Title. Origin and evolution of eukaryotic transcription factors

Author names and affiliations.
Alex de Mendoza[1,2], Arnau Sebé-Pedrós[3,4]

[1]Australian Research Council Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, Perth, WA, 6009, Australia.
[2]Harry Perkins Institute of Medical Research, Perth, WA, 6009, Australia.
[3]Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain
[4] Universitat Pompeu Fabra, Barcelona, Spain.

Corresponding author. Arnau Sebé-Pedrós (arnau.sebe@crg.es)

Abstract
Transcription factors (TFs) have a central role in genome regulation directing gene transcription through binding specific DNA sequences. Eukaryotic genomes encode a large diversity of TF classes, each defined by unique DNA-interaction domains. Recent advances in genome sequencing and phylogenetic placement of diverse eukaryotic and archaeal species are re-defining the evolutionary history of eukaryotic TFs. The emerging view from a comparative genomics perspective is that the Last Eukaryotic Common Ancestor (LECA) had an extensive repertoire of TFs, most of which represent eukaryotic evolutionary novelties. This burst of TF innovation coincides with the emergence of genomic nuclear segregation and complex chromatin organization.

Introduction
Transcription factors (TFs) are proteins that bind DNA by recognizing specific sequence motifs located at regulatory elements, such as promoters and enhancers. In turn, this TF binding controls downstream chromatin processes such as recruitment of RNA polymerases, DNA methylation, and nucleosome chemical modifications and displacement. The result is the activation or repression of gene expression. Therefore, TFs have a crucial role in interpreting genomic information and are central players in gene regulatory networks. Although TFs are present in all life forms, eukaryotes have a unique set of TF classes, as defined by class-specific DNA binding domains (DBDs) [1]. Some of these TF classes are conserved across large evolutionary distances [2,3].
Eukaryotic genomes tend to be larger than those of prokaryotes. Furthermore, eukaryotic genomic DNA is packed around histone-based nucleosomes that limit the access to genetic information and can carry epigenetic modifications, constituting a complex chromatin environment. Similarly, the origin of the nuclear envelope further changed the way proteins could access and regulate DNA. Therefore, the evolution of a new set of TF classes was likely a pivotal event in the lineage that led to the Last Eukaryotic Common Ancestor (LECA). These ancestral eukaryotic TF classes diversified into large multi-gene families like homeodomain or bHLH TFs [4]. Additionally, new TF classes appeared in specific eukaryotic lineages, further increasing the potential for regulating the genome in more sophisticated manners. This expansion was more

pronounced in plants and animals, both of which encode the most diverse and abundant TF repertoires [3].
This review discusses the emergence and diversification of eukaryotic TF classes, as well as the modes of TF acquisition and the evidence of conserved TF functionality across eukaryotes.

Revisiting Transcription Factor diversity across the tree of life

The continuously growing availability of genome sequence data from key branches of the tree of life is transforming our understanding of the evolution of major eukaryotic gene families. For example, several deep-branching eukaryotic species have recently been either described and/or sequenced for the first time [5–8]. Similarly, the discovery and placement of Asgard archaea as the sister group to eukaryotes reshaped our view on eukaryotic origins [9,10]. Although there is not yet a consensus on the phylogenetic root of eukaryotes, phylogenomic analyses have reduced the potential eukaryotic tree topologies to a few alternative options, which chiefly differ on the phylogenetic position of Discoba and Metamonada [5,11]. Taking advantage of these new genomic data, we reviewed the distribution of a curated list of DBDs representing 74 TF classes in 158 eukaryotic species, 265 archaea and 5,394 bacteria (Figure 1, 2) [12].
Some TF classes have pre-eukaryotic origins. For example, the basal transcription factor machinery is present in multiple archaeal species [13,14], including the TBP (TATA box binding protein), NFYB (Nuclear transcription factor Y subunit beta) and the TFIIB (Figure 2). CSD TFs are also found across all domains of life. Interestingly, some Asgard archaea also encode E2F/TDP, which is a key cell cycle regulator in eukaryotes [15]. This constitutes a new example of a gene family shared between Asgard archaea and eukaryotes but absent from other archeal lineages [9,10], thus reinforcing the view of an Asgard-like ancestor as the initial step towards eukaryogenesis. An additional group of TFs are found in a small number of bacterial species. For example, AP2 and Myb TFs are found in 149 and 257 bacterial species respectively. There are three possible explanations for these observed distributions. First, this could indicate that these TFs have bacterial origins [13]. Second, some bacterial lineages could have acquired eukaryotic TFs through Horizontal Gene Transfer (HGT). Finally, the presence of these eukaryotic TFs in some bacterial genomes could also be explained by contaminations in the genome sequencing/assembly process.
Another possible source of eukaryotic TFs could have been viruses. In particular, giant viruses such as Marseilleviridae have been hypothesized as representatives of a fourth domain of life or as having acquired genes from proto-eukaryotic lineages, such as histone tetramers [16–18]. Intriguingly, some of these giant viruses encode for TFs such as Homeobox or HMG-box that are specific to eukaryotes. However, it is increasingly accepted that giant virus lineages originated multiple times independently, and that most of their genomic repertoire has been acquired from eukaryotic hosts [19–21].
Despite the presence of a few TF classes in non-eukaryotic lineages, the phylogenetic distribution of most other TF classes indicates that they emerged in the lineage leading to the LECA. These include major TF classes such as Homeobox, bZIP or Forkhead. Most phylogenies situate the root of eukaryotes close to Metamonads and/or Discoba . Therefore, depending on the exact topology of the deep branches of the eukaryotic tree of life, the absence of particular TF classes like GATA, bHLH, and HSF in Discoba and Metamonada could change the inferred repertoire of TFs in the LECA. Importantly, including data from free living species of Metamonads offers a

complementary view to the secondarily reduced genomes of many parasites of this lineage. This is illustrated by TFIIA or Forkhead TFs which were considered absent in Metamonads [14] but are, in fact, found in the free-living *Trimastix marina*. Overall, this highlights the need for additional efforts in sampling divergent eukaryotic lineages and to resolve the eukaryotic tree of life to reconstruct the genomic repertoire of the LECA.

