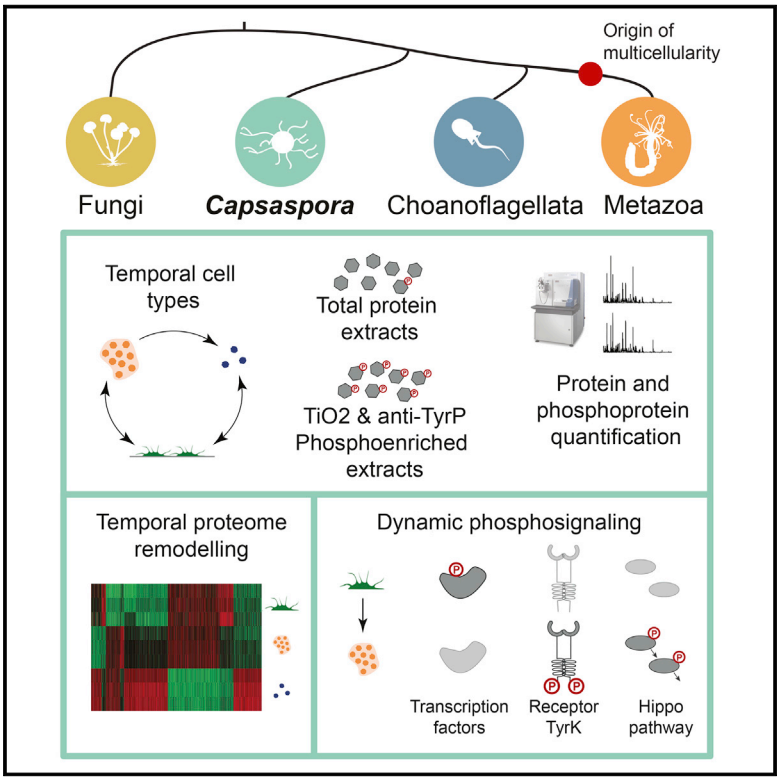
Article

Developmental Cell

# High-Throughput Proteomics Reveals the Unicellular Roots of Animal Phosphosignaling and Cell Differentiation

## Graphical Abstract




## Authors

Arnau Sebé-Pedrós, Marcia Ivonne Peña, Salvador Capella-Gutiérrez, Meritxell Antó, Toni Gabaldón, Iñaki Ruiz-Trillo, Eduard Sabidó

## Correspondence

inaki.ruiz@ibe.upf-csic.es (I.R.-T.), eduard.sabido@crg.cat (E.S.)

## In Brief

Many essential genes for animal multicellularity evolved in a unicellular context. Sebé-Pedrós et al. examine, using proteomics, temporal cell differentiation in *Capsaspora owczarzaki*, a unicellular animal relative. They provide insight into phosphoregulatory network evolution and show that *Capsaspora* differentiation is supported by extensive proteome remodeling and dynamic phosphosignaling.

## Highlights

- Proteome remodeling is linked to temporal differentiation in a unicellular context
- Dynamic phosphosignaling underlies unicellular temporal differentiation
- Parallel evolution of Ser/Thr and Tyr kinase phosphoregulatory networks
- Cell-type-specific phosphoactivation of Tyr kinases in *Capsaspora*

## Accession Numbers

PXD004567

Sebé-Pedrós et al., 2016, Developmental Cell *39*, 186–197
October 24, 2016 
http://dx.doi.org/10.1016/j.devcel.2016.09.019

CellPress

CellPress

Developmental Cell
Article

# High-Throughput Proteomics Reveals the Unicellular Roots of Animal Phosphosignaling and Cell Differentiation

Arnau Sebé-Pedrós,[1,8] Marcia Ivonne Peña,[2,4] Salvador Capella-Gutiérrez,[3,4,5] Meritxell Antó,[1] Toni Gabaldón,[3,4,6] Iñaki Ruiz-Trillo,[1,6,7,*] and Eduard Sabidó[2,4,9,*]

[1]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain
[2]Proteomics Unit, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain
[3]Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain
[4]Universitat Pompeu Fabra (UPF), Dr. Aiguader 88, 08003 Barcelona, Spain
[5]CBS Fungal Biodiversity Centre, Uppsalalaan 8, 3584 LT Utrecht, the Netherlands
[6]ICREA, Pg. Lluis Companys 23, 08010 Barcelona, Spain
[7]Departament de Genètica, Microbilogia i Estadística, Universitat de Barcelona, Avinguda Diagonal 643, 08028 Barcelona, Spain
[8]Present address: Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel
[9]Lead Contact
*Correspondence: inaki.ruiz@ibe.upf-csic.es (I.R.-T.), eduard.sabido@crg.cat (E.S.)
http://dx.doi.org/10.1016/j.devcel.2016.09.019

## SUMMARY

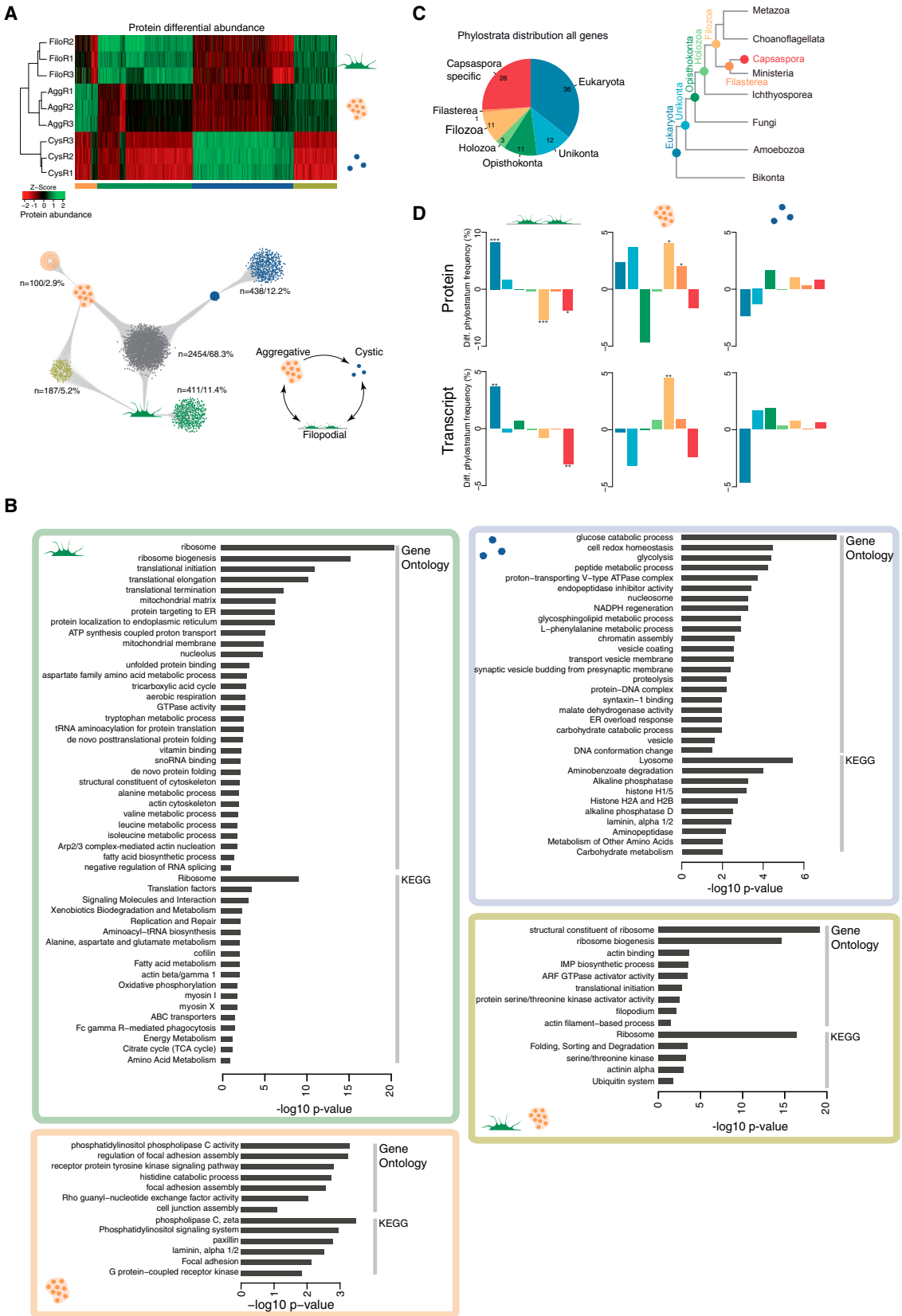
**Cell-specific regulation of protein levels and activity is essential for the distribution of functions among multiple cell types in animals. The finding that many genes involved in these regulatory processes have a premetazoan origin raises the intriguing possibility that the mechanisms required for spatially regulated cell differentiation evolved prior to the appearance of animals. Here, we use high-throughput proteomics in *Capsaspora owczarzaki*, a close unicellular relative of animals, to characterize the dynamic proteome and phosphoproteome profiles of three temporally distinct cell types in this premetazoan species. We show that life-cycle transitions are linked to extensive proteome and phosphoproteome remodeling and that they affect key genes involved in animal multicellularity, such as transcription factors and tyrosine kinases. The observation of shared features between *Capsaspora* and metazoans indicates that elaborate and conserved phosphosignaling and proteome regulation supported temporal cell-type differentiation in the unicellular ancestor of animals.**

## INTRODUCTION

A defining feature of multicellularity is the ability to generate multiple cell types, each defined by a specific profile of protein abundance and phosphorylation-regulated activity. Yet it remains unclear how the mechanisms controlling cell-specific differentiation programs evolved. Comparative genomics suggests that the unicellular ancestor of animals was genetically complex and already had many of the genes required for animal cell signaling and cell-type specification (King et al., 2008; Suga et al., 2013). A key step in the evolution of cell-specific differentiation programs in multicellular animals must therefore have been the appearance of mechanisms to regulate the abundance and activity of different proteins according to cell type. It is known that premetazoan unicellular organisms have defined cell stages that are temporally separated across the life cycle of the organism (Fairclough et al., 2013; Sebé-Pedrós et al., 2013; de Mendoza et al., 2015). However, it remains unclear whether the mechanisms that regulate protein abundance and activity in these cell types represent evolutionary precursors of those found in their animal counterparts.

The ameboid filasterean *Capsaspora owczarzaki* (herein *Capsaspora*) is one of the closest animal relatives, and has a rich repertoire of metazoan-related genes involved in cell differentiation and cellular signaling, including an expanded tyrosine kinase system (Suga et al., 2012). Moreover, *Capsaspora* differentiates into three temporally distinct cell types (Sebé-Pedrós et al., 2013): (1) a filopodiated ameba, which corresponds to the proliferative trophic stage; (2) an aggregative multicellular stage, in which the cell produces an extracellular matrix; and (3) a cystic resistance form. These temporal differentiation capabilities, together with its key phylogenetic position and its genetic toolkit, make *Capsaspora* an ideal candidate to study how animal-like cell signaling and proteome dynamics contribute to cell differentiation in a unicellular, premetazoan context.

High-throughput proteomics is revolutionizing the study of cell differentiation dynamics by allowing the systematic quantitation of thousands of proteins and the covalent modifications that modulate their activity. However, to date this approach has only been systematically applied to a handful of species (Kim et al., 2014; Nagaraj et al., 2012; Ringrose et al., 2013; Schrimpf et al., 2009; Solyakov et al., 2011; Wilhelm et al., 2014). In this study, we perform a multilayer proteome characterization of *Capsaspora* using label-free mass spectrometry (MS)-based proteomics. We quantify the steady-state levels of thousands of proteins and the dynamics of their phosphorylation sites across temporally distinct *Capsaspora* cell types. Integration of




*(legend on next page)*

these datasets with transcript expression and advanced phylogenetic analyses reveals that *Capsaspora* life stages are associated with specific proteome and phosphoproteome profiles, which display unique phylostratigraphic and functional patterns. Furthermore, gene age strongly influences multiple aspects of this observed differential regulation. We provide evidence that similar proteome regulation and phosphosignaling mechanisms underlie cell differentiation in *Capsaspora* and animals.

## RESULTS AND DISCUSSION

### Proteome Dynamics in Temporal Cell Differentiation

Protein levels are the final outcome of gene expression and ultimately determine the cellular phenotype. To analyze the proteome dynamics underlying temporally distinct cell differentiation programs, we used high-resolution MS to identify 4,372 *Capsaspora* proteins, using a false discovery rate (FDR) of less than 5% (Table S1). We then performed relative quantitation of 3,590 proteins across the three *Capsaspora* cell types, using three biological replicates per stage. Significant differences in abundance in at least one of the cell types were observed for 1,136 proteins ($p < 0.01$, ANOVA) in our dataset. Each life stage showed a specific proteome expression profile (Figure 1A) associated with exclusive cellular functions, such as integrin adhesion and tyrosine kinase signaling in the aggregative stage (Figures 1B and S1).

To analyze the evolutionary origin of these stage-specific proteins, we performed a phylostratigraphic enrichment analysis (Domazet-Loso et al., 2007). We first generated a complete set of phylogenies for 6,551 *Capsaspora* genes (phylome) (Huerta-Cepas et al., 2011), which we used to assign each gene to a particular phylostratum (Figure 1C). We then analyzed the over- or under-representation of phylostrata among the differentially expressed genes, both at the protein and transcript level, using available transcriptome data from the same cell types (Sebé-Pedrós et al., 2013) (Figure 1D). We found that the filopodia-specific proteome and transcriptome are strongly enriched ($p < 0.001$, Fisher's exact test) in genes of eukaryotic origin (i.e., ancient genes shared by most eukaryotes), whereas the cystic stage has no significant enrichments. Interestingly, the aggregative multicellular stage shows a significant enrichment in genes of filozoan origin (i.e., genes evolved at the stem of Filasterea [*Capsaspora*], Choanoflagellata, and Metazoa) (Figure 1D), at both the protein and transcript levels. This finding indicates that aggregate formation in *Capsaspora* chiefly involves genes shared with metazoans and suggests that this is not a *Capsaspora*-specific adaptation, but rather a cellular behavior that was present in the unicellular ancestor of metazoans. Thus, our data show that extensive proteome remodeling accompanies temporal cell-type transitions in *Capsaspora* and that each cell type has unique functional and gene-age patterns.

### Gene Age and Transcript-Protein Correlation

Protein abundances depend on multiple layers of regulation that go beyond transcript production and, as a result, transcript and protein levels are not always strongly correlated (Dahan et al., 2011). We explored the correlation between protein and transcript levels in *Capsaspora* using RNA sequencing (RNA-seq). The observed global correlation between averaged RNA and protein levels in *Capsaspora* was similar to that found in other species ($\rho = 0.55$; Figure 2A) (Laurent et al., 2010; Schrimpf et al., 2009; Vogel and Marcotte, 2012). This implies that transcript levels only explain part of the protein abundances and the rest is attributable to post-transcriptional regulatory mechanisms. Moreover, this transcript-protein-level correlation is not uniform across genes (Figure S2). For example, a pattern of strong anticorrelation between RNA and protein levels across stages is observed in the predicted secretome proteins (Figure S2C). It has also been shown that protein levels are under stronger evolutionary constraint than RNA levels (Artieri and Fraser, 2014; Khan et al., 2013; Schrimpf et al., 2009). Consistent with these findings, analysis of one-to-one orthologs between *Capsaspora* and two other species (*Homo sapiens* and *Saccharomyces cerevisiae*) revealed a stronger correlation among protein levels than RNA levels (Figure 2B).

To further explore the evolution of protein expression, we analyzed protein-RNA correlation in *Capsaspora* across phylostrata. We found weaker correlation in evolutionarily younger genes, especially in *Capsaspora*-specific/orphan genes (Figure 2C). Interestingly, analysis of available human and yeast datasets revealed a similar pattern in other species (Figure 2D) (Fagerberg et al., 2014; Hobson et al., 2012; Kim et al., 2014; Kuang et al., 2014; Nagaraj et al., 2012). This suggests the existence of increased post-transcriptional gene regulation in evolutionarily younger genes and an increase in gene expression constraint over evolutionary time, as recently observed for human genes (Popadin et al., 2014). We also found that younger genes have more variable expression across cell types or stages than ancient genes, at both the transcript and protein levels (Figures 2C–2E). This pattern is also found in zebrafish and fly developmental transcriptomes (Domazet-Lošo and Tautz, 2010). These findings indicate that evolutionarily younger genes have more cell-type/tissue-specific regulation, not only during animal development (Tautz and Domazet-Lošo, 2011) but also during temporal cell-type differentiation in unicellular eukaryotes.

**Figure 1. Protein-Level Dynamics across *Capsaspora* Cell Types**

(A) Heatmap (top) showing estimated protein abundances in the different biological replicates of each stage. Protein expression level was calculated as the sum of all unique normalized peptide ion abundances for a specific protein on each run. Only proteins exhibiting significant changes in abundance (ANOVA $p < 0.01$) are shown. The network (bottom) represents the proteins (nodes) assigned (edges) to each stage.

