KLF/SP Transcription Factor Family Evolution: Expansion, Diversification, and Innovation in Eukaryotes

Jason S. Presnell[1], Christine E. Schnitzler[2], and William E Browne[1,*]

[1]Department of Biology, University of Miami

[2]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health

*Corresponding author: E-mail: wbrowne@bio.miami.edu.

Accepted: July 22, 2015

Data deposition: Sequences associated with this project have been deposited at GenBank under accession numbers KJ576841 and KJ576842.

Abstract

The Krüppel-like factor and specificity protein (KLF/SP) genes play key roles in critical biological processes including stem cell maintenance, cell proliferation, embryonic development, tissue differentiation, and metabolism and their dysregulation has been implicated in a number of human diseases and cancers. Although many KLF/SP genes have been characterized in a handful of bilaterian lineages, little is known about the KLF/SP gene family in nonbilaterians and virtually nothing is known outside the metazoans. Here, we analyze and discuss the origins and evolutionary history of the KLF/SP transcription factor family and associated transactivation/repression domains. We have identified and characterized the complete KLF/SP gene complement from the genomes of 48 species spanning the Eukarya. We have also examined the phylogenetic distribution of transactivation/repression domains associated with this gene family. We report that the origin of the KLF/SP gene family predates the divergence of the Metazoa. Furthermore, the expansion of the KLF/SP gene family is paralleled by diversification of transactivation domains via both acquisitions of pre-existing ancient domains as well as by the appearance of novel domains exclusive to this gene family and is strongly associated with the expansion of cell type complexity.

Key words: C2H2 zinc fingers, domain shuffling, domain architecture, domain co-occurrence network, domain evolution, low-complexity regions.

Introduction

One of the most ancient and abundant classes of DNA binding domains (DBDs) is the C2H2 zinc finger class (Rubin et al. 2000; Ravasi et al. 2003; de Mendoza et al. 2013). The C2H2 zinc finger domain has two cysteine and two histidine residues that coordinate a zinc ion, and typically consists of the amino acid sequence C-X(2-4)-C-X(12)-H-X(3-5)-H (Brown et al. 1985; Miller et al. 1985). C2H2 zinc finger motifs are found in many transcription factors and based on their arrangement and number can be subdivided into different families (Iuchi 2001). The Krüppel-like factor and specificity protein (KLF/SP) transcription factor gene family is characterized by a highly conserved triple-C2H2 DBD located toward the C-terminus composed of three tandem zinc fingers that are evenly spaced by conserved linker regions (Iuchi 2001) and share similarity with the *Drosophila Krüppel* gene (Rosenberg et al. 1986). This C2H2 zinc finger DBD (KLF-DBD) binds to Guanine-Cytosine-rich regions and CACC elements (GT boxes) (Kadonaga et al. 1987). The more N-terminal regions of KLF/SP transcription factors are typically highly variable and consist of different combinations of transactivation/repression domains. Historically, mammalian KLFs have been divided into 3 groups based on shared domain architecture: The KLF1, 2, 4, 5, 6, and 7 groups; the KLF3, 8, and 12 groups; and the KLF9, 10, 11, 13, 14, 16 groups (McConnell and Yang 2010), whereas SPs, which differ from KLFs by the presence of the Buttonhead (Btd) box domain just 5′ of the KLF-DBD, are typically divided into 2 groups: SP1–4 and SP5–9 (Suske et al. 2005).

KLF/SP genes within each domain architecture group share similar functions based on the retention of explicit transactivation motif complements. A range of studies present a complex picture in which KLF/SP genes can be singly or combinatorially involved in temporally and spatially disparate cellular and developmental processes. For example, fly embryos mutant for *luna*, the *Drosophila* KLF6/7 ortholog, die early during

development due to mitotic defects (De Graeve et al. 2003; Weber et al. 2014), whereas *cabut*, a KLF9/13 ortholog in *Drosophila*, is required for proper dorsal closure during gastrulation (Muñoz-Descalzo et al. 2005). However, *cabut* also plays a role later in fly organ development by coordinating signaling for proper wing disc patterning (Rodriguez 2011). Among the vertebrates, KLF genes are often associated with balancing stem cell proliferation and differentiation, as well as regulating metabolic homeostasis. The most notable member is KLF4, one of the four pioneer transcription factors required to induce pluripotency in human and mouse fibroblasts (Takahashi and Yamanaka 2006; Soufi et al. 2012, 2015) and a component of a core circuit of genes that maintain self-renewal in mammalian embryonic stem cells along with KLF2 and KLF5 (Jiang et al. 2008). However, in gut epithelia, KLF4 regulates terminally differentiated cells while KLF5 is expressed in the proliferating crypt cells (McConnell et al. 2007). In mammals, KLF2 together with KLF1 and KLF13 also regulate erythrocyte maturation and differentiation as well as globin gene activity (Miller and Bieker 1993; Basu et al. 2005; Gordon et al. 2008). KLF2 in zebrafish contributes to the differentiation of ectoderm derived tissues (Kotkamp et al. 2014). In mammals, including humans, KLF11 and KLF14 play an important role in the regulation of genes associated with diabetes and metabolic syndrome phenotypes, respectively (Small et al. 2011; Lomberk et al. 2013). Similarly, complex intersections with both development and metabolism exist for members of the SP subfamily. For example, in mammals, SP1, SP3, and SP7 regulate osteoblast mineralization and differentiation (Nakashima et al. 2002; Suttamanatwong et al. 2009). SP1 is also an important regulator of metabolic genes involved in the glycolytic pathway, fatty acid synthesis, and ribosome biogenesis (Archer 2011; Nosrati et al. 2014). Overall, members of the KLF/SP gene family are known to function in a wide variety of biological processes (Black et al. 2001; Zhao and Meng 2005; Pearson et al. 2008; Wierstra 2008; McConnell and Yang 2010; Zhao et al. 2010; Tsai et al. 2014).

In contrast to the extensive studies highlighting the importance of the KLF/SP genes to core cellular processes, comparatively few studies have investigated the evolutionary relationship of KLF/SP genes in lineages outside of mammals (Kolell and Crawford 2002; Materna et al. 2006; Shimeld 2008; Chen et al. 2009; Meadows et al. 2009; Schaeper et al. 2010; Seetharam et al. 2010). A KLF gene was recently identified in the choanoflagellate *Monosiga brevicollis* genome; however, that study's conclusions were restricted to examining porcine KLF paralogy (Chen et al. 2009). A more recent study, focused on the phylogenetic distribution of C2H2 zinc finger families in eukaryotes, also showed that KLFs were present in *Monosiga* but absent in the fungal taxa surveyed (Seetharam and Stuart 2013). No study to date has examined the phylogenetic context of the different transactivation/repression domains associated with the KLF/SP gene family. Pinpointing the origin and evolutionary history of this gene family and associated domains can help determine possible relationships of the KLF/SP repertoire expansion to key innovations in the evolution of metazoan cellular diversity. Hypotheses of metazoan gene evolution are greatly aided by sampling a wide range of taxa that include nonmetazoan representatives. Here, we infer the evolutionary history of the KLF/SP gene family and their associated transactivation/repression domains across a wide range of eukaryotes. Species were chosen to represent well established groups across the Eukarya including the following: Bikonta, Amorphea (Amoebozoa + Apusozoa + Opisthokonta), Opisthokonta (Holomycota + Holozoa), and Holozoa (Ichthyosporea + Filasterea + Choanoflagellata + Metazoa) (fig. 1) (Adl et al. 2012; Derelle and Lang 2012; Paps et al. 2013). A number of transactivation/repression domains found in KLF/SPs are also present in the genomes of ancient unicellular lineages that lack KLF/SPs, lending support for domain shuffling playing a major role in the acquisition of transactivation/repression domains during the expansion and diversification of the KLF/SP gene family. We show that domain connectivity and resulting unique domain architectures among these transcription factors have become increasingly complex in metazoans. Thus a pattern of gene duplication along with domain shuffling and the rare emergence of de novo domains have collectively played a vital role in the evolution of the KLF/SP gene family.

Materials and Methods

KLF Identification Pipeline

To broadly identify C2H2 zinc finger and KLF/SP proteins associated with a diverse range of Eukaryote lineages, publicly available sequenced genomes listed in table 1 were comprehensively searched including 26 metazoans, 4 unicellular holozoans, 7 holomycotans, 1 apusozoan, 4 amoebozoans, and 6 bikonts. We used the HMMER 3.0 program (Eddy 1998), to identify proteins that contained C2H2 zinc fingers. The Hidden Markov Model (*hmm*) of the C2H2 zinc finger domain PF00096 (Punta et al. 2012) was downloaded from the Pfam database. The *hmmsearch* command was then used to search protein models of the representative 48 Eukaryote species using the PF00096 *hmm* as a query. HMMER identified all protein sequences (using default settings) that contained at least one C2H2 zinc finger corresponding to the *hmm.* This output was used for subsequent analyses. Raw outputs for this and subsequent steps of the pipeline can be found in Supplementary Material online

We then used a perl script modified from Zeng et al. (2011) to search the protein sequences of the HMMER output for the 81 amino acid triple-C2H2 zinc finger DBD conserved in KLF/SPs (KLF-DBD). The amino acid sequence is as follows: C-X_4-C-X_{12}-H-X_3-H-X_7-C-X_4-C-X_{12}-H-X_3-H-X_7-C-X_2-C-X_{12}-H-X_3-H, where X can be any amino acid. Sequences meeting the

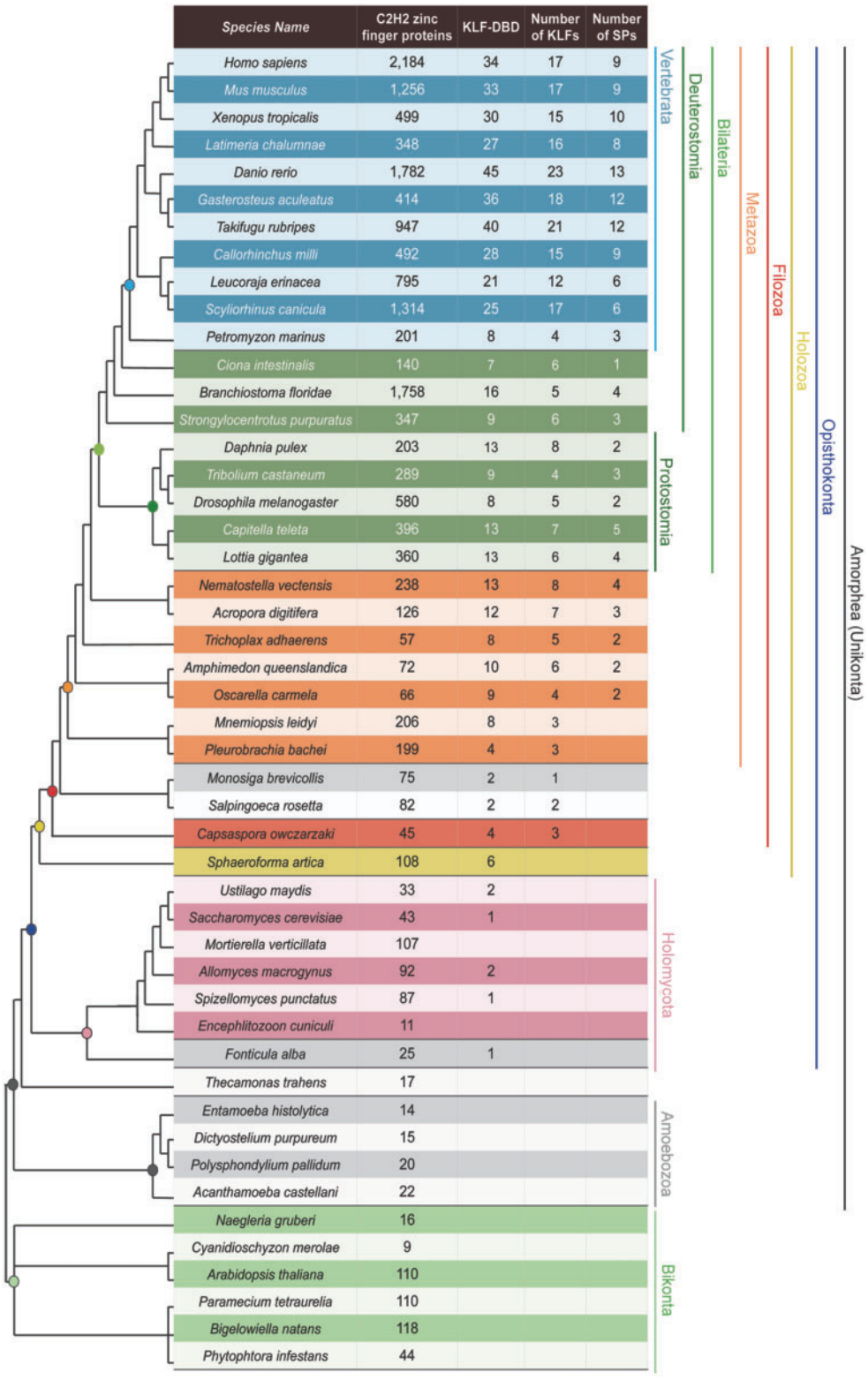

Fig. 1.—Distribution of C2H2 zinc finger proteins, KLF-DBD containing proteins, and KLF/SP proteins in representative Eukarya taxa. Rows indicate representative genomes searched. Columns indicate the total number of protein sequences that contain at least one C2H2 zinc finger using the Pfam PF00096 HMM model, the total number of protein sequences that contain the archetypical KLF-DBD, the total number of bona fide KLF sequences recovered, and the total number of SP sequences recovered. Phylogeny is based on Adl et al. (2012), Derelle and Lang (2012), Dunn et al. (2008), Ryan et al. (2013), and Sebé-Pedrós et al. (2013).

