Chapter 20

Parallel Evolution of Transcription Factors in Basal Metazoans

Krishanu Mukherjee and Leonid L. Moroz

Abstract

Transcription factors (TFs) play a pivotal role as regulators of gene expression, orchestrating the formation and maintenance of diverse animal body plans and innovations. However, the precise contributions of TFs and the underlying mechanisms driving the origin of basal metazoan body plans, particularly in ctenophores, remain elusive. Here, we present a comprehensive catalog of TFs in 2 ctenophore species, *Pleurobrachia bachei* and *Mnemiopsis leidyi*, revealing 428 and 418 TFs in their respective genomes. In contrast, morphologically simpler metazoans have a reduced TF representation compared to ctenophores, cnidarians, and bilaterians: the sponge *Amphimedon* encodes 277 TFs, and the placozoan *Trichoplax adhaerens* encodes 274 TFs. The emergence of complex ctenophore tissues and organs coincides with significant lineage-specific diversification of the zinc finger C2H2 (ZF-C2H2) and homeobox superfamilies of TFs. Notable, the lineages leading to *Amphimedon* and *Trichoplax* exhibit independent expansions of leucine zipper (BZIP) TFs. Some lineage-specific TFs may have evolved through the domestication of mobile elements, thereby supporting alternative mechanisms of parallel TF evolution and body plan diversification across the Metazoa.

Key words Ctenophora, Placozoa, Porifera, Genome, Transcription factors, *Pleurobrachia*, *Mnemiopsis*, *Amphimedon*, *Trichoplax*, Homeobox, BHLH, Neurons, Evolution of animal complexity, Transposons

1 Introduction

By specific DNA binding, transcription factors (TFs) evolved to directly interpret the genome, often acting as master regulators of gene expression. In humans alone, the genome encodes over 1600 TFs [1], controlling the expression of 20,000+ genes. Investigating the gain, loss, and lineage-specific origins of TFs during the transition from single-celled organisms to multicellular metazoans can provide valuable insights into the evolution of animal innovations. However, the origins of metazoan-specific genes are elusive, with some mechanisms emerging from comparative studies [2].

Leonid L. Moroz (ed.), *Ctenophores: Methods and Protocols*, Methods in Molecular Biology, vol. 2757,
https://doi.org/10.1007/978-1-0716-3642-8_20,

The concept of gene duplication followed by functional diversification has been proposed as a major mechanism for the origin of novel genes and genome evolution [3]. According to this hypothesis, ancestral genes within the same animal lineage gave rise to duplicated copies through gene duplication events. This idea is supported by extensive evidence of gene conservation and the corresponding regulatory networks over large evolutionary distances [2], exemplified by the discovery of the highly conserved HOX gene cluster [4, 5]. However, the absence of canonical bilaterian master gene regulators, such as the HOX cluster, in the ctenophores genomes [6, 7] challenges tracing their origin and evolution.

Some TFs might evolve de novo or be a result of lateral gene (viral) transfer [2] or activity of mobile elements [8]. The birth of the majority of metazoan-specific TFs remains largely unexplored. Moreover, the exact number and annotation of TFs encoded in basal metazoan lineages remain incomplete. The challenges include limitations of gaps, errors, polymorphisms in genomes, and the limitations of homology detection algorithms for detecting precise orthologs. Meticulous, hand-curated annotation, especially across basal metazoan genomes, should be combined with systematic phylogenetic analyses of each class/family of genes to identify taxonomically restricted or novel TFs that lack close homologs.

Here, we comprehensively analyze major TF families encoded in two ctenophore genomes as representatives of the earliest branching animal lineage, sister to the rest of Metazoa [7, 9–11]. This survey is supported by complementary manual annotation of genomes from sponges and placozoans, representing the second and third basal branches of the animal tree of life, respectively. Our findings reveal that zinc finger, homeobox, and specific basic helix–loop–helix (BHLH) gene families of TFs have undergone lineage-specific expansions in ctenophores. Independent ctenophore lineage-specific expansions of homeobox genes Antennapedia (Antp) and the six protein families, as well as in leucine zipper (BZIP), are equally evident, contributing to the parallel evolution of animal complexity within comb jellies.

This annotation can be a valuable reference platform for further exploring the genomic basis of complex cellular and tissue phenotypes in ctenophores and other basal metazoan lineages.

2 Methods

1. **Sequence searches** were conducted using the local database. In cases where the online BLAST tool was unavailable, the standalone BLAST tool from NCBI (https://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/) was utilized. The local BLAST was installed and unpacked on a UNIX platform. To prepare the BLAST database specifically for the *Pleurobrachia*

bachei genome, the "makeblastdb" command was employed (./makeblastdb -in pleurobrachia.nt -dbtype nucl). The database was then queried using the DNA-binding domain of transcription factors, and the results were saved in a text file named "test.fasta." TBLASTn searches were performed against the genome database with the following command: (./tblastn -query test.fasta -db pleurobrachia.nt -evalue 5 -out output.txt). A cutoff range of 10^{-5} to 10^{-10} was employed for the standalone BLAST to identify potential homologs [12]. The DNA-binding domains were recursively subjected to BLAST analysis until no further homologs were obtained. Blast hits were manually scrutinized to determine their potential as genuine homologs. In cases where gene models (exomes) were unavailable, the genomic sequences surrounding the coding region were extracted, and homology-based gene prediction using hidden Markov models (HMMs) was performed using FGENESH+ (www.softberry.com).

2. **Sequence searches at the online database**: TBLASTn searches were conducted to retrieve *Mnemiopsis leidyi* transcription factors using the online BLAST server at https://research.nhgri.nih.gov/mnemiopsis/sequenceserver/. For *Amphimedon queenslandica* and *Trichoplax adhaerens*, a comprehensive set of transcription factors was obtained by performing TBLASTn searches at the Ensembl genome browser (https://metazoa.ensembl.org). To search against metazoan genomes, the complete sets of human and fruit fly transcription factors were downloaded from the AnimalTFDB3.0 online database (http://bioinfo.life.hust.edu.cn/AnimalTFDB/#!/species). The default e-value cutoff for the online BLAST was utilized to retrieve all potential homologs.
3. **Protein domain identification**: To identify the DNA-binding domain and other associated domains, searches were performed using the NCBI conserved domain database (CDD) [13] and the SMART [14, 15] online database. These databases were utilized to uncover conserved structural and functional domains relevant to DNA binding and transcription factor activity.
4. **Protein multiple-domain alignment**: The sequences were aligned using the MUSCLE alignment algorithm [15]. Alignment was performed either through an online tool available at EBI (https://www.ebi.ac.uk/Tools/msa/muscle/) or using the command-line approach in UNIX with the following command: "./muscle -in input.fasta -out output.muscle". This allowed for the alignment of the sequences to generate a comprehensive and accurate alignment for further analysis.