(B) Functional enrichment of *Capsaspora* differentially expressed proteins. Bar plots represent the gene ontology and KEGG pathway enrichment analyses of stage-specific expressed proteins.

(C) Inferred evolutionary origin of all *Capsaspora* genes. The pie chart represents the frequency of each phylostratum as represented in the phylogenetic tree.

(D) Bar plot showing the phylostratigraphic enrichment analysis of the stage-specific differentially expressed genes, at the protein level and the transcript level. Asterisks indicate the level of significance of Fisher's exact test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). See also Figure S1.

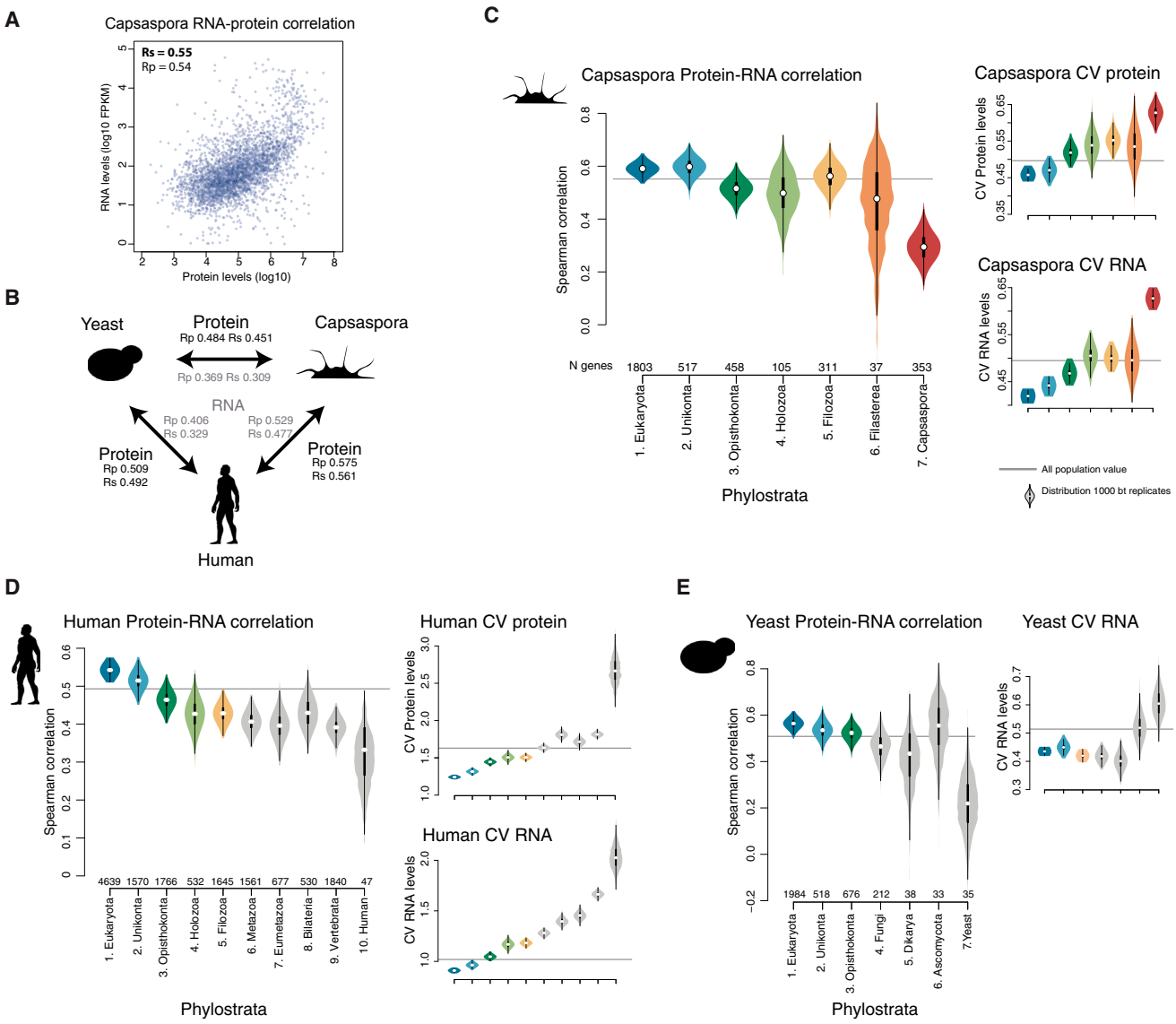



**Figure 2. Phylostratigraphic Patterns in Protein-Transcript Levels**

(A) Correlation between protein and transcript levels in *Capsaspora*. Pearson (Rp) and Spearman (Rs) correlations are indicated.

(B) Correlation between protein and RNA abundances in one-to-one orthologs of yeast, human, and *Capsaspora*.

(C) Spearman correlation between protein and RNA abundances of *Capsaspora* genes classified by phylostratum (left) and coefficients of variation of protein (top right) and RNA (bottom right) levels across *Capsaspora* cell types for genes classified by phylostratum. The violin plot represents the distribution of 1,000 bootstrap replicates. The number of genes in each phylostratum is indicated in the correlation plot (left). A horizontal line indicates the values corresponding to the whole population of genes.

(D) Same as (C) for human genes analyzed across different tissues (Fagerberg et al., 2014; Kim et al., 2014).

(E) Same as (C) for yeast genes analyzed in temporal metabolic series (Hobson et al., 2012; Kuang et al., 2014; Nagaraj et al., 2012).

See also Figure S2.

## Phosphoregulatory Landscape Evolution

Protein phosphorylation at serine/threonine (pSer/Thr) residues is the most widespread signaling mechanism in eukaryotes and is mediated by the ancient paneukaryotic family of serine/threonine kinases. In contrast, signaling through tyrosine phosphorylation (pTyr) emerged at the stem of Holozoa (i.e., Ichthyosporea, Filasterea, Choanoflagellata, and Metazoa) (Suga et al., 2012), and is one of the few examples of innovation in post-translational protein modifications in eukaryote evolution (Beltrao et al., 2013). Thus, *Capsaspora* occupies a key phylogenetic position within Holozoa in understanding the emergence of tyrosine kinase (TK) signaling, an essential animal intercellular communication mechanism, and its early coexistence with serine/threonine phosphosignaling. To study the *Capsaspora* phosphoproteome, we performed phosphopeptide enrichment using both titanium dioxide beads and a combination of three different

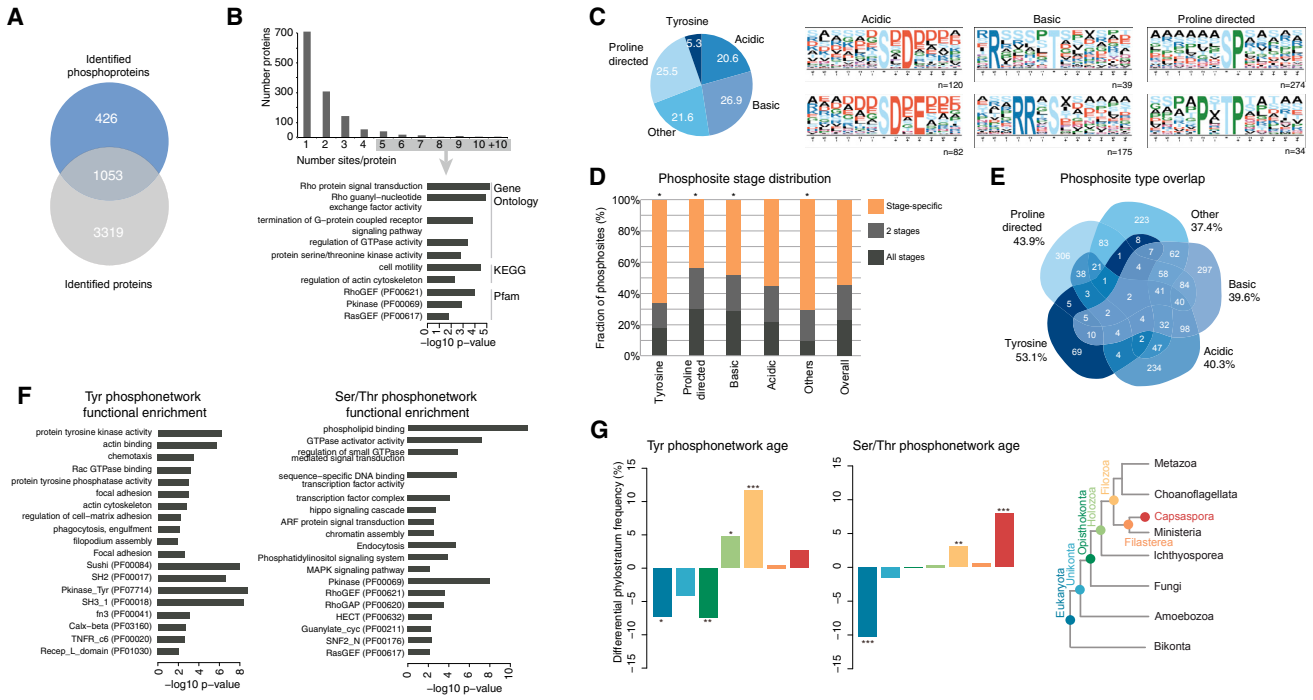



**Figure 3. The *Capsaspora* Phosphoproteome**

(A) Venn diagram of the number of all identified proteins and phosphoproteins.

(B) Bar graph showing the distribution of the number of phosphosites per protein and bar plot representing functional enrichments of genes with more than four phosphosites.

(C) Pie chart showing the frequency of different motif categories (left) and examples of phosphorylation-specific motifs identified (right). The number of sites on which each logo is based is indicated below.

(D) Bar graph showing the cell-type specificity distribution (in all three, in two, or only in one stage) of each motif category. Asterisks indicate significance of chi-square test of each category versus the overall distribution (*p < 0.01).

(E) Venn diagram showing the level of co-occurrence of different motif types in the same protein. The percentage of proteins with a unique motif type is indicated.

(F) Bar plots representing functional enrichments of pTyr-regulated genes versus pSer/Thr-regulated genes in *Capsaspora*.

(G) Bar graph showing the phylostratigraphic enrichment analysis of pTyr-regulated genes versus pSer/Thr-regulated genes as represented in the phylogenetic tree (third panel). Asterisks indicate significance of Fisher's exact test (*p < 0.05, **p < 0.01, ***p < 0.001).

anti-pTyr antibodies. We identified 2,609 nonredundant phosphorylation sites (FDR < 5%) corresponding to 1,479 phosphoproteins (Figure 3A and Table S2) and examined the distribution of number of sites per protein (Figure 3B). We found that among the proteins with multiple phosphosites (five or more), many are involved in Rho GTPase signaling, actin cytoskeleton, and protein kinases (Figure 3B), indicating a complex phosphoregulation of these genes in *Capsaspora*.

Phosphosite selectivity is often affected by neighboring amino acid positions, reflecting kinase substrate specificities (Villén et al., 2007). Hence, we searched for enriched amino acid motifs surrounding phosphorylation sites in *Capsaspora* (see Supplemental Experimental Procedures). We found 29 enriched motifs (Table S4), which we used to group phosphosites into five categories (acidic, basic, proline-directed, tyrosine-based, and others; the latter represent pSer/Thr sites with no significant motif enrichment) (Figure 3C). Interestingly, the enriched motif categories are similar to those found in animals (Huttlin et al., 2010; Villén et al., 2007) and the motif types show different levels of cell stage specificity. Basic and proline-directed motifs tended to be more globally distributed in cell stages (Figure 3D). Tyrosine motifs and "other" motifs were more frequently stage specific, in contrast to what has been observed in mouse tissues (Huttlin et al., 2010). At the single protein level, phosphorylation motifs classified as "others" co-occurred most frequently with other motifs in the same protein (Figure 3E). In contrast, tyrosine phosphosites showed the least overlap with other motifs, revealing that pTyr and pSer/Thr networks are minimally overlapping in *Capsaspora*.

Next, we compared these two systems with different evolutionary origins to explore how they coexist in *Capsaspora* (Figures 3F and 3G). First, we found strong subfunctionalization patterns. The pTyr network is enriched in TK activity (as expected from the fact that TKs are activated by autophosphorylation), filopodia, and actin cytoskeleton proteins as well as focal adhesion proteins. In contrast, the pSer/Thr network was found to be associated with guanosine triphosphatase (GTPase) signaling and transcription factors, among others (Figure 3F). Second, phylostratigraphic analysis of these two networks revealed that, despite being an ancient signaling system, pSer/Thr-regulated genes are enriched in *Capsaspora*-specific genes. This observation suggests that pSer/Thr networks evolve rapidly in a species-specific manner (Beltrao et al., 2009), also during the early stages of animal evolution (see below). In contrast,

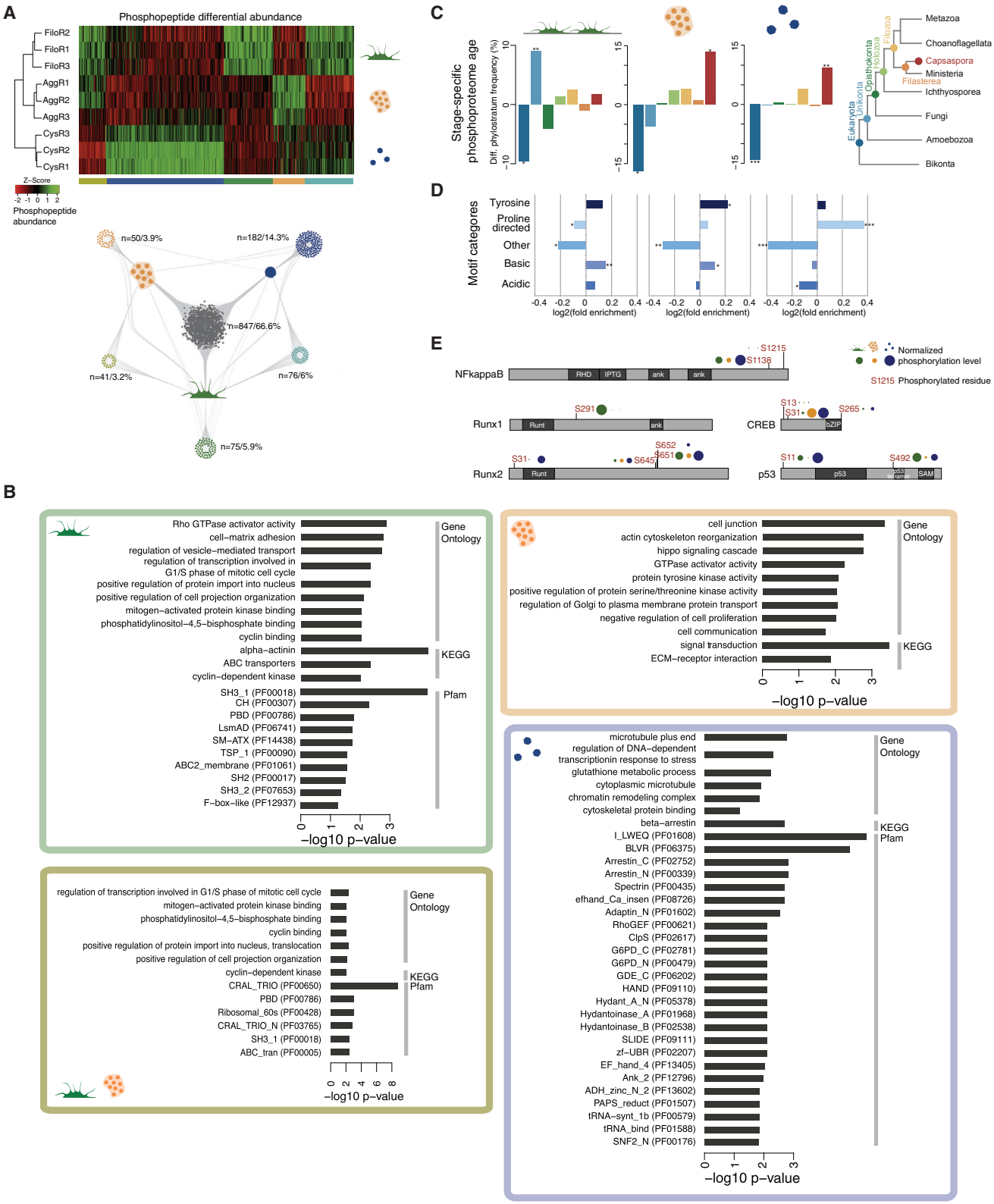



**Figure 4. Phosphoproteome Dynamics and Cell Differentiation in *Capsaspora***

(A) Heatmap (top) showing differential phosphopeptide level in the different biological replicates of each stage. Only phosphopeptides exhibiting significant changes in abundance (ANOVA $p < 0.01$) are shown. The network (bottom) represents the proteins (nodes) differentially phosphorylated in each stage (edges). (B) Bar plots represent the gene ontology, KEGG pathway, and Pfam domain enrichment analyses of stage-specific phosphorylated proteins.

