Genomics 114 (2022) 110390

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

Phylogeny of NF-YA trans-activation splicing isoforms in vertebrate evolution

Andrea Bernardini [*,1], Alberto Gallo, Nerina Gnesutta, Diletta Dolfini, Roberto Mantovani [*]

Dipartimento di Bioscienze, Università degli Studi di Milano, Via Celoria 26, 20133 Milano, Italy

ARTICLE INFO

Keywords:
Transcription factor
Alternative splicing
Transactivation domain
Intrinsically disordered region
Glutamine-rich
NFYA
Evolution

ABSTRACT

NF-Y is a trimeric pioneer Transcription Factor (TF) whose target sequence –the CCAAT box– is present in ~25% of mammalian promoters. We reconstruct the phylogenetic history of the regulatory NF-YA subunit in vertebrates. We find that in addition to the remarkable conservation of the subunits-interaction and DNA-binding parts, the Transcriptional Activation Domain (TAD) is also conserved (>90% identity among bony vertebrates). We infer the phylogeny of the alternatively spliced exon-3 and partial splicing events of exon-7 –7N and 7C– revealing independent clade-specific losses of these regions. These isoforms shape the TAD. Absence of exon-3 in basal deuterostomes, cartilaginous fishes and hagfish, but not in lampreys, suggests that the "short" isoform is primordial, with emergence of exon-3 in chordates. Exon 7N was present in the vertebrate common ancestor, while 7C is a molecular innovation of teleost fishes. RNA-seq analysis in several species confirms expression of all these isoforms. We identify 3 blocks of amino acids in the TAD shared across deuterostomes, yet structural predictions and sequence analyses suggest an evolutionary drive for maintenance of an Intrinsically Disordered Region –IDR– within the TAD. Overall, these data help reconstruct the logic for alternative splicing of this essential eukaryotic TF.

1. Introduction

Regulation of transcriptional initiation is at the heart of gene expression, and, as such, of all fundamental processes in living organisms. It is controlled by the binding of Transcription Factors – TFs – to short DNA sequences in promoters and enhancers of genes. The production of TFs represents a sizeable portion –7/8%– of the protein-coding capacity of the human genome [1]. Structurally, they are minimally composed of two domains. (i) The DNA-binding domain –DBD– required for recognition of the specific DNA element [2]; many DBDs have been structurally characterized in complex with DNA, and TFs binding sites have been systematically assessed [3]. (ii) A Transcription Activation Domain –TAD– involved in the activatory function, typically by interacting with coactivators/repressors and/or the General Transcription Factors (GTFs). Structurally, much less is known on TADs. In most cases, TF gene families are expanded across evolution, with new members acquiring neofunctionalization [1,4]. Diversification typically impacts less on the DBDs, constrained by the requirement for DNA sequence-specificity, than on other parts of proteins, including the TADs. This often concerns expression, with new members being expressed in novel territories, divergent timing or environmental conditions.

NF-Y (Nuclear Factor Y, or CBF CCAAT Binding Factor) is a TF that binds to the CCAAT box, one of the first DNA elements identified in human promoters (Reviewed in [5]). It is formed by three subunits, present in all eukaryotes. The DNA-binding parts are structurally well characterized in *fungi*, mammals and plants [6,7,8]. With respect to other TFs, the system presents several distinct features. First, unlike most TFs, three subunits are required for robust DNA-binding. NF-YB and NF-YC belong to a large family sharing the Histone Fold Domain –HFD– resembling to H2B/H2A [9]. NF-YA provides the trimer with exquisite sequence-specificity. Second, NF-YA does not form a large gene family other than in plants, where a remarkable expansion –8/15 members– and diversification with a sister family termed CCT (CONSTANS, CO-like, TOC1) is observed. CCTs maintain a similar structure with a slightly different DNA-binding specificity [10,11]. Indeed, NF-YA is one of few TFs that does not belong to an overtly expanded gene

Abbreviations: NF-YAl, NF-YA long isoform; NF-YAs, NF-YA short isoform.

* Corresponding authors.
E-mail addresses: andrea.bernardini@igbmc.fr (A. Bernardini), mantor@unimi.it (R. Mantovani).
[1] Present affiliation: Institut de Génétique et de Biologie Moléculaire et Cellulaire, 67404, Illkirch, France; Centre National de la Recherche Scientifique, UMR7104, 67404, Illkirch, France; Institut National de la Santé et de la Recherche Médicale, U964, 67404, Illkirch, France; Université de Strasbourg, 67404, Illkirch, France.

https://doi.org/10.1016/j.ygeno.2022.110390
Received 22 November 2021; Received in revised form 2 May 2022; Accepted 12 May 2022
Available online 16 May 2022

family in metazoans. Third, the asymmetric target DNA sequence –R,R, C,C,A,A,T,C/G,A/G– is well positioned in promoters at −60/−120 [12], impacting both on the Transcriptional Start Site –TSS– choice and on the establishment of nucleosome-free areas [13].

Because of variability in expression levels, NF-YA is considered the regulatory subunit of the trimer. A mouse model with complete KO of NF-YA is lethal at early embryo stages [14]. Conditional KO in various adult tissues –neurons, adipocytes, hepatocytes, hematopoietic system– lead to variable, but ultimately severe, consequences (reviewed by [15]). In humans, the NF-YA gene is composed of 10 exons, spanning ~27 kb in the short arm of chromosome 6. The protein coding sequence starts within exon-2 and ends within exon-10 (Fig. 1A). The core domain that defines NF-YA, originally named after the *Saccharomyces cerevisiae* ortholog HAP2, is composed of two distinct subdomains: the A1 α-helix, responsible for heterotrimerization with the NF-YB/NF-YC dimer, and a DNA-recognition subdomain, composed of the A2 α-helix followed by the GXGGRF motif [8]. The trimer hosts two separate TADs, located on the N-terminal of NF-YA and the C-terminal of NF-YC [16,17,18,19,20,21]. Both domains are involved in Alternative Splicing (AS).

Like the majority of protein coding genes, TF genes undergo AS, sculpturing expression patterns of specific isoforms during growth, differentiation, development, or following cellular transformation. The significance and biological implications of the different layers of regulation of the alternative products are being increasingly understood globally, *via* the sequencing of genomes and the identification of isoforms panels by RNA-seq.

Mammalian NF-YA presents different mRNA species depending on the cell-type, developmental stage and physiopathology, and alternative 3′-untranslated regions (UTR) of different lengths are annotated [19]. All protein isoforms retain the HAP2 domain, hence a similar capacity to interact with the HFD subunits and bind the CCAAT box. Historically, the first AS isoform described excludes exon-3, generating a shorter protein –NF-YAs, Nuclear Factor YA short– devoid of 28–29 aa, while the inclusion of exon-3 is translated in the long isoform NF-YAl, Nuclear Factor YA long (see [19]). This exon-3 stretch harbors glutamines, hydrophobic residues and a scarcity of charged amino acids, with a compositional bias very much like the surrounding TAD. Two additional single codon splicing variants have been reported for exon-3 [22,19], which we refer to as L2 and L5 isoforms, arising from the use of an alternative acceptor or donor splice sites at the boundaries of exon-3 (Fig. S1A). A second AS event identified in human/mouse involves the removal of a segment at the 5′ portion of exon-7 –that we name exon-7N– caused by the use of a downstream acceptor splice site located within the exon (Fig. S1A). The resulting protein lacks a valine-rich hexapeptide, VTVPVS, located in the TAD. Functional data on transactivation of the Cystathionine-β-Synthase promoter support the relevance of L2, L5 and 7N variations, at least in the NF-YAl configuration, including in synergy with Sp1 [22]. Finally, a third AS event was recently reported, involving elimination of both exon-5 and exon-3. This shorter isoform (named NF-YAx), with altered functional activity, was found in glioblastoma cells and in a specific window of mouse embryo development [23]. In summary, all AS events involve the N-terminal part of NF-YA, which contains a functional TAD.

We noticed a lack of top-down systematic characterization of NF-YA evolution, specifically on the conservation and variation of its alternatively spliced regions. The aim of our work was to investigate these aspects and, specifically, the origins of isoforms in the context of a TAD.

2. Results

2.1. NF-YA orthologues in vertebrates

A scheme of human NF-YA gene structure is shown in Fig. 1A. To explore the origins of alternatively spliced regions, we looked in animal groups for orthologous sequences, generating multiple sequence alignments (MSA). We used the 56 amino acids HAP2 domain shared by yeast, plants and metazoans (Fig. 1A), suitable to retrieve *bona fide* NF-YA genes in all eukaryotes [24,25,19,26,27]. We found protein sequences annotated as CBFB_NFYA in the Pfam database, belonging to all major groups of eukaryotes, notably 490 species of *fungi*, 376 species of metazoans, including 215 Chordata (Fig.1B). Initially, we decided to focus exclusively on sequences from vertebrates. We ended up with a total of 482 different NF-YA protein sequences of 220 vertebrate species.

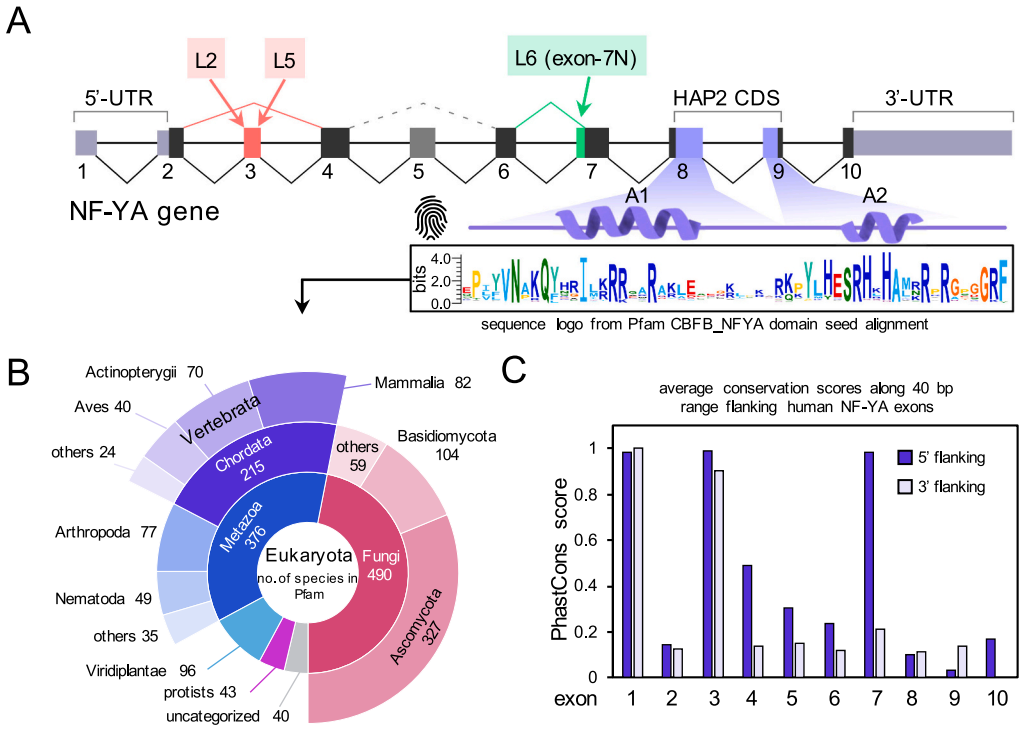


Fig. 1. *NF-YA* genes in all major eukaryotic taxa.
A. Scheme of the human *NF-YA* gene structure. Known alternative splicing events are indicated. The conserved HAP2 domain is highlighted in violet and the corresponding protein seed alignment from distantly related species in Pfam database is represented as sequence logo. The secondary structure elements found in mammalian, yeast and plant crystal structures are indicated above (A1 and A2 helices). B. Taxonomic distribution of the species retrieved in Pfam database as having a match for NF-YA HMM built from the seed alignment shown in A. The number of species for each taxon is indicated. C. Evaluation of evolutionarily conserved intronic elements flanking human *NF-YA* exons. The average PhastCons score along 40 bp upstream (5′ flanking) or downstream (3′ flanking) each exon is reported. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As shown in Fig. S1B and C, the HAP2 core domain is almost perfectly conserved in all species. The TAD is also almost identical within amniotes (mammals, birds and reptiles): the human protein differs in 1 and 6 positions from mouse and alligator orthologs (same isoform comparison). This extends the conservation previously reported in human/mouse/frog [28,19], but it makes sequence comparison in amniotes less informative. Importantly, annotations of exon-3 and exon-7 splicings (Fig. S1A) are frequently present in amniotes and fishes, but those of the exon-5 spliced NF-YAx isoform are far less common (Fig. S1C).