Table 1
Species used in this Study with Reference to Genome or Transcriptome Database

	Species	Genome/Transcriptome	Reference
Amorphea	*Homo sapiens*	ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/	Lander et al. (2001)
	Mus musculus	www.ensembl.org/Mus_musculus/	Chinwalla et al. (2002)
	Xenopus tropicalis	http://useast.ensembl.org/Xenopus_tropicalis/Info/Index	Hellsten et al. (2010)
	Latimeria chalumnae	http://useast.ensembl.org/Latimeria_chalumnae/Info/Index	Amemiya et al. (2013)
	Danio rerio	www.ensembl.org/Danio_rerio/Info/Index	Howe et al. (2013)
	Takifugu rubripes	www.ensembl.org/Takifugu_rubripes/Info/Index	Aparicio et al. (2002)
	Gasterosteus aculeatus	http://useast.ensembl.org/Gasterosteus_aculeatus/Info/Index	Jones et al. (2012)
	Callorhinchus milii	http://esharkgenome.imcb.a-star.edu.sg/download/	Venkatesh et al. (2014)
	Leucoraja erinacea	http://skatebase.org/downloads	
	Scyliorhinus canicula	http://skatebase.org/downloads	
	Petromyzon marinus	http://useast.ensembl.org/Petromyzon_marinus/Info/Index	Smith et al. (2013)
	Ciona intestinalis	http://useast.ensembl.org/Ciona_intestinalis/Info/Index	Dehal et al. (2002)
	Branchiostoma floridae	http://genome.jgi-psf.org/Brafl1/Brafl1.home.html	Putnam et al. (2008)
	Strongylocentrotus purpuratus	http://metazoa.ensembl.org/Strongylocentrotus_purpuratus/Info/Index	Sodergren et al. (2006)
	Daphnia pulex	http://genome.jgi-psf.org/Dappu1/Dappu1.download.ftp.html	Colbourne et al. (2011)
	Drosophila melanogaster	http://flybase.org	Adams et al. (2000)
	Tribolium castaneum	http://beetlebase.org	Richards et al. (2008)
	Capitella teleta	http://genome.jgi.doe.gov/Capca1/Capca1.download.ftp.html	Simakov et al. (2013)
	Lottia gigantea	http://genome.jgi-psf.org/Lotgi1/Lotgi1.download.ftp.html	Simakov et al. (2013)
	Nematostella vectensis	http://genome.jgi.doe.gov/Nemve1/Nemve1.home.html	Putnam et al. (2007)
	Acropora digitifera	http://www.compagen.org/datasets.html	Shinzato et al. (2011)
	Trichoplax adhaerens	http://genome.jgi.doe.gov/Triad1/Triad1.home.html	Srivastava et al. (2008)
	Amphimiedon queenslandica	http://spongezome.metazome.net/cgi-bin/gbrowse/amphimedon/	Srivastava et al. (2010)
	Oscarella carmela	http://www.compagen.org/datasets.html	
	Mnemiopsis leidyi	http://research.nhgri.nih.gov/mnemiopsis/	Ryan et al. (2013)
	Pleurobrachia bachei	http://neurobase.rc.ufl.edu/pleurobrachia	Moroz et al. (2014)
Metazoa	*Monosiga brevicollis*	http://genome.jgi.doe.gov/Monbr1/Monbr1.home.html	King et al. (2008)
	Salpingoeca rosetta	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html	Fairclough et al. (2013)
	Capsaspora owczarzaki	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html	Suga et al. (2013)
	Sphaeroforma arctica	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html	
Holomycota	*Ustilago maydis*	http://www.broadinstitute.org/annotation/genome/ustilago_maydis	Kamper et al. (2006)
	Saccharomyces cerevisiae	http://www.yeastgenome.org/	Goffeau et al. (1996)
	Mortierella verticillata	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html	
	Allomyces macrogynus	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html	
	Spizellomyces punctatus	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html	

(continued)

Table 1 Continued

	Species	Genome/Transcriptome	Reference
	Encephalitozoon cuniculi	http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Enccu1	Katinka et al. (2001)
	Fonticula alba	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html	
Opisthokonta	*Thecamonas trahens*	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html	
	Entamoeba histolytica	http://amoebadb.org/common/downloads/	Loftus et al. (2005)
	Dictyostelium purpureum	http://genome.jgi-psf.org/Dicpu1/Dicpu1.home.html	Sucgang et al. (2011)
	Polysphondylium pallidum	http://genomes.dictybase.org/pallidum/current	Heidel et al. (2011)
Amorphea	*Acanthamoeba castellani*	http://amoebadb.org/common/downloads/	Clarke et al. (2013)
Bikonta	*Naegleria gruberi*	http://genome.jgi-psf.org/Naegr1/Naegr1.download.ftp.html	Fritz-Laylin et al. (2010)
	Cyanidioschyzon merolae	http://merolae.biol.s.u-tokyo.ac.jp/download/	Matsuzaki et al. (2004)
	Arabidopsis thaliana	http://www.arabidopsis.org/	Arabidopsis Genome Iniative (2000)
	Paramecium tetraurelia	http://paramecium.cgm.cnrs-gif.fr/download/fasta/	Aury et al. (2006)
	Bigelowiella natans	http://genome.jgi.doe.gov/Bigna1/Bigna1.download.ftp.html	Curtis et al. (2012)
	Phytophtora infestans	http://protists.ensembl.org/Phytophthora_infestans/Info/Index	Haas et al. (2009)

following criteria were initially considered putative KLF/SPs: 1) Presence of the KLF-DBD and 2) presence of only three zinc fingers (i.e., no additional zinc fingers other than the KLF-DBD). Sequences fitting these initial criteria were only found within the Filozoa (supplementary fig. S1*A*, Supplementary Material online). We additionally found 13 nonfilozoan amorphean sequences possessing a KLF-DBD; however, in all cases these sequences also contained additional zinc fingers (Supplementary Material online). To determine the relationship between the nonfilozoan amorphean sequences and the putative filozoan KLF/SPs, all sequences were aligned and included in our phylogenetic analyses.

Where possible, partial KLF/SP gene models were manually extended by obtaining genomic scaffold regions spanning the location of incompletely annotated gene models. Briefly, the CLCBio software package was used to map partial KLF/SP nucleotide sequences to genomic scaffolds and identify associated open reading frames. Extended exonic nucleotide sequences were then translated to obtain enhanced annotation of KLF/SP amino acid sequences. Protein schematics were created using the PROSITE MyDomians Image Creator (http://prosite.expasy.org/cgi-bin/prosite/mydomains/, last accessed April 30, 2014). Scripts for the KLF/SP identification pipeline are publicly available (Supplementary Material online).

Transactivation/Repression Domain Identification and Characterization

Custom perl scripts were used to search for KLF/SP-associated transactivation/repression domains. The motifs used for the perl scripts followed these amino acid sequences: Btd box, C-X-C-P-X-C (Wimmer et al. 1993); SP box, S/T-P-L-X-ϕ-L-X-X-X-C-X-R/K-ϕ (Harrison et al. 2000); SID, D-X_{1-4}-X-A-ϕ-X-X-L-M/V/L/A-X-F/M/L/I (Zhang et al. 2001); PVDLS, P-V-D-L-S/T (Crossley et al. 1996); R2, S-V-I-R-H-T-X-D/E (Cook et al. 1999); and R3, ϕ-X-X-G-X-ϕ-ϕ-ϕ-ϕ-ϕ-P/S-Q/P (Cook et al. 1999), where X can be any amino acid and ϕ represents one of V, I, L, M, F, W, G, A, or P. To confirm the accuracy of the automated identification of transactivation/repression domains, we first searched the completely annotated human KLF/SP amino acid sequences. In all cases, our searches identified the human KLF/SP sequences that had previously been characterized and annotated as containing the specified domains. For example, the search with the Btd box model identified all 9 human SP sequences. Thus, our models are both specific and sensitive enough to capture the intended domains. To identify the nine-amino-acid transactivation domain (9aaTAD), we utilized a prediction tool from http://www.med.muni.cz/9aaTAD/index.php (last accessed April 30, 2014), using the "moderately stringent pattern" option (Piskacek et al. 2007). Compositionally biased, low complexity regions (LCRs) associated with KLF/SPs were identified using four independent assessments: The CAST algorithm, http://athina.biol.uoa.gr/cgi-bin/CAST/cast_cgi (Promponas et al. 2000); ScanProsite, http://prosite.expasy.org/scanprosite/ (last accessed April 30, 2014) (de Castro et al. 2006); the SEG algorithm, http://mendel.imp.ac.at/METHODS/seg.server.html (last accessed April 30, 2014) (Wootton 1994); and manual curation based on criteria from Sim and Creamer (2004). A putative LCR was assigned if it was identified in at least two of the four methods (supplementary table S2, Supplementary Material online).

Additionally, we searched the entire genomes from our representative 48 eukaryotes for the following transactivation/repression domains in proteins excluding identified KLF/SP genes: Btd box, SP box, SID, PVDLS, R2, and R3 domain motifs. Perl scripts for each transactivation/repression domain model were run against the complete set of protein models from each genome (Supplementary Material online).

Domain Co-occurrence Networks

To visually represent the different domain architectures of the identified KLF/SP proteins, we created co-occurrence network maps. For each species network map, all unique domain combination pairs were identified, summed, and divided by the total number of proteins (KLFs only, SPs only, or both KLF/SPs) to obtain a percentage of co-occurrence for a particular domain pair. Composite network maps reflecting an organismal clade or grade were generated by the same procedure as for individual species maps. The size of the circles (domains) reflects how often specific domains appear relative to the KLF-DBD, which has 100% representation. Lines connecting two domains represent that unique combination pair. Line weights represent the relative frequency of that specific pair combination. A given domain pair combination was only counted once. Network maps were visualized using Microsoft PowerPoint.

Phylogenetic Analysis

Putative filozoan KLF/SP and nonfilozoan amorphean KLF-DBD amino acid sequences were aligned using default settings of the MUSCLE alignment package in CLCBio (Edgar 2004). To improve statistical support in our analyses, we concatenated the transactivation/repression domains with the KLF-DBD. The highly variable 9aaTAD and LCR sequences were excluded from the alignment. Individual aligned sequences ranged from 81 to 112 amino acids in length. Each domain, including the KLF-DBD, can be distinguished as separate blocks within the alignment (Supplementary Material online). For the nonfilozoan amorphean sequences, only the core three zinc fingers corresponding to the KLF-DBD model were included in the alignment. Duplicate sequences were removed and are listed in supplementary table S5, Supplementary Material online. ProtTest v2.4 was used to determine the LG+I+G model as the best-fit model for protein evolution (Abascal et al. 2005).

Maximum likelihood (ML) analyses were performed using the MPI version of RAxMLv7.2.8 (Stamatakis 2006). We executed 300 independent ML searches on randomized maximum parsimony starting trees using the standardized RAxML search algorithm, followed by comparison of likelihood values among all 300 resulting ML trees. The final log-likelihood score of the best ML tree was −12974.002279. One hundred bootstrap replicates were computed and applied to the best scoring ML tree. ML bootstrap values are indicated on the ML tree (supplementary fig. S2, Supplementary Material online). Bayesian analyses were performed with MrBayes3.2.5 (Ronquist and Huelsenbeck 2003). We ran two independent 5 million generation runs of five chains each with default heating and with the "LG+I+G" amino acid model option. The "average standard deviation of split frequencies" between the two runs was 0.048635. This diagnostic is an indicator of how well the two runs converge. A value below 0.01 is a strong indication of convergence, while a value between 0.01 and 0.05 is typically acceptable for convergence. Additional convergence diagnostics were examined with AWTY (Nylander et al. 2008), which was used to determine if a sufficient number of generations had been completed for posterior probabilities to stabilize, and to determine the amount of burn-in. From 60,001 trees, 45,001 were sampled (25% burn-in was confirmed as adequate with Tracer v1.6; Rambaut and Drummond 2007) and used to create a consensus tree. The runs reached stationarity, and adjusting the burn-in did not affect the topology of the tree. Bayesian posterior probabilities (BPP) were calculated and are shown on the Bayesian consensus tree (supplementary fig. S3, Supplementary Material online). FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/) was used for tree visualization.

