Title: Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals

Alex de Mendoza 1,*, Hiroshi Suga 1,2, Jon Permanyer 3,4, Manuel Irimia 3,4, Iñaki Ruiz-Trillo 1,5,*

1. Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain.
2. Prefectural University of Hiroshima, 562 Nanatsuka-cho, 727-0023, Shobara City, Japan.
3. EMBL-CRG Systems Biology Unit, Centre for Genomic Regulation (CRG), 08003, Barcelona, Spain.
4. Universitat Pompeu Fabra (UPF), 08003, Barcelona, Spain.
5. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

# present address: ARC CoE Plant Energy Biology (University of Western Australia), 35 Stirling Highway, 6009, Crawley, Australia.

* For correspondence: alexmendozasoler@gmail.com (AdM), inaki.ruiz@ibe.upf-csic.es (IRT).

Abstract

Cell-type specification through differential genome regulation is a hallmark of complex multicellularity. However, it remains unclear how this process evolved during the transition from unicellular to multicellular organisms. To address this question, we investigated transcriptional dynamics in the ichthyosporean Creolimax fragrantissima, a relative of animals that undergoes coenocytic development. We find that Creolimax utilizes dynamic regulation of alternative splicing, long inter-genic non-coding RNAs and co-regulated gene modules associated with animal multicellularity in a cell-type specific manner. Moreover, our study suggests that the different cell types of the three closest animal relatives (ichthyosporeans, filastereans and choanoflagellates) are the product of lineage-specific innovations. Additionally, a proteomic survey of the secretome reveals adaptations to a fungal-like lifestyle. In summary, the diversity of cell types among protistan relatives of animals and their complex genome regulation demonstrates that the last unicellular ancestor of animals was already capable of elaborate specification of cell types.
Introduction

The process by which multicellular animals develop from a unicellular zygote is believed to mirror the first evolutionary steps that led to the origins of animal multicellularity from a unicellular species (1). Since the development of complex multicellularity is dependent upon differential genome regulation, the evolutionary onset of animal multicellularity must likewise have involved the appearance of differential genome regulatory capacities leading to distinct cell types. Many of the genes involved in the control of animal development and cell type identity, including signaling pathways and transcription factors, predate animal origins (2–5). Since these genes are known to be present in the genomes of the protistan relatives of animals, the unicellular holozoans (6–8), the evolution of complex multicellularity, with cell-type specific transcriptional programs must have involved changes in gene regulation. Therefore, a key step in understanding the evolution of multicellularity will be to infer the regulatory complexity of the last common ancestor of all living animals.

To address ancestral regulatory complexity, it will be necessary to elucidate the molecular control of cell differentiation through development in the unicellular relatives of animals. Three distinct developmental modes that lead to transient simple multicellular forms have been described in the protistan relatives of animals (Figure 1) (9–11). Colonial clonal development has been shown to involve differential regulation of a few multicellularity-related genes in the choanoflagellate Salpingoeca rosetta (8). On the other hand, the filasterean amoeba Capsaspora owczarzaki exhibits up-regulation of adhesion-related genes in its aggregative stage (9). To date, however, there has been no molecular characterization of coenocytic development, a third and completely distinct mode of development observed in the ichthyosporeans. Ichthyosporeans are the earliest branching holozoan lineage (12–14), and coenocytic development is a shared feature within the group (15,16). This developmental mode comprises a growth stage in which nuclei divide synchronously within a common cytoplasm before undergoing cellularization, followed by release of motile ameboid zoospores (10,17). These stages have distinct physiological and structural characteristics, the ameboid stage is mono-nucleated, non-diving and motile (Figure 1B), while the multinucleate stage has a cell wall, a big central vacuole and does not move (Figure 1C). Despite being quite distinct from canonic animal development, coenocytic development and multinucleate cell types are found in some animal lineages, such as in Drosophila syncytial blastoderm (10). Thus, in order to infer ancestral regulatory complexity by comparison of
premetazoan developmental modes, it will be essential to obtain a detailed molecular characterization of ichthyosporean coenocytic development.

Here we describe the transcriptome dynamics of *Creolimax fragrantissima*, which has been proposed as a model system for ichthyosporeans (10,17). We show that *Creolimax* employs complex gene regulation including alternative splicing and the use of long intergenic non-coding RNAs (lincRNA). Through analysis of the *Creolimax* secretome *in silico* and by proteomics, we also provide evidence of secondary adaptation to a specialized osmotrophic feeding mode through lateral gene transfer and gene duplication. Taken together, our results suggest that the last common unicellular ancestor of animals was already capable of implementing elaborate, cell-type specific differentiation programs.

**Results**

**Assembly and annotation of a reference genome**

As a reference genome for mapping RNA-seq data, we assembled the 45 megabase (Mb) draft genome of *Creolimax* from a combination of 454 reads and mate-paired end Illumina reads corresponding to a 75x coverage. The draft assembly comprises 82 scaffolds, with an N50 of 1.5 Mb (see Materials and Methods). The genome is twice as large as that of *Capsaspora* but in line with sequenced choanoflagellate genomes. We annotated 8695 genes, 92% of which are supported by transcriptional evidence. A diverse array of annotation pipelines rendered functional information on 78% of the predicted protein coding genes (see Material and Methods). Among those genes, many belong to gene families involved in multicellularity and development in animals, such as transcription factors and signaling pathways, as previously reported in targeted gene family analysis (18–20).

**Transcriptional dynamics reveal differences between *Creolimax* multinucleate stage and animal development**

We isolated two different stages of the *Creolimax* life cycle by taking advantage of the cell size difference between the amoeboid stage and the multinucleate growth stages (17). After filtering with a 5-μm mesh, we obtained a highly enriched culture of amoebae. The amoebae encysted and grew for at least 24 hours (Figure 1-figure supplement 1), and then multinucleate coenocytic cysts matured asynchronously, releasing new amoeboid zoospores with different cell sizes to form a heterogeneous culture. Given the drastic morphological difference between the non-motile mitotic multinucleate stage pre-dominant at 24 hours culture and the motile non-
dividing amoeba isolated after filtration, we decided to investigate their transcriptomic differences through RNA-seq (see Material and Methods).

The amoeboid and the multinucleate stage showed distinct transcriptional profiles, consistent among replicates (Figure 2A), from which we identified 956 genes as significantly differentially expressed. Functional enrichment analyses of Gene Ontologies (GOS) and PFAM domains ($P < 0.01$, Fisher’s exact test) revealed that the multinucleate stage shows up-regulation of genes associated with cell growth, including ribosome, translation, DNA replication, amino acid and RNA metabolism activities (Figure 2, Figure 2-figure supplement 1). Conversely, in the amoeboid stage we found enrichment for signaling activities (GTPases and Kinases) and an up-regulation of the actin cytoskeleton, most likely involved in the motile behavior of the amoeba (see raw GOS in Figure 2-source data 1). Strikingly, single-celled amoebas also showed up-regulation of extracellular matrix (ECM) adhesion, including the up-regulation of the integrin pathway components.

Besides signaling and adhesion activities, regulation through transcription factors is also considered to be a defining characteristic of animal development (19,21). To address whether that is the case for the multinucleate stage in *Creolimax*, we compared the stage-specific expression level of all the transcription factors encoded in the genome. Contrary to our expectation, the comparison revealed that TF expression is higher in the amoeboid stage ($P < 0.05$, Wilcoxon signed rank test, Figure 2-figure supplement 2 A). Among the 127 putative transcription factors found in the genome, only 5 are significantly up-regulated in the multinucleate stage. These 5 genes have DNA binding domains with ambiguous TF activity and include genes containing Myb_DNA-binding and CSD domains (Figure 2-figure supplement 2 B). In the amoeboid stage, we found significant up-regulation of 11 TFs, including T-box, Runx and HLH TFs, TF classes that have important roles in animal development (2). Altogether, these data indicate that, unlike in animal development, the multinucleate stage of *Creolimax* is not characterized by a tight regulation of adhesive, signaling, and transcription factor activities.

**Differential alternative splicing down-regulates specific pathways in *Creolimax***

Alternative splicing is a widespread mechanism for post-transcriptional gene regulation in animals and many other eukaryotes. To evaluate the extent to which alternative splicing expands transcriptome complexity in the intron-rich genome of *Creolimax* (6.5 introns per gene), we undertook a comprehensive analysis of intron retention and exon skipping events, and quantified their inclusion levels using previously developed pipelines (22,23) (see Material and Methods). We observed 3927 intron-retention events that affected 2172 genes. In contrast, we found only
211 genes affected by exon skipping. Despite having a two-fold higher intron density and longer
introns compared to *Capsaspora*, both holozoan species show comparable alternative splicing
profiles that are highly dominated by intron retention (9). This provides further support for the
hypothesis that exon skipping-dominated alternative splicing is a unique feature of animal
transcriptomes (9,24,25).

