Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages

Alex de Mendoza[a,b,1], Arnau Sebé-Pedrós[a,b,1], Martin Sebastijan Šestak[c], Marija Matejčić[c], Guifré Torruella[a,b], Tomislav Domazet-Lošo[c,d], and Iñaki Ruiz-Trillo[a,b,e,2]

[a]Institut de Biologia Evolutiva (Consejo Superior de Investigaciones Científicas–Universitat Pompeu Fabra), 08003 Barcelona, Spain; [b]Departament de Genètica, Universitat de Barcelona, 08028 Barcelona, Spain; [c]Laboratory of Evolutionary Genetics, Ruđer Bošković Institute, HR-10000 Zagreb, Croatia; [d]Catholic University of Croatia, HR-10000 Zagreb, Croatia; and [e]Institució Catalana de Recerca i Estudis Avançats, 08010 Barcelona, Spain

Edited by Walter J. Gehring, University of Basel, Basel, Switzerland, and approved October 31, 2013 (received for review June 25, 2013)

Transcription factors (TFs) are the main players in transcriptional regulation in eukaryotes. However, it remains unclear what role TFs played in the origin of all of the different eukaryotic multicellular lineages. In this paper, we explore how the origin of TF repertoires shaped eukaryotic evolution and, in particular, their role into the emergence of multicellular lineages. We traced the origin and expansion of all known TFs through the eukaryotic tree of life, using the broadest possible taxon sampling and an updated phylogenetic background. Our results show that the most complex multicellular lineages (i.e., those with embryonic development, Metazoa and Embryophyta) have the most complex TF repertoires, and that these repertoires were assembled in a stepwise manner. We also show that a significant part of the metazoan and embryophyte TF toolkits evolved earlier, in their respective unicellular ancestors. To gain insights into the role of TFs in the development of both embryophytes and metazoans, we analyzed TF expression patterns throughout their ontogeny. The expression patterns observed in both groups recapitulate those of the whole transcriptome, but reveal some important differences. Our comparative genomics and expression data reshape our view on how TFs contributed to eukaryotic evolution and reveal the importance of TFs to the origins of multicellularity and embryonic development.

phylotypic stage | Holozoa | LECA

Transcription factors (TFs) are proteins that bind to DNA in a sequence-specific manner (1) and enhance or repress gene expression (2–4). In response to a broad range of stimuli, TFs coordinate many important biological processes, from cell cycle progression and physiological responses, to cell differentiation and development (5, 6). Thus, TFs have a central role in the transcriptional regulation of all cellular organisms, being present in all branches of the tree of life (bacteria, archaea, and eukaryotes). There appears to be a correlation between elaborate regulation of gene expression and the complexity of organisms (7), such that the amount (as a proportion of an organism's total gene content) and diversity of TF proteins is expected to be directly correlated with this complexity (8). Indeed, TFs play a crucial role in multicellular eukaryotes. For example, TFs are the master regulators of embryonic development in embryophytes and metazoans (9), and analyses of their embryonic transcriptional profiles support the presence of a phylotypic stage in both lineages (10–14). These studies have also shown that evolutionarily younger genes tend to be expressed at earlier and later stages of development, whereas the transcriptomes of the middle stages (the phylotypic stage) are dominated by ancient genes (10, 13). It remains to be investigated how the evolutionary age and the expression patterns of the different TFs shift throughout the ontogeny of these lineages and whether TF expression profiles correlate with the general transcriptome profiles.

Previous studies have analyzed the evolutionary history and phylogenetic distribution of TFs in various branches of the tree of life (6, 15–22). However, it is not yet clear whether the evolutionary scenarios previously proposed are robust to the incorporation of genome data from key phylogenetic taxa that were previously unavailable.

In this paper, we present an updated analysis of TF diversity and evolution in different eukaryote supergroups, focusing on various unicellular-to-multicellular transitions. We report genome and/or transcriptome data from several unicellular relatives of Metazoa and Fungi (including one filasterean, five ichthyosporeans, a nucleariid, and the incertae sedis *Corallochytrium limacisporum*), and use published data from key, but previously unsampled, eukaryotic lineages, such as Glaucophyta, Haptophyta, Rhizaria, and Cryptophyta. We show that an important fraction of the metazoan TF toolkit is not novel but rather appeared in the root of Opisthokonta and/or Holozoa. Similarly, we show that many plant TFs are present in unicellular chlorophytes, and many fungal TFs are present in microsporidians and nucleariids. Finally, we analyze the TFome (i.e., the general TF repertoire) expression throughout embryonic development in embryophytes and metazoans and show that each phylostratigraphic layer of TFs differentially contributes to successive stages of embryonic development, linking the evolutionary history of TFs to organismal ontogeny.

Significance

Independent transitions to multicellularity in eukaryotes involved the evolution of complex transcriptional regulation toolkits to control cell differentiation. By using comparative genomics, we show that plants and animals required richer transcriptional machineries compared with other eukaryotic multicellular lineages. We suggest this is due to their orchestrated embryonic development. Moreover, our analysis of transcription factor (TF) expression patterns during the development of animals reveal links between TF evolution, species ontogeny, and the phylotypic stage.

Author contributions: A.d.M., A.S.-P., T.D.-L., and I.R.-T. designed research; A.d.M., A.S.-P., M.S.Š., and M.M. performed research; G.T. contributed new reagents/analytic tools; A.d.M., A.S.-P., M.S.Š., M.M., T.D.-L., and I.R.-T. analyzed data; and A.d.M., A.S.-P., T.D.-L., and I.R.-T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited with the National Center for Biotechnology Information [NCBI BioProject number PRJNA189477 (*Amoebididum parasiticum*); and NCBI SRA projects SRX096927 and SRX096925 (*Ministeria vibrans*); SRX377507 (*Pirum gemmata*); SRX377508 (*Abeoforma whisleri*); and SRX377514 (*Creolimax fragrantissima*)].

[1]A.d.M. and A.S.-P. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: inaki.ruiz@multicellgenome.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1311818110/-/DCSupplemental.

Results

Lineage-Specific and Paneukaryotic TFs. DNA-binding domains (DBDs) are univocal signatures of the presence of a particular type of TF (8, 16). Therefore, we analyzed DBDs across eukaryotes to survey the presence, abundance, and relative contribution of each TF class (71 in total) throughout various eukaryotic lineages (Fig. 1). It should be noted that we used a conservative method that tends to underestimate total counts, but minimizes false positives (*Methods*). We used a wide taxon sampling strategy optimized to have the largest possible diversity, especially around multicellular transitions.