Splicing relies on the presence of intronic regulatory motifs: one expects to find higher conservation close to alternatively spliced junctions, and lower around constitutive ones. We evaluated the conservation score of DNA sequences flanking human exons: indeed, the highest scores are found in sequences flanking exon-3 and the region upstream from exon-7 (Fig. 1C). Regions surrounding exon-1 are also conserved: this is predicted for 5′ sequences, which are part of the core promoter; less so for the 3′, potentially hinting at regulation of the 5′UTR region. These results suggest that AS of NF-YA exon-3 and exon-7N are conserved features in vertebrates.

Two additional features were apparent in fishes: duplication of NF-YA genes and the presence of additional AS involving the C-terminal of exon-7; we focus on these findings, before we present our results in isoform-centered sections.

2.2. Gene duplications and independent losses in Teleostei

Zebrafish (*Danio rerio*) has two annotated copies of *NF-YA*, named *nfya* and *nfyal* (nuclear factor Y alpha like), located on Chr11 and Chr8, respectively (Fig. 2A). Sequence analysis unequivocally confirms their paralogy, with *nfyal* showing 82% protein identity with *nfya,* and a similar gene structure. The ancestor of teleost fishes underwent a whole genome duplication event known as the teleost-specific WGD (Ts3R), followed by a rapid gene loss phase [29,30]. The distribution of *NF-YA* copies in ray-finned fish species is shown in Fig. 2B and S2. Non-teleost actinopterygians –reedfish (*Erpetoichthys calabaricus*) and spotted gar (*Lepisosteus oculatus*)– have a single NF-YA gene, while the distribution in Teleostei is more complex: single copy in bony tongues (Osteoglossiformes) and in all Neoteleostei, double copy in northern pike (*Esox lucius*) and Otophysi, to which zebrafish and nearly two-thirds of freshwater fish species belong [31]. Among Otophysi, the two species of Characiformes –red-bellied piranha (*Pygocentrus nattereri*) and Mexican tetra (*Astyanax mexicanus*)– have a single-copy *NF-YA* gene, possibly representing a secondary gene loss in this clade. At the opposite, the Cyprininae have additional paralogues: the common carp (*Cyprinus carpio*) and golden-line barbel (*Sinocyclocheilus grahami*) have three *NF-YA* paralogs, goldfish (*Carassius auratus*) five. This can be explained by the relatively recent carp-specific WGD (Cs4R) generated by an allotetraploidization event, dated about 12.4 Mya [32]. The fourth gene could have been lost in carp and retained in others (*e.g.* goldfish). Yet, the Cs4R

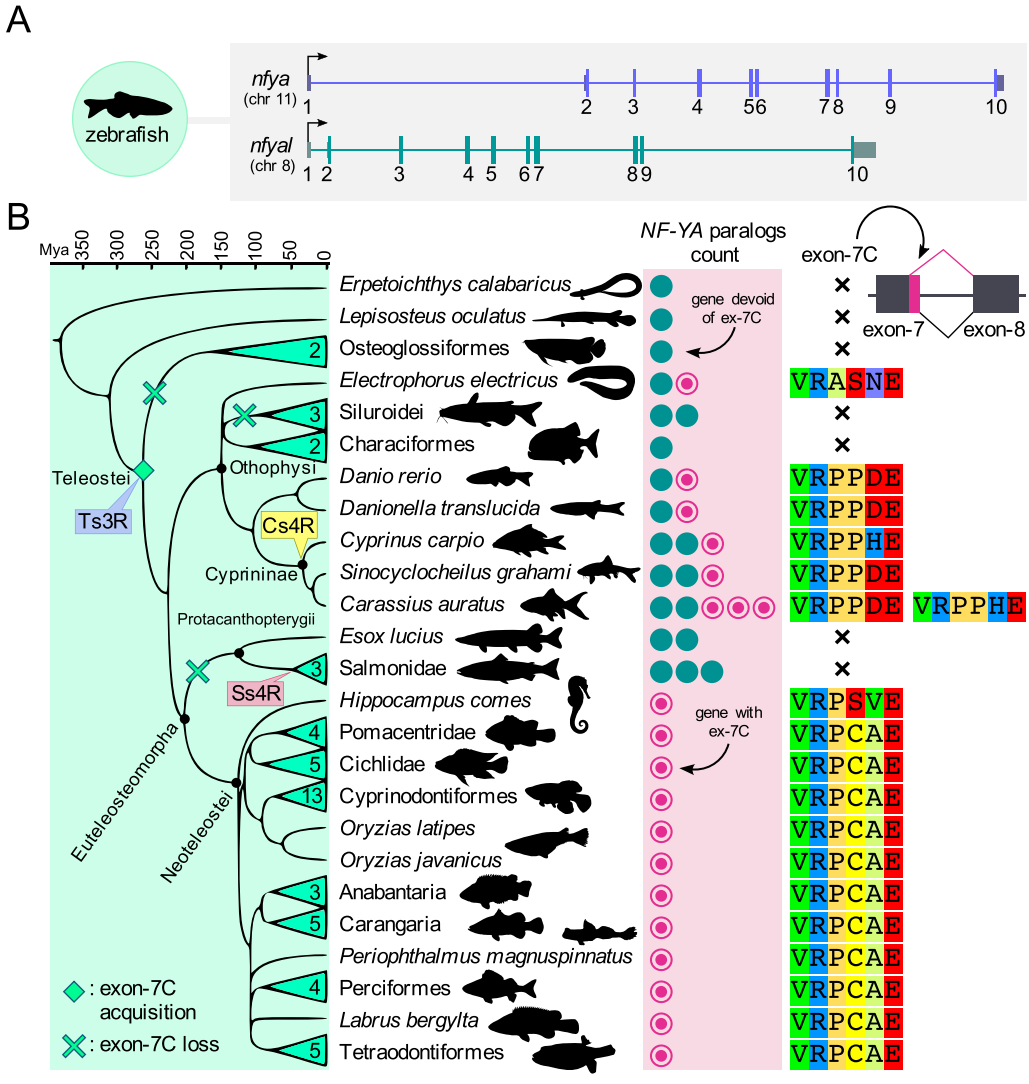


Fig. 2. *NF-YA* paralogs and exon-7C in teleost fishes.
A. Gene structure of the two *NF-YA* paralogs found in zebrafish. B. Phylogenetic time-tree of 63 species of ray-finned fishes (left panel) checked for the number of *NF-YA* paralogous genes (central red panel) and the exon-7C extension sequence (right panel). The number of species collapsed in the same leaf is indicated. Documented WGD events are indicated at the corresponding nodes: teleost-specific (Ts3R), carp-specific (Cs4R) and salmon-specific (Ss4R). The hypothetical phylogeny of exon-7C is overlaid on the tree. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

event would not explain the presence of a fifth copy in goldfish. Two goldfish paralogues located in the same chromosome (Chr8) were generated by an inverted duplication involving a ~ 250 kb region encoding at least 14 genes (Fig. S3). Finally, three paralogs are found in Salmonidae, likely representing the product of a further WGD specific to this clade (Ss4R) [33]. In summary, after the teleost specific WGD, one of the two gene copies was independently lost in bony tongues, in the vast group of Neoteleostei and in the order Characiformes. In the rest of teleost groups, two *NF-YA* paralogues were maintained. An additional gene copy, generated by independent WGD specific to ancestors of salmonids and carps was fixed in their descendants.

2.3. The exon-7C isoforms in Teleostei

The second novelty in fishes is a 6 aa extension at the end of exon-7, deriving from the usage of an alternative downstream donor splice site (Fig. 2B). We refer to this as exon-7C. None of the species analyzed here other than Teleostei possess exon-7C. In zebrafish, exon-7C is only found in *nfyal* –VRPPDE– and its expression is confirmed by cDNA clones and annotation in RefSeq data. We further defined exon-7C distribution at DNA level among fish species (Fig. 2B). Within Otophysi, exon-7C is shared among electric eel (*Electrophorus electricus*) and Cyprinidae family, in one of the two *NF-YA* paralogues, while it is absent in Siluroidei and Characiformes. Exon-7C is absent in bony tongues and in Protacanthopterygi (northern pike, salmons and trouts). It is present as VRPCAE in the single-copy gene of all species belonging to the major group of Neoteleostei (bottom clades in Fig. 2B). This short C-terminal extension of exon-7 presents an amino acid composition markedly different from the rest of NF-YA TAD, harboring charged side chains and even a cysteine in Neoteleostei, a residue otherwise absent in vertebrate NF-YA. Exon-7C likely arose at the basement of Teleostei, as a splicing alternative in one of the two NF-YA paralogs, following the fate of its host gene in some descendant lineages. In others, such as the salmonid paralogues, it was probably lost independently. In northern pike, belonging to salmonids sister group Esociformes, an incomplete exon-7 extension –VRPL– (not shown) is possibly a remnant of the loss started in the common ancestor of these two groups.