5. **Phylogeny reconstruction**: Maximum-likelihood (ML) trees were constructed using PhyML v3.0 [16, 17]. The most appropriate evolutionary model was determined using the AIC criterion as estimated by ProtTest [18]. ML phylogenies were generated using the JTT model, which accounts for rate heterogeneity, an estimated proportion of invariable sites, four rate categories, and an estimated alpha distribution parameter. This was achieved through the following command-line UNIX command: "./phyml -i input.phy -d aa -m JTT -v 0.0 -c 4 -a e -s 'BEST'". To optimize tree topology searches, both NNI (nearest-neighbor interchanges) and SPR (subtree pruning and regrafting) moves were employed [19]. Clade support was assessed using the SH-like approximate likelihood ratio test [20]. These analyses facilitated the construction of robust ML trees, capturing the evolutionary relationships among the studied transcription factors.
6. **Post-processing of tree file**: The tree file generated from the PhyML analysis was uploaded to an online tool (https://itol.embl.de/) for visualization. Once final adjustments were completed, the tree was exported in a vector-graphics format, such as SVG or EPS. The exported file was then imported into Adobe Illustrator CS6 version 16.0.0 for further refinement and the addition of labels and other annotations. Adobe Illustrator was utilized to create a polished and visually appealing representation of the final tree for presentation and publication purposes.
7. **Ortholog assessment**: To identify one-to-one orthologs across metazoan species, we employed the OrthoFinder tool [21]. Additionally, we employed a family-wise maximum-likelihood (ML) tree reconstruction approach to identify taxonomically restricted transcription factors (TFs). In addition to the four basal metazoan genomes, we incorporated the genomes of bilaterian organisms, specifically *Drosophila melanogaster* (fly) and *Homo sapiens* (humans), to assign the evolutionary conservative set of TF orthologs. Protein domain conservation served as a benchmark for evaluating the effectiveness of ortholog inference methods, ensuring robust and reliable identification of orthologous TFs across the metazoan lineage.

3 Illustrated Examples

3.1 Identification and Annotation of TFs in Basal Metazoan Lineages

We identify the nearly complete set of transcription factors (TFs) encoded in Ctenophora, Porifera, and Placozoa genomes. And representatives of each group show remarkable lineage-specific diversifications of different families of TFs, as summarized below.

Ctenophores *Pleurobrachia bachei* [6] and *Mnemiopsis leidyi* [7] encode 428 and 418 TFs, respectively (Fig. 1a, b), which exceed the previously reported number of 281 TFs in the *Mnemiopsis* genome [22]. We also found 277 TFs in *Amphimedon queenslandica* and 274 TFs in *Trichoplax adhaerens.*

The broader TF gene repertoire in ctenophores is likely associated with the greater complexity of cell types, tissues, and organs than in morphologically simpler sponges and placozoans. Still, both *Pleurobrachia* and *Mnemiopsis* have independent expansions of two major gene families: homeobox and zinc finger C2H2 (ZF-C2H2) (Fig. 1a–c). Approximately half of the total TFs encoded in ctenophore genomes are zinc finger TFs, which, unlike C2H2-ZF in tetrapods, lack KRAB and SCAN domains. The ZF-C2H2 family can be further classified into BED, THAP, and FLYWCH types derived from transposons or linked to transposon capture [23–25].

In *Amphimedon*, there is the overall expansion of different TFs: BZIP, CENPB [26, 27], FHY3 [28], HTH-Psq [29], and MADF [30] (Fig. 1a). Except for BZIP, all these TF families have transposon-derived origins or are associated with transposase capture. Similarly, the lineage-specific radiation of ZF-C2H2 in *Amphimedon* is attributed to expansions of ZF-BED and ZF-THAP (Fig. 1d), both of which have transposon origins [8].

In the third lineage leading to *Trichoplax*, the observed diversification of TFs in the genome primarily stems from BZIP, BHLH, CBFB, HMG/SOX, and specific homeobox genes (Fig. 1a, d).

In summary, these findings highlight the parallel evolution of TF families across all basal metazoan genomes, with transposons potentially playing a significant role [8].

3.2 Lineage-Specific Diversification of TFs and Their Evolutionary Implications

Phylogenetic reconstruction enables more precise identification of lineage-specific expansions within specific transcription factor [31] families across genomes. Figure 2 illustrates the expansion of the Antennapedia group in the ctenophore lineage. This group encompasses homeobox TF families, including ParaHox, EHGbox, and NK-like genes, which play crucial roles in developmental processes in bilaterians [32, 33]. The mechanisms underlying the independent expansion of Antennapedia and SIX homeobox genes in the ctenophore lineage remain unknown.

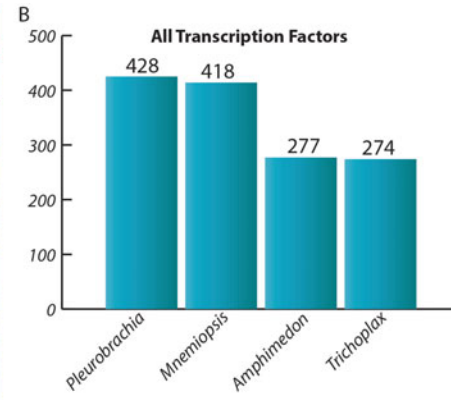
HOX genes determine the anterior–posterior body axis in bilaterians [34, 35]. Interestingly, ctenophores, sponges, and apparently placozoans do not encode recognizable HOX genes [6, 7, 36, 37], suggesting the origin of the HOX code in the common ancestor of Cnidaria+Bilateria.

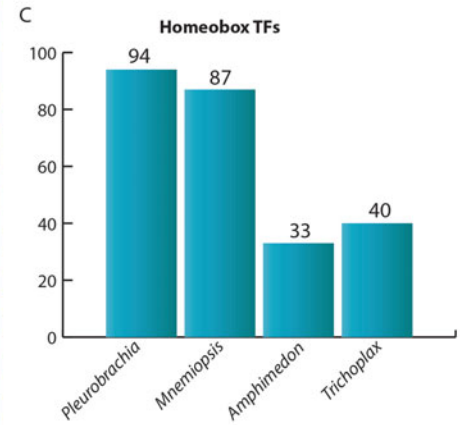
Our analysis further reveals the lineage-specific expansion of posterior HOX genes, independently occurring in sea urchins and humans (Fig. 2). The posterior HOX gene cluster within vertebrates exhibits interesting expression patterns, including HOX10 to HOX13. For instance, HOX10 is involved in patterning the