Results

C2H2 Zinc Finger and KLF Identification

To better understand the evolution of the KLF/SP gene family and to gain insight into the genetic repertoire of transactivation/repression domains known to regulate disparate aspects of metabolism, and growth and development, we comprehensively searched for, identified, and characterized KLF/SP gene family complements from 48 eukaryotic genomes using a combination of hmms and custom perl models (table 1) with the assumption that the amorphean bikont split occurred at or near the origin of the Eukarya (Derelle and Lang 2012). We found C2H2 zinc finger proteins highly represented in all 48 eukaryotic genomes. The 81 amino acid KLF-DBD contains 3 highly conserved C2H2 zinc fingers separated by 2 highly conserved linker sequences (supplementary fig. S1*A*, Supplementary Material online). We found this domain architecture to be restricted to the opisthokont lineage (fig. 1). Among a small number of nonfilozoans, represented by *Sphaeroforma* and several holomycotans, sequences containing the KLF-DBD motif were also found to possess additional zinc fingers. Further analyses revealed the presence of a conserved aspartic acid residue in the second zinc finger of the KLF-DBD at position 44 (D_{44}) among 397 putative KLF/SP filozoan sequences, the single exception being a ctenophore gene. The D_{44} residue is critical for stabilizing proper KLF/SP DNA binding (Feng et al. 1994; Schuetz et al. 2011) and is notably absent from all nonfilozoan KLF-DBD

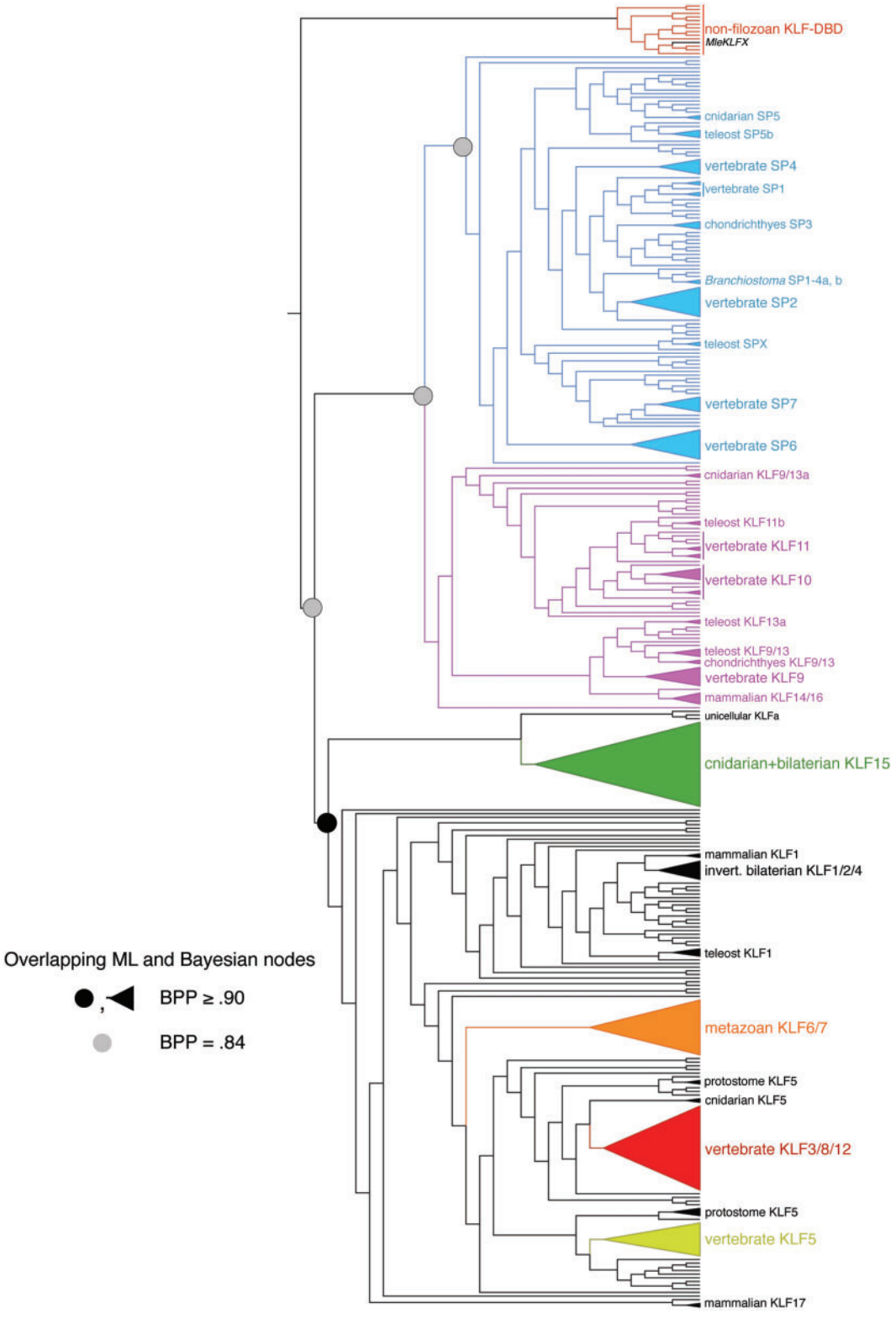

Fig. 2.—Combined gene tree estimates for the concatenated KLF/SP data set using Bayesian criterion (MrBayes) and ML criterion (RAxML). Gray node labels indicate congruent topology with BPP support = 84%. Black node labels indicate congruent topology with BPP support ≥90%. Clades collapsed to triangles indicate congruent topologies with BPP support ≥90%. The single highly divergent ctenophore *MleKLFX* sequence clusters with nonfilozoan KLF-DBD presumably due to long-branch attraction. Bayesian and ML trees with support values and branch lengths are available in supplementary figs. S2 and S3, Supplementary Material online.

Fig. 3.—Phylogenetic distribution of transactivation/repression domains and LCRs associated with KLF/SP proteins. The + indicates the presence of the corresponding domain or LCR in at least one KLF/SP protein in the indicated taxa. Only filozoan lineages containing bona fide KLF/SP proteins are shown. An asterix indicates that RNA-seq data were used for that species. Phylogeny is based on Dunn et al. (2008), Ryan et al. (2013), and Sebé-Pedrós et al. (2013).

motifs and the ctenophore sequence *MleKLFX* (supplementary fig. S1, Supplementary Material online). Our phylogenetic analyses consistently recover a moderately supported clade that represents the 397 filozoan genes possessing the canonical D_{44} aspartic acid residue in the second zinc finger of the KLF-DBD to the exclusion of the remaining 13 nonfilozoan sequences, along with *MleKLFX* (fig. 2). We then searched the human genome with a refined KLF-DBD model including the canonical D_{44} residue. We recovered the 26 human KLF/SPs plus Wilms Tumor 1 (WT1). The WT1 gene contains an extra zinc finger and is phylogenetically distinct (Shimeld 2008). Therefore the high stringency D_{44} KLF-DBD model is specific and sensitive. When the D_{44} KLF-DBD model is run against the eukaryotic genomes, we recover only putative KLF/SP protein sequences along with a small number of WT1 orthologs and WT1-like sequences that contain additional zinc fingers (supplementary fig. S1, Supplementary Material online).

Therefore, we operationally define bona fide KLF/SP orthologs as containing only 3 zinc fingers that conform to an 81 amino acid, 3 zinc finger KLF-DBD motif containing a canonical aspartic acid residue at position 44, D_{44} (supplementary fig. S1, Supplementary Material online). Thus in our analyses KLF genes first appear in the filasterean *Capsaspora* while the SP subfamily first appears near the base of the metazoan lineage in sponges and is absent from ctenophores (fig. 1). The full set of KLF/SPs identified in this study with their corresponding accession numbers or protein/transcript identifiers are

available in supplementary table S1, Supplementary Material online. Gene names given to each KLF/SP gene from this study can be found in supplementary table S4, Supplementary Material online.

Bikonts

C2H2 zinc finger proteins were identified in both unicellular and multicellular bikont species. However, the KLF-DBD was not present in any of the bikont genomes searched (fig. 1).

Unicellular Amorpheans (Unikonts)

A phylogenetically diverse range of unicellular species, including amoebas and fungi, were used in this study. Although C2H2 zinc finger proteins were identified in all unicellular amorpheans, the KLF-DBD was only found in *Fonticula* (sister to fungi), fungi (with the exceptions of *Mortierella* and *Encephalitozoon*), the ichthyosporean *Sphaeroforma*, the filasterean *Capsaspora*, and choanoflagellates (fig. 1). Among unicellular taxa bona fide KLF genes were only found in *Capsaspora* and choanoflagellates, two sister groups to the Metazoa (fig. 1).

Metazoans

Within the metazoans C2H2 zinc finger proteins along with the KLF-DBD were present in all species. However, neither representative ctenophore, *Mnemiopsis* or *Pleurobrachia*, show evidence for SP genes and both have fewer KLF genes as compared with other nonbilaterian metazoans. Sponge, placozoan, and cnidarian KLF/SP complements are similar to bilaterian protostome lineages and early branching deuterostome lineages. The total complement of KLF/SP genes substantially increases within the jawed vertebrates (fig. 1) and teleosts have an average of 33 KLF/SPs, which is slightly higher than cartilaginous fishes (22) and tetrapods (25).

Transactivation/Repression Domains

The defining characteristic of the KLF/SP family, the highly conserved KLF-DBD, is located near the C-terminal region of nearly all identified KLF/SP proteins (supplementary figs. S1 and S4, Supplementary Material online). In contrast, N-terminal regions of KLF/SP proteins are generally less conserved. These more variable N-terminal regions encode a variety of transactivation/repression domains, some of which have been well characterized in mammalian model systems (Suske et al. 2005; McConnell and Yang 2010). We generated custom perl script models for the following transactivation/repression domains to facilitate their identification across filozoan KLF/SP proteins (Supplementary Material online): SID repressor domain, PVDLS repressor domain, R2/R3 repressor domains, the Btd box, and the SP box. We screened for the presence of a conserved 9aaTAD (Piskacek et al. 2007). We also identified several nonrandom LCRs implicated in transactivation/repression that are highly biased for a particular amino acid residue (Wootton and Federhen 1993). After determining the distribution of transactivation/repression domains within KLF/SPs (fig. 3), we extended our search to include the full set of 48 eukaryotic genomes used in the study to better understand the representation of these domains in other proteins and to determine the extent domain shuffling may have played in the expansion of the KLF/SP gene family (Supplementary Material online).

Low Complexity Regions

LCRs are nonrandom regions of protein sequences that are highly biased for a particular amino acid residue (Wootton and Federhen 1993). LCRs are commonly found in transcription factors (Faux et al. 2005), have been shown to influence transcriptional regulation (Gerber et al. 1994), and are typically c-LCRs located centrally within the protein (Coletta et al. 2010). The composition of LCRs typically found in KLF/SP proteins includes serine/threonine (S/T)-rich, glutamine (Q)-rich, and proline (P)-rich regions. S/T-rich regions are generally associated with enhanced transcriptional activation, whereas P-rich regions have been associated with transcriptional repression (Hanna-Rose and Hansen 1996). Q-rich regions, which are more frequently found among members of the SP subgroup, are known to interact with $TAF_{II}110$ to activate transcription (Hoey et al. 1993; Gill et al. 1994).

We used four different algorithms (see Materials and Methods) to identify putative LCRs present in filozoan KLF/SP protein sequences (supplementary table S2, Supplementary Material online). We further required a putative LCR to be detected by a minimum of two methods for annotation as an LCR. S/T-rich LCR regions occur most frequently and are found in at least one KLF/SP family member in all filozoan taxa except the poriferan *Oscarella* (fig. 3). P-rich LCR regions have lower representation among KLF/SPs, and are found in all filozoans except for *Ciona*, *Daphnia*, *Tribolium*, *Lottia*, *Trichoplax*, *Oscarella*, and *Monosiga* (fig. 3). KLF/SPs with Q-rich LCR regions were present in all filozoans except *Ciona*, *Tribolium*, *Capitella*, *Lottia*, cnidarians, *Oscarella*, ctenophores, and *Monosiga* (fig. 3).