2.4. The exon-7N isoform predates jawed vertebrates and is lost in ray-finned fishes

The hexad peptide –VTVPVS– encoded by exon-7N can be excluded from the protein product by the recognition of a downstream alternative acceptor splice site (Fig. S1A). Exon-7N is identical in all mammals, shared by Sauropsida and amphibians with a Ser/Thr substitution (VTVPVT) (Fig. 3A). The only sarcopterygian fish included in our dataset, the coelacanth *Latimeria chalumnae*, has this hexad peptide, unlike ray-finned fishes (Actinopterygii). Chondrichthyes possess exon-

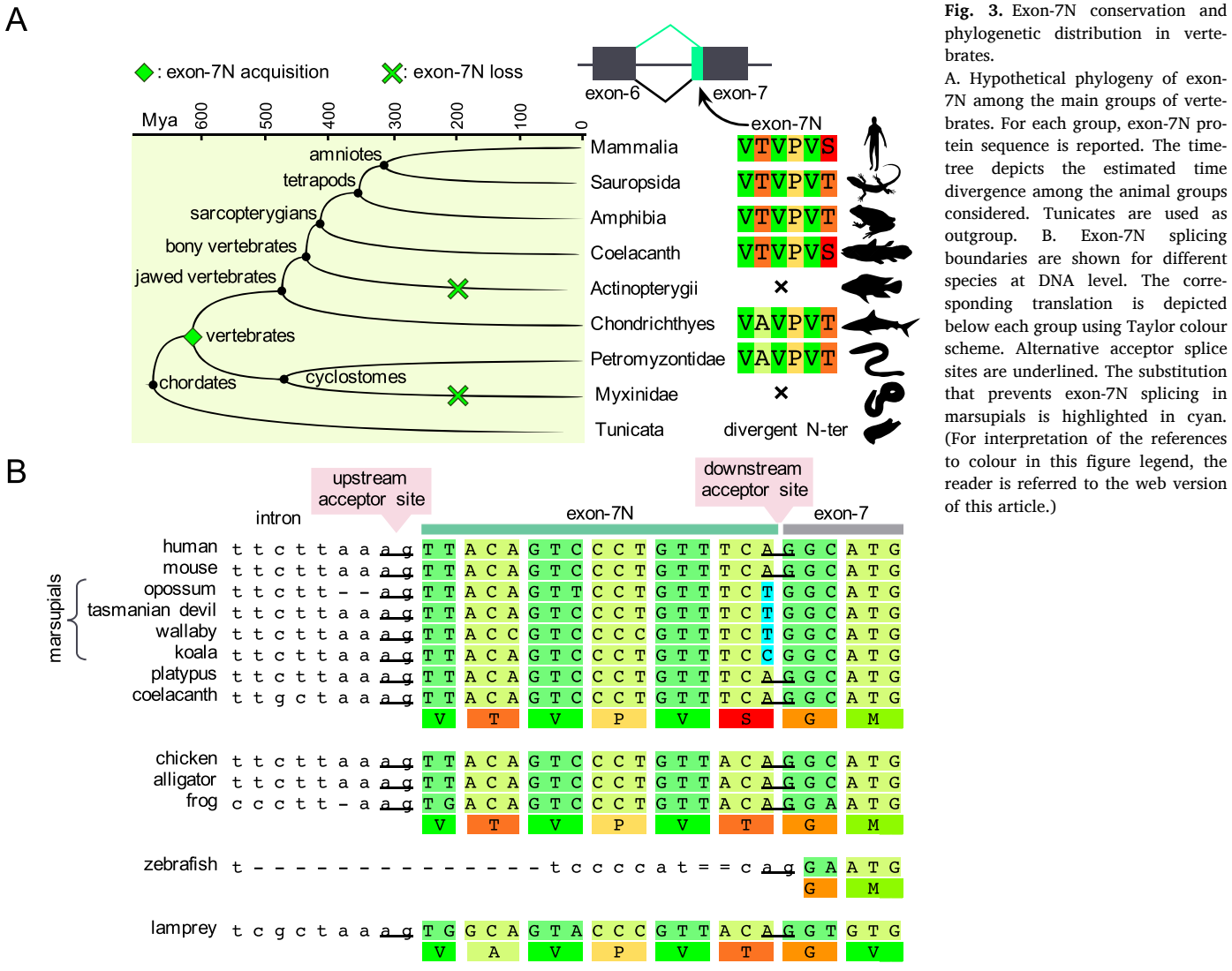


Fig. 3. Exon-7N conservation and phylogenetic distribution in vertebrates.
A. Hypothetical phylogeny of exon-7N among the main groups of vertebrates. For each group, exon-7N protein sequence is reported. The time-tree depicts the estimated time divergence among the animal groups considered. Tunicates are used as outgroup. B. Exon-7N splicing boundaries are shown for different species at DNA level. The corresponding translation is depicted below each group using Taylor colour scheme. Alternative acceptor splice sites are underlined. The substitution that prevents exon-7N splicing in marsupials is highlighted in cyan. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

7N with an Ala substitution (VAVPVT) (Fig. 3A), shared with lamprey. Myxinidae (hagfish) lack exon-7N, missing an additional acceptor splice site before an in-frame stop codon. As for expression, we found evidence of alternatively spliced exon-7N, either in the form of EST clones (not shown) or in RNA-seq data (see Section 2.8). Thus, NF-YA exon-7N peptide represents an ancient accessory feature present at the basement of vertebrate tree, and independently lost in hagfish and all ray-finned fishes (Fig. 3A).

Both splice junctions are under selective pressure, as the acceptor splice site (AG) internal to exon-7 is conserved, being composed by the third position of a Ser codon (A) and the first position of the subsequent Gly codon (G) (Fig. 3B). Thereby, for the canonical splice site to be maintained, one expects a constrained codon usage for Ser at third position (TCA) relative to the other five synonymous codons for this amino acid. Indeed, even species with the above-mentioned Ser/Thr substitution use the only Thr codon with A at third position (ACA) (Fig. 3B): this suggests selective pressure to maintain a splicing-competent exon-7N. Within the species analyzed here, the exception is represented by marsupials, where Ser is encoded by TCT or TCC codons (Fig. 3B), disrupting the acceptor splice site and forcing exon-7N inclusion in the mature mRNA. A BLAST search on translated RNA-seq datasets of opossum testis, brain and muscle using as query the protein sequence resulting from exon-7N skipping did retrieve only reads containing exon-7N. These observations suggest that marsupials lost the ability to splice exon-7N due to elimination the exon-7 acceptor splice site.

2.5. NF-YA exon-3 is absent in cartilaginous fishes

We verified the conservation of regions homologous to the alternatively spliced mammalian exon-3. The sequence logo derived from a set of species representative of the main vertebrate groups is shown in Fig. 4A. Inspection of MSA returns 100% identity in all amniotes (126 species), the only exception being a single S/T substitution in turtles (Fig. 4B). L2 and L5 isoforms splice sites, adding a single Gln at the 5′ and Val at the 3′ boundaries, are found in all amniotes (Fig. 4B). The four amphibians–*Xenopus tropicalis*, *Xenopus laevis*, *Microcaecilia unicolor* and *Rhinatrema bivittatum*– have exon-3, with a substitution of 4 aa in *Xenopus* and one in caecilian (Fig. 4B). In *Xenopus,* one of the substitutions –Gln25Pro– abolishes the L2 isoform, while L5 is preserved.

All bony fishes carry exon-3 (Fig. 4B), including the coelacanth. In

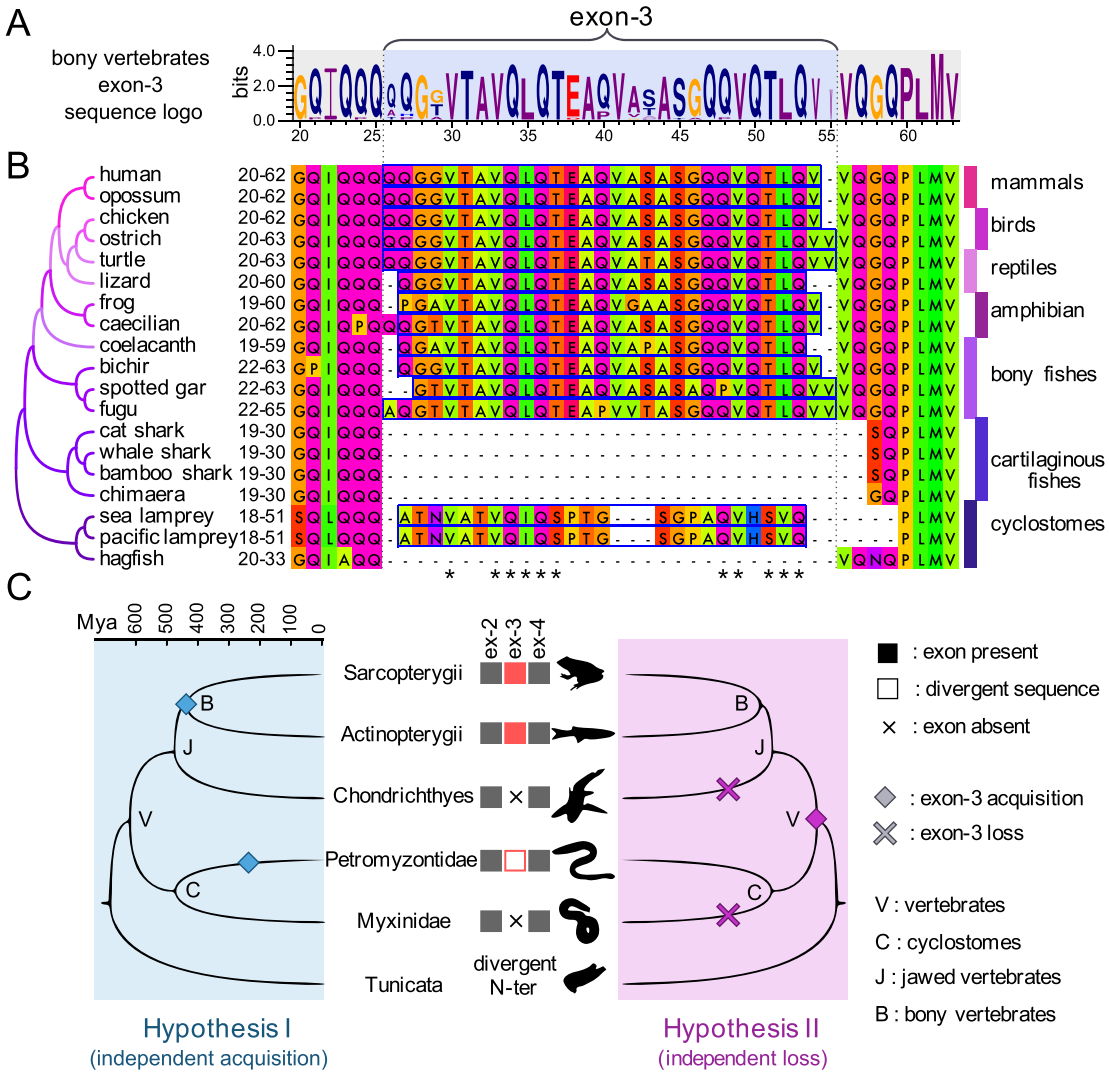


Fig. 4. Exon-3 conservation and phylogenetic distribution in vertebrates.
A. Sequence logo of NF-YA exon-3 and adjacent regions derived from a MSA of 15 species representative of the bony vertebrate groups possessing canonical exon-3. Species belonging to the following groups were chosen: mammals, sauropsids, amphibians, non-teleost and teleost fishes. B. Exon-3 MSA in different taxonomic groups. Species are arranged according to their phylogenetic relationship. Numbering indicates the range of the protein shown in the alignment. Cartilaginous fishes and hagfish are devoid of exon-3. Asterisks indicate positions with high sequence similarity with lampreys' exon-3. C. Phylogenetic distribution of exon-3 in vertebrates (central Panel) and the two hypothetical phylogenies of this exon (lateral Panels). Tunicates are used as outgroup.