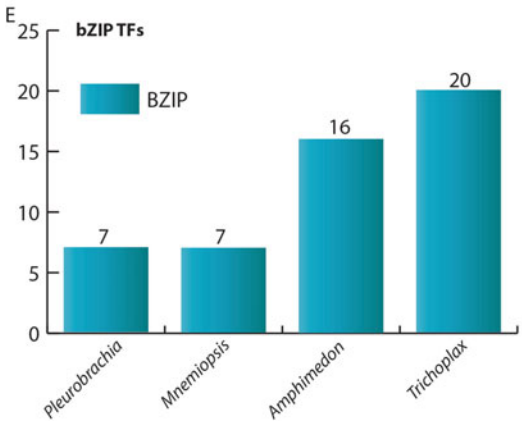
A

TF Gene Family	*Pleurobrachia*	*Mnemiopsis*	*Amphimedon*	*Trichoplax*
Previous Data		281	232	209
Total	**428**	**418**	**277**	**274**
AP2	1	1	0	1
ARID/BRIGHT	4	5	6	5
BZIP	7	7	**16**	**20**
BHLH	**25**	22	19	**28**
CBFB	2	3	2	**6**
CENPB *	2	0	**8**	**3**
churchill	0	0	1	1
CP2/Grainy Head	1	1	1	1
DM	2	5	0	2
E2F	3	3	6	4
ETS	6	10	6	7
FORKHEAD	20	14	18	18
FHY3 *	0	0	**3**	0
GATA	3	2	2	2
General TFs	10	9	9	10
HMG/SOX	12	11	10	**17**
Homeobox	**94**	87	33	**40**
HSF	1	1	1	2
HTH & HTH-Psq *	**5**	**1**	**13**	**2**
IRF	1	1	2	2
MADS	0	2	0	3
MBD	0	0	0	1
MYB/SANT	14	13	10	17
MYB/MADF *	**2**	**2**	**10**	**4**
NHR	0	2	2	**6**
NURF	1	2	1	1
RFX	3	2	3	3
RUNT	1	1	1	1
SMAD	3	3	3	4
SPT20	0	1	1	1
STAT	2	3	2	3
T-BOX	5	**7**	**8**	5
ZF-BED *	**1**	**1**	**13**	0
ZF-THAP *	**6**	**4**	**15**	**1**
ZF-FLYWCH *	**16**	**16**	0	0
ZF-C2H2	**175**	**174**	52	49
Other ZF TFs	-	2	2	4
Total	**428**	**418**	277	274

* Originated by transposon capture







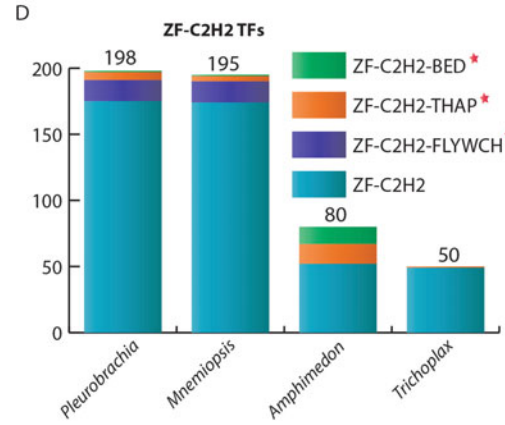


Fig. 1 Diversity and independent expansion of transcription factors in the genomes of the three earliest branching metazoan lineages: Ctenophora (*Pleurobrachia bachei* and *Mnemiopsis leidyi*), Porifera (*Amphimedon queenslandica*), and Placozoa (*Trichoplax adhaerens*). (**a**) The metazoan genomes exhibit diverse transcription factors (TFs), with each lineage undergoing independent radiations highlighted in bold and distinct color-coding within the cell. The letter "red" signifies either the absence of TFs or a significant reduction in their numbers. (**b**) The bar diagram illustrates the total number of TFs encoded in the genomes from four basal metazoan lineages, with the corresponding numbers above each bar. Notably, the ctenophore

lumbar vertebra region, an evolutionary novelty associated with locomotion adaptations. HOX11 patterns the sacrum region, which is crucial for the male and female reproductive tracts, including the development of the oviduct, uterus, and cervix [38]. The observed correlation between innovation in the lumbar region and evolutionary novelties in the reproductive system might contribute to the shift from egg-laying to embryo implantation [38].

NK genes are essential for establishing neuronal identity along the vertebrate neural tube's dorsoventral (D–V) axis [39]. Similarly, the *sine oculis* homeobox (SIX) gene family plays a vital role in *Drosophila* retinal development and contributes to constructing multiple tissues and organs in vertebrates [40]. Additionally, vertebrate homeobox genes, apart from C2H2-ZF, are prone to transposase capture, indicating the need for further analysis to understand the involvement of transposons in the generation of these lineage-specific genes.

We support the ctenophore-first hypothesis, which places this phylum as the most basally branching clade, sister to other metazoans [6, 7, 9–11]. This topology of the metazoan phylogenetic tree helps to determine the evolutionary spectrum of gene expansion or gene loss events (see also details in [6]). If a TF homolog is present in Placozoa but absent in Ctenophora or Porifera, we inferred that this TF originated within the placozoan lineage. Similarly, if a TF homolog is present in the ctenophore lineage but absent in both Porifera and Placozoa, we inferred that this TF homolog originated within the Ctenophora lineage. The presence of the same TF homologs in Ctenophora and Cnidaria or Bilateria indicates that sponge and placozoan lineages secondarily lost these genes from the common metazoan ancestor.

In Fig. 3, the presence of BHLH gene families, including Achaete-scute, Dimm, Delilah, Fer, Hes 1/4, and Ahr, is depicted in the *Trichoplax* lineage while being absent from the other two more basal metazoan lineages (Ctenophora and Porifera). This evolutionary pattern suggests the origin of these transcription factors (TFs) in the common ancestor of Placozoa and Cnidaria, with their preservation in Bilateria. Notably, Achaete-scute, Dimm, and Delilah belong to the Group A BHLH gene, also known as proneuronal BHLH factors, which play a vital role in the early stages of

Fig. 1 (continued) genomes contain significantly more TFs than sponges (*Amphimedon*) and placozoans (*Trichoplax*). (**c**) The ctenophore genomes demonstrate the independent expansion of homeobox genes, indicating the diversification of this gene family within this lineage. (**d**) The represented basal metazoan genomes display an independent diversification of ZF-C2H2 TFs, highlighting their increased abundance and diversity within this group. (**e**) The *Amphimedon* and *Trichoplax* genomes exhibit independent expansions of leucine zipper (BZIP) TFs, signifying the specific diversification of BZIP TFs within these organisms