Nine-Amino-Acid Transactivation Domain

The 9aaTAD, first identified in yeast transcription factors (Piskacek et al. 2007), is a short motif highly conserved throughout eukaryotes. The 9aaTAD has been shown to interact with TAF9 of the RNA polymerase II holoenzyme and this domain motif has been identified in many transcription factors (Piskacek 2009). Using a 9aaTAD prediction tool (Piskacek et al. 2007), we found corresponding motifs in all filozoan KLF/SPs except *Capitella*, *Lottia*, and *Salpingoeca* (fig. 3).

Buttonhead Box and SP Box

The Btd box and SP box are conserved domains that, in combination with the KLF-DBD, characterize the SP subfamily. The Btd box was first identified in the *Drosophila* gene *buttonhead* (Wimmer et al. 1993) and the associated SP box was subsequently discovered in SP5 (Harrison et al. 2000). The Btd box motif is typically found just N-terminal of the KLF-DBD, while the SP box motif is located proximal to the N-terminus of the protein. Although not definitive, there is evidence suggesting that the Btd box is involved in transactivation (Athanikar et al. 1997). We examined filozoan KLF/SP sequences for the presence of the Btd and SP boxes. Both Btd and SP box motifs are absent in unicellular KLFs, absent in both ctenophore KLFs, but present in all other metazoan phyla (fig. 3). Uniquely within the poriferans, the SP box was not found in any *Amphimedon* KLF/SP, but was identified in *Oscarella* KLF/SPs.

Using the same domain models to screen the complete set of 48 eukaryotic whole genomes, we identified Btd box motifs in a number of genes other than the KLF/SP family with the notable exception of the fungi, *Encephalitozoon* and *Saccharomyces* (Supplementary Material online). In contrast, the distribution of the SP box motif is highly restricted. This domain motif first appears coincident with the SP subfamily in poriferans. The SP box is present in one out of two SP genes in *Oscarella* but is not detected in the two SP genes identified in the *Amphimedon* genome. Excluding the ctenophores, metazoan genomes outside the poriferans have SP box domain motifs among a small set of genes to the exclusion of the SPs (fig. 6).

Sin3a Interacting Domain

The Sin3a protein acts as a transcriptional repressor, and is able to recruit and bind histone deacetylases (Laherty et al. 1997; Silverstein and Ekwall 2005). Mammalian KLF9, 10, 11, 13, 14, and 16 are known to interact with the Sin3a protein through a Sin3a interacting domain (SID) which binds the PAH domain of the Sin3a protein (Imataka et al. 1992; Blok et al. 1995; Cook et al. 1998; Song et al. 1999; Kaczynski et al. 2002). No SID motif was detected in unicellular KLF/SP sequences. Within metazoans a SID-containing KLF was identified in all taxa except ctenophores and the protostomes *Capitella*, *Lottia*, and *Drosophila* (fig. 3). To gain further insight into the evolutionary history of the SID motif, we searched the complete set of 48 whole genomes and identified a conserved SID in a number of genes in all representative eukaryote genomes (Supplementary Material online).

R2 and R3 Domains

KLF10 and KLF11 proteins contain R2 and R3 repressor domains in combination with an SID motif (Cook et al. 1999). We searched filozoan KLF/SP complements and were only able to identify R2 and R3 domains in representative vertebrate taxa (fig. 3). We then searched the complete set of 48 whole genomes for the presence of the R2 and R3 domains. Our search revealed the presence of an R3 domain in all representative eukaryotic genomes, while the R2 domain was restricted to vertebrate KLF10/11 genes (Supplementary Material online).

PVDLS Domain

KLF3, KLF8, and KLF12 can corepress transcription through interaction with the C-terminal binding protein mediated by the PVDLS domain (Crossley et al. 1996; Turner and Crossley 1998; Imhof et al. 1999; van Vliet et al. 2000). A PVDLS domain was identified in at least one KLF/SP in all jawed vertebrates and in *Drosophila* (fig. 3). To help resolve this relationship between the fly PVDLS containing KLF and the vertebrate KLF3/8/12 genes, we searched additional protostome genomes for the presence of KLF genes containing the PVDLS motif. Curiously, the association of this motif with KLF genes was not detected in other Drosophilid species, but was identified in two hymenopterans *Apis millifera* and *Nasonia vitripennis* (data not shown). Analysis of the complete set of 48 whole genomes revealed the PVDLS domain in multiple genes in all representative eukaryote species (Supplementary Material online).

Co-occurrence Networks

Co-occurrence networks are visual representations of domain pair occurrences within a given protein or protein family. Typically, domains are represented as circles, and lines connect domains that appear together within a protein. These maps can be useful for visualizing the frequencies of certain domains occurring with each other, that is, domain architectures. To explore differences in domain architecture complexity, we generated domain co-occurrence maps for different KLF/SP domain networks (fig. 4). These maps can also show the general N-terminal to C-terminal relationship between different domains (fig. 4*A*). Our co-occurrence maps indicate the frequency of unique domain pairs as well as how often an individual domain appears within a given network in extant species (fig. 4*B*–*I*). Composite networks of larger taxonomic groupings represent the consensus map of all included species. The composite networks are additive and thus do not compensate for missing data due to poor genome annotations. Importantly, these co-occurrence maps are not intended to be ancestral reconstructions. They are used to show observed domain relationships in, or among groupings of, extant species. Nonetheless, evolutionary inferences can be drawn from these network maps when combined with inferences of ancestral presence or absence of discrete transactivation/repression domains as represented in figure 6.

The unicellular KLF/SP network (*Capsaspora* + choanoflagellates) is one of the least complex with only LCR domains linked to the KLF-DBD (fig. 4*C*). Within metazoans, there

appears to be a gradient of network complexity; the nonbilaterian network (fig. 4*D*) is less complex than the invertebrate bilaterian network (fig. 4*E*), and both of these are less complex than the vertebrate network (fig. 4*F*). A small number of lineage-specific networks showed a significant departure from larger composite networks. For example, the Ctenophoran network (fig. 4*G*), in contrast to both the Poriferan network (fig. 4*H*) and larger nonbilaterian composite network (fig. 4*D*), is composed of only S/T-rich and P-rich LCRs, with the P-rich LCRs + KLF-DBD domain pair occurring at greater frequency. The Poriferan network (fig. 4*H*) bears substantially greater similarity to the more inclusive nonbilaterian composite network (fig. 4*D*). The urochordate *Ciona* also presents an interesting departure from composite networks (fig. 4*I*), diverging dramatically from the invertebrate bilaterians in having S/T-rich LCRs as the most prevalent domain linked to the KLF-DBD. The relationship of Ctenophora and Porifera to other metazoans shown in figure 4*J* and *K* are based on recent hypotheses of metazoan phylogeny (Dunn et al. 2008; Ryan et al. 2013).

KLF Domain Architecture and Phylogenetics

All KLF/SP gene family members share homology at the conserved KLF-DBD, typically located toward the C-terminus. The N-terminal regions share very little similarity across the entire family. Distinct KLF/SP subgroups can, however, be defined based on comparable structure and function (McConnell and Yang 2010). These subgroups share domain architectures defined by unique combinations of transactivation/repression domains occurring with the KLF-DBD (fig. 5; supplementary table S3, Supplementary Material online). Domain architectures have been described for mammalian KLF/SP genes (Kaczynski et al. 2003; Suske et al. 2005; McConnell and Yang 2010; Archer 2011). Historically, KLFs have been divided into five domain architectures and named according to the genes in each class: KLF1/2/4, KLF6/7, KLF3/8/12, KLF9/13, and KLF10/11. Similarly, the SPs have been divided into two architecture classes: SP1–4 and SP5–9. Using this existing classification scheme as a foundation, we determined the distribution of domain architectures for all filozoan KLF/SP genes. The 9aaTAD was not used in determining domain architecture as it does not contribute to any unique domain combination.

We performed ML and Bayesian phylogenetic analyses on an alignment of nonfilozoan amorphean and filozoan concatenated sequences that included the KLF-DBD and transactivation domains (supplementary figs. S2 and S3, Supplementary Material online). The overall topologies of the ML and Bayesian trees share significant overlap, albeit with only moderate support for deeper nodes, for example, BPP = 0.84 (fig. 2). A number of informative clades with congruence across both ML and Bayesian analyses were recovered. We primarily focused on overlapping nodes with BPP $\geq$ 0.90. Most of these nodes are consistent with explicit domain architectures and well-accepted inferences of metazoan phylogeny. Many of the well-supported orthologous sequence clades are composed of vertebrates with corresponding nonvertebrate putative orthologs diverging prior to the highly supported vertebrate nodes (fig. 2; supplementary figs. S2 and S3, Supplementary Material online).

Among the unicellular KLF representatives, the *Capsaspora* and *Salpingoeca* domain architectures consist of alternating S/T- and Q-rich LCRs (fig. 5) with scattered P-rich LCRs at low frequency (fig. 4*C*). Although the S/T- and Q-rich LCR organization superficially resembles the architecture found among SP1–4 genes, these unicellular KLFs lack the characteristic Btd box and SP box motifs that define the SPs. Therefore, we created a unicellular specific KLF architecture class (fig. 5). Although a single unicellular KLF clade composed of *CowKLFa*, *SroKLFa*, and *MbrKLFa* branch sister to a highly supported cnidarian + bilaterian KLF15 clade (BPP = 0.99), the remaining unicellular KLF sequences are divergent and characterized by long branches (fig. 2; supplementary figs. S2 and S3, Supplementary Material online).

The KLF6/7 and KLF1/2/4 classes are defined by the presence of mostly S/T-rich or P-rich LCRs, respectively. KLF6/7 domain architectures were identified in *Monosiga* and all metazoans except for *Oscarella* and *Lottia* (fig. 5). The single *Monosiga* KLF gene superficially consists of a KLF6/7 domain architecture; however, phylogenetically this sequence groups with other unicellular KLFs. *Amphimedon* sequences with KLF6/7 architecture grouped more closely with KLF1/2/4 and KLF5, albeit with low statistical support. Our phylogenetic analysis recovered a well-supported clade (BPP = 0.91) that includes bilaterian KLF6/7 along with representative cnidarian and *Trichoplax* KLF6/7. The KLF1/2/4 class was identified in all metazoans except *Oscarella*, *Trichoplax*, *Lottia*, *Drosophila*, *Daphnia*, and *Ciona* (fig. 5). Our phylogenetic analyses recovered a KLF 1/2/4 clade including invertebrate bilaterians (BPP = 0.96), a teleost-specific KLF1 clade (BPP = 1), and a mammalian KLF1 clade (BPP = 1; fig. 2).