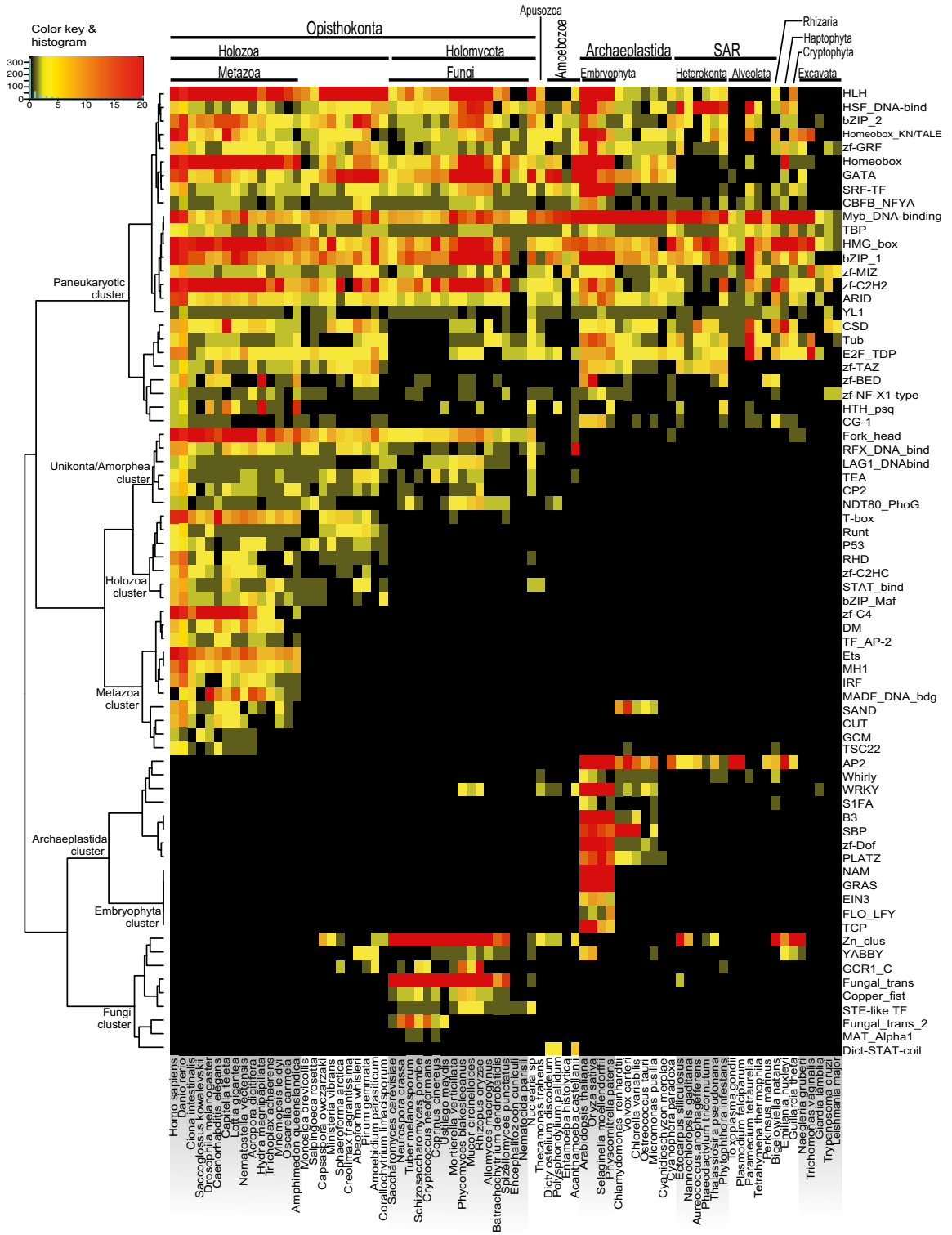

Fig. 1. Presence and abundance of transcription factors (TFs) in eukaryotes. The heat map depicts absolute TF counts according to the color scale. TFs/DBDs (rows) are clustered according to abundance and distribution, and species (columns) are grouped according to phylogenetic affinity. Major eukaryotic lineages are indicated (*Top*). Raw data are in Dataset S1. Further taxonomic information is in Dataset S2.

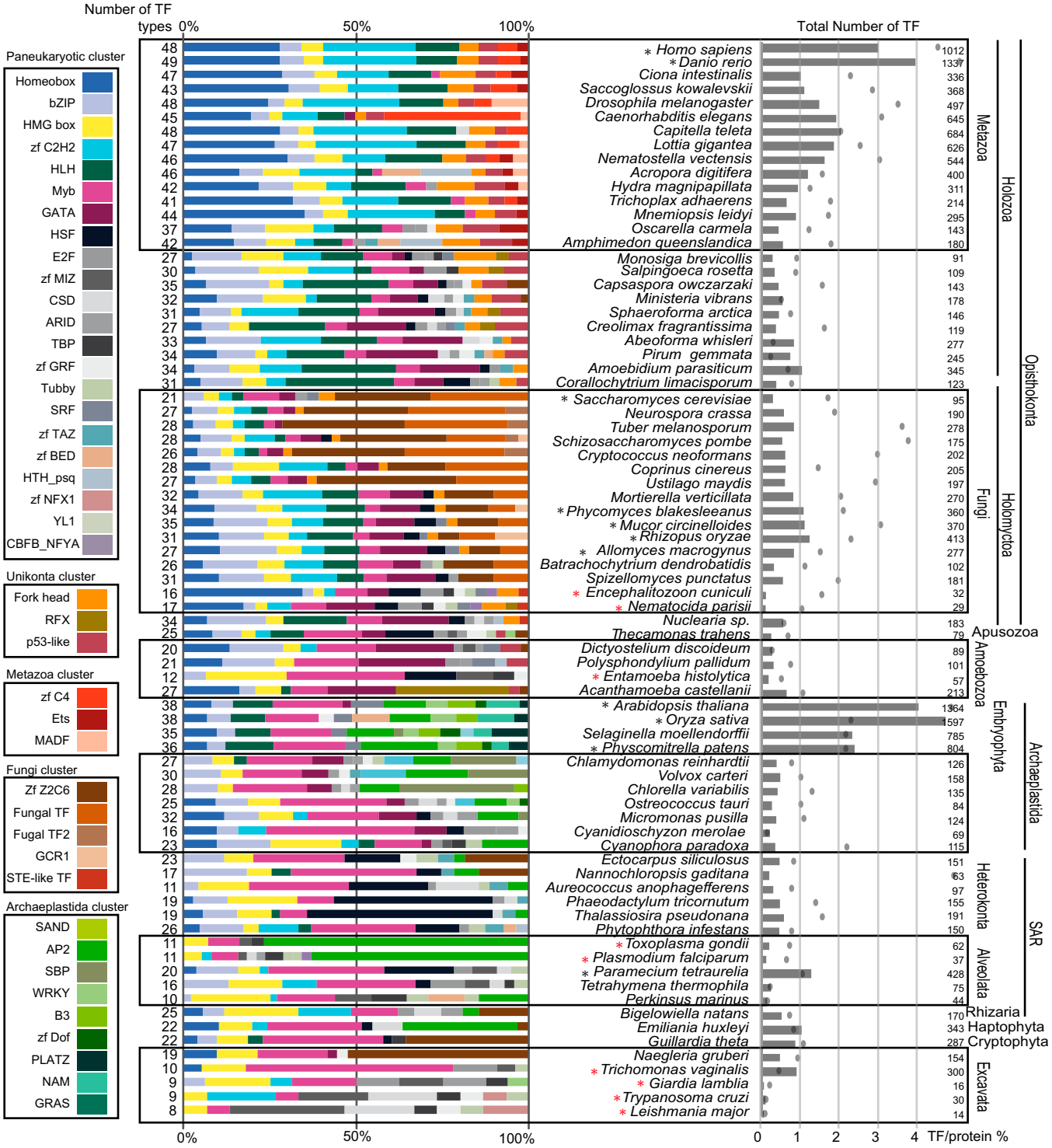

Fig. 2. Phylogenetic patterns of TFome composition across eukaryotes. For the sake of clarity and simplicity, the TF types that represent less than 2% of the corresponding TFome are not considered. For the same reason, some TF types are summarized in higher-level categories according to structural similarities. This is the case of the (*i*) Homeobox supergroup, which comprise Homeobox and Homeobox_KN/TALE; (*ii*) the bZIP supergroup, which comprise bZIP_1, bZIP_2, and bZIP_Maf; and (*iii*) the p53-like supergroup, which comprise p53, STAT, Runx, NDT80, LAG1, and RHD. To the *Left*, the total number of TF types present in each taxa and the relative abundance of each DBD type (reduced as specified in *Methods*, *Ancestral Genome Reconstruction and Enrichment*) in the TFome of every species are depicted using the color code in the legend of DBDs. To the *Right*, a bar graph indicates the total number of TFs in each species, and dots indicate the percentage of total TFs/total number of proteins. The black asterisks indicate species with WGDs. The red asterisks indicate strict parasites. Raw data are in Dataset S1. Further taxonomic information is in Dataset S2.