*(legend continued on next page)*

pTyr-regulated genes are significantly enriched in Holozoa and Filozoa phylostrata (Figure 3G), which implies that many genes regulated by TKs originated concomitantly with the pTyr signaling system (Suga et al., 2012), thus establishing a new phosphosignaling network parallel to the pre-existing pSer/Thr signaling network.

## Phosphoregulation in Temporal Cell Differentiation

To explore the role of the phosphorylation system in the differentiation of temporally distinct cell types, we analyzed the quantitative dynamics of the *Capsaspora* phosphoproteome. We found 425 phosphopeptides exhibiting significant changes ($p < 0.05$, ANOVA) across life stages (Figure 4A). Notably, each *Capsaspora* cell type has a highly specific phosphorylation profile, the most distinct being the cystic stage. These cell-type-specific phosphoproteomes are enriched in different functions and phylostrata (Figures 4B and 4C). For instance, we observed a dynamic activation of the mitogen-activated protein kinase pathway consistent with the proliferative behavior of the *Capsaspora* filopodial-stage cells (Figure S3). We also detected an enrichment of proteins involved in the Hippo signaling pathway in the aggregative-stage phosphoproteome. This kinase cascade is an evolutionary innovation present only in *Capsaspora* and animals (Sebé-Pedrós et al., 2012), and its activation is associated with the inhibition of cell proliferation. Interestingly, the Hippo pathway might repress proliferation during the aggregative stage but not in the cystic stage, which is also nonproliferative (Sebé-Pedrós et al., 2013). The observed activation of the Hippo pathway in the multicellular aggregates suggests a role in the repression of cell proliferation in this specific *Capsaspora* stage, similar to the role of the Hippo pathway in organ-size control in metazoans (Sebé-Pedrós et al., 2012).

Regarding the evolutionary origin of phosphoregulated genes, the filopodial-stage phosphoproteome is enriched in genes of ancient, unikont origin (i.e., shared exclusively by opisthokonts and amoebozoans), whereas the aggregative and cystic stages are enriched in new genes (specific to *Capsaspora*) (Figure 4C). Interestingly, the observed stage-specific phosphoproteomes show statistically significant variations in the frequencies of the different phosphosite motif categories (Figure 4D) (see above). This finding indicates that the different life stages use specific kinases to maintain specialized phosphosignaling events, as occurs in animal tissues (Huttlin et al., 2010). Consistent with this observation, analysis of the *Capsaspora* kinome revealed that 135 serine/threonine protein kinases, most of them with animal orthologs, are active throughout the *Capsaspora* life cycle, and many of them show differential patterns of abundance and/or phosphorylation (Figure S4, see also below), which supports the observed cell-type-specific phosphosignaling. Overall, our results reveal that temporal cell-type transitions in *Capsaspora* are associated with hundreds of unique phosphosignaling events supported by a highly dynamic kinome.

## *Capsaspora* Transcription Factor Phosphoregulation

*Capsaspora* has a rich repertoire of transcription factors (TFs) with roles in animal multicellularity (Sebé-Pedrós et al., 2011). Since in metazoans TF activity is often controlled by phosphorylation (Whitmarsh and Davis, 2000), we examined the phosphorylation state of animal TFs with orthologs in *Capsaspora*. We found differential phosphorylation in 25 TFs, among them Runx, nuclear factor κB (NF-κB), CREB (cyclic AMP response element binding protein), and p53 (Figure 4E). Runx TFs are important regulators of animal development, and Runx phosphorylation has been directly linked to activation of these TFs in animals (Aho et al., 2006; Ito et al., 2015). Interestingly, the two Runx paralogs (Sebé-Pedrós et al., 2011) present in *Capsaspora* show complementary patterns of phosphorylation, with Runx1 being phosphorylated in the filopodial stage and Runx2 in the cystic stage. This finding suggests that these two paralogs are subfunctionalized via differential phosphorylation and might direct specific transcriptional programs in these two stages (Figure 4E). NF-κB is an important regulator of proliferation and apoptosis (Viatour et al., 2005) and C-terminal phosphorylation blocks its activity in animals by preventing nuclear translocation (Demarchi et al., 2003). The *Capsaspora* NF-κB C-terminal domain is overphosphorylated in the cystic stage, suggesting that it would only be active in filopodial and aggregative stages (Figure 4E). CREB is a transcriptional activator involved in multiple functions in animals and activated by phosphorylation (Johannessen et al., 2004; Mayr and Montminy, 2001). The *Capsaspora* CREB ortholog has three phosphorylation sites and one of them is strongly differentially phosphorylated in the aggregative and cystic stage (Figure 4E), suggesting that CREB also has stage-specific activity. Finally, p53 is an essential regulator of animal apoptosis and cell cycle. Animal p53 activity and stability is strongly enhanced by phosphorylation (Bode and Dong, 2004), and this results in cell-cycle arrest, apoptosis, and/or DNA repair (Bode and Dong, 2004). The *Capsaspora* p53 ortholog has two phosphorylation sites, one in the N-terminal transactivation domain (Figure 4E) and another in the nuclear localization signal. Both sites are clearly overphosphorylated in the cystic stage and filopodial stage; in the filopodial stage, phosphorylation of the nuclear localization site is predominant. Although the potential conservation of these specific phosphosites in *Capsaspora* TF regulation needs experimental validation, our results indicate that differential phosphorylation plays an important role in regulating the activity of multiple animal-like TFs during cell differentiation in *Capsaspora*.

## A Premetazoan Tyrosine Kinase System in Action

Tyrosine-specific phosphorylation mediated by TKs is essential for animal cell-cell communication. Since *Capsaspora* has an extremely rich TK repertoire (Figures 5B and S5; Table S5) (Suga et al., 2012), it is an interesting system for the study of early evolution of pTyr signaling. To understand how the TK system

(C) Bar graph showing the phylostratigraphic enrichment analysis of the differentially phosphorylated proteins. Asterisks indicate significance of Fisher's exact test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

(D) Bar plot representing the enrichment or depletion of different motif categories among the stage-specific phosphopeptides. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

(E) Schematic representation of five *Capsaspora* transcription factors and their phosphorylation sites. For each site, dot sizes represent the level of phosphorylation in each cell type (color coded) normalized by the highest value of all the sites in the protein.

See also Figures S3 and S4.

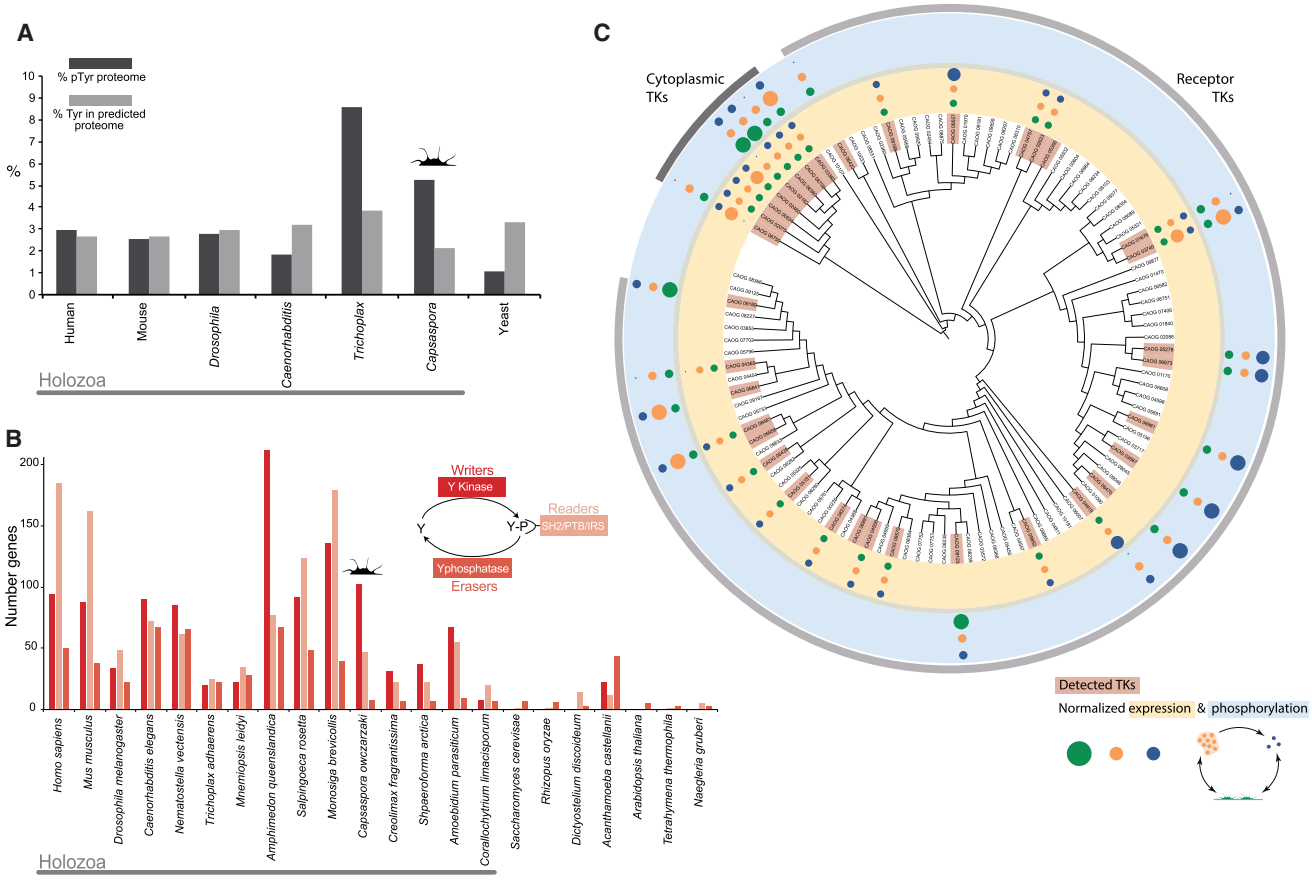



**Figure 5. Phosphotyrosine Signaling in *Capsaspora***

(A) Bar graph comparing the percentage of tyrosine residues in the genome and the percentage of detected phosphotyrosines in the proteome of different species.

(B) Bar graph showing the number of tyrosine signaling-related genes in different species, including writers (tyrosine kinases), readers (SH2, PTB, and IRS domain-containing proteins), and erasers (tyrosine phosphatases).

(C) Maximum likelihood phylogenetic tree representing all *Capsaspora* cytoplasmic and receptor tyrosine kinases. Gray shading indicates which kinases have been detected, while in the outer rings dot sizes represent the relative (*Z* score) levels of expression (inner ring, yellow) and phosphorylation (outer ring, blue) of each protein across *Capsaspora* cell types (color coded).

See also Figure S5.

works in *Capsaspora* and how the pTyr of animals might have evolved, we first examined the frequency of pTyr in the *Capsaspora* phosphoproteome (Table S3). We observed a high frequency of tyrosine phosphorylation (5.33%), indicating that the early holozoan expansion of TKs was accompanied by a burst of tyrosine phosphorylation (Figure 5A). Surprisingly, however, this high pTyr frequency is not linked to a high frequency of tyrosines in the *Capsaspora* genome (2.1%, less than in most animals) (Figure 5A). This finding is inconsistent with the hypothesis that tyrosine residues would have been progressively reduced in the genome as the number of TKs expanded during animal evolution, reducing the possibility for promiscuous, unspecific tyrosine phosphorylation (Ringrose et al., 2013; Tan et al., 2009).

Next, we examined the protein abundance and phosphorylation of *Capsaspora* TKs. We identified 32 *Capsaspora* TKs, including all the cytoplasmic TKs and 24 TK receptors with different extracellular domain architectures (Figures 5C and S5). Interestingly, many of these TKs show changes in abundance and levels of tyrosine autophosphorylation (Figure 5C), indicative of dynamic kinase activation across the *Capsaspora* life cycle. For example, TK receptors with extracellular Sushi domain repeats and receptors with leucine-rich repeats (involved in protein-protein interactions) are activated in the aggregative stage, whereas a TK receptor with two L domains (found in animal receptor TKs such as epidermal growth factor receptors and insulin receptors) is activated in the filopodial and aggregative stage (Figure S5). Among the cytoplasmic TKs, Abl is a multifunctional animal proto-oncogene (Wang, 2014) that appears to be active in the filopodial and aggregative stages of *Capsaspora*. Another example is Tec, which in animals is involved in actin cytoskeleton reorganization and inside-out signaling to integrins, promoting cell adhesion (Gomez-Rodriguez et al., 2007). Tec is strongly activated in the aggregative stage and could be linked to stimulated adhesiveness in these cells. Focal adhesion kinase (FAK) is a well-studied downstream signaling effector of integrin adhesion (Mitra and Schlaepfer, 2006; Mitra et al., 2005; Schlaepfer et al., 1999) and its levels are

increased in the aggregative stage; in addition, it appears to be phosphorylated in in both the filopodial and aggregative stages. This indicates that integrin signaling is active in both stages, with different integrin paralog pairs expressed in different life stages (Sebé-Pedrós et al., 2013), but using the same single *Capsaspora* FAK ortholog. Finally, Src is involved in many human cancers, and its activation by a range of transmembrane receptors induces cell motility and proliferation (Yeatman, 2004). The autophosphorylation activator site of one of the *Capsaspora* Src orthologs (Src1) was determined experimentally (Y541) (Schultheiss et al., 2012), showing that it is homologous to the human c-Src activation site. Interestingly, we observe the same Src1 site dynamically phosphorylated, as well as the homologous site (Y447) in Src2. These activation sites of the two *Capsaspora* Src orthologs are strongly phosphorylated in the filopodial proliferative stage and it is, therefore, tempting to speculate that Src may play an important role in stimulating cell proliferation in *Capsaspora*. Taken together, our data suggest that TKs function as a major signaling mechanism that controls temporal cell behaviors in response to environmental stimuli in *Capsaspora*.

### The Origin of Animal Phosphoregulatory Networks

To gain further insights into the evolution of animal phosphoregulation, we compared the *Capsaspora* phosphorylation network with other species for which similar proteomic phosphoenrichment datasets are available, including mouse (*Mus musculus*) (Huttlin et al., 2010), an early-branching animal *Trichoplax adhaerens* (Ringrose et al., 2013), and yeast (*Saccharomyces cerevisiae*) (Hustedt et al., 2014), as an outgroup of Holozoa (Figure 6A). We first analyzed the number of phosphosites per protein as a measure of the complexity of the different phosphoregulation networks. We observed that mouse has more phosphosites in total than any of the other taxa; most notably, it has many more phosphosites per protein. This suggests more elaborate phosphosignaling machinery in vertebrates than in the early-branching or nonanimal taxa analyzed. In contrast, *Capsaspora* and *Trichoplax* do not differ in the number of sites or in the distribution of sites per protein, with yeast showing a similar profile but with a reduced number of phosphosites (Figure 6A). These data indicate that the phosphoregulatory network did not undergo a sudden quantitative increase in complexity at the origin of Metazoa. Instead, the biggest innovation was the emergence of pTyr signaling at the stem of Holozoa (see above).

Next, we explored how these phosphoregulatory networks changed qualitatively during evolution. We used our phylome analysis to assign one-to-one orthologs and then determined whether orthologous genes are phosphorylated in different species. We then applied a parsimony criterion to reconstruct phosphorylated ortholog content at ancestral nodes, including the stems of Opisthokonta, Holozoa, and Metazoa (Figure 6B). The inferred ancestral phosphonetworks showed striking differences in functional gene content. Thus, the ancestral shared Opisthokonta phosphonetwork is enriched in basic cellular functions such as translation, DNA replication, and basal transcription machinery (Figure 6B). In contrast, the holozoan ancestral phosphorylated orthologs are enriched in, among others, GTPase signaling, actin cytoskeleton and filopodia, Hippo signaling, protein kinases, and TFs. Finally, the ancestral metazoan phosphorylation network is enriched in genes involved in multicellularity-related processes, such as development, morphogenesis, and fertilization. These results show that important qualitative, rather than quantitative, remodeling of phosphoregulatory networks occurred in the lineage, leading to animal multicellularity. In this stepwise assembly, genes involved in functions important for multicellularity were incorporated at the stem of Metazoa.