Actinopterygii, two of the most basal groups, bichirs (*E. calabaricus*) and holosteans (spotted gar), possess an intact exon-3 with two substitutions, along with accessory L5, but not the L2. A distinct feature is deletion of a stretch of 5 aa within exon-3 –VVTAS– in a group of small fresh water teleosts, the Cyprinodontiformes (Fig. S4). Being part of Neoteleostei, this group has a single copy for *NF-YA* (Fig. 2B), making the internally deleted exon-3 the sole option for the NF-YA long isoform. Positions of this deleted stretch show higher variability among the rest of vertebrates (Fig.4A).

These changes notwithstanding, the most surprising result was the lack of exon-3 in all five species of cartilaginous fishes (Fig. 4B): four elasmobranchs (bamboo shark, *Chiloscyllium punctatum*; cat shark, *Scyliorhinus torazame*; whale shark, *Rhincodon typus*; thorny skate, *Amblyraja radiata*) and one species of Holocephali (elephant shark, *Callorhinchus milii*). To rule out incomplete or erroneous exon annotation, we inspected each gene at DNA level by distinct approaches: *de novo* gene predictions with different methods [34,35], tblastn searches, analysis of 9 available RNA-seq of the elephant shark (Fig. S5, See Materials and Methods). Note that (i) surrounding sequences of the N-terminal TAD are otherwise well conserved, and (ii) inspection of *C. milii* RNA-seq data (Fig. S5) did confirm the sole expression of the NF-YAs isoform. We conclude that cartilaginous fishes are devoid of exon-3, hence unable to generate the NF-YA long isoform.

2.6. Cyclostomes present a heterogeneous arrangement for exon-3

Absence of exon-3 in cartilaginous fishes could hint at the ancestral –plesiomorphic– condition of all vertebrates, or it could be a derived trait: a genetic loss secondary to the separation of this lineage from the rest of vertebrates (Fig. 4C).

To gather insights on this point, we first turned to jawless fishes (Cyclostomata), a small group of extant jawless vertebrates, including two classes that present several, rather primitive traits: sea lampreys (*Petromyzon marinus*) and hagfishes. Gene-level analysis of the lamprey sequence revealed the presence of an intervening CDS between canonical exon-2 and exon-4, annotated as exon. We retrieved sequences from lamprey RNA-seq datasets from 14 different embryo stages and adult tissues and we annotated them on the kPetMar1 genome assembly (Not shown). In addition, we retrieved 100% identity matches to this 'exon-3' in cDNA clones and RNA-seq datasets from two other species: the arctic lamprey *Lethenteron japonicum* and brook lamprey (*Lampetra planeri*). This exon-3-like sequence encodes a 24 aa stretch, shorter and apparently divergent from the vertebrate consensus. However, by aligning sequences with the inclusion of gaps, similarity raises considerably, with two blocks of reasonably conserved positions, anchored on the Gln, hydrophobics and S/T residues (Fig. 4B). Thus, the presence of sequences reminiscent of exon-3 can be reasonably proposed for lampreys.

Hagfishes are jawless fishes with unique physio-anatomic features. In the past, they were placed outside of the vertebrate tree, but recent phylogenomics and developmental observations support their relocation in the Cyclostomata, as lampreys' sister group [36,37,38]. We reconstituted the complete NF-YA protein sequence of *Eptatretus burgeri* hagfish, the inspection of which revealed a conventional N-terminus devoid of the exon-3 sequence (Fig. 4B). The sequence included in our initial dataset lacked a considerable portion of protein, N-terminal of exon-5, possibly due to ambiguous nucleotides in the genome assembly localized ~3 kb upstream putative exon-5. We then searched the ENA browser (https://www.ebi.ac.uk/ena/data/sequence/search) using the first annotated exon as query and indeed retrieved matches for hagfish cDNAs that contain the 5' CDS corresponding to the complete N-terminus (Fig. S6A). We back-mapped the cDNA sequence on the genome assembly to define the missing exon locations: exons-2 and -4 were mapped in a different contig, separated by a short intron (81 bp) (Fig. S6A). Yet, no evidence of an intervening 'exon-3' sequence was found. In addition, no reads from *E. burgeri* tissues could be mapped to this sequence. Searches in translated RNA-seq and genomic sequences, using either lamprey or spotted gar exon-3 amino acids also did not retrieve any match. Both lampreys and hagfish are known to undergo a process of massive somatic loss of genomic material during development [39,40,41], potentially explaining failure to retrieve the hagfish sequence. The lamprey genome data are derived from germline, the hagfish from germline and somatic cells: tblastn searches in RNA-seq sequence reads repository from *E. burgeri* testis (ERX2120222, ERX2120223) or embryos (SRX2541849, SRX2541848, SRX2541847, SRX2541846) using lampreys' exon-3 as query also failed to retrieve any match. Finally, searches on a second hagfish species for which RNA-seq datasets are available –*E. cirrhatus*– were also negative (Fig. S6C). We therefore feel confident to exclude the presence of NF-YA exon-3 in the hagfish genome and, consequently, the expression of NF-YAl.

2.7. Analysis of NF-YA TAD

We decided to evaluate NF-YA according to conservation of amino acids sequence, using a method –LIST– that takes into account conservation in orthologs, as well as the taxonomic distance across species, thus providing a score of "substitution deleteriousness" for each amino acid [42]. Fig. 5A shows the results: expectedly, the HAP2 domain is the most conserved, as confirmed by the identity analysis from representative species against the human protein (Fig. S7). This region is heavily charged with a prevalence of arginines (Fig. 5C). Some variability is observed in the linker region between A1 and A2 (See also Fig. 1A). A near perfect conservation is also reported for A1 and A2 residues not directly involved in NF-YB/NF-YC or DNA-contacts, lying on the outer surface of the trimeric complex bound to DNA: this can best be explained assuming that these residues have a functional relevance, for example for providing contacts with neighboring TFs, co-activators or the General Transcription Machinery.

It has been suggested that protein domains involved in AS are often intrinsically disordered [43], and it is also known that TADs are rich in Intrinsically Disordered Regions, IDRs [44,45,46,47]. The results of several structural disorder prediction tools applied to NF-YA is shown in Fig. 5B. The consensus points at a widespread prevalence of structural disorder along the protein, except for the HAP2 A1. Overall, exons 2–6 have a high score of predicted disorder (Fig. 5B), in line with the bias in aa composition: high in disorder promoting residues Gln (24%), Gly (13%) and depleted of charged and aromatic amino acids (Fig. 5C–D). Yet, the conservation of the area coded by exons-4/5/6 is distinctly higher with respect to exon-2/3. The edges of exon-3 and the N-term of exon-7, together with the exon-10C-term portion are the most variant (Fig. 5A).

To gather further insights on potential structural motifs in the TAD, we used AlphaFold [48] to predict NF-YA structure from human and other vertebrate species. The models faithfully recapitulated the alpha-helical structure of the HAP2 domain (Fig. 5E). On the other hand, the TAD was modelled with low confidence scores and extreme conformational heterogeneity, as expected for an IDR (Fig. 5E and Fig. S8A). Nonetheless, several of the models display recurrent β-stranded secondary structure motifs, also shared by models from different species (Fig. S8B). They include a three-stranded β-sheet in exons 4–5, a β-hairpin in exon-6 and a second three-stranded β-sheet in exon-7 (Fig. S8C). Notably, both exon-3 and exon-7N are not part of these putative motifs and were invariably modelled as random coils (Fig. 5E). We conclude that there is a differential amino acid conservation score among the exons of the N-terminal TAD and that exon-3 lies within a large region predicted to be intrinsically disordered. Moreover, we speculate that the recurrent low-confidence structural motifs identified in the TAD might represent alternative, dynamic conformations which are populated upon binding to specific interactors.

The dissimilar deleteriousness scores of exon-2/3 *vs* exon-4/5/6 led us to further consider NF-YA sequences in other deuterostomes. This could also shed light on the hypotheses outlined in Fig. 4C, related to the phylogenies of exon-3. Sequence analysis of the tunicate *Ciona*