Our data divide eukaryotic TFs into two main groups according to their phylogenetic distribution, paneukaryotic and lineage-specific TFs (Fig. 1). Paneukaryotic TFs are widely distributed and were already present in the last common ancestor of eukaryotes (LECA). These paneukaryotic TFs include HLH, GATA, SRF-TF, bZIP, Homeobox, HMGbox, and zf-C2H2 (Fig. 1). Although common to most eukaryotes, the abundance of paneukaryotic TFs tends to be highly variable, with independent expansions in different lineages. A good example is the expansion of homeobox TFs in both metazoans and embryophytes (Fig. 1). A few paneukaryotic TFs, such as CFBF_NFYA, YL1, or TBP, are less prone to diversification and are often lost secondarily (Fig. 1). Regarding the lineage-specific TFs, we could define six taxonomically restricted clusters of TFs: (*i*) TFs exclusive or mostly present in unikonts; (*ii*) TFs that are specific to, or very uncommon outside, holozoans; (*iii*) metazoan-specific TFs; (*iv*) TFs mostly present in Archaeplastida; (*v*) Embryophyta-specific TFs; and (*vi*) fungal-specific TFs or TFs that are predominant in fungal taxa (Fig. 1).

Protein domain architecture provides an additional layer of TFome complexity (8, 17). For example, it is known that, during their evolution, homeobox TFs acquired extra domains that provided new binding targets and specificity (23). Therefore, we analyzed the complexity of the gene architecture of TFs across eukaryotes. Our data show that the complexity of the protein domain architecture of some paneukaryotic TFs, including Myb, zf-C2H2, Homeobox, and bHLH, is significantly enriched in both metazoans and embryophytes (Figs. S1 and S2).

Phylogenetic Patterns of TF Numbers. The total number of TFs varies markedly between taxa (Fig. 1), raising the possibility that these variations are correlated with some specific features, such as organismal complexity. Our data show that complex developing multicellular taxa (i.e., Embryophyta and Metazoa) present a dramatic increase in TF numbers compared with other eukaryotes. Moreover, morphologically simpler forms within these lineages, such as sponges within metazoans, and mosses within embryophytes, have less TFs than morphologically more complex groups (Fig. 2). In contrast, irrespective of their phylogenetic position, parasitic eukaryotes have significantly fewer TFs (Fig. 2). Notably, some species described as parasitic or endosymbiotic, such as the different Ichthyosporea or *Capsaspora owczarzaki* (24, 25), have relatively rich TF repertoires, perhaps revealing a more complex life cycle or an undescribed free-living stage. Finally, whole-genome duplication (WGD) also partly explains the existence of some particularly rich TF repertoires in groups such as the vertebrates and the embryophytes, or *Allomyces macrogynus*, the Mucoromycotina, and *Paramecium* (Fig. 2) (26, 27). Species with or without WGD that branch within the same group, such as the deuterostomes *Homo sapiens* and *Ciona intestinalis* or the fungi *Allomyces macrogynus* and the chytrids, show similar TFome profiles in terms of the proportions of each TF class, suggesting that this tendency is independent of TF type. This is consistent with the finding that TFs are one of the gene classes that are most resilient to loss after WGD (28, 29). As for the total number of TFs, the proportion of TFs in a genome (as a fraction of the total number of proteins) also varies considerably. This proportion is high in Embryophyta and to a lesser extent in Metazoa, similar to the pattern observed for the total number of TFs. In contrast, Fungi, which have low numbers of TFs, have a high proportion of TFs in their genomes. Notably, these measures are strongly affected by the quality of genome sequencing and annotation, which differs markedly between the taxa studied, and which can result in underestimation or overestimation of the total number of genes in a genome.

TFome Profiles in Eukaryotic Groups. We next evaluated the contribution of different TF types in each organism's TFome, in terms of the abundance of each TF type as a proportion of the total number of TFs in that particular genome (the TFome profile) (Fig. 2). Our analysis shows a clear phylogenetic pattern of TF diversity that is recovered by clustering based on TF content, with a few exceptions (Fig. S3).

Metazoans have a very distinctive TFome profile, with Homeobox, zf-C2H2, and bHLH representing the largest fraction, especially in Bilateria, in which these three TF classes represent more than 50% of the total number of TFs (Fig. 2). These are the same TFs that are significantly enriched in protein domain architecture in metazoans (Fig. S2). As suggested elsewhere, the main role of these types of TF is in patterning and cell type differentiation (19, 20). Conversely, p53-like TFs (including p53, runx, T-box, NDT80, and STAT) and bZIP seem to represent a larger fraction of the TFome in early-branching metazoans, but a minor fraction in bilaterians. Sponges, which are possibly the earliest branching metazoans (30–33), have some rather uncommon TFs expanded in comparison with other metazoans. For example HTH_psq represents more than 9% of the TFome of *Amphimedon queenslandica* (Fig. 2). In any case, most metazoans have similar TFome profiles, with a few DBDs representing a large percentage of their TF diversity, with some exceptions such as zf-C4 (hormone receptors) in *Caenorhabditis elegans* or MADF in *Drosophila melanogaster* (Fig. 2).

The TFome profiles of unicellular holozoans such as choanoflagellates, filastereans, ichthyosporeans and *Corallochytrium limacisporum* are different to those of metazoans and fungi, but similar to each other, with a small deviation of ichthyosporeans (from *Sphaeroforma arctica* to *Amoebidium parasiticum* in Fig. 2), which have a higher proportion of GATA and zf-TAZ than others. Despite having the smallest TF repertoire among of all holozoans, due mainly to the reduction or loss of some TF classes (19), choanoflagellates have a very similar TFome profile compared with other holozoans.

Fungi have a very particular TFome profile including some fungus-specific TFs (Fig. 2), some of which were further expanded in Dikarya. For example, this is the case in ZnClus (also known as zf-Z2-C6) and Fungal_transcription_factor, which become more abundant during fungal evolution (Fig. 1). In parallel, Dikarya presents important TF losses (Fig. 3), such as that of E2F-TDP, zf-TAZ, CSD, and Tub, and represents an interesting case of simplification and divergence. Indeed, most early-branching Fungi, like chytrids, blastocladiomycetes, and zygomycetes, show distinct TF diversity, in some cases resembling that of the unicellular holozoans. Microsporidians show a drastic reduction in their TF repertoire but conserve STE-like, one of the fungus-specific TFs (Fig. 1). *Nuclearia* sp., the unicellular sister group to Fungi (24), have typical fungus TFs such as STE-like and fungal transcription factor 1. Interestingly, *Nuclearia* sp. also has STAT and NF-κB, both secondarily lost in fungi and previously thought to be exclusive to holozoans (19).