Following the initial burst of TF innovation in LECA, novel TF classes emerged in specific eukaryotic lineages (Figure 2). Many of these innovations occurred in the Amorphean lineage and, within this group, in the Opisthokont lineage, which includes animals, fungi and their unicellular relatives. Many novel TFs emerged at the root of Holozoa, comprising animals plus choanoflagellates, filastereans and teretosporeans. This expansion of new TFs was particularly pronounced in animals, both in terms of number of TF classes and number of TFs encoded in animal genomes (Figure 1, 2) [3,22]. More recently, an expansion in TF genes has also been described in multicellular fungi [23]. A similar stepwise TF evolution is observed in the plant lineage, with specific TF classes originating at the root of Chloroplastida (plants and their algal relatives) and later innovation and expansion in the number of TFs at the root of land plants (Figure 1, 2) [24–26].

It is important to note that the observed phylogenetic patterns of TF acquisition are biased by model-system studies. Most TFs were characterized in plant, fungal or animal model species, which at least partially explains why we observe many lineage-specific TF classes in these groups. In contrast, we are very likely missing specific TF classes in other, understudied major eukaryotic lineages.

Modes of transcription factor evolution

The most widespread mechanism of TF diversification is gene duplication. Gene duplication explains the expansion of many TF classes into large multi-gene families and, in many instances, gene duplication comes in hand with novel domain acquisitions. This has been particularly well established in the animal and plant lineages [25,27,28] (Figure 1). Interestingly, some of these duplications date back to the origin of eukaryotes. For example, E2F/TDP is found in single copy in Asgard archaea but in eukaryotes two paralogs are present, E2F and DP, which are known to heterodimerize through their C-terminal domains (Pfam PF08781, Pfam PF16421) [15]. Another example of ancestral LECA paralogs are TALE and non-TALE Homeobox, distinguished by a insertion of 3 amino acids in the TALE sub-class [28].

While gene duplications can explain the expansion of TF classes, it is unclear how entirely new TF classes, with unique DBDs, first emerge. *De novo* gene origin seems to be the most likely scenario to explain the origin of many of these TFs. However, structural similarities between different DBD types might indicate evolutionary affinities obscured by rapid sequence evolution. For example, it has been proposed that Homeobox TFs are derived from Helix-Turn-Helix DBDs [4]. Another mechanism that could have fostered the origin of eukaryotic DBDs is domestication of transposable elements. For example, the plant MUSTANG and FAR/FHY families of TFs evolved from MULE DNA transposons [29,30]. Similarly, many other TFs have been proposed to have originated from transposons in animals and fungi [31,32]. However, transposons also capture sequences from host genomes [33], thus confounding the reconstruction of the evolutionary history of these transposon-derived TFs. Still, given that one of the key events in

eukaryotic history was invasion by transposable elements [34,35], ancestral gain of transposon-derived DBDs could have played an important role in the evolution of LECA.

Conserved TF functions across eukaryotes

Although many TF classes date back to the eukaryotic ancestor, we are ignorant regarding the extent to which they function in a similar manner and whether they mediate similar regulatory programs in different eukaryotic lineages. However, recent analyses of non-conventional model systems provide interesting examples of evolutionarily conserved TF functions or convergent deployment of the same TF classes in similar processes.

The best examples of conserved function across eukaryotes come from TALE Homeobox TFs. Two studies in the moss *Physcomitrella patens* and the unicellular green alga *Chlamydomonas reinhardtii* indicate a conserved role of heterodimerizing TALE homeoboxes in sexual determination in the plant lineage [36–38]. A more recent report showed that this conservation extends to the multicellular brown algae *Ectocarpus siliculosus*, where two heterodimerizing TALE TFs (Ouroboros and Samsara) control sporophyte-gametocyte transitions [39]. A previous report identified two homeobox-like heterodimerizing TFs (MatA and MatB) controlling haploid-to-diploid transitions in the amoebozoan *Dictyostelium discoideum*, although in this case these case it is unclear whether these are highly-divergent homeobox homologs or a lineage-specific TF class [40]. In any case, these results indicate that heterodimerizing TALE homeoboxes are likely linked to an ancient mode of sex determination or, at least, that this system is particularly amenable to be co-opted into this function. Interestingly, TALE homeoboxes are also known to heterodimerize with non-TALE homeoboxes: Hox in animals and MATa1 in yeast [28]. However, while the TALE homeoboxes involved in this heterodimerization are deeply conserved, the interacting non-TALE homeoboxes are later innovations within each lineage [41]. In summary, while the specific dimerization partners may vary in each lineage, the capacity of TALE homeoboxes to heterodimerize seems to be an ancient conserved mechanism present in the LECA.

Other cases of conserved roles of TFs span relatively shorter phylogenetic distances. One such example is the TF Brachyury, a member of the T-box class involved in animal gastrulation and mesoderm differentiation. Analysis of the Brachyury ortholog of the unicellular holozoan *Capsaspora owczarzaki* showed that this distant ortholog could rescue gastrulation and mesoderm specification in the frog *Xenopus*, through recognition of the same DNA binding motifs [42]. Moreover, the inferred *Capsaspora* Brachyury regulatory network and the mouse Brachyury network share target genes linked to actin-based cell motility. This indicates a possible conserved role of this TF in regulating amoeboid cell behavior across more than 800 million years of evolution and predating the origin of animal multicellularity [43].