Next, we evaluated the extent to which intron retention is differentially regulated across
*Creolimax* stages. We found 865 introns with differential retention between stages (differential
Percent Intron Retention (ΔPIR) >15), the majority of which were more retained in the amoeboid
stage (Figure 3A). RT-PCR assays confirmed differential retention for all 10 tested introns (Figure
3D and Figure 3-figure supplement 2). Despite these marked differences in intron retention
between stages, the size distribution of retained introns was similar in both stages, and comparable
to that of constitutively spliced introns (PIR<2% in all samples) (Figure 3-figure supplement
1A,B). GO enrichment analysis of genes with amoeboid-specific intron retention revealed
enrichment in spindle pole formation and other mitosis-related activities (Figure 3C), supporting
an active role of IR in down-regulating these functions in this stage. Indeed, consistent with recent
reports in vertebrates (22), genes with amoeboid-specific intron retention showed significantly
lower steady-state mRNA levels compared to the multinucleate stage (Figure 3-figure supplement
1C; *P* < 0.0001, Wilcoxon signed rank test; the converse was true for genes with multinucleate-
specific intron retention). Therefore, intron retention is a conserved mechanism for reducing
transcript levels in pathway-specific genes from unicellular holozoans to vertebrates.

In the case of exon skipping, we found 63 exons with differential inclusion levels across
stages (Figure 3B). RT-PCR assays were used to validate the differential inclusion for all 12 tested
cases (Figure 3E and Figure 3-figure supplement 2). Only a minority of these exons (14) keep the
reading frame upon skipping, and only 2 of these overlapped with functional Pfam domains,
suggesting that, similar to intron retention, exon skipping in *Creolimax* predominantly contributed
to functionally down-regulating specific genes. GO enrichment analysis revealed that
differentially spliced exons belong to genes involved in various biological processes, including
channel activity and histone modifications (Figure 3F).

**A population of lincRNAs with regulated expression in *Creolimax***

Another layer of transcriptional complexity is provided by long non-coding RNAs, which are
increasingly recognized as important players in animal development and cell-type specific genome
regulation (26–28). To characterize the repertoire of long non-coding RNAs in *Creolimax*, we
assembled a *de novo* transcriptome from the RNA-seq data. We filtered out transcripts shorter than
200 bp and mapped the rest to the genome. We then applied a pipeline to identify putative non-coding RNAs, including searches for coding potential, homology, untranslated region (UTR) mis-annotation and low expression (see Material and Methods). Restricting our search to transcripts that did not overlap with protein-coding genes, known as long-intergenic RNAs (lincRNAs), we identified 692 putative lincRNA loci in *Creolimax*. In comparison to protein-coding transcripts, lincRNAs in *Creolimax* were shorter in length, harbored fewer exons, had longer exons and had a lower GC content (Figure 4-figure supplement 1). Moreover, overall transcription levels of lincRNAs were significantly lower than those of protein-coding genes (*P* < 0.01, Wilcoxon rank sum test). Interestingly all those characteristics have been reported for animal lincRNAs (29–31).

To infer possible functions of *Creolimax* lincRNAs, we first looked at the functional annotations of the closest neighboring protein-coding genes, as animal lincRNAs are enriched near developmental genes and transcription factors (26). In *Creolimax*, the only significantly enriched GOs of lincRNA closest neighboring genes were metabolic activities (*P* < 0.01). However, when we analyzed the vicinity of all the transcription factors, we found that 23.6% had at least one neighboring lincRNA, a significant enrichment compared to the rest of the genome (14.5%, *P* = 0.0007, Fisher’s exact test). Next, we searched for putative homologs of the lincRNAs in the transcriptomes of closely related species and other unicellular holozoans, but we did not retrieve any positive hits. This pattern of rapid evolutionary sequence turnover of lincRNAs has also been described for animals (31,32), where homology detection based on sequence similarity is restricted to short evolutionary distances.

In animals, expression of lincRNAs is generally restricted to specific tissues and organs (26,33). In *Creolimax* we detected 51 lincRNA loci that were differentially expressed between the amoeba and the multinucleate stage (Figure 4). Overall, only 7% of the total detected lincRNAs are differentially expressed, compared to 10% of coding genes (*P* = 0.0059, Fisher’s exact test). Thus, in stark contrast to animals, the lincRNAs in *Creolimax* appear to be less cell-type specific than coding genes.

To test whether the transcription of lincRNAs is linked to their upstream genes, we first subdivided lincRNAs into two categories: head-to-head, where the lincRNA and the adjacent protein-coding gene have opposite orientation and may share the same promoter, and head-to-tail, where the lincRNA promoter is downstream of the adjacent protein-coding gene (Figure 4-figure supplement 2). Head-to-head orientation was significantly over-represented in the genome (485/692, 70% of the total vs the 50% expected by chance, *P* < 3.6e-14, *χ*² test). Furthermore, pairs in this head-to-head orientation showed a higher proportion of positively correlated expression levels, although this was similar to pairs of protein-coding genes in the head-to-tail
orientation. Moreover, 80% of differentially regulated lincRNAs are in head-to-tail orientation or display uncoupled transcriptional profiles in relation to their neighbor genes ($P < 0.05$ Pearson’s correlation coefficient). Although the possibility that bidirectional promoters are the main source of lincRNAs in *Creolimax* cannot be excluded, differential regulation of lincRNA is largely independent of their neighboring protein-coding genes.

Finally, we analyzed the impact of alternative splicing in lincRNAs. Although the majority of lincRNAs have a single exon, 132 (19.1%) loci have two or more exons. Among those multi-exon transcripts, we detected 183 intron-retention events (Figure 4E). Of these, 8 events differed between samples ($PIR > 15$), suggesting differential regulation of lincRNAs also at the level of splicing. Therefore, both transcriptional regulation and alternative splicing may play a role in the active regulation of lincRNAs in *Creolimax*.

**Evidence of species-specific cell type evolution in Holozoa**

Direct comparison of steady-state transcript abundance is useful to reveal cross-species molecular homologies in animals, as homologous tissues in different species show more similarity in the transcriptional profiles than do non-homologous tissues in a single species (34–36). Using a similar rationale, we sought to compare the different cell stages of unicellular holozoans. From normalized transcriptional levels of a set of 2177 one-to-one protein-coding orthologs between *Salpingoeca*, *Capsaspora* and *Creolimax*, we calculated pairwise Spearman correlation distances, which were then used to perform hierarchical clustering and neighbor-joining tree reconstructions (Figure 5). Both approaches showed consistent trees with species-specific clustering for the different stages. The signal obtained from those 2177 genes was enough to cluster samples consistent with their whole transcriptome patterns (except one sample from *Salpingoeca*).

Direct cell-type comparisons between species revealed that *Creolimax* multinucleate stage and *Capsaspora* cystic stage are the most dissimilar among all samples, whereas the amoeboid stage in *Creolimax* has a stronger correlation with the adherent and aggregative stages of *Capsaspora* (Figure 5A). In order to characterize which gene sets were responsible for those cross-species similarities, we used a principal component analysis (PCA) to distinguish between the species-specific factors and the underlying biological similarities. Species-specific variability was captured by PC1 (47.96%); however, when we plotted PC2 and PC3 we obtained different groupings of cell types independent of their species of origin (Figure 5C). PC2 grouped the multinucleate stages of *Creolimax* with the aggregative and the adherent stages of *Capsaspora*. This grouping was explained by top loading genes involved in DNA replication and spliceosomal...
activities (Figure 5D), a pattern consistent with the down-regulation of mitotic activity and cell growth in the *Capsaspora* cystic stage (9). On the other hand, PC3 grouped the amoeboid stage of *Creolimax* and the adherent and aggregative stages of *Capsaspora*. PC3 was loaded with genes involved in the integrin adhesome and amoeboid actin-based motility and signaling activities (Figure 5D). Therefore, despite the vast phylogenetic divergence among holozoan species, some functional patterns can be recovered through direct comparison of transcriptional profiles.