## Conclusions

Our integrative analysis of the *Capsaspora* proteome represents the first systematic study of the protein levels and phosphosignaling dynamics underlying temporal cell differentiation in a unicellular eukaryote. In our deep proteomic datasets we identified 4,798 proteins, representing 55.4% of the annotated *Capsaspora* genes. Quantitative analysis of these proteins across stages indicates that temporal cell-type differentiation during the life cycle of this animal relative is finely regulated at the level of protein abundance and phosphorylation. This is reflected in highly cell-type-specific proteomic and phosphoproteomic profiles, similar to those observed in different animal cell types and tissues (Huttlin et al., 2010; Kim et al., 2014; Wilhelm et al., 2014). Similarly, the *Capsaspora* phosphosignaling system resembles in many aspects the phosphorylation observed in metazoans (Huttlin et al., 2010; Ringrose et al., 2013), including an important role for tens of TKs that are activated in a cell-specific manner, and a shared network of phosphoregulated orthologs that was progressively assembled during premetazoan evolution. Our results further indicate that gene age similarly affects proteome regulation in temporal and spatial cell differentiation, with increased post-transcriptional regulation and higher cell-type specificity in evolutionarily younger genes. Gene-age distribution also differs between pTyr and pSer/Thr networks, indicating that most pTyr-regulated genes appear concomitantly with the expansion of the pTyr system at the stem of Holozoa (Beltrao et al., 2013; Suga et al., 2012).

Overall, these findings suggest that the unicellular ancestor of animals regulated the differentiation of temporally distinct cell types through the use of elaborate animal-like gene expression control and phosphosignaling systems. We hypothesize that these pre-existing mechanisms in the protistan ancestors of animals constituted the molecular basis of a transition from temporal to spatial cell differentiation at the origin of Metazoa.

## EXPERIMENTAL PROCEDURES

### *Capsaspora* Cell Cultures

*Capsaspora* strain ATCC30864 cells were grown axenically in ATCC medium 1034 at 23°C and differentiated as described in Supplemental Experimental Procedures.

### Protein Extraction and Digestion

For each cell stage and each of the three biological replicates (nine samples), total protein extracts were obtained by resuspending cells in extraction buffer containing protease and phosphatase inhibitors and sonicating on ice. Protein extracts were precipitated in cold acetone, resuspended in urea solution, and sonicated. Following reduction and alkylation, proteins were digested with trypsin. Details are provided in Supplemental Experimental Procedures.

### Phosphopeptide Enrichment

Phosphorylated peptides were enriched using an offline immobilized metal-affinity $TiO_2$ protocol. In addition, we performed a specific enrichment for

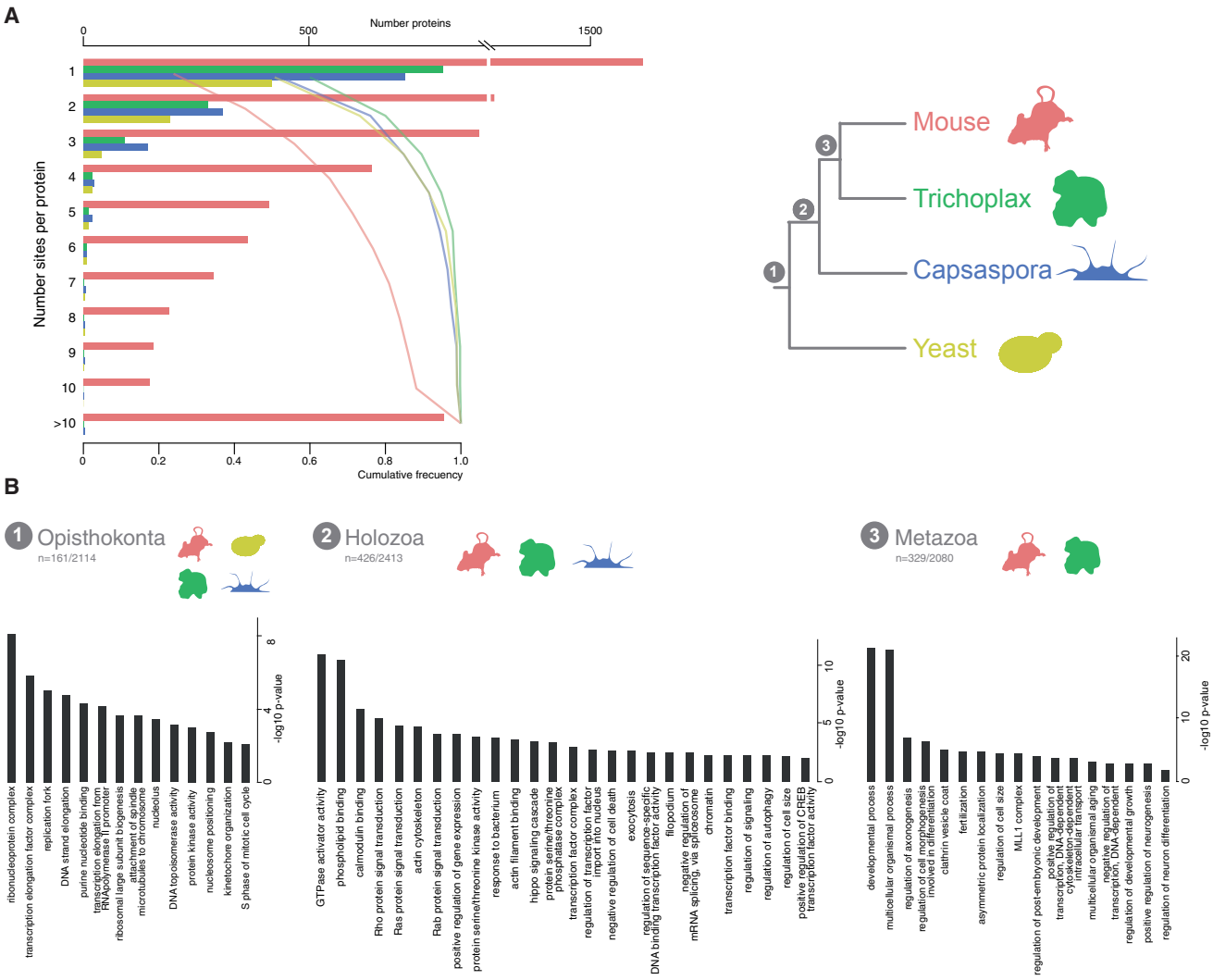



**Figure 6. Evolution of Metazoan Phosphoregulatory Networks**

(A) Right: schematic phylogenetic tree representing the evolutionary relationship between the four analyzed species: mouse, *Trichoplax*, *Capsaspora*, and yeast. Left: histogram representing the distribution of number of phosphosites per protein in each species and line plot representing the cumulative frequency of each distribution.

(B) Bar graphs representing the gene ontology enrichment analyses of the inferred ancestral Opisthokonta (1), Holozoa (2), and Metazoa (3) phosphorylation networks. In each case, the number of phosphorylated orthologs versus the background (i.e., total number of orthologs) used in the enrichment are indicated.

phosphotyrosine-containing peptides by immunoaffinity purification, using three different anti-pTyr antibodies. Details are provided in Supplemental Experimental Procedures.

### Liquid Chromatography-Tandem Mass Spectrometry Sample Acquisition and Data Analysis

Peptide and phosphopeptide samples were purified by reverse-phase columns and subsequently analyzed in an LTQ-Orbitrap Velos Pro mass spectrometer. Acquired data were analyzed using Mascot search engine (v2.3, Matrix Science) for peptide and $TiO_2$-enriched phosphopeptide identification, whereas Andromeda search engine (MaxQuant v1.4.0.5) was used for antibody-based enrichment of pTyr-containing peptides. Data were searched against an in-house generated database based on *C. owczarzaki* genome v3 and all identified peptides were filtered using a 5% FDR threshold (peptide level). Peptide and phosphopeptide (pSer, pThr) areas were extracted from the peptide features using the feature detection algorithm implemented in Progenesis (v4.1, Nonlinear Dynamics). Normalized abundances of unique peptides were used to estimate protein quantities, and significant differences in protein levels were evaluated using an ANOVA model. Details are provided in Supplemental Experimental Procedures.

### Phylome Reconstruction and Analysis

The complete phylomes of *Capsaspora* and four other species were reconstructed as described in Supplemental Experimental Procedures. We used these phylomes to infer orthology and paralogy relationships among genes from the five species and also to assign a relative gene age to each gene (phylostratigraphy). Phylome data were used in multiple downstream analyses as described in Supplemental Experimental Procedures and below.

### RNA-Protein Correlation Analysis

Previously published RNA-seq data (Sebé-Pedrós et al., 2013) were used to compare protein and RNA levels. This RNA-seq dataset contains replicates of the same *Capsaspora* life stages studied here. General and gene-age-specific RNA-protein level Spearman and Pearson correlations were calculated

using averaged cross-sample/cross-replicate expression values. A similar strategy was used to analyze publicly available human and yeast datasets. Further details are provided in Supplemental Experimental Procedures.

### Phosphorylation Motif Analysis

We generated a dataset of centered 13-mer sequences by extending six amino acids around each of the detected phosphosites, and computed the motif enriched with X motif (Schwartz and Gygi, 2005) using the whole predicted proteome as background and the default significance threshold ($p = 0.000001$). The detected motifs were classified using a previously defined decision-tree algorithm (Villén et al., 2007; Huttlin et al., 2010). Further details are provided in Supplemental Experimental Procedures.

### Ancestral Phosphorylation Network Reconstruction

We used available phosphoproteomics datasets for yeast (Hustedt et al., 2014), *T. adhaerens* (Ringrose et al., 2013), and mouse (Huttlin et al., 2010), and used our phylome dataset to determine pairwise ortholog genes that are phosphorylated in both species. We used a parsimony criterion to infer phosphorylated ortholog repertoires in ancestral nodes. Further details are provided in Supplemental Experimental Procedures.

## ACCESSION NUMBERS

MS proteomics data have been deposited to the Proteomics Identifications repository under accession number PRIDE: PXD004567. Phylome data can be accessed through PhylomeDB.org (phylomes 100–105).

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.devcel.2016.09.019.

## AUTHOR CONTRIBUTIONS

A.S.-P., M.I.P., and E.S. designed the study. M.I.P., M.A., and A.S.-P. performed experiments. M.I.P. and E.S. planned, acquired, and analyzed MS data. S.C.-G. and T.G. performed phylome analysis. A.S.-P., M.I.P., and E.S. analyzed and interpreted the data. A.S.-P, I.R.-T., and E.S. wrote the paper. All authors discussed the results and commented on the manuscript.

## ACKNOWLEDGMENTS

The CRG/UPF Proteomics Unit is part of the "Plataforma de Recursos Biomoleculares y Bioinformáticos (ProteoRed)" supported by grant PT13/0001 of Instituto de Salud Carlos III (ISCIII). A.S.-P. is supported by an EMBO Long-Term Fellowship (ALTF 841-2014). T.G. group research is funded in part by a grant from the Spanish Ministry of Economy and Competitiveness (BIO2012-37161), a grant from the Qatar National Research Fund (NPRP 5-298-3-086), a grant from the European Union FP7 FP7-PEOPLE-2013-ITN-606786, and a grant from the European Research Council under the European Union's Seventh Framework Program (FP/2007-2013)/ERC (grant agreement no. ERC-2012-StG-310325). I.R.-T. group research is supported by an Institució Catalana de Recerca i Estudis Avançats contract, a European Research Council Consolidator Grant (ERC-2012-Co-616960), and a grant (BFU-2011-23434) from Ministerio de Economía y Competitividad (MINECO). I.R.-T. also acknowledges financial support from Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (Project 2014 SGR 619 and Project 2014 SGR 678). We thank Iain Patten, Alex de Mendoza, and Xavier Grau-Bové for comments on the manuscript, and Helena Parra-Acero for technical support.

Received: October 1, 2015
Revised: July 17, 2016
Accepted: September 16, 2016
Published: October 13, 2016

## REFERENCES

Aho, T.L.T., Sandholm, J., Peltola, K.J., Ito, Y., and Koskinen, P.J. (2006). Pim-1 kinase phosphorylates RUNX family transcription factors and enhances their activity. BMC Cell Biol. *7*, 21.

Artieri, C.G., and Fraser, H.B. (2014). Evolution at two levels of gene expression in yeast. Genome Res. *24*, 411–421.

Beltrao, P., Trinidad, J.C., Fiedler, D., Roguev, A., Lim, W.A., Shokat, K.M., Burlingame, A.L., and Krogan, N.J. (2009). Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. PLoS Biol. *7*, e1000134.

Beltrao, P., Bork, P., Krogan, N.J., and van Noort, V. (2013). Evolution and functional cross-talk of protein post-translational modifications. Mol. Syst. Biol. *9*, 714.

Bode, A.M., and Dong, Z. (2004). Post-translational modification of p53 in tumorigenesis. Nat. Rev. Cancer *4*, 793–805.

Dahan, O., Gingold, H., and Pilpel, Y. (2011). Regulatory mechanisms and networks couple the different phases of gene expression. Trends Genet. *27*, 316–322.

Demarchi, F., Bertoli, C., Sandy, P., and Schneider, C. (2003). Glycogen synthase kinase-3B regulates NF-kB1/p105 stability. J. Biol. Chem. *278*, 39583–39590.

de Mendoza, A., Suga, H., Permanyer, J., Irimia, M., and Ruiz-Trillo, I. (2015). Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. Elife *4*, e08904.

Domazet-Lošo, T., and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature *468*, 815–818.

Domazet-Loso, T., Brajkovic, J., and Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. *23*, 533–539.

Fairclough, S.R., Chen, Z., Kramer, E., Zeng, Q., Young, S., Robertson, H.M., Begovic, E., Richter, D.J., Russ, C., Westbrook, M.J., et al. (2013). Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. Genome Biol. *14*, R15.

Fagerberg, L., Hallström, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., et al. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol. Cell. Proteomics *13*, 397–406.

Gomez-Rodriguez, J., Readinger, J.A., Viorritto, I.C., Mueller, K.L., Houghtling, R.A., and Schwartzberg, P.L. (2007). Tec kinases, actin, and cell adhesion. Immunol. Rev. *218*, 45–64.

Hobson, D.J., Wei, W., Steinmetz, L.M., and Svejstrup, J.Q. (2012). RNA polymerase ii collision interrupts convergent transcription. Mol. Cell *48*, 365–374.

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L.P., Denisov, I., Kormes, D., Marcet-Houben, M., and Gabaldon, T. (2011). PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Res. *39*, D556–D560.

Hustedt, N., Seeber, A., Sack, R., Tsai-Pflugfelder, M., Bhullar, B., Vlaming, H., van Leeuwen, F., Guénolé, A., van Attikum, H., Srivas, R., et al. (2014). Yeast PP4 interacts with ATR homolog Ddc2-Mec1 and regulates checkpoint signaling. Mol. Cell *57*, 273–289.

Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S.A., Villén, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. Cell *143*, 1174–1189.

Ito, Y., Bae, S.-C., and Chuang, L.S.H. (2015). The RUNX family: developmental regulators in cancer. Nat. Rev. Cancer *15*, 81–95.

Johannessen, M., Delghandi, M.P., and Moens, U. (2004). What turns CREB on? Cell. Signal. *16*, 1211–1227.

Khan, Z., Ford, M.J., Cusanovich, D.A., Mitrano, A., Pritchard, J.K., and Gilad, Y. (2013). Primate transcript and protein expression levels evolve under compensatory selection pressures. Science *342*, 1100–1104.

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

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

Kuang, Z., Cai, L., Zhang, X., Ji, H., Tu, B.P., and Boeke, J.D. (2014). High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast. Nat. Struct. Mol. Biol. *21*, 854–863.

Laurent, J.M., Vogel, C., Kwon, T., Craig, S.a, Boutz, D.R., Huse, H.K., Nozue, K., Walia, H., Whiteley, M., Ronald, P.C., et al. (2010). Protein abundances are more conserved than mRNA abundances across diverse taxa. Proteomics *10*, 4209–4212.

Mayr, B., and Montminy, M. (2001). Transcriptional regulation by the phosphorylation-dependent factor CREB. Nat. Rev. Mol. Cell Biol. *2*, 599–609.

Mitra, S.K., and Schlaepfer, D.D. (2006). Integrin-regulated FAK-Src signaling in normal and cancer cells. Curr. Opin. Cell Biol. *18*, 516–523.

Mitra, S.K., Hanson, D.A., and Schlaepfer, D.D. (2005). Focal adhesion kinase: in command and control of cell motility. Nat. Rev. Mol. Cell Biol. *6*, 56–68.

Nagaraj, N., Kulak, N.A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012). System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. Mol. Cell. Proteomics *11*, M111.013722.

Popadin, K.Y., Gutierrez-arcelus, M., Lappalainen, T., Buil, A., Steinberg, J., Nikolaev, S.I., Lukowski, S.W., Bazykin, G.A., Seplyarskiy, V.B., Ioannidis, P., et al. (2014). Gene age predicts the strength of purifying selection acting on gene expression variation in humans. Am. J. Hum. Genet. *95*, 660–674.