The conserved binding motifs observed in Brachyury and other T-box TFs across Opisthokonts represent a common theme in TF evolution. Many TF classes have highly conserved core motifs, and specific orthologues conserve identical binding properties across vast evolutionary distances [44–46]. Notable exceptions include Myb/Sant, B3 and, especially, zfC2H2 TF classes, all of which have fast diverging binding motifs [46]. Overall, the DNA sequences that define TF binding can be highly conserved in evolution and constitute a constrained regulatory lexicon. Together, these sequence motifs are essential building blocks of the genetic programs that define eukaryotic cell decisions, from physiological states to developmental processes.

Conclusions

Comparative genomics indicates that few TF classes pre-date the origin of eukaryotes, as these TFs can be found in extant archaea and/or bacterial species (Figure 3). Regardless of the different eukaryogenesis scenarios [47–49], a large number of novel TF classes emerged at the root of eukaryotes. This burst of innovation was accompanied by changes in the nuclear chromatin environment such as the emergence of nucleosomes with protruding histone tails with chemical modifications. In this context, novel TFs could have played a crucial role in LECA genome regulation, mobilizing regulatory processes such as chemical DNA and histone modifications and controlling chromatin accessibility. Later in eukaryotic evolution, additional lineage-specific TF classes emerged and TF repertoires expanded in the plant and animal lineages, concomitantly with the emergence of complex multicellularity.

The study of TF function is still heavily biased towards a handful of model species in the plant, animal and fungal lineages. Still, pioneering studies are uncovering the existence of at least some conserved features across eukaryotes, including TF dimerization networks and TF DNA binding motif preferences. We predict that the phylogenetic expansion of functional TF studies will transform our view on TF function and evolution. This transformation will be unlocked by coupling genomic data with current high-throughput approaches such as *in vitro* TF sequence binding affinity assays, genome-wide profiling of TF binding, and proteomics studies of chromatin beyond model species [50]. Additionally, the establishment of genetic manipulation tools in species representing unsampled eukaryotic lineages will crucially open the window to both targeted studies and genetic screens [51,52]. The comparative analysis and interpretation of these data will ultimately allow us to uncover general principles of TF regulation across eukaryotes and it will contribute to reconstruct the cellular and regulatory biology of the LECA.

Conflict of interest statement

The authors declare no conflict of interest.

Acknowledgements

We thank Matt Brown, Alex Tice, Guifre Torruella, Andrew Roger and Michelle Leger for advice on newly sequenced eukaryotic species and data sources. We thank Daniel J. Richter for commenting on the manuscript. AdM thanks Ryan Lister for mentorship and financial support, ASP is supported by CRG Severo Ochoa.

Figure legends

Figure 1. Distribution of transcription factor classes across eukaryotic species. (A) Barplot showing the total number of TF proteins encoded in the genome/transcriptome of different eukaryotic species. The y axis is square root transformed. (B) Heatmap showing the number of TFs of each class (rows) found in each species. TFs are identified using Pfam HMM profiles for different DNA-binding domains (DBDs) and HMMER3 *hmmsearch* (http://hmmer.org/) searches against predicted proteomes with default gathering threshold (--cut_ga). The total number of proteins encoding a given domain is reported, not the total number of domains (ie. TFs with more than one copy of a particular DBD are counted only once). Asterisks indicate those species for which the genome is not available and transcriptomes were used instead. The transcriptomes were obtained from previous assemblies (one asterisk) [5–7,53] or the publicly available Illumina

reads were downloaded from NCBI Sequence Read Archive (two asterisks) [8,54]. The later were assembled into transcripts using Trinity (https://github.com/trinityrnaseq/) and coding regions were identified using Transdecoder (https://github.com/TransDecoder/). To reduce redundancy in *de novo* transcriptomes, transcripts classified as isoforms of the same gene were counted only once. Finally, proteins that encode more than one DBD domain were counted only once, choosing the DBD with the lowest e-value from the HMM searches.

Figure 2. Transcription factors across the tree of life. Presence (blue) and absence (white) of TF classes in distinct eukaryotic lineages. Major phyletic patterns are subdivided by dashed lines. The phylogenetic relationships among species are based on [6–9,53]. Arrow indicates the TF class (E2F/TDP) shared exclusively by Asgard archaea and eukaryotes. Eukaryotic TF classes found in nucleocytoplasmic large DNA viruses are shown in red and TF classes found in a small subset of bacterial genomes are shown in yellow.

Figure 3. Transcription factors and eukaryogenesis. On the left side, TF classes found in archaea and bacteria are indicated. In the case of bacteria, the contribution to the eukaryotic TF repertoire could be ancestral and/or more recent, depending on different eukaryogenesis scenarios and the timing of different symbiotic events (mitochondria, plastid). The TF classes that were acquired during eukaryogenesis are shown in the grey arrow. A question mark indicates TFs with possible presence in LECA, depending on the exact topology of the eukaryotic tree. On the right side, examples of TF novelties in different eukaryotic groups are indicated. The proposed contribution of viral- or transposable element-derived proteins is indicated with a question mark.

Recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest

•• of outstanding interest

Lambert SA et al. *Cell* 2018 •

The most updated and complete review on human transcription factors, covering all aspects of our current understanding of TF function.

Lambert SA et al. *Nature Genetics* 2019 ••

The authors develop a novel *in silico* TF binding motif prediction strategy to systematically compare TF motifs across animals, plants and fungi. This analysis reveals TF class-specific patterns of binding motif divergence and conservation at an unprecedented scale.