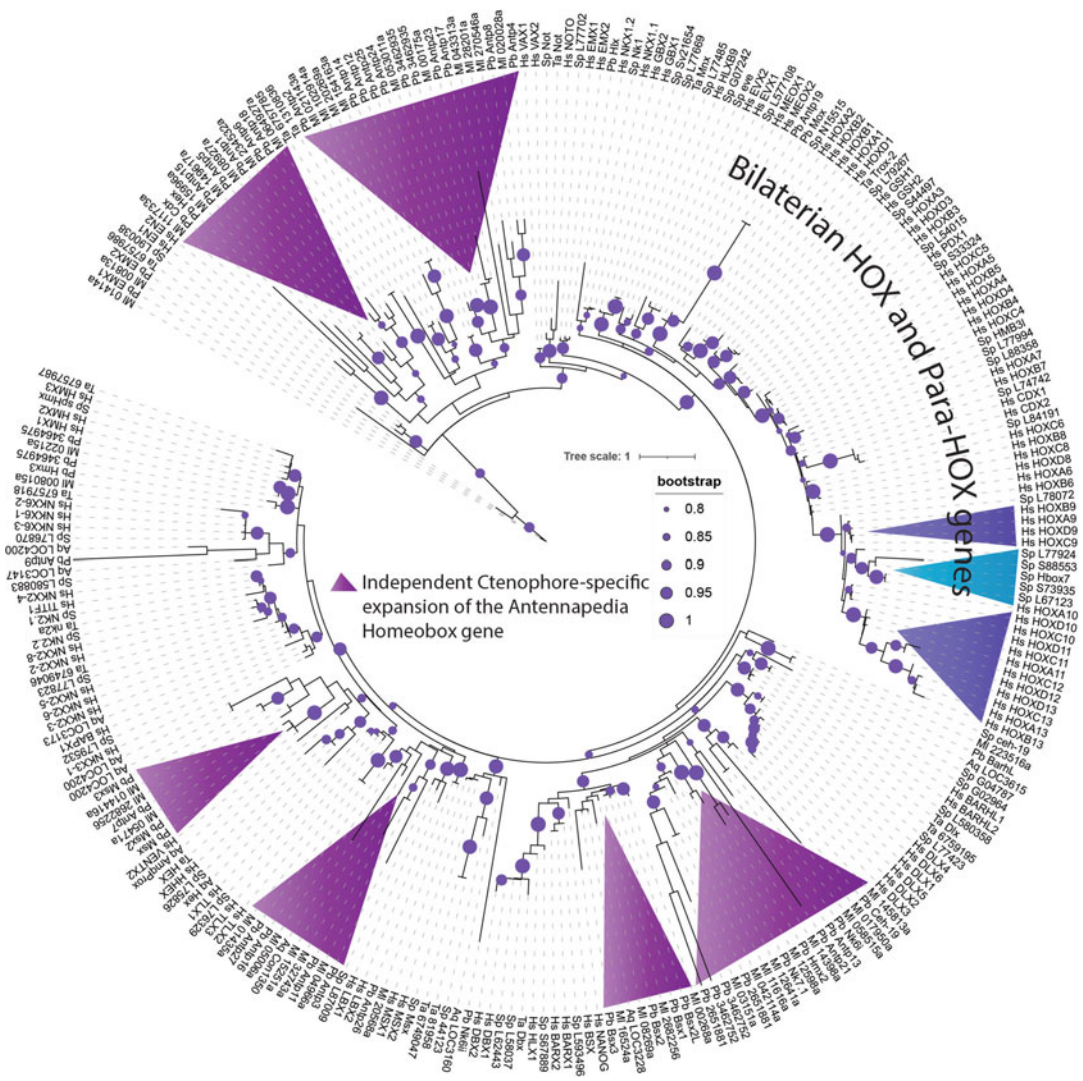


Fig. 2 Antennapedia homeobox genes have undergone independent lineage-specific expansions, particularly in the ctenophore lineage represented by *Pleurobrachia* and *Mnemiopsis*. Purple triangles denote these expanded genes and are absent from other basal metazoans, including sponge and placozoan lineages. Additionally, posterior HOX genes have independently expanded in some deuterostomes, such as sea urchins (cyan triangle) and humans (blue triangle). Abbreviations used in this and the other trees are Sci, *Sycon ciliatum*; Pb, *Pleurobrachia bachei*; ML, *Mnemiopsis leidyi*; Aq, *Amphimedon queenslandica*; Ta, *Trichoplax adhaerens*; Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Sp, *Strongylocentrotus purpuratus*. (See text and notes for details)

nervous system development [41, 42]. Given that placozoans lack neurons [43–46], it would be intriguing to investigate whether the Achaete-scute gene in *Trichoplax* is expressed in specific secretory (neuroid-like) cells [47], providing an avenue for future studies of alternative integrative systems [48, 49].

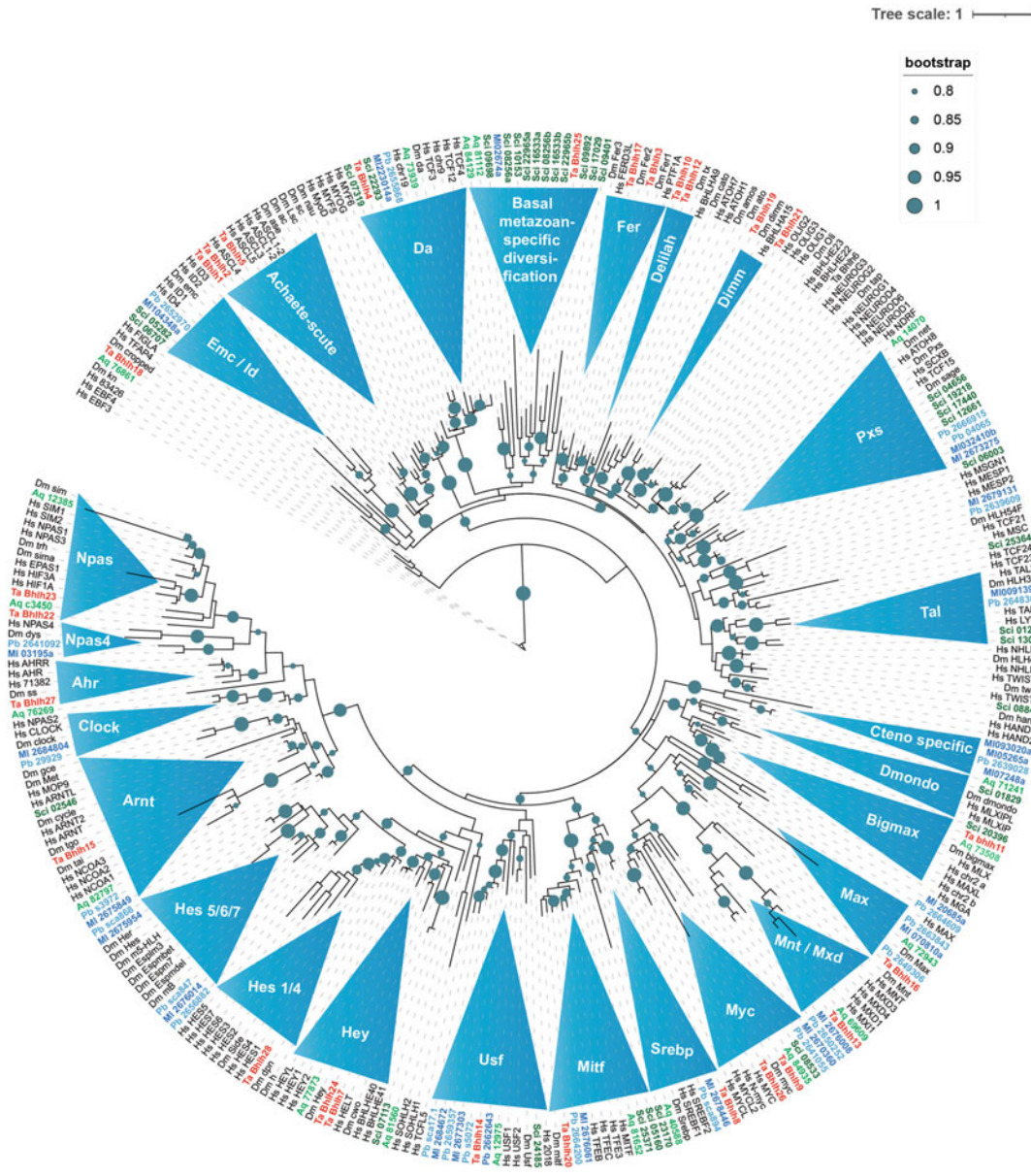


Fig. 3 Independent lineage-specific radiations of the basic helix–loop–helix (BHLH) genes. In the phylogenetic tree, the ctenophore species (*Pleurobrachia* and *Mnemiopsis*) are represented in blue color, Porifera (*Amphimedon*) is in green, and Placozoa (*Trichoplax*) is in red. The genes marked in black indicate their origin in bilaterians. Abbreviations: Sci, *Sycon ciliatum*; Pb, *Pleurobrachia bachei*; ML, *Mnemiopsis leidyi*; Aq, *Amphimedon queenslandica*; Ta, *Trichoplax adhaerens*; Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Sp, *Strongylocentrotus purpuratus*. (See text and notes for details)