A common feature of all opisthokonts is the relative abundance of Forkhead TFs compared with other eukaryotes. In contrast, the four amoebozoans analyzed have quite different TFome profiles, although this is not unexpected because genome data from Amoebozoa are still scarce. *Entamoeba histolytica* is a strict parasite and therefore most likely derived, whereas *A. castellanii*, the only nonprotostelid amoebozoan sampled to date, shows a striking increase in the number of RFX TFs, and it also has Forkhead and TEA/Sd TFs (34, 35).

The observed TF content of the group Archaeplastida, which comprises Embryophyta and their unicellular relatives (Chlorophyta, Rhodophyta, and Glaucophyta), is consistent with previous analyses (17). Our data show that Viridiplantae (Embryophyta plus Chlorophyta) share a unique set of TFs, which contrasts with the TFome profile of the glaucophyte *Cyanophora paradoxa* (36), whereas the rhodophyte *Cyanidioschyzon merolae* contains only one of those Viridiplantae-specific TFs (PLATZ). Similar to what is observed in metazoans and fungi, embryophytes are particularly

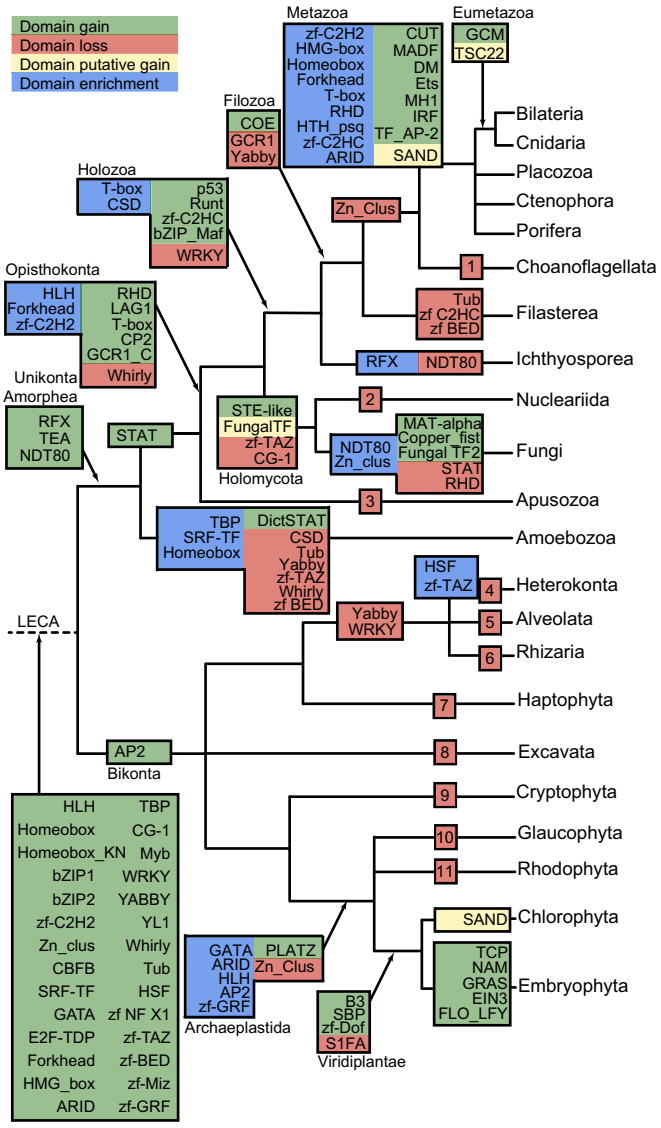

Fig. 3. TFs gains/losses and quantitative enrichments in eukaryote evolution, based on our 77-species taxonomic sampling. The reconstruction of evolutionary gains/losses of given TFs types in each node and ancestral states has been inferred using Dollo parsimony. The phylogenetic framework was taken from up-to-date eukaryote phylogenomic studies (71–75). DBDs gains are shown in green and lineage-specific DBDs secondary losses are shown in red boxes. See Table S1 for complete lists of TFs indicated as numbers. DBDs of unclear origin are shown in yellow boxes (e.g., also present in very distant, unrelated, species). Quantitative enrichments DBDs are shown in blue boxes and were obtained using Wilcoxon rank-sum tests, with *P* value threshold of <0.01.

enriched in some TF families, especially SBP, AP2, B3, GRAS, zf-Dof, and SRF/MADS. This supports the finding that the TF profiles of independent multicellular lineages are a combination of enriching particular paneukaryotic TF families and evolving new TF families.

Despite the poor taxon sampling and large evolutionary distances between the heterokont species analyzed here, there is a phylogenetic pattern in their TF profiles. The heat shock factors (HSFs) and the zf-TAZ are significantly enriched in all heterokont genomes (Fig. 3). The multicellular brown algae *Ectocarpus silicosus* (37) and its unicellular relative *Nanochloropsis gaditana* (38) are enriched not only for HSF but also for Zn_cluster TFs, which are also abundant in fungi (see above). Although brown algae have complex multicellularity, the total number of their TFs is not significantly increased, in contrast to what is found in metazoans and embryophytes. This may be explained by the presence of heterokont or brown-algae specific TFs that have not yet been described due to the paucity of functional studies in this group (39, 40). Alternatively, the fact that *Ectocarpus siliculosus* has modular growth instead of stereotypical embryonic development (39) may account for the lack of complexity in their TFome. Data from multicellular brown algae with embryonic development, such as *Fucus spiralis* (41), will be key to answering this question.

Reconstruction of TF Gain/Loss and Expansion Across Eukaryotes. We reconstructed the evolutionary history of the TFs in eukaryotes by mapping gains and losses using Dollo parsimony. We used Wilcoxon rank-sum tests to detect significant lineage-specific enrichments (Fig. 3). In terms of new gene gains, a strikingly similar pattern is found in embryophytes and metazoans, and to a lesser extent in Fungi. The emergence of these lineages with complex multicellularity is thought to be linked to a burst in their respective multicellularity toolkits, including TFs (42, 43). Our data, however, suggest an increase in TFome complexity in two steps. First, some innovation already took place in the ancestors of complex multicellular organisms given that their closest extant unicellular relatives (i.e., in holozoans and in the branch leading to Embryophyta plus Chlorophyta) already have complex TFomes. A second innovation step took place at the origin of metazoans and embryophytes. Similarly, the enrichment of some TF families follows the same two-step process, such as the enrichment of T-box and CSD TFs in holozoans. Later on, at the origin of Metazoa, Forkhead, RHD, homeobox, and T-box TFs were significantly enriched. However, GATA, ARID, HLH, AP2, and zf-GRF were enriched in Archaeplastida, whereas SRF-TFs, homeobox, and TBPs were enriched at the onset of Embryophyta. This means that a large percentage of the metazoan and embryophyte regulatory toolkits were already in place before their origin and divergence. However, TFs that are enriched in metazoans fall into ancient TF classes, whereas new domains account for more than 50% of TF diversity in Archaeplastida. The Fungi also gained several specific TF families, such as STE-like or Copper-fist TFs, although some such as STAT or zf-TAZ were also lost, in contrast to what is observed in metazoans and embryophytes. Moreover, Zn_clust and NDT80 were significantly enriched in Fungi (Fig. 3). Finally, the only significant case of depletion observed is that of Myb TFs at the stem of Opisthokonta plus Apusozoa. It is worth mentioning that this reconstruction remains tentative and is based on our current genomic taxonomic sampling. The sequencing of new genomes from additional taxa, especially in cases where a lineage is covered by only a single taxon, will help improve the resolution of the evolutionary scenario here presented.