Zaremba-Niedzwiedzka K et al *Nature* 2017 ••

The authors describe a set of new archaeal lineages from metagenomic samples. They find in these genomes key genes previously considered to be unique of eukaryotes.

Lax G et al *Nature* 2018 •

Two species of the neglected Hemimastigophora eukaryotic clade are first described in this work, using transcriptomics to reveal how this lineage likely represents a new deep branching eukaryotic supergroup.

Brown MW et al *Genome Biology and Evolution* 2018 •
Through newly sequencing multiple deep branching eukaryotes, the authors resolve major branching points in the amorphean lineage.

Strassert et al. *Molecular Biology and Evolution* 2019 •
Another recent example of newly resolved phylogenetic position of an 'orphan' eukaryotic group: Telonemia.

Arun A et al *eLife* 2019 ••
The authors discover how the life cycle transition from gametophyte to sporophyte in the brown algae *Ectocarpus siliculosus* is controlled by heterodimeric TALE homeoboxes.

Joo S et al *BMC Biology* 2018 •
A phylogenetic framework establishing deeply conserved TALE homeobox relationships across eukaryotes and additional experimental confirmation of heterodimeric TALE partners in unicellular chlorophytes.

Grau-Bove et al *eLife* 2017 •
The description of several new genomes of unicellular relatives of animals reveal new patterns of conservation across the multicellular-unicellular boundary, including key TF such as LIM homeoboxes.

References

1. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT: **The Human Transcription Factors**. *Cell* 2018, **175**:598–599.

2. Weirauch MT, Hughes TR: **A Catalogue of Eukaryotic Transcription Factor Types, Their Evolutionary Origin, and Species Distribution**. In Edited by Hughes TR. Springer Netherlands; 2011:25–73.

3. de Mendoza A, Sebé-Pedrós A, Sestak MS, Matejcic M, Torruella G, Domazet-Lošo T, Ruiz-Trillo I: **Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages**. *Proc Natl Acad Sci U S A* 2013, **110**:E4858–66.

4. Iyer LM, Anantharaman V, Wolf MY, Aravind L: **Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes**. *Int J Parasitol* 2008, **38**:1–31.

5. Derelle R, Torruella G, Klimeš V, Brinkmann H, Kim E, Vlček Č, Lang BF, Eliáš M: **Bacterial proteins pinpoint a single eukaryotic root**. *Proc Natl Acad Sci U S A* 2015, **112**:E693–9.

6. Strassert JFH, Jamy M, Mylnikov AP, Tikhonenkov DV, Burki F: **New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life**. *Mol Biol Evol* 2019, **36**:757–765.

7. Lax G, Eglit Y, Eme L, Bertrand EM, Roger AJ, Simpson AGB: **Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes**. *Nature* 2018, **564**:410–414.

8. Brown MW, Heiss AA, Kamikawa R, Inagaki Y, Yabuki A, Tice AK, Shiratori T, Ishida K-I, Hashimoto T, Simpson AGB, et al.: **Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group**. *Genome Biol Evol* 2018, **10**:427–433.

9. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, et al.: **Asgard archaea illuminate the origin of eukaryotic cellular complexity**. *Nature* 2017, **541**:353–358.

10. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG: **Complex archaea that bridge the gap between prokaryotes and eukaryotes**. *Nature* 2015, **521**:173–179.

11. He D, Fiz-Palacios O, Fu C-J, Fehling J, Tsai C-C, Baldauf SL: **An alternative root for the eukaryote tree of life**. *Curr Biol* 2014, **24**:465–470.

12. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li W, Chitsaz F, Derbyshire MK, Gonzales NR, et al.: **RefSeq: an update on prokaryotic genome annotation and curation**. *Nucleic Acids Res* 2018, **46**:D851–D860.

13. Iyer LM, Aravind L: **Insights from the architecture of the bacterial transcription apparatus**. *J Struct Biol* 2012, **179**:299–319.

14. Talbert PB, Meers MP, Henikoff S: **Old cogs, new tricks: the evolution of gene expression in a chromatin context**. *Nat Rev Genet* 2019, **20**:283–297.

15. Trimarchi JM, Lees JA: **Sibling rivalry in the E2F family**. *Nat Rev Mol Cell Biol* 2002, **3**:11–20.

16. Erives AJ: **Phylogenetic analysis of the core histone doublet and DNA topo II genes of Marseilleviridae: evidence of proto-eukaryotic provenance**. *Epigenetics Chromatin* 2017, **10**:55.

17. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie J-M: **The 1.2-megabase genome sequence of Mimivirus**. *Science* 2004, **306**:1344–1350.

18. Forterre P, Gaïa M: **Giant viruses and the origin of modern eukaryotes**. *Curr Opin Microbiol* 2016, **31**:44–49.

19. Koonin EV, Yutin N: **Multiple evolutionary origins of giant viruses**. *F1000Res* 2018, **7**.

20. Moreira D, López-García P: **Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes?** *Philos Trans R Soc Lond B Biol Sci* 2015, **370**:20140327.

21. Krupovic M, Dolja VV, Koonin EV: **Origin of viruses: primordial replicators recruiting capsids from hosts**. *Nat Rev Microbiol* 2019, doi:10.1038/s41579-019-0205-6.

22. Sebé-Pedrós A, de Mendoza A, Lang BF, Degnan BM, Ruiz-Trillo I: **Unexpected Repertoire of Metazoan Transcription Factors in the Unicellular Holozoan**

Capsaspora owczarzaki. *Mol Biol Evol* 2011, **28**:1241–1254.

23. Kiss E, Hegedis B, Varga T, Merenyi Z, Koszo T, Balint B, Prasanna AN, Krizsan K, Riquelme M, Takeshita N, et al.: **Comparative genomics reveals the origin of fungal hyphae and multicellularity**. *bioRxiv* 2019, doi:10.1101/546531.