The KLF9/13 class possesses a SID near the N-terminal region and typically contains one or more c-LCRs located between the SID and KLF-DBD. The related KLF10/11 class shares the SID with KLF9/13, but uniquely contains the R2 and R3 repressor domains situated between the SID and the KLF-DBD. The R2 and R3 repressor domains are generally flanked by one or more P-rich or S/T rich c-LCRs. Based on phylogenetic analyses, all metazoans, with the notable exception of ctenophores, have at least one KLF9/13 gene (supplementary figs. S2–S4, Supplementary Material online). Mammalian KLF14 and KLF16 form a unique, highly supported (BPP = 1) clade within the 9/13 group supporting the assignment of these orthologs to a mammalian-specific lineage (fig. 2; supplementary fig. S5, Supplementary Material online). The KLF10/11 class has a more restricted distribution and is exclusively found in vertebrates (fig. 5). All SID containing KLF sequences

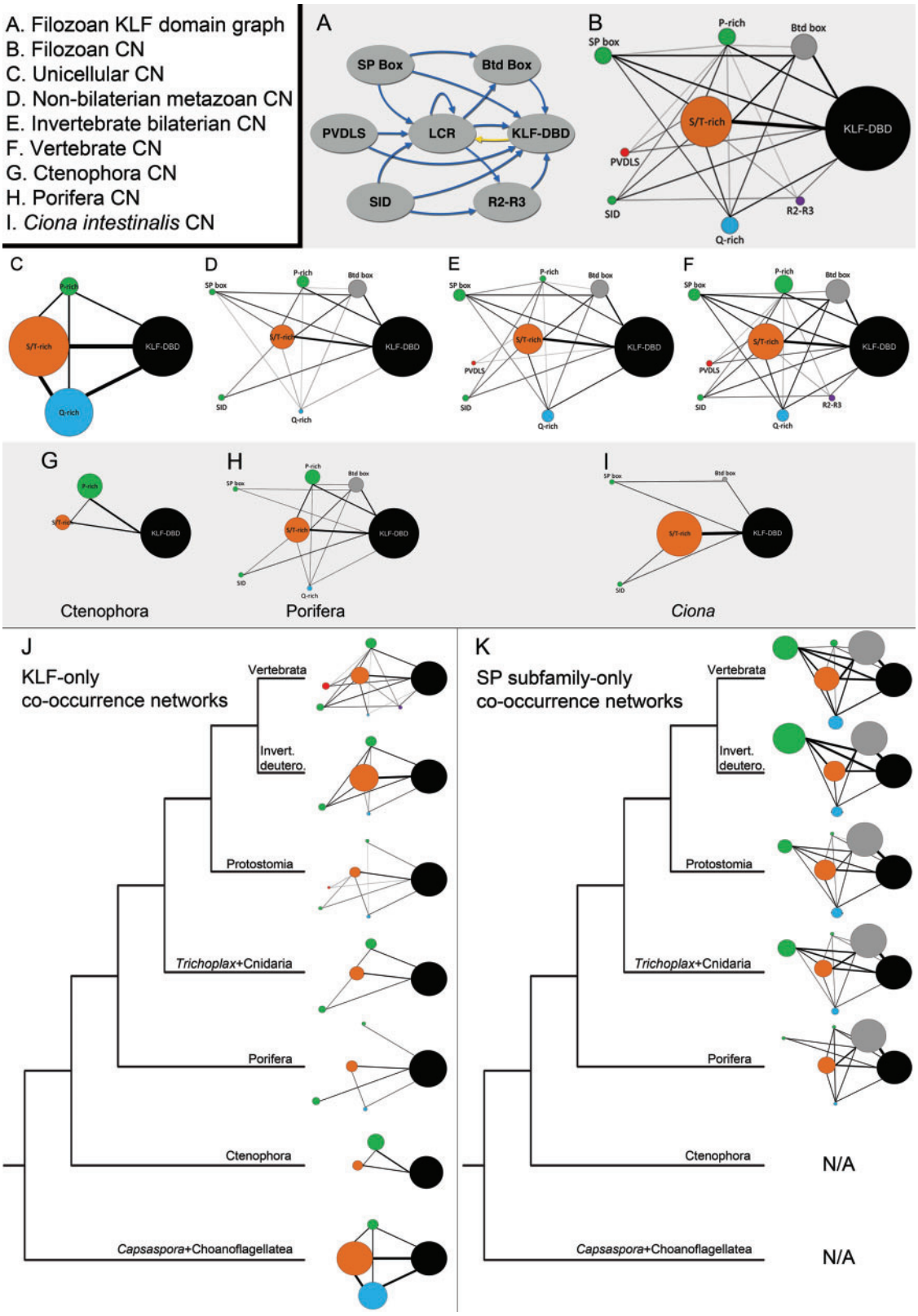

Fig. 4.—KLF/SP protein domain co-occurrence networks. In all networks, each circle represents a transactivation/repression domain or an LCR. A line connecting two domains indicates a co-occurrence of those two domains. Domains are arranged in approximately the same 5′–3′ spatial orientation as they appear encoded in KLF/SP sequences. (*A*) General network diagram showing connectivity and unidirectional spatial relationships between transactivation

form a clade with low support (BPP = 0.75) (supplementary figs. S2 and S3, Supplementary Material online).

Finally, the KLF3/8/12 class is characterized by a PVDLS domain and often has additional c-LCRs between the PVDLS and the KLF-DBD. According to domain architecture, this class is found in all the jawed vertebrates and in the Endopterygota as represented in this study by *Drosophila* (fig. 5). Our phylogenetic analyses recover a KLF3/8/12 clade with high support (BPP = 0.95; fig. 2). Notably, the *Drosophila* KLF sequence that contains a PVDLS domain (*CG42741*) falls outside of the KLF3/8/12 clade which consists exclusively of vertebrate sequences. Within this clade the KLF8 orthologs are recovered with BPP = 1. Despite poor annotation in the lamprey genome, we identified a lamprey sequence within the vertebrate KLF3/8/12 clade (supplementary figs. S2 and S3, Supplementary Material online).

The SP subfamily is defined by the presence of a Btd box 5′ proximal to the KLF-DBD and generally possesses an additional SP box located near the N-terminus. The SP factors can be further separated into two domain architecture classes, SP1–4 and SP5–9 (Bouwman and Philipsen 2002; Archer 2011). The SP1–4 class typically has S/T-rich c-LCRs adjacent to Q-rich c-LCRs. The SP5–9 class is characterized by having either predominately S/T, or with lesser frequency, P-rich c-LCRs. Notably, the SP subfamily is absent in both ctenophore genomes. The SP1–4 domain architecture is found in all metazoans except *Oscarella*, cnidarians, *Capitella*, *Lottia*, *Tribolium*, and *Ciona*. The SP5–9 domain architecture is found in all metazoan genomes excluding ctenophores, *Drosophila*, and *Daphnia* (fig. 5). Phylogenetic analyses recover moderate support (BPP = 0.84) for a metazoan SP clade (fig. 2; supplementary figs. S2 and S3, Supplementary Material online). Well-supported SP clades (BPP $\geq$ 0.9) include vertebrate SP2, SP6, and SP7.

Discussion

KLF/SP Gene Family Origins in Eukarya

The evolution of the KLF/SP family and associated transactivation/repression domains is represented in figure 6. Although C2H2 zinc finger domains are ubiquitous across a wide range of eukaryotes (de Mendoza et al. 2013), the KLF-DBD first appears in the Holomycota (figs. 1 and 6). Historically, the single criteria for defining the KLF/SP gene family has been the presence of a highly conserved KLF-DBD composed of three C2H2 zinc fingers each separated by a seven amino acid linker region (McConnell and Yang 2010). The first 2 zinc fingers are 23 amino acids long (from C to H), while the third zinc finger is only 21 amino acids long. Our analyses reveal that all filozoan KLF/SP sequences recovered except one, the highly divergent *MleKLFX* ctenophore sequence, contains a canonical aspartic acid residue in the second zinc finger, D_{44}, which is critical for proper DNA binding (supplementary fig. S1, Supplementary Material online) (Feng et al. 1994; Schuetz et al. 2011). All nonfilozoan KLF-DBD containing genes lacked this diagnostic residue and possess additional zinc fingers. Furthermore, none of the transactivation/repression domains considered in this study were found in the nonfilozoan KLF-DBD containing sequences, despite several relevant domains being well represented in all eukaryotic genomes examined (fig. 6, Supplementary Material online). Thus the nonfilozoan KLF-DBD sequences were not classified as bona fide KLF/SP orthologs. Moreover, our phylogenetic analyses recovered a moderately supported filozoan clade that possesses the canonical D_{44} residue to the exclusion of all nonfilozoan sequences and the highly divergent *MleKLFX* ctenophore sequence (fig. 2). We operationally define a bona fide KLF/SP gene as containing only 3 zinc fingers conforming to 81 amino acid, 3 zinc finger KLF-DBD motif typically found in the C-terminal region of the parent sequence and containing a canonical aspartic acid residue at position 44, D_{44}, following the general consensus sequence C-X_4-C-X_{12}-H-X_3-H-X_7-C-X_4-C-X_7-**D**-X_4-H-X_3-H-X_7-C-X_2-C-X_{12}-H-X_3-H.

Therefore, the KLF-DBD likely has its origins in the opisthokont stem lineage prior to the divergence of the Holomycota (figs. 1 and 6). In our study, KLF genes first appear in the filasterian *Capsaspora*, while the SP subfamily is restricted to metazoans, notably excluding the ctenophores. Thus we infer the origin of KLF genes in the filozoan stem lineage and a later

Fig. 4.—Continued

domains among filozoan KLF/SPs. Blue arrows represent connectivity upstream of the KLF-DBD; the gold arrow represents connectivity downstream of the KLF-DBD. (*B–I*) KLF/SP co-occurrence networks from different taxonomic groups. Circle size indicates the relative frequency of occurrence in the network, with the KLF-DBD always representing 100%. Circle color follows the same convention as seen in figure 3. Repeated domains were counted as occurring only once. Lines connecting circles indicate the presence of that specific domain pair co-occurrence in at least one KLF/SP. Line width indicates the frequency of domain pair co-occurrence. Only LCR domains which are found N-terminal of the KLF-DBD are represented in these networks (supplementary fig. S4, Supplementary Material online). (*B*) Complete filozoan KLF/SP network. (*C*) Representative unicellular KLF/SP network. (*D*) KLF/SP network from nonbilaterian metazoans. (*E*) Invertebrate bilaterian KLF/SP network. (*F*) Vertebrate KLF/SP network. (*G-H*) Representative ctenophoran and poriferan KLF/SP networks for comparison with each other and with the network in *D*. (*I*) *Ciona* KLF/SP network for comparison with the networks in *E* and *F*. (*J*, *K*) Co-occurrence network maps for the KLF subfamily and SP subfamily mapped onto the filozoan phylogeny (Dunn et al. 2008; Ryan et al. 2013) for evolutionary comparison. Each network represents a composite for the taxonomic group indicated. (*J*) Co-occurrence maps for domains found in the KLF subfamily. (*K*) Co-occurrence maps for domains found in the SP subfamily. The unicellular filozoan genomes and ctenophore genomes do not contain SP genes.

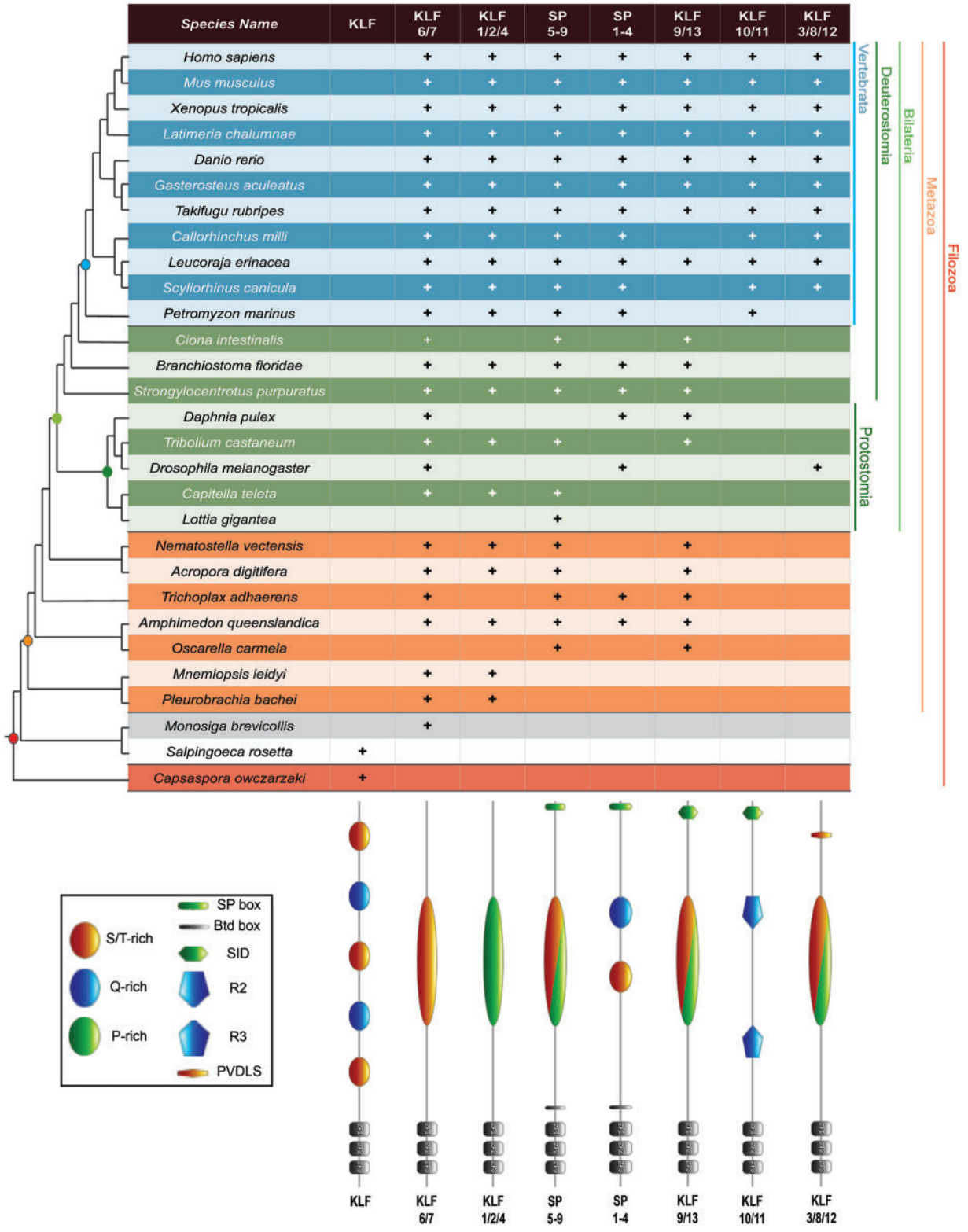

Fig. 5.—Phylogenetic distribution of explicit domain architectures represented among KLF/SP proteins. The key at lower left identifies LCRs and transactivation/repression domains used to determine domain architectures. The protein schematics along lower right represent the particular combinations of domains and LCRs with the KLF-DBD that define each specific KLF/SP protein architecture. All groups, except for the ancient unicellular KLF architecture recovered, are named according to established human KLF/SP paralogy groups that conform to each specific architecture. The three C-terminal zinc fingers of the KLF-DBD are indicated with grey boxes labeled zf1, zf2, and zf3. Architecture schematics are not to scale.

origin of the SP genes after the divergence of the metazoan lineage (figs. 1 and 6). Our phylogeny assumes an early divergence of the ctenophore lineage from the metazoan stem (Dunn et al. 2008; Ryan et al. 2013). An alternate view of early animal phylogeny (Philippe et al. 2009; Nosenko et al. 2013) would infer the origin of SP genes prior to the divergence of the poriferans and assume a subsequent loss of SP orthologs in the ctenophore lineage. Under either scenario of early metazoan lineage divergence, the appearance of KLF genes precedes the origin of the SP gene subfamily.