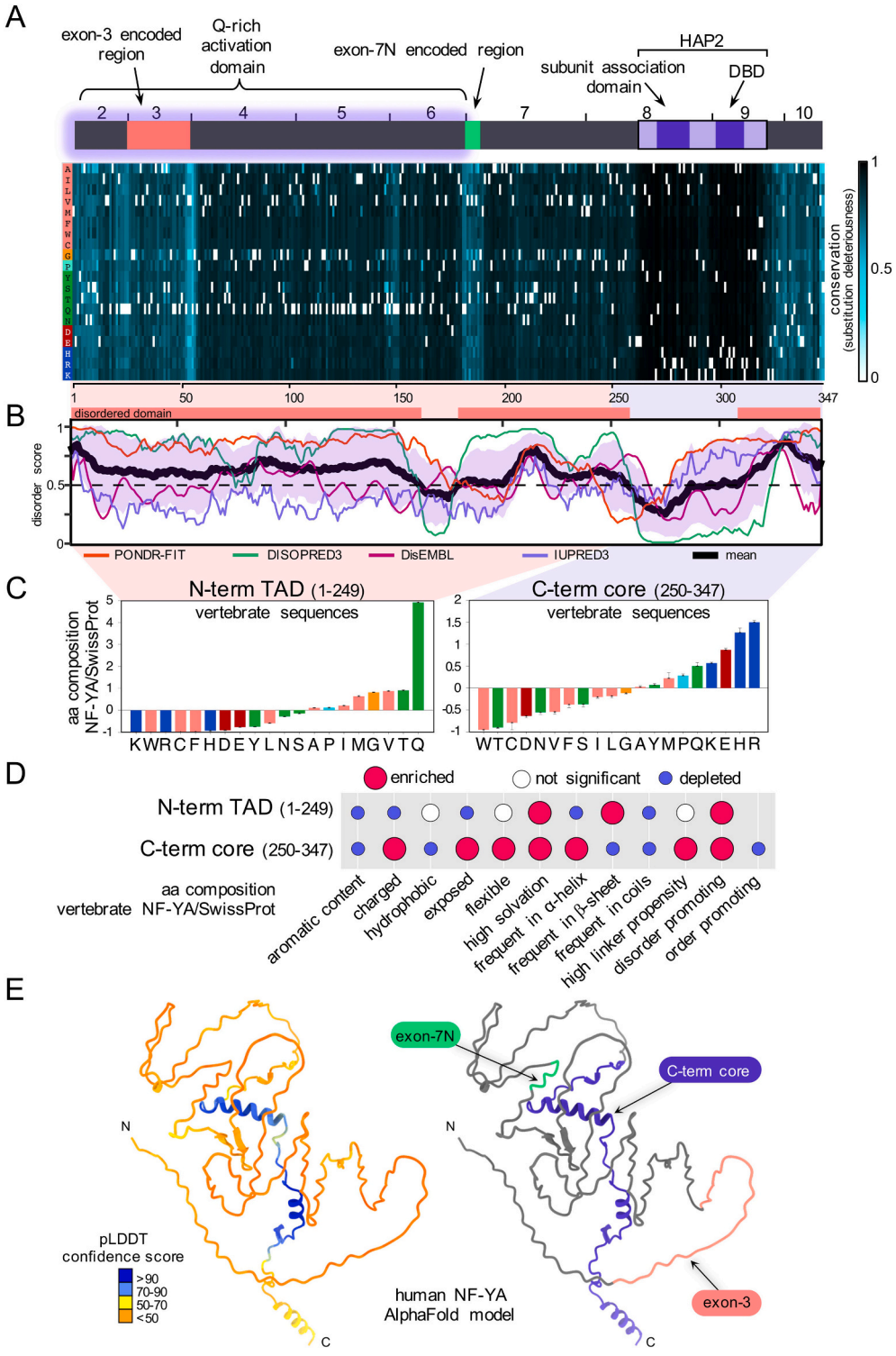


Fig. 5. NF-YA domains conservation, disorder and sequence composition.
A. NF-YA protein annotation and conservation. Exon boundaries, regions affected by AS and functional domains are indicated. The heatmap shows the deleteriousness scores computed for every possible substitution at each position along human NF-YA, as defined by LIST [42], which takes into account position conservation in orthologs and taxonomic distance across the correspondent species. The higher the score for a given substitution, the less frequently that substitution is observed in orthologs. B. Structural disorder prediction for human NF-YA protein. The plot shows the consensus probability score (black line) derived from averaging several predictors (colored lines). Regions above the 0.5 threshold are annotated as intrinsically disordered (red bars). The violet shaded range represents standard deviation. C. Analysis of the enrichment/depletion patterns of individual amino acids in the two NF-YA functional domains: N-term TAD (1–249) and C-term core (250–347, numbering according to human protein). A set of 19 vertebrate NF-YA orthologs was compared with a reference collection of SwissProt proteins. The same amino acid colour code is applied as in A. D. The same dataset was analyzed in terms of groups of amino acids classified by different physico-chemical properties. Significantly enriched or depleted groups are indicated by red or blue circles, respectively. E. Human NF-YA AlphaFold structural model colored according to confidence score (left panel). The alternatively spliced regions within the disordered TAD are indicated on the right panel. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

intestinalis was poorly informative, due to the scarce sequence conservation within the TAD (~14% identity, not shown), although the amino acid composition was similar. We considered lancelet (*Branchiostoma lanceolatum*, Cephalochordata), feather star (*Anneissia japonica,* Echinodermata) and acorn worm (*Saccoglossus kowalevskii,* Hemichordata). The gene structures are similar, 9 exons in lancelet and feather star, 8 in acorn worm (Fig. 6A). Alignments of protein sequences are shown in Fig. 6B: the amino acid composition of the N-terminal is somewhat similar, although the overall sequence identity is 37%, 33% and 27% for lancelet, acorn worm and feather star, respectively, when compared to the human sequence. An exon topologically corresponding to exon-3 is present in lancelet, but absent in feather star and acorn worm. With manual annotation, conservation patterns emerge: (i) stretches of similarity are visible in the areas corresponding to exons-4/5/6 in all species: Block I at the edge of exon-4 and 5, characterized by an alternation of glutamine and hydrophobic residues interrupted by a proline. Block II within exon-5, made of a short isoleucine-rich hydrophobic patch followed by a pair of glutamines. The 10 aa long Block III in exon-6, characterized by the pattern (E/D)GQTΦΦYQPV (Φ, hydrophobic aa). Notably, Block I and III are part of the recurrent putative structural

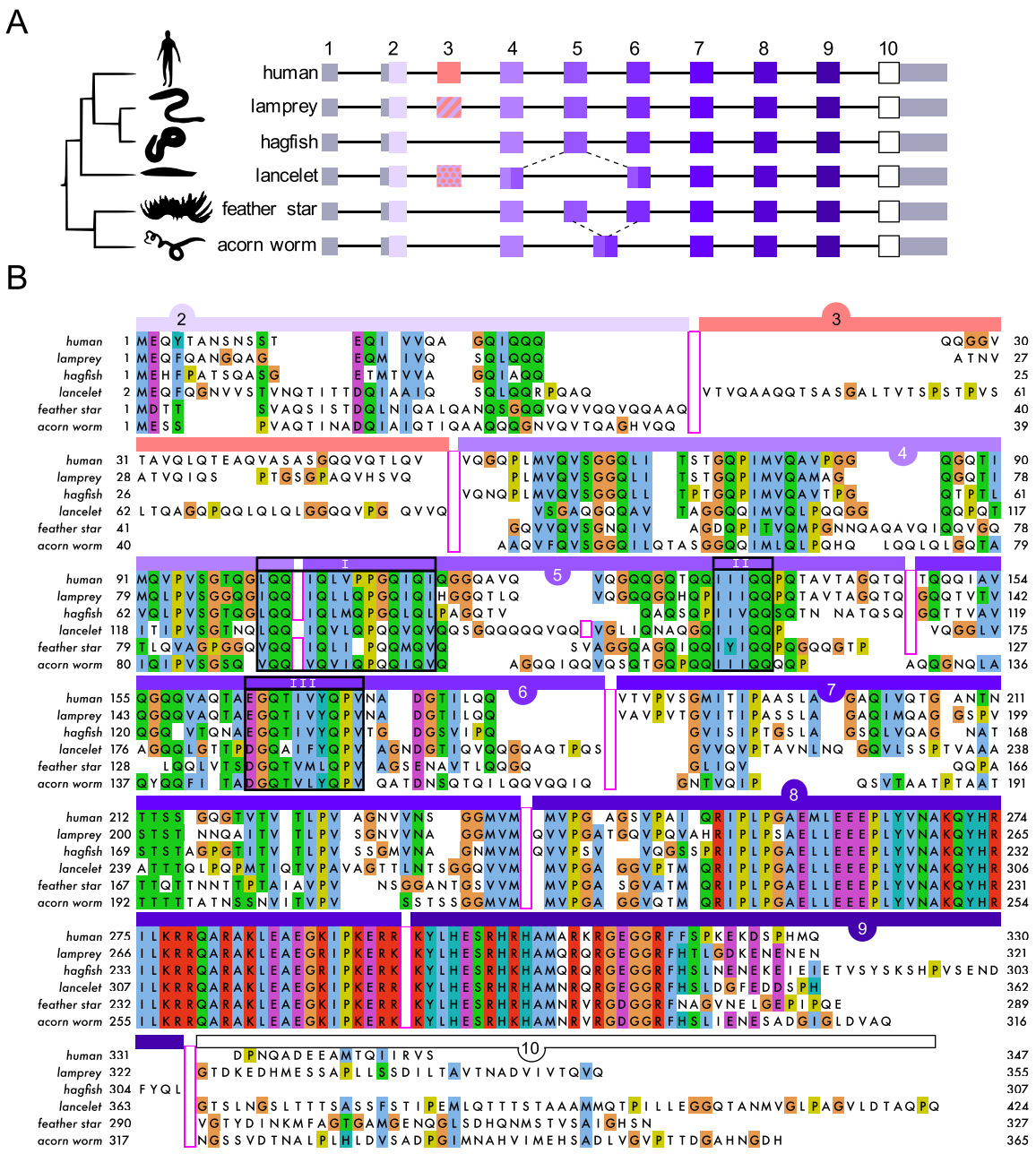


Fig. 6. NF-YA gene structure and conservation blocks in other deuterostomes.
A. Intron-exon structure of *NF-YA* gene orthologs in different species. Exon numbering is referred to human gene. Exons for which a recognizable amino acid signature could be traced among the considered species have the same colour. In lancelet 5′ and 3′ halves of canonical exon-5 seem to be part of canonical exon-4 and 6, respectively. In acorn worm canonical exon-5 and 6 sequences are part of a single exon. B. Manually edited MSA of NF-YA protein sequences from the species reported in A. The alignment was edited according to exon boundaries (indicated by the open pink boxes) and colored according to Clustal colour scheme. Conserved motifs in the TAD are highlighted in black boxes. Lamprey (*P. marinus*), hagfish (*E. burgeri*), lancelet (*B. lanceolatum*), feather star (*Anneissia japonica*), acorn worm (*Saccoglossus kowalevskii*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

motifs identified by AlphaFold. (ii) Exon-2 shows some similarities, limited to the very N-term. (iii) Exon-3 sequences are absent in feather star and acorn worm, while a 50 aa-long stretch positioned between exon-2 and exon-4 is present in amphioxus: this represents exon-3 sequences, not only semantically, but also effectively, as judged by the Q-rich content (26%), presence of hydrophobics and absence of aromatic and charged residues. Inspection of another lancelet, *Branchiostoma floridae,* confirms the presence of the 50 aa exon-3 sequences. Finally, in keeping with previously published sequences [49,50], the NF-YA of sea urchin (*Strongylocentrotus purpuratus*), another echinoderm, lacks exon-3 sequences (not shown).

In summary, despite primary sequences divergence, a similar genomic organization is reported in all deuterostomes; an exon-3 is present in basal chordates, but not in more distantly related deuterostomes (echinoderms and emichordates), suggesting that NF-YA short is the ancestral form of this subunit.

2.8. Expression of NF-YA isoforms in vertebrates

The data shown above suggest, but do not prove, that the genomic sequences are actually incorporated in mRNA species. A systematic analysis of NF-YA isoforms expression in different organisms has never

been performed. We therefore analyzed RNA-seq of 17 species for which datasets of three adult tissues –brain, liver, skeletal muscle– are available. To avoid confusion with the annotations, we mapped all reads from the raw NGS data to each of the *NF-YA* exons, focusing on exon-3, 7N and, in fish, 7C. Note that it was not possible to compute L2 and L5, due to technical limitations to quantify the presence of the single triplets at the N- and C- terminal of exon-3. Fig. 7 shows the results. NF-YAl, as determined by exon-3 reads, is expressed in brain of all species, and comparison of exon-2 and exon-4 reads suggests that it is predominant; in muscle, NF-YAl also predominates, with lower expression in goldfish and Atlantic cod (*Gadus morhua*). A relevant exception is Axolotl (*Ambystoma mexicanum*), in which NF-YAs is the only isoform present. In liver, only in cat and pig exon-3 reads are scored, in line with NF-YAs being the predominant isoform. All these data are in agreement with what is known for expression of the two major isoforms in mouse and man.

As for exon-7N, it is also generally present in brain and muscle, with lower expression in alligator and toad. It is also expressed in liver, being predominant in many species. Note that there is no immediate correlation between expression of exon-3 and 7N. Again, Axolotl appears to be the only species in which exon-7N sequences are missing. Exon-7C in fish species showed variable levels of expression in the tissues considered, with the exception of Atlantic cod, in which the isoform is absent. In summary, the expression data of distantly related vertebrate species concur that NF-YAs and NF-YAl, 7N and 7C (in fishes) are expressed in adult tissues; exon-7N appears to be regulated independently from exon-3.

Finally, because of the presence of a 50 aa exon-3 in amphioxus, we wished to ascertain whether this would also be alternatively spliced, an indication that the mechanism would be introduced in the common ancestor of chordates. We searched for isoforms in available RNA-seq datasets of lancelet: analysis of TPMs coverage of all exons only detected one isoform and no evidence for splicing of the 50 aa exon-3 (Fig. S9). Hence, it seems reasonable to conclude that AS of NF-YA TAD appeared in the common ancestor of vertebrates.

