampling frequency of 100 and two parallel runs. Runs were stopped when the average standard deviation of split frequencies of the two parallel runs was <0.01, usually at around 18,000,000 generations. The two LnL graphs were checked and an appropriate burn-in length established. Bayesian posterior probabilities (BPP) were used to assess the confidence values of each bipartition.

Protein Binding Microarrays. Details of the design and use of universal PBMs has been described elsewhere (27, 28). Here, we used two different universal PBM array designs, designated 'ME' and 'HK', after the initials of their designers (45, 46). The T-box DNA binding domain of all analysed T-box genes (see Table S1), along with 50 amino acid "pads" flanking either side, were cloned as SacI–BamHI fragment into pTH5325, a modified T7-driven GST expression vector. We used 150 ng of plasmid DNA in a 15 µl in vitro transcription/translation reaction using a PURExpress In Vitro Protein Synthesis Kit (New England BioLabs) supplemented with RNase inhibitor (Invitrogen) and 50 µM zinc acetate. After a 2-h incubation at 37ºC, 12.5 ml of the mix was added to 137.5 ml of protein-binding solution for a final mix of PBS/2% skim milk/0.2 mg per ml BSA/50 µM zinc acetate/0.1% Tween-20. This mixture was added to an array previously blocked with PBS/2% skim milk and washed once with PBS/0.1% Tween-20 and once with PBS/0.01% Triton-X 100. After a 1-h incubation at room temperature, the array was washed once with PBS/0.5% Tween-20/50 mM zinc acetate and once with PBS/0.01% Triton-X 100/50 mM zinc acetate. Cy5-labeled anti-GST antibody was added, diluted in PBS/2% skim milk/50 mM zinc acetate. After a 1-h incubation at room temperature, the array was washed three times with PBS/0.05% Tween-20/50 mM zinc acetate and once with PBS/50 mMzinc acetate. The array was then

imaged using an Agilent microarray scanner at 2 mM resolution. Image spot intensities were quantified using ImaGene software (BioDiscovery). A position frequency matrix (PFM) motif was created from the PBM data by aligning all 8mers with E-scores > 0.45 (27, 28) using ClustalW (47), trimming the alignment by restricting to positions present in at least half of the sequences in the alignment, and converting each remaining position to frequencies. Plasmid sequences, array intensity data, E-scores, Z-scores, and PWMs will be available in the Cis-BP database (Weirauch et al., in prep), and are also available upon request.

Author contributions: A.S.-P., JL.G.-S., T.R.H. and I.R-T. designed research; A.S.-P., A.A.-C. and A.Y. performed research; S.L., G.T., M.A. and M.A. contributed new reagents/analytic tools; A.S.-P., A.A.-C., M.T.W., T.R.H., I.R-T, and JL.G.-S. analysed data; and A.S.-P., M.W., T.R.H., JL.G.-S. and I.R.-T. wrote the paper.

Acknowledgments

We thank Joint Genome Institute and Broad Institute for making data publicly available. We thank Ignacio Maeso, Alex de Mendoza and other members of the multicellgenome lab for useful insights. We thank Ana Gilles and Ulrich Technau (University of Viena) for providing the N. vectensis Brachyury clon. This work was supported by an Institució Catalana per a la Recerca i Estudis Avançats contract, a European Research Council Starting Grant (ERC-2007-StG- 206883), a grant (BFU2011-23434) from Ministerio de Economía y Competitividad (MINECO) to I. R.-T.. A.S-P. was supported by a pregraduate Formacion Profesorado Universitario grant from MICINN and a grant from MICINN to perform a research stay at JL.G.-S. lab. JL.G.-S. thank the Spanish and Andalusian Governments for grants (BFU2010-14839, CSD2007-00008 and Proyecto de Excelencia CVI-3488) for funding this study. M.A., M.A. and S.L. acknowledge funding from the core budget of the Sars International Centre for Marine Molecular Biology. MTW was supported by fellowships from CIHR and the Canadian Institute for Advanced Research (CIFAR) Junior Fellows Genetic Networks Program.

1. Papaioannou VE, Silver LM (1998) The T-box gene family. BioEssays 20:9–19.
2. Smith J (1999) T-box genes: what they do and how they do it. Trends in Genetics 15:154–158.
3. Wilson V, Conlon FL (2002) The T-box family. Genome biology 3:3008.1–3008.7.
4. Marcellini S, Technau U, Smith J, Lemaire P (2003) Evolution of Brachyury proteins: identification of a novel regulatory domain conserved within Bilateria. Developmental biology 260:352–361.
5. Scholz CB, Technau U (2003) The ancestral role of Brachyury: expression of NemBra1 in the basal cnidarian Nematostella vectensis (Anthozoa). Development genes and evolution 212:563–570.
6. Showell C, Binder O, Conlon FL (2004) T-box genes in early embryogenesis. Developmental dynamics 229:201–218.
7. Yamada A, Martindale MQ, Fukui A, Tochinai S (2010) Highly conserved functions of the Brachyury gene on morphogenetic movements: Insight from the early-diverging phylum Ctenophora. Developmental biology 339:212–222.
8. Papaioannou VE (2001) T-box genes in development: from hydra to humans. International review of cytology 207:1–70.
9. Technau U (2001) Brachyury, the blastopore and the evolution of the mesoderm. BioEssays 23:788–94.
10. Martinelli C, Spring J (2005) T-box and homeobox genes from the ctenophore Pleurobrachia pileus: comparison of Brachyury, Tbx2/3 and Tlx in basal metazoans and bilaterians. FEBS letters 579:5024–8.
11. Adell T, Grebenjuk V a, Wiens M, Müller WEG (2003) Isolation and characterization of two T-box genes from sponges, the phylogenetically oldest metazoan taxon. Development genes and evolution 213:421–34.
12. Manuel M, Parco Y Le, Borchiellini C (2004) Comparative analysis of Brachyury T-domains, with the characterization of two new sponge sequences, from a hexactinellid and a calcisponge. Gene 340:291–301.
13. Larroux C et al. (2008) Genesis and expansion of metazoan transcription factor gene classes. Molecular biology and evolution 25:980–96.
14. Holstien K et al. (2010) Expansion, diversification, and expression of T-box family genes in Porifera. Development genes and evolution 220:251–62.
15. King N et al. (2008) The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 451:783–8.

16. Sebé-Pedrós A, Mendoza A de, Lang BF, Degnan BM, Ruiz-Trillo I (2011) Unexpected repertoire of metazoan transcription factors in the unicellular holozoan Capsaspora owczarzaki. Molecular biology and evolution 28:1241–54.
17. Torruella G et al. (2012) Phylogenetic Relationships within the Opisthokonta Based on Phylogenomic Analyses of Conserved Single-Copy Protein Domains. Molecular biology and evolution 29:531–44.
18. Bielen H et al. (2007) Divergent functions of two ancient Hydra Brachyury paralogues suggest specific roles for their C-terminal domains in tissue fate induction. Development 134:4187.
19. Grigoriev I, Cullen D, Goodwin S, Hibbett D (2011) Fueling the future with fungal genomics. Mycology 2:192–209.
20. Charoensawan V, Wilson D, Teichmann S a (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. Nucleic acids research 38:7364–77.
21. Itzkovitz S, Tlusty T, Alon U (2006) Coding limits on the number of transcription factors. BMC genomics 7:239.
22. Philippe H et al. (2009) Phylogenomics revives traditional views on deep animal relationships. Current biology 19:706–12.
23. Pick KS et al. (2010) Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Molecular biology and evolution 27:1983–7.
24. Conlon FL, Fairclough L, Price BM, Casey ES, Smith JC (2001) Determinants of T box protein specificity. Development 128:3749–58.
25. Messenger NJ et al. (2005) Functional specificity of the Xenopus T-domain protein Brachyury is conferred by its ability to interact with Smad1. Developmental cell 8:599–610.
26. Marcellini S (2006) When Brachyury meets Smad1: the evolution of bilateral symmetry during gastrulation. BioEssays 28:413–420.
27. Berger MF et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature biotechnology 24:1429–35.
28. Berger MF, Bulyk ML (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nature protocols 4:393–411.
29. Kispert A (1993) The Brachyury gene encodes a novel DNA binding protein. The EMBO journal 12:3211–3220.
30. Casey ES, O'Reilly M a, Conlon FL, Smith JC (1998) The T-box transcription factor Brachyury regulates expression of eFGF through binding to a non-palindromic response element. Development 125:3887–94.
31. Garnett AT et al. (2009) Identification of direct T-box target genes in the developing zebrafish mesoderm. Development 136:749–60.
32. Badis G et al. (2009) Diversity and complexity in DNA recognition by transcription factors. Science (New York, N.Y.) 324:1720–3.
33. Slattery M et al. (2011) Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. Cell 147:1270–1282.
34. Ruiz-Trillo I et al. (2004) Capsaspora owczarzaki is an independent opisthokont lineage. Current biology 14:R946–R947.
35. Turner DL, Weintraub H (1994) Expression of achaete-scute homolog 3 in Xenopus embryos converts ectodermal cells to a neural fate. Genes & Development 8:1434–1447.
36. Harland R, Weintraub H (1985) Translation of mRNA injected into Xenopus oocytes is specifically inhibited by antisense RNA. The Journal of Cell Biology 101:1094–1099.
37. Coffman CR, Skoglund P, Harris WA, Kintner CR (1993) Expression of an extracellular deletion of Xotch diverts cell fate in Xenopus embryos. Cell 73:659–671.
38. Conlon FL, Sedgwick SG, Weston KM, Smith JC (1996) Inhibition of Xbra transcription activation causes defects in mesodermal patterning and reveals autoregulation of Xbra in dorsal mesoderm. Development 122:2427–2435.
39. Jones CM, Smith JC (1999) Wholemount In Situ Hybridization to Xenopus Embryos. Molecular Embryology 97:635–640.
40. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755.
41. 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 research 30:3059.
42. Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics (Oxford, England) 21:2104–5.
43. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688.
44. Huelsenbeck JPP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–5.

45. Mintseris J, Eisen MB (2006) Design of a combinatorial DNA microarray for protein-DNA interaction studies. BMC Bioinformatics 7:429.
46. Philippakis A a, Qureshi AM, Berger MF, Bulyk ML (2008) Design of compact, universal DNA microarrays for protein binding microarray experiments. Journal of computational biology 15:655–65.
47. Chenna R, Sugawara H, Koike T (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic acids 31:3497–3500.
48. Shalchian-Tabrizi K et al. (2008) Multigene phylogeny of choanozoa and the origin of animals. PloS one 3:e2098.

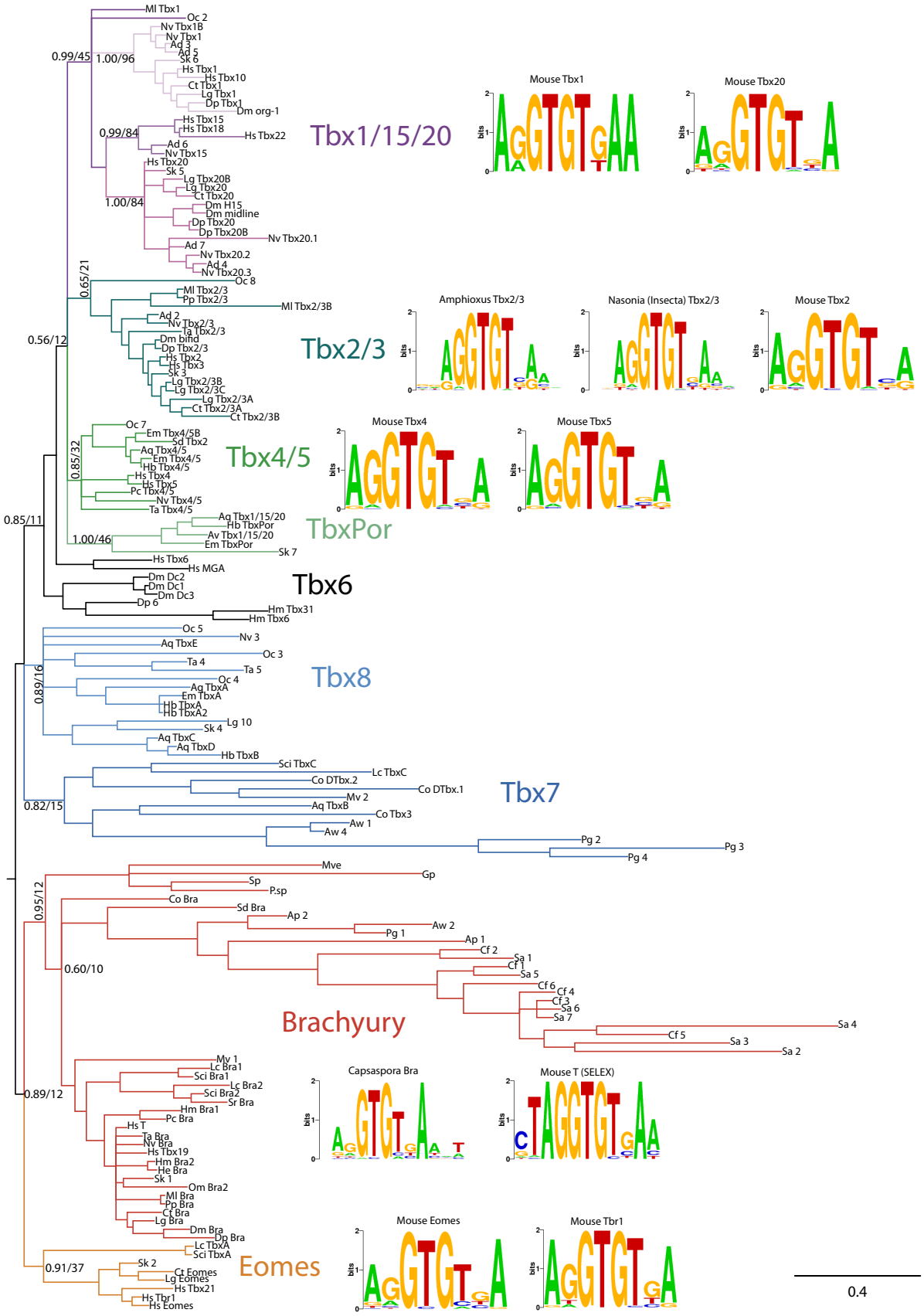


Fig. S1. Bayesian inference tree of T-box domains showing the different T-box families. The tree is rooted using the midpoint-rooted tree option. Statistical support values indicate bayesian posterior probablities (BPP) and 1,000 ML bootstrap replicates (BV). Colors correspond to different T-box families (same as in Fig. 1). Taxa include Ad (*Acropora digitifera*), Aq (*Amphimedon queenslandica*), Av (*Axinella verrucosa*), Ct (*Capitella teleta*), Co (*Capsaspora owczarzaki*), Dm (*Drosophila melanogaster*), Dp (*Daphnia pulex*), Em (*Ephydatia muelleri*), Gp (*Gonapodya prolifera*), Hb (*Halichondria bowerbanki*), He (*Hydractinia echinata*), Hm (*Hydra magnipapillata*), Hs (*Homo sapiens*), Lc (*Leucosolenia complicata*), Lg (*Lottia gigantea*), Ml (*Mnemiopsis leydi*), Mv (*Ministeria vibrans*), Mve (*Mortierella verticillata*), Nv (*Nematostella vectensis*), Oc (*Oscarella carmela*), Om (*Oopsaca minuta*), Pc (*Podocoryne carnea*), Pp (*Pleurobrachia pileus*), P.sp (*Pyromices* sp.), Sci (*Sycon cilitaum*), Sd (*Suberites domuncula*), Sk (*Saccoglossus kowalevskii*), Sp (*Spizellomyces punctatus*), Sr (*Sycon raphanus*), and Ta (*Trichoplax adhaerens*). Co DTbx.1 and Co DTbx.2 are the two T-box domains of the same T-box *C. owczarzaki* gene (for further details, see main text). PBM-based (except Mouse T, based on a SELEX experiment) DNA binding motifs for several members of different classes are shown (see Methods).

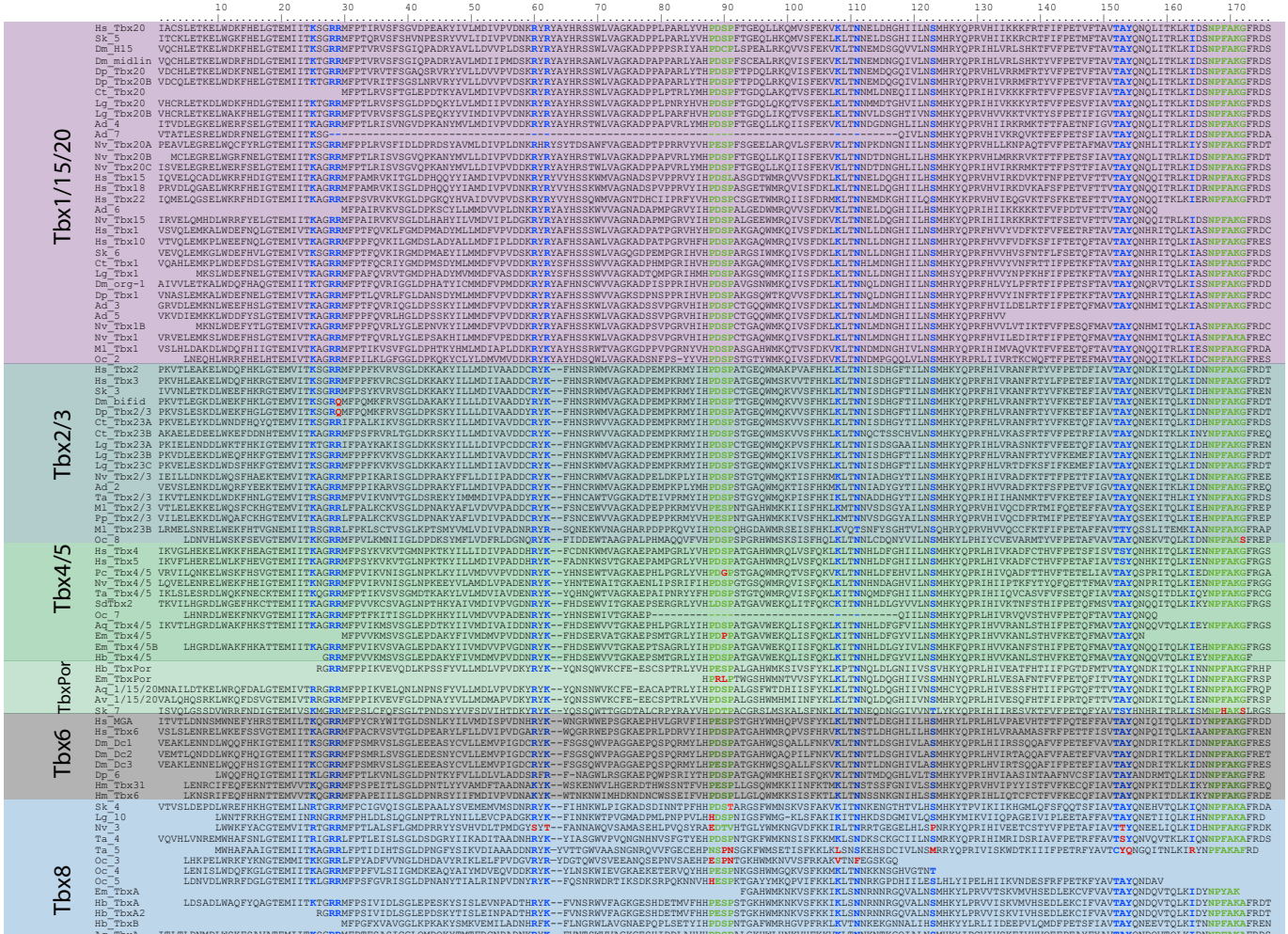


Fig. S2. Alignment of the T-box domain with the different families shown in distinct colors (same as in Fig. 1 and Fig. S1). Key DNA binding amino acids are highlighted in blue and dimerization aminoacids in green. Non-conservative amino acidic changes are depicted in red. Taxa included are the same as in Fig. S1. An arrowhead indicates Lysine149 after Conlon et al. 2001.

DNA binding
Dimerization
TRD Differences

T-box Class	TF_Species	TF_Name	TF_ID	DBID	Motif_Type	Data Source
Brachyury	Mus_musculus	T	ENSMUSG00000062327	MA0009.1	SELEX	JASPAR database
Brachyury	Capsaspora_owczarzaki	CoBra	CoBra	pEX0020	PBM	This work
Eomes	Mus_musculus	Eomes	ENSMUSG00000032446	Eomes_0921	PBM	Badis et al 2009
Tbx1/15/20	Mus_musculus	Tbx1	ENSMUSG00000009097	pTH3822	PBM	This work
Tbx1/15/20	Mus_musculus	Tbx20	ENSMUSG00000031965	pTH3777	PBM	This work
Tbx1/15/20	Mus_musculus	Tbr1	ENSMUSG00000035033	pTH2659	PBM	This work
Tbx2/3	Mus_musculus	Tbx2	ENSMUSG00000000093	pTH3751	PBM	This work
Tbx2/3	Mus_musculus	Tbx3	ENSMUSG00000018604	pTH3998	PBM	This work
Tbx2/3	Nasonia_vitripennis	nasonia_NCBI_hmm584344	nasonia_NCBI_hmm584344	pTH9336	PBM	This work
Tbx2/3	Branchiostoma_floridae	estExt_fgenesh2_pg.C_1570024	estExt_fgenesh2_pg.C_1570024	pTH9244	PBM	This work
Tbx4/5	Mus_musculus	Tbx4	ENSMUSG00000000094	pTH3973	PBM	This work
Tbx4/5	Mus_musculus	Tbx5	ENSMUSG00000018263	pTH3775	PBM	This work

Table S1. List of genes included in Fig. 4 and Fig. S1, including source and motif type.

Species	Gene_ID/Source	Short_name(Fig.S1)	Class
Abeoforma whisleri	Transcriptome data	Aw_1	Tbx7
Abeoforma whisleri	"	Aw_2	Bra
Abeoforma whisleri	"	(Aw_3)	
Abeoforma whisleri	"	Aw_4	Tbx7
Abeoforma whisleri	"	(Aw_5)	
Acropora digitifera	aug_v2a.01412.t1 (from http://marinegenomics.oist.jp/genomes/gallery)	(Ad_1)	
Acropora digitifera	aug_v2a.11828.t1	Ad_2	Tbx2/3
Acropora digitifera	aug_v2a.04553.t1	Ad_3	Tbx1
Acropora digitifera	aug_v2a.20651.t1	Ad_4	Tbx20
Acropora digitifera	aug_v2a.04552.t1	Ad_5	Tbx1
Acropora digitifera	aug_v2a.22872.t1	Ad_6	Tbx15
Acropora digitifera	aug_v2a.20650.t1	Ad_7	Tbx20
Acropora digitifera	aug_v2a.14571.t1	(Ad_8)	
Acropora digitifera	aug_v2a.05404.t1	(Ad_9)	
Acropora digitifera	aug_v2a.20650.t2	(Ad_10)	
Amoebidium parasiticum	Transcriptome data	Ap_1	Bra
Amoebidium parasiticum	"	Ap_2	Bra
Capitella teleta	jgiICapca1l149486ljgiICapca1l149486lestExt_Genewise1.C_1960038:31-459 (from http://genome.jgi.doe.gov/)	Ct_Bra	Bra
Capitella teleta	jgiICapca1l110717ljgiICapca1l110717le_gw1.78.45.1:12-197	(Ct_2)	
Capitella teleta	jgiICapca1l223644ljgiICapca1l223644lestExt_fgenesh1_pg.C_190074:171-345	Ct_Eomes	Eomes
Capitella teleta	jgiICapca1l152028ljgiICapca1l152028lestExt_Genewise1.C_12410001:51-244	Ct_Tbx2/3A	Tbx2/3
Capitella teleta	jgiICapca1l163410ljgiICapca1l163410lestExt_Genewise1.C_2960003:15-201	Ct_Tbx2/3B	Tbx2/3
Capitella teleta	jgiICapca1l226618ljgiICapca1l226618lestExt_fgenesh1_pg.C_2380027:165-367	Ct_Tbx1	Tbx1
Capitella teleta	jgiICapca1l168974ljgiICapca1l168974lestExt_Genewise1Plus.C_7420011:27-168	Ct_Tbx20	Tbx20
Creolimax fragantissima	New genome data	Cf_1	Bra
Creolimax fragantissima	"	Cf_2	Bra
Creolimax fragantissima	"	Cf_3	Bra
Creolimax fragantissima	"	Cf_4	Bra
Creolimax fragantissima	"	Cf_5	Bra
Creolimax fragantissima	"	Cf_6	Bra
Daphnia pulex	jgiIDappu1l44522ljgiIDappu1l44522le_gw1.7.111.1:17-203 (from http://genome.jgi.doe.gov/)	Dp_Bra	Bra
Daphnia pulex	jgiIDappu1l43085ljgiIDappu1l43085le_gw1.5.167.1:66-247	Dp_Tbx2/3	Tbx2/3
Daphnia pulex	jgiIDappu1l59316ljgiIDappu1l59316le_gw1.80.119.1:61-253	Dp_Tbx20	Tbx20
Daphnia pulex	jgiIDappu1l59423ljgiIDappu1l59423le_gw1.80.86.1:10-202	Dp_Tbx20B	Tbx20
Daphnia pulex	jgiIDappu1l3881ljgiIDappu1l3881lgw1.11.80.1:10-193	Dp_Tbx1	Tbx1
Daphnia pulex	jgiIDappu1l16384ljgiIDappu1l16384lgw1.53.185.1:1-174	Dp_Tbx6	Tbx6
Ephydatia muelleri	From Holstein et al (2010)	Em_Tbx4/5A	Tbx4/5
Ephydatia muelleri	"	Em_Tbx4/5B	Tbx4/5
Ephydatia muelleri	"	Em_TbxA	Tbx8
Ephydatia muelleri	"	Em_TbxPor	TbxPor
Gonapodya prolifera	jgiIGanpr1l143538lestExt_fgenesh1_pg.C_100037 (from http://genome.jgi.doe.gov/programs/fungi/index.jsf)	Gp	
Halichondria bowerbanki	From Holstein et al (2010)	Hb_Tbx4/5	Tbx4/5
Halichondria bowerbanki	"	Hb_TbxA1	Tbx8
Halichondria bowerbanki	"	Hb_TbxA2	Tbx8
Halichondria bowerbanki	"	Hb_TbxB	Tbx8
Halichondria bowerbanki	"	Hb_TbxPor	TbxPor
Leucosolenia complicata	New genome data	Lc_Bra2	Bra
Leucosolenia complicata	"	Lc_Bra1	Bra
Leucosolenia complicata	"	Lc_Eomes	Eomes
Leucosolenia complicata	"	(Lc_TbxB)	
Leucosolenia complicata	"	Lc_TbxC	
Lottia gigantea	jgiILotgi1l154800ljgiILotgi1l154800lfgenesh2_pg.C_sca_6000303:45-245 (from http://genome.jgi.doe.gov/)	Lg_Bra	Bra
Lottia gigantea	jgiILotgi1l171118ljgiILotgi1l171118lfgenesh2_pg.C_sca_109000058:8-196	(Lg_2)	
Lottia gigantea	jgiILotgi1l129911ljgiILotgi1l129911le_gw1.67.191.1:17-197	Lg_Eomes	Eomes
Lottia gigantea	jgiILotgi1l117236ljgiILotgi1l117236le_gw1.25.117.1:25-206	Lg_Tbx2/3A	Tbx2/3
Lottia gigantea	jgiILotgi1l116991ljgiILotgi1l116991le_gw1.25.6.1:36-217	Lg_Tbx2/3B	Tbx2/3
Lottia gigantea	jgiILotgi1l117095ljgiILotgi1l117095le_gw1.25.31.1:43-231	Lg_Tbx2/3C	Tbx2/3
Lottia gigantea	jgiILotgi1l179359ljgiILotgi1l179359lfgenesh2_pm.C_sca_63000005:8-196	Lg_Tbx15	Tbx15
Lottia gigantea	jgiILotgi1l104372ljgiILotgi1l104372le_gw1.2.972.1:2-181	Lg_Tbx1	Tbx1
Lottia gigantea	jgiILotgi1l129083ljgiILotgi1l129083le_gw1.63.220.1:51-237	Lg_Tbx15	Tbx15
Lottia gigantea	jgiILotgi1l76774ljgiILotgi1l76774lgw1.19.257.1:1-172	Lg_Tbx8	Tbx8
Ministeria vibrans	Transcriptome data	Mv_1	Bra
Ministeria vibrans	"	Mv_2	Tbx7
Mnemiopsis leidyi	Repredicted	Ml_TbxE	
Mnemiopsis leidyi	Newly predicted from genome data	Ml_Tbx2/3B	
Mortierella verticillata	MVEG_02729 (from http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html)	Mve	
Oscarella carmela	From Holstein et al (2010)	(Oc_1)	
Oscarella carmela	"	Oc_2	Tbx1/15/20
Oscarella carmela	"	Oc_3	Tbx8
Oscarella carmela	"	Oc_4	Tbx8
Oscarella carmela	"	Oc_5	Tbx8
Oscarella carmela	"	(Oc_6)	
Oscarella carmela	"	Oc_7	Tbx4/5
Oscarella carmela	"	Oc_8	Tbx2/3
Piromyces sp.	jgiIPirE2_1l16062lgm1.15008_g (from http://genome.jgi.doe.gov/programs/fungi/index.jsf)	P.sp	
Pirum gemmata	Transcriptome data	Pg_1	Bra
Pirum gemmata	"	Pg_2	Tbx7
Pirum gemmata	"	Pg_3	Tbx7
Pirum gemmata	"	Pg_4	Tbx7
Sphaeroforma arctica	SARC_02039 (from http://www.broadinstitute.org/annotation/genome/multicellularity_project/)	Sa_1	Bra
Sphaeroforma arctica	SARC_02912	Sa_2	Bra
Sphaeroforma arctica	SARC_13371	Sa_3	Bra
Sphaeroforma arctica	SARC_04223	Sa_4	Bra
Sphaeroforma arctica	SARC_04473	Sa_5	Bra
Sphaeroforma arctica	SARC_15599	Sa_6	Bra
Sphaeroforma arctica	Newly predicted from genome data	Sa_7	Bra
Sycon ciliatum	New genome data	Sc_Bra1	Bra
Sycon ciliatum	"	Sc_Bra2	Bra
Sycon ciliatum	"	Sc_TbxA	Eomes
Sycon ciliatum	"	(Sc_TbxB)	
Sycon ciliatum	"	Sc_TbxC	Tbx7
Saccoglossus kowaleskii	Assembled from NCBI data	Sk_1	Bra
Saccoglossus kowaleskii	"	Sk_2	Eomes
Saccoglossus kowaleskii	"	Sk_3	Tbx2/3
Saccoglossus kowaleskii	"	Sk_4	Tbx8
Saccoglossus kowaleskii	"	Sk_5	Tbx20
Saccoglossus kowaleskii	"	Sk_6	Tbx1
Saccoglossus kowaleskii	"	Sk_7	TbxPor?
Saccoglossus kowaleskii	"	(Sk_8)	

Table S2. List of newly annotated genes in Fig. S1, including source, sequence and classification.

Results R5

Premetazoan Origin of the Hippo Signaling Pathway.

RESUM ARTICLE R5: Origen anterior als metazous de la via de senyalització Hippo

La multicel·lularitat no agregativa requereix un estricte control del nombre de cèl·lules. La via de senyalització Hippo coordina la proliferació i l'apoptosi i és un regulador essencial de la mida dels òrgans en animals. Estudis recents mostraren la presència d'elements de la via Hippo en animals no-bilaterals, però no n'identificaren cap fora dels animals. A través d'anàlisis genòmics comparats d'holozous recentment seqüenciats, en aquest treball demostrem que components de la via de senyalització Hippo, com ara les quinases Hippio i Warts, el co-activador Yorkie, i el factor de transcripció Scalloped, estaven presents en els ancestres unicel·lulars dels animals. A més, anàlisis funcionals dels components de la via Hippo de l'holozou ameboide *Capsapsora owczarzaki*, duts a terme en *Drosophila melanogaster*, demostren que l'activitat de control del creixement de la via Hippo està conservada en aquest llinatge unicel·lular. Les nostres troballes mostren que la via Hippo evolucionà molt abans de l'origen dels animals i remarca la importància de la senyalització per Hippo com a un mecanisme del desenvolupament clau anterior a l'origen dels animals.

Cell Reports
Report

Premetazoan Origin of the Hippo Signaling Pathway

Arnau Sebé-Pedrós,[1,5] Yonggang Zheng,[2,5] Iñaki Ruiz-Trillo,[1,3,*] and Duojia Pan[2,4,*]
[1]Institut de Biologia Evolutiva (UPF-CSIC), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain
[2]Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[3]Institució Catalana per a la Recerca i Estudis Avançats (ICREA) and Universitat de Barcelona, Barcelona, Spain
[4]Howard Hughes Medical Institute
[5]These authors contributed equally to this work
*Correspondence: inaki.ruiz@icrea.es (I.R.-T.), djpan@jhmi.edu (D.P.)
DOI 10.1016/j.celrep.2011.11.004

SUMMARY

Nonaggregative multicellularity requires strict control of cell number. The Hippo signaling pathway coordinates cell proliferation and apoptosis and is a central regulator of organ size in animals. Recent studies have shown the presence of key members of the Hippo pathway in nonbilaterian animals, but failed to identify this pathway outside Metazoa. Through comparative analyses of recently sequenced holozoan genomes, we show that Hippo pathway components, such as the kinases Hippo and Warts, the coactivator Yorkie, and the transcription factor Scalloped, were already present in the unicellular ancestors of animals. Remarkably, functional analysis of Hippo components of the amoeboid holozoan *Capsaspora owczarzaki*, performed in *Drosophila melanogaster*, demonstrate that the growth-regulatory activity of the Hippo pathway is conserved in this unicellular lineage. Our findings show that the Hippo pathway evolved well before the origin of Metazoa and highlight the importance of Hippo signaling as a key developmental mechanism predating the origin of Metazoa.

INTRODUCTION

The emergence of multicellularity represents one of the most important transitions in animal evolution. Nonaggregative multicellularity requires strict control over cell differentiation, proliferation and survival and this is attained by sophisticated cell-cell communication systems. A startling revelation from decades of developmental genetic studies is that these cell communications are largely mediated by just a handful of signaling pathways in metazoans such as Notch, Wnt, TGF-β, Hedgehog, and receptor tyrosine kinase (RTK) (Pires-daSilva and Sommer, 2003). A long-standing question in evolutionary developmental biology concerns the genetic mechanisms underlying the transition from unicellular eukaryotes to metazoa, in particular the relationship between multicellularity and the evolutionary origin of the major metazoan signaling pathways (King et al., 2003; Adamska et al., 2007; King et al., 2008; Srivastava et al., 2010; Sebé-Pedrós et al., 2010). Comparative genomic analyses have so far confirmed the absence of Notch, Wnt, TGF-β, and Hedghog signaling in any unicellular organisms, supporting the view that these signaling pathways are metazoan synapomophies (Pires-daSilva and Sommer, 2003). In contrast, the discovery of RTKs in the closest unicellular relatives of animals (choanoflagellates) suggests that RTKs may serve as preadaptations in the metazoans' unicellular ancestors for co-option into the multicellular lifestyle (King et al., 2003; Manning et al., 2008).

The most recent addition to the metazoan signaling "toolkit" is the Hippo signaling pathway. The Hippo pathway was first discovered in *Drosophila melanogaster* as a critical regulator of imaginal disc growth, and more recent studies have implicated a conserved function of this pathway in organ-size control in mammals (Zhao et al., 2010; Pan, 2010; Halder and Johnson, 2011; Zeng and Hong, 2008; Harvey and Tapon, 2007; Reddy and Irvine, 2008; Badouel et al., 2009). The core of the Hippo pathway is a functionally conserved kinase cascade leading from the Ste20-like kinase Hippo (Hpo) (Mst1/Mst2 in mammals) and the NDR family kinase Warts (Wts) (Lats1/Lats2 in mammals) to the transcription factor complex formed by the coactivator Yorkie (Yki) (YAP/TAZ in mammals) and its major DNA-binding partner Scalloped (Sd) (TEAD1/TEAD2/TEAD3/TEAD4 in mammals) (Figure 1A). The Sd-Yki transcription factor complex, in turn, regulates an array of target genes involved in cell proliferation and cell survival, such as the cell-death inhibitor *diap1*. Diverse upstream inputs into the core kinase cascade have been identified in *Drosophila*. These include an apical protein complex composed of the WW- and C2-domain-containing protein Kibra, and two FERM-domain containing proteins, Expanded (Ex) and Merlin (Mer); the Fat signaling module composed of the atypical cadherins Fat and its effectors, such as Four-jointed and Dachs; the apical-basal polarity regulators Crumbs (Crb), atypical Protein Kinase C (aPKC) and the WD40 scaffold protein Lethal giant larvae (Lgl) (Figure 1A). With the exception of Kibra and Mer, these upstream inputs have not been functionally linked to Hippo signaling in mammals (Zhao et al., 2010; Pan, 2010; Halder and Johnson, 2011; Zeng and Hong, 2008; Harvey and Tapon, 2007; Reddy and Irvine, 2008; Badouel et al., 2009).

Despite its essential role in animal development, the evolutionary history of the Hippo signaling pathway has been unresolved. Two recent comparative analyses reported key components of the pathway in nonbilaterian animals, but failed to identify any of these components outside Metazoa (Srivastava et al., 2010; Hilman and Gat, 2011). These two analyses

Cell PRESS
Open ACCESS

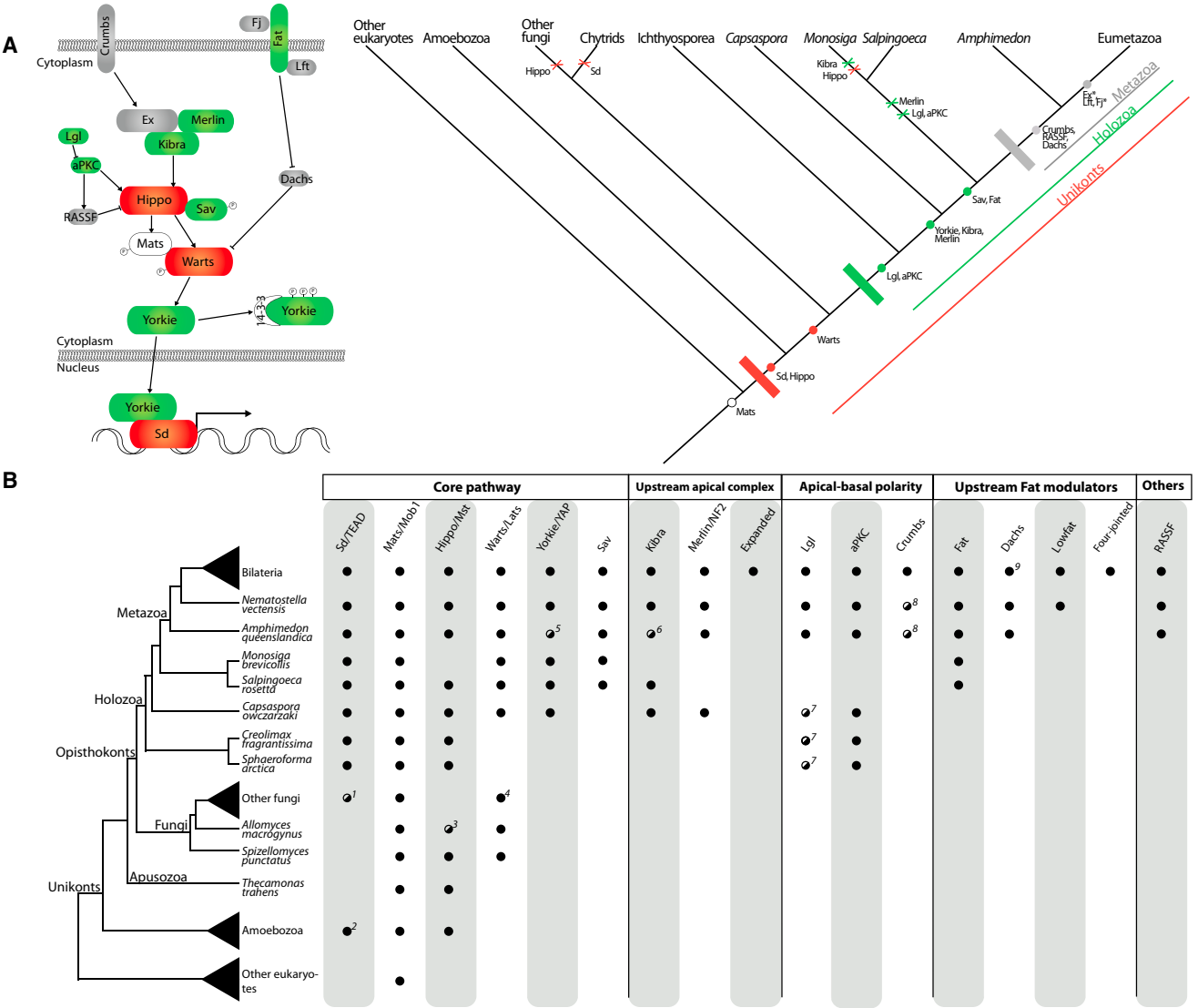


Figure 1. Evolution of the Hippo Signaling Pathway

(A) Schematic representation of the Hippo pathway evolution. The canonical metazoan Hippo pathway is shown on the left. The colors correspond to the three main steps in the evolution of the pathway, as shown in the cladogram (white, eukaryotes; red, unikonts; green, Holozoa; gray, Metazoa). Dots indicate origin and crosses indicate losses. Asterisks in Expanded (Ex) and Four-jointed (Fj) indicate that these proteins are exclusive to Bilateria.

(B) Schematic representation of the eukaryotic tree of life showing the distribution of the different components of the Hippo pathway. A black dot indicates the presence of clear orthologs, while a striped white-black dot indicates the presence of putative or degenerate orthologs. Absence of a dot indicates that an ortholog is lacking in that taxon. The taxon sampling for Bilateria includes *Homo sapiens, Drosophila melanogaster, Daphnia pulex* and *Capitella teleta*; other fungi includes the Ascomytoca *Neurospora crassa* and the Basidiomycota *Ustilago maydis*; Amoebozoa includes *Acanthamoeba castellanii* and *Dictyostelium discoideum*; other eukaryotes includes *Arabidopsis thaliana, Chlamydomonas reinhardtii, Naegleria gruberi, Trichomonas vaginalis, Thalassiosira pseudonana*, and *Tetrahymena thermophila*. Footnotes are as follows: [1] Fungi Sd orthologs do not have the C-terminal Y460 residue. [2] Sd/TEAD is present in the amoebozoan *A. castellanii* (whose ortholog includes the C-terminal Y460 residue), but not in *D. discoideum*. [3] A. macrogynus Hippo ortholog does not contain the SARAH domain. [4] *N. crassa* does not encode any ortholog of Warts/Lats, although other Ascomycota such as *Schizosaccharomyces pombe* and *Aspagillus niger* do encode this gene. [5] Putative *A. queenslandica* Yorkie ortholog contains just one, instead of two, WW protein domains. [6] Putative *A. queenslandica* Kibra ortholog contains an extra N-terminal PDZ domain. [7] *C. owczarzaki, C. fragrantissima* and *S. arctica* have proteins with the LLGL protein domain that in phylogenetic analysis appear as sister-group to a clade of the LLGL-containing Tomosyn and Lgl proteins. [8] Protein domain architecture is aberrant compared to bilaterian orthologs. [9] Absent in *H. sapiens*.

concluded that key components of the pathway were metazoan innovations and that the Hippo pathway originated in the last common ancestor of cnidarians and bilaterians (Hilman and Gat, 2011) or sometime within the early metazoan evolution (Srivastava et al., 2010). Through comparative genomic analysis of several recently sequenced holozoan genomes coupled with functional genetic characterization, we provide compelling evidence that an active Hippo signaling pathway was already

present in the unicellular ancestors of Metazoa, thus significantly pushing back the origin of this important cell-signaling mechanism.

RESULTS

Comparative Genomic Analysis Reveals a Premetazoan Origin of the Hippo Signaling Pathway

To trace the evolutionary origin of Hippo signaling, we performed an extensive search of pathway components in several recently sequenced holozoan genomes (Ruiz-Trillo et al., 2007), including ichthyosporeans, filastereans, and choanoflagellates, the closest unicellular relatives of Metazoa (Torruella et al., 2011), as well as in other eukaryotes. Our comparative genomic analysis reconstructed with unprecedented detail the evolutionary history of the Hippo signaling pathway and allowed us to trace the birth of Hippo signaling well before the origin of Metazoa (Figure 1A).

We have identified clear Yki orthologs in two independent nonmetazoan lineages, the filastereans (*Capsaspora owczarzaki*, hereafter called "*Capsaspora*") and the choanoflagellates (*Monosiga brevicollis* and *Salpingoeca rosetta*) (Figure 1B). Phylogenetic analysis clusters them unequivocally with metazoan Yki orthologs with high nodal support and well differentiated from the WWP1 and other ubiquitin ligases that also contain WW domains (Figure S1 available online). Importantly, all of these nonmetazoan Yki orthologs contain highly conserved functional sites such as the Hippo-pathway-responsive phosphorylation site S168/127 and the N-terminal homology region that is critical for interaction with the Sd/TEAD transcription factor (Figure 2A). Indeed, these holozoan species contain orthologs of Sd/TEAD with the C-terminal Y460 residue known to be important for YAP-TEAD interaction (Li et al., 2010; Chen et al., 2010).

Our searches further identified orthologs of Hpo, defined by the presence of a Ste20-like kinase domain and a SARAH domain (Scheel and Hofmann, 2003), in amoebozoans, apusozoans, and most opisthokonts, with the exception of *M. brevicollis* and nonchytrid fungi, most likely because of secondary loses (Figures 1A and 1B). Likewise, Wts orthologs are present along all opisthokonts, with the possible exception of the ichthyosporeans, for which the incompleteness of genome data makes it difficult to ensure its absence. The adaptor protein Mats is present in all eukaryotes, whereas the other adaptor of the core pathway, Salvador, is present only in choanoflagellates and Metazoa. Besides these critical components of the core Hippo kinase cascade, the amoeboid *Capsaspora* encodes several upstream regulators of Hpo, such as Kibra, Mer, aPKC, and Lgl but not Ex, Crb, or the Fat signaling module (Figures 1A and 1B).

Thus, our data have pinpointed with unprecedented detail the evolutionary history of all members of the Hippo pathway and show that a well-constituted Hippo pathway was present well before the origin of Metazoa, acting in a unicellular context. Moreover, given that *Capsaspora* encodes both Kibra and Mer and the apical-basal polarity proteins Lgl and aPKC, the level of upstream regulatory complexity of the Hippo/YAP pathway in *Capsaspora* is potentially very high. Although little is known about the receptors that lead to activation of the Hippo pathway in Metazoa, Hippo signaling is known to be activated in a cell-density-dependent manner (Zhao et al., 2010). In this regard, Mer has indeed been shown to directly mediate contact inhibition of proliferation in cell cultures (Okada et al., 2007; McClatchey and Giovannini, 2005), where Mer is known to engage reciprocal signaling with key effectors of the integrin signaling and adhesion machinery (Pugacheva et al., 2006). Interestingly, *Capsaspora* is so far the only analyzed nonmetazoan organism known to harbor all of the components of the integrin-mediated adhesion and signaling system present in metazoans (Sebé-Pedrós et al., 2010). Moreover, *Capsaspora* also encodes some genes known to be downstream of the Hippo pathway, such as Myc (Sebé-Pedrós et al., 2011) and cyclin E (data not shown). This suggests that the regulatory complexity of cell-proliferation control in the closest unicellular relatives of animals is remarkably high. Thus, a possible function of the Hippo pathway in this unicellular context could be the control of cell proliferation in a cell-density- and/or cell-adhesion-dependent manner.

The Sd-Yki Transcription Factor Complex from the Unicellular Amoeboid *Capsaspora Owczarzaki* Promotes Tissue Growth and Hippo Target Gene Expression in *Drosophila*

To test the functional relevance of our evolutionary analysis, we assayed the activities of *Capsaspora owczarzaki* (Co) Hippo pathway components in *Drosophila* (see Figures S2, S3, and S4 for sequence alignment of Yki, Sd, Hpo, Mats and Wts orthologs among *Capsaspora*, *Drosophila*, and humans). Given their critical roles in *Drosophila* and mammalian Hippo signaling, we first examined the *Capsaspora* orthologs of Sd and Yki (Figure 2A). We have shown previously that overexpression of *Drosophila melanogaster* (Dm) Yki by the GMR-Gal4 driver (GMR > Dm-Yki) leads to increased eye size (Huang et al., 2005) (Figure 2C), whereas overexpression of DmSd by the same Gal4 driver (GMR > Dm-Sd) results in smaller eye size (Figure 2D), most likely because of a dominant-negative effect whereby overexpressed Dm-Sd titrates (or squelches) certain endogenous Sd cofactor(s) (Wu et al., 2008). We found that overexpression of Co-Sd (GMR > Co-Sd) in *Drosophila* did not result in an appreciable change in eye size (Figure 2G), suggesting a reduced ability of Co-Sd to squelch endogenous Dm-Sd cofactors. Surprisingly, we found that unlike its *Drosophila* counterpart, overexpression of Co-Yki (GMR > Co-Yki) did not result in any tissue overgrowth, but rather caused a small and rough eye phenotype (Figure 2F). While the exact reason for this rough eye phenotype is unclear, the failure of Co-Yki overexpression to promote *Drosophila* eye growth suggests that Co-Yki has greatly diminished ability to interact productively with endogenous Dm-Sd to drive tissue overgrowth. Alternatively, Co-Yki may not possess intrinsic ability to drive tissue overgrowth (for example, due to its lack of general or specific coactivator activity to turn on growth-promoting genes), even if Co-Yki can form a transcription factor complex with Sd.

To distinguish between these models, we examined pairwise combinatorial overexpression between Co-Sd/Co-Yki and Dm-Sd/Dm-Yki. As shown previously, coexpression of Dm-Sd

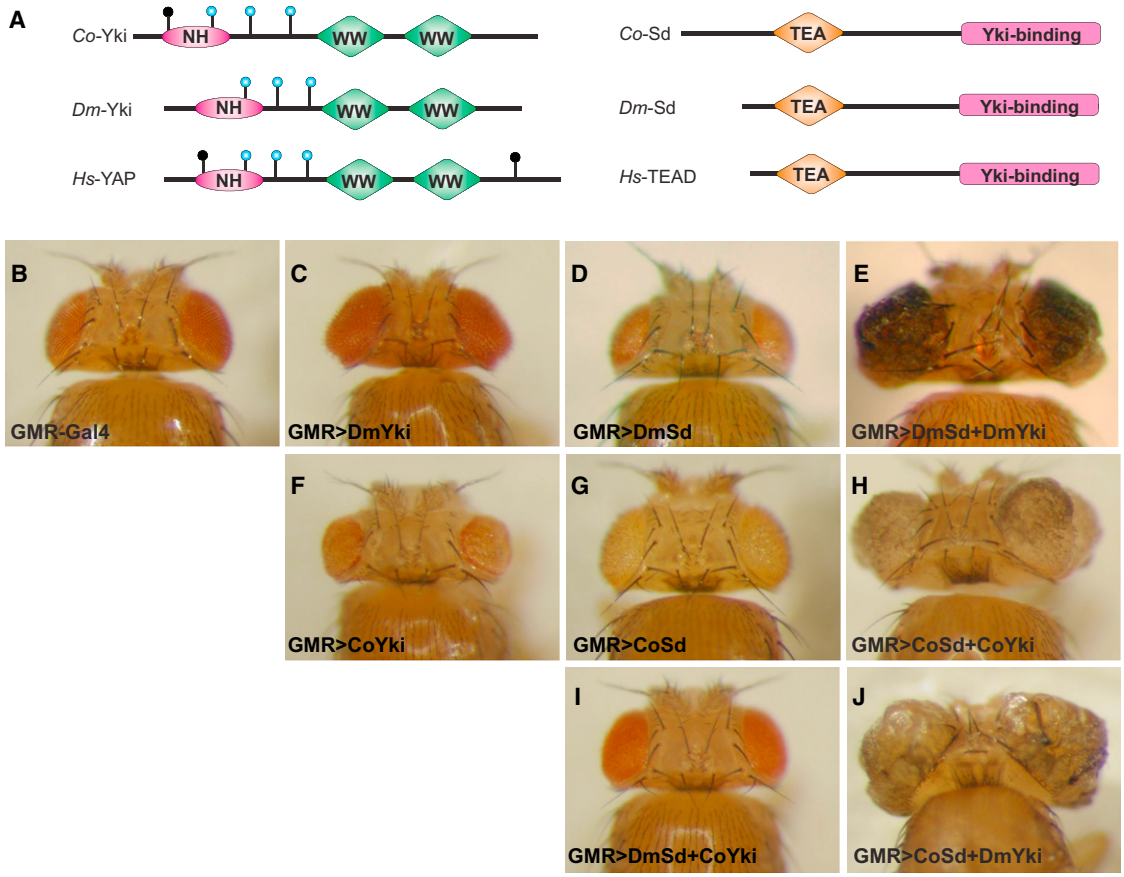


Figure 2. The Sd-Yki Transcription Factor Complex from *Capsaspora* Promotes Tissue Growth in *Drosophila*

(A) Schematic structures of Yki (left) and Sd (right) orthologs from *Capsaspora owczarzaki* (Co), *Drosophila melanogaster* (*Dm*), and *Homo sapiens* (*Hs*). Wts phosphorylation motifs (HxRxxS/T) in each Yki ortholog are indicated by vertical lines ending with circles, with the blue circles indicating the three conserved Wts phosphorylation motifs. "NH" refers to Yki's *N*-terminal *H*omology domain that binds to Sd/TEAD. "TEA" refers to the DNA-binding domain of the Sd orthologs.