24. Catarino B, Hetherington AJ, Emms DM, Kelly S, Dolan L: **The Stepwise Increase in the Number of Transcription Factor Families in the Precambrian Predated the Diversification of Plants On Land**. *Mol Biol Evol* 2016, **33**:2815–2819.

25. Wilhelmsson PKI, Mühlich C, Ullrich KK, Rensing SA: **Comprehensive Genome-Wide Classification Reveals That Many Plant-Specific Transcription Factors Evolved in Streptophyte Algae**. *Genome Biol Evol* 2017, **9**:3384–3397.

26. Bowman JL, Kohchi T, Yamato KT, Jenkins J, Shu S, Ishizaki K, Yamaoka S, Nishihama R, Nakamura Y, Berger F, et al.: **Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome**. *Cell* 2017, **171**:287–304.e15.

27. Grau-Bové X, Torruella G, Donachie S, Suga H, Leonard G, Richards TA, Ruiz-Trillo I: **Dynamics of genomic innovation in the unicellular ancestry of animals**. *Elife* 2017, **6**.

28. Bürglin TR, Affolter M: **Homeodomain proteins: an update**. *Chromosoma* 2016, **125**:497–521.

29. Joly-Lopez Z, Hoen DR, Blanchette M, Bureau TE: **Phylogenetic and Genomic Analyses Resolve the Origin of Important Plant Genes Derived from Transposable Elements**. *Mol Biol Evol* 2016, **33**:1937–1956.

30. Lin R, Ding L, Casola C, Ripoll DR, Feschotte C, Wang H: **Transposase-derived transcription factors regulate light signaling in Arabidopsis**. *Science* 2007, **318**:1302–1305.

31. Babu MM, Iyer LM, Balaji S, Aravind L: **The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons**. *Nucleic Acids Res* 2006, **34**:6505–6520.

32. Feschotte C: **Transposable elements and the evolution of regulatory networks**. *Nat Rev Genet* 2008, **9**:397–405.

33. de Mendoza A, Bonnet A, Vargas-Landin DB, Ji N, Li H, Yang F, Li L, Hori K, Pflueger J, Buckberry S, et al.: **Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons**. *Nat Commun* 2018, **9**:1341.

34. Koonin Eugene V.: **Viruses and mobile elements as drivers of evolutionary transitions**. *Philos Trans R Soc Lond B Biol Sci* 2016, **371**:20150442.

35. Lee G, Sherer NA, Kim NH, Rajic E, Kaur D, Urriola N, Martini KM, Xue C, Goldenfeld N, Kuhlman TE: **Testing the retroelement invasion hypothesis for the emergence of the ancestral eukaryotic cell**. *Proc Natl Acad Sci U S A* 2018, **115**:12465–12470.

36. Lee J-H, Lin H, Joo S, Goodenough U: **Early sexual origins of homeoprotein heterodimerization and evolution of the plant KNOX/BELL family**. *Cell* 2008, **133**:829–840.

37. Joo S, Wang MH, Lui G, Lee J, Barnas A, Kim E, Sudek S, Worden AZ, Lee J-H: **Common ancestry of heterodimerizing TALE homeobox transcription factors across Metazoa and Archaeplastida**. *BMC Biol* 2018, **16**:136.

38. Horst NA, Katz A, Pereman I, Decker EL, Ohad N, Reski R: **A single homeobox gene triggers phase transition, embryogenesis and asexual reproduction**. *Nat Plants* 2016, **2**:15209.

39. Arun A, Coelho SM, Peters AF, Bourdareau S, Pérès L, Scornet D, Strittmatter M, Lipinska AP, Yao H, Godfroy O, et al.: **Convergent recruitment of TALE homeodomain life cycle regulators to direct sporophyte development in land plants and brown algae**. *Elife* 2019, **8**.

40. Hedgethorne K, Eustermann S, Yang J-C, Ogden TEH, Neuhaus D, Bloomfield G: **Homeodomain-like DNA binding proteins control the haploid-to-diploid transition in Dictyostelium**. *Sci Adv* 2017, **3**:e1602937.

41. Hudry B, Thomas-Chollier M, Volovik Y, Duffraisse M, Dard A, Frank D, Technau U, Merabet S: **Molecular insights into the origin of the Hox-TALE patterning system**. *Elife* 2014, **3**:e01939–e01939.

42. Sebé-Pedrós A, Ariza-Cosano A, Weirauch MT, Leininger S, Yang A, Torruella G, Adamski M, Adamska M, Hughes TR, Gómez-Skarmeta JL, et al.: **Early evolution of the T-box transcription factor family**. *Proceedings of the National Academy of Sciences* 2013, **110**:16050–16055.

43. Sebé-Pedrós A, Ballaré C, Parra-Acero H, Chiva C, Tena JJ, Sabidó E, Gómez-Skarmeta JL, Di Croce L, Ruiz-Trillo I: **The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity**. *Cell* 2016, **165**:1224–1237.

44. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al.: **Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity**. *Cell* 2014, **158**:1431–1443.

45. Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, et al.: **Conservation of transcription factor binding specificities across 600 million years of bilateria evolution**. *Elife* 2015, **4**:1–20.

46. Lambert SA, Yang AWH, Sasse A, Cowley G, Albu M, Caddick MX, Morris QD, Weirauch MT, Hughes TR: **Similarity regression predicts evolution of transcription factor sequence specificity**. *Nat Genet* 2019, **51**:981–989.

47. Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG: **Archaea and the origin of eukaryotes**. *Nat Rev Microbiol* 2017, **15**:711–723.