3. Discussion

This study sheds light on the phylogenetic history of NF-YA, notably of the N-terminal TAD and its splicing isoforms. Alternative splicing is a key event that variegates the compendium of gene products in eukaryotes. Although exceptions exist, AS events in TFs are more often found in domains outside of the DBDs, involved in greater evolutionary divergence. Our findings are summarized in Fig. 8: gene expansion is observed in teleost fishes and AS events in the TAD. AS appear in vertebrates, involving the N- and C-term ends of exon-7 and, especially, exon-3. They are producing independently regulated mRNA isoforms in various tissues of the species examined (Fig. 7). We further identified three blocks of conserved stretches within the TAD, shared by deuterostomes.

3.1. The "original" NF-YA

Our data indicate that the two major isoforms in vertebrates are NF-YAl and NF-YAs. We were particularly intrigued by the question as to which was the "original" isoform of the primordial vertebrate: to find the answer, we inquired further back, in other deuterostomes. The absence of exon-3 sequences in echinoderms and hemichordates, and presence in amphioxus, supports the hypothesis that NF-YAs was present in the common ancestor of all extant deuterostomes. A relevant point is the homology of exon-3 sequences of amphioxus and lampreys to that of bony vertebrates: they differ in length – 50 and 24 aa *vs* 28/29 aa, respectively– and sequence. However, they are similar in aa composition. The alignment of lampreys' exon-3 shows a sequence identity of ~29% with human (Fig. 4), falling within the so-called 'twilight zone' of protein alignment to infer homology [51] and showing two stretches of high sequence similarity; in addition, a stretch of identity –GQQVxxQVVQ– is observed in amphioxus at the C-terminal of exon-3, including the beginning of human exon-4 (29% identity to human sequence). Note that a certain degree of sequence flexibility is permitted even in bony vertebrates, as proven by an internal deletion of exon-3 found in Cyprinodontiformes (Fig. S4), by variability in the extremities and by the presence of internal substitutions observed in several species, including amphibians.

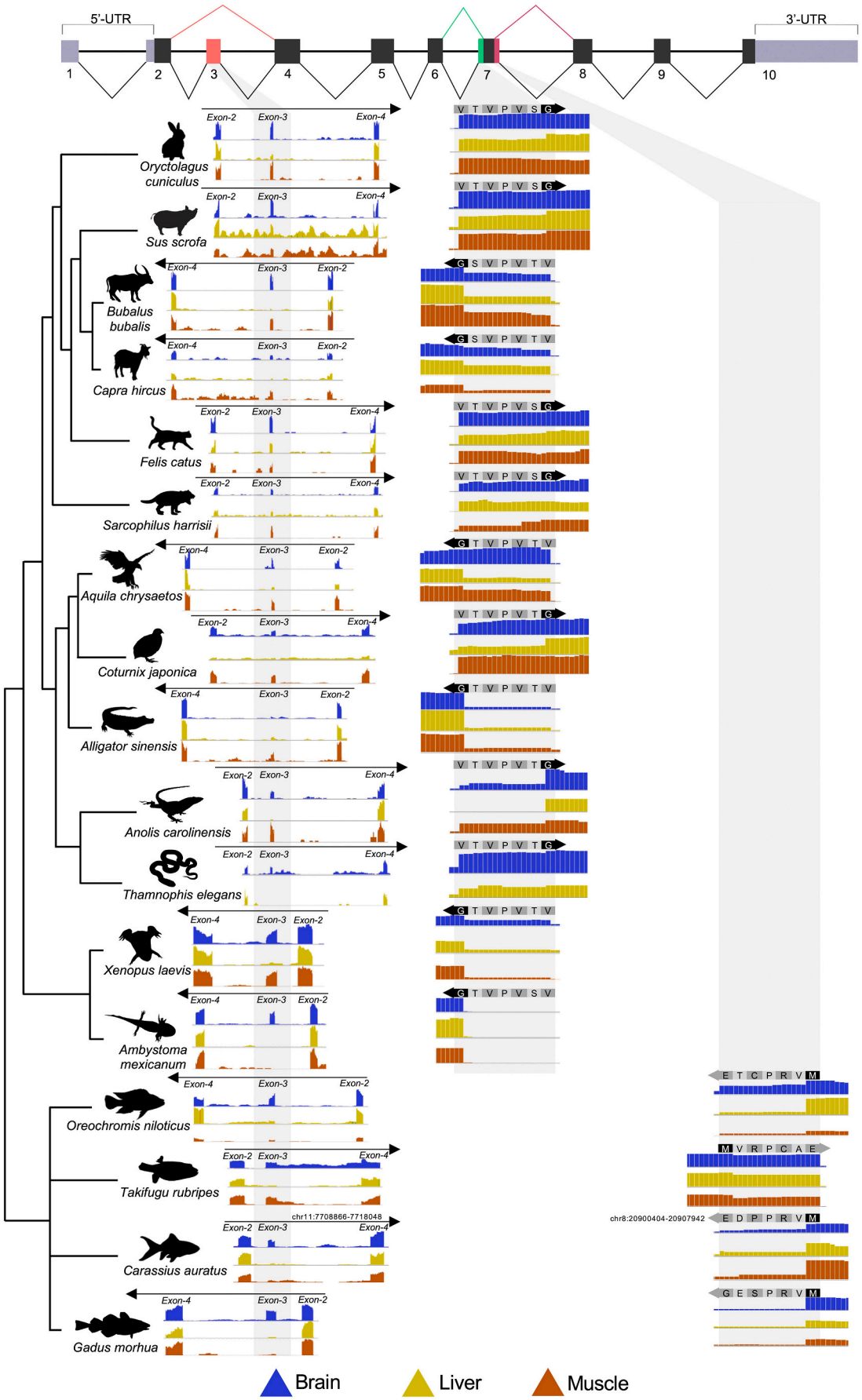
Another important point is that AS of NF-YA starts with vertebrates, since lancelet lacks exon-3 splicing, as well as variability at exon-7C or exon-7N. There is ample evidence of AS in amphioxus, including in TF genes, as documented for *Pax2/5/8* [52,53] and *Hif1a* [54]: this makes the lack of AS in NF-YA remarkable.

In summary, exon-3 is a molecular innovation acquired in the common ancestor of chordates, unspliced in non-vertebrates, alternatively spliced in vertebrates and independently lost in hagfish and cartilaginous fishes. Hagfish is reported as an outlier in terms of median intron length among basal vertebrates and chordates [55], having one of the largest median gene lengths, yet the intron that separates exon-2 and exon-4 homologs is merely 81 bp long: a deletion within this intron could have led to the loss of exon-3. Importantly, a secondary loss of an alternatively spliced region in hagfish happened also for exon-7N (Fig. 3A and Fig. 8), effectively abolishing NF-YA AS events in this group.

As for cartilaginous fishes, many phenotypic traits are derived (newly acquired), rather than ancestral. This has been linked to secondary loss of genetic elements: exon-3 could indeed represent one such case, either neutrally by the loss of functional interacting genes, or by adapting to selective pressure. A hint in this regard might come from genes coding for secreted calcium-binding phosphoproteins –SCPP– involved in the mineralization of hard tissues, such as perichondral bones and teeth. The two SCPP ancestor genes –*Sparc* and *Sparcl1*– are present in all metazoan and jawed vertebrates, respectively [56,57], but absent in Chondrichthyes. Importantly, inactivation of the zebrafish homologue *spp1* led to reduced bone formation, possibly explaining the differences in mineralization and formation of hard tissues between bony fishes and Chondrichthyes [58,59]. The promoters of SCPP genes in human and in zebrafish are shown in Fig. S10: the NF-Y matrix is absent in the SPARC and SPARCL1 genes, including in *sparcb* in lamprey, but present in the majority of human (12/22) and zebrafish (6/9) genes. The CCAAT position is canonical, between −50 and −130 from the TSS: under these circumstances, the functionality of the element has been proven in bashing experiments of >150 promoters (Reviewed in [12]). Indeed, the mouse SCPP *Dspp* gene was experimentally validated as NF-Y target [60]. We can speculate that bone mineralization and expansion of SCPP genes is coupled to the acquisition of CCAAT in promoters and presence of NF-YAl: the loss of NF-YA exon-3 might have generated an altered, or insufficient expression of SCPP genes, potentially being positively selected in the bony ancestor of Chondrichthyes, entailing a progressive loss of the genes and, consequently, of bone formation.

3.2. Conservation and evolution of the Q-rich TAD

The unstructured state of IDRs underlies a release from the selective pressures usually operating on structurally constrained domains. In general, a vast study on 1121 proteins from eukaryotes, bacteria and archaea indicates that DNA- and RNA-binding proteins show very high levels of disorder [47]. Related to TF domains, only some of the DBDs are highly disordered, while the degree of disorder in TADs is vastly generalized [45]. This latter unbiased study ranked NF-YA (NF-YAl specifically) at the top of intrinsically disordered TFs, with a prediction of 96,25% residues being in this condition. This prediction included not only the TAD, but every part other than the subunits-interacting A1. In theory, conservation of IDRs is expected to be lower than that of structured domains, deemed to recognize highly defined structures, such as specific DNA sequences. Yet a high level of heterogeneity is found for



(caption on next page)

Fig. 7. *NF-YA* isoforms expression in different vertebrate species.
Expression of *NF-YA* exon-3, exon-7N, and exon-7C in 17 vertebrate species, assessed by RNA-seq mapped read coverage, in adult brain (blue track), liver (yellow) and muscle (orange). In the latter case, samples from skeletal muscle were preferentially selected; when not available, samples from the heart were included in the analysis, instead. In goldfish, the coverage from the exon-3 and exon-7C regions was associated to two distinct paralogs, as indicated by their coordinates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