Ringrose, J.H., van den Toorn, H.W.P., Eitel, M., Post, H., Neerincx, P., Schierwater, B., Altelaar, A.F.M., and Heck, A.J.R. (2013). Deep proteome profiling of *Trichoplax adhaerens* reveals remarkable features at the origin of metazoan multicellularity. Nat. Commun. *4*, 1408.

Schlaepfer, D.D., Hauck, C.R., and Sieg, D.J. (1999). Signaling through focal adhesion kinase. Prog. Biophys. Mol. Biol. *71*, 435–478.

Schrimpf, S.P., Weiss, M., Reiter, L., Ahrens, C.H., Jovanovic, M., Malmström, J., Brunner, E., Mohanty, S., Lercher, M.J., Hunziker, P.E., et al. (2009). Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. PLoS Biol. *7*, e48.

Schwartz, D., and Gygi, S.P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. *23*, 1391–1398.

Schultheiss, K.P., Suga, H., Ruiz-Trillo, I.I., and Miller, W.T. (2012). Lack of Csk-mediated negative regulation in a unicellular Src kinase Lack of Csk-mediated negative regulation in a unicellular Src kinase. Biochemistry *51*, 8267–8277.

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.

Sebé-Pedrós, A., Zheng, Y., Ruiz-Trillo, I., and Pan, D. (2012). Premetazoan origin of the Hippo signaling pathway. Cell Rep. *1*, 13–20.

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

Solyakov, L., Halbert, J., Alam, M.M., Semblat, J.-P., Dorin-Semblat, D., Reininger, L., Bottrill, A.R., Mistry, S., Abdi, A., Fennell, C., et al. (2011). Global kinomic and phospho-proteomic analyses of the human malaria parasite *Plasmodium falciparum*. Nat. Commun. *2*, 565.

Suga, H., Dacre, M., de Mendoza, A., Shalchian-Tabrizi, K., Manning, G., and 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., Chen, Z., de Mendoza, A., Sebé-Pedrós, A., Brown, M.W., Kramer, E., Carr, M., Kerner, P., Vervoort, M., Sánchez-Pons, N., et al. (2013). The Capsaspora genome reveals a complex unicellular prehistory of animals. Nat. Commun. *4*, 1–9.

Tan, C.S.H., Pasculescu, A., Lim, W.A., Pawson, T., Bader, G.D., and Linding, R. (2009). Positive selection of tyrosine loss in metazoan evolution. Science *325*, 1686–1688.

Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. Nat. Rev. Genet. *12*, 692–702.

Viatour, P., Merville, M.P., Bours, V., and Chariot, A. (2005). Phosphorylation of NF-kB and IkB proteins: implications in cancer and inflammation. Trends Biochem. Sci. *30*, 43–52.

Villén, J., Beausoleil, S.A., Gerber, S.A., and Gygi, S.P. (2007). Large-scale phosphorylation analysis of mouse liver. Proc. Natl. Acad. Sci. USA *104*, 1488–1493.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.

Wang, J.Y.J. (2014). The capable ABL: what is its biological function? Mol. Cell. Biol. *34*, 1188–1197.

Whitmarsh, a J., and Davis, R.J. (2000). Regulation of transcription factor function by phosphorylation. Cell Mol. Life Sci. *57*, 1172–1183.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582–587.

Yeatman, T.J. (2004). A renaissance for SRC. Nat. Rev. Cancer *4*, 470–480.

**Developmental Cell, Volume *39***

## Supplemental Information

# High-Throughput Proteomics Reveals the Unicellular Roots of Animal Phosphosignaling and Cell Differentiation

**Arnau Sebé-Pedrós, Marcia Ivonne Peña, Salvador Capella-Gutiérrez, Meritxell Antó, Toni Gabaldón, Iñaki Ruiz-Trillo, and Eduard Sabidó**

Inventory of Supplementary Materials:

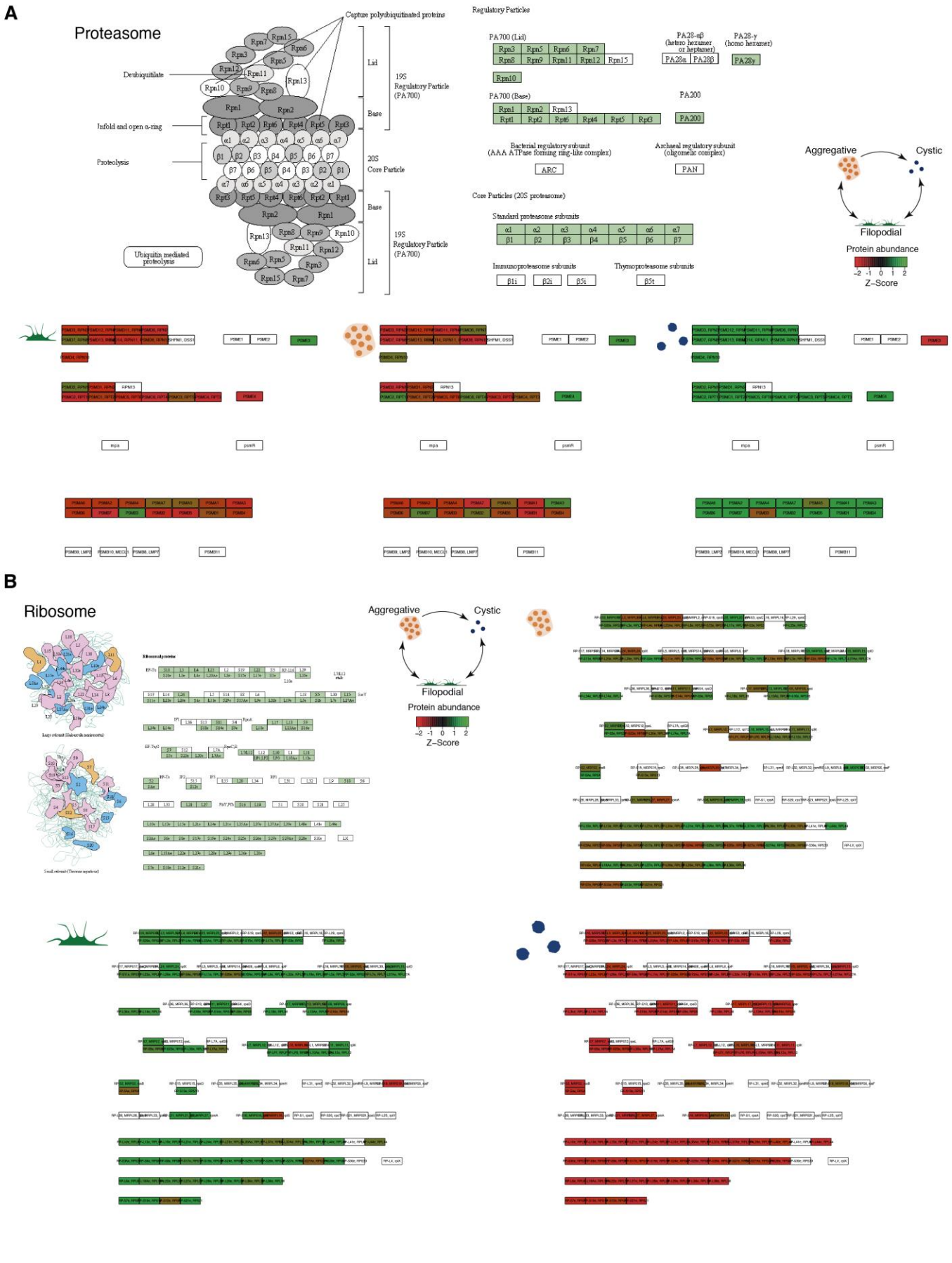

A
Proteasome
Capture polyubiquitinated proteins
Deubiquitilate
Unfold and open α-ring
Proteolysis
Ubiquitin mediated proteolysis
Lid
Base
19S Regulatory Particle (PA700)
20S Core Particle
Regulatory Particles
PA700 (Lid)
PA700 (Base)
PA28-αβ (hetero hexamer or heptamer)
PA28-γ (homo hexamer)
PA200
Bacterial regulatory subunit (AAA ATPase forming ring-like complex)
Archaeal regulatory subunit (oligomeric complex)
ARC
PAN
Core Particles (20S proteasome)
Standard proteasome subunits
Immunoproteasome subunits
Thymoproteasome subunits
Aggregative
Cystic
Filopodial
Protein abundance
Z–Score
B
Ribosome
Aggregative
Cystic
Filopodial
Protein abundance
Z–Score

**Figure S1. Concerted Differential Expression of the Proteasome and Ribosomal KEGG Pathways, related to Figure 1.** In each case, the schematic representation of the pathway is shown and the genes present in the *Capsaspora* genome are highlighted in light green. A color-coded pathway for each cell type indicates the normalized level of expression (Z-score) of each protein. Proteins absent in the *Capsaspora* genome or undetected in our proteomics dataset are shown in white. Notice the concerted overexpression of the proteasome in the cystic stage (except the PSME3 regulatory subunit), suggesting extensive proteome remodeling and protein recycling in this stage. In contrast, ribosomal proteins are overexpressed in the filopodial stage, the trophic proliferative phase of the *Capsaspora* life cycle, and are strongly repressed in the quiescent cystic stage.

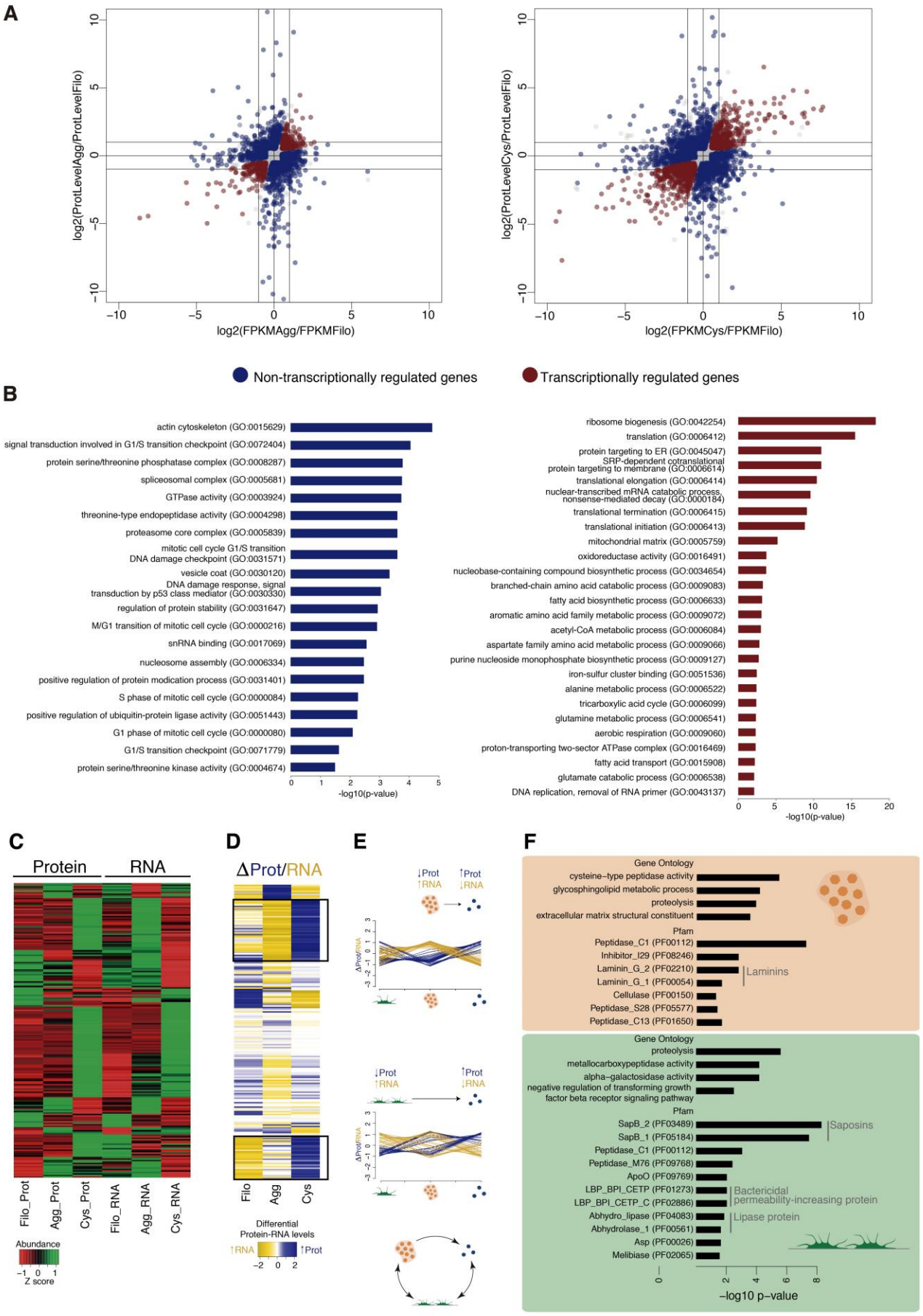

A
log2(ProtLevelAgg/ProtLevelFilo)
log2(FPKMAgg/FPKMFilo)
log2(ProtLevelCys/ProtLevelFilo)
log2(FPKMCys/FPKMFilo)
Non-transcriptionally regulated genes
Transcriptionally regulated genes
B
actin cytoskeleton (GO:0015629)
signal transduction involved in G1/S transition checkpoint (GO:0072404)
protein serine/threonine phosphatase complex (GO:0008287)
spliceosomal complex (GO:0005681)
GTPase activity (GO:0003924)
threonine-type endopeptidase activity (GO:0004298)
proteasome core complex (GO:0005839)
mitotic cell cycle G1/S transition DNA damage checkpoint (GO:0031571)
vesicle coat (GO:0030120)
DNA damage response, signal transduction by p53 class mediator (GO:0030330)
regulation of protein stability (GO:0031647)
M/G1 transition of mitotic cell cycle (GO:0000216)
snRNA binding (GO:0017069)
nucleosome assembly (GO:0006334)
positive regulation of protein modification process (GO:0031401)
S phase of mitotic cell cycle (GO:0000084)
positive regulation of ubiquitin-protein ligase activity (GO:0051443)
G1 phase of mitotic cell cycle (GO:0000080)
G1/S transition checkpoint (GO:0071779)
protein serine/threonine kinase activity (GO:0004674)
-log10(p-value)
ribosome biogenesis (GO:0042254)
translation (GO:0006412)
protein targeting to ER (GO:0045047)
SRP-dependent cotranslational protein targeting to membrane (GO:0006614)
translational elongation (GO:0006414)
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (GO:0000184)
translational termination (GO:0006415)
translational initiation (GO:0006413)
mitochondrial matrix (GO:0005759)
oxidoreductase activity (GO:0016491)
nucleobase-containing compound biosynthetic process (GO:0034654)
branched-chain amino acid catabolic process (GO:0009083)
fatty acid biosynthetic process (GO:0006633)
aromatic amino acid family metabolic process (GO:0009072)
acetyl-CoA metabolic process (GO:0006084)
aspartate family amino acid metabolic process (GO:0009066)
purine nucleoside monophosphate biosynthetic process (GO:0009127)
iron-sulfur cluster binding (GO:0051536)
alanine metabolic process (GO:0006522)
tricarboxylic acid cycle (GO:0006099)
glutamine metabolic process (GO:0006541)
aerobic respiration (GO:0009060)
proton-transporting two-sector ATPase complex (GO:0016469)
fatty acid transport (GO:0015908)
glutamate catabolic process (GO:0006538)
DNA replication, removal of RNA primer (GO:0043137)
-log10(p-value)
C
Protein
RNA
Filo_Prot
Agg_Prot
Cys_Prot
Filo_RNA
Agg_RNA
Cys_RNA
Abundance
Z score
D
ΔProt/RNA
Filo
Agg
Cys
Differential Protein-RNA levels
↑RNA
↑Prot
E
↓Prot ↑RNA
↑Prot ↓RNA
ΔProt/RNA
F
Gene Ontology
cysteine−type peptidase activity
glycosphingolipid metabolic process
proteolysis
extracellular matrix structural constituent
Pfam
Peptidase_C1 (PF00112)
Inhibitor_I29 (PF08246)
Laminin_G_2 (PF02210)
Laminin_G_1 (PF00054)
Cellulase (PF00150)
Peptidase_S28 (PF05577)
Peptidase_C13 (PF01650)
Laminins
Gene Ontology
proteolysis
metallocarboxypeptidase activity
alpha−galactosidase activity
negative regulation of transforming growth factor beta receptor signaling pathway
Pfam
SapB_2 (PF03489)
SapB_1 (PF05184)
Peptidase_C1 (PF00112)
Peptidase_M76 (PF09768)
ApoO (PF09769)
LBP_BPI_CETP (PF01273)
LBP_BPI_CETP_C (PF02886)
Abhydro_lipase (PF04083)
Abhydrolase_1 (PF00561)
Asp (PF00026)
Melibiase (PF02065)
Saposins
Bactericidal permeability-increasing protein
Lipase protein
−log10 p−value

**Figure S2. Transcriptionally versus Non-transcriptionally Regulated Proteins and *Capsaspora* Predicted Secretome, related to Figure 2.** (A) Dot plots comparing the log2 fold changes in protein (y-axis) versus transcript (x-axis) levels between aggregative and filopodial stages (left) and between cystic and filopodial stages (right). Black lines indicate a fold change of 2 (log2FC=1). Each dot corresponds to a gene and dots are colored according to the degree of correlation (see Methods). (B) Barplots representing the Gene Ontology enrichment analyses of transcriptionally and non-transcriptionally regulated gene lists, pooled from the two pairwise comparisons in a. Genes with strong transcript-protein level correlation are mostly involved in translation, protein sorting and metabolic activity; these genes also have higher correlation between species(Schrimpf et al., 2009). Genes with weak couple of transcript-protein level changes include actin cytoskeleton proteins, GTPases and regulators of the cell cycle. (C) Heatmap showing the normalized (Z-score) protein and transcript levels in each stage of 198 out of the 394 predicted secreted genes in the *Capsaspora* genome (see Materials & Methods). (D) Heatmap of the difference between normalized protein and RNA levels. (E) Line plots representing the individual profiles of RNA and protein levels across stages of the two clusters highlighted with black boxes in b. The strong anti-correlation observed between RNA and protein levels between successive cell stages suggests active transcriptional activation of the genes in the first stage not matched by protein levels because of active secretion (which drastically reduces the amount of protein inside the cell). In the next stage, transcription is repressed and remnant proteins accumulate inside the cell. (F) Bar plots representing the Gene Ontology and Pfam domain enrichment analyses of predicted secreted proteins in the aggregative stage (top) and in the filopodial stage (bottom). Notice the enrichment in laminin proteins in the predicted aggregative secretome and the presence of multiple proteins involved in bacterial prey capture (such as lipases, LBPs and saposins) in the filopodial secretome.