3.3 Lineage-Specific Expansion of Leucine Zipper (BZIP) TFs

Figure 4 illustrates the phylogenetic reconstruction of the BZIP TF families. Creb1, originating early in the common ancestor of all metazoans, is conserved across all animal lineages, including ctenophores. However, Creb2 and Cebp originated in the common ancestor of sponges (*Amphimedon*, Aq) and are preserved in

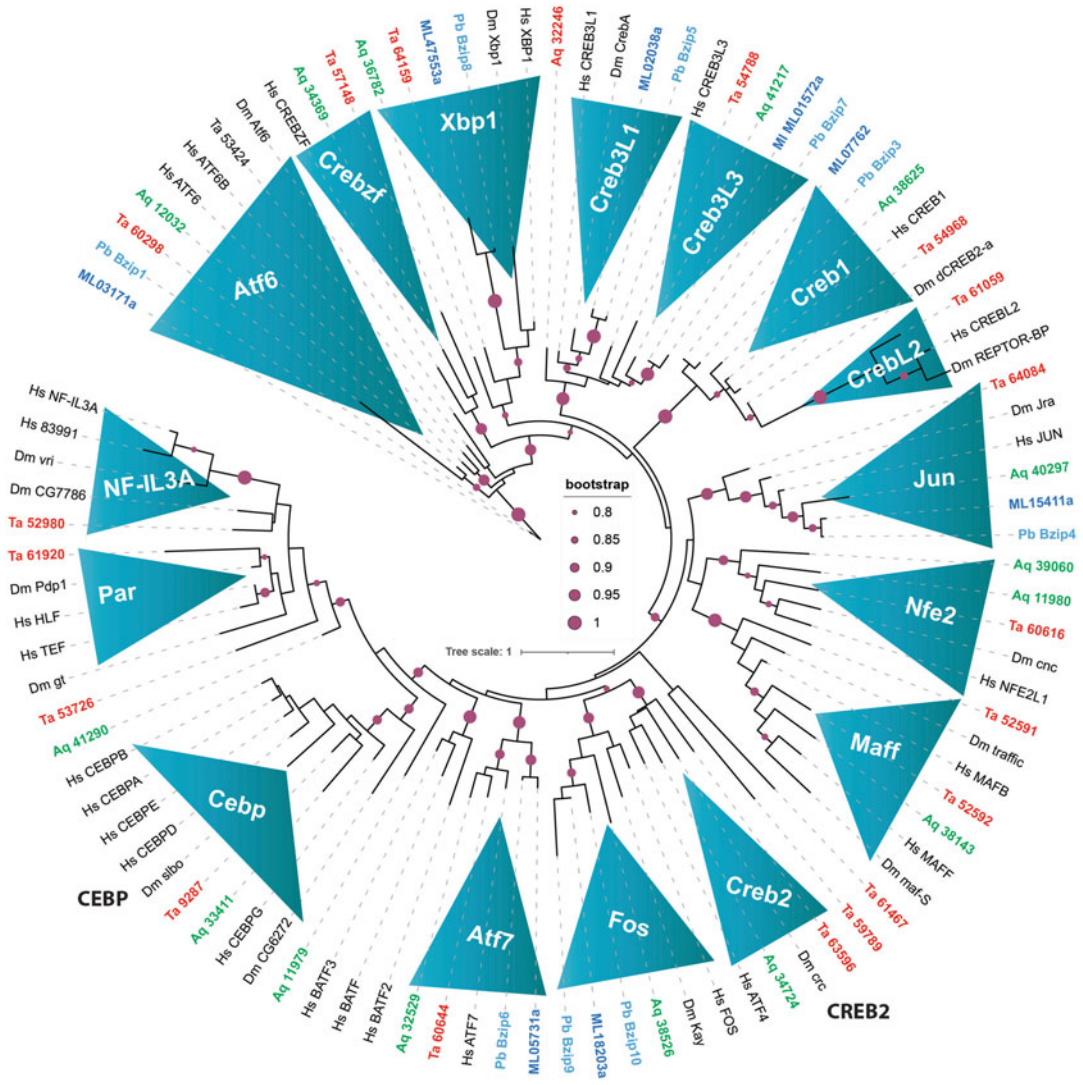


Fig. 4 Independent diversification and origin of leucine zipper BZIP gene family in the basal metazoan genome. Both Creb2 and Cebp are absent in the ctenophore lineage. On the other hand, the early gene Fos is conserved in the ctenophore and sponge lineages but is missing from the *Trichoplax* genome. Likewise, the early gene Jun is conserved across all basal metazoan lineages studied here. The beta zip gene families, including the canonical CrebL2, NFIL-3A, and Par, present in *Trichoplax* and are apparently absent from ctenophores and sponges. However, they are conserved in cnidarian and bilaterian lineages [77]. Abbreviations: Sci, *Sycon ciliatum*; Pb, *Pleurobrachia bachei*; ML, *Mnemiopsis leidyi*; Aq, *Amphimedon queenslandica*; Ta, *Trichoplax adhaerens*; Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Sp, *Strongylocentrotus purpuratus*. (See text and notes for details)

Trichoplax, but they are absent in the ctenophore genomes (*Pleurobrachia* (Pb) and *Mnemiopsis* (Ml)). Our analysis also suggests that these genes are conserved in protostomes, lophotrochozoans, deuterostomes, and cnidarian genomes. This presents two possible scenarios. If the Porifera-first hypothesis is truthful [50, 51], the

ctenophore lineage may have lost Creb2 and Cebp. Alternatively, and more likely, if the Ctenophora-first hypothesis is correct [9–11, 52], Creb2 and Cebp may have been primarily absent in stem ctenophores. Both cAMP response element-binding protein (Creb2) and CCAT enhancer-binding protein (Cebp) are involved in tissue-specific gene expression, proliferation, differentiation [53], learning [54], and long-term memory formation [55]. Thus, their absence in ctenophores raises intriguing questions regarding the mechanisms of neuroplasticity, stress response, and other cAMP-dependent processes in early-branching metazoans.

As described above, similar evolutionary distribution patterns for Creb2 and Cebp have been identified for particular beta zip subfamilies (BZIP). Specifically, Nfe2, Maff, and Crebzf are present in the *Amphimedon* and *Trichoplax* genomes as well as in cnidarians and bilaterians but absent in the *Pleurobrachia* and *Mnemiopsis* genomes, suggesting their 'later' origin (before the divergence of the sponge lineage). The functions of these BZIP TFs are currently unknown and require further analysis.

The beta zip subfamily CrebL2 presents in *Trichoplax* and is absent in all other prebilaterian metazoans but conserved among cnidarians and bilaterians. We interpret this phylogeny as a primary absence of this subfamily in ctenophores and sponges, originating in the common ancestor of the Placozoa+Cnidaria+Bilateria clade. The *Drosophila* ortholog of the human CrebL2 gene, named REPTOR-BP, functions downstream of mTORC1 and plays crucial roles in organismal metabolism, life span, and stress [56]. Additionally, human CrebL2 is a metabolic regulator in muscle and liver cells [57].