We used the same approach to compare RNA-seq data from a wide range of human cell types and tissues with those of *Creolimax* (Figure 6A). Certain human cell types showed differential grades of similarity in the expression profiles with each of the *Creolimax* stages. The human cell types with a higher positive correlation with the multinucleate stage were those with high proliferation rates, including embryonic stem cells, induced pluripotent stem cells and transformed cell lines (293T, HeLA, K562) (Figure 6B). Similar patterns were obtained using PCA on the normalized steady-state transcriptional levels from both species. PC1 explained most of the variability due to species-specific transcriptomic profiles, whereas PC2 distinguished between highly proliferative cell types and the rest in both species (Figure 6C). Consistent with this observation, GO enrichment analysis of the top 125 genes that most contributed to PC2 showed highly significant enrichment for genes involved in genome replication and proliferation (Figure 6D). These results suggest that a signal from the evolutionarily conserved machinery for cell proliferation in eukaryotes (37, 38) can be detected from direct expression pattern comparisons between the multinucleate stage of *Creolimax* and highly proliferative human cell types.

**Conserved ancient co-regulated gene modules: the Integrin adhesome**

To obtain more specific insights into the evolution of co-regulated gene programs across holozoans, we investigated several gene modules with key roles in animal multicellularity. Despite gene co-regulation providing indirect evidence for shared functionality in a given species (39,40), comparative analysis of transcriptional co-regulated gene modules in different species may offer additional insights into the functional evolution of the animal multicellular toolkit.

First, we focused on the integrin adhesome, which is crucial for ECM adhesion in animals. The core components of the integrin adhesome have been identified in *Capsaspora* (41). Interestingly, these components are significantly up-regulated in the aggregative stage (9). In contrast, most integrin adhesome components have been lost in choanoflagellates (41). In the case of *Creolimax*, we found a nearly complete repertoire of the integrin adhesome components (Figure 7D), and most of them are significantly up-regulated in the amoeboid stage (Figure 7A). To test whether the integrin adhesome components constitute a co-regulated gene module in the three
unicellular holozoans (ichthyosporeans, filastereans and choanoflagellates), we calculated the transcriptional Pearson correlation coefficients between all of the genes encoding the integrin adhesome for each species (Figure 7A,B,C). In the case of Creolimax, we observed a remarkable co-regulation among all of the components. The only exception is the Src tyrosine kinase (Figure 7A), suggesting that tyrosine kinase signaling is not connected to the integrin signaling pathway in Creolimax, consistent with the absence of a Focal Adhesion Kinase in the ichthyosporean lineage (18). In the filasterean Capsaspora, we identified a more complex pattern of co-regulation, grouped in two submodules that are associated with specific paralogs of both alpha and beta integrins (Figure 7B). This complex pattern of co-regulation could be affected by the number of Capsaspora’s samples, higher than those analysed in Creolimax. In contrast, we did not detect any co-regulation among the few conserved integrin adhesome components in the choanoflagellate Salpingoeca (Figure 7C), suggesting that gene loss in the choanoflagellate lineage was accompanied by dismantling of this ancient co-regulatory module.

In animals, integrins mediate cell-adhesion through binding to ECM proteins, such as laminins and fibronectins (42). In Capsaspora, laminin-containing genes with predicted secretion signals are up-regulated in the aggregative stage (9). However, despite showing a co-regulated integrin adhesome, we did not find any laminin-containing genes in Creolimax (Figure 7E). The only ECM-related domains that we found in the genome were fibronectins, but those are fused to metabolic domains unrelated to the ECM and none had a secretion signal peptide (Figure 7F). Therefore, it seems unlikely that integrins are involved in adhesion to an endogenous animal-like ECM in Creolimax.

Next, we investigated the filopodial machinery (43). We observed strong positive co-regulation of a core set of components in all unicellular holozoans (Figure 7-figure supplement 1), consistent with the use of an ancestral cellular machinery for filopodial formation. In stark contrast to the situation for cell replication machinery, we did not observe a single co-regulated module in any unicellular species for genes involved in animal neuronal pre-synaptic and post-synaptic processes, including most pairs of genes that are known to directly interact in animals (with the exception of Syntaxin, Synaptobrevin and Synaptogryrin, which are involved in secretory vesicle formation (44); Figure 7-figure supplement 2). In fact, a lack of interaction has been shown for some of the core proteins involved in the post-synaptic scaffold in Salpingoeca (45). Thus, our results suggest that some molecular complexes directly involved in cell morphology and behavior already formed co-regulated gene modules in unicellular holozoans, whereas other complexes involved in unique animal cell types were assembled later, despite having conserved orthologs.
The secretome of Creolimax shows convergent adaptations to a specialized osmotrophic lifestyle

Ichthyosporeans are an interesting case of convergent evolution of fungal-like traits; in fact, they were once thought to be fungi based on their lifestyle and morphology (15,16). For instance, both groups have a cell wall, similar parasitic lifestyles and a specialized osmotrophic feeding mode, unlike any of the other holozoan lineages (Figure 1). Specialized osmotrophs are characterized by their highly adapted secretomes, which are key to the external digestion of complex compounds. Therefore, we analyzed the secretome of Creolimax to investigate the convergent evolution of this specialized osmotrophic feeding mode in the Ichthyosporeans. In addition, secretome analysis should provide further insights into the production and modification of ECM and signaling ligands.

To characterize the secretome, we performed a combination of in vivo and in silico approaches. First, using high-throughput proteomics of an in vivo secretome sample in culture conditions, we identified 91 proteins. Next, we applied an in silico approach to predict 453 proteins that are likely to be secreted (see Material and Methods). Interestingly, only 48 proteins were common to both datasets, indicating that proteins without a canonical signal peptide or with additional transmembrane domains could also be found in the in vivo secretome (Figure 8A).

Among the 43 non-canonically secreted proteins, we detected some with transmembrane domains that could be the product of shedding, as proposed for other species (46). For instance, we found peptides corresponding to the extracellular region of both alpha and beta integrins. Although shedding of integrins has been observed during the inflammatory response in animals (47), the functional implications of integrin shedding in Creolimax remain elusive.

Both secretome datasets were highly enriched in peptidase and proteolytic activities (Figure 8B), whereas only the in silico dataset showed further enrichments in other catabolic functions such as lipase and hydrolase activities. Analysis of Pfam-domain enrichment consistently revealed proteolytic and peptidase domains in both datasets (Figure 8C). In contrast, an in-depth search for candidate ECM-related proteins and diffusible ligands based on Pfam domains (using known animal domains as EGF, DSL or laminins) retrieved no positive hits. Thus, the proteolytic activity of the Creolimax secretome does not seem to be related to modifying the endogenous ECM, but possibly to the adaptation to a specialized osmotrophic feeding mode.

Osmotrophic lifestyles are characterized by the external digestion of complex polymers (48). Therefore, the enrichment in proteolytic activities observed in the secretome strongly suggests that proteins and peptides are the main food source of Creolimax, at least under culture conditions.
Such external digestion requires a coupled mechanism for nutrient uptake (48). Consistent with this requirement, we found 38 genes with 4 distinct Pfam domains (PF00324, PF01490, PF03169 and PF00854) that are predicted to be involved in the amino acid and oligopeptide transporter activity.

The most-abundant gene family in *Creolimax* secretome was the trypsins (15 proteins in the *in vivo* proteomic dataset out of the 31 genes found in the genome). Trypsins are usually found in multiple copies in animal genomes, and they have important roles in the digestive system as serine proteases and also as ECM remodellers. To further complement these observations, we profiled a wide range of eukaryotes and found independent expansions of trypsins in other osmotrophic lineages (Figure 9A). Among these, two fungal species (*Coemansia reversa* and *Conidiolobus coronatus*) and an oomycete (*Saprolegnia parasitica*) are also animal-dwelling parasites, thus sharing a similar lifestyle to the known ichthyosporeans. Phylogenetic analyses of *Creolimax* trypsins (Figure 9B) revealed that most are the product of rapid lineage-specific gene duplications, a common source of molecular adaptation (49). Moreover, we found very distinct patterns of transcript and protein abundance (measured by number of unique peptides identified) across the trypsin paralogs (Figure 9B), suggesting a rapid process of functional diversification. Thus, trypsins seem to be a recurrently used effector gene family in osmotrophic eukaryotes, and have evolved rapidly in *Creolimax*.