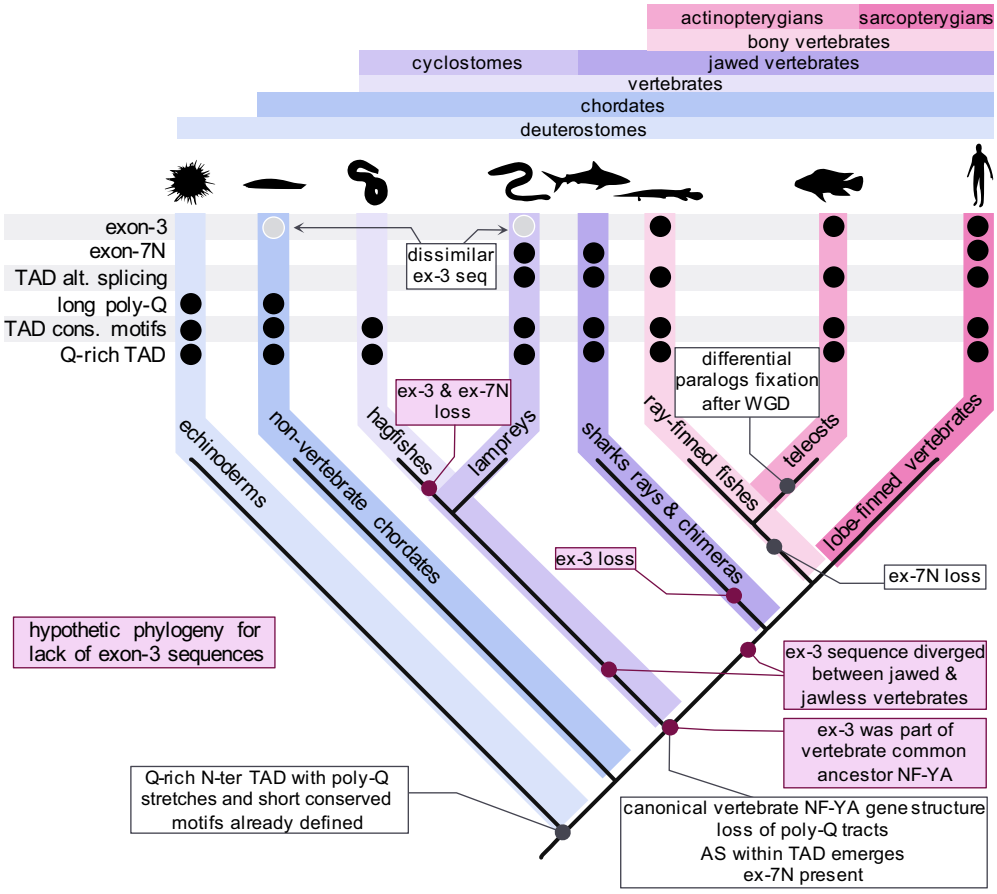


Fig. 8. Summary of NF-YA evolution in vertebrates.
The major events characterizing NF-YA transactivation domain (TAD) evolution in vertebrates are summarized. Lancelet and echinoderms are used as outgroups to infer ancestral patterns. The top panel summarizes the presence of different protein features across the groups considered (black circles). The main events describing exon-3 independent loss hypothetic phylogeny are reported on the cladogram.

conservation in IDRs [61]. Note that conservation in these regions is not trivial to quantify, leading to a paucity of dedicated studies. This is due to the expanded nature of most TFs families, with the presence of several paralogs making unambiguous identification of orthologous sequences difficult. A report comparing 380 human/mouse TFs orthologues found ~96% sequence identity for DBDs *versus* ~87% for the remaining regions [46]. Our previous human/mouse comparison on NF-YA gave 100% identity for the DBD and 99.65% (1 substitution) for the remaining regions [19]. Although we confirm lower sequence conservation in regions outside the DBD, we find the conservation of vertebrate NF-YA TAD remarkable (Fig. S7). This could be due to the lack of NF-YA homologs in vertebrates, a condition unmet by most other TFs. Consultation of The Human Transcription Factors database (http://humantfs.ccbr.utoronto.ca/index.php) returns only 7 –out of 77– DBD families with a single member, among which NF-YA. Moreover, the essentiality of NF-YA in early development [14] is likely contributing to the low rate of evolution in vertebrates.

The level of conservation within the TAD drops moving outside of vertebrates (Fig. S7). Importantly, we could isolate few –rather strong– conservation blocks (Fig. 6B). These motifs likely represent ancient molecular features at the core of TAD function, potentially involved in direct protein-protein interactions with molecular partners, likewise conserved. Indeed, we have shown that mutations of glutamines and, to a lesser extent, isoleucines within Block II –IIIQQ– led to decreased function in trans-activation assays in human cells [21]. The glutamine-rich nature of the TAD, as its overall amino acids composition, is ancestral. The mechanism of action of Q-rich TADs and the 'flavour' of their disordered state are still poorly appreciated; this is different from acidic TADs, for which high-throughput screenings in yeast [62,63] and structural characterization [64,65,66,67] have been successfully performed.

Our data also impact on the interpretation of the recent discovery of the short NF-YAx isoform, lacking both exon-3 and exon-5, identified in human neuroblastoma cells and in the head of late mouse embryos [23]. The Authors argued that this isoform serves as Dominant Negative –DN– for the functions of the two major isoforms during the expansion of neuronal progenitors. This is in line with data on artificial DN versions produced either by mutating the A2 [68], or by ablating the Q-rich N-terminal [17]. NF-YAs was shown to be transcriptionally competent, in GAL4-based trans-activation assays [19,20] and upon overexpression [16,22]. Thus, exon-3 sequences are less crucial than those of exon-5. Our finding of two conserved Blocks in this exon, including Block II validated by mutagenesis, supports this idea. Indeed, NF-YAx was unable to interact with Sp1, a Zn-finger, Q-rich TF known to be a widespread NF-Y partner, based on genome-wide locations [69,70,71].

As for the hexapeptide of exon-7N lost in ray-finned fishes, it is within another area of variability –for NF-YA standards– marking a clear boundary between the TAD (exon-6) and the S/T-rich "intermediate"

domain, which includes exon-7 and part of exon-8. The function of 7N is less clear, but overexpression assays on the Cystathionine-β-Synthase promoter indicate that the hexapeptide contributes to the function of NF-YAl and to synergism with Sp1 [22]. Overall, our analyses on protein intrinsic disorder and evolution of alternatively-spliced exons support the observation that most AS events in eukaryotes impinge on IDRs, by avoiding the disruption of structurally well-defined domains and boosting functional diversity [72].

3.3. The two major isoforms in development, and disease

Although many reports showed variation in expression, the logic of the relative abundance of the isoforms in different mammalian cell lineages, tumor cell lines, cell-cycle phases, growth/differentiation conditions was not obvious. In mouse Embryonic Stem cells –mESCs– NF-YAs is more abundant, with NF-YAl rising following differentiation [73,74,75]. Differential effects of isoforms overexpression on CCAAT-driven genes were noticed in mESCs [73]. NF-Y/CBF was identified as a key TF for ectodermal expression of the sea urchin CCAAT-dependent *spec2* gene, by monitoring the fate of *spec2*-GFP reporters introduced in fertilized eggs and harvested at the blastula stage. 98% of aboral ectodermal cells were positive with a wt construct, whereas mutation of CCAAT led to a dramatic increase in mesodermal cells [49]. We confirm that sea urchin NF-YA is devoid of exon-3 sequences, as previously reported on cDNA cloning [49,50]: therefore, the exon-3-less isoform is responsible for the ectodermal-driving activity observed.

Hematopoietic Stem cells –HSCs– express NF-YAs, in differentiated cells NF-YAl prevails [76]. NF-YAs overexpression in HSCs *ex vivo* leads to an increase in engraftment upon bone marrow transplantation, a sign of HSCs expansion. In myoblasts, only NF-YAl is expressed, declining upon terminal differentiation to myotube [77]. Genetic ablation of exon-3 by genome editing in mouse C2C12 myoblasts leads to cells growing normally and with an apparent normal phenotype, establishing that cells expressing exclusively NF-YAs are fully viable [78]. Yet, a decrease/loss of muscle commitment results in failure to form myotubes, implying a role of NF-YAl in differentiation. The only species examined in which NF-YAl –and 7N– appears to be absent in muscle is Axolotl: this salamander is a model system for tissue regeneration, including limbs, which leads to the intriguing possibility that NF-YAs-expressing cells maintain an indefinite, stem-like potential in this species.

The inferred role of NF-YAl in the expression of SCPPs in mineralogenic cells of bony vertebrates adds to the list of lineages of mesenchymal origin –HSC, myoblasts– relying on the expression of NF-YAl. A further hint at a specific role for NF-YAl in mesenchymal cells comes from systematic assessment of expression levels in large RNA-seq datasets of human cancers: NF-YAs predominates in epithelial tumors [79–81,82,83,84], but breast, lung, oral and gastric carcinomas have subsets of Epithelial-to-Mesenchymal Transition –EMT– cells featuring higher expression of NF-YAl [80,81,83,85]. NF-YAs is associated to target gene categorizations with *cell-cycle* and *metabolism* terms, NF-YAl with *differentiation*. Altogether, these data lead us to propose that NF-YAs helps maintain an epithelial "primordial" identity, NF-YAl to produce/maintain a mesenchymal one. Further genetic experiments of ablation of exon-3 will be required to lend definitive support to this hypothesis.

Finally, NF-YA binds to DNA and functions as a trimeric TF, together with the HFD subunits: NF-YB is not apparently involved in AS, at least in mammals, but NF-YC has multiple isoforms, resulting from AS –exon skipping and donor/acceptor events– impacting on the Q-rich TAD located at the C-terminal: we feel that more can be learnt from a systematic phylogenetic study of this subunit.

4. Materials and methods

4.1. Retrieval of NF-YA vertebrate orthologs and sequence filtering

Vertebrate NF-YA protein sequences were retrieved from the Pfam subsection of InterPro database (https://www.ebi.ac.uk/interpro/) under the accession number IPR001289. It corresponds to the Pfam family CBFB_NFYA (PF02045), whose members are identified by searching primary sequence databases with a hidden Markov model (HMM) built on a curated seed alignment of representative members of the family. Specifically, identification was based on a ~ 50 aa core domain responsible for subunit association and DNA-recognition, conserved in all eukaryotes and referred as HAP2 domain (see Fig.1A). Sequences were divided in the main groups in the taxonomy tab and vertebrate sequences were downloaded in FASTA format. The number of species reported in Fig.1B refers to the status of the database as of May 2020.

The 245 protein sequences from mammals were aligned in Jalview [86,87] using Muscle [88] with default settings and the resulting MSA was manually edited. The dataset was filtered based on visual inspection of the alignment by removing sequences not starting with Met, sequences derived from erroneously annotated fusions with nearby genes or sequences missing critical regions within the HAP2 domain, resulting in 223 protein sequences belonging to 76 species of mammals. We performed the same procedure with the rest of Pfam-derived vertebrate sequences (281 sequences, devoid of mammals), resulting in 238 sequences belonging to 123 different vertebrate species. Additionally, we manually retrieved 20 NF-YA sequences for phylogenetically relevant species not present in the automatically derived dataset by dedicated searches in NCBI Gene (https://www.ncbi.nlm.nih.gov/gene/) or ENSEMBL (https://www.ensembl.org/index.html) browsers. All sequences used in the study are listed in Supplementary File 1, for a total of 482 sequences of 220 different species, including 9 non-vertebrate deuterostomes. MSA-derived sequence logos were built using WebLogo [89]. Fasta format MSAs used for sequence logos are in Supplementary Files 2 and 3. MSAs corresponding to Fig. S1B-1C and Fig. 6B are in Supplementary Files 4, 5 and 6 respectively.