Phylostratigraphic Analysis of TFome Expression in Multicellular Ontogenies. Our results show that complex multicellular organisms with embryonic development (i.e., embryophytes and metazoans) have the richer TFomes that evolved through two bursts of innovation from the ancestral eukaryotic repertoire: one at the stem shared with their unicellular relatives and the other during the transition to multicellularity. To gain further insights into the similarities and differences between embryophytes and metazoans, we analyzed the deployment of TFome during the development of three model organisms, zebrafish (*Danio rerio*), fruit fly (*Drosophila melanogaster*), and mouse-ear cress (*Arabidopsis thaliana*).

We defined the phylostratigraphy of the TFome of these three species as previously described (10) (Fig. S4). Consistent with the domain-based analysis, we recovered a similar evolutionary history of the TFome: two major steps of diversification for both embryophytes and metazoans (Fig. S4). Metazoans show overrepresented gains in TFs in phylostrata ps2, ps5, ps6, and ps8 (Eukaryota, Opisthokonta, Holozoa, and Metazoa, respectively)

and embryophytes in phylostrata ps2, ps3, ps4, ps5, and ps6 (Eukaryota, Bikonta, Archaeplastida, Viridiplantae, Embryophyta). Moreover, founder-gene analysis suggests that the evolution of the TFome was marked by periods of novel protein emergence as well as phases of extensive duplication (Fig. S4).

Next, we used previously published gene expression datasets covering a series of developmental stages for these three model species (10, 44–46) to describe the average expression of the whole TFome during development (Fig. 4 and Figs. S5 and S6). However, the data used for *D. melanogaster* and *A. thaliana* have less staging density than *D. rerio* data, increasing the possible noise. *D. rerio* has two peaks of TF activity during embryogenesis, one during gastrulation and a less pronounced one in the pharyngula stage, known as the vertebrate phylotypic stage (Fig. 4*A*). *D. melanogaster* has a similar TFome profile during development to that observed in *D. rerio*, with a peak of TF expression during gastrulation and a subsequent peak during metamorphosis (Fig. S5*A*). In both these metazoan taxa, TF expression increases after early development and decreases in adult stages, when most structures have already developed. As *A. thaliana* datasets were much poorer in stages, we used two different expression datasets (45, 46). Both datasets show an increase in TF expression in the mature stage, in contrast to what is observed in metazoans (Fig. S6*A*). This can be explained by the fact that embryophytes have an indeterminate development (47), in which organogenesis and the formation of new structures take place later in development, even during the adult stage. Nevertheless, variability in *A. thaliana* datasets is still high, and there is lack of data of later developmental stages, so conclusions from *A. thaliana* should be taken with caution.

To evaluate the contribution of the phylostrata that showed statistically significant signals in the overrepresentation analyses (Fig. S4), we decomposed the data and calculated relative expressions of each phylostrata along development (10). *D. rerio* and *D. melanogaster* show similar expression patterns in early development, with a peak of opisthokont genes (ps5) before gastrulation and a peak of metazoan genes (ps9) during gastrulation (Fig. 4*C* and Fig. S5*C*), a defining developmental process that is essential for metazoan embryogenesis. The relative expression of genes from these two phylostrata (opisthokont and metazoan) again shows an increase in expression in later stages of development. Despite *D. melanogaster* being poorer in staging density, the patterns seem to be consistent with those in *D. rerio*. In contrast to metazoans, *A. thaliana* shows a very different pattern, in which eukaryotic TFs (ps2) predominate early development, followed by bikont TFs (ps3), and finally plant-specific TFs in the mature stage (ps4–6) (Fig. S6*C*). To evaluate general trends of TF age along development we used the transcriptome age index (TAI), from which we identified complementary patterns to the contribution of specific phylostrata, with newer genes being more predominant in later development of all three species (Fig. 4*B* and Figs. S5 and S6). The younger genes in later development of *D. rerio* may have been influenced by the presence of the reproductive tract in the samples, as numerous TF innovations have been described in the reproductive tract (48).

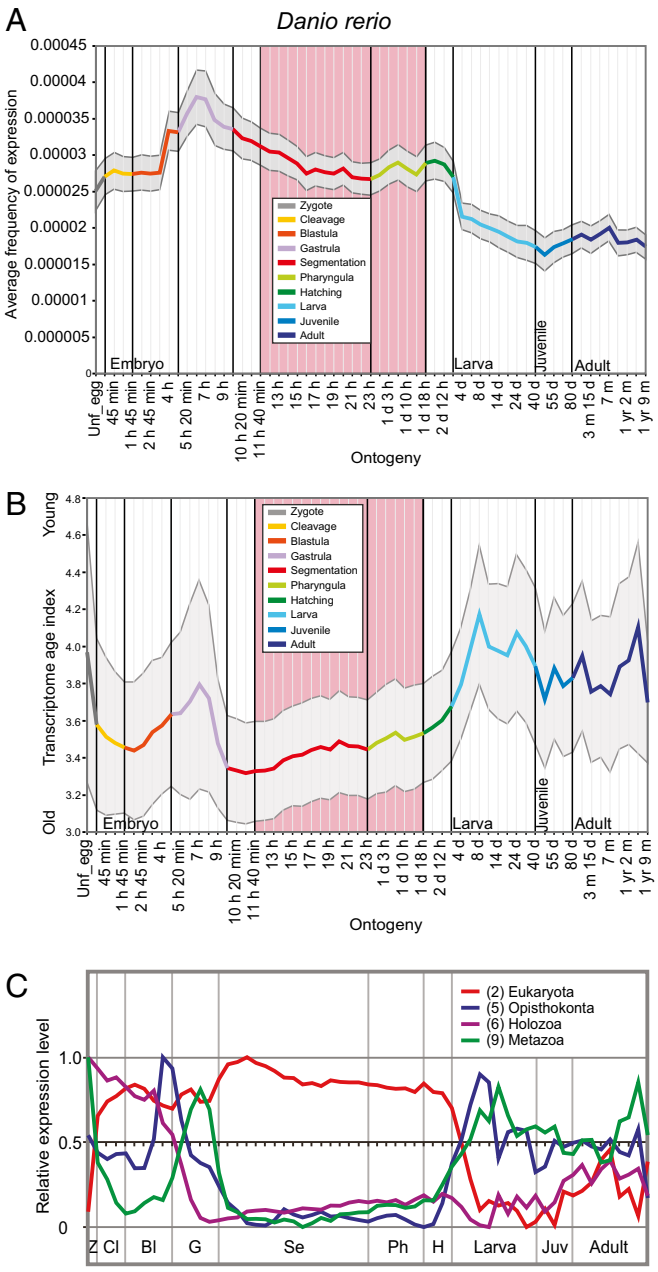

Fig. 4. (*A*) Average expression level of the TFome during development in *Danio rerio*. The area shaded in red is the proposed phylotypic stage (10). (*B*) Transcriptome age index (TAI) of the TFome during development in *Danio rerio*. The higher the TAI, the younger the TFs that are expressed, and vice versa. The gray area around TAI values corresponds to ±1 SEM. (*C*) Relative expression of TFs sorted by significantly enriched phylostrata during development in *Danio rerio*. For easier comparison, relative expression is shown in relation to the highest (0) and lowest (1) expression values across developmental stages (*Methods*). Bl, blastula; Cl, cleavage; G, gastrula; H, hatching; Juv, juvenile; Ph, pharyngula; Se, segmentation; Z, zygote.