48. Martin WF, Garg S, Zimorski V: **Endosymbiotic theories for eukaryote origin**. *Philos Trans R Soc Lond B Biol Sci* 2015, **370**:20140330.

49. López-García P, Moreira D: **Open Questions on the Origin of Eukaryotes**. *Trends Ecol Evol* 2015, **30**:697–708.

50. Marinov GK, Kundaje A: **ChIP-ping the branches of the tree: functional genomics and**

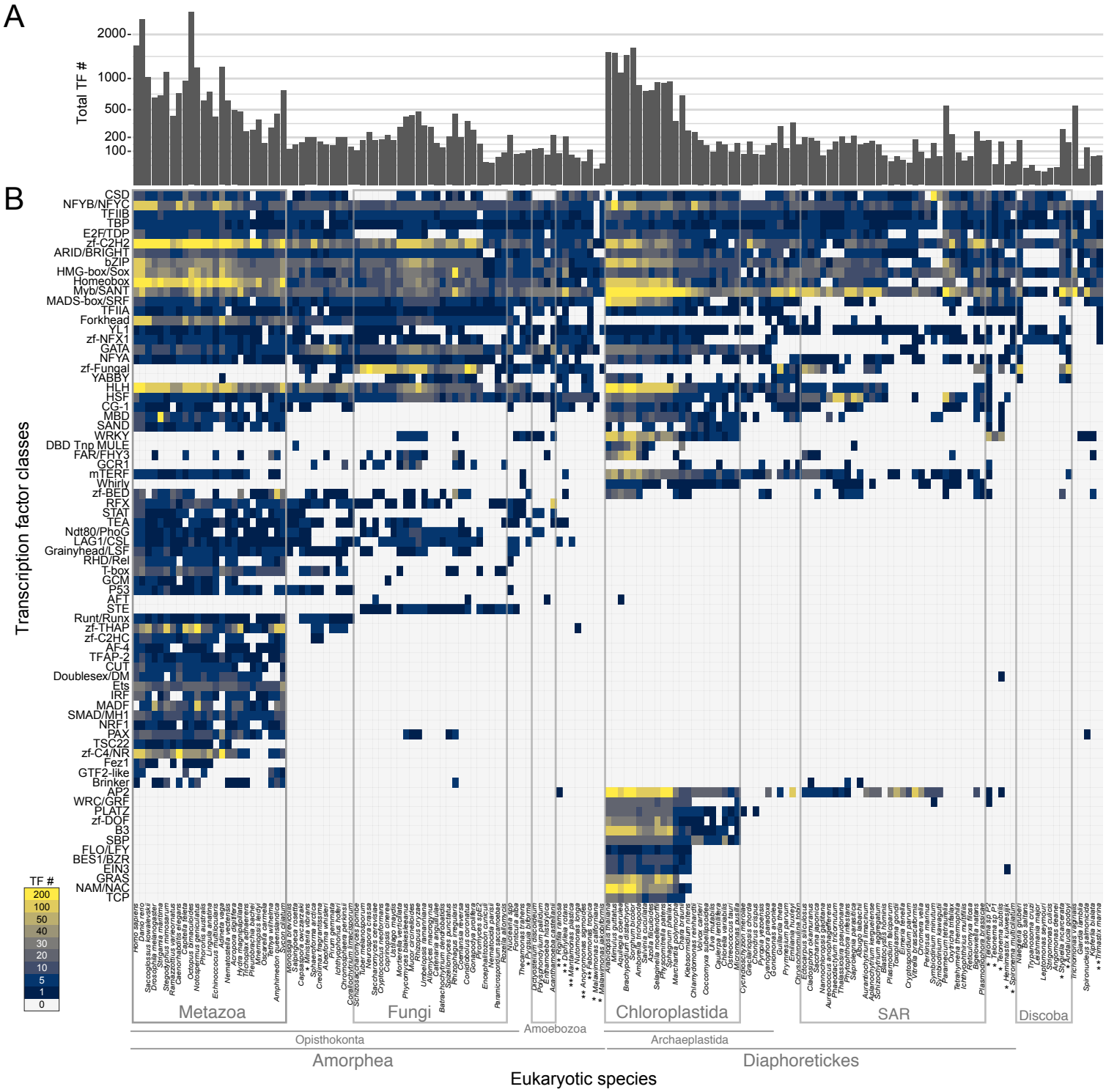
the evolution of eukaryotic gene regulation. *Brief Funct Genomics* 2018, **17**:116–137.