NF-IL3A also evolved in the common ancestor of Placozoa +Cnidaria+Bilateria. NF-IL3A is a human T-cell transcription factor transacting the interleukin-3 promoter [58]. The *Drosophila* ortholog of human NF-IL3A is vri, which controls cell growth and proliferation by regulating the actin skeleton [57]. Similarly, the canonical Par gene family, named after the *Drosophila* gene Pdp1 (PAR domain protein 1), which plays a crucial role in regulating muscle gene transcription, likely evolved in the common ancestor of *Trichoplax*, cnidarians, and bilaterians, with primary absence in sponges and ctenophores. The functions of many genes in early branched metazoans, including the examples mentioned above, remain largely unknown, and further studies on ctenophores and placozoans are crucial for shedding light on their evolutionary significance in Metazoa.

3.4 Conclusion and Future Directions

The lineage-specific expansion of transcription factors (TFs) is a widespread phenomenon observed in various TF families, including those that have potentially captured transposase domains independently. This hypothesis is one of many reasons in investigating the mechanisms underlying TF origins and evolution. Notably, zinc finger TFs exhibit significant lineage-specific expansions following

the independent capture of transposase domains in the tetrapod lineage.

A similar trend is observed in the ctenophore lineage, where the FLYWCH transposon appears to be the primary contributor to the expansion of zinc finger TFs [8]. In contrast, the expansion of zinc finger TFs in the *Amphimedon* lineage seems to involve THAP and BED transposons but not FLYWCH, suggesting a preference for transposon recruitment in the independent diversification of zinc finger TFs in sponges and metazoans as a whole.

Further studies are required to understand the expansion of the homeobox Antennapedia and SIX genes in the ctenophore lineage. These genes play critical roles in the developmental control of numerous cell types associated with different organs and tissues. However, the functional understanding of transposon-domesticated FLYWCH genes is limited. In *C. elegans*, FLYWCH genes function as repressors of embryonic expression of microRNA genes [59], which regulate gene expression post-transcriptionally. Ctenophores do not possess recognized microRNAs [6], thus highlighting the need for future analysis of the radiation of FLYWCH genes in the ctenophore lineage.

Recent studies have demonstrated that FLYWCH1 in humans suppresses the nuclear beta-catenin pathway [59], which is responsible for cell fate decisions. Similarly, the BED zinc finger TF ZBED3 has been shown to modulate the Wnt/beta-catenin pathway by binding to Axin [60]. These findings underscore the potential regulatory roles of FLYWCH and BED zinc finger TFs in crucial signaling pathways.

Further investigations require understanding the functional significance of lineage-specific TF expansions in ctenophores and their involvement in their lineage-specific signaling pathways. This critical line of research would provide valuable insights into the evolutionary and regulatory mechanisms shaping the diversity of transcription factors and body plans across basal metazoans.

In summary, we propose the following hypotheses:

1. The lineage-specific diversification of TFs triggers the convergent evolution of diverse cell types in major basal metazoan clades, including events of independent origins of neurons, mesoderm, and muscles [47, 61–64].
2. Generating a broader spectrum of morphologically similar cell types might occur convergently across phyla.
3. Lineage-specific expansions of certain TF classes can be associated with the activity and domestication of transposase-captured genes.

Notably, most TFs described here, and subjects to lineage-specific origins and/or diversification, have no association with recognizable transposase-capture events, which needs further analyses of respective mechanisms (e.g., duplications, lateral gene

transport, etc.). Notes below summarize various scenarios and examples of the evolution of TFs.

4 Notes

Illustrative examples for the role of mobile elements in the TF diversification and evolution.

1. Pax6, a homeobox gene, is a universal master control gene in bilaterian eye morphogenesis [65, 66]. The origin of the PAX gene can be attributed to the domestication of the transposable element Tc1/mariner, which is widely distributed across metazoans, certain unicellular eukaryotes, and plants [67]. The Pax family emerged through the fusion of a transposase domain with another gene, likely facilitated by exon-shuffling [68]. However, the precise timing and extent of transposase capture remain unclear due to the deep ancestry of the Pax genes. Additionally, it is currently unknown whether the capture of the transposase domain occurred once or multiple times during evolution.
2. Recent studies have provided evidence that the mobility of DNA transposons can facilitate exon-shuffling, allowing functional domains to be inserted into new genomic contexts. This phenomenon can lead to the formation of host-transposase fusion (HTF) genes through the splicing of these domains [69]. Cosby et al. demonstrated that transposase domains have been captured, primarily through alternate splicing, resulting in the generation of 106 distinct fusion proteins, mainly transcription factors. These events occurred independently around 106 times over approximately 350 million years during tetrapod evolution [69]. Remarkably, 77% of these HTF genes possess a DNA-binding domain, highlighting their role as transcription factors. These findings suggest that the fusion of host and transposase domains can give rise to novel regulatory genes, including deeply conserved and lineage-specific transcription factors.
3. The Krüppel-associated box zinc finger protein (KRAB-ZF) stands out as the largest and most prevalent family of zinc finger transcription factors (TFs), primarily found in tetrapod genomes but exhibiting a remarkable expansion in mammalian genomes [70]. Within the human TF repertoire of over 1600 genes, the KRAB domain-containing zinc finger family accounts for the highest abundance, encompassing around 423 TFs in the genome. Notably, a subset of KRAB-ZF proteins also possesses an additional SCAN domain. Interestingly, the KRAB and SCAN domains have been observed as

prominent targets for transposase fusion events [69]. Across the tetrapod phylogeny, the KRAB and SCAN domains have been involved in 32 and 19 independent fusion events, accounting for approximately 50% of such occurrences. Furthermore, 2 other transcription factor domains, C2H2-ZF and homeodomain, have undergone independent transposase fusion events 12 and 5 times, respectively, within the tetrapod lineage [69]. These analyses indicate that zinc finger and homeobox are among the most abundant transcription factor protein families, characterized by multiple independent instances of transposase capture throughout evolution.

4. See also Fig. 2 for independent diversification, origin, and losses of BHLH gene families in the basal metazoan genomes.
5. Notes about particular subfamilies of TFs relevant to their evolution.

Achaete-scute—The Achaete-scute gene complex, which plays a crucial role in forming neural precursors and subsequent differentiation of specific neuronal lineages in *Drosophila* [71], is found in the *Trichoplax* genome. However, it is absent in the sequenced genomes of sponges (*Sycon* [72], *Amphimedon* (Aq)), and ctenophores (*Pleurobrachia* (Pb), and *Mnemiopsis* (Ml)). On the other hand, it is present in major bilaterian lineages. In addition to its role in neuronal development, Achaete-scute is also expressed in a cluster of mesodermal cells and specific cells in the gut [73], highlighting its involvement in multiple cell patterning processes.

Dimm—The *Trichoplax* genome contains the BHLH gene Dimm, which is not found in the sequenced prebilaterian genomes of *Sycon* [74], *Amphimedon* (Aq), *Pleurobrachia* (Pb), and *Mnemiopsis* (Ml). However, it is present in bilaterians. Dimm is known for its role as a master regulator of secretory phenotypes in neuroendocrine cells. It is involved in the combinatorial code that regulates the terminal differentiation of peptidergic neurons [75].