Discussion

Phylogenetic Inertia, Lifestyle, and Genome Structure Influence the TF Repertoire of Eukaryotic Taxa. Our study reconstructs with unprecedented detail the phylogenetic distribution of TF classes during eukaryote evolution, based on the presence or absence of their specific DBDs. We provide evidence of a paneukaryotic TF complement that is present in almost all eukaryotic groups analyzed, and that was further expanded in some specific lineages, such as metazoans and embryophytes. These expansions were mainly due to gene duplications, and, in some lineages, to diversification in protein domain architecture. In contrast, other TF classes present sharp phylogenetic boundaries, to the point that some well-defined eukaryotic lineages can be defined by their shared and specific TFome profiles. This suggests an important role for phylogenetic inertia in defining the TFome composition of a given clade (49).

Our results suggest a relationship between TFome content, lifestyle, genome dynamics, and multicellularity. For example, strict parasitism, which has been shown to be linked to genome reduction and extensive gene loss (50), is associated with a

marked reduction of TFs. The opposite, however, is not always true, in that free-living species do not always have rich TFomes. Instead, WGDs are associated with rich TFomes. Moreover, the relative abundance of each TF class as a proportion of overall TFome of each species further highlights a strong phylogenetic patterning, probably as a result of the combined effects of phylogenetic inertia and system-level adaptation. In fact, each particular clade has a characteristic TFome profile, although a few lineages, such as Metazoa, Embryophyta, and Dikarya, show drastic shifts in their TFome profiles compared with their ancestors. Unicellular holozoans (Choanoflagellata, Filasterea, and Ichthyosporea) and early-branching fungi (Mucoromycota, Blastocladiomycota, and Chytridiomycota) have rather homogeneous TF profiles despite being paraphyletic. Conversely, the crown groups of these lineages, metazoans and Dikarya, respectively, changed their ancestral TF profile, suggesting that a genome-wide reshaping of the TF cellular function occurred at the boundaries of key evolutionary transitions followed by a posterior stasis.

Metazoa and Embryophyta Have the Richer TFome Among Eukaryotes. Notably, our data show that embryophytes and metazoans have the most complex TF repertoires of all eukaryotes. This contrasts with the relatively less complex TF repertoire of other multicellular lineages, such as Fungi, the brown algae *E. siliculosus*, or even the red alga *Chondrus crispus* (51). There exists the possibility that our analysis has missed as-yet-undescribed TFs that are unique to those lineages, as TFs that are restricted to small clades may be more prevalent than expected (52). In any case, fungi and brown and red algae did not expand their repertoire of paneukaryotic TFs (Homeoboxes, bHLH, and bZIPs) nor their protein domain architectures, in contrast to embryophytes or metazoans. We suggest that this difference is due to the fact that both embryophytes and metazoans go through a complex embryonic development, in contrast to the modular development of fungi and brown and red algae. The orchestrated embryonic development of both metazoans and embryophytes may need a more complex regulation and, thus, a more complete transcriptional regulation molecular toolkit.

Preadaptive Expansion of the TF Repertoire in the Unicellular Ancestors of Both Metazoa and Embryophyta. Moreover, our data show that the TF toolkit of metazoans and embryophytes was assembled in a stepwise manner from the ancestral eukaryotic TFome, with bursts of TF innovation in (*i*) their unicellular relatives, and (*ii*) at the origin of the metazoan or embryophyte lineages. Notably, both unicellular holozoans and chlorophytes already have complex TFomes, both with regards the number of genes and in the complexity of the protein domain architectures. Further expansions, however, occurred at the origin of both metazoans and embryophytes.

Analysis of TF Expression Patterns Highlight Differences on the Ontogeny of Both Metazoa and Embryophyta. The role of the different TFs in the ontogeny of the organisms analyzed varies according to the evolutionary origin of the different TFs in the tree of life (their phylostrata). Thus, the relative importance of TFs corresponding to different phylostrata changes between developmental stages. For example, gastrulation shows remarkably high expression of TFs of metazoan origin. In general, both in *D. rerio* and *D. melanogaster*, evolutionarily younger TFs seem to be more important in later stages of development, somehow being added to terminal specifications, whereas evolutionarily older genes are prevalent in the earlier stages of development. Indeed, TF expression by itself marks the phylotypic stage of metazoans.

Conclusions

To sum up, we show that the evolution of the TFome is characterized by important similarities between the independent transitions to multicellularity of embryophytes and metazoans. Not only do they share a common pattern of TF origin and expansion, which is richer than other eukaryotes, but also express their TFs differentially during development, depending on their evolutionary age. This suggests that common evolutionary forces drove the unicellular-to-multicellular transition in two phylogenetically distinct lineages. We hypothesize that is due to their complex embryonic development rather than just their multicellular lifestyle or number of cell types. The success of metazoans and embryophytes in producing extensive morphological diversification is due to the specific adaptations of their genomes, with the TFome being a key aspect into the acquisition of a complex multicellularity and phenotypic plasticity.

Methods

TF Identification. We obtained data on complete proteomes from publicly available databases. We also used RNAseq data for some taxa sequenced by the Broad Institute (www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html). To have a better representation of unicellular taxa close to Metazoa, we also screened some taxa whose RNAseq were sequenced in-house. These include the nucleariid *Nuclearia* sp. (sister group of Fungi) and several unicellular holozoans (*Ministeria vibrans*, *Pirum gemmatta*, *Abeoforma whisleri*, *Amoebidium parasiticum*, and *Corallochytrium limacisporum)*. In this case, nonfiltered reads were assembled using Trinity software (53), and a six-frame translation of the assembly was generated. A PfamScan was performed on all proteomes and transcriptomes using PFAM A, version 26, and selecting the gathering threshold option as a conservative approach, which can underestimate total counts for some domains but minimizes false positives (54).

TFs were selected from two database resources (Transcriptionfactor.org, PfamTOGo website) as well as previous studies on this topic (6, 16, 55). In all cases, we defined a univocal one-to-one relationship between TF class and DBD class. DBDs that appeared just in combination with other DBDs were neglected to avoid an overestimation of TF numbers in some genomes.

We counted the number of genes containing a given DBD, and the number of different associated domain architectures associated with each DBD in each species using custom Perl scripts (Fig. 1 and Fig. S1). In cases in which two or more DBDs were found in the same gene, the evolutionarily older DBD, or the larger when age was identical, was considered the defining of the TF type. To avoid overestimating architecture numbers due to problems with detecting small repeated domains, consecutively repeated domains were counted only once. Therefore, domain architectures were defined on the basis of the number and order of their domains, but not by the number of repeats of each domain.

Ancestral Genome Reconstruction and Enrichment. Statistical analyses were performed using custom R scripts. We tested for enrichment of TF numbers using the Wilcoxon rank-sum test, with a significance threshold of $P < 0.01$. Cluster analysis was performed using the pvclust R package, using the absolute value of sample correlation, complete hierarchical clustering, and 100,000 replications.

Analysis of gain and loss of TFs and reconstruction of ancestral state were performed using Mesquite (56), and the most parsimonious assumption was taken, except in the cases highlighted in Fig. 3. In cases when a DBD's secondary loss was inferred based on its absence in a single species (lineages represented by a single taxon are Rhizaria, Haptophyta, Cryptophyta, Glaucophyta, and Rhodophyta), we confirmed its absence by performing tblastn against EST and nucleotide collection (nr/nt) databases at the National Center for Biotechnology Information (NCBI) website, which include other taxa from those lineages. The results are indicated in Table S1.