Apart from the KLF-DBD, the most common sequence features are the presence of one or more S/T-, P-, or Q-rich c-LCRs (figs. 3–5; supplementary fig. S4, Supplementary Material online). Our results highlight promiscuous variation in LCR composition and length between KLF/SP gene family members (supplementary fig. S4, Supplementary Material online). In contrast to earlier studies with more restricted sampling, we are unable to discriminate phylogenetic relationships between KLF/SP architectural classes composed of explicit c-LCR compositional types linked to the KLF-DBD. For example, many representative genes nested within the KLF1/2/4 and KLF6/7 clades include all three predominant c-LCR compositional types (figs. 2 and 5; supplementary fig. S4, Supplementary Material online). Variability, expansion, and extinction of LCRs have been associated with gene conversion due to mismatch repair of DNA heteroduplexes (Radding 1982) and higher rates of recombination relative to flanking sequences due to unequal crossing over and replication slippage (DePristo et al. 2006). Both of these core genetic mechanisms would contribute to our observations of significant LCR homoplasy within and between orthologous domain architectures.

Our analyses of the frequency of appearance of explicit compositionally biased c-LCRs and their frequency of co-occurrence with flanking transactivation domains suggest functional consequences. For example, S/T- and Q-rich LCRs are typically acidic and associated with transcriptional activation, whereas P-rich regions, with their bulky side chains, are typically neutral and associated with context-dependent repression. The high frequency of S/T-rich c-LCRs suggests an ancestral KLF that was likely involved in transcriptional activation. Despite high rates of gene conversion and unequal crossover associated with LCRs, we observe P-rich c-LCRs at low frequency. Conversely, among the KLF/SP genes that retain P-rich c-LCRs, our analyses also show an increase in the frequency of connectivity with other repressor motifs in metazoans (fig. 4). This suggests the evolution of coordinated multidomain repression during the expansion of the KLF/SP gene family. For example, known repressor domains such as the SID and PVDLS motifs are typically located 5′ of c-LCRs and show a similar pattern of increased frequency during metazoan lineage diversification (figs. 4 and 6). This pattern suggests that within the KLF/SP gene family, despite a selective preference for acidic c-LCRs as evidenced by their high frequency, the evolution of a repertoire of transcriptional repressor combinations through the differential pairing of c-LCRs with repressor motifs occurred. Our analyses further suggest that these shuffling events may be associated with cell type diversification in metazoans. For example, chordates possess >100 distinct cell types, whereas earlier diverging metazoan lineages range from 4 to 59 distinct cell types (Chen et al. 2014). This pattern mirrors the increased frequency of repressor domain connectivity observed across members of the KLF/SP gene family in metazoans (figs. 4–6).

The total complement of KLF/SP genes across nonbilaterian metazoans, protostomes, and invertebrate deuterostomes shows only a ~2-fold variance of 6–12 genes. Vertebrates have an average number of 24 KLF/SPs, likely due to the 2 rounds of whole genome duplication (WGD) in the vertebrate stem lineage (supplementary fig. S5, Supplementary Material online). Despite the additional round of WGD in the teleosts, their KLF/SP gene complement is not substantially greater than other jawed vertebrates. It seems that KLF/SP gene duplicates were not necessarily preferentially retained, which has been shown to be the case with other transcription factors (de Mendoza et al. 2013). The lamprey *Petromyzon marinus* has an atypically depauperate KLF/SP complement as compared with other vertebrates. It has been shown that during embryogenesis lampreys can lose ~20% of germline DNA in somatic tissues due to genomic rearrangements (Smith et al. 2009, 2012). The lamprey genome was derived completely from somatic tissue, thus the true *P. marinus* KLF/SP complement may be underrepresented in our analyses (Smith et al. 2013).

Ctenophores have a reduced KLF/SP complement compared with other nonbilaterian metazoans and lack members of the SP subfamily altogether (fig. 1). This pattern of gene underrepresentation in ctenophores relative to other metazoans has been observed in a number of studies (Ryan et al. 2010; Maxwell et al. 2012; Moroz et al. 2014; Sebé-Pedrós et al. 2013). Among the filozoan KLF orthologs, a single member, the highly divergent ctenophore gene *MleKLFX*, possesses atypical domain architecture in which a KLF-DBD lacking the canonical D_{44} residue is located in the N-terminal region of the protein instead of the C-terminal region (supplementary fig. S4, Supplementary Material online). This gene consistently groups with the nonfilozoan amorphean KLF-DBD sequences presumably due to the effects of long-branch attraction (fig. 2; supplementary figs. S2 and S3, Supplementary Material online).

The Btd box is found in SPs and along with the KLF-DBD represents the defining characteristic of the SP subfamily. The Btd box was initially described in the *Drosophila btd* gene which has high similarity to mouse SP1 and SP3 (Wimmer et al. 1993). Most of the amino acid sequence conservation resides in the zinc fingers, and *btd* is often classified as an SP homolog in flies. However, the second zinc finger of the *btd* gene is slightly different than the corresponding zinc finger in the KLF-DBD. The second zinc finger of *btd* is only 21 amino

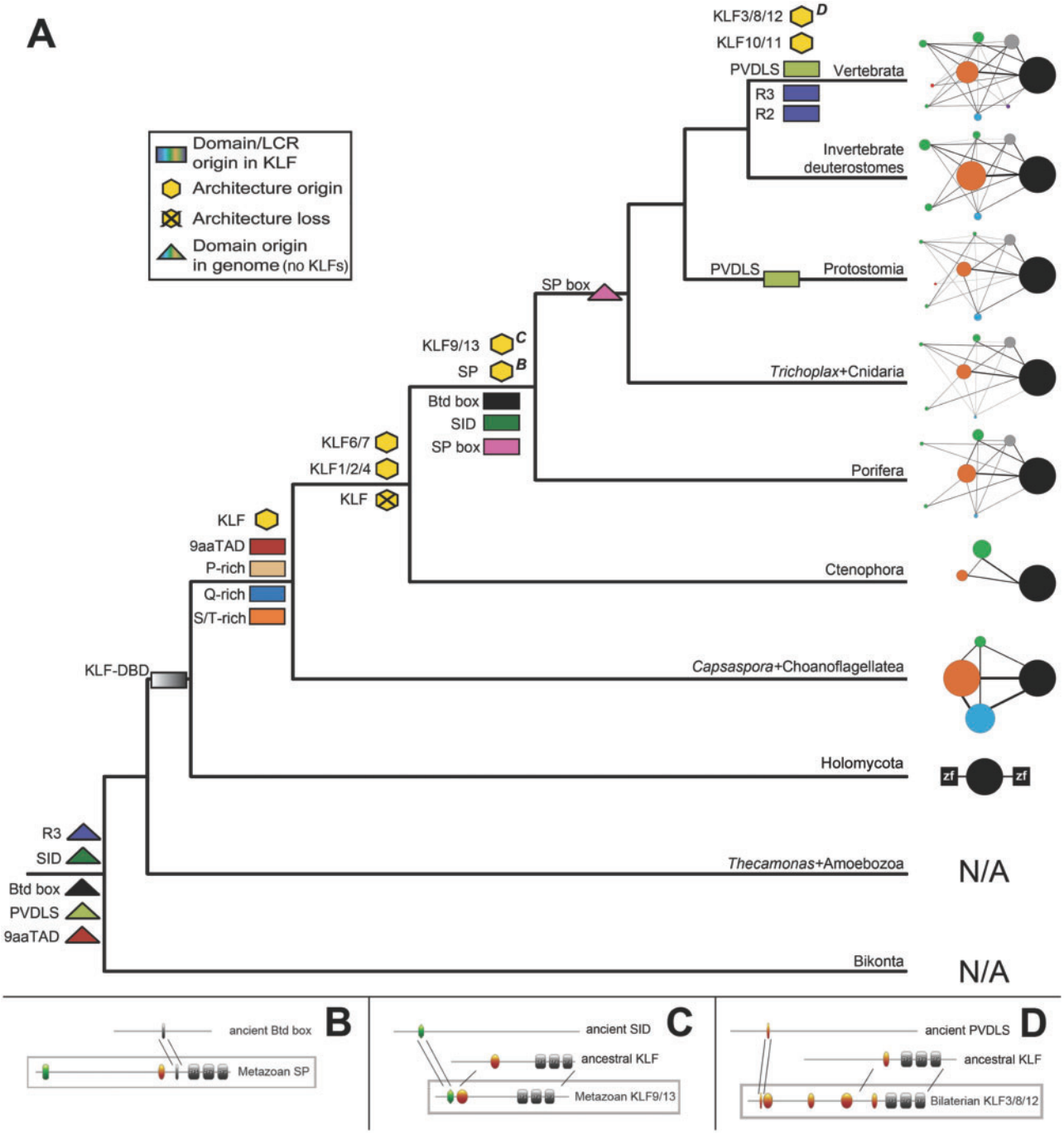

Fig. 6.—Inferred relationships between key events during the evolution and expansion of the KLF/SP gene family. Symbol key is at upper left. Colored rectangles represent the origin of particular transactivation/repression domains or LCRs co-occurring with the KLF-DBD (fig. 4). Yellow hexagons represent the origin of specific KLF/SP domain architectures (fig. 5). A black X over a hexagon represents the loss of specific domain architecture. Colored triangles represent the presence of specific transactivation domain motifs within whole eukaryote genomes to the exclusion of the KLF/SP gene family. (*A*) We infer the origin of the KLF-DBD in the opisthokont stem lineage prior to the divergence of the Holomycota. However, bona fide KLF gene architectures do not appear until the divergence of the filozoan lineage (KLF origin). The ancient unicellular KLF domain architecture is not recovered in metazoan lineages. The ancient PVDLS, SID, Btd box, and R3 domains were recovered, to the exclusion of KLF/SPs, in all eukaryote genomes searched. Notably, the Btd box was not recovered in *Saccharomyces* and *Encephalitozoon* fungal genomes. Our analysis suggests that the origin of the SP subfamily is in the metazoan stem lineage prior to the divergence of the poriferans; it is not present in the ctenophorans. The SP box motif only appears in SP genes in poriferans and is not found in additional genes until the divergence of *Trichoplax*. The R2 repressor domain appears to be a de novo innovation restricted to KLF genes in the vertebrate stem lineage, contributing to the KLF10/11 architecture class. Composite domain co-occurrence maps for each taxonomic group are shown to the right of

acids long, lacking 2 residues between the 2 C residues. According to this structural difference, *btd* would be classified in the Zif268/EGR-1 family of zinc finger transcription factors (Iuchi 2001). However, phylogenetic analysis of *btd* shows that it is more similar to SPs than to EGR-1 family genes and *btd*-like orthologs were not found in other organisms (data not shown). These suggest that at some point along the stem lineage leading to *Drosophilia*, there was a lineage-specific deletion of 2 amino acid residues in the second zinc finger of *btd*. Notably, the loss of these two residues does not hinder *btd* from binding transcription factor binding sites similar to SP factors (Wimmer et al. 1993).