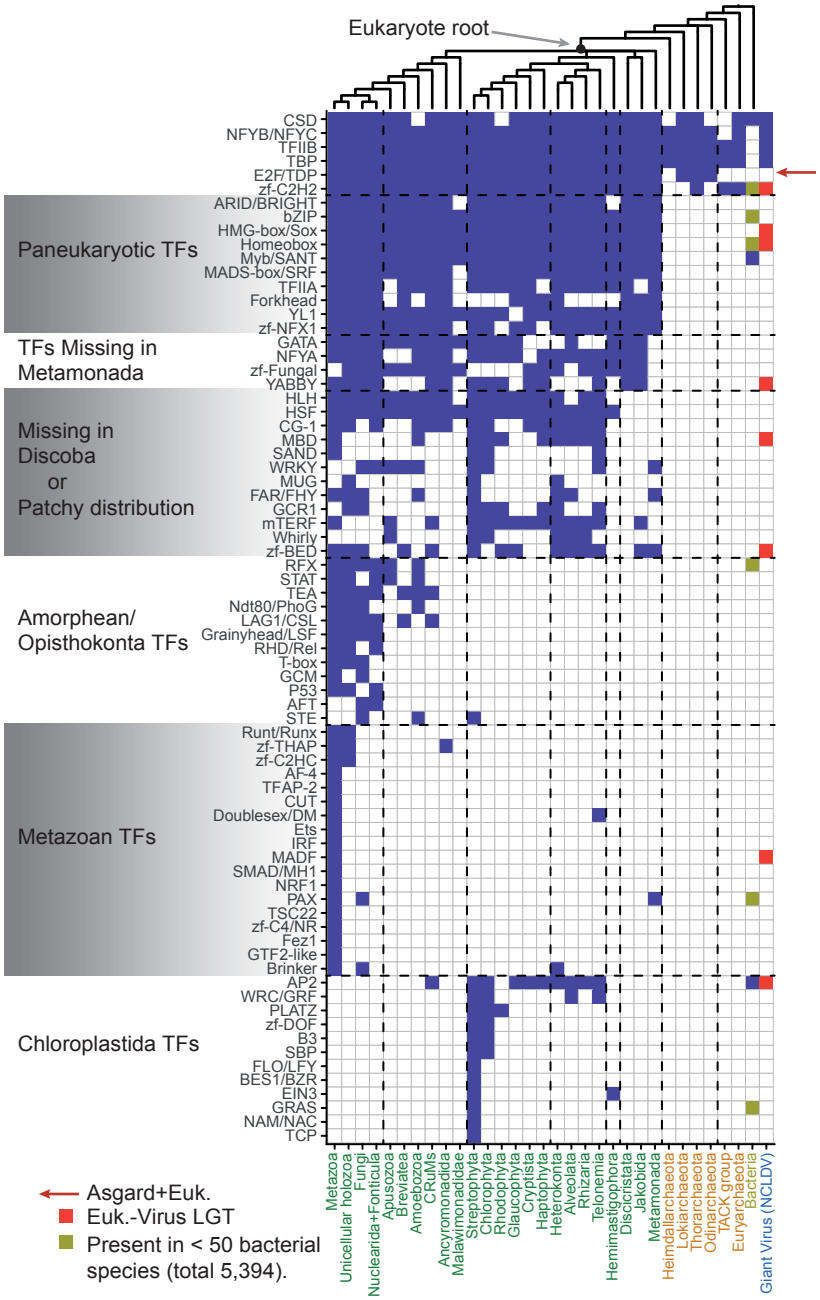
51. Lynch M, Field MC, Goodson HV, Malik HS, Pereira-Leal JB, Roos DS, Turkewitz AP, Sazer S: **Evolutionary cell biology: two origins, one objective**. *Proc Natl Acad Sci U S A* 2014, **111**:16990–16994.

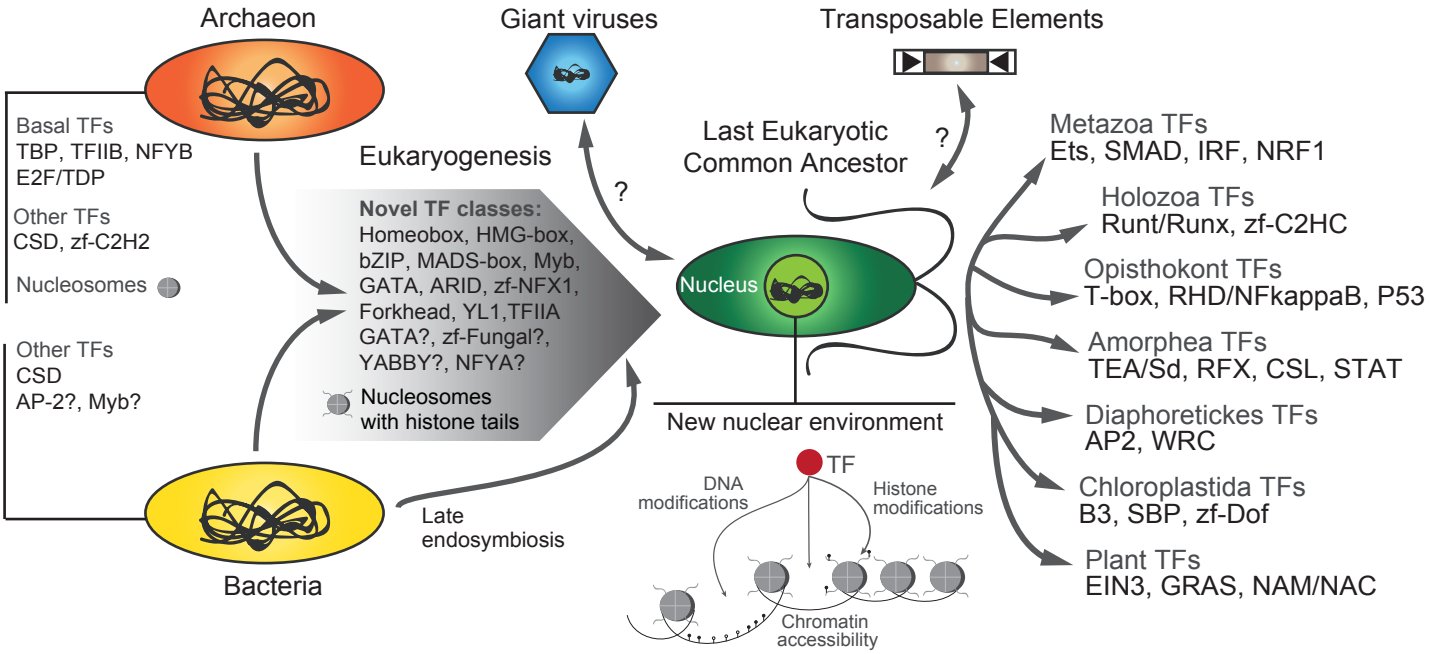
52. Waller RF, Cleves PA, Rubio-Brotons M, Woods A, Bender SJ, Edgcomb V, Gann ER, Jones AC, Teytelman L, von Dassow P, et al.: **Strength in numbers: Collaborative science for new experimental model systems**. *PLoS Biol* 2018, **16**:e2006333.