**Recurrent prokaryotic lateral gene transfer events shaped the *Creolimax* secretome**

Lateral gene transfer (LGT) is a rare but recurrent mechanism for the acquisition of osmotrophy-related genes in both fungi and oomycetes (48). We therefore investigated LGT in *Creolimax* by building automatic taxon-rich phylogenies. Manual inspection of these trees yielded a list of 163 genes that confidently branched within bacterial clades, supporting bacterial origins. Of these, 35 (21.5%) were found only in *Creolimax* among all eukaryotes sampled, 50 (30.7%) were shared with the close relative *Sphaeroforma arctica* and 71 (43.6%) were shared with other ichthyosporeans (Figure 10-figure supplement 2). Although most LGT genes (102 out of 163) were intron-less, the vast majority (143, 87.7%) showed transcriptional support from our polyA-selected RNA-seq, minimizing the possibility of bacterial contaminations. LGT from bacteria accounts for 1.8% of the total proteome, which is slightly higher than in other eukaryotic genomes (50). Importantly, we found 6 genes acquired by LGT in the *in vivo* secretome and 18 more in the *in silico* secretome (Figure 10-figure supplement 2). Therefore, similar to other eukaryotic lineages (51,52), ichthyosporeans have enriched their osmotrophy-related gene complement through prokaryotic LGT.
Among the 6 genes of prokaryotic origin that we found in the in vivo secretome, 3 belonged to the spore coat homology (CotH) family (PF08757). CotH proteins were first described as being fundamental for spore coat formation in the bacteria *Bacillus subtilis* (53), but they have been recently characterized as critical factors for host invasion in the fungus *Rhizopus oryzae* (54). Our phylogenetic reconstruction of CotH family evolution revealed that the presence of CotH homologs in *R. oryzae* and other fungi (only belonging to the class mucorales) originated from an LGT event that was independent of those found in ichthyosporeans (Figure 10 and Figure 10 - figure supplement 1). Interestingly, both lineages have expanded their CotH family members after LGT acquisition. Moreover, CotH genes were also found in other eukaryotic and archaeal lineages, suggesting a complex history of inter-domain LGT similar to that recently described for other gene families (55,56). Active gene duplication and domain shuffling characterized the evolutionary history of CotH in the Ichthyosporea, where we could observe variable transcriptional levels and peptides among distinct paralogs, as well as the acquisition of an N-terminal LTD domain (PF00932) (Figure 10B). Active transcription and secretion of *Creolimax* CotH genes in axenic culture conditions underscore their putative role in host invasion and highlight the importance of LGT in effector gene acquisition across different osmotrophic lineages.

**Discussion**

In this study, we show how different layers of genome regulation shape the coenocytic development and life cycle of the ichthyosporean *Creolimax*, a unicellular relative of animals. These regulatory layers include complex mechanisms that also play a role in animal development and cellular differentiation. In *Creolimax*, differential genome regulation is not only limited to the transcriptional control of protein coding genes, but also includes cell-stage specific alternative splicing and lincRNA expression. Despite some global similarities to animal regulation, we observed significant differences in how these regulatory layers are deployed in *Creolimax* and animals. For example, the population of lincRNAs that we describe in *Creolimax* does not show a major enrichment near developmental genes nor a highly specific cell-type dependent expression, both hallmarks of metazoan lincRNAs (26,31). The differential processing of the intron-rich genes during alternative splicing in *Creolimax* is dominated by intron retention and not exon skipping. Although this is similar to most other eukaryotic groups including sponges (57), it contrasts with alternative splicing in more complex animals, which predominantly involves exon skipping (9,24). Therefore, while alternative splicing in *Creolimax* is likely to contribute to the regulation of gene
expression, it does not provide a greater expansion of the proteome by generating multiple protein isoforms from individual genes as occurs in animals (25).

Intriguingly, while this complex genome regulation in *Creolimax* often involves similar gene toolkits to those employed in animal development, it is the unicellular amoeboid stage and not the multinucleate stage that is defined by multicellularity-related activities. This is the case for most differentially expressed transcription factors, signaling pathways and adhesion molecules, which are characteristically associated with animal development and multicellularity. In fact, our results indicate that the transcriptional profile of the multinucleate stage of *Creolimax* is more similar to those of highly proliferative cell types in humans, despite the obvious differences in cell morphology and cell division strategy (cell division vs. coenocytic nuclear division). Thus, we consider that the multinucleate stage of *Creolimax* can be regarded as a highly specialized proliferative cell type. This cell type can be considered to function in a manner analogous to stem cells, with undifferentiated nuclei dividing in the multinucleate cell before differentiation into amoeboid cells closes the life cycle. This would suggest that separation of functions as crucial as self-renewal and differentiation can occur in a unicellular context in a temporal manner, which pre-dates the exclusive ability of multicellular organisms to engulf both functionally distinct cell types within a single entity (58,59). Drastic differences in cell structure, morphology, function and molecular signatures found between stages in protistan relatives of animals indicate that cell stages can be considered cell types according to established definition (57 and see Author’s response section for an in depth discussion on this topic).

In addition to assessing the impact of genome regulation in *Creolimax*, our multiple genome-wide approaches reveal other novel aspects of ichthyosporean biology. For example, we show that *Creolimax* has undergone secondary adaptation to a specialized osmotrophic feeding mode, shaping its secretome and genome through both gene duplication and acquisition of bacterial genes by LGT. These observations highlight the uniqueness of the different holozoan lineages, each representing derived specializations from an ancestral state. However, these specializations of unicellular holozoans are achieved through a largely common genetic toolkit shared with animals, including several signaling pathways, transcription factors and adhesion molecules (5,7). Moreover, many of the components of this toolkit are assembled in co-regulated gene modules preserved since the common ancestor of all holozoans, suggesting that recurrent recruitment of full co-regulated gene programs underlies the evolution of lineage-specific cell types and developmental modes (60). Further complementary insights on the evolution of developmental modes will be provided by studying the immediate out-group of the Holozoa, the Holomycota (Fungi + Discicristoidea) (14). The Holomycota show a wide variety of developmental modes,
ranging from aggregative fruiting body formation to several modes of coenocytic development (61,62).

We have shown that the diversity of cellular behaviors and morphologies observed in holozoan life cycles is likely to have evolved from lineage-specific specializations. A widely accepted hypothesis states that the origin of animals involved a clonally dividing organism, similar to choanoflagellate colonies, that subsequently evolved specialized cellular differentiation (1,63). A competing hypothesis, however, suggests that pre-existing cell types and their associated molecular mechanisms were integrated in the spatiotemporal developmental dynamics of the last common ancestor of all animals (64). Our data suggests a third mixed model, in which the capacity to build differentiated cell types and transient multicellular entities was not a rare feature in pre-metazoan evolution. Nevertheless, the rise of animal multicellularity is not directly homologous to any ancestral developmental mode, and it may be seen as another derived specialization, involving the integration of ancestral molecular modules and their associated cell behaviors into one single multicellular entity. Consequently, our results reveal the importance of obtaining complementary data from multiple lineages before significant insights can be gained into the organism that took the first steps on the road towards complex animal multicellularity.

Materials and Methods

Cell culture and nucleic acid extraction

_Creolimax fragrantissima_ cells (available from the Canadian Centre for the Culture of Microorganisms under accession numbers CCCM 101—107) were grown axenically in liquid medium (Marine Broth Difco 2216) at 12ºC. To obtain biological replicates for the RNA extraction, three independent cell lines were isolated from distinct colonies grown on a Marine Agar plate (Marine Agar Difco 2216). Those initial isolates were then grown in liquid medium (Marine Broth Difco 2216). After one pass, new 1/10 subcultures (10 mL) were initiated and grown statically for 5 days in 25 mL flasks. When cells became confluent on the fifth day, cells were scratched and passed into 50 mL flasks with an additional 25 mL of fresh medium. These 50 mL flasks were then grown for 48 hours with gentle agitation (150 rpm), letting the mature coenocytes form clumps. Then, the 50 mL flasks were filtered using a 5.0 μm Isopore membrane filter (Millipore) and collected into a 50 mL Falcon tube. As only amoebas pass through the 5 μm filter step, filtered cells were then immediately pelleted by centrifugation at 1500 rcf for 3 minutes and harvested to get the RNA from the amoeboid stage. For the multinucleate-stage RNA, filtered cells were re-cultured in a new 50 mL flask, grown for 24 hours and harvested by centrifugation at
3000 rcf for 3 minutes. For all cell lines and stages, the RNA was extracted using Trizol reagent (Life Technologies) with a further step of DNaseI (Roche) to avoid gDNA contamination, and then purified using RNeasy columns (Qiagen).

**Genome sequencing, assembly and annotation**

We generated 1.7 million 454 single reads and 34 million Illumina 5kb mate-pair reads (both after trimming, totaling 3.4G bp). Those were combined and preassembled with a Newbler 2.7 assembler (Roche). The mitochondrial DNA sequence was removed before the assembly. Using the pairing information of the Illumina mate-pair reads, the 846 pre-scaffolds were broken at unreliable positions found by REAPR 1 (65) and re-assembled by SSPACE 2 (66). Some of the N-stretches within the scaffolds were filled by Gapfiller (67). This array of improvement tools assembled the pre-scaffolds into 196 sequence pieces. We then used the pairing information of the mate-pair reads for further assembly improvement, breaking and re-connecting the scaffolds. Finally, we obtained an assembly with 82 final scaffolds, of which 29 are short (<1000 bp) fragments.