Recently, three SPs in the cnidarian *Nematostella* and the placozoan *Trichoplax* were identified and grouped into three separate clades: SP1–4, SP5/*btd*, and SP6-9 (Schaeper et al. 2010) leading to a proposed ancestral complement of three SPs in the Metazoa. In our study, we recovered two SPs in sponges, two SP genes in *Trichoplax*, and four SP genes in *Nematostella*. The *Trichoplax btd* gene and SP6–9 gene from Schaeper et al. (2010) match with two *Trichoplax* SPs from our study (supplementary table S1 and supplementary fig. S4, Supplementary Material online). However, the *Trichoplax* SP1–4 gene from Schaeper et al. (2010) matches with the *Trichoplax* KLF9/13 gene in our study. We were able to extend this *Trichoplax* sequence and identified a SID motif in the previously unannotated 5′ end of this gene model thus confirming the placement of this sequence within the KLF9/13 group (supplementary fig. S4, Supplementary Material online). Therefore, our analyses suggest an ancestral complement of a single SP1–4 ortholog and a single SP5–9 ortholog in the early metazoan stem lineage followed by an expansion after the divergence of the Placozoa (fig. 1).

Our analyses also suggest that two KLF9/13 genes were present early in Metazoa (figs. 5 and 6; supplementary figs. S2 and S3, Supplementary Material online). In mammals, two additional genes with similar domain architecture, KLF14 and KLF16, are present and a recent study provided evidence that KLF14 evolved from KLF16 through a retrotransposon event (RTE) within the mammalian stem lineage (Parker-Katiraee et al. 2007). Thus, we infer that KLF16 arose by a tandem duplication of one of the two ancestral KLF9/13 genes in the mammalian stem lineage followed by KLF14 evolution by RTE of KLF16 (supplementary fig. S5, Supplementary Material online).

Transactivation/Repression Domains Show Unique Evolutionary History

Our results highlight that individual transactivation/repression domains associated with KLF/SP transcription factors have unique evolutionary histories (fig. 6). The Btd box, PVDLS, R3, and SID are ancient domain motifs present in other genes in all representative eukaryote genomes in this study. The notable exceptions being the *Saccharomyces* and *Encephalitozoon* fungal genomes in which no Btd box domains were detected. These four ancient domains first appear in differential combinations with the KLF-DBD in KLF/SP genes after the divergence of the metazoans and correlate strongly with the observed expansion of domain architecture repertoires (fig. 5). Given the phylogenetic distribution of domain architectures associated with the putative acquisition of ancient motifs by domain shuffling, we infer that particular domain architectures stem from unique independent ancestral shuffling events along the metazoan stem lineage (figs. 5 and 6).

Our analyses also infer two instances of de novo domain origin within the KLF/SP gene family (fig. 6). The SP subfamily appears early in metazoan evolution in the Poriferan lineage (fig. 1) coincident with the appearance of the SP box motif (fig. 3). However, in contrast to the Btd box motif, the SP box motif has a very limited genomic and phylogenetic distribution. In sponges, the SP box motif is uniquely associated with the SP genes. In later diverging lineages of metazoans, the SP box is associated with a small number of genes in addition to the KLF/SP family. Thus the SP genes provide an example of a metazoan-specific multidomain protein that consists of both ancient domains, including the Btd box and the KLF-DBD, coupled with the de novo origin of a metazoan-specific domain motif, the SP box. Another example of de novo domain motif origin within the KLF/SP family is the vertebrate-specific KLF10/11 genes (fig. 6; supplementary fig. S5, Supplementary Material online). This architectural class is composed of ancient SID, R3, and KLF-DBD domains combined with a vertebrate-specific R2 domain. Our exhaustive search uncovered no R2 domains in any invertebrate genome or outside of the KLF10/11 genes within the vertebrate genomes in this study. Interestingly, the R3 domain is present in all eukaryote genomes but there is no evidence for it being shuffled into KLFs until much later than the SID motif (fig. 6).

The PVDLS domain represents an intriguing case of putative convergence. In our analyses, the PVDLS domain appears in vertebrate KLFs defining the 3/8/12 architecture group and in

Fig. 6.—Continued

the tree. Representative examples of putative domain shuffling events during the evolution and expansion of the KLF/SP gene family. (*B*) An ancient Btd box and a metazoan SP gene may have contributed to the origin of the SP gene subfamily early in metazoan evolution. (*C*) An ancient SID likely combined with a pre-existing ancestral KLF gene to form the KLF9/13 group, also early in metazoan evolution. (*D*) An ancient PVDLS domain combined with a pre-existing ancestral KLF gene to form the KLF3/8/12 group. We infer an independent convergent acquisition of the PVDLS domain within a KLF gene in the Protostomia lineage (see Discussion). Domain icon colors are the same as figure 5.

a lone *Drosophila* KLF (*CG42741*). Our phylogenetic analyses suggest, however, that the fly sequence is not nested within the highly supported KLF3/8/12 clade (supplementary figs. S2 and S3, Supplementary Material online). To help elucidate the incongruous relationship between the representative fly gene containing a PVDLS motif and the PVDLS containing vertebrate KLF3/8/12 class, we searched a number of additional protostome genomes for the presence of KLF genes containing the PVDLS motif (data not shown). Our search yielded only two other instances, both within the hexapods, of a KLF gene also containing a PVDLS motif: The hymenopterans *A. mellifera* and *N. vitripennis*. Exhaustive searches of 12 other Drosophilid species did not uncover any PVDLS containing KLFs. Based on these additional results, we infer that a putative ancestral KLF3/8/12 gene most likely evolved early in the vertebrate stem lineage, whereas the PVDLS motif was likely also convergently acquired in the hexapod lineage leading to the Endopterygota.

Conclusions

Our analysis across 48 eukaryotic genomes illuminates the origin and evolutionary history of the KLF/SP gene family. We also identify and characterize associated transactivation/repression domains, including LCRs, enabling us to develop models of KLF/SP domain co-occurrence evolution. By extending our domain search to include entire proteomes, we find evidence for a complex intersection of domain shuffling, gene duplication, and de novo domain evolution as the primary mechanisms for the diversification of the KLF/SP gene family across the Metazoa. Our results uncover a pattern of an increased frequency of repressive domain connectivity repertoires (P-rich LCRs, SID, R2/R3, and PVDLS domains) in the KLF/SP gene family among metazoans suggesting a role in mediating diverse transcriptional repression activity. Our phylogenetic results further suggest that the expansion of the KLF/SP gene family mirrors increased cell type diversity during metazoan lineage diversification. The expansion and diversification of the KLF/SP gene family within the Metazoa may thus reflect the accumulation of differential transcriptional repression strategies associated with the development of extensive repertoires of cell types required to support complex tissues.

Supplementary Material

Supplementary figs. S1–S5 and tables S1–S5 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/). Additional online supplementary materials can be viewed at https://goo.gl/dbdBil.

Acknowledgments

This work was supported by startup funds from the University of Miami College of Arts and Sciences to W.E.B. J.S.P. was supported by the University of Miami College of Arts and Sciences. C.E.S. was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. The authors thank James Baker, Allie Graham, Isaac Skromne, Lauren Vandepas, members of the Browne laboratory, and two anonymous reviewers for their thoughtful comments and critical reading of the manuscript.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. Prottest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.
Adams MD, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195.
Adl SM, et al. 2012. The revised classification of eukaryotes. J Eukaryot Microbiol. 59:429–514.
Amemiya CT, et al. 2013. The *African coelacanth* genome provides insights into tetrapod evolution. Nature 496:311–316.
Aparicio S, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310.
Arabidopsis Genome Iniative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815.
Archer MC. 2011. Role of sp transcription factors in the regulation of cancer cell metabolism. Genes Cancer 2:712–719.
Athanikar JN, Sanchez HB, Osborne TF. 1997. Promoter selective transcriptional synergy mediated by sterol regulatory element binding protein and Sp1: a critical role for the Btd domain of sp1. Mol Cell Biol. 17:5193–5200.
Aury J-M, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444:171–178.
Basu P, et al. 2005. Klf2 is essential for primitive erythropoiesis and regulates the human and murine embryonic beta-like globin genes in vivo. Blood 106:2566–2571.
Black AR, Black JD, Azizkhan-Clifford J. 2001. Sp1 and Krüppel-like factor family of transcription factors in cell growth regulation and cancer. J Cell Physiol. 188:143–160.
Blok LJ, Grossmann ME, Perry JE, Tindall DJ. 1995. Characterization of an early growth response gene, which encodes a zinc finger transcription factor, potentially involved in cell cycle regulation. Mol Endocrinol. 9:1610–1620.
Bouwman P, Philipsen S. 2002. Regulation of the activity of Sp1-related transcription factors. Mol Cell Endocrinol. 195:27–38.
Brown RS, Sander C, Argos P. 1985. The primary structure of transcription factor tfiiia has 12 consecutive repeats. FEBS Lett. 186:271–274.
Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. 2014. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Mol Biol Evol. 31:1402–1413.
Chen Z, Lei T, Chen X, Zhang J. 2009. Porcine KLF gene family: structure, mapping, and phylogenetic analysis. Genomics 95:111–119.
Chinwalla AT, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
Clarke M, et al. 2013. Genome of *Acanthamoeba castellanii* highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. Genome Biol. 14:R11.
Colbourne JK, et al. 2011. The ecoresponsive genome of *Daphnia pulex*. Science 331:555–561.
Coletta A, et al. 2010. Low-complexity regions within protein sequences have position-dependent roles. BMC Syst Biol. 4:43.
Cook T, Gebelein B, Belal M, Mesa K, Urrutia P. 1999. Three conserved transcriptional repressor domains are a defining feature of the TIEG subfamily of Sp1-like zinc finger proteins. J Biol Chem. 274:29500–29504.