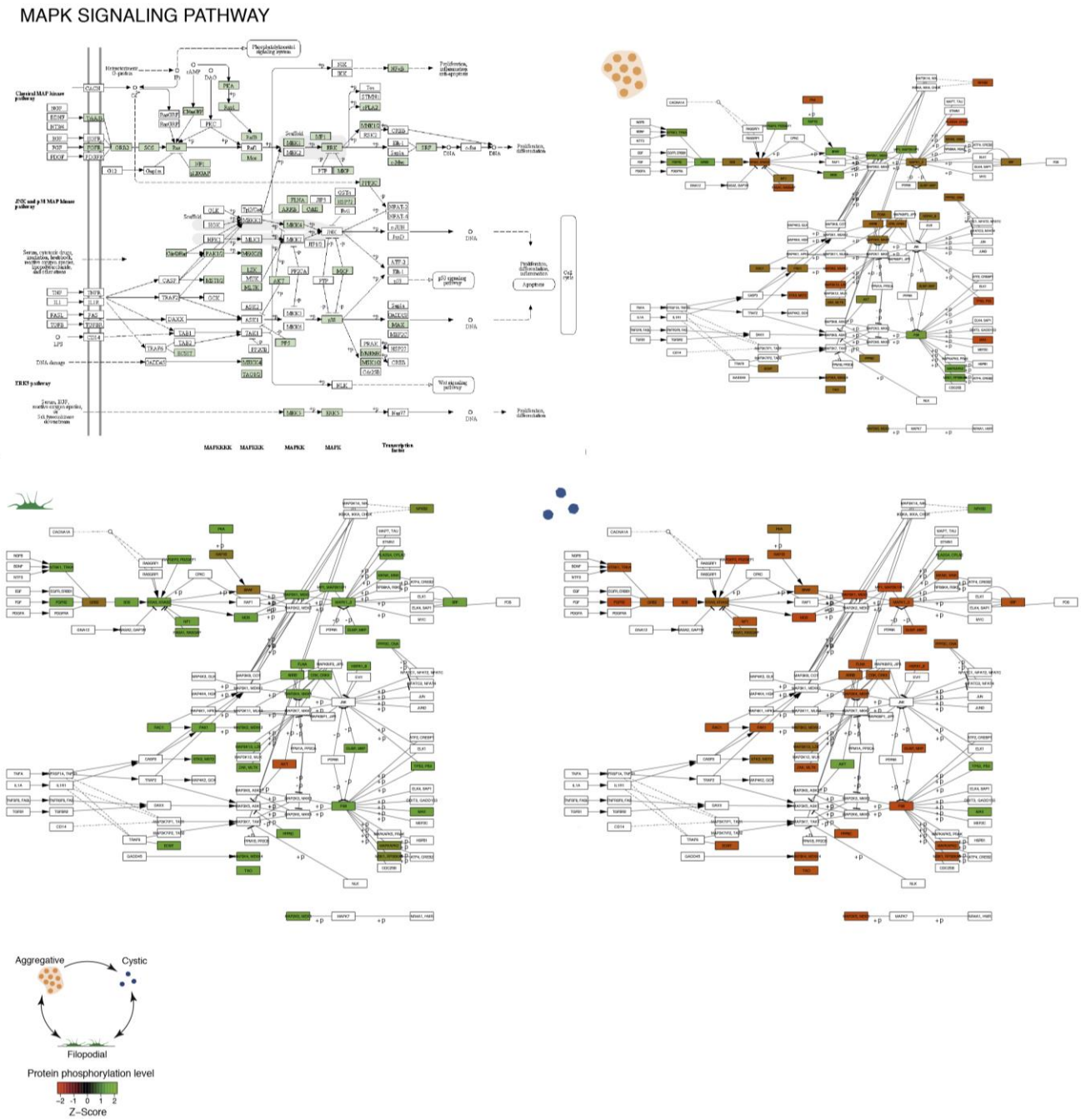



**Figure S3. Concerted Differential Phosphorylation of the MAPK KEGG Pathway, related to Figure 4.** A schematic representation of the pathway is shown and the genes present in the *Capsaspora* genome are highlighted in light green. A color-coded pathway for each cell type indicates the normalized level of phosphorylation (Z-score) of each protein. Proteins absent in the *Capsaspora* genome or undetected in our proteomics dataset are shown in white. Notice the coherent phospho-activation of the MAPK pathway in the filopodial proliferative stage and, in contrast, the repression in the quiescent cystic stage. The main exception is the Akt/PKB protein, a negative regulator of the MAPK pathway, which follows an inverse pattern of phosphorylation.

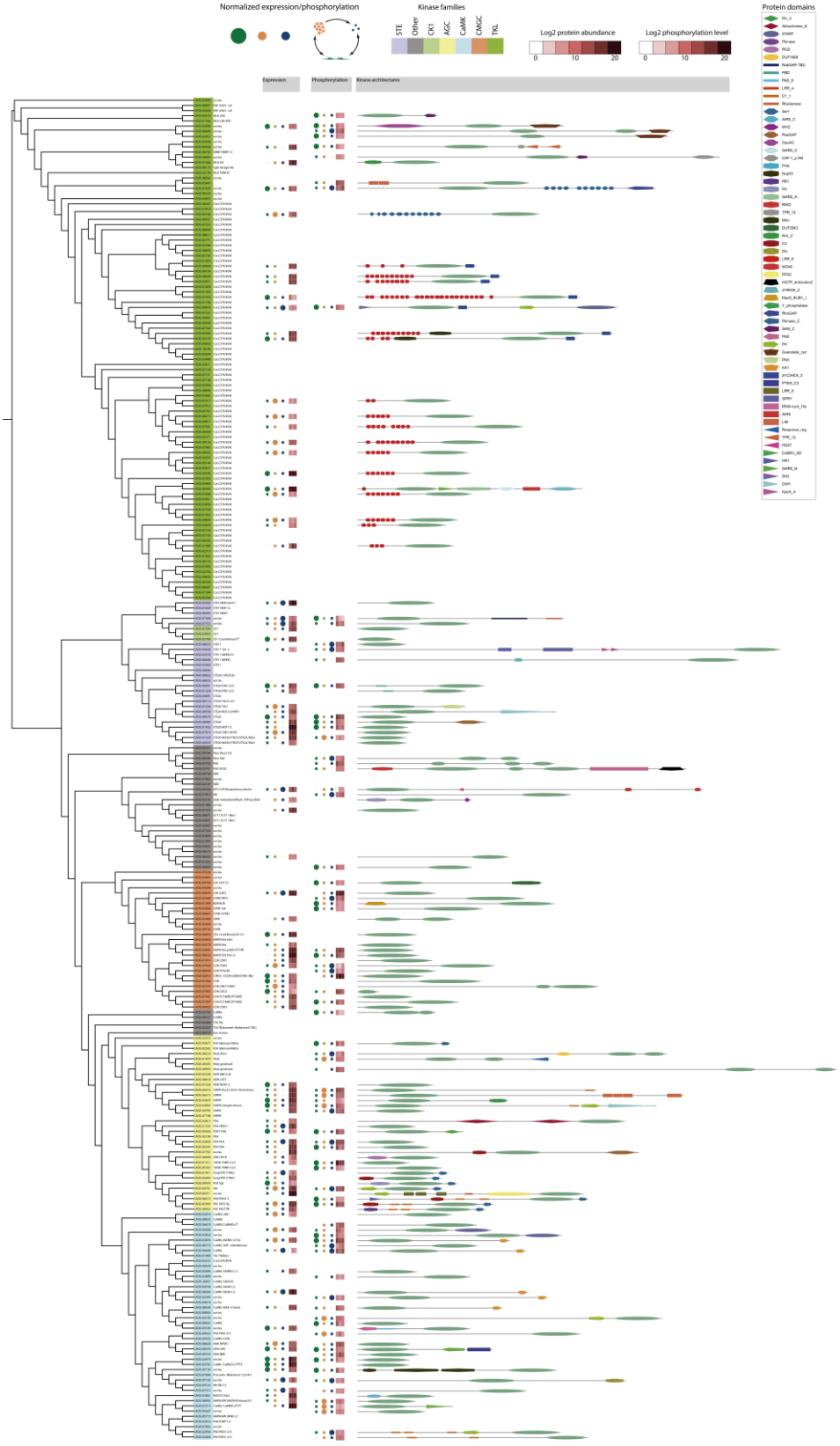

Normalized expression/phosphorylation
Kinase families
STE
Other
CK1
AGC
CaMK
CMGC
TKL
Log2 protein abundance
0
10
20
Log2 phosphorylation level
0
10
20
Protein domains
Expression
Phosphorylation
Kinase architectures

**Figure S4. The *Capsaspora* Kinome, retated to Figure 4.** Maximum likelihood phylogenetic tree of *Capsaspora* serine/threonine kinases. Tree leaves (Gene accession number) are colored according to kinase family classification (as defined in Suga et al 2013 (Suga et al., 2013)), and the subfamily classification is indicated next to each gene ID (if known) (Suga et al., 2013). The first column of data refers to protein expression. The dot sizes represent the relative (Z-score) level of expression of each kinase in each cell type (to be interpreted only horizontally) and the heatmap represents absolute expression levels (in order to allow vertical comparisons between different kinases). The second column refers to protein phosphorylation, and similar to the first column, dots and heatmaps indicate relative and absolute levels of phosphorylation. The third column represents the structure of Pfam domains of each detected protein kinase. The length of the representation is proportional to the length of the protein.

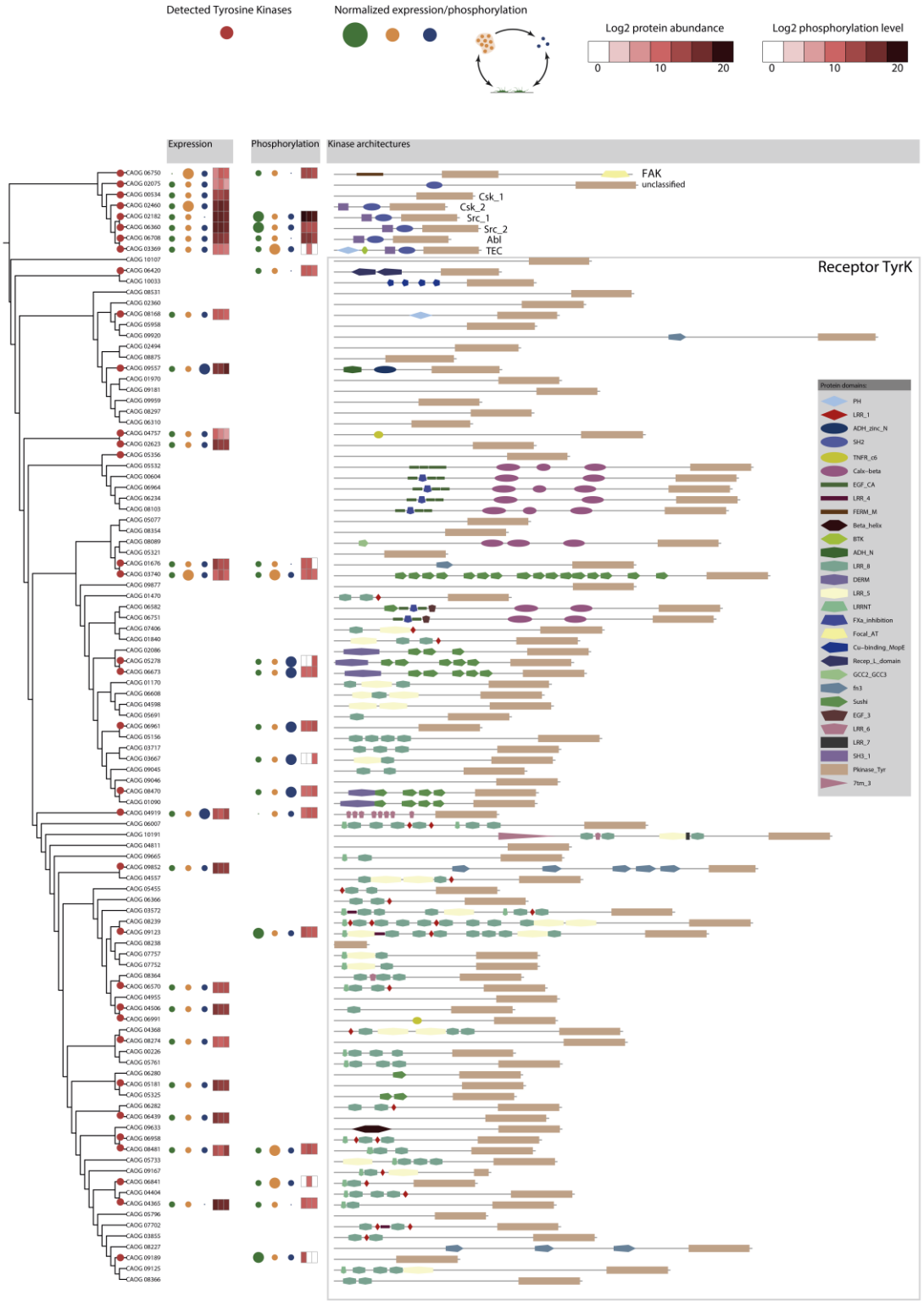

Detected Tyrosine Kinases
Normalized expression/phosphorylation
Log2 protein abundance
0
10
20
Log2 phosphorylation level
0
10
20
Expression
Phosphorylation
Kinase architectures
FAK
unclassified
Csk_1
Csk_2
Src_1
Src_2
Abl
TEC
Receptor TyrK
Protein domains:
PH
LRR_1
ADH_zinc_N
SH2
TNFR_c6
Calx-beta
EGF_CA
LRR_4
FERM_M
Beta_helix
BTK
ADH_N
LRR_8
DERM
LRR_5
LRRNT
FXa_inhibition
Focal_AT
Cu-binding_MopE
Recep_L_domain
GCC2_GCC3
fn3
Sushi
EGF_3
LRR_6
LRR_7
SH3_1
Pkinase_Tyr
7tm_3

**Figure S5. The *Capsaspora* Tyrosine Kinome, related to Figure 5.** Maximum likelihood phylogenetic tree of *Capsaspora* tyrosine kinases. A red dot indicates that the tyrosine kinase has been identified. The first column of data refers to protein expression. The dot sizes represent the relative (Z-score) level of expression of each tyrosine kinase in each cell type (to be interpreted only horizontally) and the heatmap represents absolute expression levels (in order to allow vertical comparisons between different tyrosine kinases). The second column refers to protein phosphorylation, and similar to the first column, dots and heatmaps indicate relative and absolute levels of phosphorylation. The third column represents the structure of Pfam domains of each detected protein kinase. The length of the representation is proportional to the length of the protein.

**Table S1. Proteomics Data, related to Figure 1.** (A) List of identified proteins across *Capsaspora* life stages. (B) List of identified peptide-spectrum matches in all *Capsaspora* life stages. (C) Protein quantitation results for all *Capsaspora* life stages. (D) Feature quantitation results for all *Capsaspora* life stages.