To predict the protein-coding genes from the whole genome sequence, we used Augustus 2.7 (68) combined with RNA-seq data (see details on this data in “RNA-seq and differential expression analysis” section). We followed the protocol described here: [http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=IncorporatingRNAseq.Tophat](http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=IncorporatingRNAseq.Tophat). Briefly, we pooled all the RNA-seq samples and mapped them to the genome using Tophat2 (69), using the resulting introns to train Augustus *ab initio* predictions in an iterative process. The resulting predictions were manually screened in a genome browser and compared to the spliced-aligned reads resulting from Tophat2. We further validated our predicted annotation comparing the data to a set of genes that we had previously cloned by RT-PCR and RACE PCR, including highly expressed genes (e.g. Histone 2B, Tubulin beta) and lowly expressed genes (e.g. Myc, Grainyhead, p53, Src Tyrosine Kinase).

Moreover, we used the mapped transcriptome to perform a genome-guided Trinity assembly (70). Those transcripts were then used to annotate the UTRs of the protein-coding genes resulting from the Augustus annotation step. The elongation of the UTRs was done using PASA (71). The transcripts that did not overlap with Augustus annotations were then searched against the NCBI non-redundant protein database using tBlastX. Those that retrieved significant hits (e-value < 10e-3) and had clear open reading frames were then manually annotated as protein-coding transcripts. The resulting annotation retrieved 8695 genes, which are available here: [figshare.com/s/339648def79f11e49f3c06ec4bbcf141](http://figshare.com/s/339648def79f11e49f3c06ec4bbcf141).
To functionally annotate the genes, we used Blast2GO (72), searching the protein sequences against the NCBI non-redundant database using a BLASTP threshold of 10e-6 and the default InterProScan settings. We also performed a PfamScan search with PFAM A database version 26 using the default Gathering Threshold parameters (73). As a result, 6814 genes were functionally annotated.

RNA-seq and differential expression analysis

100-base paired-end libraries were constructed using the TruSeq Stranded mRNA Sample Prep kit (Illumina, San Diego, CA, USA). The libraries were sequenced in 2 lanes of an Illumina HiSeq2000 instrument at the CRG genomics unit (Barcelona, Spain). We obtained 417 million reads that were then mapped to the genome using Tophat2 (69), resulting in an average mapping of 82%. Raw gene counts and FPKM (fragments per kilobase of exon per million fragments mapped) values were obtained using Cufflinks2 (74). Differential expression analysis was performed by comparing the three replicates from each stage using DEseq (75) (threshold 5e-5), EdgeR (76) (threshold 5e-5), Cuffdiff2 (74) (threshold 5e-5) and NOIseq (77) (threshold 0.8). Only the genes that were identified as differentially expressed with at least 3 methods were taken, resulting in 956 genes. Data can be downloaded from GEO GSE68616.

Gene Ontology enrichments were obtained using the Topology-Weighted method in Ontologizer (78) taking a $P$ value lower than 0.01 as a threshold (see full list in Figure 2-source data 1). The resulting Gene Ontology (GO) enrichments were then visualized as a network in Cytoscape using the Enrichment map plug-in (79). Enrichment map plug-in connects GO terms according to gene annotation; therefore connected GOs belong to the same set of genes with multiple associated GOs. We used network connectivity between enriched GO terms as a criterion to collapse GO redundancy, shading general GO groups in more inclusive categories as seen in Figure 2. Inclusive categories complementary relied in functional similarity between GOs based in GO definitions (e.g. distinct unconnected aminoacid metabolic pathways are collapsed in “Aminoacid metabolism” inclusive category). Additionally, PFAM domain enrichments were calculated using a Fisher’s exact test implemented in R, taking a threshold of 0.01. PFAM enrichments were also visualized as a network using the Enrichment map plug-in, where connected nodes reflect domain presence in the same genes.

Genome-wide analysis of alternative splicing
Identification and quantification of exon skipping (including events with single or multiple cassette exons and microexons of 3 to 15 nucleotides) and intron retention (IR) were performed as previously described (22,23). For exon skipping, we used two complementary approaches. First, we implemented a “splice site-based module”, which utilizes the joining of all hypothetically possible exon-exon junction (EEJ) forward combinations from annotated and de novo splice sites (as described in(80)). To identify splice sites de novo, for each annotated splice site donor/acceptor, we scanned two downstream/upstream introns for potential splice-site acceptors/donors that would constitute a novel EEJ. Next, we mapped our RNA-seq data to this library of all potential EEJs, and considered “novel splice sites” those supported by at least five reads mapped to multiple positions of the EEJ. Then, we implemented our recently described “microexon module” (23), which also includes de novo searching of pairs of donor and acceptor splice sites in intronic sequences to detect novel, very short (3-15 nt) microexons. For intron retention (IR), we used our recently described pipeline (22), which employs a comprehensive set of reference sequences comprising exon-intron junctions (EIJs), intron sequences (if introns were longer than 200 nt, only a mid-intron segment of 200 nt was used), and EEJs formed by intron removal. Introns were classified as retained when there was a balanced accumulation of reads mapping to 5’ and 3’ EIJs and the intron body sequence, relative to the EEJ sequence. The level of retention was calculated based on “Percent Intron Retention” (PIR), which is the percentage of transcripts from a given gene in which the intron sequence is present.

In all modules, quantification of alternative sequence inclusion in the transcripts is derived only from junction reads (either EEJs or EIJs). To increase the fraction of mapping junction reads within each RNA-Seq sample, each read was first split into 50-nucleotide (nt) read groups using a sliding window of 25 nt. Therefore, each 100-nt (replicates a and b) and 125-nt (replicate c) read produces 3 and 4 overlapping reads, respectively. In addition, both read mates from the paired-end sequencing were pooled. These 50-nt split reads were then mapped to the genome using Bowtie (81) with –m 1 –v 2 parameters (unique mapping with no more than two mismatches). Reads that mapped to the genome were discarded for exon skipping quantifications. For quantification, only one random count per read group (i.e. all sub-reads coming from the same original read) was considered to avoid multiple counting of the same original sequenced molecule. In addition, for all modules and alternative splicing types, final read counts were corrected for the number mappable positions in each EEJ or EIJ following the formula:

\[
Corrected_{EEJ\_count} = EEJ_{count} \times \frac{Maximum_{mappability}}{EEJ_{mappability}}
\]
where EEJ<sub>count</sub> is the number of read groups mapped to the EEJ, Maximum<sub>mappability</sub> the maximum number of mapping positions that any EEJ can have for reads of length 50 nt (i.e. 35 positions), and EEJ<sub>mappability</sub> the number of positions that can be mapped uniquely to the EEJ using specific bowtie parameters (–m 1 –v 2), and thus EEJ<sub>mappability</sub> ≤ Maximum<sub>mappability</sub> (see (34,80) for details).

The different modules to detect and quantify AS have been integrated into vast-tools (https://github.com/vastgroup/vast-tools; species key “Cfr”). Associated files can be downloaded at http://vastdb.crg.eu/libs/vastdb.cfr.31.1.15.tar.gz.

We used a threshold of ≥20 PIR in at least one stage for positive intron retention events. As a threshold for exon skipping events we used skipping rates below 90% (measured using the metric Percent Spliced In, PSI) in at least one stage. In both cases, the minimum coverage allowed was 20 reads per splice junction. To evaluate differential alternative splicing, we used differences over 15 PSI or PIR between stages, allowing a standard deviation of < 10 between replicates.