53. Torruella G, de Mendoza A, Grau-Bové X, Antó M, Chaplin MA, del Campo J, Eme L, Pérez-Cordón G, Whipps CM, Nichols KM, et al.: **Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi**. *Curr Biol* 2015, **25**:2404–2410.

54. Leger MM, Kolisko M, Kamikawa R, Stairs CW, Kume K, Čepička I, Silberman JD, Andersson JO, Xu F, Yabuki A, et al.: **Organelles that illuminate the origins of Trichomonas hydrogenosomes and Giardia mitosomes**. *Nat Ecol Evol* 2017, **1**:0092.


A
Total TF #
2000
1000
500
200
100
B
Transcription factor classes
CSD
NFYB/NFYC
TFIIB
TBP
E2FTDP
zf-C2H2
ARID/BRIGHT
bZIP
HMG-box/Sox
Homeobox
Myb/SANT
MADS-box/SRF
TFIIA
Forkhead
YL1
zf-NFX1
GATA
NFYA
zf-Fungal
YABBY
HLH
HSF
CG-1
MBD
SAND
WRKY
DBD Tnp MULE
FAR/FHY3
GCR1
mTERF
Whirly
zf-BED
RFX
STAT
TEA
Ndt80/PhoG
LAG1/CSL
Grainyhead/LSF
RHD/Rel
T-box
GCM
P53
AFT
STE
Runt/Runx
zf-THAP
zf-C2HC
AF-4
TFAP-2
CUT
Doublesex/DM
Ets
IRF
MADF
SMAD/MH1
NRF1
PAX
TSC22
zf-C4/NR
Fez1
GTF2-like
Brinker
AP2
WRC/GRF
PLATZ
zf-DOF
B3
SBP
FLO/LFY
BES1/BZR
EIN3
GRAS
NAM/NAC
TCP
TF #
200
100
50
40
30
20
10
5
1
0
Metazoa
Fungi
Amoebozoa
Chloroplastida
SAR
Discoba
Opisthokonta
Archaeplastida
Amorphea
Diaphoretickes
Eukaryotic species


Eukaryote root
CSD
NFYB/NFYC
TFIIB
TBP
E2F/TDP
zf-C2H2
ARID/BRIGHT
bZIP
HMG-box/Sox
Homeobox
Myb/SANT
MADS-box/SRF
TFIIA
Forkhead
YL1
zf-NFX1
GATA
NFYA
zf-Fungal
YABBY
HLH
HSF
CG-1
MBD
SAND
WRKY
MUG
FAR/FHY
GCR1
mTERF
Whirly
zf-BED
RFX
STAT
TEA
Ndt80/PhoG
LAG1/CSL
Grainyhead/LSF
RHD/Rel
T-box
GCM
P53
AFT
STE
Runt/Runx
zf-THAP
zf-C2HC
AF-4
TFAP-2
CUT
Doublesex/DM
Ets
IRF
MADF
SMAD/MH1
NRF1
PAX
TSC22
zf-C4/NR
Fez1
GTF2-like
Brinker
AP2
WRC/GRF
PLATZ
zf-DOF
B3
SBP
FLO/LFY
BES1/BZR
EIN3
GRAS
NAM/NAC
TCP
Paneukaryotic TFs
TFs Missing in Metamonada
Missing in Discoba or Patchy distribution
Amorphean/ Opisthokonta TFs
Metazoan TFs
Chloroplastida TFs
Metazoa
Unicellular holozoa
Fungi
Nucleariida+Fonticula
Apusozoa
Breviatea
Amoebozoa
CRuMs
Ancyromonadida
Malawimonadidae
Streptophyta
Chlorophyta
Rhodophyta
Glaucophyta
Cryptista
Haptophyta
Heterokonta
Alveolata
Rhizaria
Telonemia
Hemimastigophora
Discicristata
Jakobida
Metamonada
Heimdallarchaeota
Lokiarchaeota
Thorarchaeota
Odinarchaeota
TACK group
Euryarchaeota
Bacteria
Giant Virus (NCLDV)
Asgard+Euk.
Euk.-Virus LGT
Present in < 50 bacterial species (total 5,394).


Archaeon
Giant viruses
Transposable Elements
Basal TFs
TBP, TFIIB, NFYB
E2F/TDP
Other TFs
CSD, zf-C2H2
Nucleosomes
Other TFs
CSD
AP-2?, Myb?
Eukaryogenesis
Novel TF classes:
Homeobox, HMG-box,
bZIP, MADS-box, Myb,
GATA, ARID, zf-NFX1,
Forkhead, YL1,TFIIA
GATA?, zf-Fungal?,
YABBY?, NFYA?
Nucleosomes
with histone tails
?
Last Eukaryotic
Common Ancestor
?
Nucleus
New nuclear environment
Late
endosymbiosis
Bacteria
TF
DNA
modifications
Histone
modifications
Chromatin
accessibility
Metazoa TFs
Ets, SMAD, IRF, NRF1
Holozoa TFs
Runt/Runx, zf-C2HC
Opisthokont TFs
T-box, RHD/NFkappaB, P53
Amorphea TFs
TEA/Sd, RFX, CSL, STAT
Diaphoretickes TFs
AP2, WRC
Chloroplastida TFs
B3, SBP, zf-Dof
Plant TFs
EIN3, GRAS, NAM/NAC