Phylostratigraphic Analyses. The theoretical basis of genomic phylostratigraphy and detailed phylostratigraphic procedures have been described previously (10, 13, 57–59). Protein coding sequences for *Danio rerio* (25,638 genes), *Drosophila melanogaster* (13,413 genes), and *Arabidopsis thaliana* (27,148 genes) were retrieved from the Ensembl database (version 69) (60). We compared these protein sequences against the nonredundant (nr) database from NCBI using the BLASTP algorithm (BLAST program) with an E-value cutoff of 10^{-3} (61). This database contains the most exhaustive set of known proteins across all organisms and is therefore the most suitable dataset for phylostratigraphic analysis. Before performing sequence simi-

larity searches, we excluded all sequences of viral or unknown taxonomic origin, as well as those from metazoan taxa with a currently unreliable phylogenetic position (Myxozoa, Mesozoa, Chaetognatha, and Placozoa). Following this cleanup procedure, we complemented the nr database with sequenced genomes that were not present in the database but were otherwise available in other public repositories. The final database contained 8,393,531 protein sequences.

Using the BLAST output obtained above, we selected the gene IDs from the TFs obtained from PfamScan analysis for each species separately and mapped them onto the consensus phylogeny. We used the phylogenetically most distant BLAST match using an E-value cutoff of 10^{-3} as the criterion for assigning the evolutionary origin of a gene. This is a quite conservative method for sorting genes that aims to detect novelty in the protein sequence space (10, 57, 62). The number and choice of internodes in the phylogeny are a result of balancing the robustness of these internodes in phylogenetic studies (32, 63–68), the availability of sequence data for the sequence similarity searches, and the importance of evolutionary transitions (Table S2). Our consensus phylogenies span 16 evolutionary levels (phylostrata) for *Arabidopsis thaliana*, 18 for *Danio rerio*, and 20 for *Drosophila melanogaster*, starting from the origin of cellular organisms (ps1).

We also calculated the number of founder TFs in each phylostratum by self-comparing sequences within a phylostratum by BLAST analysis (E-value cutoff 1e 10^{-3}). For instance, we compared all of the TFs in the first phylostratum (ps1) to TFs genes from phylostratum 1 (ps1). The number of founder TFs in each phylostratum, G_f, was calculated from the number of hits (H) obtained for every TF using Eq. **1**:

$$G_f = \sum_{i=1}^{G} \frac{1}{H_i}, \qquad [1]$$

where G represents the number of TFs in the phylostratum and $1 \leq H \leq G$. The lowest possible value of G_f is 1, indicating that all genes in the phylostratum are related, and the highest possible value is G, indicating that all genes in the phylostratum are founders. However, gene founder analysis encompass full sequence length not only domains that are signatures of TFs, providing a complementary view on the DBD approach. The additional domains present in a TF gene may be responsible for the overestimation of phylostratum 1 counts in *Arabidopsis thaliana* TFs phylostratigraphy.

Overrepresentation Analysis and Statistics. We analyzed overrepresentation of TFs both at the level of total genome and at the level of founder genes. In both cases, and for every group of TFs of interest, we performed overrepresentation analyses by comparing the frequency of TFs in each phylostratum to that of all genes in that phylostratum (expected frequency) (57, 58). The deviations obtained were expressed as log-odds ratios, and their significance was tested using a two-tailed hypergeometric test (69), corrected for multiple comparisons using a false-discovery rate of 0.05 (70) (Fig. S4).

TAI. TAI is a measure that reflects the evolutionary age of a transcriptome at a given ontogenetic stage (10). TAI is the weighted mean of phylogenetic ranks (phylostrata) and is calculated for every ontogenetic stage s as follows (Eq. **2**):

$$\mathrm{TAI}_s = \frac{\sum_{i=1}^{n} \mathrm{ps}_i \mathrm{e}_i}{\sum_{i=1}^{n} \mathrm{e}_i}, \qquad [2]$$

where ps_i is an integer representing the phylostratum of gene i (e.g., 1, the oldest; 18, the youngest for *D. rerio*), e_i is the microarray signal intensity value of gene i, which acts as weighting factor, and n is the total number of genes analyzed. Note that TAI = 1 indicates that all expressed genes in a specific stage coming from ps1 (origin of cellular organisms), and TAI = 18 (zebrafish) indicates that a given stage expresses only *D. rerio*-specific genes. Therefore, lower TAI values correspond to a phylogenetically older transcriptome.

ACKNOWLEDGMENTS. We thank Joint Genome Institute and Broad Institute for making these data publicly available. We thank Andy Baxevanis for sharing Mnemiopsis *leidyi* sequences, Scott A. Nichols for sharing *Oscarella carmela* sequences, Carlos Palacín for his help, and Hiroshi Suga for helpful insights in the bioinformatics analyses and sharing Creolimax *fragrantissima* sequences. This work was supported by a contract from the Institució Catalana de Recerca i Estudis Avançats, European Research Council Starting Grant ERC-2007-StG-206883, and Ministerio de Economía y Competitividad (MINECO) Grant BFU2011-23434 (to I.R.-T.). A.d.M. was supported by a Formación Personal Investigador grant from MINECO. A.S.-P. was supported by a pregraduate Formación Profesorado Universitario grant from MINECO.

1. Wunderlich Z, Mirny LA (2009) Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet* 25(10):434–440.
2. Latchman DS (1997) Transcription factors: An overview. *Int J Biochem Cell Biol* 29(12): 1305–1312.
3. Latchman DS (2008) *Eukaryotic Transcription Factors* (Elsevier/Academic Press, Amsterdam), 5th Ed.
4. Papavassiliou A (1997) *Transcription Factors in Eukaryotes* (Landes Bioscience, Austin, TX).
5. Spitz F, Furlong EEM (2012) Transcription factors: From enhancer binding to developmental control. *Nat Rev Genet* 13(9):613–626.
6. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: Function, expression and evolution. *Nat Rev Genet* 10(4):252–263.
7. Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424 (6945):147–151.
8. Charoensawan V, Wilson D, Teichmann SA (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res* 38(21): 7364–7377.
9. Meyerowitz EM (2002) Plants compared to animals: The broadest comparative study of development. *Science* 295(5559):1482–1485.
10. Domazet-Lošo T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468(7325):815–818.
11. Kalinka AT, et al. (2010) Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468(7325):811–814.
12. Irie N, Kuratani S (2011) Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat Commun* 2:248.
13. Quint M, et al. (2012) A transcriptomic hourglass in plant embryogenesis. *Nature* 490(7418):98–101.
14. Levin M, Hashimshony T, Wagner F, Yanai I (2012) Developmental milestones punctuate gene expression in the *Caenorhabditis* embryo. *Dev Cell* 22(5):1101–1108.
15. Charoensawan V, Wilson D, Teichmann SA (2010) Lineage-specific expansion of DNA-binding transcription factor families. *Trends Genet* 26(9):388–393.
16. Weirauch MT, Hughes TR (2011) *A Handbook of Transcription Factors*, ed Hughes TR (Springer Netherlands, Dordrecht), pp 25–73.
17. Lang D, et al. (2010) Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: A timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol Evol* 2:488–503.
18. Shelest E (2008) Transcription factors in fungi. *FEMS Microbiol Lett* 286(2):145–151.
19. Sebé-Pedrós A, de Mendoza A, Lang BF, Degnan BM, Ruiz-Trillo I (2011) Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki*. *Mol Biol Evol* 28(3):1241–1254.
20. Degnan BM, Vervoort M, Larroux C, Richards GS (2009) Early evolution of metazoan transcription factors. *Curr Opin Genet Dev* 19(6):591–599.
21. Rayko E, Maumus F, Maheswari U, Jabbari K, Bowler C (2010) Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytol* 188(1):52–66.
22. Shiu SH, Shih MC, Li WH (2005) Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiol* 139(1):18–26.
23. Bürglin TR (2011) *A Handbook of Transcription Factors*, ed Hughes TR (Springer, Dordrecht, The Netherlands), pp 95–122.
24. Ruiz-Trillo I, et al. (2007) The origins of multicellularity: A multi-taxon genome initiative. *Trends Genet* 23(3):113–118.
25. Glockling SL, Marshall WL, Gleason FH (2013) Phylogenetic interpretations and ecological potentials of the Mesomycetozoea (Ichthyosporea). *Fungal Ecol* 6:237–247.
26. Aury J-M, et al. (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444(7116):171–178.
27. Albertin W, Marullo P (2012) Polyploidy in fungi: Evolution after whole-genome duplication. *Proc R Soc B Biol Sci* 279(1738):2497–509.
28. De Smet R, et al. (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci USA* 110(8): 2898–2903.
29. Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. *Nat Rev Genet* 10(10):725–732.
30. Srivastava M, et al. (2010) The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* 466(7307):720–726.
31. Pick KS, et al. (2010) Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol Biol Evol* 27(9):1983–1987.
32. Philippe H, et al. (2009) Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* 19(8):706–712.
33. Nosenko T, et al. (2013) Deep metazoan phylogeny: When different genes tell different stories. *Mol Phylogenet Evol* 67(1):223–233.
34. Sebé-Pedrós A, Zheng Y, Ruiz-Trillo I, Pan D (2012) Premetazoan origin of the hippo signaling pathway. *Cell Rep* 1(1):13–20.
35. Clarke M, et al. (2013) Genome of *Acanthamoeba castellanii* highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome Biol* 14(2):R11.
36. Price DC, et al. (2012) *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science* 335(6070):843–847.