**lincRNA annotation**

From the genome-guided Trinity assembly (see Genome sequencing, assembly and annotation) we further analyzed the transcripts that did not retrieve any significant TBLASTX hit against the NCBI non-redundant database (e-value > e-3) and were more than 200 bp long. To avoid lineage-specific protein-coding genes, we performed an additional TBLASTX search against 6 frame translations of the de novo assembled transcriptomes of several closely related species (Sphaeroforma arctica, Ichthyophonus hoferi, Pirum gemmata, Amoebidium parasiticum, Abeoforma whisleri, Corallochytrium limacisporum, Capsaspora owczarzaki, Salpingoeca rosetta, Monosiga brevicollis) and the protein coding genes of Creolimax, filtering out the positive hits (e-value < e-3). From the remaining transcripts, we performed a RfamScan_2 search against RFAM 11.0 (82) to annotate all known ncRNAs (e.g. U6, 18S, 28S). Additionally we used Coding Potential Calculator (CPC) (83) to discard all those transcripts with putative coding potential (coding potential score < -0.5). With the final list of transcripts, we selected those that did not overlap with gene+UTRs annotations or were close to uncertain assembly regions (multi-N stretches). For all those transcripts that were in a head-to-tail orientation regarding protein-coding genes, we manually inspected those that had an intergenic distance shorter than 1000 bp to the nearby gene to filter out misannotated UTRs. Additionally, we discarded transcripts overlapping repetitive regions of the genome. Finally we collapsed all the remaining transcripts into single loci and quantified their expression level using Cuffcompare and Cufflinks (74). From the resulting
To avoid noisy transcription, we filtered out all those that did not have at least 5 FPKM in at least one sample, and over 1 FPKM in any other sample.

To detect putative homology, lincRNAs were searched using BLASTN against the same list of closely related species described above. Differential expression analysis of the lincRNAs was done using the same parameters as for the coding genes (see RNA-seq and differential expression analysis). Consequently, we only accepted lincRNA loci as differentially expressed when they were identified by at least 3 out of 4 methods. We validated 6 out of 6 lincRNA using RT-PCR (see below). Finally, we used the same pipeline to detect alternative splicing events in coding genes, using a minimum coverage of 10 reads for each splice junction.

**RT-PCR**

To validate lincRNAs and alternative splicing events, RNA samples obtained as described in Cell culture, gDNA and RNA extraction were reverse transcribed to cDNA using SMARTer cDNA kit (CloneTech). For both stages, the same amount of initial purified RNA was used (1 μg). Pairs of primers with melting temperatures close to 60ºC were designed to capture the lincRNA and the alternative splicing events, and the PCR was performed using Expand high-fidelity Taq polymerase (Roche). Validations of IR and ES events were performed using primer pairs spanning the neighboring constitutive exons. Quantification of alternative sequence inclusion levels from gel band intensity was done using ImageJ software (84).

**Comparative transcriptomic cross-species clustering**

For the cross-species comparison, we first identified one-to-one orthologs in the proteomes of Creolimax, Capsaspora and Salpingoeca using the Multiparanoid pipeline (85). We trimmed all RNA-seq datasets into the same length (50 bp) and only mapped the left reads when paired-end data was available. We then obtained cRPKM (corrected by mappability) values only for the subset of orthologs, transformed them to log2(cRPKM +1) and further normalized the expression data using quantile normalization. Hierarchical clustering (‘complete’ method) of samples was obtained by comparing pairwise distances based on Spearman correlation coefficients in R. To obtain the neighbor-joining trees and bootstrap supports across the samples, we used the ‘ape’ package in R (86). For the Creolimax/Human comparison we followed the same methodology, using the data detailed in Figure 6-source data 1. The principal component analysis of the expression data was performed as implemented in R ‘prcomp’ function. Gene Ontology enrichments were obtained as described in RNA-seq and differential expression analysis section.
Secretome proteomics and in silico prediction

To obtain the secretome sample for proteomics, we cultured *Creolimax* cells in liquid medium (Marine Broth Difco 2216) at 12°C for 5 days and allowed the cells to attach to the bottom of the flask. The medium was then replaced with artificial seawater to avoid excessive protein contamination, and the culture was incubated for another 24 hours. Then, we collected the medium by gently tilting the flask to avoid collecting attached cells. The medium was immediately centrifuged at 10,000 rcf for 2 minutes, and the supernatant was collected and filtered twice through a 0.2 μm filter. The filtered medium was concentrated by ultrafiltration using a MWCO membrane (Vivaspin 6, 3000 MW; Sartorius, Gottingen, Germany) and quantified using the BCA Protein Quantification Kit (Thermo Fisher Scientific). The resulting protein extract was digested with 5 µg of trypsin (Promega, cat # V5113) (overnight, 37°C). Finally, 2 µg of the sample was analyzed using an LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) coupled to an EasyLC (Thermo Fisher Scientific (Proxeon), Odense, Denmark) at the CRG proteomics unit (Barcelona, Spain). All data were acquired with Xcalibur software v2.2. Proteome Discoverer software suite (v1.4, Thermo Fisher Scientific) and the Mascot search engine (v2.5, Matrix Science (87)) were used for peptide identification and quantification. The data were searched against a database containing *Creolimax* proteome, a list of common contaminants, and all the corresponding decoy entries. Resulting data files were filtered for FDR < 0.05. Finally we discarded the contaminants and all the proteins that were identified by less than 2 unique peptides, resulting in a list of 91 proteins.

To obtain the list of *in silico* secretome components, we performed an initial search step using SignalP 3.0 (88) (D-cutoff=0.450) to identify all the proteins with a canonical signal peptide. We then performed a search step with TMHMM v.2.0 (89) to discard all those proteins with a transmembrane domain downstream of the first 60 amino acids. We also filtered out proteins tagged to the mitochondria using TargetP v.1.1 (90), and proteins with endoplasmic reticulum retention signal with c-terminal motifs KDEL or HDEL and GPI-anchored proteins using PredGPI (91). The resulting list comprised 453 proteins. PFAM and Gene Ontology enrichment analyses were performed as described in *RNA-seq and differential expression analysis*.

Ortholog identification, phylogeny and LGT detection

In order to obtain orthology assignments for the various gene families analyzed in this study, we used a phylogeny-based pipeline. First, we used PFAM domain information to obtain all the members of a gene family across a database comprising 108 eukaryotic proteomes. Then, proteins...
were aligned using MAFFT software with L-INS-i parameters (92). The resulting alignments were automatically trimmed using trimAl v1.2 (93) (-gt 0.7) and phylogenies were obtained using RAxML v8.0 (94) (LG model, gamma distribution, 100 bootstrap supports) and Phylobayes 3 (95) (LG model, ran until two chains converged). Tree visualization and annotation was performed using iTol v2 (96).

This pipeline was slightly modified to detect LGT cases. Instead of using PFAM, we gathered close orthologs of all the proteins in the genome by performing a BLASTP search against the NCBI non-redundant database plus the 108 eukaryotic genomes. Only those proteins that retrieved lower e-values for bacterial/archaeal hits were selected for downstream analysis. Those proteins were then separately searched using BLASTP against all Bacteria in nr, all Archaea in nr and the 108 eukaryotic proteomes. We selected those sequences that had at least 25 hits with an e-value under e-10, selecting a maximum of 50 proteins for bacteria, 25 for archaea and eukaryotes. We got rid of redundancy using CD-HIT (97) by filtering for 0.95 identity and then performed alignment, trimming and phylogeny as described in the general pipeline above. The resulting trees were analyzed manually, taking LGT positives when Creolimax (and other ichthyosporean) sequences branched within bacterial clades with nodal bootstrap supports over 70%. Finally, we manually checked that the resulting LGT genes were found in distinct parts of the genome and not in genomic singletons. For those LGT candidates without introns and not found in any other Ichthyosporean (therefore, Creolimax specific) we further checked if they were located in scaffolds with other genes containing introns. To further discard bacterial contaminations, the neighboring genes to LGT candidates were blasted against NCBI nr to verify their eukaryotic origins. When the immediate neighbor did not retrieve any hit the following gene was searched.

**Acknowledgements**

We would like to thank the proteomics unit in the CRG for help in the analysis of the secretome, especially Eduard Sabidó and Guadalupe Iglesias, and the genomics unit of CRG for help in the RNA-seq library preparation and sequencing. We also thank Arnau Sebé-Pedrós, Xavi Grau-Bové and Ozren Bogdanovic for critical reading and discussion of the manuscript. Finally, we are grateful to Meritxell Antó for technical support provided throughout the project.
References


958 Through Phylogenomics and Relaxed Molecular Clock Analyses. Evolutionary Transitions

964 \textbf{FIGURES}

967 \textbf{Figure 1. Evolution of developmental and feeding modes across holozoans.} (A) The
968 cladogram represents known phylogenetic relationships among holozoans (12,14). Each lineage is
969 represented by the species proposed as a model system with a schema of its developmental mode
970 on the right. The evolution of specialized osmotrophy is shown as a blue triangle in the cladogram,
971 while the putative ancestral phagotrophic feeding mode of opisthokonta is shown as an orange
972 circle (98). Divergence times of the lineages shown in this figure range between 700 Mya
973 (considered the latest estimates of animal origins) and 1200 Mya (earliest estimates of
974 Opisthokont origins) (99). Micrographs depicting the (B) amoeboid stage and (C) multinucleate
975 stage of \textit{Creolimax fragrantissima} are shown. Scale bars = 10 μm. Choanoflagellate adapted from
976 Mark Dayel (CC BY-SA 3.0) www.dayel.com/blog/2010/10/07/ choanoflagellate-illustration.