**Table S2. Phosphoproteomics Data, related to Figure 3.** (A) List of identified peptide-spectrum matches in all *Capsaspora* life stages after enrichment of phosphorylated peptides. (B) Quantitation results of the phosphopeptide features for all *Capsaspora* life stages.

**Table S3. Phosphotyrosine Proteomics Data, related to Figure 3.** (A) List of identified peptide-spectrum matches in all *Capsaspora* life stages after enrichment of phosphotyrosine-containing peptides. (B) Extracted ion chromatogram peak areas (MS1) corresponding to the identified phosphotyrosine containing peptides.

**Table S4. Phosphorylation Sites Motif Enrichment Analysis, related to Figure 3.**

**Table S5. Phosphotyrosine System Gene Counts and Percentage of Tyrosine Residues in the Genome, related to Figure 3.**

# Supplemental Experimental Procedures

## Cell Culture and Differentiation Conditions

*Capsaspora owczarzaki* cells were grown axenically in 5-ml flasks with ATCC medium 1034 (modified PYNFH medium) in a 23°C incubator. Three biological replicates (three independent clonal cell lines) were generated by subculturing from a single-founding cell and grown for 2 months. Adherent filopodiated cells were obtained by initiating a new 1/100 sub-culture (from an approximately $5 \times 10e^6$ cells/ml initial culture) and, after 3–4 days, cells were scratched from the substrate. Aggregate formation was induced by initiating a new 1/250 sub-culture (from an approximately $5 \times 10e^6$ cells/ml initial culture) and by gentle agitation at 60 rpm during 4–5 days. Finally, floating cystic cells were obtained from a 14-day-old culture, starting from a new 1/100 sub-culture (from an approximately $5 \times 10e^6$ cells/ml initial culture).

## Protein Extraction

Between 8x10e8 and 25x10e8 cells were collected by centrifugation 10 min at 5000g at 4ºC and washed once in sterile PBS1x. The cell pellet was resuspended in 1.5ml of *Capsaspora* protein extraction buffer (50mM TrisHCl pH8, 150mM NaCl, 0.1% SDS, 0.5% deoxycholic acid, 5mM EDTA, 1mM EGTA, 1% NP-40, 1mM MgCl2, 1mM CaCl2) to which protease inhibitors and phosphatase inhibitors where freshly added (including 1mM PMSF, pepstatin 1 ug/ml, 1 Complete Mini Tablet (Roche, Basel, Switzerland) and 1 PhoSTOP tablet (Roche)). Samples were incubated on ice for 15 min and sonicated three times for 30 seconds at 10% amplitude in a Branson Sonifier 250 with a microtip (Branson Ultrasonics, Danbury, USA). Samples were centrifuged for 20 min at 20.000g at 4ºC in order to remove cellular debris and the supernatants were recovered and stored at -80ºC.

## Sample Digestion

Cell lysates were precipitated with six volumes of cold acetone at -20 ºC overnight. Pellets were resuspended with 6 M Urea 200 mM $NH_4HCO_3$ and sonicated in cycles of 30 sec for 15 min (in a Bioruptor system). Proteins were reduced with DL-Dithiothreitol (ratio 3 nmol:1µg protein) 1 h at 37ºC, alkylated with iodoacetamide (ratio of 6 nmol:1µg protein) 30 min at room temperature in dark. The resulting protein extract was then diluted 1:6 with 200 mM $NH_4HCO_3$ and digested

with sequencing grade modified trypsin (Promega, enzyme/substrate ratio 1:10; w:w) at 37 ºC overnight. Trypsin activity was quenched by the addition of formic acid (FA) to a final concentration of 10%.

**Phosphopeptide Enrichment**

Phosphorylated peptides were enriched using an offline immobilized metal affinity $TiO_2$ protocol. Briefly, microcolumns were prepared by placing a small plug of $C_8$ material (3M Empore $C_8$ extraction disk) into the end of a GeLoader tip (Eppendorf, Hamburg, Germany). $TiO_2$ beads were rinsed with washing buffer (WB: 0.1 % trifluoroacetic acid + 80 % acetonitrile in water), then suspended in loading buffer (LB: 6 % trifluoroacetic acid + 80 % acetonitrile in water) in a concentration 10 µg/µl. A slurry-suspension of $TiO_2$ (50 µl) was packed in the GeLoader tip by centrifugation at 100 x *g*. Then, 250 µg of peptides were dissolved in LB (100 µl) and they were loaded into the $TiO_2$ microcolumn by centrifugation (100 x *g*). Microcolumns were rinsed with LB (50 µl) and subsequently with WB (50 µl). Phospho-enriched peptides were then eluted from the microcolumns with 5 % $NH_{3(aq)}$ (30 µl), followed by an additional elution step with 2 % FA + 80 % acetonitrile in water. Eluted peptides were acidified with 30 µl of 20 % FA in water.

Phosphotyrosine-containing peptides were enriched by immunoaffinity purification. Briefly, protein G bound to agarose beads (30 µl, Calbiochem) was incubated with a mix of three anti-pTyr antibodies (PT66, Sigma; PY100, Cell Signaling Technologies; and 4G10, Millipore; 4 µg each) and 250 µl of IP buffer (100 mM Tris, 1% Nonidet P-40, pH 7.4) for 8 h at 4 ºC. Antibody-conjugated protein G agarose was rinsed three times with 400 µl IP buffer. Then, 3 mg of tryptic peptides dissolved in 1.5 ml of IP buffer were added to the loaded agarose beads, and incubated with rotation overnight at 4ºC. Conjugated protein G was rinsed with 400 µl IP buffer and four times with 400 µl Rinse Buffer (100mM Tris pH 7.4). Enriched phosphotyrosine-containing peptides were eluted with 100 µl 100 mM glycine pH 2. A further elution step was performed with 100 µl 0.1% TFA.

**LC-MSMS Sample Acquisition**

Preceding the LC-MSMS analysis both peptide and phosphopeptide samples were purified by reverse phase columns (Hypersep $C_{18}$ from Thermo; or MicroSpin $C_{18}$ from The Nest Group). The resulting peptide mixtures were dried using a speed-vac system.

Samples were analyzed in a LTQ-Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) coupled to an EasyLC (Proxeon, Odense, Denmark) equipped with a reversed-phase chromatography 25-cm column with an inner diameter of 75 μm, packed with 1.9 and 3 μm $C_{18}$ particles (Nikkyo Technos Co., Ltd. Japan). Chromatographic gradients from 93 % buffer A, 7 % buffer B to 65% buffer A, 35% buffer B in 360 min for the proteome and phosphoproteome, and in 120 min for phosphotyrosine samples. In both cases the flow rate was set at 250 nl/min. Buffer A: 0.1% formic acid in water. Buffer B: 0.1% formic acid in acetonitrile.

The mass spectrometer was operated in positive ionization mode with nanospray voltage set at 2.5 kV and source temperature at 200 °C. Ultramark 1621 for the FT mass analyzer was used for external calibration prior the analyses. Moreover, an internal calibration was also performed using the background polysiloxane ion signal at m/z 445.1200. The instrument was operated in DDA mode and full MS scans with 1 micro scans at resolution of 60,000 were used over a mass range of m/z 350-1500 with detection in the Orbitrap. Auto gain control (AGC) was set to 1E6, and dynamic exclusion (60 seconds) and charge state filtering disqualifying singly charged peptides were both activated. In each cycle of DDA analysis, following each survey scan the twenty most intense multiple charged ions were selected for fragmentation at normalized collision energy of 35% (ion count threshold = 5000). Fragment ion spectra produced via collision-induced dissociation (CID) were acquired in the ion trap, AGC was set to 5E4, isolation window of 2.0 m/z and maximum injection time of 50 ms was used. Neutral loss masses of 32.7, 49, 65.4 and 98.0 m/z among the three most intense fragment ions, was set to trigger a multi-stage activation acquisition in the phosphorylation database. For the particular case of phosphotyrosine-enriched samples neutral losses of 80, 40 and 26.7 m/z were also considered. All data was acquired with Xcalibur software v2.2.

**Mass Spectrometry Data Analysis**

Acquired data was analyzed using Mascot search engine (v2.3, Matrix Science) for peptide and TiO2-enriched phosphopeptide identification, and Andromeda search engine (MaxQuant v1.4.0.5) for antibody-based enrichment of pTyr-containing peptides. Data was searched against an in-house generated database based on *Capsaspora owczarzaki* genome v3. A precursor ion mass tolerance of 10 ppm at the MS1 level was used, and up to three miscleavages for trypsin were allowed. The fragment ion mass tolerance was set to 0.5 Da. Oxidation of methionine and

protein acetylation at the N-terminal were defined as variable modification. In the case of the phospho-enriched samples, phosphorylation on Ser, Thr and Tyr were also considered as variable modifications. Carbamidomethylation on cysteines was set as a fix modification in all cases. All identified peptides were filtered using a 5 % FDR threshold (peptide level). Identified pTyr-containing peptides were quantified using the area under the curve of extracted ion chromatograms (MS1) (Schilling et al., 2014). In the analysis of Tyrosine Kinase (TK) phosphorylation, only pTyr residues present in the TyrK catalytic domain were considered, as a proxy of TK activation by autophosphorylation.

Peptide and phosphopeptide (pSer, pThr) areas were extracted from the peptide features using the feature detection algorithm implemented in Progenesis (v4.1, Nonlinear Dynamics Inc.). Normalized abundances of unique peptides were used to estimate protein quantities, and significant differences in protein levels were evaluated using an ANOVA model. Heatmaps of the proteins and phosphopeptides exhibiting significant differences in abundance (ANOVA q-value<0.01 and q-value<0.05, respectively) where plotted using 'heatmap.2' R function, from the gplots library (Warnes et al., 2015). To draw these heatmaps we performed a hierarchical clustering of the normalized (Z-score) data with distances based on Pearson correlation and using the 'ward' agglomeration method as implemented in the 'hclust' R function. Finally, the visualization networks of differentially expressed and phosphorylated genes per stages were produced using Cytoscape v3.1.1 (Cline et al., 2007).

**Phylome Reconstruction and Analysis**

In order to gain insight into the evolutionary dynamics of the genome of *Capsaspora*, compare it with representative species of major eukaryotic lineages, and establish the relative age of genes, we reconstructed several phylomes (Sicheritz-Pontén and Andersson, 2001), *i.e.* the complete collections of evolutionary histories of all genes encoded in a genome and their homologs across other fully-sequenced species.

Proteins encoded in 18 fully sequenced eukaryotic genomes were downloaded from various sources (Table 1). Additionally, a proteome based on RNA-Seq data for *Ministeria vibrans* was included. Five phylomes were reconstructed with these species but using 5 alternative species as seed (species marked in bold in Table 1). We were able to reconstruct 6,598 (75.49%) single-gene trees using as seed the 8,740 unique protein sequences from *Capsaspora*. We were also able to reconstruct 18,538 (87.91%), 20,020 (90.67%), 4,441 (68.08%), and 10,199 (88.55%) single-gene

trees in the context of the *Homo sapiens* (Human), *Mus musculus* (Mouse), *Saccharomyces cerevisiae* (Yeast), and *Trichoplax adhaerens* phylomes, respectively.

| Species Name | Unique Sequences | Genes | Source | As in |
|---|---|---|---|---|
| **Capsaspora owczarzaki ATCC 30864** | **8,740** | **8,759** | **Broad Institute** | **11/2013** |
| **Homo sapiens** | **21,088** | **23,660** | **Ensembl - 73** | **09/2013** |
| **Mus musculus** | **22,080** | **22,563** | **Ensembl - 77** | **11/2014** |
| **Saccharomyces cerevisiae** | **6,523** | **6,523** | **YGOB** | **10/2013** |
| **Trichoplax adhaerens** | **11,518** | **11,518** | **UniProt** | **01/2015** |
| Amphimedon queenslandica | 29,761 | 29,883 | Ensembl Metazoa | 07/2012 |
| Caenorhabditis elegans | 20,345 | 20,532 | Ensembl - 73 | 11/2013 |
| Cryptococcus neoformans | 6,244 | 6,244 | Quest for Orthologs - 04/2011 | 12/2012 |
| Dictyostelium discoideum | 12,714 | 13,073 | Quest for Orthologs - 05/2012 | 07/2012 |
| Drosophila melanogaster | 13,789 | 13,937 | Ensembl - 73 | 11/2013 |
| Ministeria vibrans | 19,506 | 19,506 | Genome Project - RNA-Seq data | 11/2013 |
| Monosiga brevicollis | 9,180 | 9,180 | JGI | 06/2011 |
| Naegleria gruberi | 15,634 | 15,634 | UniProt | 11/2013 |
| Nematostella vectensis | 24,424 | 24,424 | Quest for Orthologs - 04/2011 | 10/2011 |
| Rhizopus oryzae | 16,971 | 17,459 | Broad Institute | 06/2011 |
| Salpingoeca sp. ATCC 50818 | 11,614 | 11,624 | Broad Institute | 07/2012 |
| Schizosaccharomyces pombe | 5,072 | 5,665 | Quest for Orthologs - 05/2012 | 07/2012 |
| Sphaeroforma arctica JP610 | 18,207 | 18,213 | Broad Institute | 07/2012 |
| Spizellomyces punctatus | 8,942 | 8,952 | Broad Institute | 07/2012 |

**List of the 19 species used in the comparative genomics analysis of this study.** Columns indicate i) the scientific name of each species, ii) the number of unique longest transcripts, iii) the number of genes predicted for each species, iv) the data source, and v) when data was acquired. Note that the number of unique longest transcripts can be smaller than the number of

genes, since two longest transcripts with identical sequence are collapsed in a single entry. Species marked in bold were used as seed for reconstructing the phylomes used in this project.

The phylome reconstruction was performed as follows. For each encoded protein in the seed genome a Smith-Waterman (Smith and Waterman, 1981) search was used to retrieve homologs using an e-value cut-off of 1e-3, and considering only sequences that aligned with a continuous region representing more than 50% of the query sequence. Then, selected homologous sequences were aligned using three different programs: MUSCLE v3.8 (Edgar, 2004), MAFFT v6.861b (Katoh and Toh, 2008), and KAlign v2.04 (Lassmann et al., 2009). Alignments were performed in forward and reverse direction (i.e using the Head or Tail approach (Landan and Graur, 2007)), and the six resulting alignments were combined using M-Coffee (Wallace et al., 2006). The resulting combined alignment was subsequently trimmed with trimAl v1.4 (Capella-Gutierrez et al., 2009), using a consistency score cut-off of 0.1667 and a gap score cut-off of 0.1, to remove poorly aligned regions.

Phylogenetic trees based on Maximum Likelihood (ML) approach were inferred from these alignments. ML trees were reconstructed using the best-fitting evolutionary model. The selection of the evolutionary model best fitting each protein family was performed as follows: A phylogenetic tree was reconstructed using a Neighbour Joining (NJ) approach as implemented in BioNJ (Gascuel, 1997); The likelihood of this topology was computed, allowing branch-length optimisation, using eight different models (JTT, WAG, MtREV, VT, LG, Blosum62, Dayhoff, and DCMut), as implemented in PhyML v3 (Guindon et al., 2010). The two evolutionary models best fitting the data were determined by comparing the likelihood of the used models according to the AIC criterion. Then, ML trees were derived using these two models, using the default tree topology search method NNI (Nearest Neighbour Interchange), and the one with the best likelihood was used for further analyses. A similar approach based on NJ topologies to select the best-fitting model for a subsequent ML analysis has been shown previously to be highly accurate (Huerta-Cepas et al., 2011). Branch support was computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution, as implemented in PhyML v3.0. In all cases, a discrete gamma-distribution with four rate categories plus invariant positions was used, estimating the gamma parameter and the fraction of invariant positions from the data.

Next, orthology and paralogy relationships among genes from the 5 species used as seed and those encoded by the other genomes included in the different phylomes were inferred using a phylogenetic approach (Gabaldon, 2008). In brief, a species-overlap algorithm, as implemented in

ETE v2 (Huerta-Cepas et al., 2010), was used to label each node in the phylogenetic tree as duplication or speciation depending on whether the descendant partitions have, at least one, common species or not (i.e. using a Species Overlap Score of 0). The resulting orthology and paralogy predictions can be accessed through PhylomeDB.org (Huerta-Cepas et al., 2011). These predictions have been used in subsequent analyses such as orthology-based functional annotation or the relative dating of protein-coding genes.