(B–J) Dorsal view of adult heads from the indicated genotypes. All images were taken under the same magnification.

(B) GMR-Gal4/+. Wild-type control.

(C) GMR-Gal4 UAS-DmYki/+. Overexpression of DmYki resulted in an increase in eye size (compare C to B).

(D) GMR-Gal4/UAS-DmSd. Overexpression of DmSd caused a decrease in eye size (compare D to B).

(E) GMR-Gal4 UAS-DmYki/UAS-DmSd. The eye tissue was massively overgrown and folded.

(F) GMR-Gal4/UAS-CoYki. Overexpression of CoYki resulted in small and rough eyes (compare F to B).

(G) GMR-Gal4 UAS-CoSd/+. The eye size was similar to that of the wild-type control (compare G to B).

(H) GMR-Gal4 UAS-CoSd/UAS-CoYki. The eye tissue was massively overgrown and folded.

(I) GMR-Gal4 UAS-DmSd/UAS-CoYki. The eye size was similar to that of the wild-type control (compare I to B).

(J) GMR-Gal4 UAS-CoSd/UAS-DmYki. The eye tissue was massively overgrown and folded.

and Dm-Yki (GMR > Dm-Sd+Dm-Yki) resulted in tremendous overgrowth of eye tissue (Figure 2E), consistent with the well-established role of the Sd-Yki complex in promoting tissue growth. In agreement with the inability of Co-Yki alone to drive tissue overgrowth, coexpression of Co-Yki and Dm-Sd (GMR > Dm-Sd+Co-Yki) failed to drive eye overgrowth (Figure 2I). We noted that the GMR > Dm-Sd+Co-Yki eyes were larger than GMR > Dm-Sd or GMR > Co-Yki eyes, suggesting that when both proteins were overexpressed at high levels, they may interact with each other, albeit in a greatly attenuated manner. Most strikingly, despite the inability of Co-Yki or Co-Yki+Dm-Sd to induce tissue overgrowth, coexpression of Co-Yki and Co-Sd (GMR > Co-Sd+Co-Yki) resulted in massive tissue overgrowth resembling that caused by coexpression of their *Drosophila* counterparts (compare Figures 2H and 2E). A similar and massive tissue overgrowth was also observed when Co-Sd was co-expressed with Dm-Yki (GMR > Co-Sd+Dm-Yki) (Figure 2J). Thus, despite the greatly attenuated cross-species interactions between Co-Yki and Dm-Sd, the Sd-Yki complex evolves as a functional entity—it is the function of the Sd-Yki complex rather than the individual subunit that is pivotal to growth control.

To understand the molecular mechanism by which coexpression of Co-Sd and Co-Yki induces tissue overgrowth, we examined the expression of *diap1* and *ex*, two well-characterized Hippo/Yki target genes. Third instar eye imaginal discs of

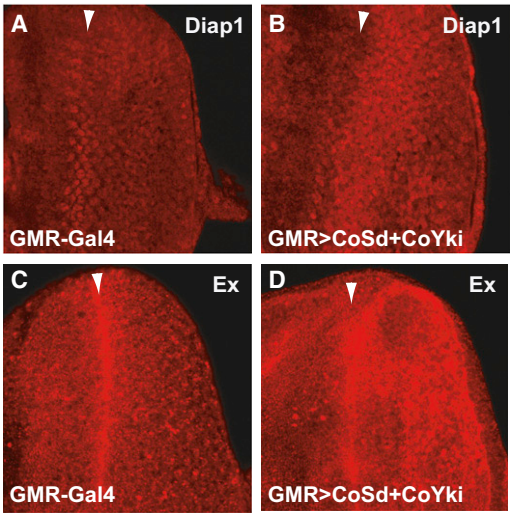


Figure 3. The Sd-Yki Transcription Factor Complex from *Capsaspora* Activates Hippo Target Genes in *Drosophila*

Confocal images of third instar eye imaginal discs from wild-type control (GMR-Gal4) (A and C) and animals with GMR-Gal4-mediated co-overexpression of Co-Sd and Co-Yki (GMR > CoSd+CoYki) (B and D). Arrowheads mark the position of the morphogenetic furrow (MF), and all eye discs are oriented anterior to the left.

(A and B) Eye imaginal discs showing Diap1 immunostaining (red). Note the elevated Diap1 expression posterior to the MF in GMR > CoSd+CoYki eye discs (compare B to A).

(C and D) Eye imaginal discs showing Ex immunostaining (red). Note the elevated Ex expression posterior to the MF in GMR > CoSd+CoYki eye discs (compare D to C).

GMR > Co-Sd+Co-Yki animals showed a marked upregulation of Diap1 and Ex staining posterior to the morphogenetic furrow (where the GMR-Gal4 driver is active) (Figure 3). Thus, despite their enormous evolutionary distance from each other, the Sd-Yki complex from a unicellular holozoan still retains the ability to promote tissue growth and to activate transcriptional targets similar to those of its *Drosophila* counterpart.

The Unicellular Amoeboid *Capsaspora Owczarzaki* Contains an Active Hippo Kinase Cascade Leading from Hpo to the Sd-Yki Complex

Our transgenic experiment predicted that Co-Sd and Co-Yki should physically interact with each other. Indeed, epitope-tagged Co-Sd and Co-Yki immunoprecipitated with each other in *Drosophila* S2R+ cells (Figure 4A), demonstrating their ability to form a protein complex. Using a well-characterized luciferase reporter driven by the minimal Hippo-Responsive Element (HRE) derived from the Hippo target gene *diap1* (Wu et al., 2008), we found that coexpression of Co-Sd and Co-Yki stimulated the transcription of the HRE-luciferase reporter in *Drosophila* S2R+ cells (Figure 4B). Together with the synergistic effect of Co-Sd and Co-Yki in inducing tissue overgrowth (Figure 2H) and Diap1 expression (Figures 3A and 3B) in vivo, these data demonstrate the ability of Co-Sd and Co-Yki to form a functional transcription factor complex with striking specificity to activate target genes similar to those of its *Drosophila* counterpart.

Next, we tested the functionality of Co-Hpo in inducing the phosphorylation of Co-Yki or Dm-Yki by coexpression of the respective constructs in *Drosophila* S2R+ cells. We found that Co-Hpo significantly inhibited Co-Sd/Co-Yki-mediated activation of the *diap1* HRE-luciferase reporter in S2R+ cells (Figure 4B), suggesting that Co-Hpo can negatively regulate the transcriptional activity of the Co-Yki/Co-Sd complex. Consistent with this finding, expression of Co-Hpo induced phosphorylation of Co-Yki in S2R+ cells, and this phosphorylation was further enhanced by coexpression of Dm-Wts (Figure 4C). Interestingly, Co-Hpo also stimulated the phosphorylation of Dm-Wts and Dm-Yki, as revealed by phospho-specific antibodies against P-Dm-Wts-T1077 and P-Dm-Yki-S168, respectively (Figure 4D). Thus, Co-Hpo can engage a canonical kinase cascade through the phosphorylation of the intermediary kinase Wts and the ultimate phosphorylation target Yki.

To corroborate the cell-based assays described above in a more physiological setting, we used a transgenic overexpression assay to examine the activity of Co-Hpo in vivo. Overexpression of Co-Hpo by the GMR-Gal4 driver (GMR > Co-Hpo) resulted in a small eye phenotype (Figures 4E and 4F) reminiscent of that caused by overexpression of its *Drosophila* counterpart, suggesting that the growth-inhibitory activity of Hpo is conserved in the unicellular *Capsaspora*. The GMR > Co-Hpo animals also allowed us to examine the influence of Co-Hpo on endogenous Yki phosphorylation in vivo. Using a phospho-specific antibody against the critical Hippo-responsive Ser168 phosphorylation site (Dong et al., 2007), we found that protein extracts from GMR > Co-Hpo fly heads showed increased Yki-S168 phosphorylation in comparison to control extracts (Figure 4G). Thus, Co-Hpo not only possesses growth-suppressing activity but also functionally activates a signaling cascade leading to the phosphorylation of endogenous Yki in *Drosophila*.

DISCUSSION

In conclusion, our study demonstrates that key components of the Hippo pathway are encoded in the genomes of unicellular relatives of metazoans. We provide compelling evidence that the amoeboid *Capsaspora* contains functional orthologs of the core Hippo kinase cascade leading from the tumor-suppressor protein Hpo to the transcriptional coactivator Yki, suggesting the existence of an active Hippo kinase cascade well before the origin of Metazoa. In particular, our data show that a well-constituted Hippo pathway originated within the Holozoa, before the divergence of filastereans, choanoflagellates, and Metazoa. Most remarkably, we demonstrate that despite the enormous evolutionary divergence, the growth-promoting and gene-regulatory activity and specificity of the Sd-Yki complex, as well as the growth-inhibitory activity of Hpo, are still retained in the unicellular *Capsaspora*. Our findings further pinpoint the Sd-Yki complex, rather than each subunit of this transcription factor complex, as a critical functional entity in the evolution of growth-control mechanisms.

The surprising conservation of biochemical functionality for different Hippo pathway elements in such phylogenetically divergent species (bilaterian metazoan versus amoeba) could probably be explained by strong functional constraints because of

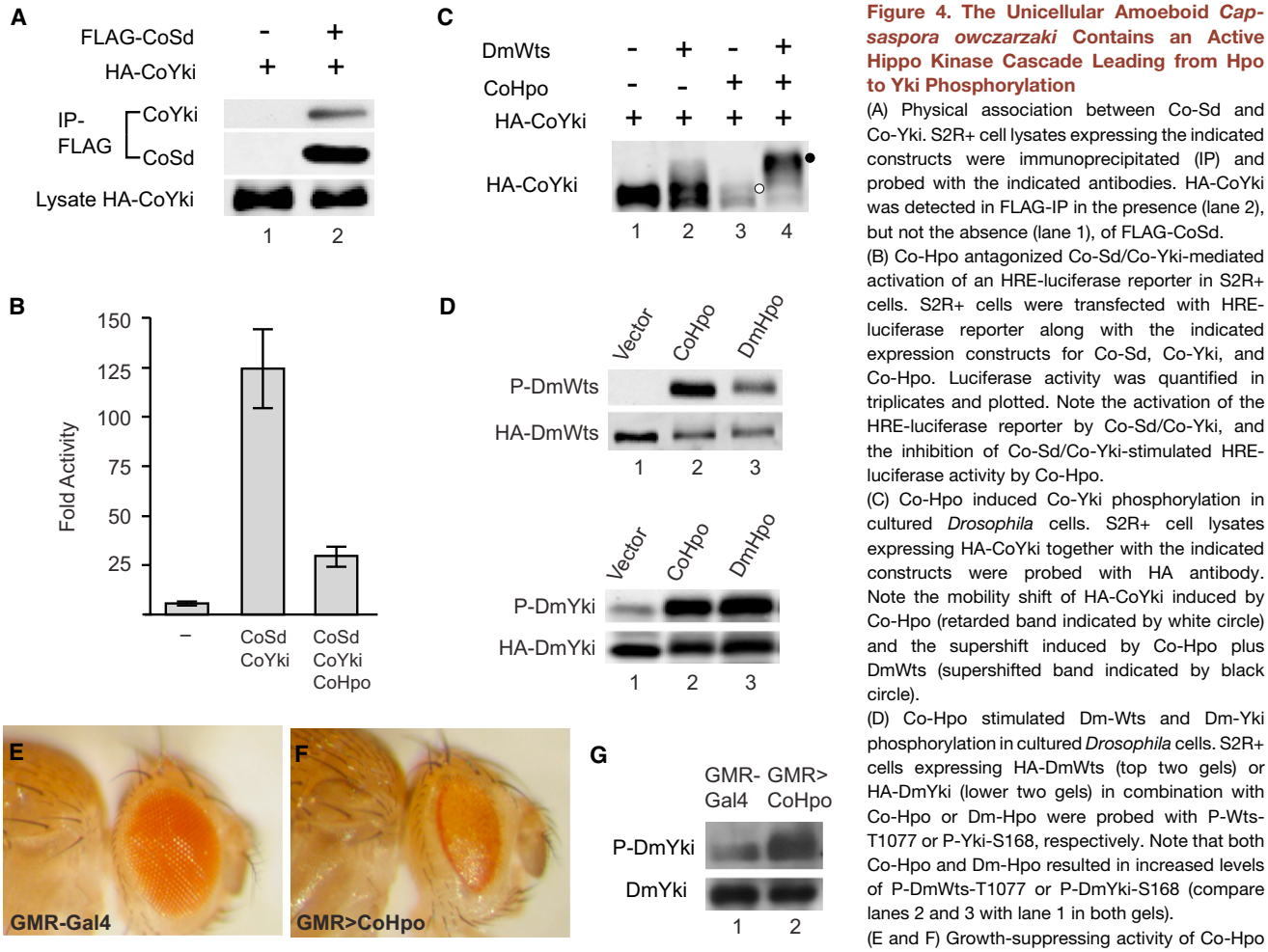


Figure 4. The Unicellular Amoeboid *Capsaspora owczarzaki* Contains an Active Hippo Kinase Cascade Leading from Hpo to Yki Phosphorylation

(A) Physical association between Co-Sd and Co-Yki. S2R+ cell lysates expressing the indicated constructs were immunoprecipitated (IP) and probed with the indicated antibodies. HA-CoYki was detected in FLAG-IP in the presence (lane 2), but not the absence (lane 1), of FLAG-CoSd.

(B) Co-Hpo antagonized Co-Sd/Co-Yki-mediated activation of an HRE-luciferase reporter in S2R+ cells. S2R+ cells were transfected with HRE-luciferase reporter along with the indicated expression constructs for Co-Sd, Co-Yki, and Co-Hpo. Luciferase activity was quantified in triplicates and plotted. Note the activation of the HRE-luciferase reporter by Co-Sd/Co-Yki, and the inhibition of Co-Sd/Co-Yki-stimulated HRE-luciferase activity by Co-Hpo.

(C) Co-Hpo induced Co-Yki phosphorylation in cultured *Drosophila* cells. S2R+ cell lysates expressing HA-CoYki together with the indicated constructs were probed with HA antibody. Note the mobility shift of HA-CoYki induced by Co-Hpo (retarded band indicated by white circle) and the supershift induced by Co-Hpo plus DmWts (supershifted band indicated by black circle).

(D) Co-Hpo stimulated Dm-Wts and Dm-Yki phosphorylation in cultured *Drosophila* cells. S2R+ cells expressing HA-DmWts (top two gels) or HA-DmYki (lower two gels) in combination with Co-Hpo or Dm-Hpo were probed with P-Wts-T1077 or P-Yki-S168, respectively. Note that both Co-Hpo and Dm-Hpo resulted in increased levels of P-DmWts-T1077 or P-DmYki-S168 (compare lanes 2 and 3 with lane 1 in both gels).

(E and F) Growth-suppressing activity of Co-Hpo in *Drosophila*. Side views of adult heads of control (GMR-Gal4/+) (E) and flies that overexpressed Co-Hpo in the eye (GMR-Gal4/UAS-CoHpo) (F). Note the reduced eye size of GMR > CoHpo flies (compare F to E).

(G) Overexpression of Co-Hpo stimulated phosphorylation of endogenous Yki in *Drosophila*. Protein extracts from control (GMR-Gal4) or GMR > CoHpo adult heads were probed with antibodies against endogenous DmYki and P-DmYki-S168. Note the increase in P-DmYki signal in GMR > CoHpo adult head extracts (compare lane 2 to lane 1).

the varied network of interactions of these components. In theory, these different network elements could have independently coevolved and, therefore, not be functional within another species context. The fact that the different Hippo pathway elements of *Capsaspora* are indeed functional within the multicellular *Drosophila* background strongly supports a functional homology between the unicellular and multicellular Hippo pathway. This suggests that the current function of the Hippo signaling pathway might be somehow similar within these two biological contexts.

How the ancestral Hippo kinase cascade is used in a unicellular organism such as *Capsaspora* remains a mystery at present. We speculate that this pathway might be used to coordinate cell proliferation in response to cell density or cell polarity (e.g., upon substrate adhesion), given the established roles of Merlin, Kibra, aPKC, and Lgl (all of which are encoded in *Capsaspora* genome) in these biological processes. The absence of other developmental signaling pathways, such as Notch, Hedgehog, Wnt, or BMP, in *Capsaspora* or any other analyzed unicellular holozoan emphasizes the relevance of Hippo signaling as a key developmental mechanism predating the origin of Metazoa. The exaptation of this pathway may have easily provided a mechanism for strict control of cell proliferation in early metazoans, an essential property for any integrated multicellular entity. The presence of a functional and highly conserved Hippo pathway in *Capsaspora* accentuates not only the importance of analyzing the unicellular prehistory of animals to understand their origin, but also the role that gene co-option may have played in the unicellular-to-multicellular transition.

EXPERIMENTAL PROCEDURES

Gene Searches and Phylogenetic Analysis

Genes were searched by with the use of BLAST (blastp, blastn, and tblastn) and BLAST reverse, with several sequences used as queries, as described

previously (Sebé-Pedrós et al., 2010). Some genes could be identified only through phylogenetic analyses. A full list of the genes and sequences used, including the newly annotated ones, is shown in Table S1. The taxon sampling used is shown in Figure 1B and includes new genome sequences obtained by the UNICORN genome project (see http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html) (Ruiz-Trillo et al., 2007). Domain arrangements were confirmed by Pfam and SMART. Alignments were constructed with the MAFFT online server (http://mafft.cbrc.jp/alignment/server/) (Katoh et al., 2002) and then manually inspected and edited in Geneious. Only those species and those positions that were unambiguously aligned were included in the final analysis. Maximum likelihood (ML) phylogenetic trees were estimated by Raxml (Stamatakis, 2006) through the use of the PROTGAMMAIWAG+ Γ+I model of evolution. Nodal supports were assessed by the performance of 100-bootstrap replicates with the same evolutionary model. Bayesian analysis was performed with MrBayes 3.1 (Ronquist and Huelsenbeck, 2003) through the use of the WAG+Γ+I model of evolution, four chains, and two parallel runs. Runs were stopped when the average SD of split frequencies of the two parallel runs was < 0.01, usually around 1,000,000 generations, and burn-in length was established by checking of the two LnL graphs; stationarity of the chain typically occurred after ~15% of the generations.

The annotation of *Capsaspora* Yorkie (Co-Yki) was further checked by the use of 5′ and 3′ rapid amplification of cDNA ends (RACE) PCR under standard conditions. An Excel file with all the sequences of the genes used in this study can be downloaded from http://www.multicellgenome.com.

Drosophila Genetics and Cell Culture

Capsaspora Hpo, Yki, and Sd cDNAs were amplified by PCR and inserted into pUAST vector to generate pUAST-CoHpo, pUAST-CoYki, and pUAST-CoSd constructs, respectively. Transgenic flies were made by P-element-mediated germline transformation pUAST constructs. Flies were raised on standard cornmeal medium at 25°C. Eye imaginal discs of wandering third instar larvae were fixed and stained as previously described (Yu et al., 2010) with the use of α-Ex (1:5000) (gift of R.G. Fehon) (Maitra et al., 2006) and α-Diap1 (1:600) (gift of B. A. Hay) (Yoo et al., 2002). For analysis of endogenous Yki phosphorylation in fly tissues, 20 fly heads from control (GMR-Gal4) or Co-Hpo-transgenic (GMR > CoHpo) animals were smashed in 50 μl 2× SDS loading buffer and then boiled for 5 min. After centrifugation at 12,000 rpm for 5 min, 10 μl supernatants were separated on 8% SDS-PAGE and transferred to an Immobilon-P PVDF membrane. The Western blots were probed with rabbit α-P-S168-Yki and rabbit α-Yki antibodies (Dong et al., 2007).

Drosophila S2R+ cells were propagated in Schneider's medium (GIBCO) supplemented with 10% FBS and antibiotics. Luciferase assay was carried out with the use of the HRE-luciferase reporter as described previously (Wu et al., 2008), with the use of pUAST-CoHpo, pUAST-CoYki, and pUAST-CoSd constructs in combination with pAc-Gal4 construct. Expression constructs for HA-DmWts and HA-DmYki have been described previously (Ling et al., 2010; Huang et al., 2005). Yki and Wts phosphorylation was probed with the use of rabbit α-P-S168-Yki (Dong et al., 2007) and rabbit α-P-Wts-T1077 (Yu et al., 2010).

For immunoprecipitation, HA-CoYki and FLAG-CoSd were constructed in the pAc5.1/V5-HisB vector by the addition of the respective epitope at the N terminus of each protein. S2R+ cells transiently transfected with these constructs were lysed in lysis buffer (50 mM Tris [Ph7.4], 150 mM NaCl, 1 mM EDTA, 0.5% Triton X-100) supplemented with protease inhibitor cocktail (Roche) and phosphatase inhibitor cocktail (20 μM sodium fluoride, 4 mM sodium orthovanadate, 4 mM sodium pyrophosphate, 12 mM β-glycerophosphate). Lysate was cleared by centrifugation at 14,000 rpm for 5 min. Supernatant was incubated with ANTI-FLAG M2 Affinity Gel (Sigma-Aldrich) at 4°C for 2 hr, followed by centrifugation and washing as described previously (Yu et al., 2010).

ACCESSION NUMBERS

The GenBank accession number for the *Capsaspora* Yorkie sequence reported in this paper is JN202490.

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and one table and can be found with this article online at doi:10.1016/j.celrep.2011.11.004.

LICENSING INFORMATION

ACKNOWLEDGMENTS

The genome sequences of *C. owczarzaki*, *S. rosetta*, *A. macrogynus*, *S. punctatus*, and *T. trahens* are being determined by the Broad Institute of MIT/Harvard University under the auspices of the National Human Genome Research Institute (NHGRI) and within the UNICORN initiative. We thank Joint Genome Institute, Broad Institute, and Baylor College of Medicine for making data publicly available. We also thank Dr. Kim Worley and her colleagues in the Human Genome Sequencing Center of Baylor College of Medicine for allowing us to use the *A. castellanii* genome sequence. We are grateful to Drs. Rick Fehon and Bruce Hay for providing antibodies used in this study. This work was supported by an Institució Catalana per a la Recerca i Estudis Avançats contract, a European Research Council Starting Grant (ERC-2007-StG-206883), a grant (BFU2008-02839/BMC) from Ministerio de Ciencia e Innovación (MICINN) to I. R.-T., and a National Institutes of Health grant (R01 EY015708) to D.P. A.S-P. was supported by a pregraduate Formacion Profesorado Universitario grant from MICINN. D.P. is an investigator of the Howard Hughes Medical Institute.

Received: September 23, 2011
Revised: November 7, 2011
Accepted: November 18, 2011
Published online: December 15, 2011

REFERENCES

Adamska, M., Matus, D.Q., Adamski, M., Green, K., Rokhsar, D.S., Martindale, M.Q., and Degnan, B.M. (2007). The evolutionary origin of hedgehog proteins. Curr. Biol. *17*, R836–R837.

Badouel, C., Garg, A., and McNeill, H. (2009). Herding Hippos: regulating growth in flies and man. Curr. Opin. Cell Biol. *21*, 837–843.

Chen, L., Chan, S.W., Zhang, X., Walsh, M., Lim, C.J., Hong, W., and Song, H. (2010). Structural basis of YAP recognition by TEAD4 in the hippo pathway. Genes Dev. *24*, 290–300.

Dong, J., Feldmann, G., Huang, J., Wu, S., Zhang, N., Comerford, S.A., Gayyed, M.F., Anders, R.A., Maitra, A., and Pan, D. (2007). Elucidation of a universal size-control mechanism in Drosophila and mammals. Cell *130*, 1120–1133.

Halder, G., and Johnson, R.L. (2011). Hippo signaling: growth control and beyond. Development *138*, 9–22.

Harvey, K., and Tapon, N. (2007). The Salvador-Warts-Hippo pathway - an emerging tumour-suppressor network. Nat. Rev. Cancer *7*, 182–191.

Hilman, D., and Gat, U. (2011). The evolutionary history of YAP and the hippo/YAP pathway. Mol. Biol. Evol. *28*, 2403–2417.

Huang, J., Wu, S., Barrera, J., Matthews, K., and Pan, D. (2005). The Hippo signaling pathway coordinately regulates cell proliferation and apoptosis by inactivating Yorkie, the Drosophila Homolog of YAP. Cell *122*, 421–434.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066.

King, N., Hittinger, C.T., and Carroll, S.B. (2003). Evolution of key cell signaling and adhesion protein families predates animal origins. Science *301*, 361–363.

King, N., Westbrook, M.J., Young, S.L., 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.

Li, Z., Zhao, B., Wang, P., Chen, F., Dong, Z., Yang, H., Guan, K.L., and Xu, Y. (2010). Structural insights into the YAP and TEAD complex. Genes Dev. *24*, 235–240.

Ling, C., Zheng, Y., Yin, F., Yu, J., Huang, J., Hong, Y., Wu, S., and Pan, D. (2010). The apical transmembrane protein Crumbs functions as a tumor suppressor that regulates Hippo signaling by binding to Expanded. Proc. Natl. Acad. Sci. USA *107*, 10532–10537.

Maitra, S., Kulikauskas, R.M., Gavilan, H., and Fehon, R.G. (2006). The tumor suppressors Merlin and Expanded function cooperatively to modulate receptor endocytosis and signaling. Curr. Biol. *16*, 702–709.

Manning, G., Young, S.L., Miller, W.T., and 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. USA *105*, 9674–9679.

McClatchey, A.I., and Giovannini, M. (2005). Membrane organization and tumorigenesis–the NF2 tumor suppressor, Merlin. Genes Dev. *19*, 2265–2277.

Okada, T., You, L., and Giancotti, F.G. (2007). Shedding light on Merlin's wizardry. Trends Cell Biol. *17*, 222–229.

Pan, D. (2010). The hippo signaling pathway in development and cancer. Dev. Cell *19*, 491–505.

Pires-daSilva, A., and Sommer, R.J. (2003). The evolution of signalling pathways in animal development. Nat. Rev. Genet. *4*, 39–49.

Pugacheva, E.N., Roegiers, F., and Golemis, E.A. (2006). Interdependence of cell attachment and cell cycle signaling. Curr. Opin. Cell Biol. *18*, 507–515.

Reddy, B.V., and Irvine, K.D. (2008). The Fat and Warts signaling pathways: new insights into their regulation, mechanism and conservation. Development *135*, 2827–2838.

Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572–1574.

Ruiz-Trillo, I., Burger, G., Holland, P.W., King, N., Lang, B.F., Roger, A.J., and Gray, M.W. (2007). The origins of multicellularity: a multi-taxon genome initiative. Trends Genet. *23*, 113–118.

Scheel, H., and Hofmann, K. (2003). A novel interaction motif, SARAH, connects three classes of tumor suppressor. Curr. Biol. *13*, R899–R900.

Sebé-Pedrós, A., Roger, A.J., Lang, F.B., King, N., and Ruiz-Trillo, I. (2010). Ancient origin of the integrin-mediated adhesion and signaling machinery. Proc. Natl. Acad. Sci. USA *107*, 10142–10147.

Sebé-Pedrós, A., de Mendoza, A., Lang, B.F., Degnan, B.M., and Ruiz-Trillo, I. (2011). Unexpected repertoire of metazoan transcription factors in the unicellular holozoan Capsaspora owczarzaki. Mol. Biol. Evol. *28*, 1241–1254.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E., Mitros, T., Richards, G.S., Conaco, C., Dacre, M., Hellsten, U., et al. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. Nature *466*, 720–726.

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

Torruella, G., Derelle, R., Paps, J., Lang, B.F., Roger, A.J., Shalchian-Tabrizi, K., and Ruiz-Trillo, I. (2011). Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single copy protein domains. Mol. Biol. Evol. Published online July 28, 2011.

Wu, S., Liu, Y., Zheng, Y., Dong, J., and Pan, D. (2008). The TEAD/TEF family protein Scalloped mediates transcriptional output of the Hippo growth-regulatory pathway. Dev. Cell *14*, 388–398.

Yoo, S.J., Huh, J.R., Muro, I., Yu, H., Wang, L., Wang, S.L., Feldman, R.M., Clem, R.J., Müller, H.A., and Hay, B.A. (2002). Hid, Rpr and Grim negatively regulate DIAP1 levels through distinct mechanisms. Nat. Cell Biol. *4*, 416–424.

Yu, J., Zheng, Y., Dong, J., Klusza, S., Deng, W.M., and Pan, D. (2010). Kibra functions as a tumor suppressor protein that regulates Hippo signaling in conjunction with Merlin and Expanded. Dev. Cell *18*, 288–299.

Zeng, Q., and Hong, W. (2008). The emerging role of the hippo pathway in cell contact inhibition, organ size control, and cancer development in mammals. Cancer Cell *13*, 188–192.

Zhao, B., Li, L., Lei, Q., and Guan, K.L. (2010). The Hippo-YAP pathway in organ size control and tumorigenesis: an updated version. Genes Dev. *24*, 862–874.

Cell PRESS
Open ACCESS

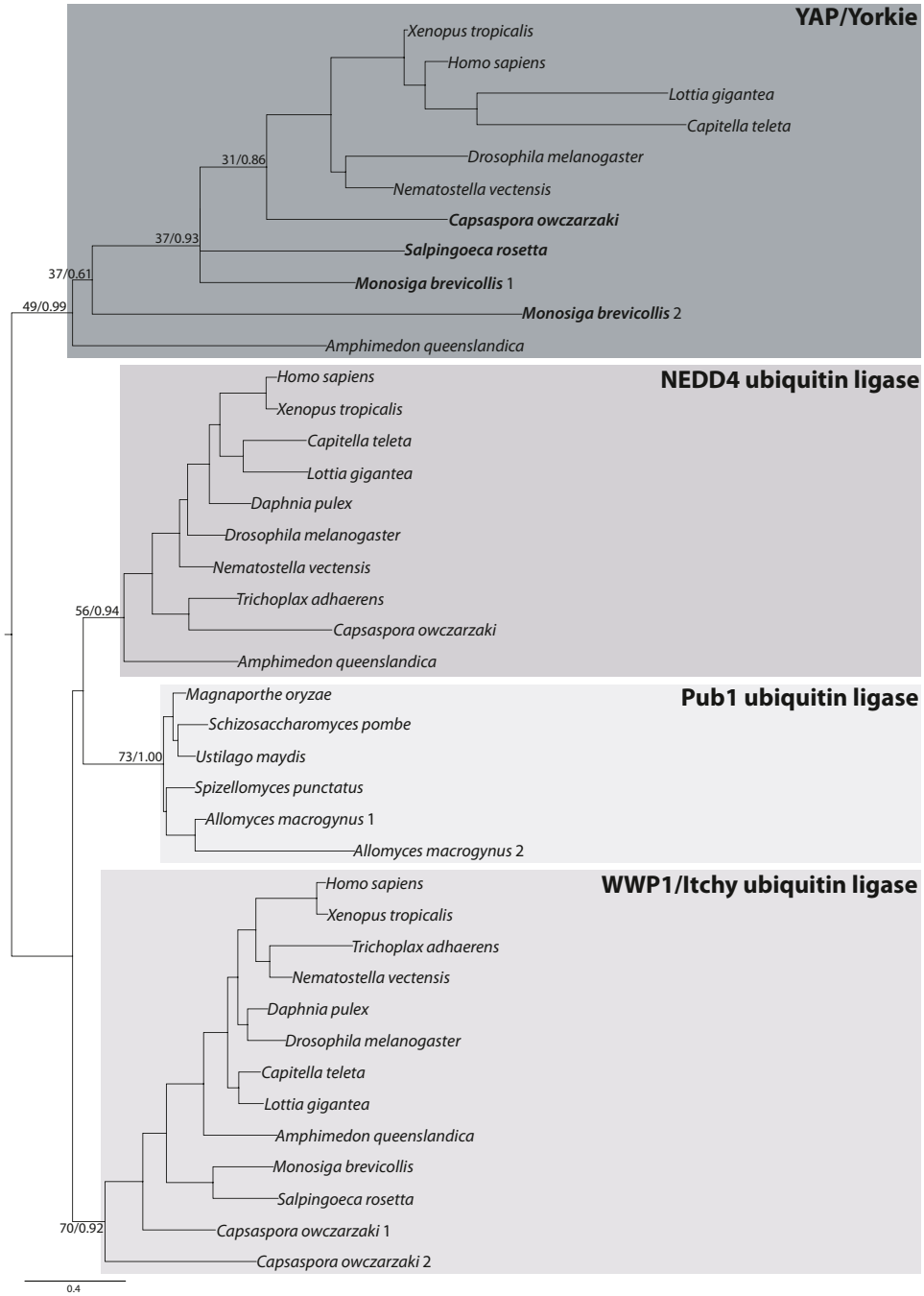


Figure S1. Maximum Likelihood Tree of the Yorkie/YAP Orthologs and Other Proteins Containing WW Domains; Related to Figure 1

Alignment has been done using the only common region among these proteins, which are two consecutive WW domains, except for the putative Yorkie ortholog of *A. queenslandica* and a second putative Yorkie ortholog in *M. brevicollis* (*M. brevicollis-2*), both of which have only one WW domain. The tree is rooted using the midpoint-rooted tree option. Statistical support was obtained by RAxML with 100-bootstrap replicates (bootstrap value, BV) and Bayesian Posterior Probabilities (BPP). Both values are only shown for some external key branches. The general topology is the same for Bayesian and ML analyses. Note that the Yorkie orthologs in *Capsaspora*, *Salpingoeca* and *Monosiga* are clearly clustered with Yorkie orthologs in Metazoa. The clustering of the one-WW-domain *A. queenslandica* and *M. brevicollis-2* Yorkie orthologs with the two-WW-domain Yorkie orthologs suggests that the former are putative or degenerate orthologs.

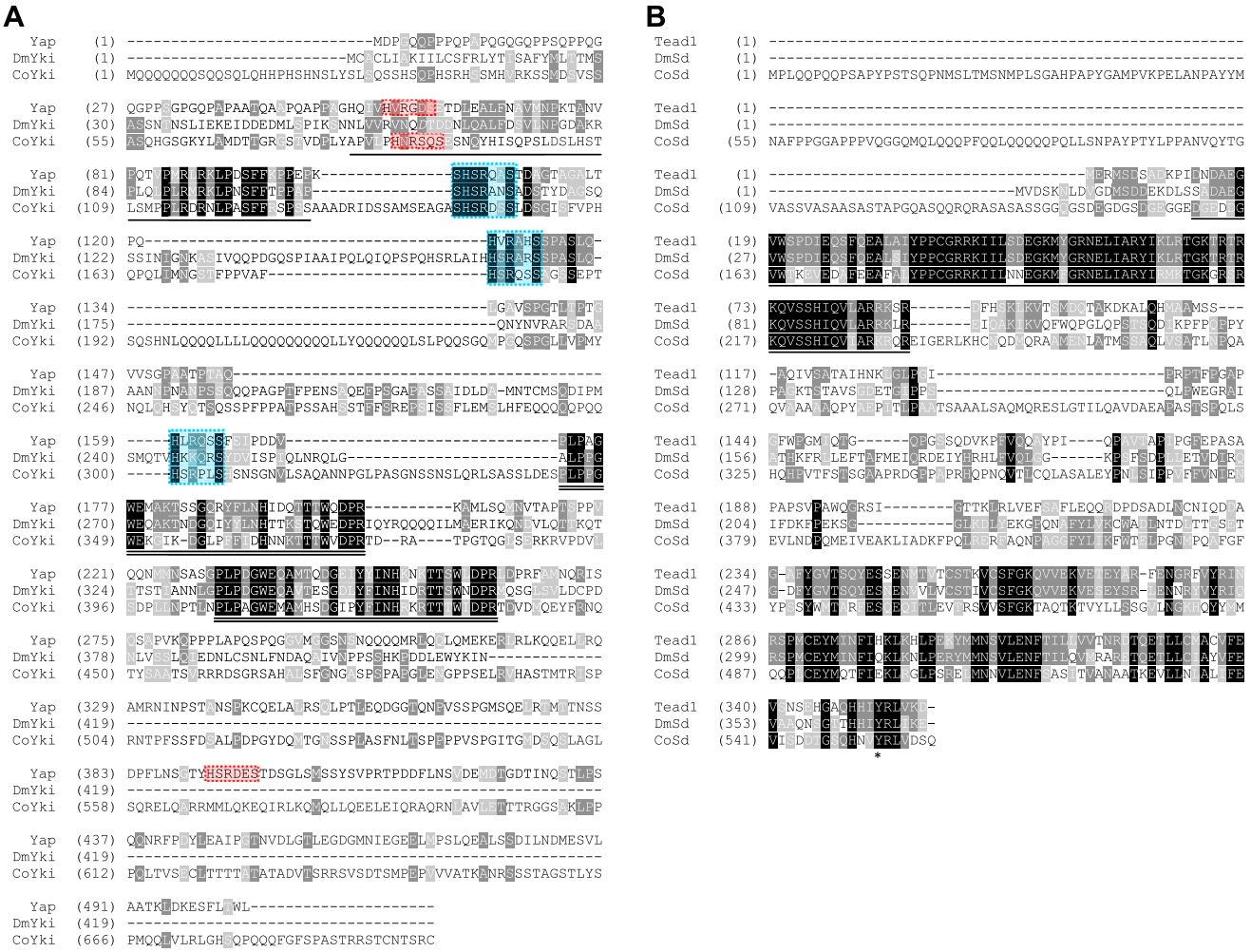


Figure S2. Schematic Sequence Alignment of Hippo Pathway Components Yki/Yap and Sd/Tead; Related to Figure 1

The sequence alignment is from *Homo sapiens* (top), *Drosophila melanogaster* (middle) and *Capsaspora owczarzaki* (bottom).

(A) Alignment of Yki/Yap. The Sd-binding domain is underlined and the two WW domains are double underlined. Three conserved Wts phosphorylation motifs are marked by blue box and non-conserved phosphorylation motifs are marked by red box. Note that Yap, DmYki and CoYki have 5, 4 and 3 Wts phosphorylation motifs, respectively.

(B) Alignment of Sd/Tead. The DNA-binding TEA domain is underlined, and the conserved Tyr residue crucial for Yki/Sd binding is marked by an asterisk.

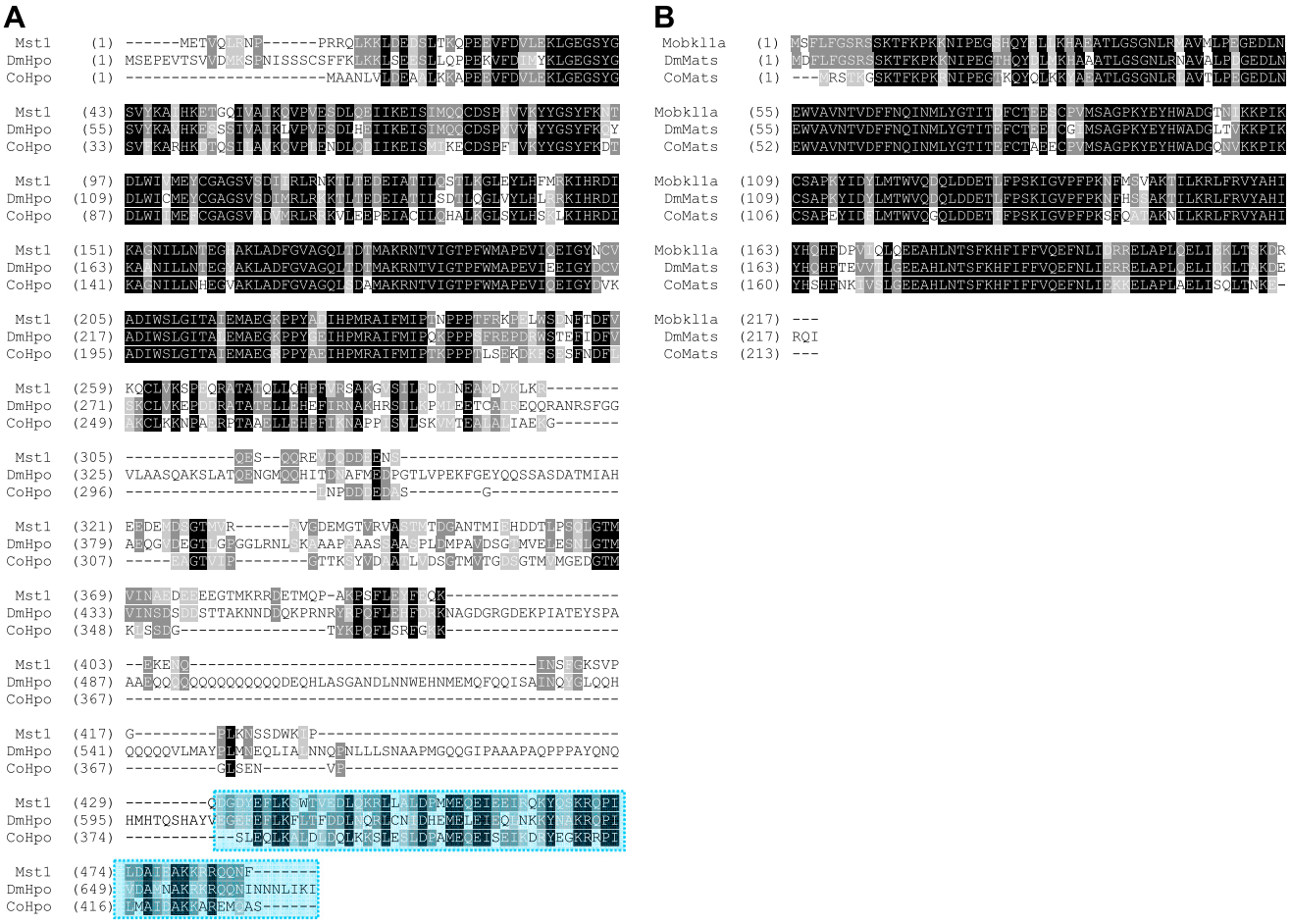


Figure S3. Schematic Sequence Alignment of Hippo Pathway Components Hpo/Mst and Mats/Mobkl1; Related to Figure 1

The sequence alignment is from *Homo sapiens* (top), *Drosophila melanogaster* (middle) and *Capsaspora owczarzaki* (bottom).

(A) Alignment of Hpo/Mst. The SARAH domain is marked by blue box.

(B) Alignment of Mats/Mobkl1.

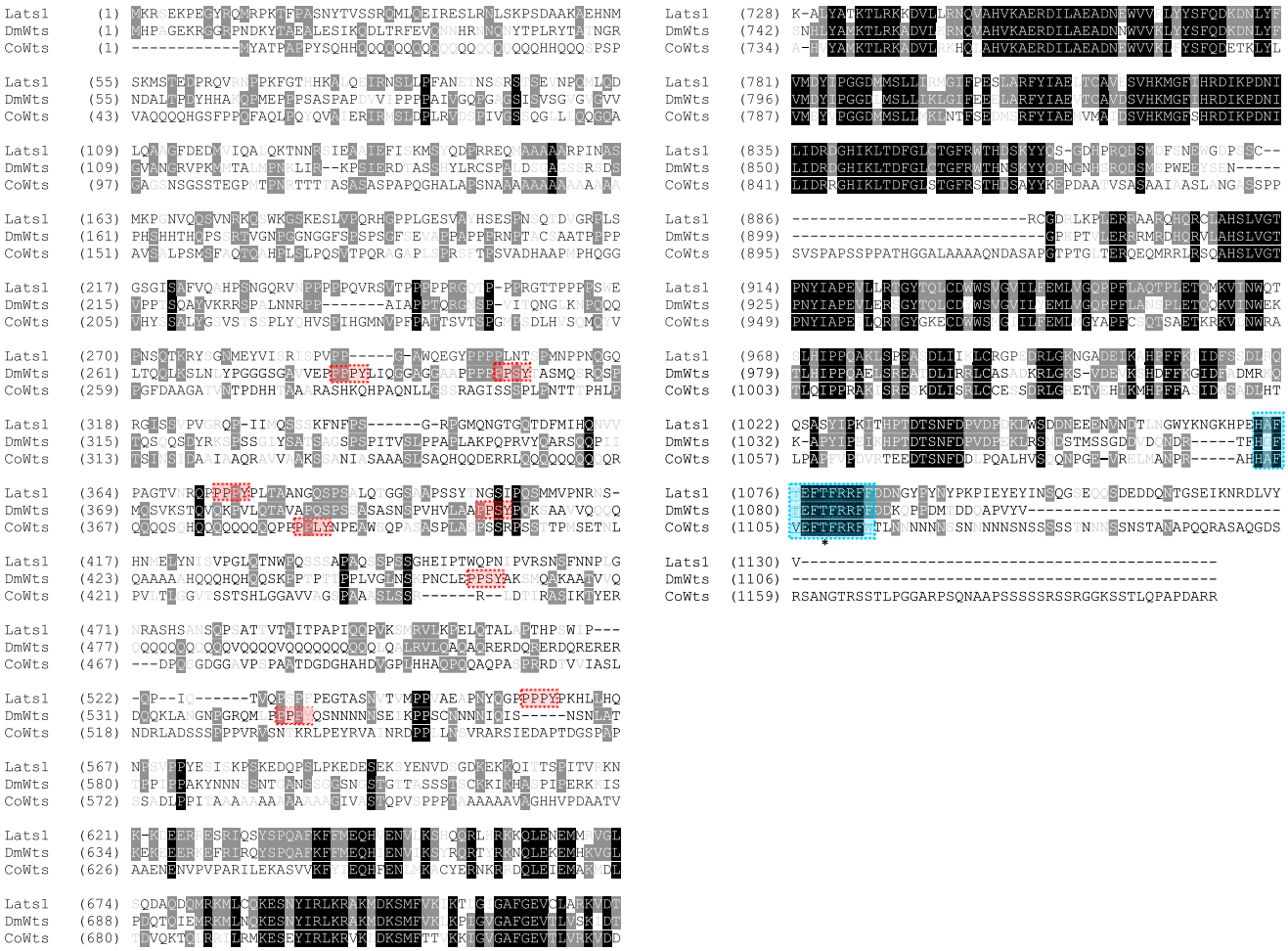


Figure S4. Schematic Sequence Alignment of Hippo Pathway Component Wts/Lats; Related to Figure 1

The sequence alignment is from *Homo sapiens* (top), *Drosophila melanogaster* (middle) and *Capsaspora owczarzaki* (bottom). PPXY motifs are marked by red box, and the hydrophobic motif is marked by blue box with the Hpo phosphorylation site indicated by an asterisk. Note that Lats1, DmWts and CoWts have 2, 5 and 1 PPXY motifs, respectively.

Results R6

Insights into the origin of metazoan filopodia and microvilli.

RESUM ARTICLE R6: Perspectives sobre l'origen dels fil·lopodis i microvil·lis a metazous

Els fil·lopodis, fines projeccions cel·lulars fetes de microfilaments d'actina i que s'usen en la motilitat cel·lular i el sondeig de l'ambient, són orgànuls essencials per a les cèl·lules animals. En aquest estudi, reconstruïm l'origen dels fil·lopodis i microvil·lis animals. En primer lloc, analitzem l'assemblatge evolutiu de la maquinària molecular del fil·lopodi i mostrem que gens homòlegs de molts dels components del fil·lopodi a animals, incloent fascina i miosina X, ja eren presents en els progenitors unicel·lulars o colonials dels animals. A més a més, trobem que la proteïna específica de fil·lopodi fascina localitza als fil·lopodi i estructures relacionades, anomenades microvil·lis, en el coanoflagel·lat *Salpingoeca rosetta*. Per últim, homòlegs dels gens fil·lopodials en l'holozou *Capsaspora owczarzaki* són sobre-expressats en cèl·lules fil·lopodiades, en comparació amb cèl·lules sense fil·lopodis. Per tant, els nostres resultats suggereixen que proteïnes essencials per als fil·lopodis animals estan funcionalment conservades en holozous unicel·lulars i colonials i que l'ancestre comú dels animals tenia una complexa i específica maquinària fil·lopodial.

Insights into the origin of metazoan filopodia and microvilli

Arnau Sebé-Pedrós[1,4*], Pawel Burkhardt[2*], Núria Sánchez-Pons[1], B. Franz Lang[3], Nicole King[2#] & Iñaki Ruiz-Trillo[1,4,5#]

[1] Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Catalonia, Spain. [2] Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720 [3] Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, 2900 Boulevard Edouard Montpetit, Montréal (Québec) H3C 3J7, Canada [4] Departament de Genètica, Universitat de Barcelona, Avinguda Diagonal 643, 08028 Barcelona, Spain [5] Institució Catalana per a la Recerca i Estudis Avançats (ICREA), Barcelona, Spain

*These authors contributed equally to this work. #corresponding authors

Filopodia, fine actin-based cellular projections used for both cell motility and environmental sensing, are essential organelles for metazoan cells. In this study we reconstruct the origin of metazoan filopodia and microvilli. We first report on the evolutionary assembly of the filopodial molecular toolkit and show that homologs of many metazoan filopodial components, including fascin and myosin X, were already present in the unicellular or colonial progenitors of metazoans. Furthermore, we find that the filopodia actin-cross-linking protein fascin localizes to filopodia and related cellular structures, called microvilli, in the choanoflagellate *Salpingoeca rosetta*. In addition, homologs of filopodial genes in the holozoan *Capsaspora owczarzaki* are upregulated in filopodia-bearing cells relative to those that lack them. Therefore, our findings suggest that proteins essential for metazoan filopodia are functionally conserved in unicellular and colonial holozoans and that the last common ancestor of metazoans bore a complex and specific filopodial machinery.

Myosin evolution – Fascin – choanoflagellate – Capsaspora – Cdc42 – Formin evolution – Gelsolin evolution – microvilli - filopodia

Introduction

A dynamic cytoskeletal and membrane system is a hallmark of the eukaryotic cell. It allows cells to change cell shape to carry out motility, phagocytosis, and other key functions (Fletcher and Mullins 2010). Cell motility, in particular, is a common feature among eukaryotes that often requires specialized organelles. There are two main classes of cellular structures responsible for cell motility in eukaryotes: tubulin-based cilia and flagella, conspicuous in eukaryotes as diverse as choanoflagellates, ciliates and dinoflagellates, and actin-based filopodia and lamellipodia, which allow cells to crawl along a surface through amoeboid movement (Soldati and Meissner 2004).

Filopodia are typically based upon 10 to 30 parallel bundled actin filaments, whose growing/barbed ends orient toward the filopodial tip. Here, many proteins accumulate and form the so-called tip complex, which controls monomer addition to filament ends (Small et al. 2002; Bohil et al. 2006; Faix and Rottner 2006; Gupton and Gertler 2007; Mattila and Lappalainen 2008; Lundquist 2009; Mellor 2010; Nambiar et al. 2010). In addition to their roles in cell motility, filopodia also contribute to cell adhesion and sensing of the extracellular milieu (Schäfer et al. 2010). In metazoans, filopodia mediate many essential, metazoan-specific phenomena, including growth cone guidance, wound-healing, embryonic development, angiogenesis and they serve as precursors for dendritic spines in neurons (Magie et al. 2007; Mattila and Lappalainen 2008; Mellor 2010).