978 \textbf{Figure 1-figure supplement 1.} \textit{Creolimax synchronized stages.} (A) Culture after 5-μm
979 filtering. (B) Culture grown for 24 hours after 5 μm filtering. Scale bars = 20 μm.
Figure 2. Differential gene expression in Creolimax. (A) Diagram of the amoeboid and multinucleate stages, and heatmap showing the significantly differentially expressed genes across biological replicates in the pair-wise stage comparison. (B) Gene set enrichment analysis for the two stages. Orange represents enrichment in the amoeboid stage and blue represents enrichment in the multinucleate stage, color intensity depicts level of significance ($P$ value). Node size represents the total number of genes in each GO, and edge width represents the total number of genes shared between each enriched GO category. Functionally related GOs are manually circled in grey shades according to functional and genic redundancy established by network connectivity. Complete list of GOs and inclusive groupings are found in Figure 2-source data 1.

Figure 2-source data 1. GOs enrichments and groupings from Figure 2 B.

Figure 2-figure supplement 1. Pfam domain set enrichments in differentially expressed genes. Orange represents domains enriched in the amoeboid stage and blue represents domains enriched in the multinucleate stage, color intensity depicts level of significance ($P$ value, Fisher’s exact test). Node size represents the total number of genes containing a Pfam domain, and edge width represents the total number of genes sharing two distinct Pfam domains. Functionally related Pfams are manually circled in grey shades primary based on the information gathered in the Pfam database for each domain (including Pfam2Go annotations). Additional criteria to include a given domain to a functionally related category included: checking the list of GOs of the statistically differentially expressed domain-containing genes in a given stage and using a network connectivity redundancy between GO and Pfam categories in a mixed network (including both Pfam and GO annotation) done with Enrichment Map plugin in Cytoscape (79).

Figure 2-figure supplement 2. Differential TFome expression. (A) Distribution of expression values (FPKM) for all transcription factors (TFs) in the genome in the amoeboid and the multinucleate stages ($P$ value, Wilcoxon signed rank test). (B) Heatmap showing the expression levels of all the TFs with a two-fold change in expression level between stages. Those with the gene id in bold have statistically significant differential expression according to at least 3 different differential expression pipelines (see Material and Methods). To the right, the domain
Architectures of the TFs, showing the DNA binding domain in dark blue and orange according to the stage where they are up-regulated. In the multinucleate stage: CSD (PF00313), Myb (PF00249), CBFB_NFYA (PF02045) and TPR (PF13414). In the amoeboid stage: bZIP_1 (PF00170), HLH (PF00010), T-box (PF00907), RFX_DNA_binding (PF02257), Runt (PF00853).

Figure 3. Regulated alternative splicing modes in Creolimax. (A) Heatmap showing Percent Intron Retention (PIR) inclusion levels of differentially retained introns. (B) Heatmap showing the percent spliced in (PSI) levels of differentially skipped exons. (C) GO enrichment activities of the genes showing differential intron retention (IR). Bar length indicates the significance of the enrichment, orange indicates those with higher inclusion levels in the amoeboid stage and blue those with higher inclusion levels in the multinucleate stage. (D-E) RT-PCR validations of selected intron retention and exon skipping events. The values correspond to relative intensity of the alternative isoform (retained intron or skipped exon) bands in the RT-PCR and the proportions observed for the inclusion values in the RNA-seq. (F) GOs enrichment of genes with differential exon skipping (ES), in blue those with higher exon inclusion levels in the multinucleate stage.

Figure 3-figure supplement 1. Intron size and transcriptional levels of differentially retained introns. (A) Relationship between intron length and retention level. The barplot shows the percentage of retained introns (PSI > 20) among the total number introns of a given size. (B) Cumulative frequency plot showing that the intron size distribution is the same for constitutive introns (PSI<2), highly retained introns (PSI > 50 both stages) and differentially retained introns. (C) Distribution of expression values (FPKM) in the amoeboid and the multinucleate stage for genes differentially retained introns.
Figure 3-figure supplement 2. Validations of intron retention and exon skipping events. The RT-PCR gels show the different splice variants for each gene in the amoeboid stage (left) and multinucleate stage (right). RT-PCR values indicate the levels of inclusion of the alternative isoform (retained intron or skipped exon) compared to the canonically spliced form obtained with ImageJ; RNA-seq values are based on read coverage for each event. The scatter plot shows the differences between the RT-PCR measures and the RNA-seq-based values for all the validated examples, overall showing a high correlation.

Figure 4. Transcriptional and post-transcriptional regulation of lincRNAs in Creolimax. (A) Heatmap showing transcriptional levels of significantly differentially expressed lincRNAs across biological replicates of amoeboid and multinucleate stages. (B) Example of genomic region where two lincRNA loci are found in tail-to-tail orientation surrounded by two protein-coding genes. (C) RT-PCR validations of the lincRNA loci. (D) Barplot depicting the average gene expression of those lincRNA in each stage. (E) Alternative splicing isoforms of lincRNAs showing various degrees of intron retention.

Figure 4-figure supplement 1. Genomic architecture of lincRNAs compared to protein-coding genes. (A) Kernel density plot showing transcript length distribution. Protein coding genes are shown in blue, lincRNA in red. (B) Density plot showing exon number distribution. (C) Kernel density plot showing exon length distribution; multi-exonic lincRNAs are shown in green. (D) Kernel density plot showing GC content distribution. The dashed line indicates the total genome GC content. (E) Cumulative frequency plot of expression levels obtained from log10 (FPKM).
**Figure 4-figure supplement 2.** Gene orientation and transcriptional co-regulation of neighboring genes. Distribution of Pearson correlation values between a gene and its upstream neighbor subdivided in 4 categories. Head-to-head oriented neighbors tend to be more co-expressed than head-to-tail genes, independently of coding potential.

**Figure 5. Holozoan cross-species comparison of transcriptional profiles.** (A) Symmetrical heatmap of the pair-wise Spearman correlation coefficients for the gene expression profiles of each cell stage. For each sample, log2(cRPKM+1) expression levels were obtained for 2177 one-to-one orthologs in the three species analyzed (see Materials and Methods). Dashed-line squares highlight the direct comparisons for 1) Cfra multinucleate stage replicates against Cowc Cystic stage replicates and 2) Cfra amoeboid stage replicates against Cowc aggregate and adherent stage replicates. (B) Neighbor-joining tree of the species cell stages based on the aforementioned Spearman correlation distances matrix. Filled circles represent > 95% bootstrap replicate nodal support. (C) The cell types plotted according to the values of the principal components 2 and 3 from a principal component analysis of a dataset of 3030 1-to-1 orthologs between *Capsaspora* and *Creolimax*. (D) The significant GO enrichments for the top positive loading genes (> 0.03) of the principal component 2 and 3. Cfra – *Creolimax fragrantissima*, Cowc – *Capsaspora owczarzaki*, Sros – *Salpingoeca rosetta*.

**Figure 6. Comparison of human and Creolimax cell-types and tissues.** (A) Symmetrical heatmap of the pair-wise Spearman correlation coefficients for the gene expression profiles of each cell type or tissue. For each sample log2(FPKM+1) expression levels were obtained for 2272 one-to-one orthologs between *Creolimax* and human (see Materials and Methods). (B) The human cell types sorted by the difference of Spearman correlation between the amoeboid and the multinucleate cell stages. Highlighted in grey those that displayed the major differences (> 0.05). (C) The cell types plotted according to values of the principal components 1 and 2 from a principal component analysis of the same transcriptional dataset of 2272 orthologs. The dots in grey are the human cell lines highlighted in the previous section. (D) The significant GO enrichments for the top positive loading genes of the principal component 2 (> 0.04). Sampled human cell types described in Figure 6-source data 1.
Figure 6-source data 1. Human RNA-seq datasets used in this analysis.

Figure 7. Co-regulation of the integrin adhesome in holozoans. Heatmaps depicting expression levels of integrin adhesome orthologs (red-green) and their pair-wise Pearson correlation coefficients (white-blue) obtained from genome-wide FPKM transcriptional levels in the ichthyosporean Creolimax (A), the filasterean Capsaspora (B) and the choanoflagellate Salpingoeca (C). (D) Diagram of integrin adhesome components, those in green are found in C. fragrantiissima and those in white are absent. In grey, a tyrosine kinase receptor with extracellular EGF domains encoded Creolimax genome that could be interacting with an ECM component. (E) Repertoire of animal ECM domains in the three unicellular holozoan genomes; green = presence, white = absence. (F) Pfam domain architectures of Fibronectin-domain containing genes in Creolimax.