Finally, we assigned a relative gene age to each gene (phylostratigraphy). To this end, we scanned the five available phylomes to date protein-coding genes which have been used as seed for single-gene trees. We used a previously-described algorithm (Huerta-Cepas and Gabaldon, 2011) to detect speciation and duplication events in which the seed protein is involved. We looked for orthologous and paralogous counterpart proteins which will set the relative age of each seed protein (see Figure 1B). In the case of *Capsaspora* we define seven relative ages: 1) Capsaspora-specific, 2) Filasterea, 3) Filozoa, 4) Holozoa, 5) Opisthokonta, 6) Unikonta, and 7) Eukaryota. We first looked for the furthest orthologous sequence age of each seed protein. If this information was not available, then we looked at the age of the most ancient duplication in which the seed protein. If such information was not available, we then looked for the most distant homologous sequence when considering BLAST results filtered only by the coverage (minimum overlapping region between query and target protein: 50%) and e-value (1e-3). Table 2 shows the results of the dating process for the different phylomes and strategies used.

| Seed Species | Annotated genes | Furthest ortholog | Most ancient duplicated | Blast-based annotation |
|---|---|---|---|---|
| *C. owczarzaki* | 8,754 / 8,759 | 3,896 | 2,750 | 2,108 |
| *H. sapiens* | 23,220 / 23,660 | 17,641 | 2,991 | 2,588 |
| *S. cerevisiae* | 6,523 / 6,523 | 2,642 | 1,852 | 2,029 |
| *M. musculus* | 22,459 / 22,563 | 17,644 | 2,843 | 1,972 |
| *T. adhaerens* | 11,516 / 11,518 | 7,173 | 3,026 | 1,317 |

**Number of seed protein dated using different strategies**. Columns indicate the species name, the total number of dated protein-coding genes, the number of dated proteins using the furthest orthologous sequence age, the number of dated proteins using the age of the most ancient duplication for which the seed protein was involved, and the number of dated protein using the

most distant homologous sequence relative age.

**RNA-protein Correlation Analysis**

RNA-protein level comparisons in *Capsaspora* were performed using RNA-seq data previously published (Sebé-Pedrós et al., 2013) (http://www.ncbi.nlm.nih.gov/biosample/?term=txid595528%5BOrganism:noexp%5D). This dataset corresponds to exactly the same life stages analyzed here and, similarly to this proteome dataset, included three biological replicates per stage.

The correlation between RNA and protein levels was calculated using the 'cor' function in R, both for Person and Spearman correlations, with averaged cross-sample/cross-replicate expression values (*ie*. obtaining a single expression value for each gene). All correlations reported are highly significant ($p<1e-05$) in correlation tests calculated using 'cor.test' R function. For the phylostratigraphic analysis of RNA-protein levels (Figure 2), genes were grouped by phylostrata and Spearman correlation was calculated, as well as a population of 1000 bootstrap values using 'boot' R library (Canty and Ripley, 2015; Davison and Hinkley, 1997) with default parameters. In a different analysis, coefficients of variation of RNA or protein levels across tissues (human) (Fagerberg et al., 2014; Kim et al., 2014), across temporal cell types (*Capsaspora*) (Sebé-Pedrós et al., 2013) or across a metabolic time series (yeast, only RNA CV) (Kuang et al., 2014), were calculated using R and, again, a population of 1000 bootstrap values using 'boot' R library. All bootstrap values were represented in violin plots using 'vioplot' R library (Adler, 2005).

In order to determine which genes of *Capsaspora* show a strong correlation between RNA and protein levels, we analyzed the samples pair-wise (Aggregative vs Filopodial and Cystic vs Filopodial) in order to determine the relationships between RNA and protein fold changes. We considered as "transcriptionally regulated" (red dots Figure S3) those genes with an absolute log2(Fold Change) in RNA and protein levels greater than 0.5 (in order to select only genes with enough change in expression to evaluate if the RNA and protein changes correlate or not) and in which the two fold changes are similar (log2FC RNA/log2FC protein < 3 and vice versa). Inversely, we considered as "non-transcriptionally regulated" genes with an absolute log2(Fold Change) in RNA and protein levels greater than 0.5 and in which the two fold changes are different (log2FC RNA/log2FC protein > 3 or vice versa). Finally, we merged the genes from both pair-wise comparisons in order to preform functional enrichment analysis of the two populations of genes (Figure S3).

**Secretome Prediction**

In order to predict the *Capsaspora* secretome *in silico*, we first used SignalP3.0 (Bendtsen et al., 2004) (D-cutoff=0.45) to retrieve all proteins with a signal peptide. From this initial list of candidates we discarded:

1. Proteins that contain, after the first 60 amino acids, a transmembrane domain, identified using TMHMM v2.0 (Krogh et al., 2001).
2. Proteins with an endoplasmatic reticulum retention signal with the C-terminal motifs KDEL or HDE[LF].
3. Glycosylphosphatidylinisotol (GPI) anchored proteins, identified using PredGPI (Pierleoni et al., 2008).
4. Proteins with a mitochondrial location tag, identified using TargetP v1.1 (Emanuelsson et al., 2007).

The resulting list comprised 384 proteins, of which 194 where detected in our proteomics dataset. We then analyzed the relationship between RNA and protein levels of these 194 genes. We normalized (Z-score) both protein and RNA levels and subtracted the normalized values (Protein – RNA levels). The resulting matrix was represented in a heatmap (Figure S4B) by performing a hierarchical clustering using distances based on Pearson correlation. Two clear clusters were visually identified and the individual gene RNA and protein profiles represented in line plots (Figure S4C) to further evaluate anti-correlation relationships.

**Phosphorylation Motif Analysis**

We first generated a dataset of centered 13-mer sequences by extending 6 amino acids around each phosphosite. We used this dataset as input for X-motif (Schwartz and Gygi, 2005) to find enriched motifs. We used the whole *Capsaspora* proteome as background, the default significance threshold ($p = 0.000001$) and a sequence occurrence threshold corresponding to 1% of the input sequences (External Database S4). We found 20 Serine-centered motifs and 8 Threonine-centered motifs, while no specific enrichments were found around Tyrosine phosphosites (these were, thus, simply classified as “Tyrosine”). The motifs found are very similar to those found in mouse by Huttlin et al. (Huttlin et al., 2010) and Villen et al. (Villén et al., 2007) and we used the same decision-tree algorithm to classify them into 5 categories:

1. If the phosphorylated residue is a tyrosine: “Tyrosine.”
2. If there is a Proline in the +1 position: “Proline-directed.”

3. If there are 5 or more Asp and Glu residues in positions +1 to +6: “Acidic.”

4. If there is an Arg or a Lys residue at position −3: “Basic.”

5. If there is an Asp or a Glu at +1, +2, or +3 positions: “Acidic.”

6. If there are 2 or more Arg or Lys residues between −6 and −1 positions: “Basic.”

7. If the site doesn't match any previous classes, classify as “Other.”

Enrichment of motifs categories per stages was done based on the different phosphopeptides identified in each stage (pooling the three replicates of IMAC-enriched phosphoproteome and of the two replicates of the phosphor-tyrosine enriched phosphoproteome).

**Functional and Phylostratigraphic Enrichment Analyses**

A gene ontology of all *Capsaspora* genes was generated using Blast2GO (Conesa et al., 2005) and GO enrichments of the different lists of differentially expressed genes (see above) were analyzed using Ontologizer (Bauer et al., 2008) using the Topology-Weighted method. A p-value threshold of 0.01 was used. KEGG pathway annotation of all *Capsaspora* genes was generated using the WebMGA (Wu et al., 2011) and KEGG pathway enrichments were calculated using Hoea (http://hoea.sourceforge.net/). Pfam domains of all genes were analyzed using Pfamscan v26 (a HMMER search-based algorithm) with default Gathering Threshold (Punta et al., 2012), and counts were generated using custom Perl scripts. Fisher’s exact tests were performed using custom R scripts. This Pfam annotation was also used to predict the domain architecture of the kinases shown in Figures S7 and S8. GO, Pfam and KEGG enrichments for any particular subset of proteins were performed using the total identified proteins as a background.

For the phylostratigraphic enrichment analyses, we assigned each to an evolutionary origin (phylostratum) based on the phylome data (see above): Eukaryota, Unikonta, Opisthokonta, Holozoa, Filozoa, Filastera or *Capsaspora*-specific (orphan genes). We determined the frequency distribution of each phylostratum in the analyzed gene set and used the distribution of all genes seen in the proteome as a background distribution to perform a Fisher exact test of enrichment as implemented in R (‘fisher.test’ function). In the case of the transcriptome differential enrichment gene set (Figure 1C), we used all the expressed genes (FPKM > 2) as the background distribution.

**Tyrosine Kinase Signaling Repertoire Counts**

We used HMMER 3.1 (Eddy, 2011) with domain profiles from Pfam and a Gathering Threshold as defined in Pfam database (Punta et al., 2012) to search for the number of components of the tyrosine kinase signaling system in the different species included in Figure 4.

We then counted the number of erasers (number of proteins containing Y_phosphatase domain (PF00102)) and readers (number of proteins containing SH2 (PF00017), IRS (PF02174) or PTB (PF08416) domains).

The number of writers (Tyrosine kinases) cannot be inferred with the previous approach, as no domain univocally identifies Tyrosine kinases. For this reason, writer counts were obtained from available literature where Tyrosine kinases have been annotated phylogenetically (Clarke et al., 2013; Fairclough et al., 2013; Manning et al., 2008; Srivastava et al., 2010; Suga et al., 2012, 2014). For two relevant species of early-branching animals (*Trichoplax adhaerens* and *Mnemiopsis leyidi*) there was not data available. In these two cases, we constructed a maximum likelihood phylogeny of all tyrosine kinase candidates of the two species together with well annotated kinomes from 'kinase.com' database: *Monosiga brevicollis* and *Strongylocentrotus purpuratus*. The sequences were aligned using the Mafft L-INS-i algorithm, optimized for local sequence homology (Katoh and Toh, 2008; Katoh et al., 2002). The alignment was then manually inspected and edited in Geneious software. This resulted in a matrix containing 107 amino acid residues. The Maximum likelihood (ML) phylogenetic tree was estimated by RaxML (Stamatakis, 2006) using the PROTGAMMALG model, which uses the Le and Gascuel (LG) model of evolution (Le and Gascuel, 2008) and accounts for between-site rate variation with a four category discrete gamma approximation. The best model was determined using ProtTest3 (Darriba et al., 2011).

The same approach was used to construct phylogenetic trees for *Capsaspora* kinases (Figure S7) and tyrosine kinases (Figures 4 and S8). All tree visualizations were produced using Interactive Tree Of Life tool (Letunic and Bork, 2007).

**Ancestral Phosphorylation Network Reconstruction**

One-to-one orthologs between the different species were extracted from our phylome analysis (see above). The only exception was mouse, in which we allowed up to one-to-four orthologs in order to account for the double whole-genome duplication occurred at the stem of vertebrata

(Dehal and Boore, 2005; Huminiecki and Heldin, 2010) and which results in many gene families having multiple orthologs. We then did pair-wise comparisons between species to determine lists of ortholog genes that are phosphoregulated in both species. We used available phosphoproteomics datatasets for yeast (Hustedt et al., 2014), *Trichoplax adhaerens* (Ringrose et al., 2013) and mouse (Huttlin et al., 2010) generated with a similar approach than ours. Finally, we used a parsimony criterion to infer phosphorylated ortholog repertoires at three ancestral nodes: Opisthokonta (shared by yeast and at least *Capsaspora*, *Trichoplax* or mouse), Holozoa (shared by *Capsaspora* and at least mouse or *Trichoplax,* excluding those present in yeast) and Metazoa (shared by *Trichoplax* and mouse, excluding those proteins found also in yeast or in *Capsaspora*).

# Supplemental References


Adler, D. (2005). vioplot: Violin plot.

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

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. *340*, 783–795.

Canty, A., and Ripley, B.D. (2015). boot: Bootstrap R (S-Plus) Functions.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

Clarke, M., Lohan, A.J., Liu, B., Lagkouvardos, I., Roy, S., Zafar, N., Bertelli, C., Schilde, C., Kianianmomeni, A., Burglin, T.R., et al. (2013). Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. Genome Biol. *14*, R11.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. *2*, 2366–2382.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics *21*, 3674–3676.

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

Davison, A.C., and Hinkley, D. V (1997). Bootstrap Methods and Their Applications (Cambridge: Cambridge University Press).

Dehal, P., and Boore, J.L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. *3*, e314.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. *2*, 953–971.

Fagerberg, L., Hallström, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., et al. (2014). Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. Mol. Cell. Proteomics *13* , 397–406.

Fairclough, S.R., Chen, Z., Kramer, E., Zeng, Q., Young, S., Robertson, H.M., Begovic, E., Richter, D.J., Russ, C., Westbrook, M.J., et al. (2013). Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. Genome Biol. *14*, R15.

Gabaldon, T. (2008). Large-scale assignment of orthology: back to phylogenetics? Genome Biol. *9*, 235.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. *14*, 685–695.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Huerta-Cepas, J., and Gabaldon, T. (2011). Assigning duplication events to relative temporal scales in genome-wide studies. Bioinformatics *27*, 38–45.

Huerta-Cepas, J., Dopazo, J., and Gabaldon, T. (2010). ETE: a python Environment for Tree Exploration. BMC Bioinformatics *11*, 24.

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L.P., Denisov, I., Kormes, D., Marcet-Houben, M., and Gabaldon, T. (2011). PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Res. *39*, D556–D560.

Huminiecki, L., and Heldin, C.H. (2010). 2R and remodeling of vertebrate signal transduction engine. BMC Biol. *8*, 146.

Hustedt, N., Seeber, A., Sack, R., Tsai-Pflugfelder, M., Bhullar, B., Vlaming, H., van Leeuwen, F., Guénolé, A., van Attikum, H., Srivas, R., et al. (2014). Yeast PP4 Interacts with ATR Homolog Ddc2-Mec1 and Regulates Checkpoint Signaling. Mol. Cell *57*, 273–289.

Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S. a, Villén, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. Cell *143*, 1174–1189.

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. *9*, 286.

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

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567–580.

Kuang, Z., Cai, L., Zhang, X., Ji, H., Tu, B.P., and Boeke, J.D. (2014). High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast. Nat Struct Mol Biol *21*, 854–863.

Landan, G., and Graur, D. (2007). Heads or tails: a simple reliability check for multiple sequence alignments. Mol. Biol. Evol. *24*, 1380–1383.

Lassmann, T., Frings, O., and Sonnhammer, E.L.L. (2009). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res. *37*, 858–865.

Le, S.Q., and Gascuel, O. (2008). An improved general amino acid replacement matrix. Mol. Biol. Evol. *25*, 1307–1320.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinforma. *23* , 127–128.

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. U. S. A. *105*, 9674–9679.

Pierleoni, A., Martelli, P.L., and Casadio, R. (2008). PredGPI: a GPI-anchor predictor. BMC Bioinformatics *9*, 392.

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

Ringrose, J.H., van den Toorn, H.W.P., Eitel, M., Post, H., Neerincx, P., Schierwater, B., Altelaar, a F.M., and Heck, A.J.R. (2013). Deep proteome profiling of Trichoplax adhaerens reveals remarkable features at the origin of metazoan multicellularity. Nat. Commun. *4*, 1408.

Schilling, B., MacLean, B.X., D'Souza, A., Rardin, M.J., Shulman, N.J., MacCoss, M.J., and Gibson, B.W. (2014). CHAPTER 7 MS1 Label-free Quantification Using Ion Intensity Chromatograms in Skyline (Research and Clinical Applications). In Quantitative Proteomics, (The Royal Society of Chemistry), pp. 154–174.

Schrimpf, S.P., Weiss, M., Reiter, L., Ahrens, C.H., Jovanovic, M., Malmström, J., Brunner, E., Mohanty, S., Lercher, M.J., Hunziker, P.E., et al. (2009). Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. PLoS Biol. *7*, e48.

Schwartz, D., and Gygi, S.P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. *23*, 1391–1398.

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

Sicheritz-Pontén, T., and Andersson, S.G. (2001). A phylogenomic approach to microbial evolution. Nucleic Acids Res. *29*, 545–552.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol. Biol. *147*, 195–197.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E. a., 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.

Suga, H., Dacre, M., de Mendoza, A., Shalchian-Tabrizi, K., Manning, G., and 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.

Suga, H., Chen, Z., de Mendoza, A., Sebé-Pedrós, A., Brown, M.W., Kramer, E., Carr, M., Kerner, P., Vervoort, M., Sánchez-Pons, N., et al. (2013). The Capsaspora genome reveals a complex unicellular prehistory of animals. Nat. Commun. *4*, 1–9.

Suga, H., Torruella, G., Burger, G., Brown, M.W., and Ruiz-Trillo, I. (2014). Earliest holozoan expansion of phosphotyrosine signaling. Mol. Biol. Evol. *31*, 517–528.

Villén, J., Beausoleil, S. a, Gerber, S. a, and Gygi, S.P. (2007). Large-scale phosphorylation analysis of mouse liver. Proc. Natl. Acad. Sci. U. S. A. *104*, 1488–1493.

Wallace, I.M., O'Sullivan, O., Higgins, D.G., and Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. *34*, 1692–1699.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W.H.A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., et al. (2015). gplots: Various R Programming Tools for Plotting Data.

Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC Genomics *12*, 444.