Filopodia-like thin actin-based cellular protrusions (we use the term filopodia here to refer to them all) are known not only in cells of metazoans, but also in cells from diverse other eukaryotic lineages. For example, among bikonts filopodia-like protrusions are found in excavates (e.g. *Naegleria gruberi* (Preston and King 2005)), stramenopiles (Pawlowski 2008; Brown et al. 2012), and rhizarians (Cavalier-Smith 2003; Pawlowski 2008; Ota et al. 2011), in which filopodia form one of the defining morphological characteristics of the group (Brown et al. 2012). However, filopodia are most abundant and diversified in unikonts, in which filopodial-like structures have been reported in amoebozoans, apusozoans (Cavalier-Smith and Chao 2010) and several opisthokont lineages, including nucleariids (the sister group of fungi) (Mikrjukov and Mylnikov 2001; Zettler et al. 2001), filastereans (Cavalier-Smith 2003), choanoflagellates (Leadbeater and Morton 1974) and metazoans (see Fig. 1 for their phylogenetic relationships). Moreover, microvilli, which are another fine actin-based cell protrusions, are present

in holozoan lineages; a good example is the apical collar of microvilli in sponge choanocytes (Gonobobleva and Maldonado 2009) and in choanoflagellates (Karpov and Leadbeater 1998; Leadbeater et al. 2009).

Two possible scenarios may account for the patchy distribution of filopodia in the eukaryotic tree. One possibility is that filopodial structures and their specific molecular components evolved independently several times during eukaryotic evolution (Pawlowski 2008). Alternatively, filopodia-like structures were present in the ancestral eukaryotes and secondarily lost in multiple lineages. In this latter case, all eukaryotes may share the same molecular toolkit for filopodia formation. To discern among these two options, it is critical to determine the evolutionary history of proteins required for metazoan filopodia formation. The molecular architecture of filopodia in metazoans is well known (Gupton and Gertler 2007; Mattila and Lappalainen 2008; Mellor 2010) and includes a wide diversity of proteins. However, with the exception of the amoebozoan *Dictyostelium discoideum* (Faix and Rottner 2006), the molecular components of filopodia in non-metazoans remains largely unknown and its definition as filopodia is based simply on morphological characters and/or actin content (Adl et al. 2012; Cavalier-Smith 2012).

By deciphering the evolutionary history of the inventory of filopodial genes, as well as by experimentally analysing the expression and subcellular localization of some filopodial components in non-metazoans, we aim to address the ancestry of the molecular toolkit for filopodia formation in metazoans (Gupton and Gertler 2007).

We generated a consensus of proteins that are known to be involved in filopodia formation in metazoans, including filopodial proteins and others that can also act in different contexts but are essential for filopodia formation and analyzed the presence of homologs of these genes in the genomes of several close unicellular relatives of Metazoa, including the filasterean *Capsaspora owczarzaki* and the choanoflagellates *Salpingoeca rosetta* and *Monosiga brevicollis* (King et al. 2008; Ruiz-Trillo et al. 2008), together with a broad sampling of other eukaryotic lineages, including unpublished genome data from the apusozoan *Thecamonas trahens* and the early-branching fungi *Spizellomyces punctatus* and *Allomyces macrogynus*. Our data show that, unlike diverse outgroups of holozoans, both choanoflagellates and *C. owczarzaki* have a complex filopodial toolkit. Thus, although some filopodial components are ancient, most metazoan filopodial genes emerged relatively recently, within the holozoan stem lineage. Moreover, we show that filopodia are a pervasive feature of unicellular holozoans and that the filopodia marker protein fascin localizes specifically to filopodia and the microvillar collar in the choanoflagellate *S. rosetta*. Finally, gene transcription analyses of the filopodial toolkit genes in *S. rosetta* and *C. owczarzaki* show subfunctionalization of some of the components in different life history stages in *S. rosetta* and a strong correlation between gene expression and the morphological presence of filopodia in *C. owczarzaki*. Taken together, these data show not only that the origin of several key components of filopodia formation pre-dates the origin of metazoans, but also suggest that those molecules perform similar functions in unicellular and colonial relatives of metazoans.

Results & Discussion

Origin of the filopodial genetic toolkit

To investigate the origin and evolutionary history of proteins required for metazoan filopodia formation, we performed a taxon-rich genomic survey of the metazoan filopodial toolkit. To classify as many proteins as possible, we performed similarity searches and, when possible, phylogenetic analyses.

Actin-crosslinking proteins

Mechanical cohesion of filopodia is achieved by actin cross-linking proteins, namely fascin, espin, fimbrin (also known as plastin), alpha-actinin, and ERM (Ezrin-Radixin-Moesin) proteins. Our data show that fimbrin and alpha-actinin, both present not only in filopodia, but also in other actin-based structures (Mellor 2010), are present in most eukaryotes. Both ERM proteins, which link the actin cytoskeleton to the membrane (Bretscher et al. 2002; Hoeflich and Ikura 2004; Niggli and Rossy 2008), and fascin, a critical filament-bundling protein in metazoan filopodial structures, are specific to holozoans (Fig. 1, Fig. S2B) (Ruiz-Trillo 2008). Espin, on the other hand, is restricted to metazoans, where it is found in only a limited number of cell-types and specialized filopodial structures, including sterocilia (Mellor 2010) (Fig. 1). Thus, our data suggest that microvilli evolved concomitantly with ERM proteins, which are indeed crucial for the formation of microvilli (Nambiar et al. 2010), in the last common ancestor of choanoflagellates and metazoans (Fig. S1, Fig. S2A).

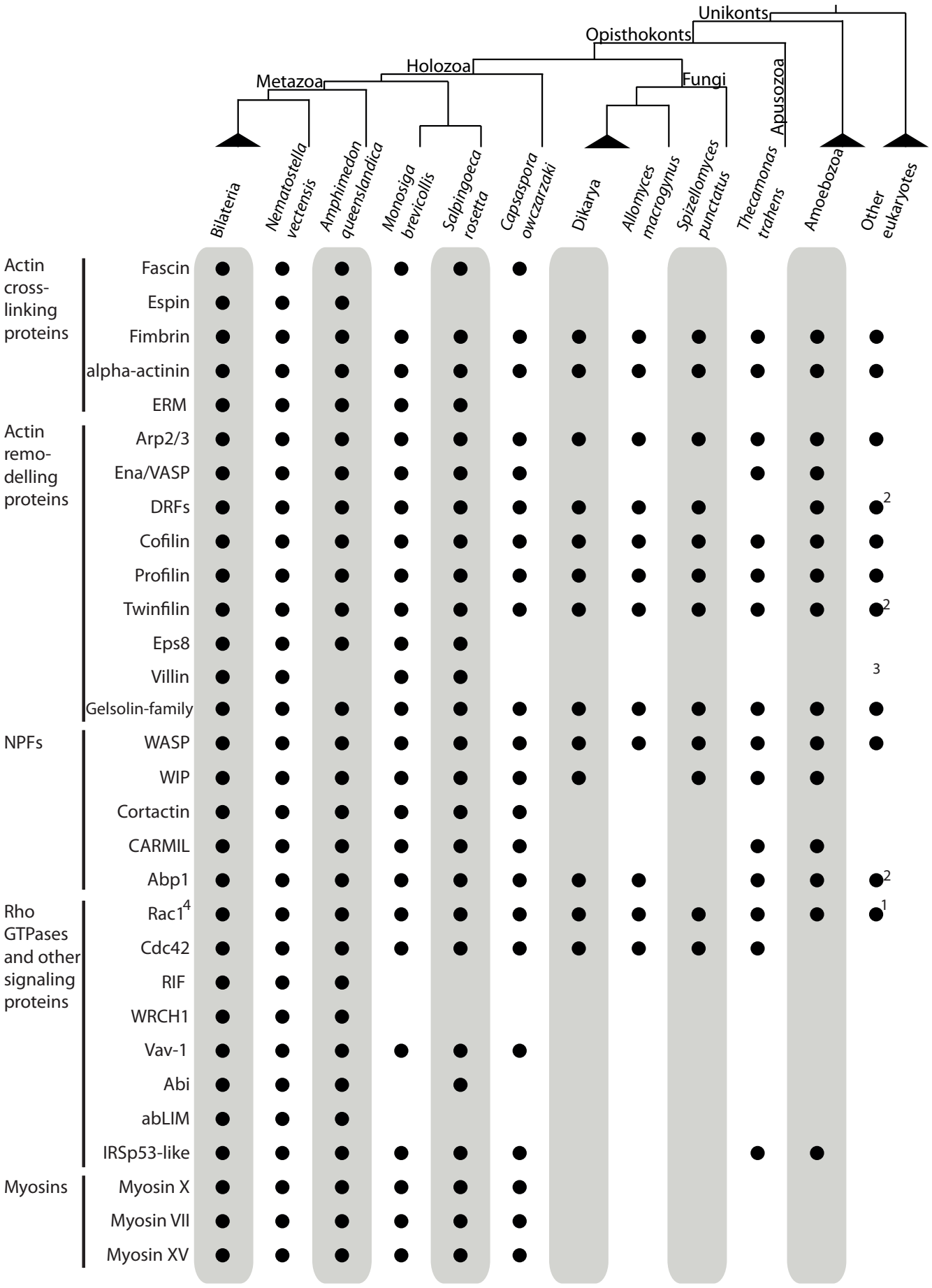


Figure 1. **Schematic representation of the eukaryotic tree of life showing the distribution of the different filopodial proteins.** A black dot indicates the presence of clear homologs, whereas absence of a dot indicates that a homolog was not detected in that taxon. [1]Based on Boureux et al. 2007 [2]Only present in the excavates *Naegleria gruberi* and *Trichomonas vaginalis*. [3]Plants have a villin-like gene, with the same domain architecture but not phylogenetically related with holozoan villin (Fig. S5). [4]Rac1 RhoGTPases are key filopodia-inducers in amoebozoans, but not in metazoans.

Actin-remodelling proteins

Most actin-remodelling proteins are widespread among eukaryotes (Fig. 1). This is the case with the Arp2/3 complex, a major actin-remodelling factor (Pollard 2007), which convergently elongates actin filaments from the cortical actin meshwork and is regulated, in metazoans, by WASP and cortactin (Weaver et al. 2003). The seven subunits of this complex are present in almost all eukaryotes, suggesting this is an ancient protein network.

Formins are also involved in actin filament formation, but instead of bundling an existent actin meshwork, formins nucleate actin filaments de novo. Our phylogenetic analyses show that Diaphanous-related formins (DRFs) (Chalkia et al. 2008), those with domain structure GBD-FH3-FH2-DAD, are present in all unikonts investigated as well as in Excavata (Fig. S3, Fig. S4). Interestingly, Ena/VASP, a multifaceted actin-regulatory protein with essential roles in filopodia formation and elongation that is essential for DRF-based de novo actin nucleation (Schirenbeck et al. 2006), is exclusive to unikonts and appears to have been secondarily lost in Fungi (Fig. 1). Therefore, although DRF-like formins are present in some bikonts (Excavata) it is unclear whether actin nucleation based on formins is truly an ancestral mechanism in eukaryotes, as they lack Ena/VASP.

Other proteins involved in actin remodelling include cofilin, which has depolymerizing activity, profilin, which sequesters actin monomers (Revenu et al. 2004), and several proteins, such as Eps8, Twinfilin, Villin and other Gelsolin family proteins (Fig. S5), that are involved in capping (Mattila and Lappalainen 2008) (i.e., they stabilize filament ends by binding them and inhibiting actin monomer association or dissociation). Most of these other actin regulators are widespread among eukaryotes or unikonts (Fig. 1), except for the capping protein Eps8, which is specific to metazoans and choanoflagellates, and Villin, that is restricted to holozoans, although plants have a villin-like protein. Eps8 is a direct binding partner of ERM proteins (see above) and, together, they stimulate formation of microvilli (Zwaenepoel et al. 2012). Villin is a multi-faceted actin-remodelling protein that, in metazoans, is usually associated with microvilli formation, especially in intestinal cells (Silacci et al. 2004; Khurana and George 2008; Nambiar et al. 2010).

Nucleation Promoting Factors

NPFs activate the Arp2/3 complex (Goley and Welch 2006). Some important NPFs in metazoan filopodia are WASP (and the associated WASP interacting protein (WIP)) (Veltman and Insall 2010; Kollmar et al. 2012) and Cortactin (Weaver et al. 2001; Goley and Welch 2006). In metazoans, WASP is a direct target of Cdc42 RhoGTPase, mediating the activation of the Arp2/3 complex (Takenawa and Miki 2001; Small et al. 2002; Antón et al. 2007; Faix et al. 2009; Mellor 2010), which is also activated by Cortactin (Kinley et al. 2003; Ren et al. 2009), while WIP is responsible for inactivating WASP (Antón et al. 2007).

Our data show that WASP is widespread among eukaryotes (Fig. 1, Fig. S6), while WIP is restricted to unikonts and Cortactin to holozoans (Fig. 1). It is worth mentioning that a major regulator of Cortactin, c-Src tyrosine kinase (Weaver et al. 2001; Weaver et al. 2003), is a component of the metazoan integrin adhesome that also originated in the holozoan lineage (Sebé-Pedrós et al. 2010; Suga et al. 2012).

Rho GTPases and other signaling proteins

RhoGTPases are key regulators of actin dynamics and play important roles as major switches in filopodia formation (Ridley 2006; Ladwein and Rottner 2008; Faix et al. 2009). They act through two types of actin nucleators: WASP and DRFs (Ridley 2006). Cdc42 RhoGTPase, thought to be exclusive to opisthokonts (Boureux et al. 2007), appears to be the primary filopodia-inducing RhoGTPase in metazoans (Ridley 2006). In *Dictyostelium*, which has filopodia but lacks Cdc42, Rac1 GTPases induce filopodia (Vlahou and Rivero 2006) whereas in plants, the Rac1-related Rop GTPases can also stimulate actin polymerization (Boureux et al. 2007). We find that Rac1-type RhoGTPases are ancestral within unikonts, and, interestingly, that Cdc42 is not specific to opisthokonts, since it is also present in the apusozoan *T. trahens*, sister group of opisthokonts (Fig. S7 and Fig. 1). RIF and WRCH1, which induce filopodia in specific metazoan cell types (Faix et al. 2009), are specific to Metazoa (Fig. 1).

Beside RhoGTPases, there are other signalling proteins involved in filopodia, such as abLIM, Abi, Vav-1, and IRSp53-like proteins. They evolved during different phases of eukaryotic evolutionary history. IRSp53-like proteins being ancestral within unikonts (Fig. S9), Vav-1 being specific to holozoans, abLIM specific to metazoans (Fig. S8) and Abi specific to metazoans and choanoflagellates (Fig. 1).

Motor proteins: myosins

Myosins are a large protein family present in all eukaryotes and essential for cell trafficking along actin filaments (Richards and Cavalier-Smith 2005;

Odronitz and Kollmar 2007). The MyTH4-FERM domain myosins (named according to their tail domain composition) have special relevance for filopodia function and formation. This is the case of the metazoan myosin X (Tuxworth et al. 2001; Berg and Cheney 2002; Nagy et al. 2008; Nambiar et al. 2010), which is essential for filopodia formation (Bohil et al. 2006) and for the transport of protein such as integrins (Breshears et al. 2010) or Ena/VASP protein to the filopodial tip (Zhang et al. 2004; Sousa and Cheney 2005). Two other MyTH4-FERM myosins, myosin VII and myosin XV, are also important for filopodia function (Breshears et al. 2010). Our analysis shows that these three myosins (X, VII and XV) emerged at the origin of Holozoa (Fig. 1). Amoebozoan myosin VII-like, known to be essential for filopodia formation in *Dictyostelium* (Sousa and Cheney 2005), likely evolved independently given the vast phylogenetic distance between it and the myosin VII present only in metazoans and choanoflagellates.

Filopodia in unicellular and colonial relatives of metazoans

The richness of the filopodial toolkit in unicellular and colonial holozoans prompted us to investigate the presence, abundance, and distribution of filopodia in these close relatives of metazoans. We stained for polymerized actin (using phalloidin) and tubulin (using anti-tubulin antibodies) in *C. owczarzaki* and *S. rosetta*. In *C. owczarzaki* multiple 1-20 μm long bundles of actin microfilaments can be found in the adherent stage (Fig. 2A). Scanning electron microscopy confirms the presence of multiple long filopodia-like structures in this cell stage of *C. owczarzaki* (Fig. 2B), in contrast with the naked cystic cell stage (Fig. 2C). In the choanoflagellate *S. rosetta* actin microfilaments were detected in two distinct sites: in the apical collar of actin-filled microvilli and in basally-positioned 1-10μm long cellular protrusions that resemble filopodia (Fig. 2D) (Leadbeater 1979; Dayel et al. 2011). Supporting the inference that the basal actin microfilaments are associated with filopodia, transmission electron microscopy of thin sections through *S. rosetta* cells shows the presence of multiple basally-positioned cellular processes (Fig. 2E and F, black rectangle). Thus, our data show the presence of multiple long, actin-filled cellular projections that resemble filopodia in two close relatives of metazoans.

Fascin localizes to filopodia and actin-filled microvilli in *S. rosetta*

In metazoans, Fascin functions as a filament-bundling protein that localizes to filopodia, and in some species also to microvilli. Given that we identified clear fascin homologs in *S. rosetta, M. brevicollis* and *C. owczarzaki* (Fig. 1, Fig. S2B) (see above), we next investigated whether the choanoflagellate fascin homologs might function in filopodia and microvilli as they do in metazoans (DeRosier and Tilney 2000; Kureishy et al. 2002). Western blot analysis shows that a commercially available fascin antibody, which was originally raised against the human protein fascin, recognizes a single band of approximately 55 kDa when used to probe the *S. rosetta* lysate (The S. rosetta genome encodes two fascin paralogs with predicted molecular weights of 54.3 and 54.6 kDa). Thus, we performed immuno-localization studies of fascin in *S. rosetta*. Interestingly, fascin localizes to the basal filopodia-like structures and to the actin filled collar of *S. rosetta* (Fig. 3A, B). These data suggest a functional conservation of the actin-cosslinking protein fascin between choanoflagellates and Metazoa.

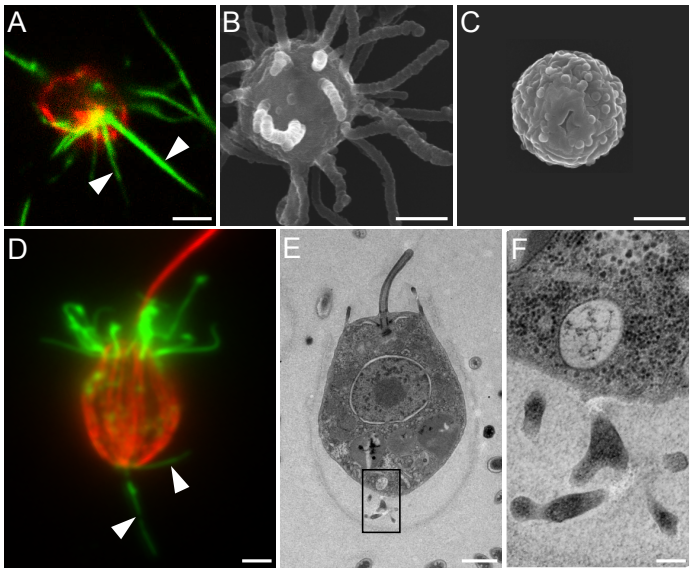


Figure 2. **Filopodia in close relatives of metazoans.** (A) *C. owczarzaki* cells were stained with phalloidin and antibodies against beta-tubulin. Colabeling of tubulin allows for identifying the cell's cytoskeleton/architecture. In *C. owczarzaki*, adherent cells bear multiple long bundles of actin microfilaments. (B, C) Scanning electron microscopy shows the presence of multiple long filopodia-like structures in *C. owczarzaki* adherent cells, but not in floating cells. (D) *S. rosetta* cells were stained with phalloidin and antibodies against beta-tubulin. In *S. rosetta*, attached cells bear actin microfilaments in two distinct sites: in the apical collar of actin-filled microvilli and in basally-positioned long cellular protrusions. (E) Transmission electron microscopy of thin sections through a choanoflagellate shows the presence of basally-positioned cellular processes (black rectangle and magnification in (F). Scale bars (A-E: 1 μm, F: 200 nm).

Expression of the filopodial toolkit in *C. owczarzaki* and *S. rosetta*

To further investigate the putative functional homology of filopodial genes between metazoans and their unicellular relatives, we analysed the expression levels of diverse filopodial and other actin-related genes between different life history stages of *C. owczarzaki* and *S. rosetta*.

C. owczarzaki can differentiate in, at least, two cell types, an attached cell type that has filopodia (Fig. 2B) and a naked, non-filopodial form that is not attached to the substrate (Fig. 2C). We investigated the expression of filopodial gene homologs in these two cell types. Homologs of most of the genes involved in metazoan filopodia, such as fascin, myosin X, Cortactin, and all subunits of the Arp2/3 complex, are upregulated in the adherent filopodial form of *C. owczarzaki* (Fig. 4A). The differential expression of homologs of filopodial genes in filopodial and non-filopodial life stages of *C. owczarzaki* is consistent with the hypothesis that there is functional homology between *C. owczarzaki* and metazoan filopodia.

S. rosetta can differentiate into at least five distinct cell types, including three solitary cell types (slow swimmers, fast swimmers, and substrate attached cells) and two colonial forms (rosettes and chains) (Dayel et al. 2011). Both attached cells and colonial cells have been previously reported to produce filopodia (Leadbeater 1979; Dayel et al. 2011). In attached cells, filopodia may mediate the attachment to environmental substrates both by searching the environment for suitable attachment sites and by contributing to the construction of a goblet-shaped attachment structure called a theca. In colonies, filopodia extend from the basal pole of cells in most, but not all, rosette colonies and may contribute to colony formation or stabilization. When we compared the expression of homologs of filopodial genes between attached cells and colonies (chains and rosettes; Fig. 4B), most were not differentially expressed, consistent with the hypothesis that cells in both life history stages form filopodia. Surprisingly, however, some of the filopodial gene homologs were differentially expressed, suggesting that the molecular composition of filopodia in different life stages might be specialized. For example, one *S. rosetta* fascin homolog, Fascin1, is upregulated in attached cells, whereas Fascin2 is upregulated in colonies, suggesting subfunctionalization. Other genes upregulated in colonies are Diaphanous-like, Vav-1 and Abi, whereas Villin and MyosinXV are upregulated in attached cells. This raises the possibility that the different patterns of expression in different types of filopodiated cells in *S. rosetta* may contribute to cell differentiation.

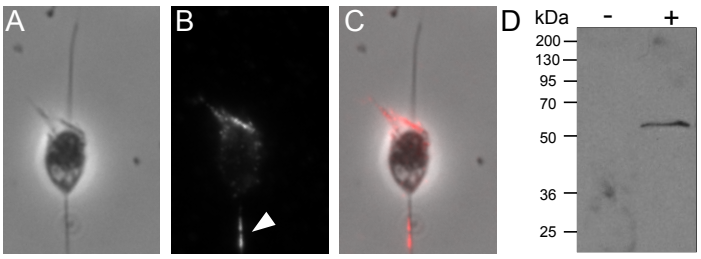


Figure 3. **Subcellular localization of Fascin in *S. rosetta*.** (A) Phase contrast microscopy shows the morphology of fixed S. rosetta cells. (B and C) Immunolocalization studies reveal that Fascin localizes to basal filopodia-like structures (arrowhead) and to the apical actin filled collar. Staining was performed with antibodies to human Fascin. (D) Western blot analysis shows that Fascin antibodies probed with *S. rosetta* cell lysate detect a single band of approximately 55 kDa (+). No signal was detected when primary Fascin antibody was omitted (-).

Evolutionary assembly of the metazoan filopodial toolkit

Our evolutionary reconstruction reveals the gradual assembly of the metazoan filopodial toolkit (Fig. 5). Many actin remodeling and cross-linking proteins, such as fimbrin, alpha-actinin, profilin, cofilin and twinfilin, are ancient. It is likely that Arp2/3-WASP-DRF based filopodia formation (with DRF as the anti-capping agent instead of VASP), coupled with RhoGTPase regulation, was the ancestral eukaryotic mechanism, rather than the formin-based mechanism, since demonstration of DRF-based filopodia formation (independent of Arp2/3 and without the presence of VASP protein, known to be essential for formin-based filopodia formation (Schirenbeck et al. 2006), see above) awaits demonstration in non-unikont taxa. Later, the Ena/VASP protein evolved in the unikont clade, where two independent mechanisms of filopodia formation, Arp2/3-WASP-VASP and DRFs-VASP, have been demonstrated in amoebozoans and in metazoans.

It was only later, in the stem of the holozoan lineage, that the metazoan filopodia-specific toolkit was established. This includes Cdc42 signalling as an initiator of filopodia formation and also the control of filopodia formation through Tyrosine kinase signalling (involving Src and Abl cytoplasmic TyrK). The metazoan filopodia toolkit also includes fascin as the main actin-bundling protein, specific motor proteins myosin X, VII and XV and other proteins such as cortactin, Vav-1 and Abi. Expression data suggest that this complex complement is, indeed, functionally conserved in unicellular holozoans, as most filopodial genes are clearly overexpressed in *C.*

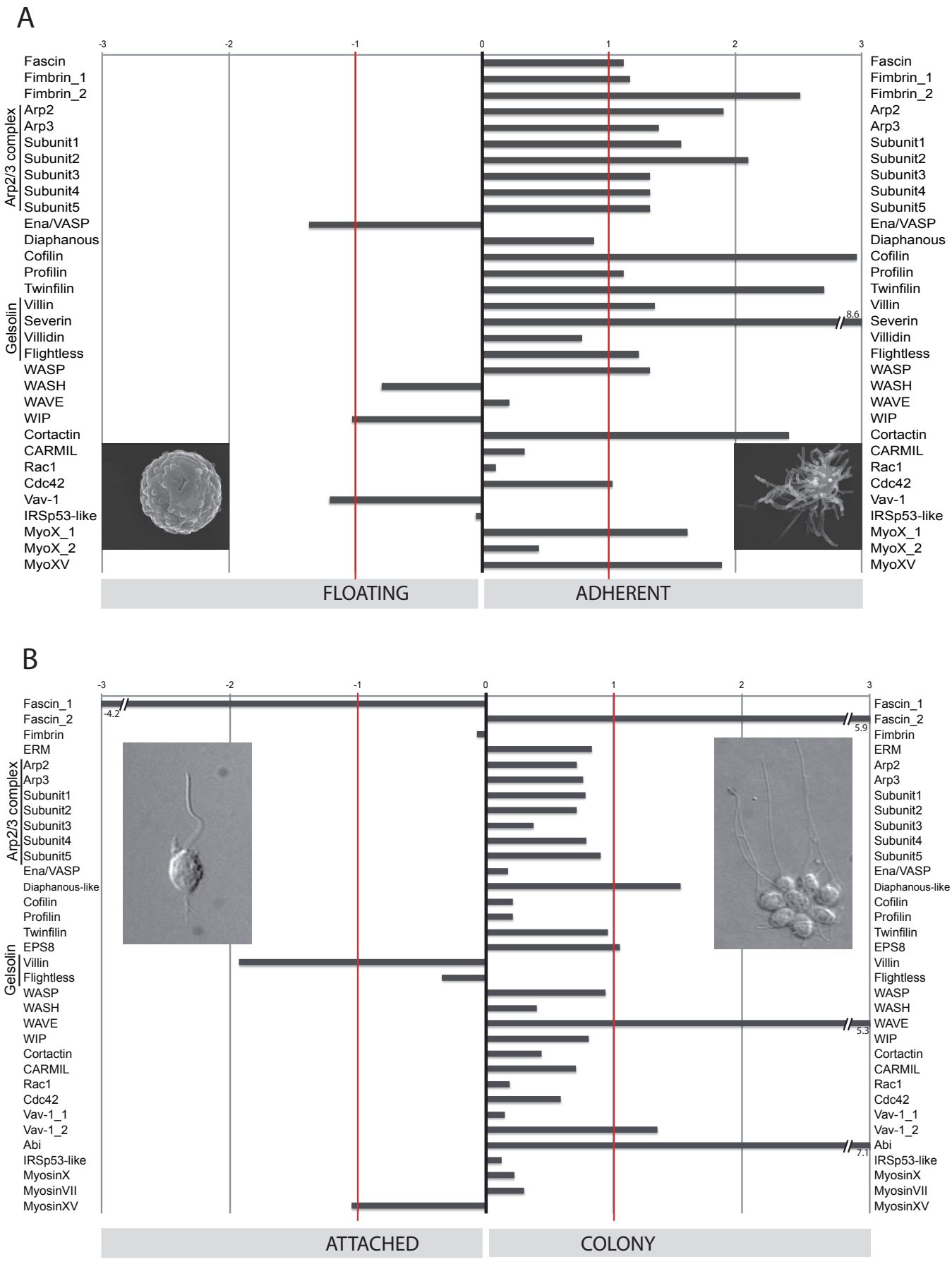


Figure 4. **Expression of filopodial genes in unicellular Holozoa.** (A) Log2-fold expression of *C. owczarzaki* filopodial genes between adherent and floating stages. (B) Log2-fold expression of *S. rosetta* filopodial genes between attached an colonial stages. Red lines highlight 2-fold expression differences.

owczarzaki filopodial cells. Moreover the main actin-bundling protein fascin localizes to filopodia and microvilli in *S. rosetta*, suggesting a conserved role of fascin already in choanoflagellates. Interestingly, the fact that choanoflagellate microvilli use fascin as its actin-crosslinking protein, in contrast with some metazoan cell-types that use other proteins like espin (DeRosier and Tilney 2000; Nambiar et al. 2010; Sedeh et al. 2010), suggests that fascin is the ancestral microvilli actin-crosslinking protein and reinforces the view of microvilli as filopodia-related structures, reusing part of the toolkit, together with the addition of new proteins like ERM and Eps8.

In the common ancestor of choanoflagellates and metazoans, the complexity of the filopodial apparatus was further expanded as filopodial specialization in the form of the microvillar collar evolved. This structure is crucial in many metazoan cell-types (e.g. epithelia) and as a feeding structure in choanoflagellates. Our data show that ERM and Eps8, two proteins that are crucial for microvilli formation, evolved in the last common ancestor of choanoflagellates and metazoans. Finally, in metazoans the toolkit further expanded, particularly with the evolution of new RhoGTPases that are known to act instead of Cdc42 in specific cell-types.

Based on our data it is possible that other non-metazoan eukaryotic lineages evolved their specific filopodial toolkits based on an ancient molecular machinery that included core actin linking proteins (profilin, twinfilin, fimbrin, cofilin and others) and an ancestral filopodia formation mechanism (Arp2/3-WASP-DRFs). We hypothesize this mechanism was deployed under control of different signalling triggers (for example, Cdc42 in metazoans and Rac1 in amoebozoans) and together with different specific co-actors. The recently reported NET superfamily (Deeks et al. 2012), a plant specific membrane-actin cytoskeleton adaptor protein, exemplifies this idea of independently-evolved genes for solving similar problems (in this case, the interaction between the membrane and the actin cytoskeleton). Therefore, we infer that metazoan-type filopodia originated at the stem of Holozoa, built upon some ancient and many evolutionarily new molecular components.

In any case, the study of the cell biology and genome content of filopodiated chlorarachniophytes (Rhizaria) (Ota and Vaulot 2012), labyrinthulomycetes, other filopodiated stramenopiles (Tsui et al. 2009; Gómez et al. 2011) and filopodiated Excavata, such as *Naegleria gruberi* (Preston and King 2005), will be crucial to gain new insights into the question of

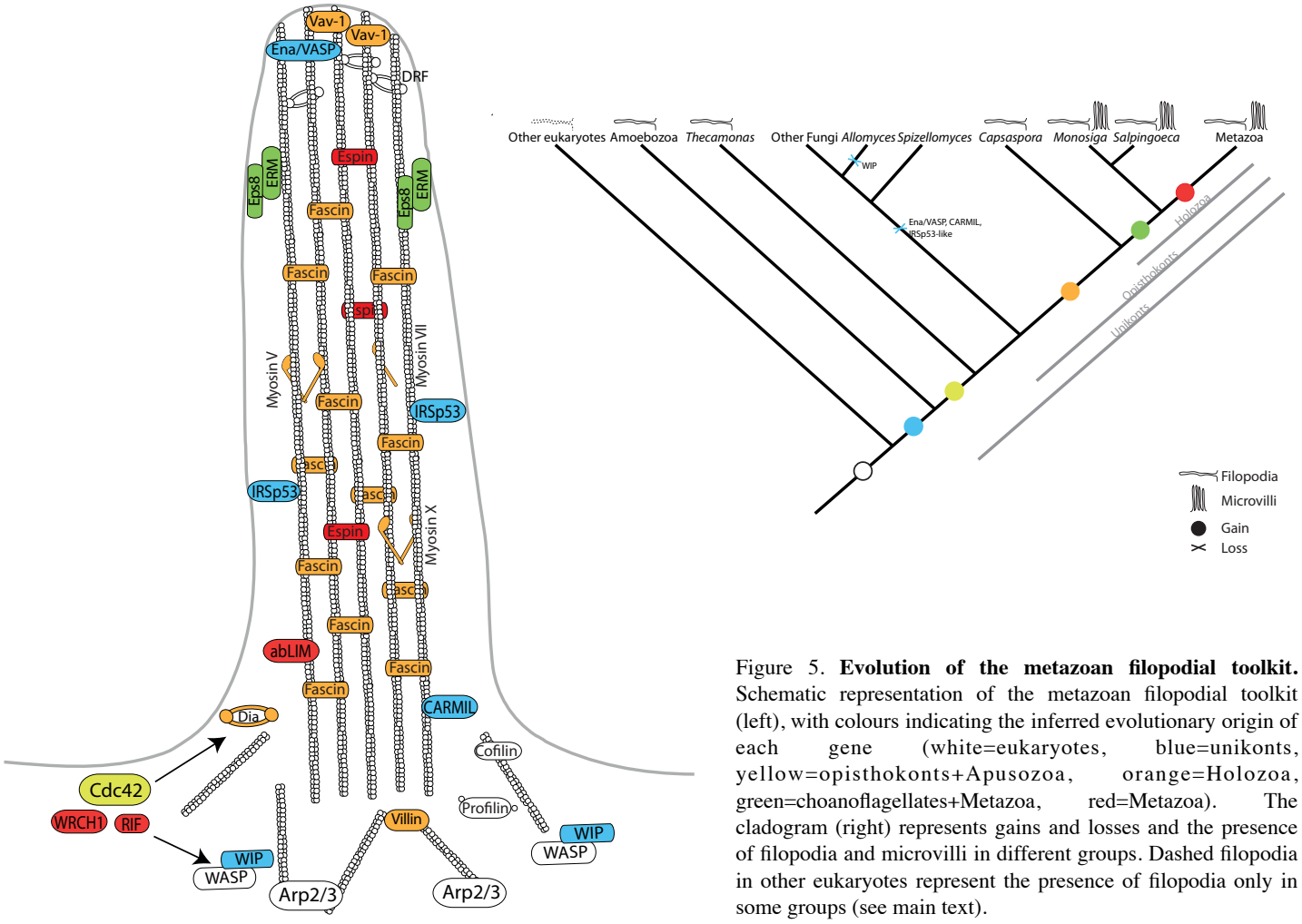


Figure 5. **Evolution of the metazoan filopodial toolkit.** Schematic representation of the metazoan filopodial toolkit (left), with colours indicating the inferred evolutionary origin of each gene (white=eukaryotes, blue=unikonts, yellow=opisthokonts+Apusozoa, orange=Holozoa, green=choanoflagellates+Metazoa, red=Metazoa). The cladogram (right) represents gains and losses and the presence of filopodia and microvilli in different groups. Dashed filopodia in other eukaryotes represent the presence of filopodia only in some groups (see main text).

whether there is a common, functionally homologous, molecular toolkit behind all eukaryotic filopodia.

Conclusion

Our study reconstructs in detail the assembly of the metazoan filopodia molecular toolkit. Some components of the metazoan filopodial toolkit are paneukaryotic, while other elements appeared at the origin of unikonts. Finally, there was a burst of evolution of filopodial components at the onset of Holozoa..

We further demonstrate that fascin is expressed both in filopodia and microvilli in the choanoflagellate *S. rosetta* and that filopodial genes are differentially upregulated in *C. owczarzaki*'s filopodial cell-stage. This suggests functional conservation of the metazoan filopodial toolkit in both choanoflagellates and *C. owczarzaki*. Given the functional homology of filopodia between metazoans and their unicellular relatives, we hypothesize that the pre-existence of this complex filopodial toolkit in the ancestors of metazoans may have contributed to the origin of metazoans, by being co-opted to function in cell-cell and cell-matrix adhesion functions within a multicellular context.

Materials & Methods

Gene searches & Phylogenetic analysis

A primary search was performed using the basic local alignment sequence tool (BLAST: BlastP and TBlastN) using *Homo sapiens* and *C. owczarzaki* proteins as queries against Protein and Genome databases with the default BLAST parameters and an e-value threshold of e-5 at the National Center for Biotechnology Information (NCBI), the Joint Genome Institute (JGI), the Broad Institute (for *Salpingoeca rosetta* and *Spizellomyces punctatus*), as well as the *A. queenslandica* genome database (www.metazome.net/amphimedon). For some proteins, we also performed Hmmer searches using HMMER3.0b2 (Eddy 1998) to confirm that we were retrieving all orthologs.

We performed searches using the following taxon sampling: seven metazoans (*Homo sapiens, Drosophila melanogaster, Daphnia pulex, Capitella teleta, Lottia gigantea, Nematostella vectensis, Amphimedon queenslandica*), two choanoflagellates (*Monosiga brevicollis, Salpingoeca rosetta*), one filasterean (*Capsaspora owczarzaki*), four fungi (*Laccaria bicolor, Saccharomyces cerevisiae, Allomyces macrogynus, Spizellomyces punctatus*), one apusozoan (*Thecamonas trahens*), two amoebozoans (*Acanthamoeba castellanii, Dictyostelium discoideum*), three viridiplantae (*Arabidopsis thaliana, Ostreococcus taurii, Chlamydomonas reinhardtii*), two excavates (*Trichomonas vaginalis , Naegleria gruberi*), and three chromalveolates (*Thalassiosira pseudonana, Tetrahymena thermophila , Toxoplasma gondii*).

Alignments were constructed using the MAFFT v.6 online server (Katoh et al. 2002) and then manually inspected and edited using Geneious software. Only those species and those positions that were unambiguously aligned were included in the final analyses. Maximum likelihood (ML) phylogenetic trees were estimated by RaxML (Stamatakis 2006) using the PROTGAMMAWAG+Γ+I model, which uses the WAG amino acid exchangeabilities and accounts for among-site rate variation with a four category discrete gamma approximation and a proportion of invariable sites. Statistical support for bipartitions was estimated by performing 100-bootstrap replicates using RaxML with the same model. Bayesian analyses were performed with MrBayes3.2 (Huelsenbeck and Ronquist 2001), using the LG+Γ+I model of evolution, with four chains, a subsampling frequency of 100 and two parallel runs. Runs were stopped when the average standard deviation of split frequencies of the two parallel runs was <0.01, usually at around 18,000,000 generations. The two LnL graphs were checked and an appropriate burn-in length established. Bayesian posterior probabilities (BPP) were used to assess the confidence values of each bipartition.

Cell Culture and Microscopy

S. rosetta cultures enriched for attached cells were maintained in artificial sea water and split 1:5 every 3 days. For immunofluorescence, the cells were grown to a density of 10^6 cells/mL and carefully scraped off from the surface of the culture flasks. Cells were then pelleted by spinning for 10 min at 4000 x g and resuspended in a small volume of artificial seawater. Approximately 0.4 mL of the cells were applied to poly-L-lysine coated coverslips, left to attach for 30 min.
C. owczarzaki cells were grown on coverslides in ATCC medium 1034 (modified PYNFH medium) for two days and directly fixed.
For both *S. rosetta* and *C. owczarzaki*, cells were fixed for 5 min with 6% aceton and for 15 min with 4% formaldehyde. The coverslips were washed gently four times with PEM (100 mM Pipes at pH 6.9, 1 mM EGTA, and 0.1 mM MgSO4), incubated

for 30 min in blocking solution (PEM+: 1% BSA, 0.3% Triton X-100), 1 h in primary antibodies solution (in PEM+), and after further washes (PEM+), 1 h in the dark with fluorescent secondary antibodies (1:100 in PEM+, Alexa Fluor 488 goat anti-mouse, and Alexa Fluor 568 goat anti-rabbit; Invitrogen) and washed again four times (PEM). To visualize F-actin coverslips were incubated for 15 min in the dark with with rhodamine phalloidin (6 U/ml in PEM; Molecular Probes). After 3 washes (PEM), coverslips were mounted onto slides with Fluorescent Mounting Media (4 µL; Prolong Gold Antifade, Invitrogen). The following primary antibodies have been used: mouse monoclonal antibody against β-tubulin (E7, 1:400; Developmental Studies Hybridoma Bank); mouse monoclonal antibody against Fascin (ab78487, 1:100; Abcam). Images were taken with a 100x oil immersion objective on an inverted Leica microscope. For scanning electron microscopy (SEM) C. owczarzaki cells were fixed with 1h 2'5% glutaraldehyde and 1h with 1% osmium tetroxide, followed by sequential dehydration with ethanol. Next, drying critical point was performed and samples were coated with carbon. Samples were observed in a Hitachi S-4100 microscope.
For transmission electron microscopy (TEM) choanoflagellate cells were concentrated by gentle centrifugation, loaded into 100 μm-deep specimen carriers and high pressure frozen in a Bal-Tec HPM 010 high pressure freezer (Bal-Tec AG, Liechtenstein). Freeze-substitution was performed over 2 hours by the SQFS method of McDonald and Webb (2011)(McDonald and Webb 2011), then infiltrated with Eponate 12 resin and polymerized in a Pelco Biowave research microwave oven (Ted Pella, Inc., Redding, CA) over a period of 2 hours. Sections were cut at 70 nm thickness, post-stained with uranyl acetate and lead citrate, and viewed in a Tecnai 12 transmission EM (FEI Inc., Hillsboro, OR) operating at 120 kV. Images were recorded on a Gatan Ultrascan 1000 CCD camera.

Gene expression analyses
Total RNA from *C. owczarzaki*'s described life stages was extracted using Trizol reagent. Libraries were sequenced with 76 base paired-read on an Illumina HiSeq instrument (Illumina). mRNA was isolated from *S. rosetta* cultures enriched for colonial and attached cells using the RNAeasy (Qiagen) and Oligotex (Qiagen) kits. Libraries were sequenced with 68 base paired-end reads on an Illumina GAII instrument (Illumina) following manufacturer's recommendations. In both cases, fragments per kilobase per million reads mapped (FPKM) per CDS was calculated and colonial and attached values averaged and log2 transformed.

Acknowledgments

We thank Joint Genome Institute and Broad Institute for making data publicly available. This work was supported by an Institució Catalana per a la Recerca i Estudis Avançats contract, 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. A.S-P. is supported by a pregraduate Formacion Profesorado Universitario grant from MICINN. P.B. is supported by a DFG postdoctoral fellowship. B.F.L. thanks for financial support through the Canadian Research Chair program.

Adl SM, Simpson AGB, Lane CE, et al. 2012. The revised classification of eukaryotes. J Eukaryot Microbiol 59:429–514.

Antón IM, Jones GE, Wandosell F, Geha R, Ramesh N. 2007. WASP-interacting protein (WIP): working in polymerisation and much more. Trends Cell Biol 17:555–62.

Berg JS, Cheney RE. 2002. Myosin-X is an unconventional myosin that undergoes intrafilopodial motility. Nat Cell Biol 4:246–250.

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.

Boureux A, Vignal E, Faure S, Fort P. 2007. Evolution of the Rho family of ras-like GTPases in eukaryotes. Mol Biol Evol 24:203–16.

Breshears LM, Wessels D, Soll DR, Titus M a. 2010. An unconventional myosin required for cell polarization and chemotaxis. Proc Natl Acad Sci U S A 107:6918–23.

Bretscher A, Edwards K, Fehon RG. 2002. ERM proteins and merlin: integrators at the cell cortex. Nat Rev Mol Cell Biol 3:586–99.

Brown MW, Kolisko M, Silberman JD, Roger AJ. 2012. Aggregative Multicellularity Evolved Independently in the Eukaryotic Supergroup Rhizaria. Curr Biol 22:1–5.

Cavalier-Smith T, Chao EE. 2010. Phylogeny and Evolution of Apusomonadida (Protozoa: Apusozoa): New Genera and Species. Protist 161:549–576.

Cavalier-Smith T. 2003. Phylogeny of Choanozoa, Apusozoa, and other Protozoa and early eukaryote megaevolution. J Mol Evol 56:540–563.

Cavalier-Smith T. 2012. Early evolution of eukaryote feeding modes, cell structural diversity, and classification of the protozoan phyla Loukozoa, Sulcozoa, and Choanozoa. Eur J Protistol. [Epub ahead of print]

Chalkia D, Nikolaidis N, Makalowski W, Klein J, Nei M. 2008. Origins and evolution of the formin multigene family that is involved in the formation of actin filaments. Mol Biol Evol 25:2717–33.

Dayel MJ, Alegado R a, Fairclough SR, Levin TC, Nichols S a, McDonald K, King N. 2011. Cell differentiation and morphogenesis in the colony-forming choanoflagellate Salpingoeca rosetta. Dev Biol 357:73–82.

Deeks MJ, Calcutt JR, Ingle EKS, et al. 2012. A Superfamily of Actin-Binding Proteins at the Actin-Membrane Nexus of Higher Plants. Curr Biol 22:1–6.

DeRosier DJ, Tilney LG. 2000. F-actin bundles are derivatives of microvilli: What does this tell us about how bundles might form? J Cell Biol 148:1–6.

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

Faix J, Breitsprecher D, Stradal TEBTEBTEB, Rottner K. 2009. Filopodia: Complex models for simple rods. Int J Biochem Cell Biol 41:1656–1664.

Faix J, Rottner K. 2006. The making of filopodia. Current Opinion in Cell Biology 18:18–25.

Fletcher D a, Mullins RD. 2010. Cell mechanics and the cytoskeleton. Nature 463:485–92.

Goley ED, Welch MD. 2006. The ARP2/3 complex: an actin nucleator comes of age. Nat Rev Mol Cell Biol 7:713–26.

Gonobobleva E, Maldonado M. 2009. Choanocyte ultrastructure in Halisarca dujardini (Demospongiae, Halisarcida). Journal of Morphology 270:615–27.

Gupton SL, Gertler FB. 2007. Filopodia : The Fingers That Do the Walking. Sci Signal 2007:re5.

Gómez F, Moreira D, Benzerara K, López-García P. 2011. Solenicola setigera is the first characterized member of the abundant and cosmopolitan uncultured marine stramenopile group MAST-3. Environ Microbiol 13:193–202.

Hoeflich KP, Ikura M. 2004. Radixin: cytoskeletal adopter and signaling protein. Int J Biochem Cell Biol 36:2131–6.

Huelsenbeck JPP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–5.

Karpov S, Leadbeater B. 1998. Cytoskeleton structure and composition in choanoflagellates. J Euk Micro 45:361–367.

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.

Khurana S, George SP. 2008. Regulation of cell structure and function by actin-binding proteins: villin's perspective. FEBS Lett 582:2128–2139.

King N, Westbrook M, Young S, Kuo A. 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 21:4300–4305.

Kinley AW, Weed S a, Weaver AM, Karginov A V, Bissonette E, Cooper J a, Parsons JT. 2003. Cortactin interacts with WIP in regulating Arp2/3 activation and membrane protrusion. Curr Biol 13:384–93.

Kollmar M, Lbik D, Enge S. 2012. Evolution of the eukaryotic ARP2/3 activators of the WASP family: WASP, WAVE, WASH, and WHAMM, and the proposed new family members WAWH and WAML. BMC Res Notes 5:1–23.

Kureishy N, Sapountzi V, Prag S, Anilkumar N, Adams JC. 2002. Fascins, and their roles in cell structure and function. Bioessays 24:350–61.

Ladwein M, Rottner K. 2008. On the Rho'd: the regulation of membrane protrusions by Rho-GTPases. FEBS Lett 582:2066–74.

Leadbeater BSC, Morton C. 1974. A microscopical study of a marine species of Codosiga James-Clark (Choanoflagellata) with special reference to the ingestion of bacteria. Biol J Linn Soc Lond 6:337–347.

Leadbeater BSC, Yu Q, Kent J, Stekel DJ, B PRS. 2009. Three-dimensional images of choanoflagellate loricae. Proc R Soc Lond B Biol Sci 276:3–11.

Leadbeater BSC. 1979. Developmental and Ultrastructural Observations on Two Stalked Marine Choanoflagellates, Acanthoecopsis spiculifera Norris and Acanthoeca spectabilis Ellis. Proc R Soc Lond B Biol Sci 204:57–66.

Lundquist E a. 2009. The finer points of filopodia. PLoS Biol 7:e1000142.

Magie CR, Daly M, Martindale MQ. 2007. Gastrulation in the cnidarian Nematostella vectensis occurs via invagination not ingression. Dev Biol 305:483–497.

Mattila PK, Lappalainen P. 2008. Filopodia: molecular architecture and cellular functions. Nat Rev Mol Cell Biol 9:446–454.

McDonald KL, Webb RI. 2011. Freeze substitution in 3 hours or less. J Microsc 243:227–233.

Mellor H. 2010. The role of formins in filopodia formation. Biochim Biophys Acta 1803:191–200.

Mikrjukov KA, Mylnikov AP. 2001. A study of the fine structure and the mitosis of a lamellicristate amoeba, Micronuclearia podoventralis gen. et sp. nov.(Nucleariidae, Rotosphaerida). Eur J Protistol 37:15–24.

Nagy S, Ricca BL, Norstrom MF, Courson DS, Brawley CM, Smithback P a, Rock RS. 2008. A myosin motor that

selects bundled actin for motility. Proc Natl Acad Sci U S A 105:9616–20.

Nambiar R, McConnell RE, Tyska MJ. 2010. Myosin motor function: the ins and outs of actin-based membrane protrusions. Cell Mol Life Sci 67:1239–1254.

Niggli V, Rossy J. 2008. Ezrin/radixin/moesin: versatile controllers of signaling molecules and of the cortical cytoskeleton. Int J Biochem Cell Biol 40:344–9.

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.

Ota S, Eikrem W, Edvardsen B. 2011. Ultrastructure and Molecular Phylogeny of Thaumatomonads (Cercozoa) with Emphasis on Thaumatomastix salina from Oslofjorden, Norway. Protist 163:560–573.

Ota S, Vaulot D. 2012. Lotharella reticulosa sp. nov.: a highly reticulated network forming chlorarachniophyte from the Mediterranean Sea. Protist 163:91–104.

Pawlowski J. 2008. The twilight of Sarcodina : a molecular perspective on the polyphyletic origin of amoeboid protists. Protistology 5:281–302.

Pollard TD. 2007. Regulation of actin filament assembly by Arp2/3 complex and formins. Annu Rev Biophys Biomol Struct 36:451–77.

Preston T, King C. 2005. Locomotion and Phenotypic Transformation of the Amoeboflagellate Naegleria gruberi at the Water-Air Interface. J Euk Micro 50:245–251.

Ren G, Crampton MS, Yap AS. 2009. Cortactin: Coordinating adhesion and the actin cytoskeleton at cellular protrusions. Cell Motil Cytoskeleton 66:865–73.

Revenu C, Athman R, Robine S, Louvard D. 2004. The co-workers of actin filaments: from cell structures to signals. Nat Rev Mol Cell Biol 5:635–46.

Richards TA, Cavalier-Smith T. 2005. Myosin domain evolution and the primary divergence of eukaryotes. Nature 436:1113–8.

Ridley AJ. 2006. Rho GTPases and actin dynamics in membrane protrusions and vesicle trafficking. Trends Cell Biol 16:522–9.

Ruiz-Trillo I, Roger AJAJ, Burger G, Gray MWMW, Lang BFF. 2008. A Phylogenomic Investigation into the Origin of Metazoa. Mol Biol Evol 25:664–72.

Schirenbeck A, Arasada R, Bretschneider T, Stradal TEB, Schleicher M, Faix J. 2006. The bundling activity of vasodilator-stimulated phosphoprotein is required for filopodium formation. Proc Natl Acad Sci U S A 103:7694–9.

Schäfer C, Born S, Möhl C, Houben S, Kirchgeßner N, Merkel R. 2010. Dependence of adhesion, actin bundles, force generation and transmission on filopodia. Cell Adhesion & Migration 4:2:215–225.

Sebé-Pedrós A, Roger A, Lang FB, King N, Ruiz-Trillo I. 2010. Ancient origin of the integrin-mediated adhesion and signaling machinery. Proc Natl Acad Sci U S A 107:10142–7.

Sedeh RS, Fedorov A a, Fedorov E V, Ono S, Matsumura F, Almo SC, Bathe M. 2010. Structure, evolutionary conservation, and conformational dynamics of Homo sapiens fascin-1, an F-actin crosslinking protein. J Mol Biol 400:589–604.