Figure 7-figure supplement 1. Co-regulation of the filopodial molecular toolkit genes in holozoans. (A) Diagram of the filopodial machinery components as described in Sebé-Pedrós et al 2013 (43); those in green are the genes found in unicellular holozoans. Heatmaps depicting expression levels of filopodial component orthologs (red-green) and their pair-wise Pearson correlation coefficients (white-blue) obtained from genome-wide FPKM transcriptional levels in the ichthyosporean Creolimax (B), the filasterean Capsaspora (C) and the choanoflagellate Salpingoeca (D). Highlighted in red are those genes that appear outside the module in at least two lineages.

Figure 7-figure supplement 2. Co-regulation of the pre- and post- synaptic genes in holozoans. (A) Diagram of the pre-synaptic (paler green) and post-synaptic (darker green) found in unicellular holozoans. Heatmaps depicting expression levels of pre- and post- synaptic orthologs (red-green) and their pair-wise Pearson correlation coefficients (white-blue) obtained from genome-wide FPKM transcriptional levels from the ichthyosporean Creolimax (B), the filasterean Capsaspora (C) and the choanoflagellate Salpingoeca (D). Flotillin genes, despite not being directly related to the post-synaptic scaffold, have been shown to interact with Homer in choanoflagellates and animals (45).

Figure 8. Functional enrichments of Creolimax secretome. (A) Venn diagram showing the number of genes identified in the Creolimax secretome through an in silico approach (see
Material and Methods) and an in vivo proteomics approach. A circle diagram describes the features of genes only identified in the in vivo approach, lacking a signalP or having TM domains. (B) GO categories and (C) PFAM domains enriched in the secretome; in dark blue those enriched in the in vivo dataset, in pale blue those enriched in the in silico dataset, in grey the total amount of PFAM-domain containing genes in the genome.

**Figure 8-source data 1. In vivo proteomics of Creolimax secretome.**

**Figure 9. Trypsin evolution.** (A) Barplot showing the total number of Trypsin proteins (PF00089) found in the genomes of diverse eukaryotes. Branches are color-coded according to the taxonomy shown in the legend. (B) Maximum-likelihood phylogenetic tree based on the amino acid sequence of the Trypsin domain from *Creolimax* and *Sphaeroforma arctica*. Expression levels obtained from genome-wide FPKM calculation are shown. Number of unique peptides obtained from the in vivo secretome proteomic dataset are also shown. In red are those genes that do not present a signal peptide according to SignalP.

**Figure 10. CotH evolution.** (A) Maximum-likelihood phylogenetic tree of the CotH domain (PF08757). Nodal support is shown in key branches (100 maximum likelihood replicates bootstrap values and Bayesian Posterior Probabilities). Color code indicates taxon distribution in each clade as depicted in the legend; for a detailed tree, see Figure 9-figure supplement 1. (B) Detail of the phylogenetic tree depicting ichthyosporean CotH sequences, covering *Creolimax, Sphaeroforma* and *Amoebidium parasiticum*. Expression levels obtained from genome-wide FPKM calculation and the number of unique peptides obtained from the in vivo secretome proteomic dataset are shown. Domain configurations obtained from a PfamScan analysis. Gene identifiers in red are those that do not present a signal peptide according to SignalP.

**Figure 10-figure supplement 1. CotH extended phylogeny.** Maximum-likelihood phylogenetic tree of the CotH domain (PF08757). Nodal support is shown in key branches (100 maximum likelihood replicates bootstrap values and Bayesian Posterior Probabilities). Color code indicates taxon distribution in each clade as depicted in the legend. Domain configurations obtained from a PfamScan analysis.
Figure 10-figure supplement 2. Features of prokaryotic lateral gene transfer. (A) Phylostratigraphy depicting origins of LGT events found in Creolimax. Each node represents the total number of horizontally acquired genes in Creolimax found in any of the remaining ichthyosporean transcriptomes/genomes. Phylogenetic relationships among ichthyosporeans and out-groups obtained from (14). (B) Boxplot showing intron number distribution according to LGT phylostratigraphic age. (C) Venn diagram of horizontally acquired genes found in the in silico and the in vivo secretome datasets.
Animals

Salpingoea rosetta

Capsaspora owczarzaki

Creolimax fragrantissima

Developmental mode

clonal

aggregative

coenocyte

Phagotrophy

Osmotrophy

Holozoa

Animals

Choanoflagellates

Salpingoea rosetta

Filastereans

Capsaspora owczarzaki

Ichthyosporeans

Creolimax fragrantissima

Fungi

Other eukaryotes
Protein Kinase Activity
Actin cytoskeleton
Cell-ECM adhesion
GTPase activity
G protein signaling
Peptidase activity
Integrin-Adhesion
MAPK kinase
Ribosome biogenesis
Mitochondrial targeting
RNA metabolism
tRNA biogenesis
Translational initiation
Amino acid metabolism
Methyltransferase activity

Differentially expressed genes

<table>
<thead>
<tr>
<th>Amoeboid_1</th>
<th>Amoeboid_2</th>
<th>Amoeboid_3</th>
<th>Multinucleate_1</th>
<th>Multinucleate_2</th>
<th>Multinucleate_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=490</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B Gene Set Enrichment Map: Amoeboid vs. Multinucleate

Protein Kinase Activity
Actin cytoskeleton
Cell-ECM adhesion
GTPase activity
Integrin-Adhesion
MAPK kinase
Mitochondrial targeting
RNA metabolism
tRNA biogenesis
Translational initiation
Amino acid metabolism
DNA replication
Methyltransferase activity
G protein signaling
Ribosome biogenesis

Enrichment amoeboid
Enrichment multinucleate
Multinucleate Amoeboid Creolimax stages Human cell-types and tissues

Spearman correlation

0.4 0.8

Multinucleate vs Amoeboid

PC1 (61.6%)

GO enrichment from top 125 top loading genes of PC2

DNA-duplex unwinding
DNA geometric change
DNA replication initiation
regulation of DNA replication
spindle
replication fork
base-excision repair
chromosome condensation
A. Creolimax fragrantissima
B. Capsaspora owczarzaki
C. Salpingoeca rosetta

D. ECM components
E. Animals
F. Choanoflagellates

- Laminin
- Integrin α
- Integrin β
- FAK
- Act
- Tyre
- Csk
- Paxillin

1. CFRG7187
2. CFRG7768

RNAseq FPKM
Creolimax secretome

A

B

GO enrichments
- serine-type endopeptidase activity
- serine-type peptidase activity
- endopeptidase activity
- collagen catabolic process
- extracellular space
- metalloexopeptidase activity
- extracellular matrix organization
- glycogen metabolic process
- triglyceride lipase activity
- lipase activity
- serine-type endopeptidase activity
- endopeptidase activity
- carboxylic ester hydrolase activity
- serine-type peptidase activity
- extracellular region
- nutrient reservoir activity
- proteolytic activity
- cell redox homeostasis
- proteolysis
- hydrolase activity
- hydrolyase activity
- glycerol ether metabolic process
- beta-N-acetylhexosaminidase activity
- oxidoreductase activity
- metalloexopeptidase activity

C

PFAM domain enrichments
- Trypsin
- Peptidase_M14
- CarboxypepD_reg
- Peptidase_C1
- Peptidase_S28
- Inhibitor_I29
- Transferrin
- CotH
- LTD
- Alk_phosphatase
- Lipase_3
- Kazal_1
- Kazal_2
- Pyr_redox_dim
- TSP_1
- SapA
- Cupin_1
- Ricin_B_lectin
- CBM_2
- Glyco_hydro_20
- Glyco_hydro_20b
- Glyco_transf_8
- Peroxidase
- Thioredoxin
- Thioredoxin_6
A. Ichthyosporean Trypsins

- Domain containing genes

B. Trypsin distribution in eukaryotes

- Taxonomy
  - Metazoa
  - Fungi
  - Unicellular holozoa
  - Rozella+Microsporidia
  - Discocystidiales
  - Apusozoa
  - Amoebozoa
  - Embryophyta
  - Chlorophyta
  - Rhodophyta
  - Glaucochyta
  - Heterokonta
  - Alveolata
  - Rhizaria
  - Haptophyta
  - Cryptophyta
  - Excavata

- Unique peptides

- FPKM

- 0 50 100 1500 3000 4500

- 0.1