Cook T, Gebelein B, Mesa K, Mladek A, Urrutia R. 1998. Molecular cloning and characterization of TIEG2 reveals a new subfamily of transforming growth factor-β-inducible Sp1-like zinc finger-encoding genes involved in the regulation of cell growth. J Biol Chem. 273:25929–25936.
Crossley M, et al. 1996. Isolation and characterization of the cDNA encoding BKLF/TEF-2, a major CACCC-box-binding protein in erythroid cells and selected other cells. Mol Cell Biol. 16:1695–1705.
Curtis BA, et al. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature 492:59–65.
de Castro E, et al. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 34:W362–W365.
De Graeve F, et al. 2003. Identification of the *Drosophila progenitor* of mammalian Krüppel-like factors 6 and 7 and a determinant of fly development. Gene 314:55–62.
de Mendoza A, et al. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc Natl Acad Sci U S A. 110:E4858–E4866.
Dehal P, et al. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science 298:2157–2167.
DePristo MA, Zilversmit MM, Hartl DL. 2006. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. Gene 378:19–30.
Derelle R, Lang BF. 2012. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. Mol Biol Evol. 29:1277–1289.
Dunn CW, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.
Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Fairclough S, et al. 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. Genome Biol. 14:R15.
Faux NG, et al. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res. 15:537–551.
Feng WC, Southwood CM, Bieker JJ. 1994. Analyses of beta-thalassemia mutant DNA interactions with erythroid Krüppel-like factor (EKLF), an erythroid cell-specific transcription factor. J Biol Chem. 269:1493–1500.
Fritz-Laylin LK, et al. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell 140:631–642.
Gerber HP, et al. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263:808–811.
Gill G, Pascal E, Tseng ZH, Tjian R. 1994. A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the *Drosophila* TFIID complex and mediates transcriptional activation. Proc Natl Acad Sci U S A. 91:192–196.
Goffeau A, et al. 1996. Life with 6000 genes. Science 274:546–567.
Gordon AR, et al. 2008. Splenomegaly and modified erythropoiesis in KLF13−/− mice. J Biol Chem. 283:11897–11904.
Haas BJ, et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature 461:393–398.
Hanna-Rose W, Hansen U. 1996. Active repression mechanisms of eukaryotic transcription repressors. Trends Genet. 12:229–234.
Harrison SM, Houzelstein D, Dunwoodie SL, Beddington RSP. 2000. Sp5, a new member of the Sp1 family, is dynamically expressed during development and genetically interacts with Brachyury. Dev Biol. 227:358–372.
Heidel AJ, et al. 2011. Phylogeny-wide analysis of social amoeba genomes highlights ancient origins for complex intercellular communication. Genome Res. 21:1882–1891.
Hellsten U, et al. 2010. The genome of the western clawed frog *Xenopus tropicalis*. Science 328:633–636.
Hoey T, et al. 1993. Molecular cloning and functional analysis of *Drosophila* TAF110 reveal properties expected of coactivators. Cell 72:247–260.
Howe K, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503.
Imataka H, et al. 1992. Two regulatory proteins that bind to the basic transcription element (BTE), a GC box sequence in the promoter region of the rat P-4501A1 gene. EMBO J. 11:3663–3671.
Imhof A, et al. 1999. Transcriptional regulation of the AP-2α promoter by BTEB-1 and AP-2rep, a novel wt-1/egr-related zinc finger repressor. Mol Cell Biol. 19:194–204.
Iuchi S. 2001. Three classes of C2H2 zinc finger proteins. Cell Mol Life Sci. 58:625–635.
Jiang J, et al. 2008. A core KLF circuitry regulates self-renewal of embryonic stem cells. Nature Cell Biol. 10:353–360.
Jones FC, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55–61.
Kaczynski JA, Cook T, Urrutia R. 2003. Sp1- and Krüppel-like transcription factors. Genome Biol. 4:206.
Kaczynski JA, et al. 2002. Functional analysis of basic transcription element (BTE)-binding protein (BTEB) 3 and BTEB4, a novel Sp1-like protein, reveals a subfamily of transcriptional repressors for the BTE site of the cytochrome P4501A1 gene promoter. Biochem J. 366:873–882.
Kadonaga JT, Carner KR, Masiarz FR, Tijian R 1987. Isolation of cDNA-encoding transcription factor Sp1 and functional-analysis of the DNA-binding domain. Cell 51:1079–1090.
Kamper J, et al. 2006. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. Nature 444:97–101.
Katinka MD, et al. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414:450–453.
King N, et al. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature 451:783–788.
Kolell KJ, Crawford DL. 2002. Evolution of Sp transcription factors. Mol Biol Evol. 19:216–222.
Kotkamp K, Mössner R, Allen A, Onichtchouk D, Driever W. 2014. A Pou5f1/Oct4 dependent Klf2a, Klf2b, and Klf17 regulatory sub-network contributes to EVL and ectoderm development during zebrafish embryogenesis. Dev Biol. 385:433–447.
Laherty CD, et al. 1997. Histone deacetylases associated with the mSin3 corepressor mediate mad transcriptional repression. Cell 89:349–356.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Loftus B, et al. 2005. The genome of the protist parasite *Entamoeba histolytica*. Nature 433:865–868.
Lomberk G, et al. 2013. Krüppel-like factor 11 regulates the expression of metabolic genes via an evolutionarily conserved protein interaction domain functionally disrupted in maturity onset diabetes of the young. J Biol Chem. 288:17745–17758.
Materna SC, Howard-Ashby M, Gray RF, Davidson EH. 2006. The C2H2 zinc finger genes of *Strongylocentrotus purpuratus* and their expression in embryonic development. Dev Biol. 300:108–120.
Matsuzaki M, et al. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature 428:653–657.
Maxwell E, Ryan J, Schnitzler C, Browne W, Baxevanis A. 2012. Micrornas and essential components of the microrna processing machinery are not encoded in the genome of the ctenophore *Mnemiopsis leidyi*. BMC Genomics 13:714.
McConnell BB, Ghaleb AM, Nandan MO, Yang VW. 2007. The diverse functions of Krüppel-like factors 4 and 5 in epithelial biology and pathobiology. BioEssays 29:549–557.
McConnell BB, Yang VW. 2010. Mammalian Krüppel-like factors in health and diseases. Physiol Rev. 90:1337–1381.

Meadows SM, Salanga MC, Krieg PA. 2009. Krüppel-like factor 2 cooperates with the ETS family protein ERG to activate Flk1 expression during vascular development. Development 136:1115–1125.
Miller I, Bieker J. 1993. A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Krüppel family of nuclear proteins. Mol Cell Biol. 13:2776–2786.
Miller J, McLachlan AD, Klug A. 1985. Repetitive zinc-binding domains in the protein transcription factor IIIa from xenopus oocytes. EMBO J. 4:1609–1614.
Moroz LL, et al. 2014. The ctenophore genome and the evolutionary origins of neural systems. Nature 510:109–114.
Muñoz-Descalzo S, Terol J, Paricio N. 2005. Cabut, a C2H2 zinc finger transcription factor, is required during *Drosophila* dorsal closure downstream of JNK signaling. Dev Biol. 287:168–179.
Nakashima K, et al. 2002. The novel zinc finger-containing transcription factor osterix is required for osteoblast differentiation and bone formation. Cell 108:17–29.
Nosenko T, et al. 2013. Deep metazoan phylogeny: when different genes tell different stories. Mol Phylogenet Evol. 67:223–233.
Nosrati N, Kapoor NR, Kumar V. 2014. Combinatorial action of transcription factors orchestrates cell cycle-dependent expression of the ribosomal protein genes and ribosome biogenesis. FEBS J. 281:2339–2352.
Nylander JA, Wilgenbusch JC, Warren DL, Swofford DL. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581–583.
Paps J, Medina-Chacón LA, Marshall W, Suga H, Ruiz-Trillo I. 2013. Molecular phylogeny of unikonts: new insights into the position of apusomonads and ancyromonads and the internal relationships of opisthokonts. Protist 164:2–12.
Parker-Katiraee L, et al. 2007. Identification of the imprinted KLF14 transcription factor undergoing human-specific accelerated evolution. PLoS Genet. 3:e65.
Pearson R, Fleetwood J, Eaton S, Crossley M, Bao S. 2008. Krüppel-like transcription factors: a functional family. Int J Biochem Cell Biol. 40:1996–2001.
Philippe H, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr Biol. 19:706–712.
Piskacek M. 2009. Common transactivation motif 9aaTAD recruits multiple general co-activators TAF9, MED15, CBP and p300. Nature Precedings 12:30.
Piskacek S, et al. 2007. Nine-amino-acid transactivation domain: establishment and prediction utilities. Genomics 89:756–768.
Promponas VJ, et al. 2000. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Bioinformatics 16:915–922.
Punta M, et al. 2012. The Pfam protein families database. Nucleic Acids Res. 40:D290–D301.
Putnam NH, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1071.
Putnam NH, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94.
Radding CM. 1982. Homologous pairing and strand exchange in genetic recombination. Annu Rev Genet. 16:405–437.
Rambaut A, Drummond AJ. 2007. Tracer v1. 4. Available from http://beast.bio.ed.ac.uk/Tracer.
Ravasi T, et al. 2003. Systematic characterization of the zinc-finger-containing proteins in the mouse transcriptome. Genome Res. 13:1430–1442.
Richards S, et al. 2008. The genome of the model beetle and pest *Tribolium castaneum*. Nature 452:949–955.
Rodriguez I. 2011. *Drosophila tieg* is a modulator of different signalling pathways involved in wing patterning and cell proliferation. PLoS One 6:e18418.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
Rosenberg UB, et al. 1986. Structural homology of the product of the *Drosophila krüppel* gene with *Xenopus* transcription factor IIIa. Nature 319:336–339.
Rubin GM, et al. 2000. Comparative genomics of the eukaryotes. Science 287:2204–2215.
Ryan J, et al. 2010. The homeodomain complement of the ctenophore *Mnemiopsis leidyi* suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. Evodevo 1:9.
Ryan JF, et al. 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science 342:1242592.
Schaeper ND, Prpic NM, Wimmer EA. 2010. A clustered set of three Sp-family genes is ancestral in the Metazoa: evidence from sequence analysis, protein domain structure, developmental expression patterns and chromosomal location. BMC Evol Biol. 10:88.
Schuetz A, et al. 2011. The structure of the Klf4 DNA-binding domain links to self-renewal and macrophage differentiation. Cell Mol Life Sci. 68:3121–3131.
Sebé-Pedrós A, et al. 2013. Early evolution of the T-box transcription factor family. Proc Natl Acad Sci. 110:16050–16055.
Seetharam A, Bai Y, Stuart G. 2010. A survey of well conserved families of C2H2 zinc-finger genes in Daphnia. BMC Genomics 11:276.
Seetharam A, Stuart G. 2013. A study on the distribution of 37 well conserved families of C2H2 zinc finger genes in eukaryotes. BMC Genomics 14:420.
Shimeld SM. 2008. C2H2 zinc finger genes of the Gli, Zic, Klf, Sp, Wilms' tumour, Huckebein, Snail, Ovo, Spalt, Odd, Blimp-1, Fez and related gene families from *Branchiostoma floridae*. Dev Genes Evol. 218:639–649.
Shinzato C, et al. 2011. Using the *Acropora digitifera* genome to understand coral responses to environmental change. Nature 476:320–323.
Silverstein R, Ekwall K. 2005. Sin3: a flexible regulator of global gene expression and genome stability. Curr Genet. 47:1–17.
Sim KL, Creamer TP. 2004. Protein simple sequence conservation. Proteins 54:629–638.
Simakov O, et al. 2013. Insights into bilaterian evolution from three spiralian genomes. Nature 493:526–531.
Small KS, et al. 2011. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat Genet. 43:561–564.
Smith JJ, Antonacci F, Eichler EE, Amemiya CT. 2009. Programmed loss of millions of base pairs from a vertebrate genome. Proc Natl Acad Sci. 106:11212–11217.
Smith JJ, Baker C, Eichler EE, Amemiya CT. 2012. Genetic consequences of programmed genome rearrangement. Curr Biol. 22:1524–1529.
Smith JJ, et al. 2013. Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nat Genet. 45:415–421, 421e1–2.
Sodergren E, et al. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. Science 314:941–952.
Song A, Chen Y-F, Thamatrakoln K, Storm TA, Krensky AM. 1999. RFLAT-1: a new zinc finger transcription factor that activates RANTES gene expression in T lymphocytes. Immunity 10:93–103.
Soufi A, Donahue G, Zaret KS. 2012. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell 151:994–1004.
Soufi A, et al. 2015. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell 161:555–568.
Srivastava M, et al. 2008. The Trichoplax genome and the nature of placozoans. Nature 454:955–960.
Srivastava M, et al. 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature 466:720–726.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
Sucgang R, et al. 2011. Comparative genomics of the social amoebae *Dictyostelium discoideum* and *Dictyostelium purpureum*. Genome Biol. 12:R20.
Suga H, et al. 2013. The Capsaspora genome reveals a complex unicellular prehistory of animals. Nat Commun. 4:2325.
Suske G, Bruford E, Philipsen S. 2005. Mammalian Sp/KLF transcription factors: bring in the family. Genomics 85:551–556.
Suttamanatwong S, et al. 2009. Sp proteins and Runx2 mediate regulation of matrix gla protein (MGP) expression by parathyroid hormone. J Cell Biochem. 107:284–292.
Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126:663–676.
Tsai M-Y, et al. 2014. Modulation of p53 and met expression by Krüppel-like factor 8 regulates zebrafish cerebellar development. Dev Neurobiol. doi: 10.1002/dneu.22258.
Turner J, Crossley M. 1998. Cloning and characterization of mCtBP2, a co-repressor that associates with basic Krüppel-like factor and other mammalian transcriptional regulators. EMBO J. 17:5129–5140.
Venkatesh B, et al. 2014. Elephant shark genome provides unique insights into gnathostome evolution. Nature 505:174–179.
van Vliet J, Turner J, Crossley M. 2000. Human Krüppel-like factor 8: a CACCC-box binding protein that associates with CtBP and represses transcription. Nucleic Acids Res. 28:1955–1962.
Weber U, Rodriguez E, Martignetti J, Mlodzik M. 2014. Luna, a *Drosophila* KLF6/KLF7, is maternally required for synchronized nuclear and centrosome cycles in the preblastoderm embryo. PLoS One 9:e96933.
Wierstra I. 2008. Sp1: emerging roles—beyond constitutive activation of TATA-less housekeeping genes. Biochem Biophys Res Commun. 372:1–13.
Wimmer EA, Jackle H, Pfeifle C, Cohen SM. 1993. A *Drosophila* homolog of human Sp1 is a head-specific segmentation gene. Nature 366:690–694.
Wootton JC. 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem. 18:269–285.
Wootton JC, Federhen S. 1993. Statistics of local complexity in amino-acid-sequences and sequence databases. Comput Chem. 17:149–163.
Zeng V, et al. 2011. De novo assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean *Parhyale hawaiensis*. BMC Genomics 12:581.
Zhang J-S, et al. 2001. A conserved α-helical motif mediates the interaction of Sp1-like transcriptional repressors with the corepressor mSin3a. Mol Cell Biol. 21:5041–5049.
Zhao CT, Meng AM. 2005. Sp1-like transcription factors are regulators of embryonic development in vertebrates. Dev Growth Differ. 47:201–211.
Zhao X, et al. 2010. Klf 6/copeb is required for hepatic outgrowth in zebrafish and for hepatocyte specification in mouse ES cells. Dev Biol. 344:79–93.

Associate editor: Mar Alba