Silacci P, Mazzolai L, Gauci C, Stergiopulos N, Yin HL, Hayoz D. 2004. Gelsolin superfamily proteins: key regulators of cellular functions. Cell Mol Life Sci 61:2614–23.

Small JV, Stradal T, Vignal E, Rottner K. 2002. The lamellipodium: where motility begins. Trends Cell Biol 12:112–120.

Soldati D, Meissner M. 2004. Toxoplasma as a novel system for motility. Curr Opin Cell Biol 16:32–40.

Sousa AD, Cheney RE. 2005. Myosin-X: a molecular motor at the cell's fingertips. Trends Cell Biol 15:533–9.

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

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.

Takenawa T, Miki H. 2001. WASP and WAVE family proteins: key molecules for rapid rearrangement of cortical actin filaments and cell movement. J Cell Sci 114:1801–9.

Tsui CKM, Marshall W, Yokoyama R, Honda D, Lippmeier JC, Craven KD, Peterson PD, Berbee ML. 2009. Labyrinthulomycetes phylogeny and its implications for the evolutionary loss of chloroplasts and gain of ectoplasmic gliding. Mol Phylogenet Evol 50:129–40.

Tuxworth RI, Weber I, Wessels D, Addicks GC, Soll DR, Gerisch G, Titus MA. 2001. A role for myosin VII in dynamic cell adhesion. Curr Biol 11:318–329.

Veltman DM, Insall RH. 2010. WASP family proteins: their evolution and its physiological implications. Mol Biol Cell 21:2880.

Vlahou G, Rivero F. 2006. Rho GTPase signaling in Dictyostelium discoideum: insights from the genome. Eur J Cell Biol 85:947–59.

Weaver AM, Karginov AV, Kinley AW, Weed SA, Li Y, Parsons JT, Cooper JA. 2001. Cortactin promotes and

stabilizes Arp2 / 3-induced actin filament network formation. Curr Biol 11:370–374.

Weaver AM, Young ME, Lee W-L, Cooper J a. 2003. Integration of signals to the Arp2/3 complex. Current Opinion in Cell Biology 15:23–30.

Zettler LAA, Nerad TA, O'Kelly CJ, Sogin ML. 2001. The Nucleariid Amoebae: More Protists at the Animal-Fungal Boundary. J Euk Micro 48:293–297.

Zhang H, Berg JS, Li Z, Wang Y, Lang P, Sousa AD, Bhaskar A, Cheney RE, Stromblad S. 2004. Myosin-X provides a motor-based link between integrins and the cytoskeleton. Nat Cell Biol 6:523–531.

Zwaenepoel I, Naba A, Da Cunha MML, Del Maestro L, Formstecher E, Louvard D, Arpin M. 2012. Ezrin regulates microvillus morphogenesis by promoting distinct activities of Eps8 proteins. Mol Biol Cell 23:1080–94.

Supplementary material

All branches in the phylogenetic trees follow the same taxonomic colour-code: Metazoa-Black (e.g. *Homo sapiens, Drosophila melanogaster, Daphnia pulex, Capitella teleta, Lottia gigantea, Nematostella vectensis, Amphimedon queenslandica*), Choanoflagellata-Green (*Monosiga brevicollis, Salpingoeca rosetta*), Filasterea-Red (*Capsaspora owczarzaki*), Fungi-Orange (e.g. *Laccaria bicolor, Saccharomyces cerevisiae, Allomyces macrogynus, Spizellomyces punctatus*), Apusozoa-Yellow (*Thecamonas trahens*), Amoebozoa-Blue (*Acanthamoeba castellanii*, *Dictyostelium discoideum*), Bikonta-Grey (*Arabidopsis thaliana*, *Ostreococcus taurii, Chlamydomonas reindhartii* (Plantae), *Trichomonas vaginalis*, *Naegleria gruberi* (Excavata), *Thalassiosira pseudonana, Tetrahymena thermophila*, *Toxoplasma gondii* (Chromalveolata)).

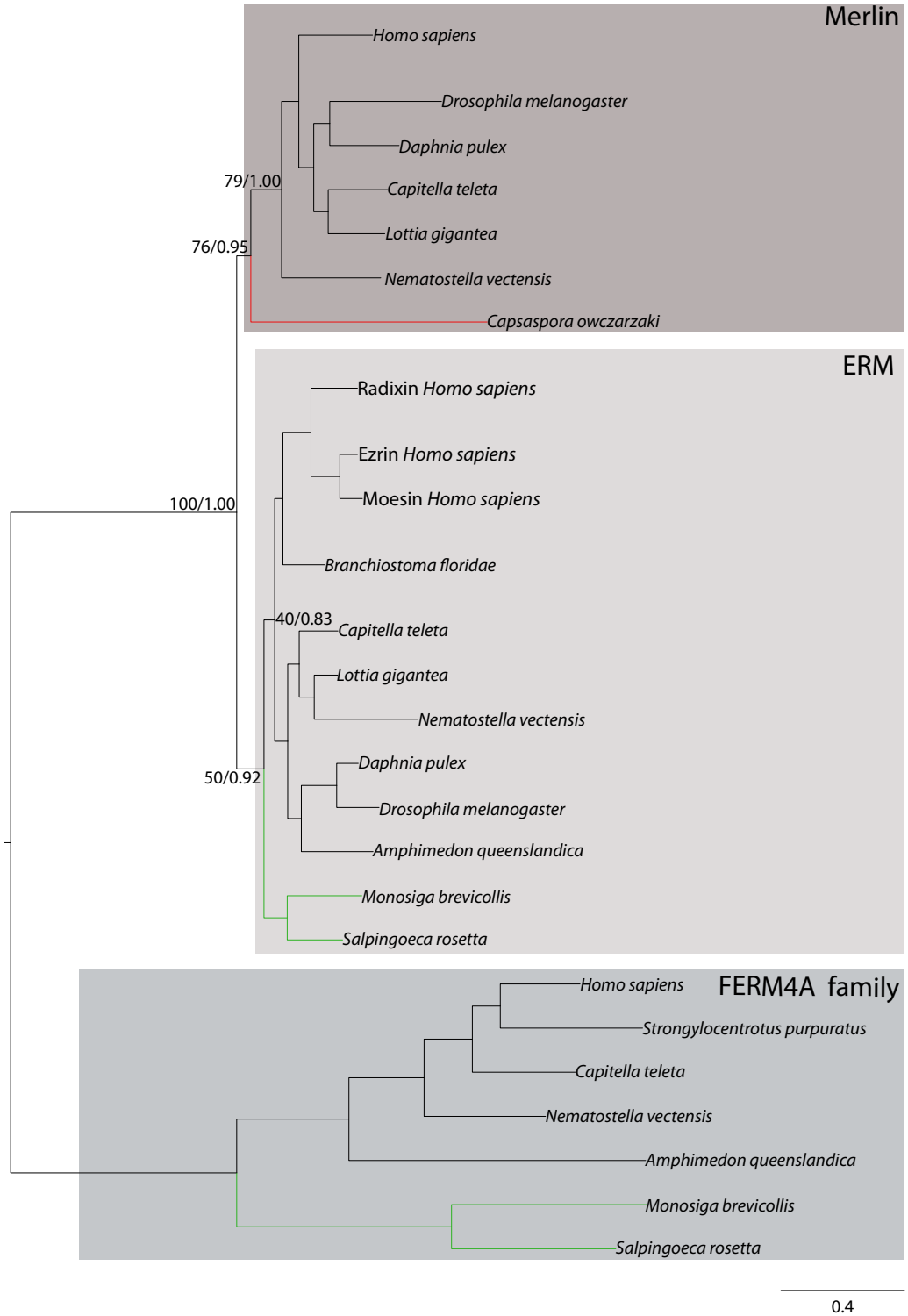


Figure S1. Maximum likelihood phylogenetic tree of FERM-domain containing genes, including ERM, Merlin and FERM4A. Alignment is based on the FERM domain and contains 408 amino acid positions. The tree was inferred by RAxML using the WAG+Γ +I model of evolution. The tree is rooted using FERM4A as an outgroup. Statistical support was obtained by RAxML with 100 bootstrap replicates (BV) and Bayesian Posterior Probabilities as inferred with MrBayes. PP. Both values are shown on key branches.

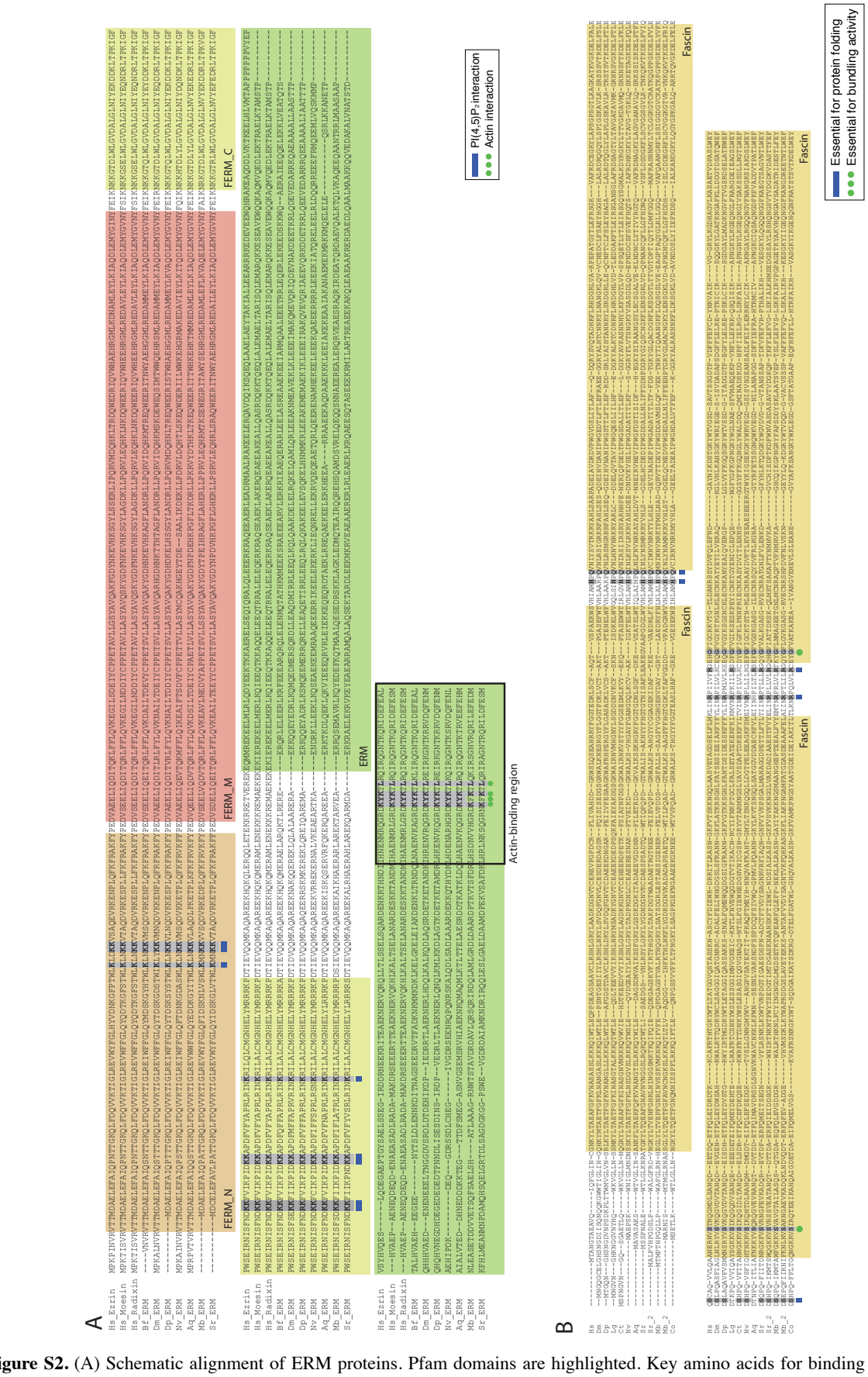


Figure S2. (A) Schematic alignment of ERM proteins. Pfam domains are highlighted. Key amino acids for binding to the membrane lipid phosphatidylinositol 4,5-bisphosphate ($PI(4,5)P_2$), involved in regulation, are highlighted in blue. Key amino acids for actin-binding are highlighted in green. Key amino acids after Turunen et al. 1994 and Niggli et al. 2007. Taxa include Aq (*Amphimedon queenslandica*), Bf (*Branchiostoma floridae*), Co (*Capsaspora owczarzaki*), Dm (*Drosophila melanogaster*), Dp (*Daphnia pulex*), Hs (*Homo sapiens*), Mb (*Monosiga brevicollis*), Nv (*Nematostella vectensis*) and Sr (*Salpingoeca rosetta*). (B) Schematic alignment of fascin proteins, showing the four consecutive fascin protein domains. Key amino acids for protein folding are highlighted in blue. Key amino acids for actin-bundling activity are highlighted in green. Key amino acids after Sedeh et al. (2010) and Zanet et al. (2012).Taxa include Aq (*Amphimedon queenslandica*), Co (*Capsaspora owczarzaki*), Ct (*Capitella teleta*), Dm (*Drosophila melanogaster*), Dp (*Daphnia pulex*), Hs (*Homo sapiens*), Lg (*Lottia gigantea*), Mb (*Monosiga brevicollis*), Nv (*Nematostella vectensis*) and Sr (*Salpingoeca rosetta*).

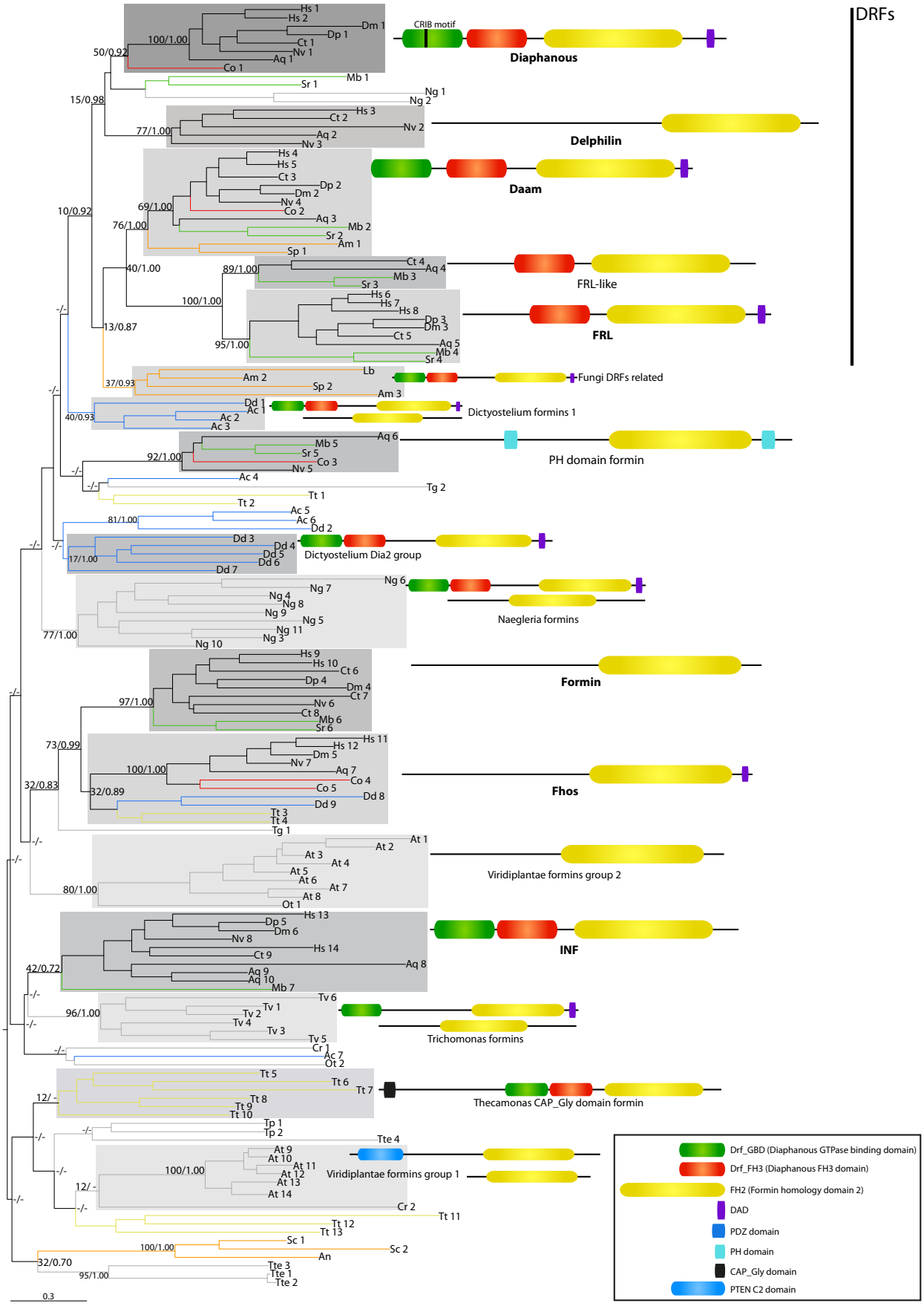


Figure S3. Maximum likelihood phylogenetic tree of formin proteins. The alignment is based on the FH2 domain and comprises 308 amino acid positions. The tree was inferred by RAxML using the WAG+Γ+I model of evolution. The tree is rooted using the midpoint-root option. Statistical support was obtained by RAxML with 100 bootstrap replicates (BV) and by BPP. Both values are shown on key branches. Domain architectures for each class are shown.Taxa include Ac (*Acanthamoeba castellanii*), Am (*Allomyces macrogynus*), Aq (*Amphimedon queenslandica*), At (*Arabidopsis thaliana*), Co (*Capsaspora owczarzaki*), Cr (*Chlamydomonas reindhartii*), Ct (*Capitella teleta*), Dd (*Dictyostelium discoideum*), Dm (*Drosophila melanogaster*), Dp (*Daphnia pulex*), Hs (*Homo sapiens*), Lb (*Laccaria bicolor*), Lg (*Lottia gigantea*), Mb (*Monosiga brevicollis*), Ng (*Naegleria gruberi*), Nv (*Nematostella vectensis*), Ot (*Ostreococcus tauri*), Sc (*Saccharomyces cerevisiae*), Sp (*Spizellomyces punctatus*), Sr (*Salpingoeca rosetta*), Tg (*Toxoplasma gondii*), Tp (*Thalassiosira pseudonana*), Tte (*Tetrahymena thermophila*),Tt (*Thecamonas trahens*), Tv (*Trichomonas vaginalis*).

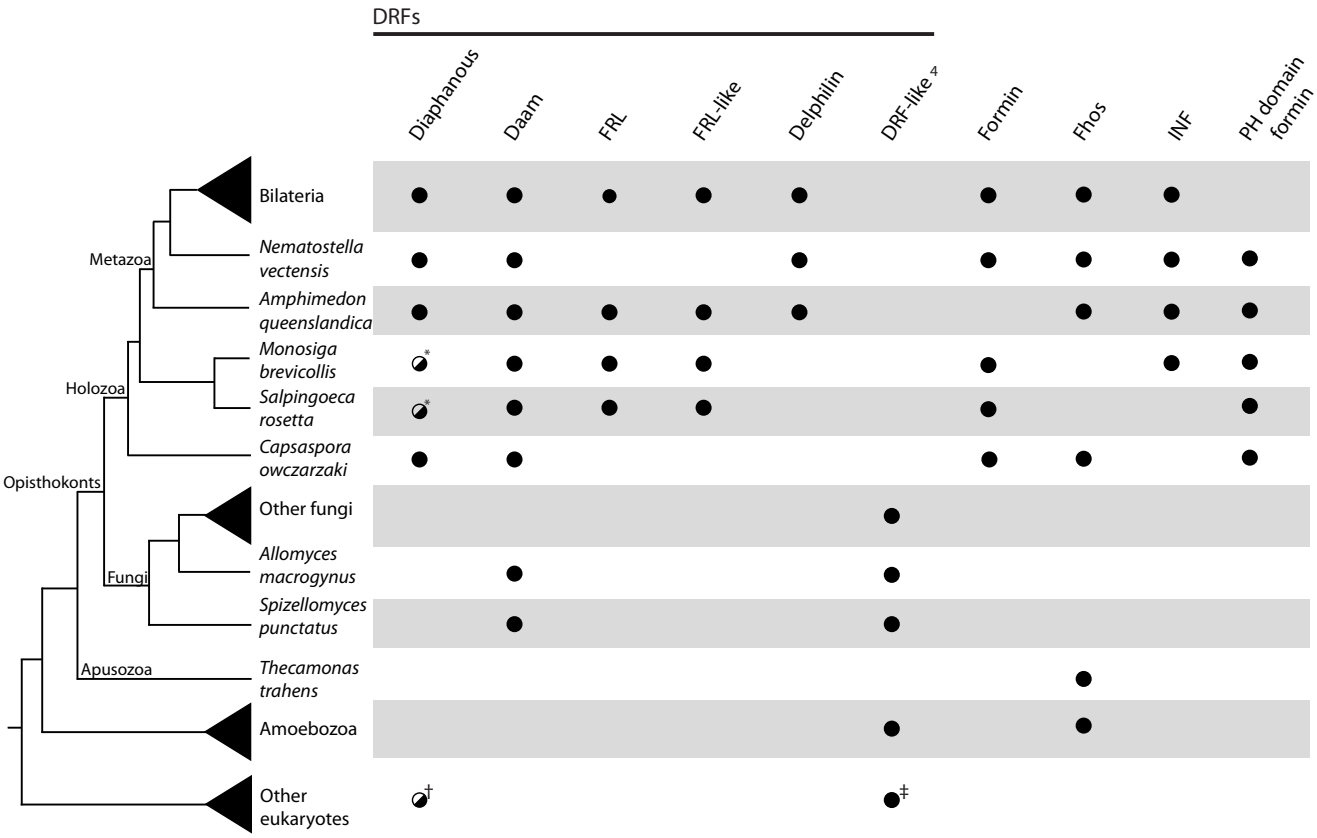


Figure S4. Schematic representation of the eukaryotic tree of life showing the distribution of the different formin classes. A black dot indicates the presence of clear homologs, whereas a dashed dot indicates the presence of putative or degenerate homologs. Absence of a dot indicates that an homolog is lacking in that taxon. *Choanoflagellate Diaphanous homologs lack a clear CRIB domain. †Only two *Naegleria gruberi* proteins cluster with Diaphanous, but they do not have a canonical DRF structure (Drf_GBD-Drf_FH3-FH2-DAD). ‡Among other eukaryotes, only the excavates *Naegleria gruberi* and *Trichomonas vaginalis* have bona fide DRF proteins. [4]Defined as having the Drf_GBD-Drf_FH3-FH2-DAD domain architecture.

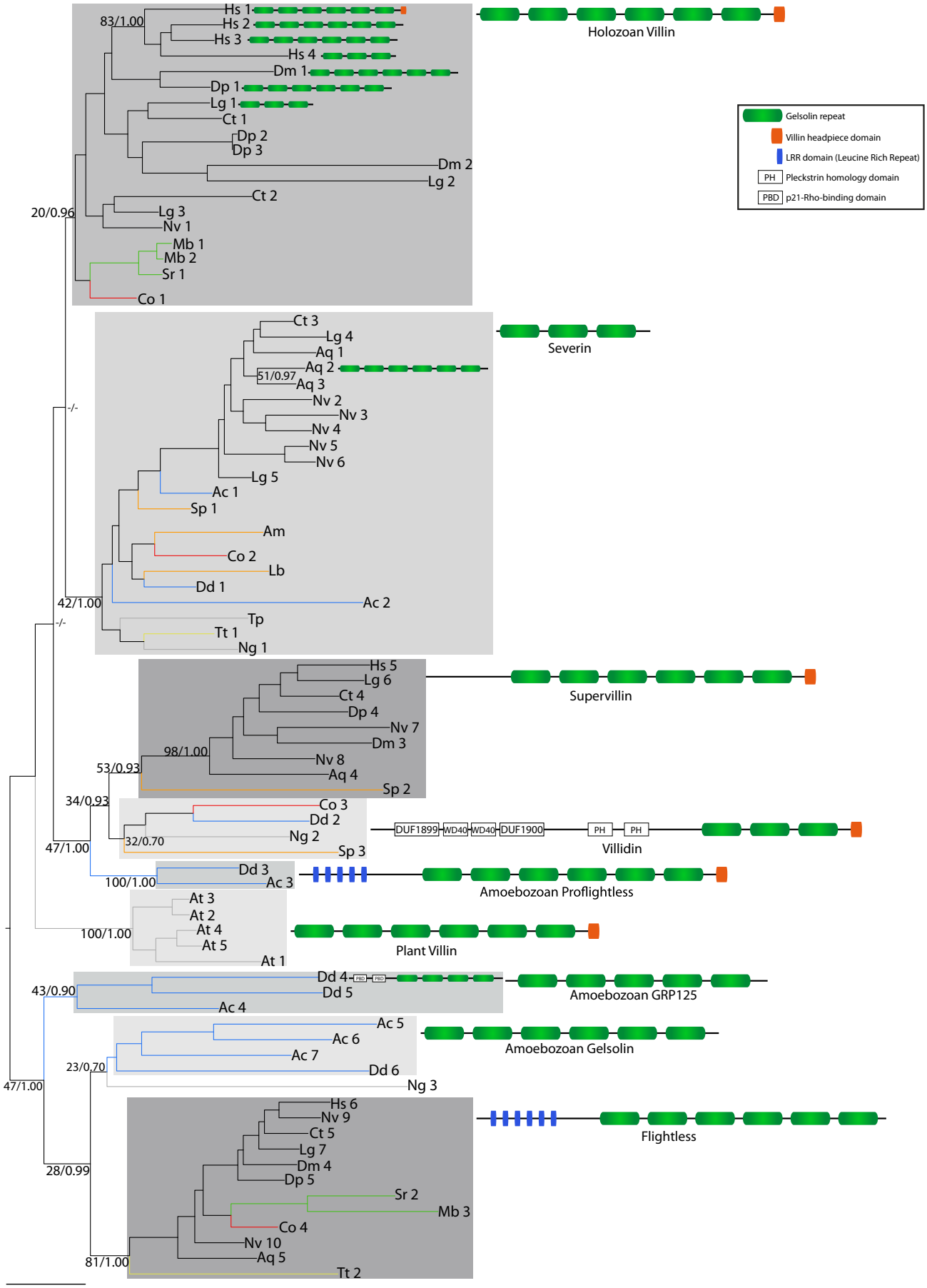


Figure S5. ML of gelsolin proteins. The alignment is based on the gelsolin domains and comprises 270 amino acid positions. The tree was inferred by RAxMlLusing the WAG+Γ+I model of evolution. The tree is rooted using the midpoint-root option. Statistical support was obtained by RAxML with 100 bootstrap replicates (BV) and BPP. Both values are shown on key branches. Domain architectures for each class are shown.Taxa include Ac (*Acanthamoeba castellanii*), Am (*Allomyces macrogynus*), Aq (*Amphimedon queenslandica*), At (*Arabidopsis thaliana*), Co (*Capsaspora owczarzaki*), Ct (*Capitella teleta*), Dd (*Dictyostelium discoideum*), Dm (*Drosophila melanogaster*), Dp (*Daphnia pulex*), Hs (*Homo sapiens*), Lg (*Lottia gigantea*), Mb (*Monosiga brevicollis*), Ng (*Naegleria gruberi*), Nv (*Nematostella vectensis*), Sp (*Spizellomyces punctatus*), Sr (*Salpingoeca rosetta*), Tp (*Thalassiosira pseudonana*), Tt (*Thecamonas trahens*).

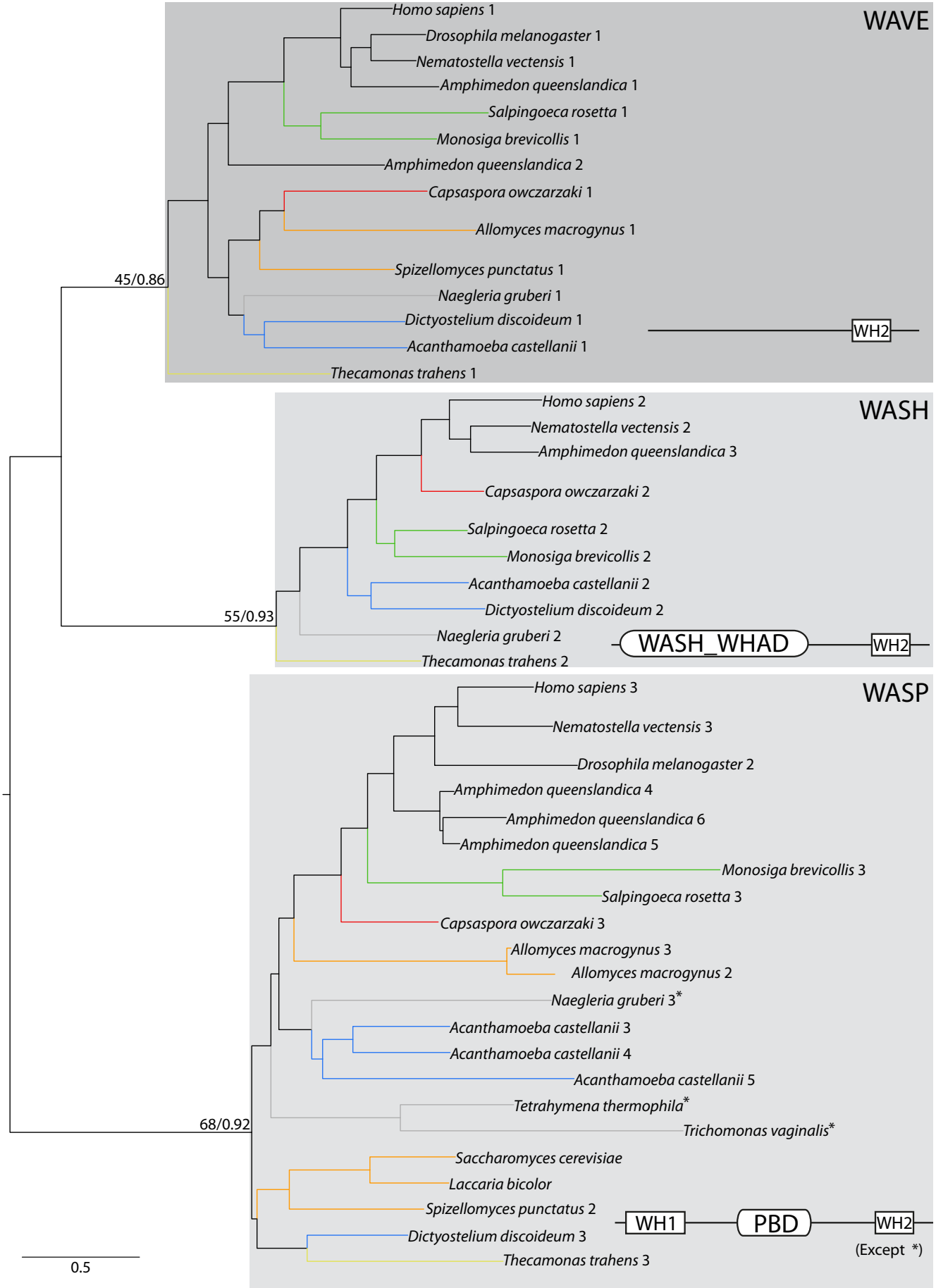


Figure S6. Maximum likelihood phylogenetic tree of WASP proteins. The alignment comprises 139 amino acid positions. The tree was inferred by RAxML using the WAG+Γ+I model of evolution. The tree is rooted using the midpoint-root option. Statistical support was obtained by RAxML with 100 bootstrap replicates (BV) and BPP. Both values are shown on key branches. Domain architectures for each class are shown.

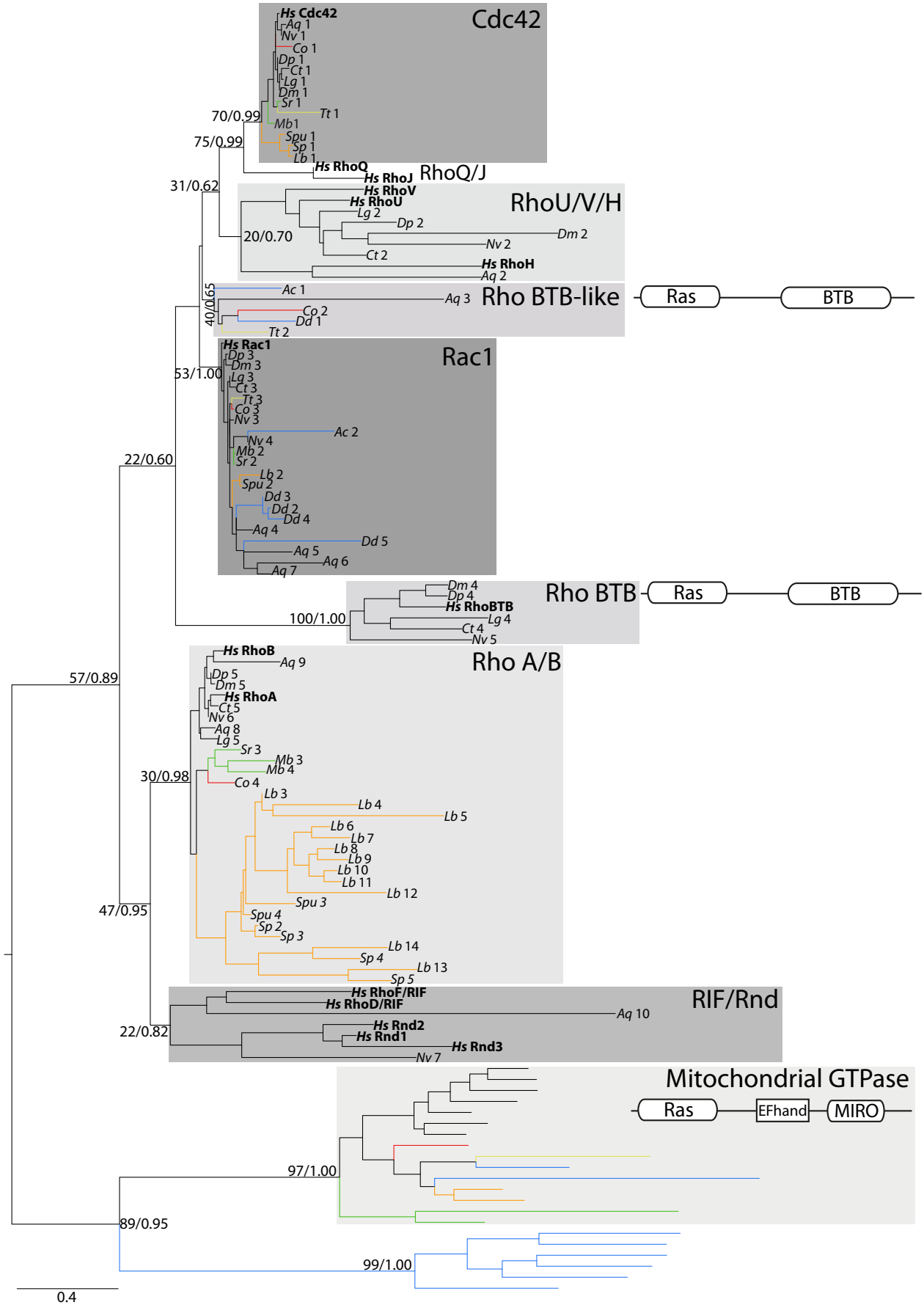


Figure S7. Maximum likelihood phylogenetic tree of RhoGTPase proteins from unikont taxa. The alignment is based on the Ras domain and comprises 153 amino acid positions. The tree was inferred by RAxML using the WAG+Γ+I model of evolution. The tree is rooted using the MIRO GTPases as an outgroup. Statistical support was obtained by RAxML with 100 bootstrap replicates (BV) and BPP. Both values are shown on key branches. Domain architectures for each class are shown. Taxa include Ac (*Acanthamoeba castellanii*), Am (*Allomyces macrogynus*), Aq (*Amphimedon queenslandica*), Co (*Capsaspora owczarzaki*), Ct (*Capitella teleta*), Dd (*Dictyostelium discoideum*), Dm (*Drosophila melanogaster*), Dp (*Daphnia pulex*), Hs (*Homo sapiens*), Lb (*Laccaria bicolor*), Lg (*Lottia gigantea*), Mb (*Monosiga brevicollis*), Nv (*Nematostella vectensis*), Sc (*Saccharomyces cerevisiae*), Sp (*Spizellomyces punctatus*), Sr (*Salpingoeca rosetta*), Tt (*Thecamonas trahens*).

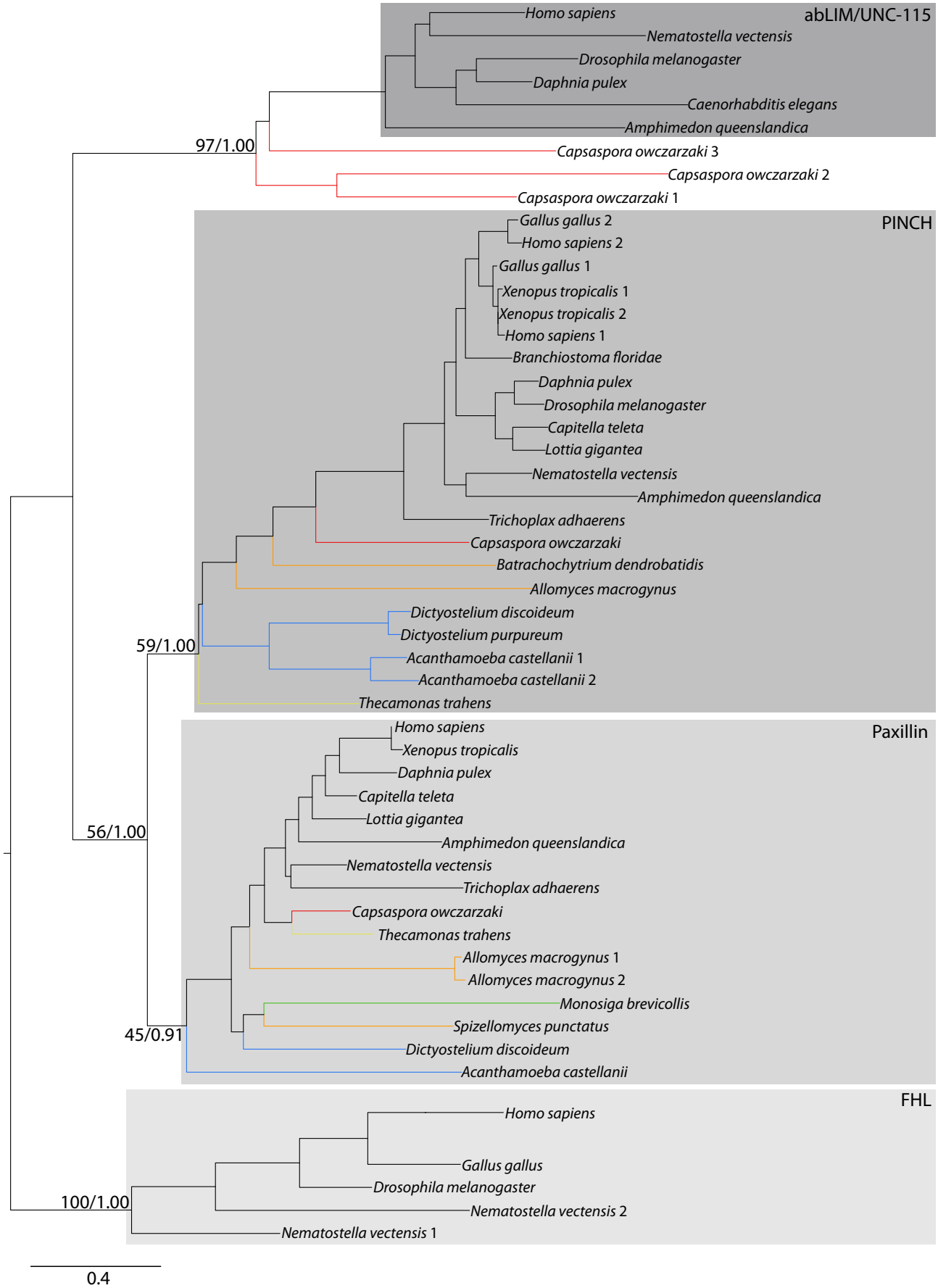


Figure S8. Maximum likelihood phylogenetic tree of abLIM and other LIM-domain containing proteins. The alignment comprises 239 amino acid positions. The tree was inferred by RAxML using the WAG+Γ+I model of evolution. The tree is rooted using the midpoint-root option. Statistical support was obtained by RAxML with 100 bootstrap replicates (BV) and BPP. Both values are shown on key branches.

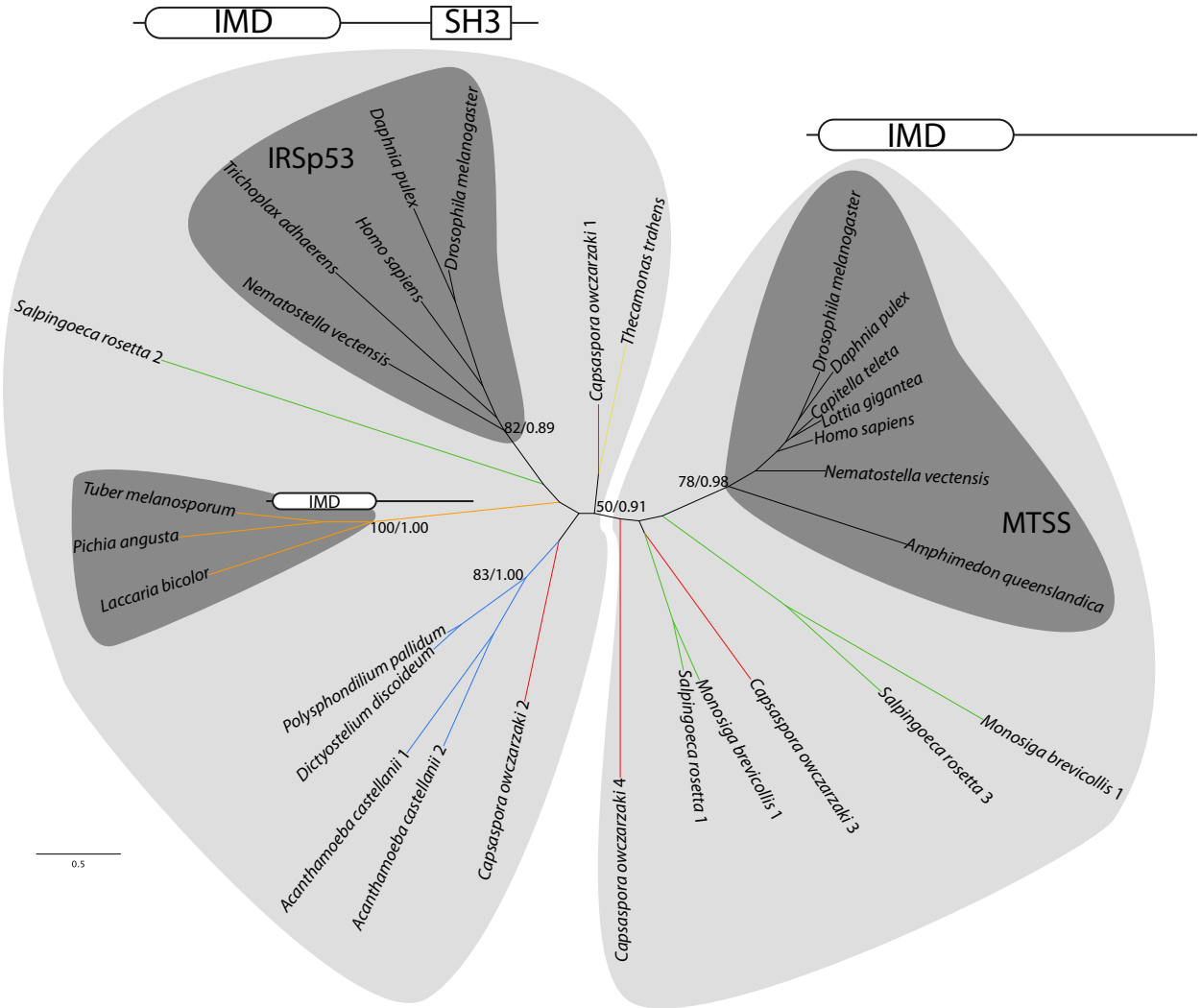


Figure S9. Maximum likelihood phylogenetic tree of IMD-domain containing proteins. The alignment is based on the IMD domain and comprises 180 amino acid positions. The tree was inferred by RAxML using the WAG+Γ+I model of evolution. The tree is rooted using the midpoint-root option. Statistical support was obtained by RAxML with 100 bootstrap replicates (BV) and BPP. Both values are shown on key branches. Domain architectures for each class are shown.

Results R7

Transcriptome remodelling during aggregative multicellularity in a close unicellular relative of Metazoa.

RESUM ARTICLE R7: Remodelatge del transcriptoma durant la multicel·lularitat agregativa en un parent unicel·lular proper dels metazous

Capsaspora owczarzaki és un dels parents unicel·lulars més propers dels animals. No obstant, a pesar d'aquesta privilegiada posició filogenètica per a entendre l'origen de la multicel·lularitat animal, la seva biologia cel·lular i molecular són en gran part desconegudes. En aquest treball, descrivim el cicle vital de *Capsaspora owczarzaki* i com aquest cicle és regulat transcripcionalment a nivell d'expressió gènica i també d'*splicing* alternatiu. Demostrem l'existència d'un estadi agregatiu multicel·lular, el primer d'aquest tipus en un parent proper dels animals. Les transicions cap a i des de aquest i altres estadis estan estrictament regulades a nivell transcripcional, afectant categories funcionals clau i suggerint que la maquinària molecular de la multicel·lularitat animal (per exemple, la maquinària d'integrina) és usada tan en un context de multicel·lularitat clonal com en un de multicel·lularitat agregativa. També trobem centenars de casos de retenció d'introns regulats de forma específica en cada estadi. Els nostres resultats, doncs, amplien la diversitat coneguda de tipus cel·lulars i comportaments cel·lulars en els parents més propers dels animals, proporcionant un nou context funcional on la maquinària molecular dels animals podria haver evolucionat, i també mostren una estricta regulació de la diferenciació temporal en un organisme unicel·lular.

Transcriptome remodelling during aggregative multicellularity in a close unicellular relative of Metazoa

Arnau Sebé-Pedrós[1,4], Manuel Irimia[2], Javier del Campo[1], Helena Parra-Acero[1], Carsten Russ[3], Brian J. Haas[3], Ben J.Blencowe[2], Chad Nusbaum[3] & Iñaki Ruiz-Trillo[1,4,5*]

[1]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta, 37-49, 08003 Barcelona, Spain. [2]Banting and Best Department of Medical Research, Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada. [3]Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [4]Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Avinguda Diagonal 643, 08028 Barcelona, Spain. [5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. *Correspondence to: Iñaki Ruiz-Trillo (inaki.ruiz@multicellgenome.org)

The origin of metazoans from their unicellular ancestor represents one of the most important evolutionary events in the history of life. Recent genomic data from both metazoans and their closest unicellular relatives have begun to clarify the genomic basis of this transition. However, little is known about the molecular and cell biology of some of the closest unicellular relatives of Metazoa, such as the filasterean *Capsaspora owczarzaki*. Here, we describe *C. owczarzaki*'s life cycle and its complex transcriptomic regulation at the level of gene expression and alternative splicing. We show the existence of an aggregative multicellular stage, the first such instance in a close relative of metazoan. The transitions in and out of this and other stages are tightly regulated at the transcriptomic level, impacting key conserved gene functional categories. We also found hundreds of intron retention events regulated in a stage-specific manner. Our results expand our knowledge on the diversity of cell types and behaviours found in the closest animal relatives and show how tight temporal regulation of the unicellular life cycle could have evolved into spatial cell differentiation in metazoans.

Living organisms emerge from the integration of multiple levels of organisation. These levels are shaped by both physicochemical constraints and historical circumstances, the later being more important in more complex systems (Jacob 1977). Therefore, it is important to identify the phylogenetic inertia (*sensu* Burt 2001) imposed by the raw starting material in order to properly understand major evolutionary transitions, such as the origin of metazoan multicellularity (Knoll 2011). Examination of both the genetic repertoire (King 2004; Ruiz-Trillo et al. 2007; Rokas 2008) and the cell types present in the immediate unicellular ancestors of metazoans can thus provide insights into this evolutionary transition, as they reveal the historical constraints in early metazoan evolution.

The selective advantages of multicellularity have long been suggested (Bonner 1998; Grosberg and Strathmann 2007) and multi-level selection theory has proposed that complex multicellular life is more likely to arise through clonal development than by simple aggregation of genetically diverse cells. In the later case, intraorganismal competition creates such strong fitness challenges that the aggregate is predicted to be evolutionary unstable (Queller 2000; Grosberg and Strathmann 2007; Michod 2007; Aanen et al. 2008; Newman 2012). Indeed, the eukaryotic lineages that attained the most complex multicellular lifestyles (i.e., green algae and plants, brown and red algae, and metazoans) did so through clonal cell division (Grosberg and Strathmann 2007). In contrast, the cell aggregation found in a few eukaryotes represents transient stages of their life cycle. This is the case of dictyostelids (Amoebozoa) (Schaap 2011), acrasid amoebas (Heterolobosea, Discicristata, Discoba) (Brown et al. 2011; Adl et al. 2012), *Guttulinopsis vulgaris* (Cercozoa, Rhizaria) (Brown et al. 2012), the genus *Sorogena* (Ciliata, Alveolata) (Lasek-Nesselquist and Katz 2001), the nucleariid *Fonticula alba* (Nucleariidae, Opisthokonta) (Brown et al. 2009) and the genus *Sorodiplophrys* (Labyrinthulomycetes, Heterokonta) (Dykstra and Olive 1975). Within the opisthokont clade, which comprises Metazoa, Fungi and their unicellular relatives (Cavalier-Smith 2003; Steenkamp et al. 2006; Ruiz-Trillo et al. 2008), so far only a single taxon has been described to have aggregative behaviour. This is the case of the nucleariid *F. alba*, a sister group of Fungi (Brown et al. 2009). On the other hand, the colonies of both choanoflagellates and ichthyosporeans, both close relatives of Metazoa (Ruiz-Trillo et al. 2008; Shalchian-Tabrizi et al. 2008; Torruella et al. 2012), show clonal development (Jøstensen et al. 2002; Marshall et al. 2008; Dayel et al. 2011; Suga and Ruiz-Trillo 2013). Within metazoans, despite their general clonal development, some cells show aggregative behaviours; for example, mesenchymal (O'Shea 1987) and germ line cells

(Savage and Danilchik 1993) during development, sponge cells after cell dissociation (Wilson 1907) and arthropod blood cells through active amoeboid movement (Loeb 1903; Loeb 1921).

We describe here for the first time aggregative behaviour in a close unicellular relative of Metazoa, the filasterean *Capsaspora owczarzaki*, and explore how this cell aggregation is regulated at the transcriptomic level. *Capsaspora owczarzaki* belongs to the clade Filasterea, the sister-group of Metazoa and choanoflagellates (Ruiz-Trillo et al. 2008; Shalchian-Tabrizi et al. 2008; Paps et al. 2012; Torruella et al. 2012). Although isolated decades ago and described as an endosymbiont of a fresh-water snail (Stibbs et al. 1979; Owczarzak et al. 1980), little is known about its basic cell biology. We analysed *C. owczarzaki*'s life cycle and its regulation using electron microscopy, flow cytometry and next generation transcriptome sequencing (RNA-Seq). Through these analyses, we demonstrate that the molecular toolkit for multicellularity, many of whose elements predate the origin of Metazoa (King et al. 2008; Sebé-Pedrós et al. 2010), can function either in aggregative or in clonal multicellularity and in different phylogenetic contexts, as previously hypothesised (Newman 2012).

Under initial culture conditions, *C. owczarzaki* differentiates into an amoeba that crawls over the substrate (Movie S1), surveying the environment with its long filopodia. At this stage active DNA replication occurs (with >10% of the cells in S-phase) and within 48 hours *C. owczarzaki* cells enter in an exponential growth phase (Fig. 1A, Fig. S1). At this moment, in a cell-density dependent manner, cells start to detach from the surface and begin to retract their filopodia and encyst (Fig. 1F), a process which (after an initial period where floating cells still actively divide) ultimately leads to a cystic floating stage (Fig. 1E) in which cell division is highly reduced (Fig. S1). After eight days, no attached amoebas are left and the growth is stabilised (Fig. S1). An alternative path to this process is the active formation of cell aggregates (Movie S2). Cells attach to each other and progressively produce cohesive extracellular material that joins them (Fig. 1B) until a compact cell aggregate, in which cells no longer bear filopodia, is formed (Fig. 1C). Transmission electron microscopy demonstrates the presence of a thick, unstructured, extracellular material (Fig. 1D) within the *C. owczarzaki* aggregates and that cells are not in direct contact. Aggregates have homogeneous sizes, and their formation occurs randomly in normal culture conditions. Nevertheless, the process can be readily induced by agitation.