Delilah—The *Trichoplax* genome contains the Delilah gene, which is not found in the sequenced genomes of ctenophores *Pleurobrachia* and *Mnemiopsis* and sponges *Sycon* and *Amphimedon*. However, it is present in many bilaterians. Delilah is absent in vertebrates but has been found in the *Branchiostoma* genome. This gene also exhibits almost exclusive expression in insect apodemes, specialized structures involved in insect locomotion and attachment [76].

Fer—The *Trichoplax* genome contains the Fer gene, which is not found in *Sycon*, *Amphimedon*, *Pleurobrachia*, and *Mnemiopsis* sequenced genomes. However, it is present

in bilaterians. In *Drosophila*, Fer plays a crucial role in the development of a specific subset of circadian pacemaker neurons and dopaminergic neurons in the photocerebral anterior media [77]. It is also involved in the development of the photocerebral anterior lateral (PAL) clusters of the brain, which are essential for the survival of PAM cluster dopaminergic neurons during adulthood and oxidative stress response [12].

Dmondo—The Dmondo gene is found in sponges *Sycon* and *Amphimedon* but is absent from the sequenced genomes of *Pleurobrachia*, *Mnemiopsis*, and *Trichoplax*. Dmondo serves as a nutrient sensor, playing a role in sensing and responding to nutrient availability [52, 78].

Hes 1/4—The *Trichoplax* genome contains the Hes 1/4 gene, which is absent from the sequenced genomes of ctenophores (*Pleurobrachia* and *Mnemiopsis*) and sponges (*Sycon* and *Amphimedon*). However, it is present in bilaterians. Hes 1/4 functions as a repressor of differentiation by targeting genes involved in the Notch signaling pathway. This pathway regulates lineage specification decisions during the development of various tissues [78].

Hes 5/6/7—The Hes6 gene is found in the sequenced genomes of *Pleurobrachia* and *Mnemiopsis* but is absent from the sequenced genomes of *Sycon*, *Amphimedon*, and *Trichoplax*. It is present in protostomes but absent in the deuterostome genomes, including humans. Hes6 plays a role in promoting neuronal differentiation, contributing to the development of neurons [79].

Clock—The clock gene is present in the genomes of *Pleurobrachia* and *Mnemiopsis* but absent from the sequenced genomes of *Sycon*, *Amphimedon*, and *Trichoplax*. It is found in bilaterians. The clock gene is a key regulator of sleep, stress, learning, and memory, as studied extensively in mice [80].

Ahr—The Ahr gene is present in the *Trichoplax* genome but absent from the sequenced genomes of ctenophores *Pleurobrachia* and *Mnemiopsis* and sponges *Sycon* and *Amphimedon*. It is found in bilaterians. Ahr is involved in binding to a range of endogenous and exogenous chemicals, including TCDD, and contributes to immune responses and the regulation of development and pathology [81].

Acknowledgments

This work was supported in part by the Human Frontiers Science Program (RGP0060/2017) and National Science Foundation (IOS-1557923) grants to LLM. Research reported in this publication was also supported in part by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under Award Number R01NS114491 (to LLM). The content is solely the authors' responsibility and does not necessarily represent the official views of the National Institutes of Health.

References

1. Lambert SA et al (2018) The human transcription factors. Cell 175(2):598–599
2. de Mendoza A, Sebe-Pedros A (2019) Origin and evolution of eukaryotic transcription factors. Curr Opin Genet Dev 58–59:25–32
3. Ohno S, Wolf U, Atkin NB (1968) Evolution from fish to mammals by gene duplication. Hereditas 59(1):169–187
4. McGinnis W et al (1984) A conserved DNA sequence in homoeotic genes of the *Drosophila* Antennapedia and bithorax complexes. Nature 308(5958):428–433
5. Scott MP, Weiner AJ (1984) Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of *Drosophila*. Proc Natl Acad Sci U S A 81(13):4115–4119
6. Moroz LL et al (2014) The ctenophore genome and the evolutionary origins of neural systems. Nature 510(7503):109–114
7. Ryan JF et al (2013) The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science 342(6164): 1242592
8. Mukherjee K, Moroz LL (2023) Transposon-derived transcription factors across metazoans. Front Cell Dev Biol 11:1113046
9. Schultz DT et al (2023) Ancient gene linkages support ctenophores as sister to other animals. Nature 618(7963):110–117
10. Whelan NV et al (2015) Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci U S A 112(18): 5773–5778
11. Whelan NV et al (2017) Ctenophore relationships and their placement as the sister group to all other animals. Nat Ecol Evol 1(11): 1737–1746
12. Tas D et al (2018) Parallel roles of transcription factors dFOXO and FER2 in the development and maintenance of dopaminergic neurons. PLoS Genet 14(3):e1007271
13. Marchler-Bauer A et al (2011) CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res 39 (Database issue):D225–D229
14. Letunic I, Bork P (2018) 20 years of the SMART protein domain annotation resource. Nucleic Acids Res 46(D1):D493–D496
15. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5): 1792–1797
16. Guindon S et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59(3):307–321
17. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52(5):696–704
18. Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21(9):2104–2105
19. Hordijk W, Gascuel O (2005) Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. Bioinformatics 21(24):4338–4347
20. Anisimova M et al (2011) Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst Biol 60(5): 685–699
21. Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20(1):238
22. Sebe-Pedros A et al (2018) Early metazoan cell type diversity and the evolution of multicellular gene regulation. Nat Ecol Evol 2(7): 1176–1188