37. Cock JM, et al. (2010) The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* 465(7298):617–621.
38. Radakovits R, et al. (2012) Draft genome sequence and genetic transformation of the oleaginous alga *Nannochloropis gaditana*. *Nat Commun* 3:686.
39. Peters AF, et al. (2008) Life-cycle-generation-specific developmental processes are modified in the immediate upright mutant of the brown alga *Ectocarpus siliculosus*. *Development* 135(8):1503–1512.
40. Coelho SM, et al. (2011) OUROBOROS is a master regulator of the gametophyte to sporophyte life cycle transition in the brown alga *Ectocarpus*. *Proc Natl Acad Sci USA* 108(28):11518–11523.
41. Bouget FY, Berger F, Brownlee C (1998) Position dependent control of cell fate in the Fucus embryo: Role of intercellular communication. *Development* 125(11):1999–2008.
42. King N (2004) The unicellular ancestry of animal development. *Dev Cell* 7(3):313–325.
43. Rokas A (2008) The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annu Rev Genet* 42:235–251.
44. Arbeitman MN, et al. (2002) Gene expression during the life cycle of *Drosophila melanogaster*. *Science* 297(5590):2270–2275.
45. Xiang D, et al. (2011) Genome-wide analysis reveals gene expression and metabolic network dynamics during embryo development in *Arabidopsis*. *Plant Physiol* 156(1): 346–356.
46. Belmonte MF, et al. (2013) Comprehensive developmental profiles of gene activity in regions and subregions of the *Arabidopsis* seed. *Proc Natl Acad Sci USA* 110(5): E435–E444.
47. Valentine JW, Tiffney BH, Sepkoski JJ, Jr. (1991) Evolutionary dynamics of plants and animals: A comparative approach. *Palaios* 6:81–88.
48. Turner LM, Hoekstra HE (2008) Causes and consequences of the evolution of reproductive proteins. *Int J Dev Biol* 52(5-6):769–780.
49. Burt DB (2001) Evolutionary stasis, constraint and other terminology describing evolutionary patterns. *Biol J Linn Soc Lond* 72:509–517.
50. Kurland CG, Collins LJ, Penny D (2006) Genomics and the irreducible nature of eukaryote cells. *Science* 312(5776):1011–1014.
51. Collén J, et al. (2013) Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. *Proc Natl Acad Sci USA* 110(13):5247–5252.
52. Lohse MB, et al. (2013) Identification and characterization of a previously undescribed family of sequence-specific DNA-binding domains. *Proc Natl Acad Sci USA* 110(19): 7660–7665.
53. Grabherr MG, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7):644–652.
54. Punta M, et al. (2012) The Pfam protein families database. *Nucleic Acids Res* 40(Database issue):D290–D301.
55. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA (2008) DBD—taxonomically broad transcription factor predictions: New content and functionality. *Nucleic Acids Res* 36(Database issue):D88–D92.
56. Maddison WP, Maddison DR (2011) *Mesquite: A Modular System for Evolutionary Analysis, Version 275*. Available at http://mesquiteproject.org. Accessed October 15, 2012.
57. Domazet-Loso T, Brajković J, Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet* 23(11): 533–539.
58. Domazet-Loso T, Tautz D (2010) Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol* 8:66.
59. Carvunis A-R, et al. (2012) Proto-genes and de novo gene birth. *Nature* 487(7407): 370–374.
60. Flicek P, et al. (2011) Ensembl 2011. *Nucleic Acids Res* 39(Database issue):D800–D806.
61. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
62. Tautz D, Domazet-Lošo T (2011) The evolutionary origin of orphan genes. *Nat Rev Genet* 12(10):692–702.
63. Blair JE, Hedges SB (2005) Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol* 22(11):2275–2284.
64. Delsuc F, Brinkmann H, Chourrout D, Philippe H (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* 439(7079):965–968.
65. Delsuc F, Tsagkogeorga G, Lartillot N, Philippe H (2008) Additional molecular support for the new chordate phylogeny. *Genesis* 46(11):592–604.
66. Telford MJ, Bourlat SJ, Economou A, Papillon D, Rota-Stabelli O (2008) The evolution of the Ecdysozoa. *Philos Trans R Soc Lond B Biol Sci* 363(1496):1529–1537.
67. Budd GE, Telford MJ (2009) The origin and evolution of arthropods. *Nature* 457(7231):812–817.
68. Heimberg AM, Cowper-Sal-lari R, Sémon M, Donoghue PC, Peterson KJ (2010) microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc Natl Acad Sci USA* 107(45):19379–19383.
69. Rivals I, Personnaz L, Taing L, Potier MC (2007) Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics* 23(4):401–407.
70. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B* 57(1):289–300.
71. Hampl V, et al. (2009) Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc Natl Acad Sci USA* 106(10):3859–3864.
72. Derelle R, Lang BF (2012) Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol Biol Evol* 29(4):1277–1289.
73. Burki F, Okamoto N, Pombert J-F, Keeling PJ (2012) The evolutionary history of haptophytes and cryptophytes: Phylogenomic evidence for separate origins. *Proc Biol Sci* 279(1736):2246–2254.
74. Torruella G, et al. (2012) Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Mol Biol Evol* 29(2):531–544.
75. Brown MW, Kolisko M, Silberman JD, Roger AJ (2012) Aggregative multicellularity evolved independently in the eukaryotic supergroup Rhizaria. *Curr Biol* 22(12): 1123–1127.