To assess whether the formation of these distinct cell stages are differentially regulated at the transcriptomic level, we compared the transcriptomic profiles obtained by RNA-Seq of the three cell stages described: adherent (filopodiated amoeba), aggregative (formation of cell aggregates) and floating (cysts) (Fig. 2). A total of 4486 genes showed statistically significant differential regulation (see Methods) in at least one of the pair-wise comparisons: 1354 between adherent and aggregate stages, 3227 between adherent and floating stages, and 3096 between aggregate and floating stages. Moreover, when performing one-versus-all comparisons, each cell stage had a specific transcriptomic profile (Fig. 2), indicating that the cycle is tightly regulated at the level of gene expression. Using both pairwise and one-versus-all comparisons, we identified significantly enriched Gene Ontology (GO) categories (Fig. 2) and Pfam domains (Fig. S2) in each set of differentially transcribed genes (both up and down-regulated), compared with the total population of 8637 genes in *C. owzarzaki*'s genome ($p<0.01$ for each significant category; Fisher's exact test). Genes up-regulated in the adherent stage were enriched in signalling functions, such as tyrosine kinase activity and G-protein coupled receptor activity, as well as in transcription factors, especially of the Basic Leucine Zipper Domain (bZIP) superfamily. In addition, genes involved in protein synthesis and DNA replication were significantly upregulated in this stage, consistent with the rapid cell proliferation observed by flow cytometry (Fig. S1), and suggesting a high metabolic rate. In contrast, the aggregative stage showed strong upregulation of the components of the integrin adhesome and other cell-adhesion related proteins (Fig 2), such as the LamininG domain-containing protein CAOG_07351 (which contains a N-terminal signal peptide and is therefore likely to be secreted) (Fig. S3). This suggests that the integrin adhesome of *C. owzarzaki* (Sebé-Pedrós et al. 2010) may play a crucial role in the formation of the aggregates. We also observed an upregulation of genes involved in tubulin cytoskeleton machinery (e.g. kinesins) in this stage. Finally, cystic cells had a very distinct transcriptomic profile compared to adherent and aggregative cells (Fig. S4A). We found that genes involved in myosin transport, translation, DNA replication and metabolic activities (especially mitochondrial energy production) were significantly downregulated in the cystic stage. On the other hand, genes involved in vesicle transport and autophagy – recycling of intracellular components that is usually triggered by starvation or other adverse conditions, and plays an essential role in cell survival in many eukaryotes (Kiel 2010) – are significantly enriched in this stage. Moreover, we found that genes involved in

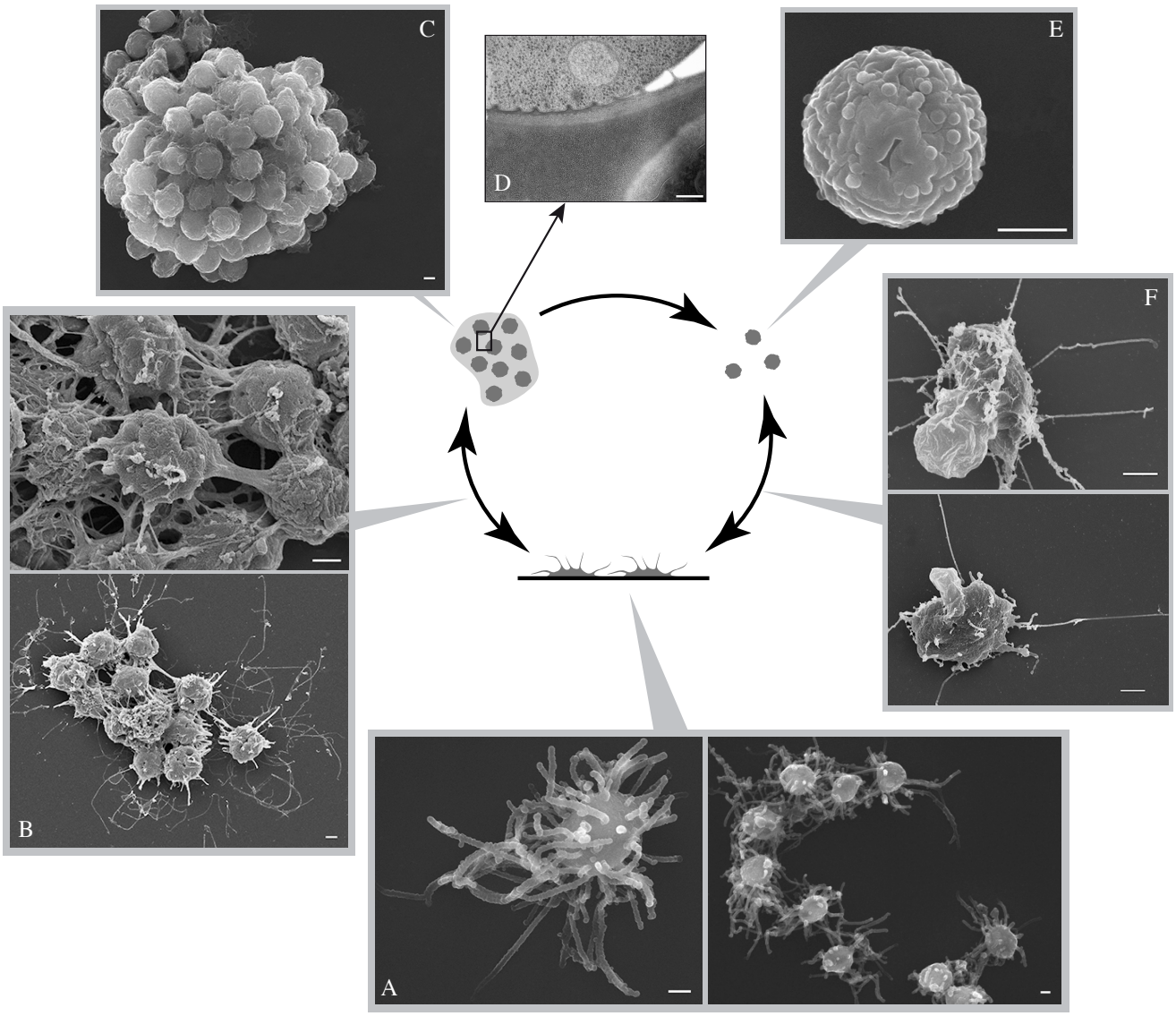


Figure 1. *C. owzarzaki* life cycle. A) Adherent stage cells, amoebas with long filopodia; B) Transition from adherent to aggregative stage, cells attach to each other and an extracellular matrix appears; C) Mature aggregate; D) TEM micrography showing to adjacent cells in the aggregate with the matrix in-between; E) Floating stage cells are rounded cysts without filopodia and slightly smaller than adherent cells; F) Transition from adherent to floating stage, cells retract filopodia. Arrows indicate the observed stage inter-conversions. All scale bars= 1 μm, except D (200nm).

the ubiquitin pathway (e.g., UQ_con, zf-RING2 and Cullin domains (Fig. S2)) and in synaptic cell-cell communication, such as SNARE, synaprobrevin and syntaxin (Bennett et al. 1993; Lang and Jahn 2008), as well as some specific transcription factor families (e.g., bHLH transcription factors), are also significantly upregulated in the cystic cells. Altogether, these results suggest that major cytosolic rearrangement and protein turnover occur at this stage.

We next studied the transcriptomic diversity generated by differential processing of introns and exons (i.e. alternative splicing, AS), and the extent to which it is regulated across the different *C. owczarzaki* cellular stages. First, we examined the presence of exon skipping, the hallmark of AS in metazoans (Nilsen and Graveley 2010). By mapping RNA-Seq reads to all possible forward exon-exon junction combinations (see Methods), we identified 191 cassette exons with inclusion levels lower than 95% in at least one stage (39 exons showed inclusion levels lower than 85%). Although most of these AS events are likely to correspond to cases of inaccurate splicing, we identified 29 exons that showed more than 15% inclusion differences in pairwise comparisons between cell stages, usually skipped only in the adherent stage (Table S1). RT-PCRs confirmed skipping in 7 out of 8 cases (Fig. 3A and S5).

We also investigated intron retention (IR), the most widespread form of AS in non-metazoan eukaryotes (McGuire et al. 2008). By mapping reads to intron-exon and exon-exon junctions, we could identify thousands of highly retained introns. For example,

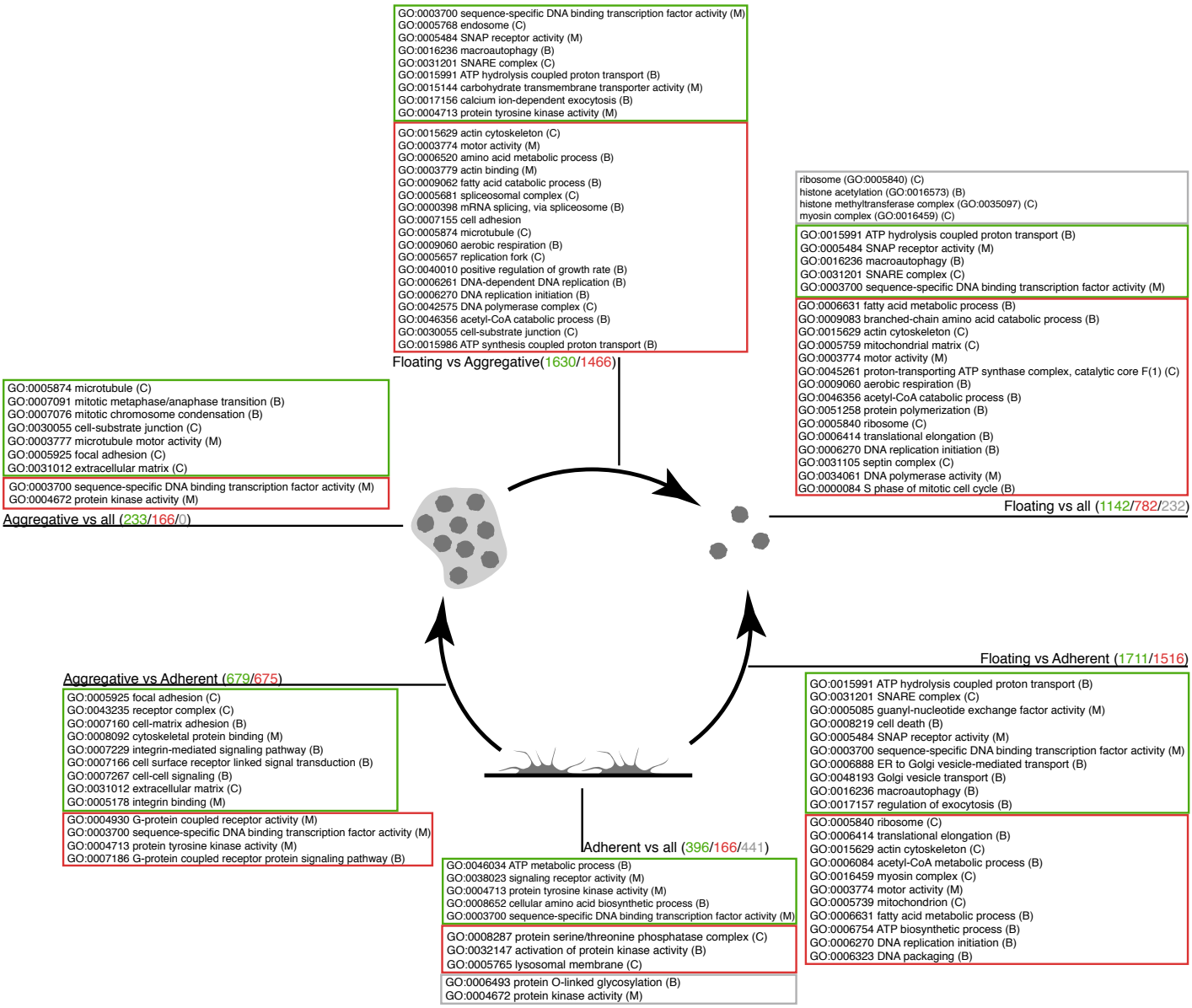


Figure 2. GO enrichment in sets of differentially expressed genes. Pairwise (Aggregative vs Adherent, Floating vs Aggregative and Floating vs Adherent) and one-versus-all comparisons are indicated. The significantly overrepresented GO categories (see Methods) are shown for sets of overexpressed (green) and downregulated (red) genes and for genes with differential intron retention (grey). The number of genes included in each set is indicated with the same colour-code.

2986 out of the 25677 (11.6%) introns with enough mapping RNA-Seq reads in the three stages (see Methods) showed ≥20% inclusion in at least one stage, and approximately a third of the genes had at least one such IR event. Interestingly, however, the three cellular stages show remarkable differences in the extent of IR (Fig. 3B). While adherent and cystic cells have relatively high fractions of retained introns (~8% and ~5% of introns are included in ≥20% of the transcripts, respectively), the aggregative multicellular stage showed extremely low IR (<1% of introns had inclusion ≥20%), a difference highly consistent across three biological replicates. These differences suggest that IR may be differentially regulated between the different cell stages. Consistently, among genes expressed in all three stages, we identified 797 (in 441 genes) and 259 (in 232 genes) differentially retained introns (dRIs; at least 25% higher retention than in the other two stages) in adherent and floating stages, respectively, and none in aggregative. Most of tested cases (12 out of 15, 80%) were validated by RT-PCR (Fig. 3 and S6). GO enrichment analysis for the two sets of dRIs showed distinct functional enrichments (Fig. 2), suggesting that IR plays different roles in the adherent and floating stages. Low fractions of read-through introns (with sizes multiple of three and no in-frame STOP codons) suggest that these dRIs may be acting by reducing the level of properly spliced mRNA that is translated into protein. Strikingly, we found that a large fraction of multi-intronic genes with dRIs retained multiple of their introns in the same stage-specific manner, particularly in the adherent stage. Over 73% and 29% of multi-intronic genes with

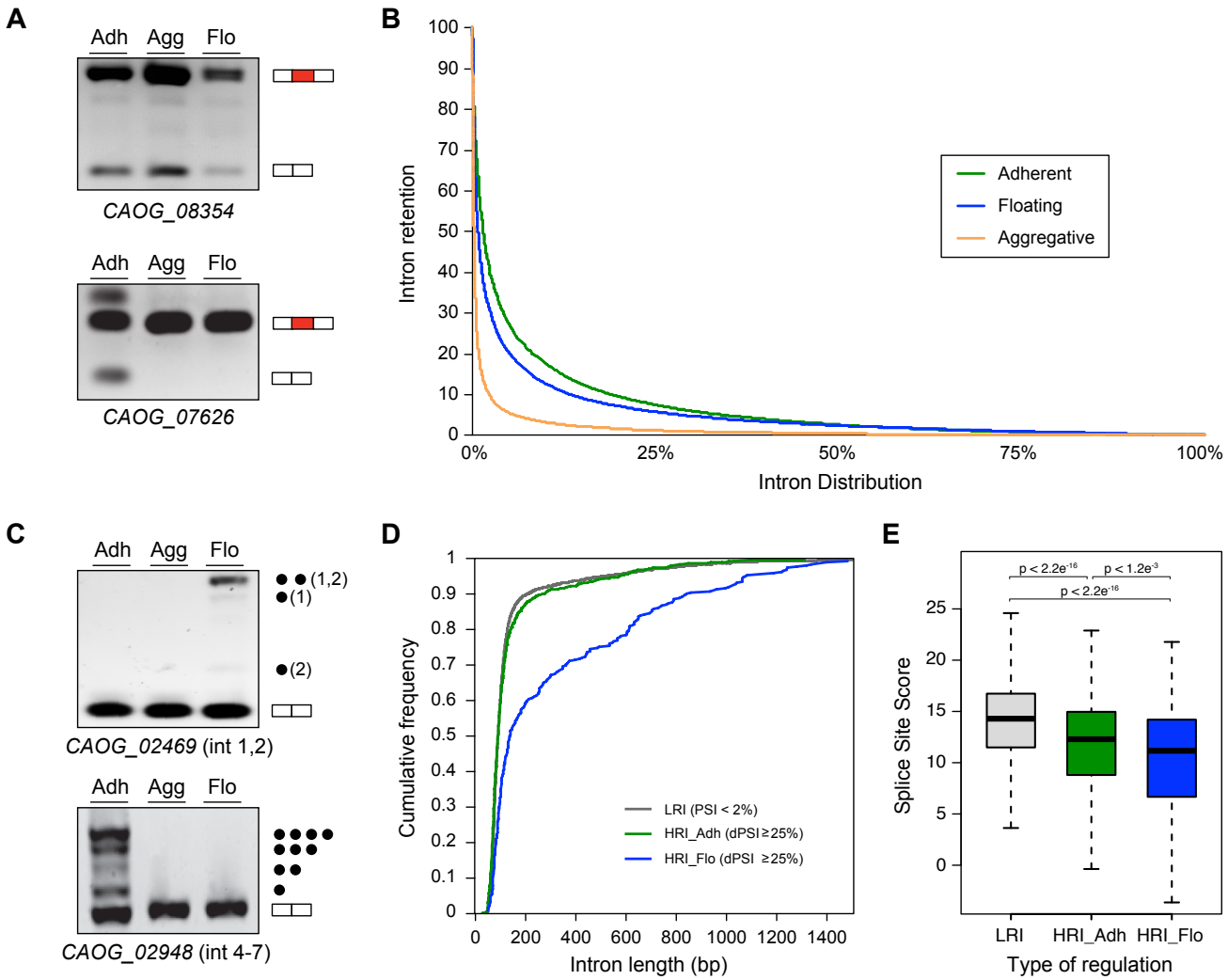


Figure 3. Alternative splicing. A) Examples of some exon skipping events. B) Plot of percentage of intron inclusion by intron in rank order for the three studied cellular stages. Adherent (green) and floating (blue) stages showed much higher IR levels than the aggregative stage (orange). C) Examples of stage-specific IR of multiple introns within the same gene (above: introns 1 and 2 of CAOG_02469 in floating; below: introns 4 to 7 of CAOG_02948 in adherent). D) Intron length distributions for differentially retained introns in floating (blue) or adherent (green), and lowly retained introns (grey). E) Box-plots showing splice site score distributions in the same three sets of introns. Floating (blue) has the weakest splice site signals.

adherent- or floating-specific dRIs, respectively, had at least another highly retained intron in that stage, and 22% and 5% showed evidence of high retention for all its introns. This pattern was consistent across the three replicates, and specific for a given gene compared to its neighbours, ruling out partial genome contamination. Furthermore, RT-PCR analyses and mate information from paired-end reads (see Methods) suggest that multiple IR often occurs in a combinatorial manner (Fig. 3C and S6), increasing the impact of IR-mediated mRNA downregulation.

Finally, we looked at different features of differentially retained introns that may shed light on their stage-specific regulation. First, we compared intron lengths. Strikingly, while adherent-specific dRIs have a similar length distribution than lowly retained introns (inclusion lower than 2% across all stages), floating stage-specific introns were significantly longer (Figure 3D, p-value=$1.7e^{-14}$ Wilcoxon rank sum test). In line with this observation, genome-wide average IR increased steadily for longer introns only in the floating stage (Figure S7). Furthermore, floating stage-specific retained introns harboured significantly weaker canonical 5' and 3' splice site signals than other intron sets (Figure 3E, $p < 0.0013$ Wilcoxon rank sum test for all comparisons). In total, these data suggest that differential intron retention in the floating stage may be achieved by mis-splicing of a subset of suboptimal introns (i.e. long and with weak splice sites) that are properly spliced in the other cell stages. In the case of the adherent-specific dRIs, analyses of sequence motif enrichment and comparisons of expression levels of splicing factors suggest a potential role for an Elav-like ortholog in their regulation (Fig. S8), which awaits to be functionally tested.

In summary, we found that stage-regulated AS patterns in *C. owczarzaki* are not metazoan-like, but more similar to most characterised unicellular eukaryotes and plants, dominated by extensive IR with only a few cases of exon skipping. A similar pattern has been shown in choanoflagellates (Westbrook 2011), further supporting that metazoans and their unicellular relatives have contrasted modes of AS. This suggests that an important step in the origin of metazoans may have been the transition from IR-based to exon skipping dominated AS.

The aggregative multicellularity described here for *C. owczarzaki* redefines our current view of the evolution of metazoan multicellularity. To date, it was thought that simple clonal multicellularity was the only type present among unicellular relatives of Metazoa, for example the colony formation in choanoflagellates and the sporangia formation by hypertrophic syncytial growth in ichthyosporeans. The presence of aggregative cell behaviour in *C. owczarzaki* therefore expands the potential starting raw material available for the evolution of metazoan multicellularity. Our data, together with the knowledge in choanoflagellate and ichthyosporean cell biology, show that the toolkit for multicellularity (i.e., genes involved in cell adhesion, cell differentiation and cell signalling) can equally be deployed in different phylogenetic contexts both for clonal and for aggregative multicellularity, even if only one (clonal), would have eventually given rise to stable multicellular forms (Hernández-Hernández et al. 2012).

The highly regulated life cycle of *C. owczarzaki* (both in terms of mRNA levels and AS) reported here is in line with its complex genetic toolkit (refs). This complex repertoire of genes and functions suggest that metazoans may have integrated different cell types/behaviours present in their different unicellular ancestors (including aggregative behaviour, flagellar motility, amoeboid movement, clonal colony formation, etc.) in a single multicellular entity, by means of controlling cell differentiation spatially, rather than temporally. These ancestral cell stages could have been the basis of the cell type diversification observed in animal evolution (Arendt 2008). Therefore, we propose that the cell behaviours of the unicellular ancestors of Metazoa, as well as and their associated gene machineries, were most likely co-opted into the multicellular animals, producing an spatial, rather than phyletic, cell type distribution.

Methods

Scanning Electron microscopy. C. owczarzaki cells of the corresponding stage were fixed for one hour with 2'5% glutaraldehyde (Sigma-Aldrich, St. Louis, MO, USA), and for another hour with 1% osmium tetroxide (Sigma-Aldrich), followed by dehydration in a graded ethanol series (25%, 50%, 70%, 99%) for 15 min per step, followed by three 15-min rinses in 100% ethanol. Samples were critical-point dried in liquid CO2 using a BAL-TEC CPD 030 critical-point drying apparatus. They were subsequently glued to SEM stubs with colloidal silver, sputter-coated with gold-palladium, and examined with a Hitachi S-3500N (Hitachi High-Technologies Europe GmbH).

Transmission electron microscopy. Cell aggregates were loaded into the copper tubes and immediately cryoimmobilized using a Self-Pressurized Freezing System (EM SPF) (Leica-Microsystems, Vienna). Cells were then stored in liquid nitrogen until further use. Peeled copper tubes were freeze-substituted in anhydrous acetone containing 2% osmium tetroxide and 0.1% uranyl acetate at –90 °C for 72 hours and warmed to room temperature, following a 2ºC increase per hour in five consecutive steps (-60ºC, -30ºC, 0ºC, 4ºC and room temperature) being a total of 8h in each temperature and using a EM AFS (Leica-Microsystems, Vienna). After several acetone rinses, samples were infiltrated with Epon resin during 7 days and embedded in resin and polymerised at 60ºC during 48 hours. Ultrathin sections were obtained using a Leica Ultracut UC6 ultramicrotome (Leica-Microsystems, Vienna) and mounting on Formvar-coated copper grids. Sections were stainned with 2% uranyl acetate in water and lead citrate, and were observed in a Tecnai Spirit 120 kv electron microscope (FEI Company, Eindhoven, The Netherlands) equipped with a Megaview III CCD camera.

RNA-Seq and analysis. C. owczarzaki cells were grown in 5 ml flasks with ATCC medium 1034 (modified PYNFH medium) in a 23ºC incubator. To have three biological replicates, three cell lines were generated from a single-founding cell and were grown for two months. Total RNA from each cell stage was extracted using Trizol reagent (Life Technologies). Adherent cells were simply scratched from a 3-4 days old culture. Homogenous aggregate formation was induced by growing the cells in gentle agitation at 60 rpm. Floating cystic cells were obtained from a 14 days old culture. Libraries were sequenced with 76 base paired-read using HiSeq 2000 instrument (Illumina). Reads were aligned to the reference genome using Tophat (Trapnell et al. 2012), rendering an average mapping of 90%. Significantly differential expression was calculated by performing pairwise comparision with DESeq (threshold $1e^{-05}$) (Anders and Huber 2010), EdgeR (threshold $1e^{-05}$) (Robinson et al. 2010), CuffDiff (threshold

1e^{-05}) (Trapnell et al. 2012) and NOISeq ((threshold 0.8) (Tarazona et al. 2012) and only genes that appear to be significant at least in three out of the four methods taken as differentially expressed (see Fig. S4B). Quality control analyses (Fig. S4) of the data were performed using cummeRbund R package (Trapnell et al. 2012). These include count vs dispersion plot (Fig. S4C) to estimate over-dispersion, density plot to assess the distributions of FPKM scores across samples (Fig. S4D) and squared coefficient of variation plot to check for cross-replicate variability (Fig. S4E). A gene ontology of *C. owczarzaki*'s 8637 genes was generated using Blast2GO (Conesa et al. 2005) and GO enrichment analyzed using Ontologizer (Bauer et al. 2008). Pfam domains of all genes were analyzed using Pfamscan, counts were generated using custom Perl scripts and Fisher's exact tests performed using custom R scripts.

Alternative splicing analysis. Exon skipping and IR were analysed as previously described (Barbosa-Morais et al. 2012; Curtis et al. 2012). In short, for exon skipping analyses, multifasta libraries of exon-exon junctions were built by combining all forward annotated splicing donor and acceptors. A minimum of eight base pairs was required at each boundary to assure specificity. Next, the number of effective mappable positions was calculated for each exon-exon junction, as previously described (Barbosa-Morais et al. 2012; Labbé et al. 2012). Then, RNA-seq reads (previously trimmed to 50 nucleotides and combining each three replicates to increase read depth) were aligned to these sequences using Bowtie, with –m 1 –v 2 parameters (single mapping and two or fewer mismatches). Percentage of exon inclusion was calculated and a minimal read coverage was required, as previously described (Khare et al. 2012). For IR, a similar approach was taken for each contiguous intron-exon and exon-exon junction and percentage of intron inclusion was calculated as previously described (Curtis et al. 2012). For comparisons among cellular stages, only events with enough read coverage in the three samples were considered (either (i) ≥15 reads in the exon-exon junction or (ii) ≥15 reads in one intron-exon junction and ≥10 in the other), and intron showing >95% inclusion in the three samples were discarded. In order to assess whether differentially retained introns in the same genes were included in a coordinated or in a combinatorial manner, mate information of read pairs was used. Basically, if each side of a read mapped to two different IR events, each side may be providing support for retention of both introns or splicing of both introns (coordinated regulation) or retention of one and splicing of another (combinatorial IR). For the 555 pairs of highly retained introns that had read mate information, 196 (35.3%) had evidence for combinatorial regulation. Finally, for sequence motif enrichment analyses, full intron sequences were compared with MEME (Bailey et al. 2009).

RT-PCRs. In order to validate some of the AS analysis predictions, the three stages were induced (see RNA-Seq and analysis section) and RNA extracted using Trizol reagent (Life Technologies). In order to eliminate genomic DNA, total RNA was treated with DNAse I (Roche) and purified using RNeasey columns (Qiagen). For each stage, cDNA was produced from 1 ug of total RNA using SuperScript III reverse transcriptase (Life Technologies). Pairs of primers of similar melting temperature (60ºC) and spanning the putative alternatively spliced segments were designed using Geneious software. PCR was performed using ExpandTaq polymerase (Roche).

Flow cytometry. C. owczarzaki cells were grown during 10 to 15 days, sampling every day from both the supernatant (to obtain cystic floating cells) and the scratched flask (to obtain adherent cells). Thus, two samples were obtained dayly, for cystic and adherent cells. For DNA-content analysis, a sample was fixed using EtOH 70% and stored at -20ºC for one month. Later, they were fixed and stained with Propidium Iodide (as described in (Darzynkiewicz and Huang 2004)) and DNA content estimated using FACScalibur flow cytometer. For cell counting, 1ml of fresh sample (one for the supernatant and one for the flask surface) was mixed in a BD Trucount Tube, with a known number of beads, so absolute cell number counts could be calculated, using an LSR Fortessa flow cytometer. Two replicate experiments (R1 and R2) where performed independently in order to confidently establish the growth dynamics. Two measures were calculated from the DNA-content analysis. First, the proliferation rate, which indicates the proportion of number of cells in S and G2/M phases vs the number of cells in G0/G1. Second, the percentage of cells in S-phase.

Aanen DK, Debets AJM, De Visser JAGM, Hoekstra RF. 2008. The social evolution of somatic fusion. Bioessays 30:1193–203.

Adl SM, Simpson AGB, Lane CE, et al. 2012. The revised classification of eukaryotes. J Eukaryot Microbiol 59:429–514.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol 11:R106.

Arendt D. 2008. The evolution of cell types in animals: emerging principles from molecular studies. Nat Rev Genet 9:868–882.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res 37:W202–W208.

Barbosa-Morais NL, Irimia M, Pan Q, et al. 2012. The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. Science 338:1587–1593.

Bauer S, Grossmann S, Vingron M, Robinson PN. 2008. Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24:1650–1651.

Bennett MK, Garcia-Arrarás J, Elferink LA, Peterson K, Fleming AM, Hazuka CD, Scheller RH. 1993. The syntaxin family of vesicular transport receptors. Cell 74:863–873.

Bonner JT. 1998. The origins of multicellularity. Integrative Biology 1:27–36.

Brown MW, Kolisko M, Silberman JD, Roger AJ. 2012. Aggregative Multicellularity Evolved Independently in the Eukaryotic Supergroup Rhizaria. Curr Biol 22:1–5.

Brown MW, Silberman JD, Spiegel FW. 2011. A contemporary evaluation of the acrasids (Acrasidae, Heterolobosea, Excavata). Eur J Protistol 48:103–23.

Brown MW, Spiegel FW, Silberman JD. 2009. Phylogeny of the "forgotten" cellular slime mold, Fonticula alba, reveals a key evolutionary branch within Opisthokonta. Mol Biol Evol 26:2699–2709.

Burt DB. 2001. Evolutionary stasis, constraint and other terminology describing evolutionary patterns. Biol J Linn Soc Lond 72:509–517.

Cavalier-Smith T. 2003. Phylogeny of Choanozoa, Apusozoa, and other Protozoa and early eukaryote megaevolution. J Mol Evol 56:540–563.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–6.

Curtis BA, Tanifuji G, Burki F, et al. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature 492:59–65.

Darzynkiewicz Z, Huang X. 2004. Analysis of Cellular DNA Content by Flow Cytometry. Current Protocols in Immunology.

Dayel MJ, Alegado R a, Fairclough SR, Levin TC, Nichols S a, McDonald K, King N. 2011. Cell differentiation and morphogenesis in the colony-forming choanoflagellate Salpingoeca rosetta. Dev Biol 357:73–82.

Dykstra MJ, Olive LS. 1975. Sorodiplophrys: An Unusual Sorocarp-Producing Protist. Mycologia 67:873–879.

Grosberg RK, Strathmann RR. 2007. The Evolution of Multicellularity: A Minor Major Transition? Annu Rev Ecol Evol Syst 38:621–654.

Hernández-Hernández V, Niklas KJ, Newman S a, Benítez M. 2012. Dynamical patterning modules in plant development and evolution. The International Journal of Developmental Biology 56:661–74.

Ince-Dunn G, Okano Hirotaka J., Jensen KB, et al. 2012. Neuronal Elav-like (Hu) Proteins Regulate RNA Splicing and Abundance to Control Glutamate Levels and Neuronal Excitability. Neuron 75:1067–1080.

Jacob F. 1977. Evolution and tinkering. Science 196:1161–1166.

Jøstensen J, Sperstad S, Johansen S, Landfald B, Jųstensen J. 2002. Molecular-phylogenetic, structural and biochemical features of a cold-adapted, marine ichthyosporean near the animal-fungal divergence, described from in vitro. Eur J Protistol 104:93–104.

Khare T, Pai S, Koncevicius K, et al. 2012. 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. Nat Struct Mol Biol 19:1037–1043.

Kiel J a KW. 2010. Autophagy in unicellular eukaryotes. Philos Trans R Soc Lond B Biol Sci 365:819–30.

King N, Westbrook M, Young S, Kuo A. 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 21:4300–4305.

King N. 2004. The unicellular ancestry of animal development. Dev Cell 7:313–25.

Knoll AH. 2011. The Multiple Origins of Complex Multicellularity. Annual Review of Earth and Planetary Sciences 39:217–239.

Labbé RM, Irimia M, Currie KW, et al. 2012. A Comparative Transcriptomic Analysis Reveals Conserved Features of Stem Cell Pluripotency in Planarians and Mammals. Stem Cells 30:1734–1745.

Lang T, Jahn R. 2008. Core proteins of the secretory machinery. Handb Exp Pharmacol 184:107–27.

Lasek-Nesselquist E, Katz L. 2001. Phylogenetic position of Sorogena stoianovitchae and relationships within the class Colpodea (Ciliophora) based on SSU rDNA sequences. J Eukaryo Microb 48:604–607.

Loeb L. 1903. On the Coagulation of the Blood of Some Arthropods and on the Influence of Pressure and Traction on the Protoplasm of the Blood Cells of Arthropods. Biological Bulletin 4:301–318.

Loeb L. 1921. Amœboid Movement, Tissue Formation and Consistency of Protoplasm. Science 53:261–262.

Marshall WL, Celio G, McLaughlin DJ, Berbee ML. 2008. Multiple isolations of a culturable, motile Ichthyosporean (Mesomycetozoa, Opisthokonta), Creolimax fragrantissima n. gen., n. sp., from marine invertebrate digestive tracts. Protist 159:415–33.

McGuire AM, Pearson MD, Neafsey DE, Galagan JE. 2008. Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biol 9:R50.

Michod RE. 2007. Evolution of individuality during the transition from unicellular to multicellular life. Proc Natl Acad Sci U S A 104 Suppl:8613–8.

Newman S a. 2012. Physico-Genetic Determinants in the Evolution of Development. Science 338:217–219.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature 463:457–63.

Owczarzak a, Stibbs HH, Bayne CJ. 1980. The destruction of Schistosoma mansoni mother sporocysts in vitro by amoebae isolated from Biomphalaria glabrata: an ultrastructural study. Journal of Invertebrate Pathology 35:26–33.

O'Shea KS. 1987. Differential deposition of basement membrane components during formation of the caudal neural tube in the mouse embryo. Development 99:509–519.

Paps J, Medina-Chacón L a., Marshall W, Suga H, Ruiz-Trillo I. 2012. Molecular Phylogeny of Unikonts: New Insights into the Position of Apusomonads and Ancyromonads and the Internal Relationships of Opisthokonts. Protist. [Epub ahead of print]

Queller DC. 2000. Relatedness and the fraternal major transitions. Philos Trans R Soc Lond B Biol Sci 355:1647–55.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 :139–140.

Rokas A. 2008. The origins of multicellularity and the early history of the genetic toolkit for animal development. Annu Rev Genet 42:235–51.

Ruiz-Trillo I, Burger G, Holland PWH, King N, Lang BF, Roger AJ, Gray MW. 2007. The origins of multicellularity: a multi-taxon genome initiative. Trends Genet 23:113–8.

Ruiz-Trillo I, Roger AJAJ, Burger G, Gray MWMW, Lang BFF. 2008. A Phylogenomic Investigation into the Origin of Metazoa. Mol Biol Evol 25:664–72.

Savage RM, Danilchik M V. 1993. Dynamics of Germ Plasm Localization and Its Inhibition by Ultraviolet Irradiation in Early Cleavage Xenopus Embryos. Dev Biol 157:371–382.

Schaap P. 2011. Evolutionary crossroads in developmental biology: Dictyostelium discoideum. Development 138:387–96.

Sebé-Pedrós A, Roger A, Lang FB, King N, Ruiz-Trillo I. 2010. Ancient origin of the integrin-mediated adhesion and signaling machinery. Proc Natl Acad Sci U S A 107:10142–7.

Shalchian-Tabrizi K, Minge M a, Espelund M, Orr R, Ruden T, Jakobsen KS, Cavalier-Smith T. 2008. Multigene phylogeny of choanozoa and the origin of animals. PLoS One 3:e2098.

Steenkamp ET, Wright J, Baldauf SL. 2006. The protistan origins of animals and fungi. Mol Biol Evol 23:93–106.

Stibbs HH, Owczarzak A, Bayne CJ, DeWan P. 1979. Schistosome sporocyst-killing amoebae isolated from Biomphalaria glabrata. Journal of Invertebrate Pathology 33:159–170.

Suga H, Ruiz-Trillo I. 2013. Development of ichthyosporeans sheds light on the origin of metazoan multicellularity. Developmental biology. [Epub ahead of print]

Tarazona S, Furio-Tari P, Ferrer A, Conesa A. 2012. NOISeq: Exploratory analysis and differential expression for RNA-seq data. R package version 1.0.0.

Torruella G, Derelle R, Paps J, Lang BF, Roger AJ, Shalchian-Tabrizi K, Ruiz-Trillo I. 2012. Phylogenetic Relationships within the Opisthokonta Based on Phylogenomic Analyses of Conserved Single-Copy Protein Domains. Mol Biol Evol 29:531–44.

Trapnell C, Roberts A, Goff L, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578.

Westbrook MW. 2011. Introns and alternative splicing in choanoflagellates. PhD Thesis. University of California.

Wilson H V. 1907. On some phenomena of coalescence and regeneration in sponges. Journal of Experimental Zoology 5:245–258.

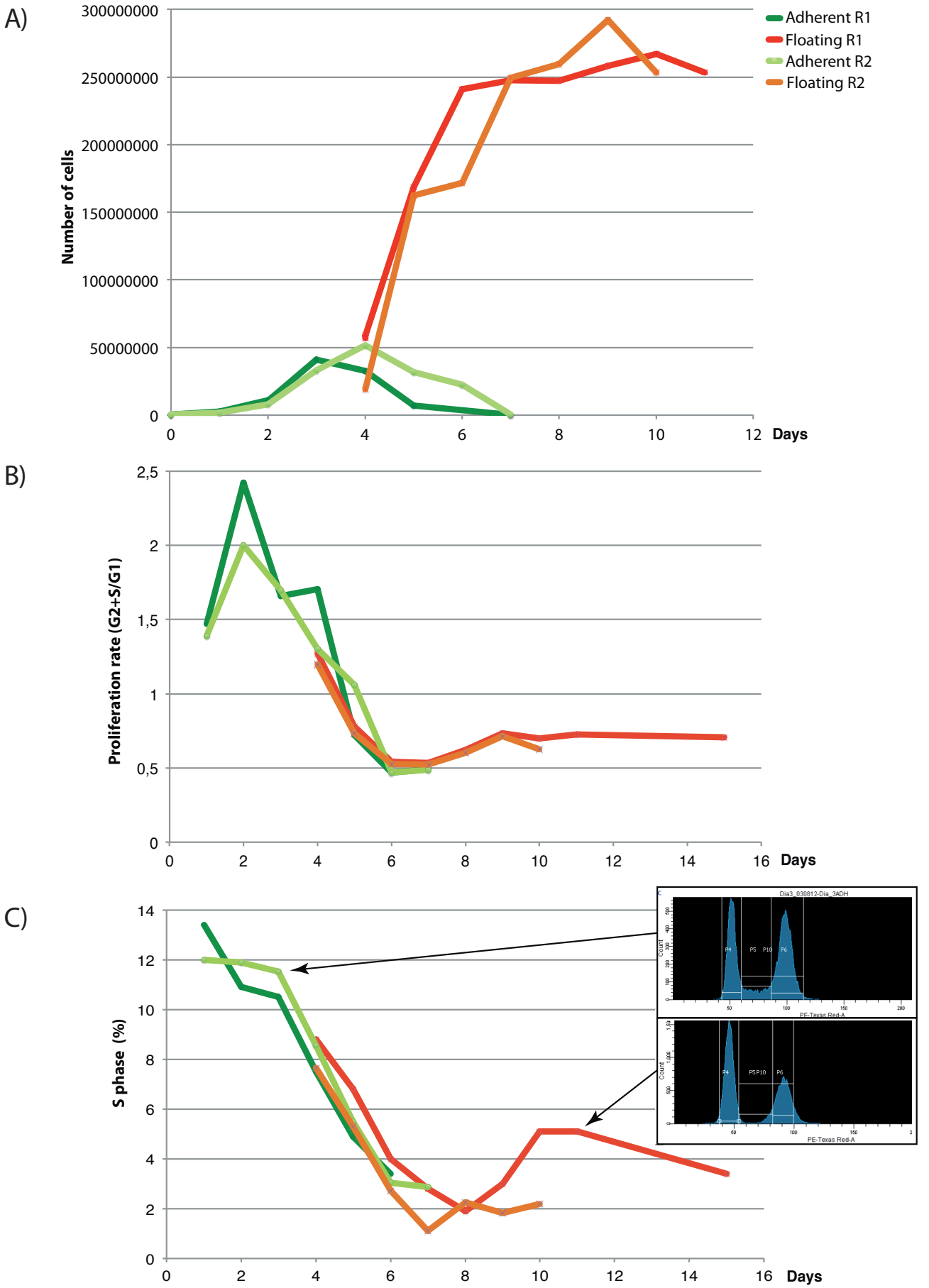


Figure S1. Analysis of *C.owzarzaki* cell cycle. A) Total number of cells per day in each fraction (adherent and floating). B) Proliferation rate per day. C) Percentage of cells in S-phase per day and two examples of DNA-content profiles obtained from days 3 and 11. Notice the reduction in the number G2/M cells (second peak) and the drastic reduction in S-phase cells (the area between the two peaks). Data from adherent cells (see Methods) is shown in greenish colours and data from floating cells in reddish colours. R1 and R2 refer to replicate experiments.

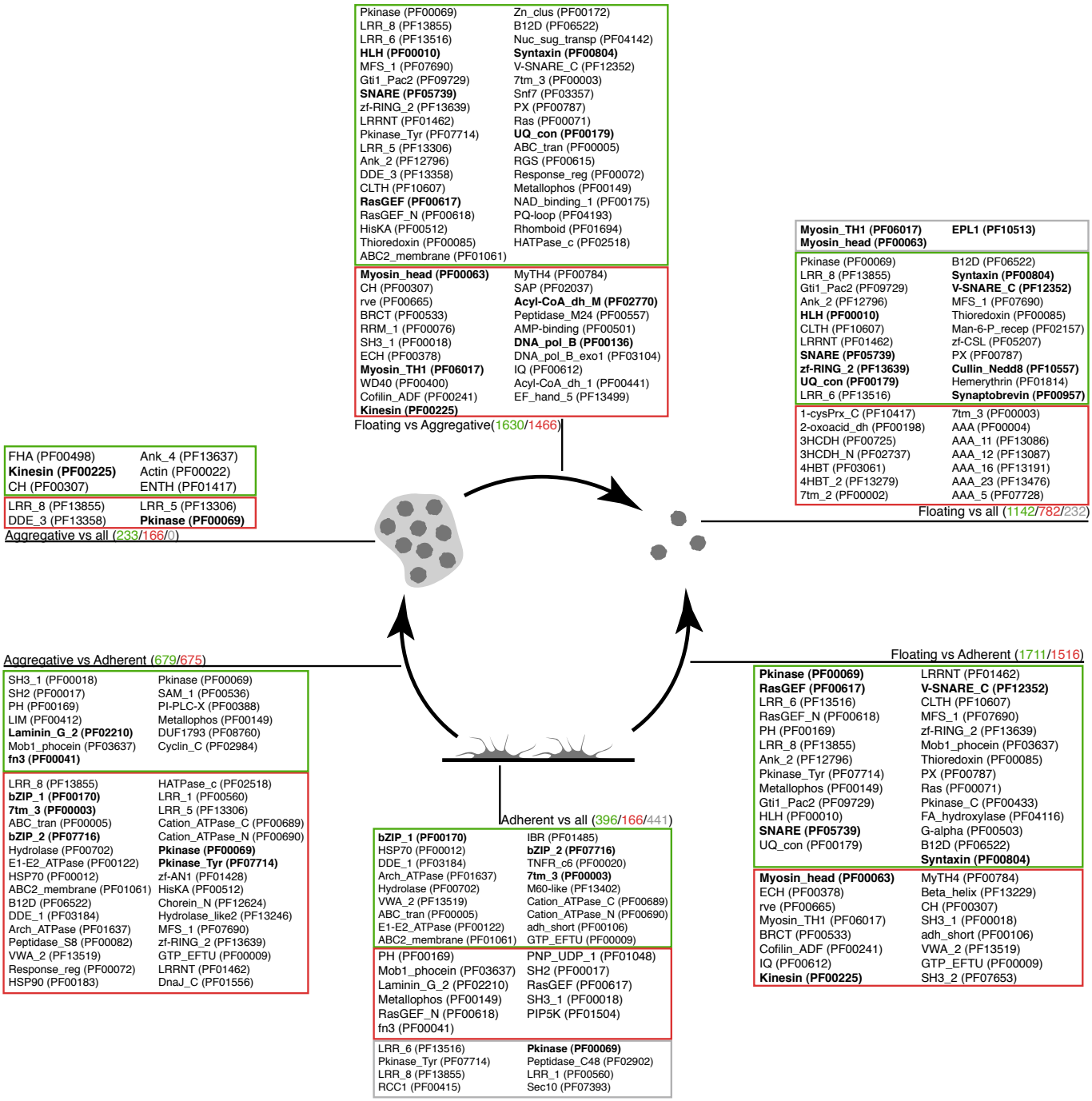


Figure S2. Pfam domain enrichment in sets of differentially expressed genes. Parwise (Aggregative vs Adherent, Floating vs Aggregative and Floating vs Adherent) and one-versus-all comparisions are indicated. The significantly overrepresented Pfam domains (see Methods) are shown for sets of overexpressed (green) and downregulated (red) genes and for genes in with differential intron retention (grey). The number of genes included in each set is indicated with the same colour-code.

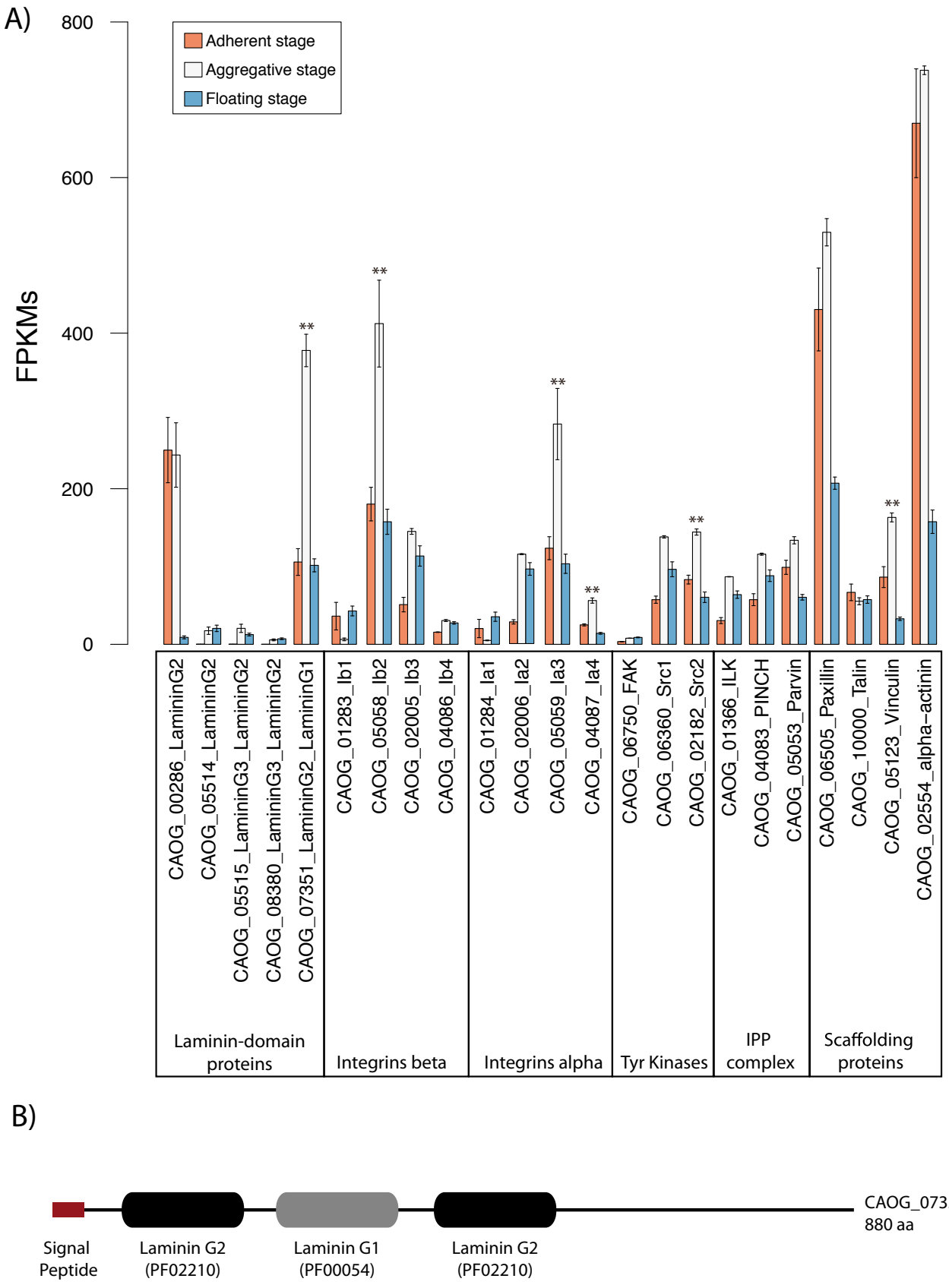


Figure S3. The deployment of the integrin adhesome. A) FPKMs of *C.owczarzaki* laminin-domain containing genes and integrin adhesome genes by stage. B) Domain architecture of the overexpressed laminin gene in aggregates. Asterisks show significantly overexpressed genes in Aggregative vs all comparisons. Bars show standard errors.

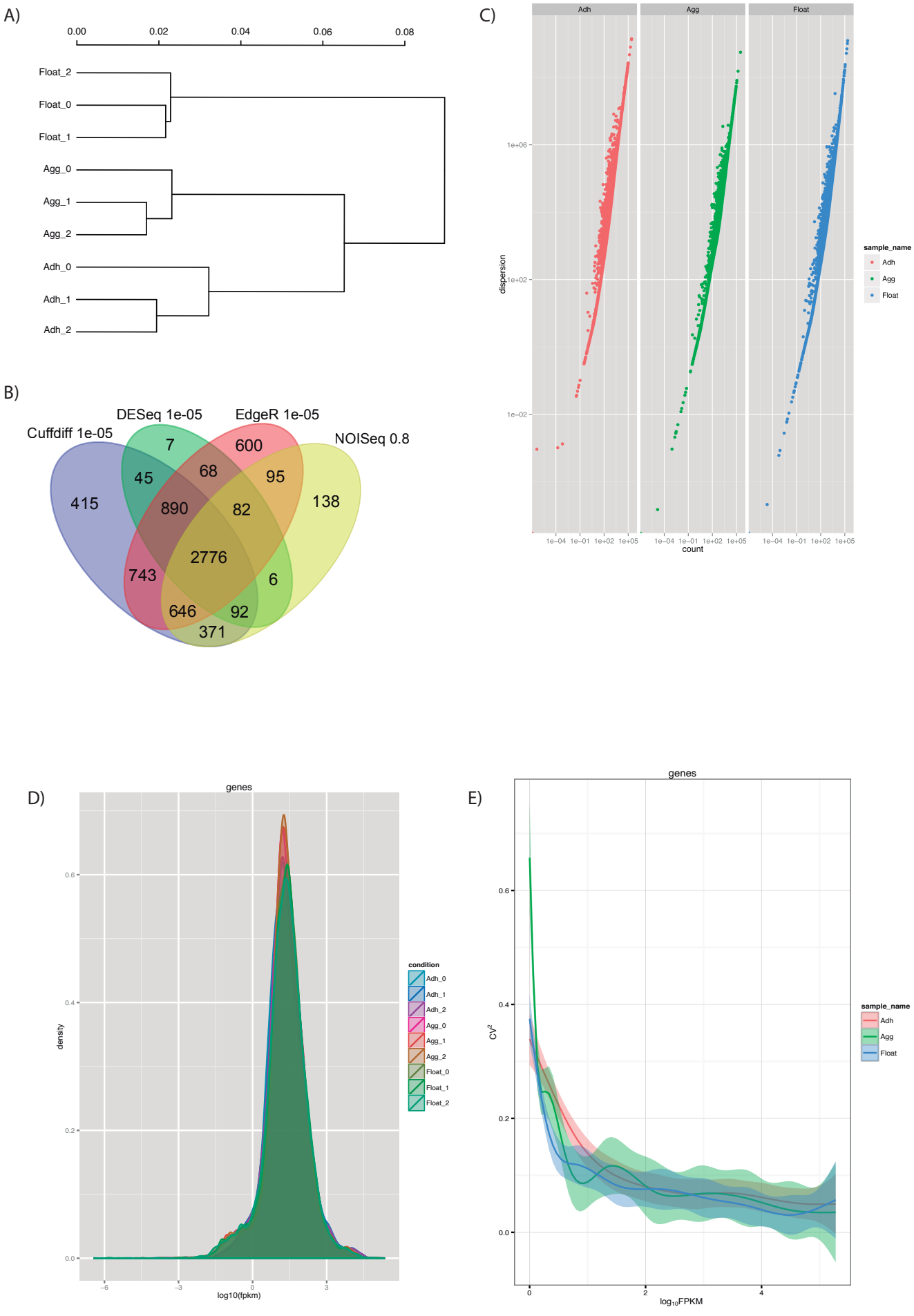


Figure S4. A) Dendrogram of Jensen-Shannon distances between replicates of each cell stage B) Venn diagram of the number of differentially expressed genes between different methods. Cut-off for each method is indicated. C) Count vs dispersion plot by stage for all genes D) Density plot of the different replicates of each cell stage E) Squared coefficient of variation plot of the different stages. Adh: Adherent stage, Agg: Aggregative stage, Float: Floating stage.