23. Hayward A et al (2013) ZBED evolution: repeated utilization of DNA transposons as regulators of diverse host functions. PLoS One 8(3):e59940
24. Roussigne M et al (2003) The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. Trends Biochem Sci 28(2):66–69
25. Marquez CP, Pritham EJ (2010) Phantom, a new subclass of Mutator DNA transposons found in insect viruses and widely distributed in animals. Genetics 185(4):1507–1517
26. Mateo L, Gonzalez J (2014) Pogo-like transposases have been repeatedly domesticated into CENP-B-related proteins. Genome Biol Evol 6(8):2008–2016
27. Casola C, Hucks D, Feschotte C (2008) Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. Mol Biol Evol 25(1):29–41
28. Hudson ME, Lisch DR, Quail PH (2003) The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant J 34(4):453–471
29. Siegmund T, Lehmann M (2002) The *Drosophila* Pipsqueak protein defines a new family of helix-turn-helix DNA-binding proteins. Dev Genes Evol 212(3):152–157
30. Kapitonov VV, Jurka J (2004) Harbinger transposons and an ancient HARBI1 gene derived from a transposase. DNA Cell Biol 23(5): 311–324
31. Lein ES et al (2007) Genome-wide atlas of gene expression in the adult mouse brain. Nature 445(7124):168–176
32. Beck F (2002) Homeobox genes in gut development. Gut 51(3):450–454
33. Harvey RP (1996) NK-2 homeobox genes and heart development. Dev Biol 178(2):203–216
34. McGinnis W, Krumlauf R (1992) Homeobox genes and axial patterning. Cell 68(2): 283–302
35. Hughes CL, Kaufman TC (2002) Hox genes and the evolution of the arthropod body plan. Evol Dev 4(6):459–499
36. Larroux C et al (2007) The NK homeobox gene cluster predates the origin of Hox genes. Curr Biol 17(8):706–710
37. Jakob W et al (2004) The Trox-2Hox/ParaHox gene of *Trichoplax* (Placozoa) marks an epithelial boundary. Dev Genes Evol 214(4):170–175
38. Wellik DM, Capecchi MR (2003) Hox10 and Hox11 genes are required to globally pattern the mammalian skeleton. Science 301(5631): 363–367
39. McMahon AP (2000) Neural patterning: the role of Nkx genes in the ventral spinal cord. Genes Dev 14(18):2261–2264
40. Kumar JP (2009) The sine oculis homeobox (SIX) family of transcription factors as regulators of development and disease. Cell Mol Life Sci 66(4):565–583
41. Baker NE, Brown NL (2018) All in the family: proneural bHLH genes and neuronal diversity. Development 145(9)
42. Hartenstein V, Stollewerk A (2015) The evolution of early neurogenesis. Dev Cell 32(4): 390–407
43. Romanova DY et al (2021) Hidden cell diversity in Placozoa: ultrastructural insights from Hoilungia hongkongensis. Cell Tissue Res 385(3):623–637
44. Grell KG, Ruthmann A (1991) Placozoa. In: Harrison FW (ed) Microscopic anatomy of invertebrates. Wiley-Liss, New York, pp 13–27
45. Smith CL et al (2021) Microscopy studies of Placozoans. Methods Mol Biol 2219:99–118
46. Smith CL et al (2014) Novel cell types, neurosecretory cells, and body plan of the early-diverging metazoan *Trichoplax adhaerens*. Curr Biol 24(14):1565–1572
47. Moroz LL (2021) Multiple origins of neurons from secretory cells. Front Cell Dev Biol 9: 669087
48. Moroz LL, Romanova DY (2022) Alternative neural systems: what is a neuron? (Ctenophores, sponges and placozoans). Front Cell Dev Biol 10:1071961
49. Moroz LL, Romanova DY, Kohn AB (1821) Neural versus alternative integrative systems: molecular insights into origins of neurotransmitters. Philos Trans R Soc Lond Ser B Biol Sci 2021(376):20190762
50. Redmond AK, McLysaght A (2021) Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. Nat Commun 12(1):1783
51. Telford MJ, Moroz LL, Halanych KM (2016) Evolution: a sisterly dispute. Nature 529(7586):286–287
52. Li Y et al (2021) Rooting the animal tree of life. Mol Biol Evol 38(10):4322–4333
53. Tsukada J et al (2011) The CCAAT/enhancer (C/EBP) family of basic-leucine zipper (BZIP) transcription factors is a multifaceted highly-regulated system for gene regulation. Cytokine 54(1):6–19

54. Amar F et al (2021) Rapid ATF4 depletion resets synaptic responsiveness after cLTP. eNeuro 8(3):ENEURO.0239
55. Mirisis AA, Kopec AM, Carew TJ (2021) ELAV proteins bind and stabilize C/EBP mRNA in the induction of long-term memory in *Aplysia*. J Neurosci 41(5):947–959
56. Tiebe M et al (2015) REPTOR and REPTOR-BP regulate organismal metabolism and transcription downstream of TORC1. Dev Cell 33(3):272–284
57. Tiebe M et al (2019) Crebl2 regulates cell metabolism in muscle and liver cells. Sci Rep 9(1):19869
58. Zhang W et al (1995) Molecular cloning and characterization of NF-IL3A, a transcriptional activator of the human interleukin-3 promoter. Mol Cell Biol 15(11):6055–6063
59. Muhammad BA et al (2018) FLYWCH1, a novel suppressor of nuclear beta-catenin, regulates migration and morphology in colorectal cancer. Mol Cancer Res 16(12):1977–1990
60. Chen T et al (2009) Identification of zinc-finger BED domain-containing 3 (Zbed3) as a novel Axin-interacting protein that activates Wnt/beta-catenin signaling. J Biol Chem 284(11):6683–6689
61. Moroz LL (2009) On the independent origins of complex brains and neurons. Brain Behav Evol 74(3):177–190
62. Moroz LL (2014) The genealogy of genealogy of neurons. Commun Integr Biol 7(6): e993269
63. Moroz LL (2015) Biodiversity meets neuroscience: from the sequencing ship (Ship-Seq) to deciphering parallel evolution of neural systems in Omic's era. Integr Comp Biol 55(6): 1005–1017
64. Moroz LL, Kohn AB (2016) Independent origins of neurons and synapses: insights from ctenophores. Philos Trans R Soc Lond Ser B Biol Sci 371(1685):20150041
65. Hill RE et al (1991) Mouse small eye results from mutations in a paired-like homeobox-containing gene. Nature 354(6354):522–525
66. Ton CC et al (1991) Positional cloning and characterization of a paired box- and homeobox-containing gene from the aniridia region. Cell 67(6):1059–1074
67. Garcia-Fernandez J et al (1993) Infiltration of mariner elements. Nature 364(6433):109–110
68. Breitling R, Gerber JK (2000) Origin of the paired domain. Dev Genes Evol 210(12): 644–650
69. Cosby RL et al (2021) Recurrent evolution of vertebrate transcription factors by transposase capture. Science 371(6531)
70. Bellefroid EJ et al (1993) Clustered organization of homologous KRAB zinc-finger genes with enhanced expression in human T lymphoid cells. EMBO J 12(4):1363–1374
71. Bertrand N, Castro DS, Guillemot F (2002) Proneural genes and the specification of neural cell types. Nat Rev Neurosci 3(7):517–530
72. Hubbard T et al (2005) Ensembl 2005. Nucleic Acids Res 33(Database issue):D447–D453
73. Tepass U, Hartenstein V (1995) Neurogenic and proneural genes control cell fate specification in the *Drosophila* endoderm. Development 121(2):393–405
74. Fortunato SAV et al (2016) Conservation and divergence of bHLH genes in the calcisponge *Sycon ciliatum*. EvoDevo 7:23
75. Liu YT, Luo JN, Nassel DR (2016) The Drosophila transcription factor dimmed affects neuronal growth and differentiation in multiple ways depending on neuron type and developmental stage. Front Mol Neurosci 9
76. Armand P et al (1994) A novel basic helix-loop-helix protein is expressed in muscle attachment sites of the *Drosophila* epidermis. Mol Cell Biol 14(6):4145–4154
77. Pereira JF et al (2013) Boto, a class II transposon in *Moniliophthora perniciosa*, is the first representative of the PIF/harbinger superfamily in a phytopathogenic fungus. Microbiology 159(Pt 1):112–125
78. Iso T, Kedes L, Hamamori Y (2003) HES and HERP families: multiple effectors of the notch signaling pathway. J Cell Physiol 194(3): 237–255
79. Bae S et al (2000) The bHLH gene Hes6, an inhibitor of Hes1, promotes neuronal differentiation. Development 127(13):2933–2943
80. Bolsius YG et al (2021) The role of clock genes in sleep, stress and memory. Biochem Pharmacol 191:114493
81. Zhu K et al (2019) Aryl hydrocarbon receptor pathway: role, regulation and intervention in atherosclerosis therapy (review). Mol Med Rep 20(6):4763–4773