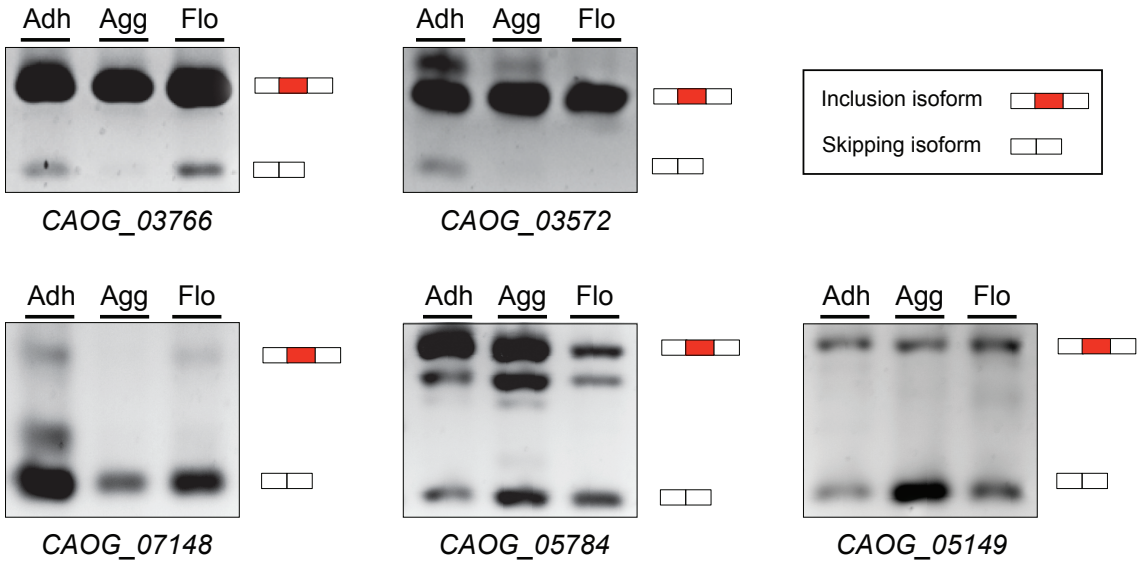


Figure S5. Exon-skipping validation. The gene ID is indicated.

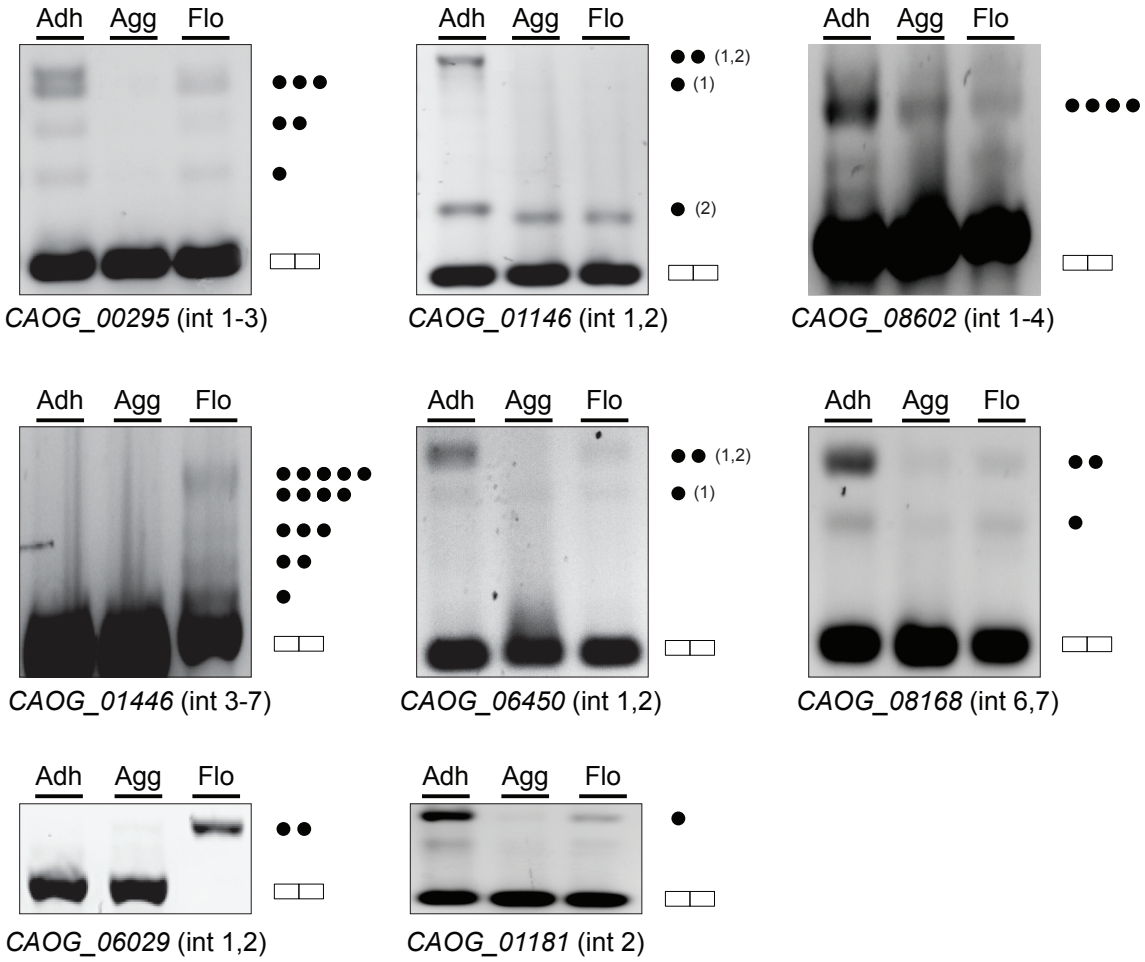


Figure S6. Intron-retention validation. The gene ID is indicated, followed by the introns spanned by the PCR primers.

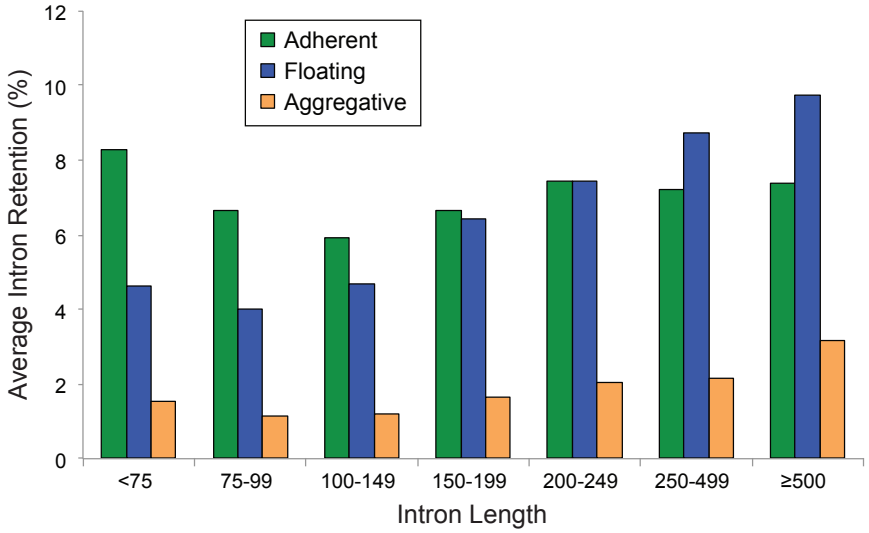
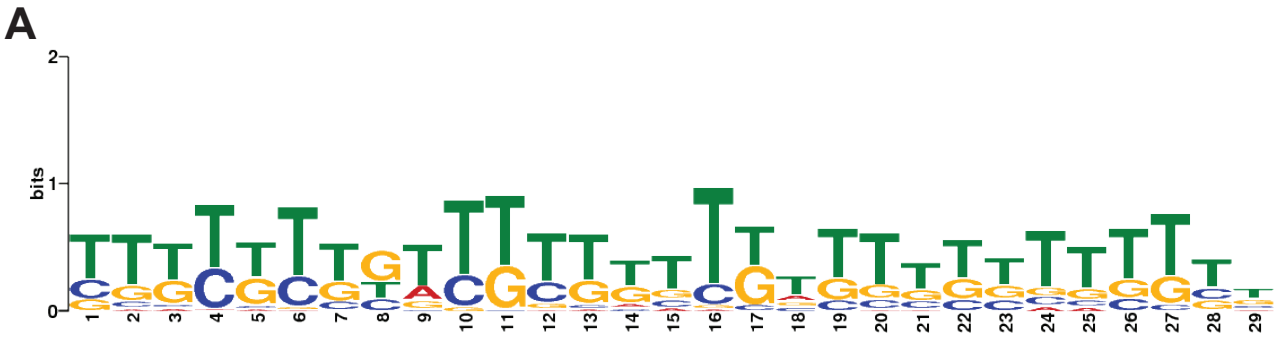
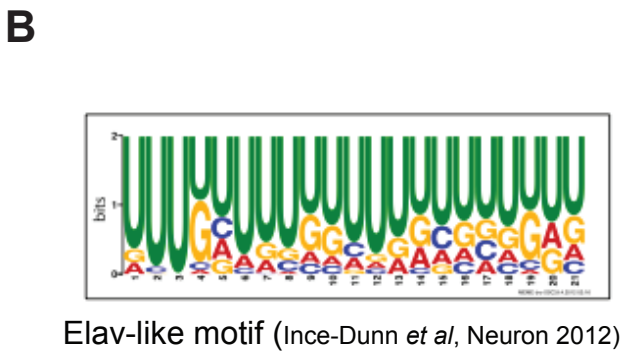


Figure S7. Relationship between intron length and retention. Percentage of average intron retention in each of the three cellular stages for different bins of intron size. In the floating stage, the percentage of intron retention increased with intron length.





Elav-like motif (Ince-Dunn *et al*, Neuron 2012)

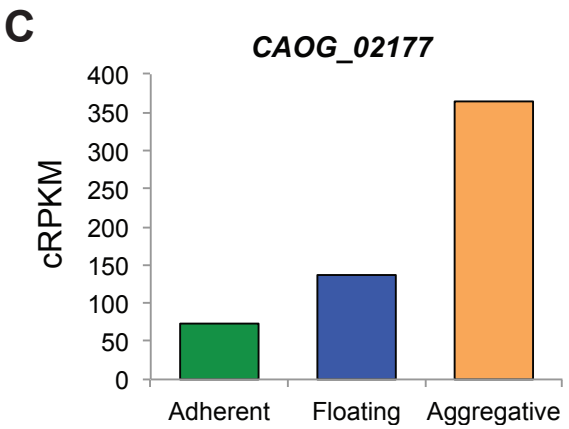


Figure S8. Possible role for a *C. owczarzaki* Elav-like ortholog in the negative regulation of adherent-specific dRIs. A) Most significantly enriched motif in adherent-specific dRIs, obtained by MEME. B) Consensus motif obtained by CLIP-Seq data for an Elav-like member in mammals by Ince-Dunn et al (2012) and that closely resembles the motif in (A). T~U. C) Expression (measured as cRPKMs) of CAOG_02177, an Elav-like ortholog from *C. owczarzaki*, shows lower expression in adherent stage.

Discussion

El accidente singular […] sacado del reino del puro azar, entra en el de la necesidad, el de la certidumbres más implacables.

Jacques Monod, *El azar y la necesidad. Ensayo sobre la filosofía natural de la biología moderna*
Tusquets Editores, 2000

4.1 A new view of *Capsaspora owczarzaki* life cycle

The life cycle of *C. owczarzaki* was unknown when my project started. However, a better understanding of the basic biology of this organism is crucial to make testable hypotheses about the functions of the molecular machineries identified in its genome. Moreover, having a catalogue of well-described phenotypes is necessary for interpreting genetic manipulations, the future research goal in *C. owczarzaki*. Thus, while analysing its genome, I also studied *C. owczarzaki* cell biology. This work has allowed us to describe its complete life cycle in laboratory culture conditions (Results R7). Interestingly, *C. owczarzaki* life cycle includes a previously unknown aggregative cell stage. The transitions from/to this stage are tightly regulated in terms of gene expression and differential alternative splicing (Results R7). As previously described (Stibbs et al. 1979; Owczarzak et al. 1980), we observe that *C. owczarzaki* has a large nucleus and a cytoplasm with multiple lipid vesicles and phagosomes (Figure 14A). Other aspects of *C. owczarzaki* cell biology that we have described include its actin and tubulin structures and how they change during its life cycle (Figure 13). *C. owczarzaki* filopodia contain filamentous actin (Figure 13A, Results R6) and they are lost during encystment (Figure 13B,C). Tubulin staining reveals the formation of a peripheral

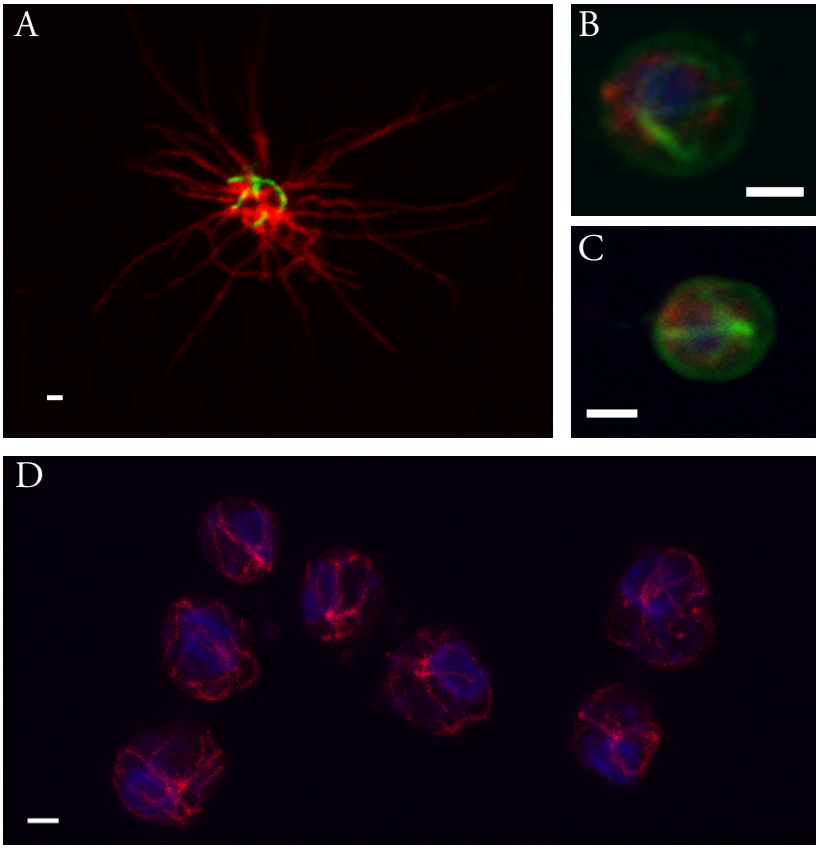


Figure 13. *C. owczarzaki* actin and tubulin structures.
A) Adherent stage cell stained with phalloidin (red, polymerised actin) and anti-tubulin antibody (green). Notice the long actin-filled filopodia and the tubulin basket.
B) C) Floating stage cells stained with phalloidin (red), anti-tubulin antibody (green) and DRAQ5 (blue, nucleus). Filopodia are absent, but the tubulin basket remains.
D) Adherent stage cells stained with anti-tubulin antibody (red) and DRAQ5 (blue, nucleus). Notice the tubulin cables extending from a tubulin-dense area, likely the MTOC, and surrounding all the cell body.
All scale bars=1 μm.

tubulin basket (Figure 13D). The centrosome is associated with the nucleus and it is nucleating a cage of microtubules that radiate away from the nucleus. A similar structure can be observed in choanoflagellates (Results R6), where it has been demonstrated that the actin microfilaments that sustain the microvilli collar are physically linked with tubulin cables that emerge from the flagellum root. This configuration helps to maintain the microvilli around the flagellum (Karpov and Leadbeater 1998). A tubulin basket is also found in flagellated eukaryotes like *Naegleria gruberi* (Heterolobosea, Excavata) and *Chlamydomonas reinhardtii* (Chlorophyta, Viridiplantae) (Zacheus Cande, pers. comm.). But its function in the non-flagellated *C.owczarzaki* remains a mystery.

The most striking ultrastructural feature of *C. owczarzaki* is the presence of long filopodia. From scanning electron micrographies, we can observe that these filopodia can be up 20 μm, allowing *C. owczarzaki* to survey the environment far away from its cell body (Figures 14B-D). Moreover, we show that these filopodia can be branched (Figure 14B) and, sometimes, they emerge from a brush-like root (Figure 14C).

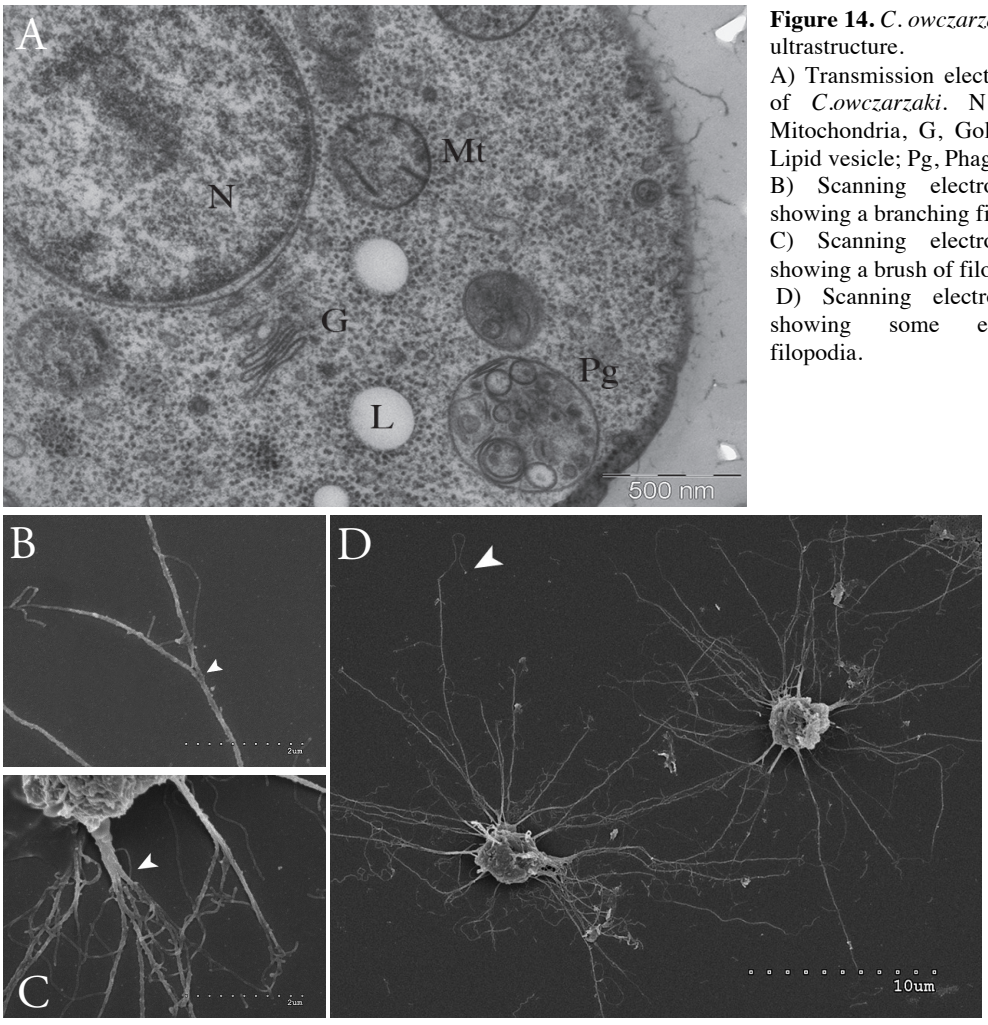


Figure 14. *C. owczarzaki* ultrastructure.
A) Transmission electron micrography of *C.owczarzaki*. N, Nucleus; Mt, Mitochondria, G, Golgi apparatus; L, Lipid vesicle; Pg, Phagosome.
B) Scanning electron micrography showing a branching filopodia.
C) Scanning electron micrography showing a brush of filopodia.
D) Scanning electron micrography showing some extremely long filopodia.

4.2 The origin of the metazoan multicellularity gene repertoire

4.2.1 Adhesion

Metazoan cell adhesion, unlike that of plants or fungi, relies on proteins that establish cell-cell and cell-extracellular matrix (ECM) contacts. A quick survey of the different types of junctions in metazoans (Figure 15) shows that focal adhesions and adherens junctions were likely the ancestral types of metazoan junctions, mediating cell-ECM and cell-cell cohesiveness, respectively (Magie and Martindale 2008). Later, occluding septate and tight junctions and cell-to-cell transport-facilitating GAP junctions evolved.

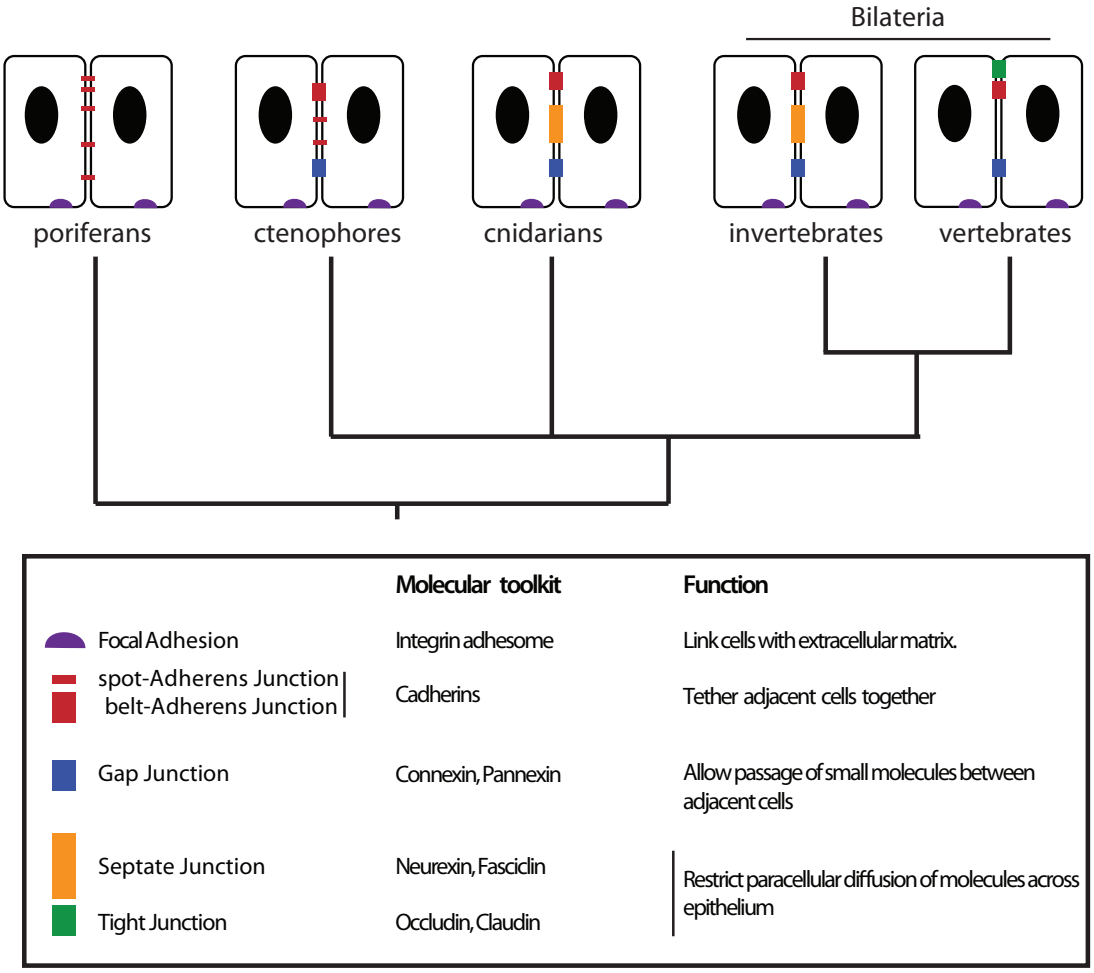


Figure 15. Metazoan cell junction types evolution. Modified from Magie et. al 2008.

In 2008, Abedin et al. reported and characterised several (23 in total) cadherin genes from the choanoflagellate *Monosiga brevicollis* (Abedin and King 2008). This number exceeds the number of cadherins found in many metazoans; for example, the sponge *Amphimedon queenslandica* and the cnidarian *Nematostella vectensis* have 17 and 16 cadherins, respectively (Nichols et al. 2012). The choanoflagellate cadherins locate in

the microvilli collar and, in contrast with classical cadherins in animals, they do not mediate homophilic interactions (Nicole King pers. comm.). For these two reasons, it has been hypothesised that, in a unicellular context, these cadherins work helping to capture bacteria or to signal when bacteria have contacted the collar. The fact that many pathogenic bacteria bind to E-cadherins and flamingo cadherins and exploit them as tethers during host cell invasion, reinforces this idea of a possible relation between cadherins and bacteria (Abedin and King 2008). *Salpingoeca rosetta*, another choanoflagellate, has also been recently reported to posses a large cadherin repertoire, 29 in total; while the filasterean *C. owczarzaki* possesses a single cadherin gene, pushing deeper in time the origin of cadherins (Nichols et al. 2012). However, these choanoflagellate cadherins represent lineage-specific expansions and none of them is homologous to the metazoan classical cadherins. The later are defined by a cytoplasmic cadherin domain (CCD), which is responsible for contacting with β-catenin to form stable cell-cell contacts (called adherens junctions) by homophilic interaction with other cadherins. Therefore, the molecular machinery for adherens junctions evolved before the advent of animal multicellularity, and was later co-opted to function in cell-cell adhesion.

Our work (Results R1 and R2) has shown that the molecular toolkit for the other ancestral metazoan cell junction, the focal adhesions, also evolved long before the origin of animals. In this case, much earlier than the cadherins: in the last common ancestor of Apusozoa and Opisthokonta. Interestingly, this toolkit was secondarily lost in both choanoflagellates and fungi. The fact that integrins, which form the focal adhesions, appear concomitantly with their associated signaling modules (first only with the IPP complex and later, in Holozoa, with the tyrosine kinases FAK and c-Src) and the fact that practically all extracellular matrix proteins (collagen, fibronectins, etc.) are absent in non-metazoan genomes (Fahey and Degnan 2010; Srivastava et al. 2010; Hynes 2012), made us hypothesise that in a unicellular context the integrin adhesome worked as a signaling system. Its dual signaling-adhesion function would have appeared only later in metazoans by co-option of this machinery. However, the transcriptomic analysis of *C. owczarzaki* life cycle (Results R7) has opened the question of whether, in this organism, the integrin adhesome plays a role in aggregate formation. In fact, several components of the adhesome and a laminin domain-containing protein (that is likely to be secreted) are significantly upregulated in the aggregate stage. Future work to

characterise the subcellular localizations of these elements in *C. owczarzaki* will help to resolve this question.

Therefore, the ground was laid for the two ancient types of metazoan cell junctions to evolve. Initially, they may have differently contributed to cell adhesion in different metazoan lineages, in a series of morphological "experiments" relying on different combinations of cell junctions (Newman and Bhat 2009). For example, among extant metazoans, sponges have a greater reliance on cell-ECM junctions to build the bulk of their body than other groups. The body of the sponges is composed basically of a highly developed mesohyl (the intercellular compartment in sponges (Simpson 1984)). The sponge cells (some of which, like the archeocytes, secrete the ECM proteins and, in some cases, mineral spicules) are embedded in it and they bind some of the mesohyl components (like collagens, laminins and fibronectins) via integrins. Indeed, it is this mesohyl and its components what basically define sponge morphology (Simpson 1984), although they also have "epithelioid" organization of cells (based on cell-cell direct contacts) in some restricted areas of their body (Fahey and Degnan 2010). In contrast, other metazoans have a much more tissue-restricted distribution of this ECM-based "mesenchymal" organization. For example, the mesoglea, a functional equivalent of the mesohyl in non-sponge metazoans, is generally less abundant and restricted. This allows the development of a more patterned morphology, in a different way than sponges do (Simpson 1984; Salazar-Ciudad 2003; Newman and Bhat 2009).

4.2.2 Transcription factors

An increasingly complex regulation of gene expression, associated with the specific expansion of the transcription factor toolkit, has long been suspected to be a major force underlying the origin of animal multicellularity (Levine and Tjian 2003). Not surprisingly, transcription factors (TFs) play essential roles in cell differentiation and control over cell proliferation, essential functions for multicellularity (Bonner 2003).

Our genomic survey of TFs in unicellular relatives of metazoans (Results R3) showed a surprisingly high diversity of TFs in *C. owczarzaki*, including several gene families previously suspected to be metazoan-specific (Larroux et al. 2008). A more detailed look at this TF toolkit showed some hints pointing to the possibility that, although having many TFs, the interactivity between TFs (in the case of dimerizing TFs), and between TFs and other cofactors, might be rather low in this unicellular context. For

example, virtually all bZIP TFs found in *C. owczarzaki* belong to families that act as homodimers, allowing little combinatorial diversity (Deppmann et al. 2006).

This possibility was further explored in our study of the evolution of T-box TF family (Results R4). The presence of a clear Brachyury ortholog in *C. owczarzaki*, a well-known TF with essential roles in gastrulation and mesoderm specification in animals (Technau 2001), was one of the most shocking findings in our TF survey. A suggesting pattern emerged when we explored its functional conservation with metazoan Brachyury homologs, by performing heterologous expression in *Xenopus laevis* and protein binding microarray (PBM) experiments. *C. owczarzaki* Brachyury can rescue embryo gastrulation in *Xenopus laevis*, but in a rather unspecific way. It does so by activating not only the specific set of known Brachyury downstream genes, but also genes that are usually activated by other T-box gene classes, but not Brachyury. The Brachyury homologs of the ctenophore *Mnemiopsis leidyi* (Yamada et al. 2010) and of the sponge *Sycon ciliatum*, in contrast, mimic the molecular behaviour of endogenous Xenopus Bra. Moreover, the results of the PBM experiments showed that these different behaviours are not due to changes in the DNA-binding motif specificity of the different Brachyury homologs, as this motif is conserved for all of them. Together, these two results strongly suggest that the T-box target specificity observed in metazoans arose through interactions with cofactors and that these interactions were established at the onset of Metazoa.

This result nicely illustrates a general principle: a major evolutionary source of innovation during the origin of Metazoa was an increase in physical, and also regulatory, interactivity between genes. This increased interactivity generated new combinatorial regulatory outputs.

The open question is how these gene regulatory networks changed, not only quantitatively (more interactions), but also qualitatively. In that sense, we hypothesise that, if we compared the downstream gene regulatory network of *C.owczarzaki* Brachyury with that of metazoan Brachyury homologs, we would of course find less downstream regulated genes. But, more importantly, we would also find a difference in the degree of hierarchical levels of these interactions (each hierarchical level defined as a master regulatory gene regulating another regulatory gene). This degree would be lower. Therefore, in a unicellular context, the regulated genes would be batteries of effector genes, rather than other regulatory genes (Erwin and Davidson 2009).

4.2.3 Signaling

Signal transduction is another key function for metazoan multicellularity. Several signaling pathways are conserved across Metazoa and not found in non-metazoans; for example Hedghog, Wnt, TFGβ or Notch pathways (King et al. 2008; Srivastava et al. 2010). In other cases, similar signaling receptors exist outside Metazoa. The best-studied case is that of the receptor tyrosine kinases (RTKs). It has been shown that both *C.owczarzaki* and the choanoflagellates have dozens of RTKs, but none of them is homologous to any metazoan RTK (neither are homologous those between *C.owczarzaki* and choanoflagellates). Metazoan RTKs are the result of yet another (the third) independent expansion of RTKs (Manning et al. 2008; Suga et al. 2012).

Our article about the origin of the Hippo pathway showed that key elements of this pathway are functionally conserved between *C. owczarzaki* and *Drosophila melanogaster* (Results R5). This result exemplifies a general principle of the evolution of some metazoan signaling pathways: while the pathways are either completely new (like Wnt or Hedgehog) or the product of independent expansions (like RTKs), in some cases the intracellular signal transduction cascades are conserved. In the particular case of the Hippo pathway, the upstream receptors Crumbs and Fat complex appeared at the origin of metazoans and they recruited the pre-existing Hippo signal transduction module. This pathway worked under unknown upstream triggers in a unicellular context, likely different environmental conditions. One possibility, in the case of *C. owczarzaki* and its Hippo pathway, is that this pathway may be activated by the integrin adhesome, as both systems coexist in this organism and the relation between them in metazoans has been suggested (Pugacheva et al. 2006).

Therefore, we hypothesise that some pre-existent signal transduction modules were plugged into new upstream receptors, activating similar behaviours (in this case cell proliferation) but under different cues.

4.3 Genetic sources of innovation in Metazoa

The results presented in this thesis, together with studies from other researchers, allow making some generalizations about the possible mechanisms through which the metazoan molecular toolkit evolved (Figure 16).

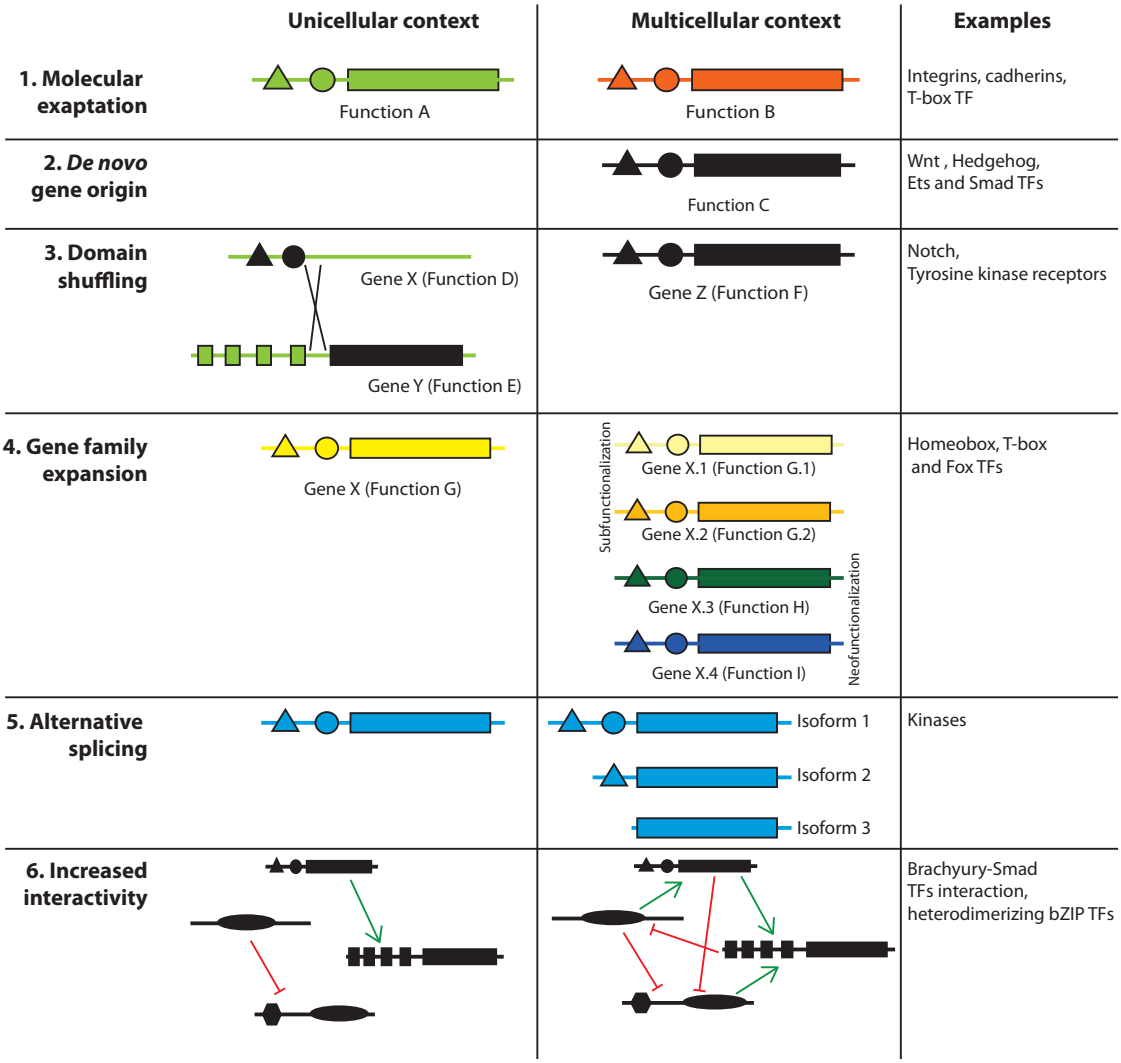


Figure 16. Genetic mechanisms of innovation in the transition to metazoan multicellularity.

4.3.1 Molecular exaptation

A large fraction of the metazoan genes are of ancient eukaryotic origin (Domazet-Loso et al. 2007). Thus, they evolved long before the origin of animals, within a unicellular context, and their current functions in metazoans are somehow similar to the functions they were performing in the unicellular ancestors. This is the case of metabolic genes; many cytoskeletal proteins and other cell structural components; the core replication, transcription and translation machineries; and others. However, there is another fraction of genes that originated before the divergence of metazoans and that changed their

function in a multicellular context. This is known as gene co-option or molecular exaptation, and we believe it was a major source of innovation for metazoans.

By tracing back the origin of many genes previously thought to be metazoan-specific and involved in functions associated with multicellularity (see above), we are making the reasonable assumption that they had a different (unknown) function in a unicellular context (although, ultimately, this must be experimentally demonstrated). Gene co-option associated with the origin of multicellularity has long been hypothesised (King 2004; Michod 2007) and even some experimental studies have dissected the function of particular genes in a unicellular and a multicellular context in green algae and plants, respectively (Nedelcu and Michod 2006; Lee et al. 2008). But only a few cases of gene co-option were hypothesised to be associated with the origin of Metazoa (King et al. 2008; Rokas 2008). In this thesis, we provide plenty of detailed new examples of different machineries essential for animal multicellularity that originated before the metazoan divergence. In some cases, we have demonstrated functional conservation between unicellular and multicellular gene homologs. Although this does not identify the ultimate function of those pathways in a unicellular context, it does show an unexpected high degree of homology, to the point that the unicellular homolog is able to physically interact with an exogenous molecular toolkit in a heterologous complex multicellular system such as *Xenopus laevis* or *Drosophila melanogaster*.

It is important to notice the conceptual difference between molecular exaptation and the rest of mechanisms I will expose next. Molecular exaptation derives from a functional definition: a product of adaptation and selection, but where the particular adaptation represents a secondary use of a trait already present for other reasons (Gould and Vrba 1982). Co-option is not, by itself, a molecular source of genetic change. The genetic basis of gene co-option can be, for example, subfunctionalization after gene expansion, changes in the protein domain architecture by domain shuffling or changes in regulatory interactions. All of them are events that involve changes in the DNA molecule (in the coding sequence, in cis-regulatory sequences, etc.).

4.3.2 *De novo* gene origin

Our comparative genomic analysis have provided not only examples of genes involved in metazoan multicellularity present in their unicellular relatives, but also a better understanding of which machineries are, under our current taxon sampling, specific to Metazoa. Thus, and even though it is a somehow reductionist view (as we shall see

below, is not all about gene content), we can now more precisely define the genetic innovations of animals.

For example, in our study of metazoan TFs evolution, we identified several TFs that are, indeed, metazoan innovations, such as Nuclear Receptor, Ets or Smad (Figure 18).

Regarding the signaling pathways, we have not found outside Metazoa six out of the seven canonical metazoan developmental signaling pathways (Pires-daSilva and Sommer 2003): wingless related (Wnt), transforming growth factor-β (TGF-β), Hedgehog (Hh), Janus kinase (JAK)/signal transducer and activator of transcription (STAT), Notch and nuclear receptor pathways (Suga et al. in prep). And the seventh, receptor tyrosine kinases (RTK), is present in *C. owczarzaki* and choanoflagellates, but none of these genes is homologous to any metazoan RTKs (Suga et al. 2012). Therefore, there was a great innovation in signaling systems at the origin of Metazoa.

In contrast with the amazing diversity of RTKs that we encounter in animal unicellular relatives (in terms of numbers and of different domain architectures), the RTK system is simplified in metazoans. We hypothesise that there was a homogenization of signaling systems at the origin of Metazoa by the establishment of a few evolutionary conserved pathways (Pires-daSilva and Sommer 2003). These pathways mediate similar developmental processes in diverse metazoan phyla (proximo-distal and anterio-posterior patterning, lateral inhibitio, etc.) and standardise the communications between metazoan cells. Cells are no longer facing an ever changing environment, but a rather stable and controlled extracellular milieu.

4.3.3 Domain shuffling

What we have just said about the origin of the receptors and ligands associated to the major metazoan signaling pathways is not exactly true in all cases. This is because, in fact, some of its constituent protein domains existed before. These protein domains were, however, embedded in completely different domain architectures and only later, by a process known as domain shuffling (Tordai et al. 2005), the specific domain architectures of the metazoan proteins arose.

This process has indeed been proposed to be extremely important during the origin of Metazoa (King et al. 2008), especially for signaling genes such as Notch (Figure 17A) and Hedgehog. Another case is Delta domain, which in *C. owczarzaki* is located in the extracellular portion of a Tyrosine kinase receptor (Suga et al. 2012) (Figure 17B).

Another system that evolved largely by domain shuffling is the extracellular matrix (Hynes 2012). *C. owczarzaki* has several protein domains that in metazoans configure major components of the extracellular matrix like laminins and fibronectins, but in *C. owczarzaki* these domains are associated to different domain architectures (Results R7).

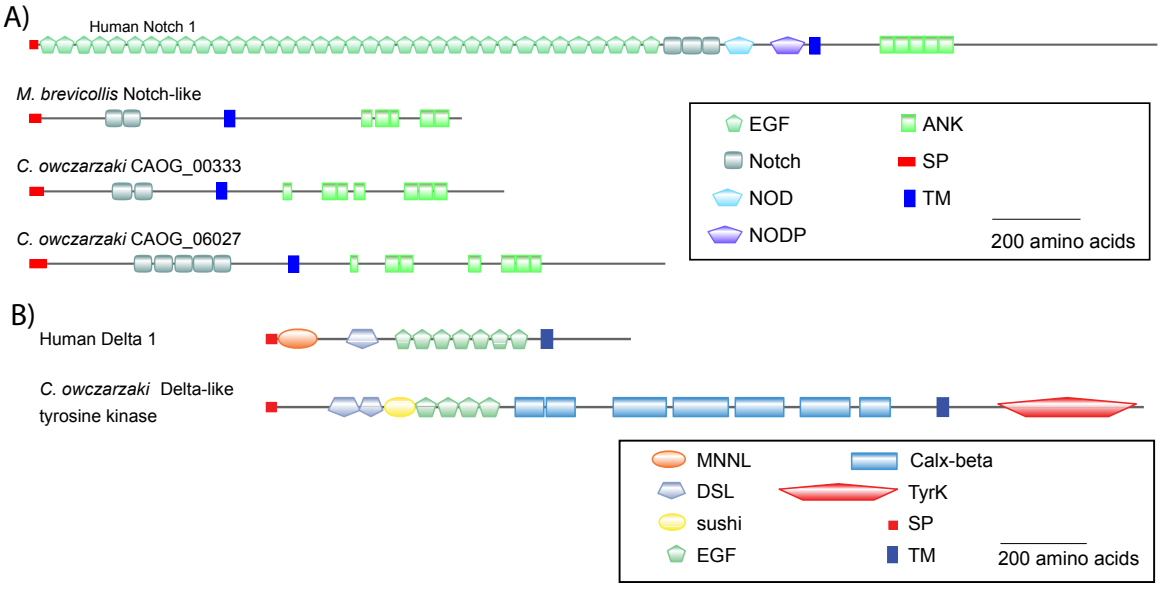


Figure 17. Domain shuffling and the origin of metazoan signaling receptors. Two examples are illustrated of the presence of protein domains that conform two metazoan receptors, Notch (A) and Delta (B), in completely different proteins in *C.owczarzaki* and the choanoflagellate *M.brevicollis,* with different domain architectures. From Suga et al. in prep.

4.3.4 Gene family expansion

Another source of innovation is the expansion of pre-existing gene families, which creates opportunities for new functions to evolve. Several ancient eukaryotic TF families have suffered this process, for example Forkhead (FoxI genes are an expansion found in metazoans), HMG_box (Sox genes are an expanded family in metazoans), Homeobox or zf-C2H2 (Figure 18, in red). Also, some TFs that originated in their immediate unicellular ancestors, such as T-box, NFkappaB and zf-C2HC, suffered significant expansions at the onset of metazoans (Figure 18, in red).

The case of the T-box family is an especially paradigmatic example. We can observe here a TF family that is co-opted at the origin of Metazoa and at the same time is expanded from the ancestral Brachyury-like class. And this expansion is accompanied by a strong subfunctionalization of the different family members: each family is progressively restricted to a subset of specific targets and functions. The heterologous expression of *C. owczarzaki* Brachyury in *Xenopus laevis* shows how this ancient member of the Brachyury T-box class is still "unsubfunctionalised", having the ability

to activate, in *Xenopus laevis,* targets of different T-box classes, which evolved later during the origin of metazoans.

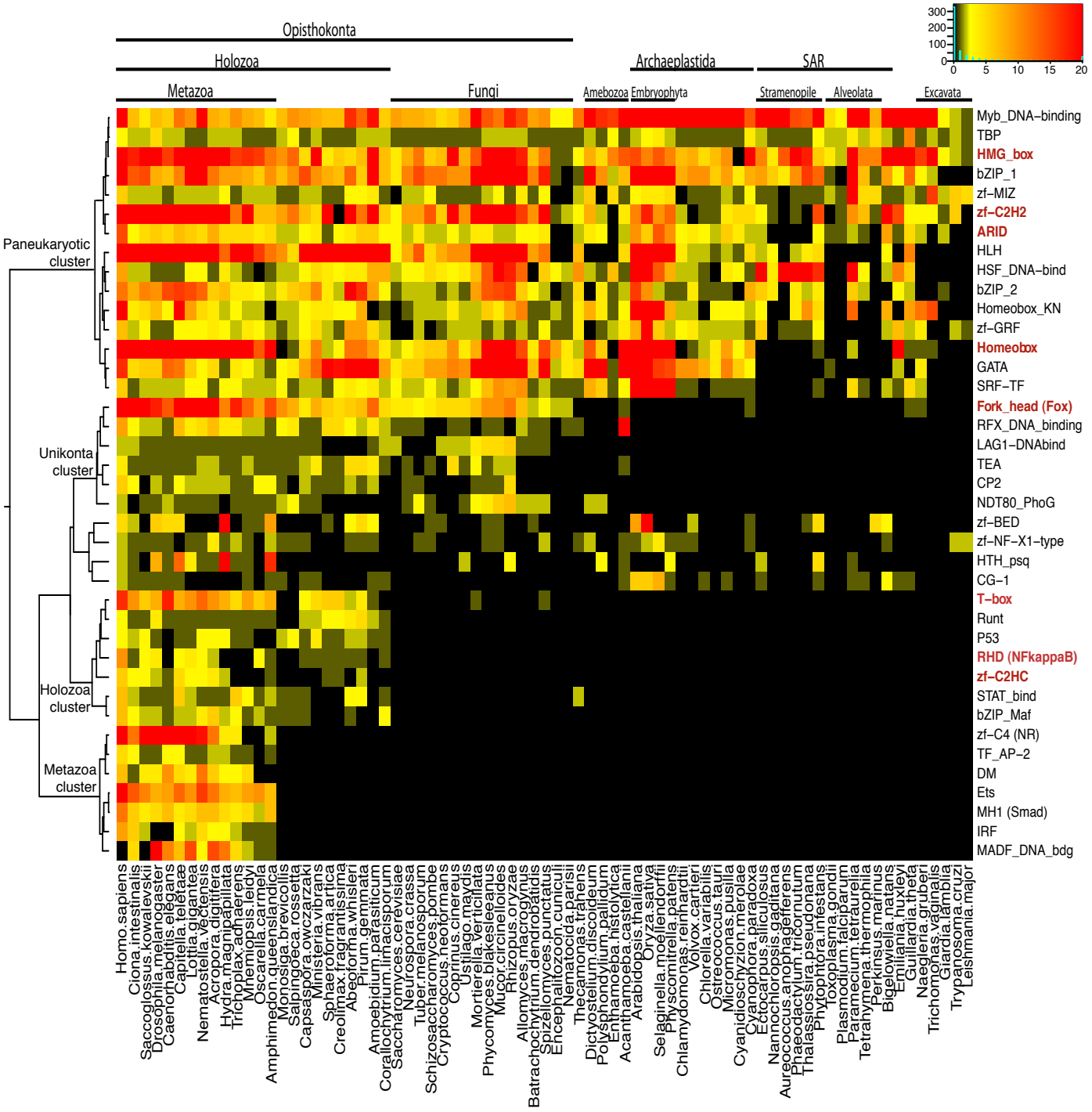


Figure 18. Transcription factor evolution in eukaryotes. The heatmap shows the absolute gene counts (from black, 0, to intense red, >20) of different TF families, clusterized according to their distribution. The taxon sampling includes all major eukaryotic lineages. In red are highlighted the names of those TF that are significantly enriched in Metazoa compared with the other eukaryotes (Wilcox test p-value<0.01).

4.3.5 Alternative splicing

The comparative transcriptome profiling of *C. owczarzaki* life cycle showed a particular pattern of regulated alternative splicing (AS) and also provided us with an idea of the relative importance of AS and the different AS modes in this organism (Results R7). Around 9% of *C. owczarzaki* genes have alternative spliced forms, but our data demonstrate that in most cases (>95%) this AS involves intron retention (IR). In contrast, only few cases of AS by exon-skipping (ES) are found. IR consists in the inclusion of intronic sequences in the final mature mRNA, which makes it useless for translation due to the presence of STOP codons and frame shifts. In contrast, ES always generates mature translatable mRNAs, but it generates different isoforms by including or excluding some exons (McGuire et al. 2008).

The analysis of gene architecture in eukaryotes (Csuros et al. 2011), measured in terms of intron length and intron density (Figure 19), shows that, although these two parameters have a tendency to increase in some metazoans (especially vertebrates), there is a high intra-phyletic variation in both of them. In this regard, *C. owczarzaki* intron composition is similar to that of most unicellular eukaryotes and not so different from that of many metazoans. In the case of choanoflagellates, we can even say that they are relatively intron rich, with intron densities higher than many metazoans

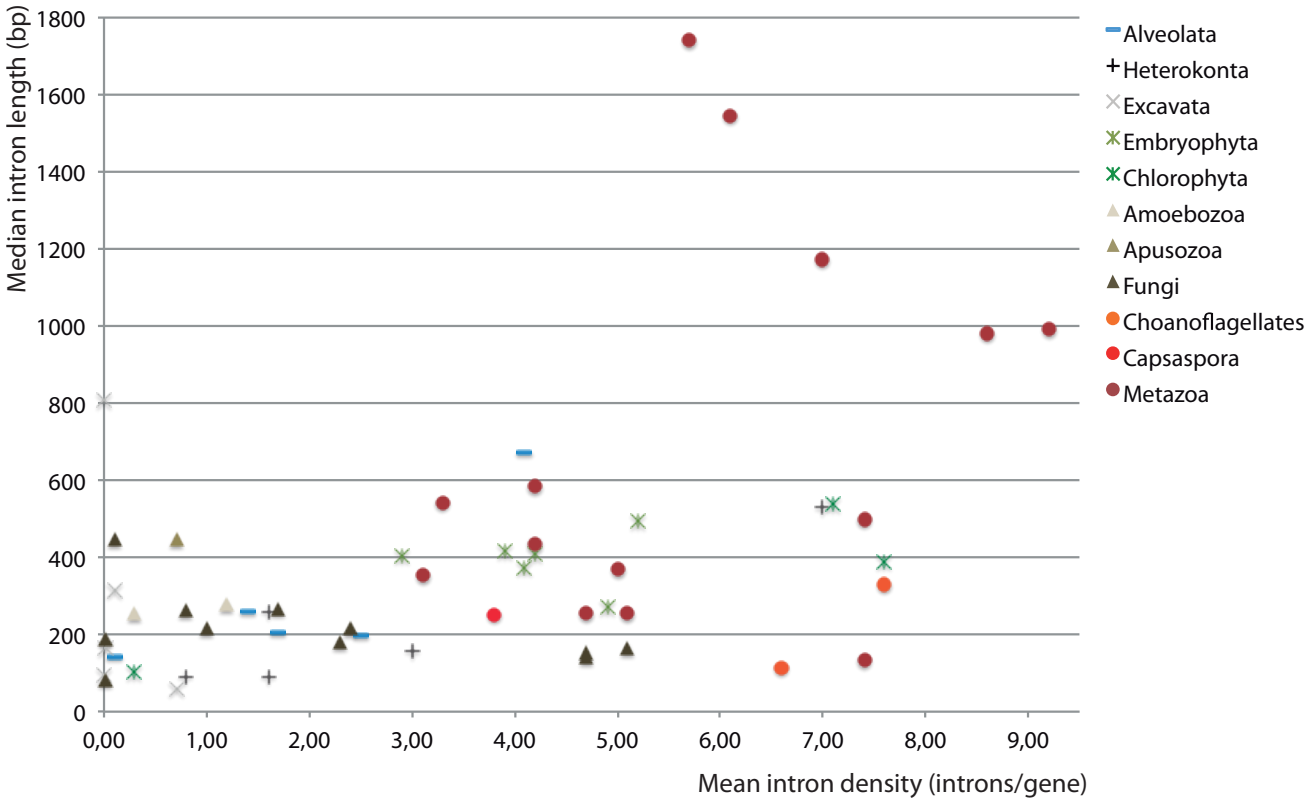


Figure 19. Intron length and intron density distribution across eukaryotes.

(including those of sponges and cnidarians).

Therefore, we can see that there is not a sharp boundary between metazoans and their unicellular relatives in terms of intron composition. However, if we compare the proportion of genes that undergo some kind of alternative splicing, the difference between metazoans and unicellular eukaryotes becomes more evident (Figure 20, right) (Loftus et al. 2005; Pan et al. 2008; Filichkin et al. 2010; Labadorf et al. 2010; Graveley et al. 2011; Ramani et al. 2011; Sorber et al. 2011; Westbrook 2011). In general, metazoans have many more alternatively spliced genes than their closest unicellular relatives (25 to 95% in metazoans versus less than 10% in *C. owczarzaki* and choanoflagellates). AS is a mechanism that allows the expansion of the proteome diversity from a finite set of genes and it has been proposed to be a key mechanism for the evolution of multicellular organisms (Romero et al. 2006; Nilsen and Graveley 2010). Thus, the results in *C. owczarzaki* and choanoflagellates corroborate the idea of an increased amount of AS at the origin of Metazoa.

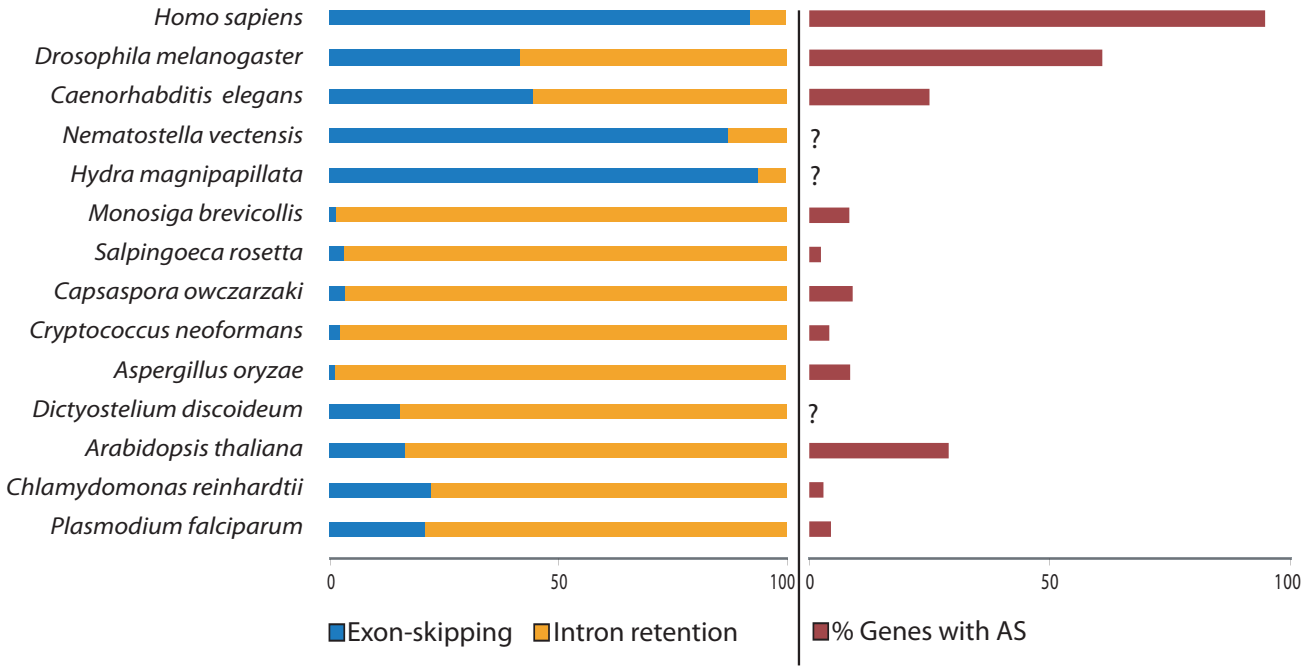


Figure 20. Alternative splicing statistics across eukaryotes. Left, the proportion of ES and IR is shown. Right, the proportion of the total number of genes in each organism that undergo some kind of AS. A quotation mark indicates that no information is available.

But this is not the only difference. When comparing the predominant mode of AS between animals and their unicellular relatives we can see that while intron retention (IR) is predominant in *C. owczarzaki* and choanoflagellates, exon-skipping (ES) is, by

far, the most common AS mechanism in metazoans (Figure 20, left). It has long been shown that IR is predominant also in other eukaryotes (McGuire et al. 2008), but our results confirm that this boundary exists between metazoans and their closest unicellular relatives. This is a crucial difference because IR does not contribute to diversify the proteome, as it generates defective mRNAs that cannot be translated; in contrast, ES does contribute to expand the diversity of the proteome (Anamika et al. 2009). Nevertheless, IR has been suggested to be an important mechanism to regulate gene expression (Yap et al. 2012). Indeed, we have shown (Results R7) that IR is tightly regulated during *C.owczarzaki* life cycle and, in some, cases its functional regulative relevance is quite clear, for example with the increased IR in myosin genes in the cystic stage, where these genes are also downregulated at the transcriptional level.

To sum up, a major shift in AS took place at the origin of animal multicellularity, both in quantitative (% of alternatively spliced genes) and qualitative (modes of AS) terms.

4.3.6 New physical and regulatory interactions

Another major change that, like AS, is not linked with changes in gene content itself (in contrast with gene expansion or *de novo* gene origin), is the establishment of more complex gene regulatory networks (GRN), both by new physical and regulatory interactions.

How GRN changed during the origin of metazoans is an issue that remains largely unexplored. The most obvious approach would be to compare downstream targets of conserved pairs of TF orthologs between animals and *C.owczarzaki* or choanoflagellates. Good candidates for this kind of studies would be Myc, p53, NFkappaB, Runx or Brachyury TFs. In this later case, we have already seen that the regulatory change was likely due new physical interactions with cofactors, which provide new DNA binding specificities (Results R5). Another example is found in signal transduction pathways, like the Hippo pathway, being plugged into new upstream receptors (like Crumbs or the Fat complex in the case of the Hippo pathway).

Changes in GRN are likely to have been of paramount importance at the origin of metazoans. The necessity to spatially (in different cell types) and temporally (during development) deploy genes (either new or co-opted) can only be met by a more complex gene regulation. Some of these major regulatory changes occurred early in metazoan history, and became virtually frozen, producing what is known as kernel GRN (Erwin and Davidson 2009); while others were incorporated later in a phylum, class or

species-specific manner, and are the basis of the morphological and functional diversity of extant metazoans (Davidson and Erwin 2006).

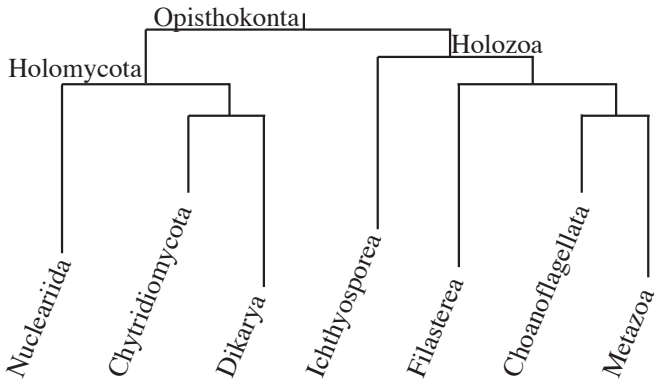
4.4 The origin of Metazoa: from phyletic to spatial cell type distribution

Whether selection can be invoked as a force shaping the origin of Metazoa or whether it was the result of pure historical contingency (see Introduction 1.3), it is incontestable that the study of metazoan unicellular relatives is crucial to understand this transition and to infer the characteristics of the unicellular Urmetazoan ancestor. These traits configure the phylogenetic inertia that laid the foundation of the extant metazoans most basic biology.

If we examine the unicellular lineages close to metazoans (Table 2), we find a mosaic of life cycles and cell structures. This makes it difficult to trace back the origin of these characteristics and to differentiate which of them are plesiomorphic from those that are the product of convergence. For example, both fungi and ichthyosporeans have evolved similar life styles. They are both cell-walled and osmotrophic, forming sporangia by synchronic cell division and having lost in some cases flagellar or amoeboid motility (or both, as in Dikarya and some ichthyosporeans) (Jøstensen et al. 2002; Marshall et al. 2008; Marshall and Berbee 2011; Suga and Ruiz-Trillo 2013). It is important to note that, in both fungi and ichthyosporeans, amoeboid or flagellar motility is only used in transient dispersal cell stages, not in trophic stages as in the case of nucleariids or filastereans (Cavalier-Smith 2012). Chytrids are early branching fungi that posses some striking intermediate characters, such as retention of both flagellar and amoeboid motility. But if we take into account the characteristics of the sister-group of opisthokonts, the Apusozoa, which have filopodia and are phagocytic (Cavalier-Smith and Chao 2010; Paps et al. 2013), we can then infer that fungi and ichthyosporeans have independently evolved fungi-like life-styles and that this was not an ancestral trait in the common ancestor of Opisthokonta, neither in Holozoa.

Filastereans and nucleariids are two independent lineages that also posses some striking similarities. They both have a cell stage consisting of a phagocytic amoeba with long filopodia. Moreover, some species in both groups (like *C. owczarzaki* or *Fonticula alba*) have the ability to form aggregates and secrete a kind of extracellular matrix

(Results R7, (Brown et al. 2009). Finally, it is likely that they independently lost the flagellum.



								Notes
Pluricellular forms	Some	Yes	Some	Yes	Some	Some	Yes	
Secondary transition to unicell	-	No	Some	No	-	?	No	1
Synchronic cell division	-	Yes	Yes	Yes	-	No	Yes	2
Syncitium formation	No	?	Yes	Some	No	No	Some	3
Hyphae formation	No	Yes	Yes	Yes	No	No	No	
Clonal multicellularity	No	Yes	Yes	Yes	No	Yes	Yes	
Spatial cell diferentiation	No	No	Some	No	No	No	Yes	4
Temporal cell diferentiation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Aggregative cell behaviour	Yes	No	No	No	Yes	No	Some	
Flagellar movement	No	Yes	No	Some	No	Yes	Yes	5
Amoeboid movement	Yes	Some	No	Some	Yes	No	Yes	5
Filopodia	Yes	No	No	Some	Yes	Yes	Yes	6
Microvilli	No	No	No	No	No	Yes	Yes	
Cyst formation	Yes	Yes	Yes	Yes	Yes	Some	Some	7
Unwalled cell stages	Yes	Some	No	Some	Yes	Yes	Yes	
Phagocytosis	Yes	No	No	No	Yes	Yes	Yes	8
Osmotrophy	No	Yes	Yes	Yes	No	No	No	8
Extracellular matrix secretion	Yes	No	No	No	Yes	No	Yes	
Undeterminate growth	-	?	Some	No	-	No	No	9

Table 2. Cell structures and behaviours in Opisthokonta. Notes:

1. No yeast-like chytrids are have been reported so far and neither any strict unicellular ichthyosporean. In the case of choanoflagellates, the phylogenetic distribution of colony-forming species makes it equally parsimonious to consider several origins of colony formation or multiple losses of coloniality from a colonial common ancestor (Carr and Leadbeater 2008).
2. See Cutter 1951; Fairclough et al. 2010; Suga and Ruiz-Trillo 2013).
3. Some ichthyosporeans develop by forming a syncytial sporangia (Suga and Ruiz-Trillo 2013). Among animals, insect blastoderm formation and hexactinellid sponge embryos (Leys et al. 2006) are examples of syncytium formation.
4. Many Dikarya are strictly unicellular. Complex multicellularity has evolved independently twice in fungi: in Agaromycotina (Basidiomycota) and in Pezizomycotina (Ascomycota) (Stajich et al. 2009).
5. Chytrid spores are flagellated and, in some cases, the spores also show amoeboid motility, usually after settling. Ichthyosporea are divided in two branches (Adl et al. 2012): in the first, Rhinosporidaceae/Dermocystida, amoeboid movement is lost but many have flagellated spores; in the second, Ichthyophonae, the flagellum has been lost, but amoeboid movement is a common mechanism of spore dispersion.
6. A single case of a pseudopodia-bearing ichthyosporean (*Psorospermium haeckeli*) has been reported so far (Vogt and Rug 1999).
7. Cyst formation has been reported in the choanoflagellate *Monosiga ovata*. Several metazoans form resistance cyst-like (but multicellular) forms, which likely involve similar cell rearrangements (autophagy, etc.).
8. Fungi and Icthysporea likely represent two independent loses of phagocytosis, associated with the formation of cell walls and osmotrophy.
9. The multicellular hypha-forming fungi don not have a determinate growth. In contrast, proliferation in ichthyosporean and choanoflagellate colonies and, specially, in animal development is limited.

Choanoflagellates have lost amoeboid movement, but they have filopodia and, more importantly, a specialised filopodia-like structure called microvilli that allows a unique feeding behaviour. We have seen that microvilli are filopodia-derived structures, as they share part of its molecular toolkit (e.g., fascin), and that microvilli evolved concomitantly with key genes for their function, for example ERM and Eps8 proteins (Results R6). Moreover, they have the ability, like the ichthyosporeans, to form colonies by clonal development.

We can infer that the Holozoa ancestor was probably a phagocytic amoeboflagellate. From this, as we have seen, the ichthyosporeans would have evolved a fungi-like life style. Filastereans would have lost the flagellar apparatus and choanoflagellates the amoeboid movement. The unicellular ancestor of metazoans probably possessed most of these characteristics, being a phagocytic amoeboflagellate (Seravin and Gudkov 2005; Mikhailov et al. 2009); with microvilli in its flagellated stages (the microvilli evolved in the common ancestor between choanoflagellates and metazoans, Results R6); and able to form clonal colonies and to aggregate in amoeboid phase. Alternatively, some or all of these characteristics may have emerged convergently in different metazoan cell types (as it seems to be the case for choanoflagellate-like cells in metazoans (Karpov and Leadbeater 1998; Maldonado 2004)), but, if that is the case, they did so by reusing a machinery that was already in place long before.

The cell types we have examined have a phyletic distribution in extant unicellular holozoans. In Metazoa, we can find these cell types integrated spatially into a single multicellular entity. Indeed, there are great similarities between some metazoan cell types, especially in the early-branching sponges, and the cell types found in the different unicellular holozoans. For example, as mentioned above, the feeding cell type of the sponges, the choanocyte, is morphologically very similar to choanoflagellates (although they have some important ultrastructural differences (Karpov and Leadbeater 1998)) and both share the same feeding mechanism.

Moreover, *C. owzarzaki* show remarkable similarities with sponge archeocytes (Simpson 1984; Funayama 2010): they both are highly motile and proliferative amoeboid cells with long filopodia; they can segregate extracellular materials; they actively phagocyte particles (in fact, archeocytes are the main players in the digestion of the captured particles in sponges (Simpson 1984)); and they both have an extremely high ability to aggregate (Sutter and Vyver 1977; Dunham and Anderson 1983).

It has been proposed that the unicellular ancestor of metazoans had many cell types in a complex life cycle (Mikhailov et al. 2009) and that early metazoans would have multifunctional cell types, integrating several of these cell behaviours in a single cell type (Arendt 2008). Only later, functions would be progressively segregated into more specialised cell types. Sponge archeocytes are an astonishing example of such versatile cell types that may have been present in early metazoans (Simpson 1984).

Further study of the cell behaviours, cell structures and genome content of the unicellular holozoan lineages will provide a better understanding of the starting raw material that conditioned the nature of the very first metazoans.

Conclusion

The main conclusions of the present work are the following:

1. The integrin adhesome originated long before the origin Metazoa, in the common ancestor of Opisthokonta and Apusozoa. It was assembled in a step-wise manner and its original function might be related with signaling rather that adhesion, although transcriptome data in *C. owczarzaki* suggest a role in aggregate formation in this organism.

2. Several transcription factors important for animal development are present in the unicellular relatives of Metazoa; including NFkappaB, T-box, Runx, Myc/Max, p53, STAT, Mef2 and others.

3. *C. owczarzaki* Brachyury is able to rescue *Xenopus* Brachyury endogenous function, including gastrulation. But it does so without the same specificity in target recognition. This specificity was established later in evolution of the T-box family, at the origin of Metazoa, by new interactions with cofactors rather than by new DNA binding motif specificities.

4. The Hippo signaling pathway, an important regulator of cell proliferation and organ growth in metazoans, is present in *C. owczarzaki*. The orthologs of the pathway of this organism are functionally conserved when expressed in *Drosophila melanogaster*.

5. Gene co-option was a more important factor in the origin of the metazoan molecular toolkit than previously recognised; this process affected, among others, the integrin adhesome, several transcription factors and the Hippo signaling pathway.

6. Cell structures and behaviours were also co-opted at the origin of Metazoa. An example is that of the microvilli, which originated in the common ancestor of choanoflagellates and metazoans and are used in a wide diversity of metazoan cell-types.

7. Metazoans and their unicellular relatives have contrasted modes of alternative splicing. This suggests that a shift from intron-retention to exon-skipping occurred at the origin of Metazoa and greatly contributed to the diversification of the animal proteome.

8. *C. owczarzaki* has a complex life cycle that includes an aggregative multicellular stage. Transcriptome data indicate that the transitions between stages are tightly regulated at the level of gene expression and differential alternative splicing. This life cycle of *C. owczarzaki* highlights the immense diversity of cell behaviours present in close unicellular relatives of metazoans and the aggregative cell stage suggests that components of the toolkit for animal multicellularity can be deployed in very different functional contexts.

References

A

Aanen DK, Debets AJM, De Visser J, Hoekstra RF. 2008. The social evolution of somatic fusion. *Bioessays* 30:1193–203.

Abedin M, King N. 2008. The premetazoan ancestry of cadherins. *Science* 319:946–8.

Adl SM, Leander BS, Simpson AGB, et al. 2007. Diversity, nomenclature, and taxonomy of protists. *Syst Biol* 56:684–9.

Adl SM, Simpson AGB, Farmer M, et al. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol* 52:399–451.

Adl SM, Simpson AGB, Lane CE, et al. 2012. The revised classification of eukaryotes. *J Eukaryot Microbiol* 59:429–514.

Aguirre J, Ríos-Momberg M, Hewitt D, Hansberg W. 2005. Reactive oxygen species and development in microbial eukaryotes. *Trends Microbiol* 13:111–8.

Anamika K, Garnier N, Srinivasan N. 2009. Functional diversity of human protein kinase splice variants marks significant expansion of human kinome. *BMC Genomics* 10:622.

Anbar AD, Knoll AH. 2002. Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *Science* 297:1137–42.

Arendt D. 2008. The evolution of cell types in animals: emerging principles from molecular studies. *Nat Rev Genet* 9:868–882.

B

Baldauf S, Palmer J. 1993. Animals and fungi are each other's closest relatives : Congruent evidence from multiple proteins. *Proc Natl Acad Sci U S A* 90:R417–22.

Beaumont NJ. 2009. Modelling the Transport of Nutrients in Early Animals. *Evol Biol* 36:256–266.

Becker B. 2012. Snow ball earth and the split of Streptophyta and Chlorophyta. *Trends Plant Sci* [published online ahead of print October 23 2012]

Bonner JT. 2003. On the origin of differentiation. *J Biosci* 28:523-528

Bonner JT. 1998. The origins of multicellularity. *Integr Biol* 1:27–36.

Boraas M, Seale D, Boxhorn J. 1998. Phagotrophy by a flagellate selects for colonial prey: A possible origin of multicellularity. *Evolutionary Ecology* 12:153–164.

Bouget FY, Berger F, Brownlee C. 1998. Position dependent control of cell fate in the Fucus embryo: role of intercellular communication. *Development* 125:1999–2008.

Boyle RA, Lenton TM, Williams HTP. 2007. Neoproterozoic "snowball Earth" glaciations and the evolution of altruism. *Geobiology* 5:337–349.

Brasier M, Green O, Shields G. 1997. Ediacarian sponge spicule clusters from southwestern Mongolia and the origins of the Cambrian fauna. *Geology* 25:303–306.

Brown MW, Kolisko M, Silberman JD, Roger AJ. 2012. Aggregative Multicellularity Evolved Independently in the Eukaryotic Supergroup Rhizaria. *Curr Biol* 22:1–5.

Brown MW, Silberman JD, Spiegel FW. 2011. A contemporary evaluation of the acrasids (Acrasidae, Heterolobosea, Excavata). *Eur J Protistol* 48:103–23.

Brown MW, Spiegel FW, Silberman JD. 2009. Phylogeny of the "forgotten" cellular slime mold, Fonticula alba, reveals a key evolutionary branch within Opisthokonta. *Mol Biol Evol* 26:2699–2709.

Butterfield NJ. 2009. Modes of pre-Ediacaran multicellularity. *Precambrian Res* 173:201–211.

C

Carr M, Leadbeater B. 2008. Molecular phylogeny of choanoflagellates, the sister group to Metazoa. *Proc Natl Acad Sci U S A* 105:16641–16646.

Cavalier-Smith T, Chao EE. 2010. Phylogeny and Evolution of Apusomonadida (Protozoa: Apusozoa): New Genera and Species. *Protist* 161:549–576.

Cavalier-Smith T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst Evol Microbiol* 52:297.

Cavalier-Smith T. 2003. Phylogeny of Choanozoa, Apusozoa, and other Protozoa and early eukaryote megaevolution. *J Mol Evol* 56:540–563.

Cavalier-Smith T. 2012. Early evolution of eukaryote feeding modes, cell structural diversity, and classification of the protozoan phyla Loukozoa, Sulcozoa, and Choanozoa. *Eur J Protistol*. [published online ahead of print October 18 2012]

Chapman JA, Kirkness EF, Simakov O, et al. 2010. The dynamic genome of Hydra. *Nature* 464:592–596.

Csuros M, Rogozin IB, Koonin E V. 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol* 7:e1002150.

Cunningham JA, Thomas C-W, Bengtson S, Kearns SL, Xiao S, Marone F, Stampanoni M, Donoghue PCJ. 2012. Distinguishing geology from biology in the Ediacaran Doushantuo biota relaxes constraints on the timing of the origin of bilaterians. *Proc R Soc Lond B Biol Sci* 279:2369–76.

Cutter VM. 1951. The Cytology of the Fungi. *Annu Rev Microbiol* 5:17–34.

D

Davidson EH, Erwin DH. 2006. Gene regulatory networks and the evolution of animal body plans. *Science* 311:796–800.

Dayel MJ, Alegado R a, Fairclough SR, Levin TC, Nichols S, McDonald K, King N. 2011. Cell differentiation and morphogenesis in the colony-forming choanoflagellate Salpingoeca rosetta. *Dev Biol* 357:73–82.

Deppmann CD, Alvania RS, Taparowsky EJ. 2006. Cross-species annotation of basic leucine zipper factor interactions: Insight into the evolution of closed interaction networks. *Mol Biol Evol* 23:1480–92.

Derelle R, Lang BF. 2012. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol Biol Evol* 29:1277–89.

Domazet-Loso T, Brajkovic J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet* 23:533–539.

Donoghue PCJ, Antcliffe JB. 2010. Origins of multicellularity. *Nature* 466:41–42.

Dunham P, Anderson C. 1983. Stimulus-response coupling in sponge cell aggregation: Evidence for calcium as an intracellular messenger. *Proc Natl Acad Sci U S A* 80:4756–4760.

Dunn CW, Hejnol A, Matus DQ, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745–749.

Dykstra MJ, Olive LS. 1975. Sorodiplophrys: An Unusual Sorocarp-Producing Protist. *Mycologia* 67:873–879.

E

El Albani A, Bengtson S, Canfield DE, et al. 2010. Large colonial organisms with coordinated growth in oxygenated environments 2.1 Gyr ago. *Science* 466:100–4.

Erwin DH, Davidson EH. 2009. The evolution of hierarchical gene regulatory networks. Nat Rev Genet 10:141–8.

Erwin DH, Laflamme M, Tweedt SM, Sperling E a, Pisani D, Peterson KJ. 2011. The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* 3:1091–7.

F

Fahey B, Degnan BM. 2010. Origin of animal epithelia: insights from the sponge genome. *Evol Dev* 12:601–17.

Fairclough SR, Dayel M, King N. 2010. Multicellular development in a choanoflagellate. *Curr Biol* 20:875–876.

Fairclough SR, Chen Z, Kramer E, et al. 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. *Genome Biol* 14:R15.

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong W-K, Mockler TC. 2010. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res* 20:45–58.

Funayama N. 2010. The stem cell system in demosponges: insights into the origin of somatic stem cells. *Dev Growth Differ* 52:1–14.

G

Gould SJ, Vrba E. 1982. Exaptation-a missing term in the science of form. *Paleobiology* 8:4–15.

Graveley BR, Brooks AN, Carlson JW, et al. 2011. The developmental transcriptome of Drosophila melanogaster. *Nature* 471:473–9.

Gromov B. 2000. Algal parasites of the genera Aphelidium, Amoeboaphelidium, and Pseudaphelidium from the Cienkovski's "monadinea" group as representatives of a new class. *Zoologichesky Zhurnal* 79:517–525.

Grosberg RK, Strathmann RR. 2007. The Evolution of Multicellularity: A Minor Major Transition? *Annu Rev Ecol Evol Syst* 38:621–654.

H

Hejnol A, Obst M, Stamatakis A, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc R Soc Lond B Biol Sci* 276:4261–70.

Hertel LA, Bayne CJ, Loker ES. 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.

Hoffman PF. 1998. A Neoproterozoic Snowball Earth. *Science* 281:1342–1346.

Huldtgren T, Cunningham JA, Yin C, Stampanoni M, Marone F, Donoghue PCJ, Bengtson S. 2011. Fossilized nuclei and germination structures identify Ediacaran "animal embryos" as encysting protists. *Science* 334:1696–9.

Hynes RO. 2012. The evolution of metazoan extracellular matrix. *J Cell Biol* 196:671–9.

J

James-Clark H. 1866. Note on the infusoria flagellata and the spongiae ciliatae. *Am J Sci* 1:113–114.

Jøstensen J, Sperstad S, Johansen S, Landfald B, Jøstensen J. 2002. Molecular-phylogenetic, structural and biochemical features of a cold-adapted, marine ichthyosporean near the animal-fungal divergence, described from in vitro. *Eur J Protistol* 104:93–104.

K

Karpov SA, Mikhailov KV, Mirzaeva GS, Mirabdullaev IM, Mamkaeva KA, Titova NN, Aleoshin VV. 2012. Obligately Phagotrophic Aphelids Turned out to Branch with the Earliest-diverging Fungi. *Protist* [published online ahead of print October 8 2012]

Karpov SA, Leadbeater B. 1998. Cytoskeleton structure and composition in choanoflagellates. *J Eukaryot Microbiol* 45:361–367.

Keim CN, Martins JL, Abreu F, Rosado AS, De Barros HL, Borojevic R, Lins U, Farina M. 2004. Multicellular life cycle of magnetotactic prokaryotes. *FEMS Microbiol Lett* 240:203–8.

King N, Hittinger CT, Carroll SB. 2003. Evolution of key cell signaling and adhesion protein families predates animal origins. *Science* 301:361–3.

King N, Westbrook M, Young S, Kuo A. 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* 21:4300–4305.

King N. 2004. The unicellular ancestry of animal development. *Dev Cell* 7:313–25.

King N. 2005. Choanoflagellates. *Curr Biol* 15:113–114.

Knoll AH, Javaux EJ, Hewitt D, Cohen P. 2006. Eukaryotic organisms in Proterozoic oceans. *Philos Trans R Soc Lond B Biol Sci* 361:1023.

Knoll AH. 2011. The Multiple Origins of Complex Multicellularity. *Annu Rev Earth Planet Sci* 39:217–239.

Kodner RB, Summons RE, Pearson A, King N, Knoll AH. 2008. Sterols in a unicellular relative of the metazoans. *Proc Natl Acad Sci U S A* 105:9897–902.

Koonin EV. 2011. Logic of Chance, The: The Nature and Origin of Biological Evolution. FT Press.

Kües U. 2000. Life History and Developmental Processes in the Basidiomycete Coprinus cinereus. *Microbiol Mol Biol Rev* 64:316–353.

L

Labadorf A, Link A, Rogers MF, Thomas J, Reddy AS, Ben-Hur A. 2010. Genome-wide analysis of alternative splicing in Chlamydomonas reinhardtii. *BMC Genomics* 11:1-10.

Lang BF, O'Kelly C, Nerad T, Gray MW, Burger G. 2002. The closest unicellular relatives of animals. *Curr Biol* 12:1773–8.

Larroux C, Luke GN, Koopman P, Rokhsar DS, Shimeld SM, Degnan BM, Shimeld M. 2008. Genesis and expansion of metazoan transcription factor gene classes. *Mol Biol Evol* 25:980–96.

Lasek-Nesselquist E, Katz L. 2001. Phylogenetic position of Sorogena stoianovitchae and relationships within the class Colpodea (Ciliophora) based on SSU rDNA sequences. *J Eukaryot Microbiol* 48:604–607.

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:R593–4.

Lee JH, Lin H, Joo S, Goodenough U. 2008. Early sexual origins of homeoprotein heterodimerization and evolution of the plant KNOX/BELL family. *Cell* 133:829–840.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. *Nature* 424:147–51.

Leys SP, Cheung E, Boury-Esnault N. 2006. Embryogenesis in the glass sponge Oopsacas minuta: Formation of syncytia by fusion of blastomeres. *Integr Comp Biol* 46:104–17.

Liu Y, Steenkamp ET, Brinkmann H, Forget L, Philippe H, Lang BF. 2009. Phylogenomic analyses predict sistergroup relationship of nucleariids and fungi and paraphyly of zygomycetes with significant support. *BMC Evol Biol* 9:1-11.

Loenarz C, Coleman ML, Boleininger A, Schierwater B, Holland PWH, Ratcliffe PJ, Schofield CJ. 2011. The hypoxia-inducible transcription factor pathway regulates oxygen sensing in the simplest animal, Trichoplax adhaerens. *EMBO Rep* 12:63–70.

Loftus BJ, Fung E, Roncaglia P, et al. 2005. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science* 307:1321–4.

Love GD, Grosjean E, Stalvies C, et al. 2009. Fossil steroids record the appearance of Demospongiae during the Cryogenian period. *Nature* 457:718–21.

M

Magie CR, Martindale MQ. 2008. Cell-cell adhesion in the cnidaria: insights into the evolution of tissue morphogenesis. *Biol Bull* 214:218–32.

Maldonado M. 2004. Choanoflagellates, choanocytes, and animal multicellularity. *Invertebrate Biology* 123:1–22.

Maloof AC, Rose CV, Beach R, Samuels BM, Calmet CC, Erwin DH, Poirier GR, Yao N, Simons FJ. 2010. Possible animal-body fossils in pre-Marinoan limestones from South Australia. *Nat Geosci* 3:653–659.

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.

Marshall WL, Berbee ML. 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.

Marshall WL, Celio G, McLaughlin DJ, Berbee ML. 2008. Multiple isolations of a culturable, motile Ichthyosporean (Mesomycetozoa, Opisthokonta), Creolimax fragrantissima n. gen., n. sp., from marine invertebrate digestive tracts. *Protist* 159:415–33.

Matz MV, Frank TM, Marshall NJ, Widder EA, Johnsen S. 2008. Giant deep-sea protist produces bilaterian-like traces. *Curr Biol* 18:1849–54.

McGuire AM, Pearson MD, Neafsey DE, Galagan JE. 2008. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol* 9:R50.

Medina M, Collins AG, Taylor JW, Valentine JW, Lipps JH, Amaral-Zettler L, Sogin ML. 2003. Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa. *International Journal of Astrobiology* 2:203–211.

Mendoza L, Taylor JW, Ajello L. 2002. The class mesomycetozoea: a heterogeneous group of microorganisms at the animal-fungal boundary. *Annu Rev Microbiol* 56:315–44.

Michod RE, Herron MD. 2006. Cooperation and conflict during evolutionary transitions in individuality. *J Evol Biol* 19:1406–9.

Michod RE. 2007. Evolution of individuality during the transition from unicellular to multicellular life. *Proc Natl Acad Sci U S A* 104:8613–8.

Mikhailov K, Konstantinova A, Nikitin M. 2009. The origin of Metazoa: a transition from temporal to spatial cell differentiation. *Bioessays* 31:758–768.

Morris SC. 1989. Burgess Shale faunas and the Cambrian explosion. *Science* 246:339–46.

Mshigeni K, Lorri W. 1977. Spore germination and early stages of development in Hypnea musciformis (Rhodophyta, Gigartinales). *Marine Biology* 42:161–164.

N

Nedelcu AM, Michod RE. 2006. The evolutionary origin of an altruistic gene. *Mol Biol Evol* 23:1460–4.

Newman SA, Bhat R. 2009. Dynamical patterning modules: a "pattern language" for development and evolution of multicellular form. *Int J Dev Biol* 53:693–705.

Newman SA. 2012. Physico-Genetic Determinants in the Evolution of Development. *Science* 338:217–219.

Newman SA, Forgacs G, Muller GB. 2006. Before programs: the physical origination of multicellular forms. *Int J Dev Biol* 50:289.

Newman SA. 2005. The pre-Mendelian, pre-Darwinian world: shifting relations between genetic and epigenetic mechanisms in early multicellular evolution. *J Biosci* 30:75–85.

Nichols SA, Roberts BW, Richter DJ, Fairclough SR, King N. 2012. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/β-catenin complex. *Proc Natl Acad Sci U S A*. 109:13046-51

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463:457–63.

O

Okasha S. 2006. Evolution and the Levels of Selection. Oxford Univ Press

Owczarzak A, Stibbs HH, Bayne CJ. 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.

P

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40:1413–5.

Paps J, Medina-Chacón LA, Marshall W, Suga H, 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

Paps J, Ruiz-Trillo I. 2010. Animals and Their Unicellular Ancestors. Encyclopedia of Life Sciences (ELS)

Peterson KJ, Cotton JA, Gehling JG, Pisani D. 2008. The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. *Philos Trans R Soc Lond B Biol Sci* 363:1435–43.

Philippe H, Derelle R, Lopez P, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* 19:706–12.

Pick KS, Philippe H, Schreiber F, et al. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol Biol Evol* 27:1983–7.

Pires-daSilva A, Sommer RJ. 2003. The evolution of signalling pathways in animal development. *Nat Rev Genet* 4:39–49.

Porter S, Meisterfeld R, Knoll A. 2003. Vase-shaped microfossils from the Neoproterozoic Chuar Group, Grand Canyon: a classification guided by modern testate amoebae. *J Paleontol* 77:409–429.

Prochnik SE, Umen J, Nedelcu AM, et al. 2010. Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science* 329:223–6.

Pugacheva EN, Roegiers F, Golemis EA. 2006. Interdependence of cell attachment and cell cycle signaling. *Curr Opin Cell Biol* 18:507–515.

Putnam NH, Srivastava M, Hellsten U, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317:86–94.

Q

Queller DC. 2000. Relatedness and the fraternal major transitions. *Philos Trans R Soc Lond B Biol Sci* 355:1647–55.

R

Raghu-kumar S. 1987. Occurrence of the Thraustochytrid, Corallochytrium limacisporum gen. et sp. nov. in the Coral Reef Lagoons of the Lakshadweep Islands in the Arabian Sea. *Botanica Marina* 30:83–89.

Rainey PB, Kerr B. 2010. Cheats as first propagules: A new hypothesis for the evolution of individuality during the transition from single cells to multicellularity. *Bioessays* 32:872–880.

Ramani AK, Calarco JA, Pan Q, et al. 2011. Genome-wide analysis of alternative splicing in Caenorhabditis elegans. *Genome Res* 21:342–348.

Retallack GJ. 2012. Ediacaran life on land. *Nature* 493:89–92.

Rokas A. 2008. The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annu Rev Genet* 42:235–51.

Romero PR, Zaidi S, Fang YY, et al. 2006. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A* 103:8390–5.

Ruiz-Trillo I, Burger G, Holland PWH, King N, Lang BF, Roger AJ, Gray MW. 2007. The origins of multicellularity: a multi-taxon genome initiative. *Trends Genet* 23:113–8.

Ruiz-Trillo I, Inagaki Y, Davis LA, Sperstad S, Landfald B, Roger AJ. 2004. Capsaspora owczarzaki is an independent opisthokont lineage. *Curr Biol* 14:R946–R947.

Ruiz-Trillo I, Roger AJAJ, Burger G, Gray MW, Lang BF. 2008. A Phylogenomic Investigation into the Origin of Metazoa. *Mol Biol Evol* 25:664–72.

S

Salazar-Ciudad I. 2003. Mechanisms of pattern formation in development and evolution. *Development* 130:2027–2037.

Sanderson M. 2003. Molecular data from 27 proteins do not support a precambrian origin of land plants. *Am J Bot* 90:954–956.

Sansjofre P, Ader M, Trindade RIF, Elie M, Lyons J, Cartigny P, Nogueira CR. 2011. A carbon isotope challenge to the snowball Earth. *Nature* 478:93–6.

Schaap P. 2011. Evolutionary crossroads in developmental biology: Dictyostelium discoideum. *Development* 138:387–96.

Seravin L, Gudkov A. 2005. Amoeboid properties of cells during early morphogenesis and the nature of a possible protozoan ancestor of Metazoa. *Zhurnal Obshchei Biologii* 66:212–223.

Shalchian-Tabrizi K, Minge MA, Espelund M, Orr R, Ruden T, Jakobsen KS, Cavalier-Smith T. 2008. Multigene phylogeny of choanozoa and the origin of animals. *PLoS One* 3:e2098.

Silberfeld T, Leigh JW, Verbruggen H, Cruaud C, De Reviers B, Rousseau F. 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.

Simpson TL. 1984. The cell biology of sponges. Springer-Verlag

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.

Sperling E a, Peterson KJ, Pisani D. 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. *Mol Biol Evol* 26:2261–74.

Srivastava M, Begovic E, Chapman J, et al. 2008. The Trichoplax genome and the nature of placozoans. *Nature* 454:955–960.

Srivastava M, Simakov O, Chapman J, et al. 2010. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* 466:720–726.

Stajich JE, Berbee ML, Blackwell M, Hibbett DS, James TY, Spatafora JW, Taylor JW. 2009. The fungi. *Curr Biol* 19:R840–5.

Steenkamp ET, Wright J, Baldauf SL. 2006. The protistan origins of animals and fungi. *Mol Biol Evol* 23:93–106.

Stibbs HH, Owczarzak A, Bayne CJ, DeWan P. 1979. Schistosome sporocyst-killing amoebae isolated from Biomphalaria glabrata. *J Invertebr Pathol* 33:159–170.

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.

Suga H, Ruiz-Trillo I. 2013. Development of ichthyosporeans sheds light on the origin of metazoan multicellularity. *Dev Biol*. [published online ahead of print January 18 2013]

Sumathi JC, Raghukumar S, Kasbekar DP, 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.

Sutter D, Vyver G. 1977. Aggregative Properties of Different Cell Types of the Fresh-Water Sponge Ephydatia fluviatilis Isolated on Ficoll Gradients. *Roux's Archives of Developmental Biology* 181:151–161.

T

Technau U. 2001. Brachyury, the blastopore and the evolution of the mesoderm. *Bioessays* 23:788–94.

Telford MJ. 2006. Animal phylogeny. *Curr Biol* 16:R981–5.

Tomitani A, Knoll AH, Cavanaugh CM, Ohno T. 2006. The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. *Proc Natl Acad Sci U S A* 103:5442–7.

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 BF, Roger AJ, Shalchian-Tabrizi K, Ruiz-Trillo I. 2012. Phylogenetic Relationships within the Opisthokonta Based on Phylogenomic Analyses of Conserved Single-Copy Protein Domains. *Mol Biol Evol* 29:531–44.

V

Velicer GJ, Vos M. 2009. Sociobiology of the myxobacteria. *Annu Rev Microbiol* 63:599–623.

Vogt G, Rug M. 1999. Life stages and tentative life cycle of Psorospermium haeckeli, a species of the novel DRIPs clade from the animal-fungal dichotomy. *J Exp Zool* 283:31–42.

W

Westbrook MW. 2011. Introns and alternative splicing in choanoflagellates. PhD Thesis. University of California, Berkeley.

Wolpert L, Szathmary E. 2002. Evolution and the egg. *Nature* 420:745.

X

Xie X, Wang G, Pan G, Gao S. 2010. Variations in morphology and PSII photosynthetic capabilities during the early development of tetraspores of Gracilaria vermiculophylla (Ohmi) Papenfuss (Gracilariales, Rhodophyta). *BMC Dev Biol* 10:1–12.

Y

Yamada A, Martindale MQ, Fukui A, Tochinai S. 2010. Highly conserved functions of the Brachyury gene on morphogenetic movements: Insight from the early-diverging phylum Ctenophora. *Dev Biol* 339:212–222.

Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev EV. 2012. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. *Genes Dev* 26:1209–23.

Yin L, Zhu M, Knoll AH, Yuan X, Zhang J, Hu J. 2007. Doushantuo embryos preserved inside diapause egg cysts. *Nature* 446:661–3.

Z

Zettler LA, Nerad TA, O'Kelly CJ, Sogin ML. 2001. The Nucleariid Amoebae: More Protists at the Animal-Fungal Boundary. *J Eukaryot Microbiol* 48:293–297.

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.

Resum en català

Introducció

L'evolució de la multicel·lularitat

La capacitat de les cèl·lules d'actuar com a individus multicel·lulars integrats és una important força que ha estructurat la biosfera terrestre, especialment durant el Fanerozoic, des de que plantes, animals i fongs aparegueren. La transició a la multicel·lularitat implica no només un cert grau de comunicació i adhesió cel·lulars, sinó també un canvi fonamental en la naturalesa de la individualitat. Les cèl·lules perden la identitat independent i, en canvi, serveixen com a part d'un ens individual major i més incloent, definit per una sèrie de propietats emergents.

Però no totes les "multicel·laritats" són iguals. En primer lloc, hem de diferenciar entre multicel·lularitat clonal i multicel·lularitat agregativa. En la primera, totes les cèl·lules del grup deriven d'una fundadora inicial que duu a terme successius cicles de divisió cel·lular, resultant en una població de cèl·lules genèticament idèntiques. En la segona, en canvi, cèl·lules genèticament diferents s'uneixen les unes a les altres (Grosberg i Strathmann 2007). En aquest cas, la competència entre les cèl·lules suposa un fort desavantatge que fa que l'agregat sigui evolutivament inestable (Aanen et al. 2008; Newman 2012). De fet, els llinatges eucariòtics on s'observa multicel·lularitat agregativa només formen agregats cel·lulars de forma transient, mai ens multicel·lulars estables. Aquest és el cas dels dictiostèl·lids (Amoebozoa) (Schaap 2011), de les amebes acrasides (Heterolobosea, Discicristata, Discoba) (Brown et al. 2011; Adl et al. 2012), *Guttulinopsis vulgaris* (Cercozoa, Rhizaria) (Brown et al. 2012), del gènere *Sorogena* (Ciliata, Alveolata) (Lasek-Nesselquist i Katz 2001), del nucleari d *Fonticula alba* (Holomycota,Opisthokonta) (Brown et al. 2009) i del gènere *Sorodiplophrys* (Labyrinthulomycetes, Heterokonta) (Dykstra i Olive 1975). En procariotes, trobem multicel·lularitat agregative en Myxobacteria (Velicer i Vos 2009). D'aquí en endavant en aquesta secció ens referirem només a la multicel·lularitat clonal.

Definida només com a cèl·lules que romanen unides després de dividir-se clonalment, el concepte "multicel·lularitat" pot ser aplicat a filaments, boles o làmines de cèl·lules originades per divisió mitòtica des d'un únic progenitor cel·lular. La diferenciació somàtica i la formació de cèl·lules reproductores s'observa en alguns casos (per exemple en les algues del gènere Volvox (Prochnik et al. 2010)), però patrons de diferenciació

més complexos són inexistents (Knoll 2011). Aquesta multicel·lularitat simple es troba inclús en bacteris; molts cianobacteris, per exemple, formen filaments que contenen dotzenes i inclús centenars de cèl·lules, algunes de les quals es diferencien en tipus cel·lulars especialitzats, incloent heterocists fixadors de nitrogen i espores de reistència (Bonner 1998; Tomitani et al. 2006). Altres llinatges bacterians que formen filaments són els bacteris verds fotosintètics no del sofre (per exemple, *Chloroflexus*), els proteobacteris oxidadors de sulfur (per exemple, *Beggiatoa* i *Thioploca*), alguns bacteris magnetotàctics (Keim et al. 2004), i una remarcable diversitat d'actinobacteris (per exemple, *Streptomyces*). No obstant, i malgrat la seva enorme importància ecològica, els procariotes multicel·lulars representen un cul-de-sac evolutiu: els grans i complexos organismes multicel·lulars són el domini exclusiu dels eucariotes (Butterfield 2009).

Hi ha, doncs, una significativa "àrea gris" en la definició de multicel·lularitat, representada per nivells críptics d'integració intercel·lular. En canvi, els organisme multicel·lulars complexos mostren evidència no només d'adhesió cèl·lula-cèl·lula, sinó també comunicació intercel·lular i, típicament, diferenciació de teixits controlada per xarxes de gens reguladors, el que coneixem com a programa de desenvolupament. Aquest criteri només és aplicable estrictament als grups Embryophyta (plantes), Metazoa (animals), Phaeophyta (algues brunes), algues vermelles dels grups Bangiales i Florideophycidae i alguns fongs (però en aquest cas, no presenten cells individualitzades i no hi ha un clar desenvolupament embrionari; molts d'ells, a més, són secundàriament unicel·lulars o bé formen filaments simples).

Des d'un punt de vista teòric, la transició a una multicel·lularitat clonal complexa implica un canvi jeràrquic en el que s'anomena individualitat darwiniana (definida com a tal quan poblacions d'aquests individus presenten variació, heretabilitat i reproducció). Durant aquest canvi l'eficàcia biològica (coneguda en anglès com a *fitness*) és re-assignada des d'unitats de baix nivell (les quals cedeixen la seva capacitat de reproduir-se com a unitats independents) a unitats més inclusives, que passen a reproduir-se com un "tot" major (Michod i Herron 2006; Michod 2007; Rainey i Kerr 2010). L'anomenada teoria de la selecció multi-nivell ha proporcionat un marc conceptual per entendre aquest transició, disseccionant-la en petits estadis analitzables. Aquests estadis inclourien (Okasha 2006; Michod 2007; Rainey i Kerr 2010):

- Avantatge inicial de la formació de grups. Grups de cèl·lules indiferenciades poden emergir donades les condicions ecològiques adequades, cooperant per a produir un major nombre de cèl·lules individuals en la generació següent. Aquí, l'eficàcia biològica del grup no és més que la mitjana o el sumatori de les eficàcies biològiques de les cèl·lules individuals que conformen el grup i, típicament, la selecció natural a nivell de grup afecta només els pocs trets que afavoreixen la formació del grup (per exemple, la producció de substàncies adhesives que mantenen les cèl·lules juntes). Aquest estadi és explicable a través de la teoria de selecció per parentesc.

- Origen de l'altruisme reproductiu dins el grup. En aquest punt, l'eficàcia biològica al nivell superior és desacoblada de l'eficàcia als nivells inferiors. Aquest és el gran salt i el més difícil d'explicar. Dues teories intenten explicar l'evolució de la reproducció col·lectiva (Rainey i Kerr 2010)
    - La primera teoria proposa que els primers propàguls reproductius s'haurien originat a partir de "tramposos" que apareixen de forma habitual en el grup, guanyant per mutació una taxa reproductiva major que la resta de cèl·lules i perdent el comportament cooperatiu. El grup col·lapsaria per l'acció d'aquestes cèl·lules tramposes i les cèl·lules (bàsicament tramposos, ja que s'hauran reproduït més) es disperaran. Després, una altra mutació revertiria el fenotip "trampós" al fenotip "col·laboratiu", tancant el cicle i produint de nou un grup cooperatiu. Aquest seria un cicle multicel·lular incipient; però aquest model té l'inconvenient de ser requerir una mutació en cada pas del cicle, diem que és un model interromput.
    - Alternativament, el model ininterromput proposa que determinats llinatges cel·lulars, degut a mutacions, s'enretirarien altruísticament de la línia reproductiva i, per contra, generarien algun benefici somàtic per al grup (Queller 2000). Aquest model tan sols requereix d'una sola mutació inicial, que produiria cèl·lules somàtiques. Un cicle tal és més fàcil d'imaginar si imaginem que la mutació altruista que elimina la cèl·lula de la línia germinal és expressada només condicionalment, ja sigui en el temps o en l'espai; aquesta condicionalitat ens duu directament al següent punt: l'adveniment un programa de desenvolupament. Un interessant

article de Nedelcu i Michod (2006) mostrà com l'altruisme reproductiu podria haver evolucionat a partir de la co-opció d'un gen regulador del cicle vital: en l'alga verda unicel·lular *Chlamydomonas reindhartii*, un senyal ambiental (en aquest cas la llum) activa el gen regA (un facto de transcripció) que indueix un canvi que maximitza la supervivència de la cèl·lula el detriment de la reproducció. L'ortòleg d'aquest gen en l'alga verda colonial *Volvox carteri* treballa sota control d'un programa del desenvolupament per tal d'especificar cèl·lules somàtiques, les quals contribueixen a la supervivència de la colònia però no es reprodueixen.

- Diferenciació cel·lular. La formació condicional de la línia germinal és la primera d'una successió de tipus i comportament cel·lulars que poden evolucionar un cop la individualitat multicel·lular ha estat establerta, especialment a mesura que el tamany del grup s'incrementa.

Finalment, el procés duu a l'establiment d'un cicle vital multicel·lular clonal, on l'organisme multicel·lular passa per un coll d'ampolla unicel·lular a cada generació des del qual, més tard, l'adult és format per un procés de desenvolupament embrionari. Això assegura que cada generació comença amb un grup de cèl·lules que comparteixen tots els seus gens per descendència. Aquest procés fa els organismes més "evolucionables" (Wolpert i Szathmary 2002), ja que la variació entre les cèl·lules que sorgeix durant l'estadi adult és distribuïda entre la descendència, en comptes de continuar acumulada com a variació intra-grup. Una predicció d'aquest model és que, almenys inicialment, aquests organismes serien petits, ja que més divisions cel·lulars signifiquen més probabilitat de mutació i més cicles de divisió en els quals els mutants poden expressar el seu avantatge replicatiu egoista (Queller 2000). Una altra raó que suporta aquesta noció és que, almenys inicialment, els recursos com l'oxigen i els nutrients haurien d'arribar a totes les cèl·lules en el grup tridimensional simplement per difusió, la qual cosa estableix un límit en la mida que pot tenir una massa compacta de cèl·lules.

Avantatges selectius i reptes de la multicel·lularitat

Diversos beneficis selectius de la multicel·lularitat han estat proposats. En primer lloc, la multicel·lularitat és una forma eficient d'incrementar la mida. Això mateix també pot ser assolit per un creixement cel·lular hipertròfic, però hi ha limitacions físico-químiques (quocient superfície-volum, coeficients de difusió en el citoplasma,...) que suposen un límit superior a la mida màxima que una sola cèl·lula pot assolir. Una de les coses que permet una mida major és escapar de predadors heterotròfics. Això fou demostrat en un experiment clàssic (Boraas et al. 1998) en el qual l'alga verda unicel·lular *Chlorella vulgaris* era cultivada amb el flagel·lat heterotròfic *Ochromonas vallescia*. En uns pocs dies, les algues començaven a formar colònies per divisió cèl·lular incompleta. Inicialment, les colònies eren desmesuradament grans, però això dificultava l'absorció de nutrients per part de les algues individualment; per la qual cosa, finalment, la mida de la colònia s'estabilitzà al voltant de vuit cèl·lules, suficient per a evitar ser depredades i al mateix temps permetent una absorció de nutrients eficient. Després d'eliminar el predador del cultiu, les algues unicel·lulars dominaren de nou. Per tant, la mera addició d'un depredador suposà una pressió selectiva que afavoria els genotips formadors de colònies. L'origen dels eucariotes heterotròfics (c. 800 Ma) podria haver estat un important motor selectiu per a l'evolució de simples formes colonials en molts llinatges eucariotes, alguns dels quals més tard evolucionarien cap a una multicel·lularitat complexa.

En segon lloc, la multicel·lularitat ajuda a solucionar compromisos cel·lulars i metabòlics. Per exemple, la motilitat flagel·lar i la mitosi competeixen per centre organitzador de microtúbuls (MTOC) (King 2004), el qual s'usa tan per a formar el cos basal que sintetitza el flagel com per a formar el fus mitòtic usat en la segregació cromosòmica. Aquest és el motiu pel qual les cèl·lules flagel·lades dels metazous (incloent els espermatozoides, estatocists,...) mai es divideixen. Esdevenint multicel·lular ambdues activitats es fan compatibles, ja que mentre algunes cèl·lules es divideixen altres poden proporcionar motilitat flagel·lar al conjunt de l'individu. Un altre exemple és el dels cianobacteris, on la fotosíntesi i la fixació de nitrogen són incompatibles i, per això, aquestes dues activitats són dutes a terme per diferents tipus cel·lulars.

En hàbitats terrestres, la dispersió d'espores és fa difícil per la manca de corrents d'aigua. Alguns eucariotes terrestres, com els dictiostèlids o els acrasids, s'agreguen per formar cossos fructífers que milloren la dispersió d'espores (Bonner 1998). Un altre avantatge de la multicel·lularitat és una alimentació més eficient, per exemple segregant conjuntament enzims digestius (Grosberg i Strathmann 2007).

Finalment, la visió neutralista defensa que no hi ha necessitat d'adduir explicacions adaptatives perquè la multicel·lularitat, igual que altres formes de complexitat, evoluciona simplement perquè "pot". És el resultat de l'acumulació irreversible de mutacions neutrals o lleugerament recessives en mides poblacions efectives petites (Koonin 2011). Aquesta teoria argumenta que les formes més complexes no són generalment més adaptatives que les formes més simples i que aquesta "roda dentada" mutacional crea l'aparent direccionalitat en l'evolució quan, de fet, no hi ha cap millora real. Un cop la multicel·lularitat és establerta, l'única pressió selectiva és per a evolucionar mecanismes que evitin el mal funcionament del sistema (Koonin 2011).

Sigui quina sigui l'explicació evolutiva per al seu origen en cada cas, hi ha diversos funcions que els organismes multicel·lulars han de dur a terme:

- Adhesió cel·lular: es pot aconseguir mitjançant substàncies extracel·lulars, com les pectines i hemicel·luloses en plantes o les pegues basades en glicoproteïnes en fongs; o bé mitjançant proteïnes transmembrana específiques que estableixen contactes cèl·lula-cèl·lula, aquest és el cas del es cadherines i altres proteïnes d'adhesió en animals.
- Senyalització cèl·lula-cèl·lula: mecanismes que coordinen les cèl·lules i transmeten senyals que, finalment, modifiquen comportaments cel·lular estimulant programes de diferenciació. Les hormones a plantes i les vies de senyalització en animals, com Notch, Hedghog o Wnt, en són exemples. Aquests sistemes estan formats per, alternativament, un lligand extracel·lular i un receptor transmembrana (per exemple, Hedgehog i Wnt en animals o la via de brassinoesteroides en plantes); de dos receptors transmembrana contactant l'un a l'altre (per exemple la via Notch en animals); o bé d'un lligand permeable al a membrana i un receptor intracel·lular (per exemple, la via de les auxines i les gibberel·lines a plantes i les hormones esteroides a

animals). En tots els casos, segueix una cascada intracel·lular de transducció del senyal.

- Diferenciació cel·lular: les xarxes reguladores de gens són establertes per controlar la proliferació de les cèl·lules en el grup (evitant l'aparició de cèl·lules no cooperatives) i també per a especificar comportaments cel·lulars de forma espacial (en contrast amb els canvis seqüencials temporals de tipus cel·lulars en un cicle vital unicel·lular). Això s'aconsegueix mitjançant factors de transcripció que control l'expressió de conjunts concrets de gens.

Els múltiples orígens de la multicel·lularitat

Els eucariotes van aparèixer fa uns 2100 Ma (Anbar i Knoll 2002), essent majoritàriament autòtrofs inicialment. Més tard, fa aproximadament 800 Ma, trobem les primeres evidències d'eucariotes heterotròfics: carcasses fossilitzades atribuïdes a amebes loboses i altres protists (Porter et al. 2003). De fet, fa 800 Ma hi hagués un pronunciat increment de la diversitat de protists (Knoll et al. 2006), una diversificació que s'hipotetitza està lligada a canvis geoquímics. Malgrat l'atmosfera era rica en oxigen degut al gran esdeveniment oxidatiu lligat a l'aparició de la fotosíntesi, durant la major part de l'Eó Proterozoic (2500-543 Ma) els oceans eren rics en sulfit (Anbar i Knoll 2002), el qual és generalment tòxic per als eucariotes. És aproximadament 800 Ma quan les masses d'aigua canviarien cap a una composició rica en oxigen (Donoghue i Antcliffe 2010). Això hauria facilitat la diversificació dels eucariotes. A més, l'increment en els nivells d'oxigen podria haver facilitat l'evolució de la multicel·lularitat en eucariotes, ja que hagués incrementat la mida màxima permissible per a un organisme multicel·lular limitat per la difusió passiva d'aquest gas (Knoll 2011).

Els fòssil més antic conegut d'un eucariota multicel·lular data de fa uns 1200 Ma i correspon a algues vermelles del llinatge Bangiomorpha (Butterfield 2009; Knoll 2011). No obstant, la gran majoria de clades eucariotes multicel·lulars apareixeren molt més tard, després de l'increment d'oxigen de fa 800 Ma. Això inclou les algues verdes macròfites (Charales, Coeocharaetales i Zygnematales) fa 750 Ma (Becker 2012; Laurin-Lemay et al. 2012), animals fa uns 600 Ma, embriòfits fa 450 Ma (Sanderson 2003), fongs multicel·lulars 300 Ma (Stajich et al. 2009) i feòfits (algues brunes multicel·lulars) fa 130 Ma (Silberfeld et al. 2010).

Independentment de quan va ocórrer, és clar que la multicel·lularitat ha evolucionat independentment en repetides ocasions durant l'evolució eucariota. Una estima conservadora calcula 16 evolucions independents (King 2004; Grosberg i Strathmann 2007), encara que estimes més recents eleven la xifra fins a 22 (Adl et al. 2007; Knoll 2011). En qualsevol cas, com prèviament hem esmentat, només en cinc casos podem parlar de multicel·lularitat complexa: embriòfits, metazous, feòfits, algues vermells dels grups Bangiales i Florideophycidae, i alguns fongs.

En el cas dels embriòfits i els metazous, el seu desenvolupament embrionari està ben descrit els mecanismes moleculars implicats són coneguts. Per contra, poc en sabem de les algues vermelles i brunes. Hi ha evidències de desenvolupament embrionari en ambdós casos (Mshigeni i Lorri 1977; Bouget et al. 1998; Xie et al. 2010); però no en sabem res del mecanismes moleculars que controlen aquest desenvolupament. La multicel·lularitat en fongs és, paradoxalment, poc coneguda i força problemàtica de categoritzar. De fet, encara que els fongs multicel·lulars complexos conformen fins al 80 o 90% de la diversitat coneguda de fongs, no sabem pràcticament res del seu desenvolupament i, inclús, tan sols uns pocs genomes de fongs no-unicel·lulars han estat seqüenciats (Kües 2000, Stajich et al. 2009).

Els modes de nutrició poden ser un factor important per explicar les diferències entre els organismes multicel·lulars complexos. Per exemple, els animals són els únics organismes fagotròfics amb multicel·lularitat complexa. Ha estat suggerit que això és un factor crucial que explica el gran salt entre metazous i no-metazous, sense formes intermèdies com les que, en canvi, sí observem en els llinatges propers a les plantes (Cavalier-Smith 2012). En autòtrofs i osmòtrofs, les cèl·lules simplement necessitarien unir-se les unes a les altres produint ciments extracel·lulars per tal d'esdevenir multicel·lulars, però per als fagòtrofs l'agregació cel·lular interfereix severament amb el mecanisme d'alimentació. L'origen dels animals fou, doncs, inherentment més difícil que en altres casos degut al requeriment de passar d'una fagotrofia unicel·lular a una fagotrofia mitjançant un sistema digestiu. El problema es podria solucionar també amb una organització corporal com la de les esponges, que mantenen un mode d'alimentació tipus "protist". En aquest sentit, la forma d'alimentació dels coanoflagel·lats, amb un flagel envoltat de microvil·lis, fou un important pas ja que aquest tipus d'organització cel·lular es pot acomodar fàcilment en una unitat pluricel·lular amb cèl·lules que

s'alimenten autònomament (com el que observem avui en dia en les colònies de coanoflagel·lats).

L'origen dels metazous

Datar l'origen dels animals ha estat durant molt de temps un tema intensament disputat i encara ho segueix essent. Els fòssils càmbrics tipus Burguess Shale han revelat una extraordinària explosió de formes animals que ocorregué fa uns 542 Ma, la majoria del es quals es poden identificar amb seguretat com a membres de fílums i classes animals actuals (Morris 1989; Erwin et al. 2011). En canvi, els estudis de rellotge molecular infereixen un origen molt anterior dels animals, fa uns 800 Ma (Peterson et al. 2008). Entre aquests dos extrem, trobem l'enigmàtica fauna Ediacara, que visqué en el Proterozoic tardà (579-565 Ma), després de la glaciació Gaskiers de mitjans del període Ediacara. Aquesta fauna ha estat temptativament classificada en 9 clades (Erwin et al. 2011), incloent Arboreomorfs, Rangeomorfs (els més antics), Triradialomorfs, Bilateralomorfs, Erniettomorfs, Kimberellomorfs, Pentaradialomorfs, Dickinsoniomorfs i Tetraradialomorfs. Alguns d'aquests fòssils han estat descrits com a pertanyents a fílums actuals. Per exemple *Cloudina* seria un cnidaria, els Kimberellomorfs els primers bilaterals (probablement mol·luscs) i els Dickinsonomorfs, serien placozous o eumetazous basals.

La presència de formes bilaterals i superficialment segmentades significaria que la maquinària molecular de la bilateralitat animal era present, almenys, fa uns 575 Ma. Però l'assignació i classificació d'aquests fòssils és encara molt controvertida. Per exemple, un treball recent suggereix que alguns fòssils ediacares, incloent *Dickinsonia*, podrien ser de fet organismes tipus liquen o colònies microbianes que viurien en terra (Retallack 2012). Un altre cas són els polèmics fòssils de Doushantuo, que daten de l'Ediacara tardà i foren inicialment descrit com a embrions en diferents fases de desenvolupament (Yin et al. 2007). No obstant, treballs recents han demostrat que, de fet, no sóm embrions animals si no probablement protists encistats d'afinitat incerta (Huldtgren et al. 2011). A més, anàlisis de sincrotró han revelat que inclús aquells exemplars descrits com a embrions gastrulant (un tret diferencial dels embrions animals) són degut a artefactes diagenètics (Cunningham et al. 2012). Així doncs, l'embriologia ediacara és encara un misteri.

El 2009, Love et al. van reportar la identificació d'abundants 24-isopropilcolestans sedimentaris (l'hidrocarbur que resta dels esterols C30 produïts per les demosponges) que dataven d'abans del final de la glaciació Marionana (635 Ma). Un estudi anterior de Kodner et al. (2008) mostrà que aquest tipus d'esterols no són produïts pels coanoflagel·lats (els parents unicel·lulars més propers dels animals), suportant la noció de que la presència de 24-isopropicolestans és un bon indicatiu de la presència de demosponges. L'origen primerenc de les demosponges (que no produeixen espícules) en comparació amb altres esponges, explicaria perquè els primers fòssils d'espícules mineralitzades d'esponges no apareixen fins molt més tard, fa uns 544 Ma (Brasier et al. 1997). Finalment, un estudi recent (Maloof et al. 2010) descriu possibles fòssils d'esponges de fa uns 660 Ma. Totes aquestes evidències suporten la idea que els metazous s'originaren, almenys, durant el Criogènic (entre 850 i 635 Ma), la qual cosa coincideix amb les estimes dels rellotges moleculars.

L'origen dels programes genètics de desenvolupament dels metazous

L'origen del desenvolupament animal és un altre tema que ha generat una considerable controvèrsia. Els gradients superfície-interior d'oxigen i nutrients, que són inherents a qualsevol estructura tridimensional, podrien haver estat una primera manera d'induir diferenciació entre cèl·lules interiors i exteriors, especialment considerant la importància que la manca de nutrients i d'oxigen tenen com a senyals per a la diferenciació en molts eucariotes (Aguirre et al. 2005, Loenarz et al. 2011). Aquests senyals podrien haver estimulats programes de diferenciació per tal de solucionar els problemes de difusió en la colònia. Una primera forma de fer-ho és establir canals proteics entre cèl·lules per facilitar la difusió. Però una forma encara més eficient és crear una cavitat interior plena de fluid on abocar-hi activament nutrients (Beaumont 2009). Més tard, teixit especialitzats (sistemes circulatoris) evolucionarien per a dur a terme aquestes funcions encara més eficientment. Així, la morfogènesi hauria estat suficient per a solucionar les inherents limitacions metabòliques dels primers animals.

La qüestió és com els programes genètics que controlen la morfogènesi apareixeren, independentment de les raons selectives que puguem adduir per les quals van aparèixer. Per Newman i col·laboradors (Newman 2005; Newman et al. 2006; Newman i Bhat 2009; Newman 2012), a l'inici de l'evolució animal la morfogènesi era un procés

altament plàstic i basat en capacitats auto-organitzatives (com ara l'adhesió cel·lular diferencial, l'inhibició lateral o l'oscil·lació bioquímica) que depenien una maquinària genètica molt bàsica. Els programes de desenvolupament regulats genèticament apareixerien més tard en l'evolució com a "accidents congelats" de selecció estabilitzadora, els quals convertiren els organismes en menys inter-convertibles i plàstics, creant un desenvolupament estereotipat. Per tant, els mecanismes de control genètic de la morfogènesi esdevingueren cada vegada més prominents durant l'evolució animal, però les arrels de la multicel·lularitat animal no es poden trobar en la complexitat de mecanismes de formació de patró que troben avui en dia en els metazous.

Una visió alternativa que sostenen Davidson i col·laboradors (Davidson i Erwin 2006) suggereix que la morfogènesi i la formació de patró espacial de les cèl·lules estava basat des de bon començament en circuits genètics específics de cada tipus cel·lular que havien evolucionat amb anterioritat. Per tant, la conservació dels plans corporals filètics serien deguts a la retenció, des de temps pre-càmbrics, d'aquests circuits regulatoris. La teoria també prediu que només elements perifèrics d'aquestes xarxes de regulació gènica canviarien durant l'evolució tardana dels metazous.

No obstant, ambdues visions són fins a cert punt complementàries. Mentre que la primera fa èmfasi en el rol dels mecanismes genètics que mobilitzen processos físics i dona un rol secundari/posterior als factors de transcripció del desenvolupament, la segona emfatitza el rol d'aquestes xarxes de factors de transcripció com a essencial per explicar la diversificació dels metazous. Es podrien considerar es circuïts regulatoris generals de Davidson com el primer producte dels "accidents congelats" de Newman.

En qualsevol cas, un cop els programes de desenvolupament multicel·lular s'establiren, foren progressivament "tancats" per l'acumulació de circuits genètics de reforç i modulació fina. La multicel·lularitat embrionària establí un nivell de complexitat irreductible, on la proliferació cel·lular incontrolada, els erros en formació de patró o diferenciació cel·lular i la mort cel·lular no programada imposen severs reduccions de l'eficàcia biològica.

Els parents unicel·lulars dels animals

Tenir un marc filogenètic robust dels metazous i els seus parents unicel·lulars és condició indispensable per a fer estudis de genòmica comparada. Però aquesta no és pas una qüestió trivial i és un àmbit en continu canvi i actualització. La nostra descripció aquí es basa en els dos anàlisis més recents i acurats sobre el tema (Torruella et al. 2012; Paps et al. 2013). Així doncs, els metazous formen part d'una sèrie d'assemblatges filogenètics més amplis, incloent (del més ampli al més restringit):

- Unikonts (Cavalier-Smith 2002): El nom "Unikonta" fa referència a l'estructura de l'aparell flagel·lar, que en els unikonts seria ancestralment només un centríol i un cili. Això en contrast amb els bikonts (el grup que inclou plantes, alveolats,...), que en tindrien dos. Una gran diversitat d'estudis filogenètics han suportat aquesta divisió (Ruiz-Trillo et al. 2008; Derelle i Lang 2012). Hi ha 3 grans llinatges dins els unikonts: Amoebozoa (que inclou amebes diverses, inloent els dictiostèlids), Apusozoa i Opisthokonta (el grup que inclou els animals, els fongs i els seus respectius parents unicel·lulars).
- Opisthokonta (Adl et al. 2005): són unikonts amb, ancestralment, un únic flagel posterior almenys en algun estadi vital (tot i que en molts casos s'ha perdut). Aquest llinatge ha estat verificat per nombrosos estudis filogenètics (Lang et al. 2002; Medina et al. 2003; Steenkamp et al. 2006; Ruiz- Trillo et al. 2008; Torruella et al. 2012) i inclou dos grans grups: Holomycota (els fongs i els seus parents unicel·lulars) i Holozoa (els animals i els seus parents unicel·lulars).
- Holozoa (Lang et al. 2002): inclou els animals i els seus parents unicel·lulars, els quals formen 3 llinatges independents: Ichthyosporea, Filasterea (on s'inclou *Capsapsora owczarzaki*) i Choanoflagellata (ordenats de més llunyans a més propers als animals).
- Filozoa (Shalchian-Tabrizi et a. 2008): seria el grup format per Filasterea, Choanoflagellata i els animals.

Els coanoflagel·lats són els parents unicel·lulars més propers dels animals, una posició que ha estat confirmada els últims anys per diversos estudis de filogènia molecular (Carr i Leadbeater 2008; Torruella et al. 2012). Són flagel·lats heterotròfics majoritàriament marins i alguns d'ells amb capacitat per formar colònies per divisió clonal (Fairclough et al. 2010; Dayel et al. 2011). Tenen un sol flagel envoltat per un

collar de microvilli d'actina que usen per atrapar les partícules atretes pel corrent generat pel flagel. Els coanoflagel·lats foren els primers holozous unicel·lulars que es van estudiar en el context de l'origen de la multicel·lularitat animal (King et al. 2003; King 2004; King 2005) i, recentment, els genomes de dues espècies (*Monosiga brevicollis* i *Salpingoeca rosetta*) han estat seqüenciats (King et al. 2008; Fairclough et al. 2013).

Els filasteris són un grup definit per filogènia molecular i inclou només dues espècies: *Capsaspora owczarzaki* i *Ministeria vibrans* (Ruiz-Trillo et al. 2004; Shalchian-Tabrizi et al. 2008). El seu nom prové dels llargs i fins tentacles que ambdues espècies presenten. *Capsaspora* és un ameboide que fou aïllat del cargol *Biomphalaria glabrata* (Stibbs et al. 1979; Owczarzak et al. 1980) i que inicialment fou anomenat *Nuclearia* sp i considerad un nucleárid (els parents unicel·lulars dels fongs) (Zettler et al. 2001). *Ministeria vibrans*, per la seva banda, són organismes marins sèssils de vida lliure.
Els ictiosporis són organismes unicel·lulars en la majoria dels casos paràsits o simbionts d'animals i alguns d'ells formen colònies sincitials/esporangis (Mendoza et al. 2002; Marshall et al. 2008; Marshall i Berbee 2011). Aquestes colònies es formen per creixement hipertròfic d'una cèl·lula inicial i posterior cel·lularització del sinciti per formar endospores. Hi ha dos grans grups d'ictiosporis: Rhinosporidiacae/Dermocystida i Ichthyophonae. Els primer són paràsits humans, tenen un flagel posterior i formen esporangis (Mendoza et al. 2002). Els segons inclouen paràstis de peixos i artròpodes i main presenten flagel; en alguns casos tenen formes ameboides i molt, per exemple *Sphaeroforma arctica* i *Creolimax fragrantissima* (Marshall et al. 2008), formen colònies. Algunes filogènies moleculars agrupen els ictiosporis amb els filasteris (Ruiz-Trillo et al. 2008), però estudis més recents mostres que Ichthyosporea és un llinatge independent (Torruella et al. 2012), el tercer llinatge més proper als animals.

Genòmica comparada i el genoma de l'Urmetazou

L'estudi de l'origen els metazous ha experimentat una autèntica revolució amb l'adveniment de les tècniques de seqüenciació automàtica de genomes. Els anàlisis de genomes de metazous basals com els cnidaris *Hydra magnipapillata* i *Nematostella vectensis* (Putnam et al. 2007; Chapman et al. 2010), el placozou *Trichoplax adhaerens* (Srivastava et al. 2008) i la demosponja *Amphimedon queenslandica* (Srivastava et al.

2010) proporcionaren la primera aproximació genòmica a la definició de la maquinària molecular comú a tots els animals (és a dir, els gens que definirien la condició d'animal i que proporcionen funcions essencials per a la multicel·lularitat). Aquesta maquinària genètica específica i comú a tots els animals incloïa gens d'adhesió com cadherines i integrines, diversos receptors de vies de senyalització (Notch, receptors tirosina quinasa,...) i cascades de transducció de senyal (la vida d'Hippo,...), així com molts factors de transcripció involucrats en el desenvolupament animal (T-box, Runx, NFkappaB, Myc/Max,...).

Però aquesta aproximació representa només la meitat de la història, la qual seria completa si comparem una mostra filogenèticament àmplia de genomes d'animals amb genomes de parents unicel·lulars d'aquests (coanoflagel·lats, filasteris i ictiosporis), només així podrem definir el contingut genòmic de l'ancestre dels animals (Ruiz-Trillo et al. 2007).

La seqüenciació del genoma del coanoflagel·lat *Monosiga brevicollis* marcà un abans i un després en l'estudi de l'origen de la multicel·lularitat animal (King et al. 2008) i demostrà la potència de la comparació de genomes d'holozous. Aquest estudi mostrà la presència en aquest organisme unicel·lular de gens que es creien exclusius dels metazous, com ara tirosina quinases i cadherines (Abedin i King 2008; Manning et al. 2008).

No obstant, i donat que la pèrdua gènica és un fenomen molt habitual en els genomes eucariotes (Zmasek i Godzik 2011), l'estudi del genome de filasteris i ictiosporis pot proporcionar una inferència més robusta del genoma de l'Urmetazou. En aquest sentit, el projecte UNICORN és una iniciativa dissenyada per a seqüenciar diversos organismes en posicions filogenètiques clau en l'arbre dels opisthokonts, incloent-hi el filasteri *Capsaspora owczarzaki* (Ruiz-Trillo et al. 2007).

El major desavantatge per a l'estudi d'aquestes espècies, en contrast amb el coanoflagel·lats (que han estat estudiats per més d'un segles (James-Clark 1866)), és la manca de coneixement disponible sobre la seva biologia bàsica.

És en aquest context que aquesta tesi s'ha desenvolupat, estudiant *Capsaspora owczarzaki* tan a nivell genòmic com a nivell de biologia cel·lular bàsica.

Capsaspora owczarzaki i el seu genoma

El 1979, una ameba simbiont fou aïllada del mol·lusc *Biomphalaria glabrata* (Stibbs et al. 1979; Owczarzak et al. 1980) a Puerto Rico, en el context de l'estudi d'aquest cargol com a hoste intermediari del trematode patogen d'humans *Schistosoma mansoni*. Es reportà que les amebes atacaven els esporocists del trematode i que tenien uns 3-5 μm amb un nucli que ocupava aproximadament 1/2-1/3 del diàmetre cel·lular, amb llargs pseudopodis no ramificats i nombrosos fagosomes, vacuoles lipídiques i grànuls de glicogen. Les cèl·lules s'encistaven en determinades condicions, formant cists d'uns 4-5 μm amb una paret doble. Aquests treballs originals no descriuen formalment l'espècie i, basant-se en caràcters morfològics, la situen dins el gènere *Nuclearia*.

No hi ha cap altra publicació sobre aquest organisme fins 20 anys més tard, el 2001, quan Zettler et al. publiquen la primera filogènia molecular on s'inclou *Capsaspora owczarzaki*, en el marc d'un estudi sobre el gènere *Nuclearia*. En el seu resultat *Capsaspora owczarzaki* no s'agrupa amb les altres espècies de *Nuclearia,* però l'afinitat amb altres opisthokonts és incerta i no la reclassifiquen. Un any més tard, *Capsaspora owczarzaki* és descrit formalment i batejat. A més, s'aïlla novament l'organisme, aquesta vegada de cargols recol·lectats al Brasil. El seu estudi mostra que *Capsaspora owczarzaki* és proper al ictiosporis. Després s'han succeït diversos estudis que, alternativament, han proposat *Capsaspora owczarzaki* com a un llinatge independent o com a grup germà dels ictiosporis. Els estudis més recents (Torruella et al. 2012) corroboren la primera opció, així doncs, *Capsaspora owczarzaki* (i per extensió el llinatge dels filasteris) és el segon grup d'organismes unicel·lulars més proper als animals.

Capsaspora owczarzaki pot ser cultivat fàcilment al laboratori en flascons de cultiu estèrils a 23ºC. Creix en el medi ATTC 1034, compost de bactopeptona, extracte de llevat, àcids nucleics de llevat, àcid fòlic, hemina, sèrum fetal boví i aigua destil·lada.

En el context del projecte UNICORN (abans esmentat) s'ha seqüenciat el genoma de la soca ATCC30684 de *Capsaspora owczarzaki* (l'original aïllat al Brasil el 1979) mitjançant seqüenciació Sanger. Els fragments seqüenciats es van poder ensamblar en un total de 84 *scaffolds*, amb un total final de 28 Mb amb un 8x de cobertura. El

genoma de *Capsaspora owczarzaki* de 8567 gens codificants per proteïna predits, la qual cosa suposa el 58'7% del genoma, i té una mida mitjana per gen de 3'2 Kb. Is, per tant, un genoma relativament compacta, comparat amb els genomes de metaous i del coanoflagel·lats *Monosiga brevicollis*.

Objectius

La seqüenciació del genoma de *Capsaspora owczarzaki* i altres espècies en posicions clau de l'arbre de la vida dels Opisthokonta (incloent fongs basals com *Spizellomyces punctatus* o l'apusozou *Thecamonas trahens,* grup germà dels Opisthokonta), en el context del projecte UNICORN, ha obert una finestra sense precedents per a l'estudi de l'origen de la multicel·lularitat animal mitjançant la genòmica comparada. La meva Tesi s'ha desenvolupat en aquest context, amb tres objectius principals:

1. Reconstruir la història evolutiva de diferents elements de la maquinària molecular associada a la multicel·lularitat animal, mitjançant tècniques de genòmica comparada.

2. Avaluar la conservació funcional d'alguns d'aquests elements entre els animals i els seus parents unicel·lulars.

3. Estudiar la biologia bàsica de *Capsaspora owczarzaki* per tal d'entendre millor el context fenotípic unicel·lular en el qual aquests elements funcionen.

Resultat i Discussió

Una nova visió sobre el cicle vital de *Capsaspora owczarzaki*

El cicle vital de *Capsaspora owczarzaki* era pràcticament desconegut quan quest projecte s'inicià. No obstant, entendre millor la biologia bàsica d'aquest organisme és clau per a entendre les possibles funcions dels gens que hem identificat en el seu genoma, molts dels quals es consideraven exclusius dels animals. A més, tenir un catàleg de fenotips ben descrits és necessari per a interpretar manipulacions genètiques, l'objectiu de futur de recerca en *C. owczarzaki*. Així, al mateix temps que analitzava el seu genoma, també he estudiat , la biologia cel·lular de *C. owczarzaki*, la qual cosa m'ha permès descriure'n el cicle vital complet en condicions de cultiu al laboratori (Article R7). Entre els trets més interessants observats, destaca la presència d'un estadi multicel·lular agregatiu, altament regulat a nivell d'expressió gènica i d'*splicing* alternatiu diferencial. Altres aspectes de la biologia cel·lular de *C. owczarzaki* que hem descrit inclouen les seves estructures d'actina i tubulina i com aquetes canvien durant el cicle cel·lular. Com s'havia descrit anteriorment (Stibbs et al. 1979; Owczarzak et al 1980), observem que *C. owczarzaki* té un nucli gran i un citoplasma ple de fagosomes i vesícules lipídiques. També té llargs fil·lopodis, els quals demostren estan constituïts per filaments d'actina (Article R6). Aquest fil·lopodis es perden durant l'encistament. El marcatge de la tubulina mostra la formació d'una cistella perifèrica de tubulina, que sembla sorgir del centrosoma associat al nucli. Els fil·lopodis són probablement el tret ultraestructural més característic de *C. owczarzaki*. Aquests fil·lopodis són llargs (fins a 20 μm), arribant a fer fins a 5 vegades la mida del cos cel·lular. Són fil·lopodis ramificats i, en ocasions, emergeixen com un manyoc del cos cel·lular.

L'origen del repertori genètic de la multicel·lularitat animal

L'adhesió cel·lular en animals, a diferència de la de plantes i fongs, depèn de proteïnes que estableixen contactes cèl·lula-cèl·lula i cèl·lula-matriu extracel·lular. Un estudi comparatiu permet de concloure que els dos tipus d'adhesions cel·lulars ancestrals en tots els animals són les adhesions focals (basades en integrines) i les unions adherents (basades en cadherines) (Magie i Martindale 2008). Les maquinàries moleculars per aquest tipus d'adhesions es consideraren durant molt de temps exclusives dels animals.
El 2008, Abedin et al. caracteritzaren la presència de fins a 23 cadherines al genoma de coanoflagel·lat *Monosiga brevicollis* (Abedin i King 2008); això és més inclús que el

nombre que en trobem en molts animals basals, com l'esponja *Amphimedon queenslandica* (17) i el cnidari *Nematostella vectensis* (16). Aquestes cadherines es localitzen al collar de microvil·lis del coanoflagel·lat i per això s'ha hipotetitzat que li serveixen per a capturar bacteris o per a senyalitzar quan un bacteri ha contactat amb el collar. Tot això, no s'identificà cap element de l'adhesoma d'integrina en el genoma del coanoflagel·lat.

A l'article R1 hem demostrat que la maquinària per a l'altre gran tipus d'adhesió cel·lular animal (les adhesions focals basades en integrines) estan presents en llinatges unicel·lulars. L'adhesoma d'integrina s'originà a l'ancestre como dels opisthokonts i els apusozous, ja que és present en el genoma de *Thecamonas trahens*. Més endavant, en l'ancestre dels holozous, s'hi agregaren mòduls de senyalització basats en tirosina quinases (FAK i Src). Aquest resultat implica també que l'adhesoma és va perdre secundàriament en fongs i coanoflagel·lats, mentre que és present a filasteris i també a ictiosporis (veure Article R2). La funció ancestral d'aquest sistema és un misteri, encara que es pot hipotetizar que funcionava, en un origen, com a sistema de senyalització més que no pas d'adhesió (Article R2). Les dades d'expressió gènica a *C. owczarzaki* semblen suggerir que, en aquest organisme, l'adhesoma d'integrina podria estar relacionat en la formació d'agregats multicel·lulars (Article R7). Així doncs, veiem que la maquinària molecular pels dos tipus principals d'adhesions cel·lulars animals ja existien abans de l'origen d'aquests.

Una altra necessitat la multicel·lularitat és el control de la proliferació i la diferenciació ordenada (en el temps i en l'espai) dels diferents tipus cel·lulars (Levine i Tjian 2003). Ambdues funcions són realitzades pels factors de transcripció, proteïnes que s'uneixen al DNA i activen o reprimeixen l'expressió gènica. El nostre estudi sobre l'origen dels factors de transcripció més importants per al desenvolupament animal (Article R3) revelà que molts d'ells (per exemple, p53, T-box, NFkappB, Myc, Runt, etc.) no són exclusiu dels animals com es creia (Larroux et al. 2008, Rokas 2008) si no que diversos van aparèixer en un context multicel·lular, ja que els trobem al genoma de *C. owczarzaki*. De nou, el nostre estudi demostrà que molts d'aquests factors es van perdre secundàriament en el llinatge dels coanoflagel·lats, fent èmfasi en la importància d'un mostreig filogenètic ampli a l'hora d'inferir la història evolutiva dels gens.

Sembla ser, però, que una diferència important entre els factors de transcripció dels animals i els dels seus ancestres unicel·lulars seria en el nivell d'interactivitat d'aquests factors entre ells i també amb altres cofactors. Així, per exemple, veiem que la majoria de factors de transcripció de la família bZIP a *C. owczarzaki* pertanyen a classes que actuen com a homodímers, mentre que en animals existeixen moltes classes que actuen com a heterodímers, incrementat així la seva diversitat per combinatòria (Deppmann et al. 2006).

El nostre estudi de la conservació funcional del factor de transcripció Brachyury (de la família T-box) de *C. owczarzaki* (CoBra) en *Xenopus laevis*, ens dona pistes encara més clares en aquest sentit (Article R4). El gen Brachyury és essencial per a la gastrulació i l'especificació del mesoderm en l'embrió (Technau 2001). En primer lloc, mitjançant aquest estudi poguérem reconstruir la història evolutiva del a família T-box, mostrant que estan presents en molts holozous unicel·lulars (diverses espècies de filasteris i ictiosporis) i en fongs basals, que Brachyury és la classe ancestral de gen T-box a partir de la qual emergiren la resta de classes i, per últim, que la família T-box es diversificà a l'origen dels animals, donant lloc a pràcticament la totalitat de classes (amb l'excepció de Tbx6). L'estudi de la funció heteròloga de CoBra, comparat amb la funció heteròloga dels ortòlegs de Brachyury d'una esponja i d'un cnidari, ens permeté comprovar que aquesta està funcionalment conservat, però que existeixen diferències en l'especificitat d'aquest gen CoBra (és més generalista a l'hora d'activar gens) i els ortòlegs de Brachyury en animals. Estudiant els motius d'unió a DNA de CoBra veiem que no hi ha diferències entre aquest i els gens T-box d'animals, així, podem concloure que les diferències d'especificitat es deuen a la interacció amb cofactors (probablement del tipus Smad). Així, l'establiment de noves interaccions reguladores, juntament amb la diversificació de la família T-box, permeté incrementar la complexitat de les xarxes regulades pels factors de transcripció T-box.

Per últim, la senyalització cel·lular és una altra funció clau per a la multicel·lularitat. Diverses vies de senyalització són exclusives dels animals (incloent Hedgehog, Wnt i Notch) i no es troben tampoc a *C. owczarzaki* (King et al. 2008; Srivastava et al. 2010; Suga et al. in prep). Els holozous unicel·lulars són especialment rics en receptors tirosina quinasa, que presenten expansions independents en filasteris, en coanoflagel·lats i en animals (Manning et al. 2008; Suga et al. 2012). El nostre article

sobre la història evolutiva de la via de transducció de senyal Hippo (Article R5), que controla la proliferació cel·lular en teixits animals, demostrà que aquesta té un origen anterior als animals i que els elements d'aquesta via estan conservats a nivell molecular entre *C. owczarzaki* i *Drosophila melanogaster*. Però els receptors coneguts que activen la via no estan presents a *C. owczarzaki* la qual cosa suggereix que aquesta via controla la proliferació, en un context unicel·lular, depenent de senyals extracel·lulars (probablement ambientals) diferents de les que actuen en animals. Així, la via fou co-optada durant l'origen de la multicel·lularitat, connectant-se a nous receptors per produïr efectes similars (control de la proliferació) però sota senyals diferents.

Fonts genètiques d'innovació en metazous

Els diferents resultats d'aquestes estudi ens proporcionen exemples de diversos mecanismes d'innovació que expliquen l'origen de la maquinària de la multicel·lularitat animal.

En primer lloc la co-opció de gens, és a dir, la utilització de gens pre-existents en un context unicel·lular per a noves funciones en el context multicel·lular. Tot i que es creu que la co-opció gènica ha estat una important font d'innovació en l'origen de la multicel·lularitat (King 2004; Michod 2007), pocs casos s'havien reportat fins ara de gens co-optats en l'origen dels animals (King et al. 2008; Rokas 2008). En aquesta tesi hem mostrat diversos exemples detallats de gens que foren co-optats a l'origen de la multicel·lularitat animal, incloent la maquinària d'integrines (Articles R1 i R2), diversos factors de transcripció (Article R3) i la via de senyalització Hippo (Article R5). La co-opció de gens és definida en termes funcionals (canvi de funció) i, per tant, en tots aquests casos estem pressuposant que la funció en un context multicel·lular era substancialment diferent de la que duu a terme en animals. Tot i que, en última instància, cal demostrar aquesta funció unicel·lular, creiem que l'assumpció és raonable, en el sentit de que es tracta de maquinàries amb funcions específicament associades a la multicel·lularitat en animals.

Per suposat, a l'origen dels metazous hi hagué nombrosos gens apareguts *de novo*. Exemples els trobem en factors de transcripció com Ets, Smad i NR (Article R3) R3) i també en vies de senyalització com Wnt, TGF o Hedgehog (Suga et al. in prep). De fet, l'estudi genòmic dels parents unicel·lulars dels animals ens permet una definició més

acurada de quines gens aparegueren a l'origen dels animals. Un mecanisme concret que permet innovar en aquest sentit és el rearranjament de dominis funcionals de proteïnes (conegut com *domain shuffling*), mitjançant el qual dominis proteics pre-existents són organitzats en combinacions totalment noves, de vegades juntament amb dominis de nova aparició. Aquesta plasticitat en les arquitectures de dominis explica, per exemple, l'origen del receptor Notch (King et al. 2008).

Algunes famílies gèniques ja existien abans de l'origen dels animals, però sofriren una important expansió en el llinatge animal. Molts exemples els trobem en factors de transcripció paneucariòtics, com Fox i Homeobox (Article R3), i inclús factors que aparegueren en els parents unicel·lulars del animals, com ara T-box, es diversificaren durant la transició (Article R4).

L'*splicing* alternatiu (SA) és un procés que permet expandir la diversitat del proteoma sense necessitat de nous gens i s'ha hipotetitzat que tingué una importància cabdal per a l'origen de la multicel·lularitat (Romero et al. 2006; Nilsen i Graveley 2010). El nostre estudi d'aquest fenomen en *C. owczarzaki* (Article R7), juntament amb el que se'n sap en coanoflagel·lats (Westbrook 2011) ens ha permès demostrat que hi ha dues diferències significatives entre els animals i els seus parents unicel·lulars més propers:
- En primer lloc, els animals tenen una major proporció de gens amb formes de SA. Així, hi ha una diferència quantitativa.
- En segon lloc, hi ha un canvi en els modes de SA entre animals i unicel·lulars. Mentre que en animals la forma de SA més habitual és l'anomenat *exon-skipping*, que contribueix a produir diversitat de isoformes, en els unicel·lulars és més habitual la retenció d'introns, que genera mRNAs defectius per a la traducció (codons STOP i canvis de pauta de lectura) (McGuire et al. 2008). El primer és una mecanisme de diversificació del proteoma, mentre que el segon es creu que contribueix a regular l'expressió gènica disminuint el nombre de mRNAs madur que poden ser traduïts (Anamika et al. 2009). Així, hi ha una diferència també qualitativa en SA.

Per últim, l'origen de la multicel·lularitat probablement anà acompanyada d'un major nombre d'interaccions físiques i reguladores entre gens, generant més xarxes complexes reguladores de gens. Els nostres de resultats sobre Brachyury (Article R5) suggereixen un canvi d'aquest tipus en els gens T-box, que diversificarien i especificarien les seves

funcions mitjançant noves interaccions amb co-factors (més que no pas per canvis en l'especificitat del seu motiu d'unió a DNA).

Conclusions

Les principals conclusions d'aquest treball són:

1. L'origen de la maquinària d'integrina s'originà abans de l'evolució dels animals, en l'ancestre comú d'Opisthokonta i Apusozoa. Fou assemblat de manera progressiva i la seva funció ancestral podria estar lligada a la senyalització més que no a l'adhesió, encara que les dades transcriptòmiques a *C. owczarzaki* suggereixen un rol en la formació d'agregats en aquest organisme.

2. Diversos factors de transcripció importants per al desenvolupament animal es troben en els seus parents unicel·lulars; incloent NFkappaB, T-box, Runx, Myc/Max, p53, STAT, Mef2 i altres.

3. *C. owczarzaki* Brachyury és capaç de rescatar la funció endògena de *Xenopus* Brachyury, incloent també la gastrulació. Però ho fa sense la mateixa especificitat en el reconeixement de gens diana. Aquesta especificitat fou establerta posteriorment en l'evolució de la família T-box, a l'origen dels metazous, mitjançant noves interaccions amb co-factors i no degut a noves especificitats en el motiu d'unió a DNA.

4. La via de senyalització Hippo, un important regulador de la proliferació en animals, està present a *C. owczarzaki*. Els ortòlegs d'aquesta via mostren conservació funcional quan són expressats en *Drosophila melanogaster*.

5. La co-opció gènica fou un procés més important en l'origen de la maquinària molecular animal del que fins ara s'havia reconegut; aquest procés afectà, entre altres, a la maquinària d'integrina, a diversos factors de transcripció i a la via d'Hippo.

6. Estructures i comportaments cel·lulars foren també co-optats a l'origen dels animals. Un exemple en són els microvil·lis, que s'originaren a l'ancestre comú d'animals i coanoflagel·lats i que són usats per una gran diversitat de tipus cel·lulars animals.

7. Els animals i els seus parents unicel·lulars tenen contrastats modes d'*splicing* alternatiu. Això suggereix que un canvi des de retenció d'introns a *exon-skipping* tingué lloc a l'origen dels metazous i contribuí enormement a la diversificació del proteoma animal.

8. *C. owczarzaki* té un cicle vital complex que inclou un estadi agregatiu multicel·lular. Les dades transcriptòmiques indiquen que les transicions entre estadis estan finament regulades a nivell d'expressió gènica i d'*splicing* alternatiu diferencial. Aquest cicle de *C. owczarzaki* posa de manifest l'enorme diversitat de comportaments cel·lulars presents en els parents unicel·lulars actuals dels animals i la presència de l'estadi agregatiu suggereix que components de la maquinària de multicel·lularitat animal poden ser usats en contexts funcionals molt diferents.

AGRAÏMENTS

Han passat més de cinc anys, quan encara estudiava a la facultat, des d'aquell dia en que en Jaume Baguñà em va parlar d'un tio esprimatxat, que acabava d'arribar del Canadà i buscava gent per fer una tesi i començar un grup de recerca. No estava pas molt clar ben bé què feia (jo volia sentir a parlar d'animals i d'evo-devo!), ni si tenia pasta, ni espai per formar un grup. Però vaig decidir anar a parlar amb ell; i allà estava, l'Iñaki, a la biblioteca del departament, amb el seu portàtil. I em va convèncer: l'origen dels animals... caram, el *quid* de la qüestió! Doncs sí, ja podia dir que treballa en evo-devo animal [sic]. Després d'uns inicis més aviats incerts (fent uns bons nyaps en el racó que la Marta Riutort amablement ens deixava al seu laboratori del departament de genètica), la cosa agafà dimensions èpiques: ERC *grant*, laboratori al Parc Científic,... Així vaig començar la tesi, la tardor del 2008, en un laboratori totalment buit, ni una sola tècnica posada apunt i terreny científic pràcticament verge per explorar. Cinc anys després, la situació no ha canviat pas molt: poques tècniques (...) i la mateixa llibertat per sorprendre'ns cada dia amb coses noves. Així doncs, gràcies en primer lloc a l'Iñaki, per haver assentat les bases d'aquest fèrtil terreny i per deixar-nos explorar-ho amb total llibertat d'idees i de treball. I per fer les bromes més estranyes i sonades del món.

El de la tesi és un període d'aprenentatge, així que gran part del meu agraïment és per aquelles persones que m'han ensenyat i influït. L'Àlex Pérez és la primera persona que em va ensenyar alguna cosa de biologia molecular (de veritat), en aquell estiu del 2007 que vaig passar amb ell al lab de l'Elvira Juan, on ell feia la tesi. Em va ensenyar, acollir i entretenir d'allò més. Coses de la vida, ens vam tornar a trobar un parell d'any després, ara ell com a *lab manager* de l'MCG. I encara me'n va ensenyar més de coses! així doncs, gràcies Àlex! La Núria en va agafar el relleu i m'ensenyà coses noves i, sobretot, una manera diferent (més *punky*) de fer biologia molecular, gràcies Ponsy! Però en fer el *punky* al lab ningú supera en Hiroshi... Thank you Hiroshi for showing me that "molecular biology is not an exact science"! Després vingué en Jordi, un mite vivent del departament de genètica, qui ens ho va ensenyar tot i més sobre filogènia i evolució, a banda de ser el tio més divertit que ha passat pel lab, gràcies Paps! Gràcies també al Javi per la ensenyar-nos la vessant ecològica i més purament protistològica de la nostra feina. I gràcies la resta de gent que ha passat pel lab (John, Lora, Majo, David i Meritxell).
Més recentment, es giraria la truita i em tocaria a mi d'ensenyar alguna cosa als nouvinguts. Però resulta que va venir a parar al lab un noi taaan bon noi i tan espavilat que de seguida han canviat de nou les tornes: ara aprenc jo d'ell. Gràcies Xavi! I también a gracias a Helena, el otro *tadpole* del lab, por su entusiasmo y alegría desbordantes (a ver cuanto te dura...)!
També vull agrair als nostres veïns i ex-veïns de laboratori per la seva paciència majúscula (en aguantar els nostres crits, en prestar-nos tot tipus de reactius i material,

coneixements,...). Gràcies a l'Eva, en *Serginho*, la Laia i en Víctor. I gràcies a l'Adriana i l'Antonella.
Malgrat no hi érem físicament, el departament de genètica sempre exercí una poderosa influència sobre nosaltres, tal contuberni de ments brillants! Especialment agraït estic als incondicionals dels *journal clubs* d'evo-devo, on vaig aprendre molt sobre evolució: Pere, Cristian, Jaume, Chema, Johannes, Nacho i Manu. A Nacho y a Manu agradecerles también por sus excelentes consejos científicos en distintas etapas de mi tesis. Gràcies també a la Susanna per haver estat, en els últims temps, el nostre nexe d'unió i integració al departament.
L'estiu de 2010 vaig coincidir en un curs a Suècia amb dos personatges realment extraordinaris, l'un va exercir *de facto* de professor privat per a mi durant curs i l'altre em fascinaria amb la seva personalitat desbordant i la seva obsessió pels acels. Gracias Chema por enseñarme todo sobre el mundo de la immuno y de la in-situ y por ser el crack que eres! Thank you Johannes for being such an exceptional specimen!
Gracias a José Luis por acogerme en su lab en Sevilla, por sus ideas y su mala leche; y a Ana Ariza por enseñarme a trabajar con los sapos y peces, con paciencia infinita y dedicación, y por ser tan currante (sin ti no habría proyecto Brachyury!).
Gràcies tots els nous companys de l'IBE, que en tot moment ens han ajudat i facilitat l'adaptació al nou entorn.
I gràcies a totes aquelles altres persones que a través dels seus articles, converses, col·laboracions,... han contribuït a l'educació del meu pensament científic.
I, *last but not least*, gràcies al senyor Alfred Owczarzak per haver trobat i aïllat aquest bitxo tan catxondo, que tantes alegries m'ha donat i que es diu Capsaspora.

Passem ara a dos individus que se'm fa impossible de classificar, a cavall entre la secció "lab" i "amics". En Guifré i en Mendoza, amics des de la facultat, amb qui he fet tota la tesi. Hem après junts aquests quasi cinc anys i, sobretot, ens hem fet uns tips de riure inenarrables. Gràcies Guifré pel teu saber fer tranquil, sistemàtic i tocant de peus a terra; per ser un bon suport quan ha calgut; per tenir el *funky in the body*; i per no defallir mai, mai, en intentar fer-me riure fins a tallar-me la respiració (inclús a centenars de quilòmetres de distància). Gràcies Mendoza per la teva brillantor i agilitat mentals; per saber trobar la perspectiva adequada, que sovint els altres no veiem, als problemes científics; pel *touch*; per ensenyar-me mil coses sobre el bon viure i, perquè no, també per *l'ennui* postmodern (és important de tenir models a evitar...). Sou els principals responsables que el de la meva tesi hagi estat un període extremadament divertit i gratificant. Mil gràcies als dos!

Saltem del món de la recerca al de les persones que fan que la resta valgui tant la pena.
Al Pau, per ser un autèntic model vital i moral; impertèrrit, ascètic i alhora afable i bonhomiós.
Al Juse pel seu neohedonisme rural (que no hedonisme neorural), per saber tantes coses impossibles, per estar de tornada de tantes coses i ensenyar-me'n.

Al Javi per ser savi en un sentit tan antic, fer-se sempre les preguntes més bàsiques i actuar sempre radicalment en conseqüència.
Al Carlos, i també de nou al Mendoza i al Javi, per les *bimbades* extremes (per desgràcia cada cop menys habituals...), viatges i anellades. Al triumvirat de limnòlegs Marc, Pau i Carlos, per interessants discussions, per la vostra voluntat èpica i per treure'm d'excursió algun cop per la muntanya.
Gràcies a la Cèlia i a la Irene per compartir cuina, sofà i menjador amb mi durant gran part de la carrera i de la tesi i per ser tan formoses! També gràcies a la Bea, pel divertit estiu que vam passar al pis de Poblet, i al Kike i a la Sri, els altres habitants de Poblet en algun moment.
Gràcies també a la Berta, a la Roser i a la Rousie, a l'Ares, a la Marta i a l'Ariadna, al Gil (el bon salvatge), i a algun més (no gaires) que potser em descuido.

Gràcies a tota la meva família. Al meu pare, qui em va inculcar la passió per la natura des de petit; que es transmutà primer en afició per l'ornitologia, després en un interès general per la biologia i en particular per l'evolució i, finalment, en... grilladura per la protistologia molecular? A la Julita que ha estat sempre un referent i un puntal per mi. A l'àvia per ser la meva *fan* més incondicional. Gràcies a la mare, en l'absència.

Gràcies a la Cris, per ser la meva quotidianitat, el meu tot.

Per acabar. Algú a qui realment aprecio em va dir un dia que en la ciència no hi ha poesia. Certament no, però jo hi he trobat molta diversió i bona companyia. Gràcies a tots.