4.2. Manual sequence annotation

Marine lamprey (*P. marinus*) NF-YA protein sequence was manually edited by removing a short fragment (RLGTMESY) encoded by a spurious mini-exon included in the automatic annotation in ENSEMBL gene (ENSPMAG00000000964), for which we found no evidence of expression (either by BLAST searches in EST databases or RNA-seq analysis). Pacific lamprey (*E. tridentatus*) sequence was retrieved from SIMRbase (https://genomes.stowers.org/, gene ID ETRf_mk00023629-RA) and exons 3 and 9, missing in the deposited annotation, were retrieved by BLAST searches using marine lamprey sequence (100% identity). To complete the hagfish (*E. burgeri*) sequence deposited in ENSEMBL (ENSEBUG00000015438), lacking a significant portion of the N-terminal region found in other vertebrates, we retrieved a *E. burgeri* leukocyte cDNA clone (ENA accession: BJ649853) matching with NF-YA canonical N-terminal region. We then back-mapped this sequence on the hagfish genome and retrieved 100% identity matches in a contig (Eburgeri_3.2 contig FYBX02000108.1) different from the one harboring the annotated NF-YA gene (Eburgeri_3.2 contig FYBX02009477.1): the gene is thus split in two separate contigs due to an assembly error and nearby ambiguous sequence regions. The manually annotated complete protein sequence was reconstituted from the translated cDNA clone and the annotated protein-coding gene in ENSEMBL. We removed a mini-exon located between canonical exon-6 and 7 –LHCQRPIDG– since there is no evidence of expression and this sequence shows no homology with other NF-YA sequences: we consider this sequence spurious. We extended the 3′ boundary of exon-6 including the conserved motif VIPQG, present in the cDNA clone. We found evidence of inclusion in

mature transcripts of a 10 aa stretch –QVVPSVVQGSSPRI– at the 5′ boundary of canonical exon-8 missing from the annotated transcript. To validate the reconstituted hagfish NF-YA protein sequence we performed tblastn searches against RNA-seq experiments from a second hagfish species (*E. cirrhatus*, SRX1134573), retrieving a full coverage of the query, including all manually-annotated regions, with a total of 7 substitutions between the two species (Fig. S6).

All cladograms with evolutionary timescales were generated with TimeTree [90] and tree topology rendered with iTOL [91].

4.3. Exon flanking-sequences conservation

To evaluate the conservation of sequences flanking human *NF-YA* exons, we calculated the mean PhastCons scores [92] for segments of 40 bp upstream or downstream each exon. Scores at the given coordinates were retrieved using *GenomicScores* package in R [93] with the annotation package *phastCons100way.UCSC.hg19*, which stores PhastCons conservation scores for human genomic positions calculated from MSA with other 99 vertebrate species.

4.4. Analysis of exons homologous to alternatively spliced human NF-YA exons

Presence of exon-3 and -7N homologous sequences was first assessed in MSAs. For species lacking exon-3 or -7N in the protein sequence used for MSAs, we checked the corresponding gene structure models for the presence of these regions either in ENSEMBL or NCBI gene repositories, along with the presence of correctly positioned donor/acceptor canonical splice sites. For species lacking exon-3 homologous sequences at protein and gene model level, (i) we performed *de novo* gene prediction using GENSCAN [34] and WebAUGUSTUS [35] algorithms using the whole gene locus as input; (ii) we launched tblastn searches within the intron where the hypothetical exon-3 should reside, employing as query either human, fish (spotted gar) or lamprey exon-3 protein sequence; (iii) for cartilaginous fishes, we analyzed several available RNA-seq datasets for elephant shark (*C. milii*) using STAR (see below).

Exon-7C presence in bony fishes was assessed individually for each species reported in Fig. S2 by inspection in the ENSEMBL browser of the corresponding gene at DNA level, including annotated gene paralogs. Presence of canonical acceptor splice sites was also assessed.

4.5. RNA-seq datasets, mapping, and mRNA expression quantification

We retrieved the FASTQ files associated to each of the datasets considered for the analyses (details in Supplementary File 7), using the SRA Explorer website (https://sra-explorer.info/). We mapped the FASTQ files using *STAR* (version 2.7.8a) [94], and mapped reads coverage was visualized by loading the BAM file corresponding to each sample into the software Integrative Genomic Viewer (IGV, version 2.10.2) [95].

4.6. Protein disorder analysis, sequence composition and modelling

For protein composition analysis, we used 19 NF-YA sequences belonging to a representative set of vertebrate species. We isolated the N-term TAD and the C-term core domains at position 250 of human protein (long isoform, Supplementary Files 8 and 9, respectively) and subjected the resulting sequence sets to compositional bias analysis using Composition Profiler [96] with SwissProt 51 as background distribution. For structural disorder analysis we subjected human NF-YA protein sequence to the following disorder prediction algorithms: PONDR-FIT [97], DISOPRED3 [98], DisEMBL coils [99] and IUPRED3 long disorder [100]. The resulting disorder-prediction scores were used to build Fig. 5B. For structural models prediction we used AlphaFold2 [48]. Human NF-YA model shown in Fig. 5E was downloaded from AlphaFold2 Protein Structure Database (https://alphafold.ebi.ac.uk/). Models shown in Fig. S8 were generated using AlphaFold2_advanced ColabFold implementation with standard settings (https://github.com/sokrypton/ColabFold, [101] Jan 1). For each species, at least 5 models were generated and inspected in UCSF ChimeraX [102].

Data availability

The data underlying this article are available in the article and in its online Supplementary material.

Autor contributions

A.B. designed the study and performed the investigation. A.G. and D. D. analyzed expression data. N.G. analyzed the data and critically read the manuscript. R.M. conceptualized the research. A.B. and R.M. wrote the manuscript.

Authors statement

The authors declare that they have no competing interests.

Acknowledgements

Authors acknowledge support from the University of Milan through the APC initiative.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2022.110390.

References

[1] S.A. Lambert, A. Jolma, L.F. Campitelli, P.K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T.R. Hughes, M.T. Weirauch, The human transcription factors, Cell 175 (2) (2018) 598–599, https://doi.org/10.1016/j.cell.2018.09.045.

[2] E. Wingender, T. Schoeps, M. Haubrock, M. Krull, J. Dönitz, TFClass: expanding the classification of human transcription factors to their mammalian orthologs, Nucleic Acids Res. 46 (D1) (2018) D343–D347, https://doi.org/10.1093/nar/gkx987.

[3] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K.R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, et al., DNA-binding specificities of human transcription factors, Cell 152 (1) (2013) 327–339, https://doi.org/10.1016/j.cell.2012.12.009.

[4] C. Larroux, G.N. Luke, P. Koopman, D.S. Rokhsar, S.M. Shimeld, B.M. Degnan, Genesis and expansion of metazoan transcription factor gene classes, Mol. Biol. Evol. 25 (5) (2008) 980–996, https://doi.org/10.1093/molbev/msn047.

[5] D. Dolfini, R. Mantovani, Targeting the Y/CCAAT box in cancer: YB-1 (YBX1) or NF-Y? Cell Death Differ. 20 (5) (2013) 676–685, https://doi.org/10.1038/cdd.2013.13.

[6] A. Chaves-Sanjuan, N. Gnesutta, A. Gobbini, D. Martignago, A. Bernardini, F. Fornara, R. Mantovani, M. Nardini, Structural determinants for NF-Y subunit organization and NF-Y/DNA association in plants, Plant J. 105 (1) (2021) 49–61, https://doi.org/10.1111/tpj.15038.

[7] E.M. Huber, D.H. Scharf, P. Hortschansky, M. Groll, A.A. Brakhage, DNA minor groove sensing and widening by the CCAAT-binding complex, Structure 20 (10) (2012) 1757–1768, https://doi.org/10.1016/j.str.2012.07.012.

[8] M. Nardini, N. Gnesutta, G. Donati, R. Gatta, C. Forni, A. Fossati, C. Vonrhein, D. Moras, C. Romier, M. Bolognesi, et al., Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination, Cell 152 (1) (2013) 132–143, https://doi.org/10.1016/j.cell.2012.11.047.

[9] N. Gnesutta, M. Nardini, R. Mantovani, The H2A/H2B-like histone-fold domain proteins at the crossroad between chromatin and different DNA metabolisms, Transcription 4 (3) (2013) 114–119, https://doi.org/10.4161/trns.25002.

[10] N. Gnesutta, R. Mantovani, F. Fornara, Plant flowering: imposing DNA specificity on histone-fold subunits, Trends Plant Sci. 23 (4) (2018) 293–301, https://doi.org/10.1016/j.tplants.2017.12.005.

[11] X. Lv, X. Zeng, H. Hu, L. Chen, F. Zhang, R. Liu, Y. Liu, X. Zhou, C. Wang, Z. Wu, et al., Structural insights into the multivalent binding of the *Arabidopsis* FLOWERING LOCUS T promoter by the CO–NF–Y master transcription factor complex, Plant Cell 33 (4) (2021) 1182–1195, https://doi.org/10.1093/plcell/koab016.

[12] D. Dolfini, F. Zambelli, G. Pavesi, R. Mantovani, A perspective of promoter architecture from the CCAAT box, Cell Cycle 8 (24) (2009) 4127–4137, https://doi.org/10.4161/cc.8.24.10240.

[13] A.J. Oldfield, T. Henriques, D. Kumar, A.B. Burkholder, S. Cinghu, D. Paulet, B. D. Bennett, P. Yang, B.S. Scruggs, C.A. Lavender, et al., NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region, Nat. Commun. 10 (1) (2019) 3072, https://doi.org/10.1038/s41467-019-10905-7.

[14] A. Bhattacharya, J.M. Deng, Z. Zhang, R. Behringer, B. de Crombrugghe, S. N. Maity, The B subunit of the CCAAT box binding transcription factor complex (CBF/NF-Y) is essential for early mouse development and cell proliferation, Cancer Res. 63 (23) (2003) 8167–8172.

[15] S.N. Maity, NF-Y (CBF) regulation in specific cell types and mouse models, Biochim. Biophys. Acta Gene Regul. Mech. 1860 (5) (2017) 598–603, https://doi.org/10.1016/j.bbagrm.2016.10.014.

[16] M. Ceribelli, P. Benatti, C. Imbriano, R. Mantovani, NF-YC complexity is generated by dual promoters and alternative splicing, J. Biol. Chem. 284 (49) (2009) 34189–34200, https://doi.org/10.1074/jbc.M109.008417.

[17] F. Coustry, S.N. Maity, S. Sinha, B. de Crombrugghe, The transcriptional activity of the CCAAT-binding factor CBF is mediated by two distinct activation domains, one in the CBF-B subunit and the other in the CBF-C subunit, J. Biol. Chem. 271 (24) (1996) 14485–14491, https://doi.org/10.1074/jbc.271.24.14485.

[18] Q. Hu, S.N. Maity, Stable expression of a dominant negative mutant of CCAAT binding factor/NF-Y in mouse fibroblast cells resulting in retardation of cell growth and inhibition of transcription of various cellular genes, J. Biol. Chem. 275 (6) (2000) 4435–4444, https://doi.org/10.1074/jbc.275.6.4435.

[19] X.Y. Li, R. Hooft van Huijsduijnen, R. Mantovani, C. Benoist, D. Mathis, Intron-exon organization of the NF-Y genes. Tissue-specific splicing modifies an activation domain, J. Biol. Chem. 267 (13) (1992) 8984–8990, https://doi.org/10.1016/S0021-9258(19)50377-5.

[20] E. Serra, K. Zemzoumi, V. Lardans, C. Dissous, A. di Silvio, R. Mantovani, Conservation and divergence of NF-Y transcriptional activation function, Nucleic Acids Res. 26 (16) (1998) 3800–3805, https://doi.org/10.1093/nar/26.16.3800.

[21] A. Silvio di, C. Imbriano, R. Mantovani, Dissection of the NF-Y transcriptional activation potential, Nucleic Acids Res. 27 (13) (1999) 2578–2584, https://doi.org/10.1093/nar/27.13.2578.

[22] Y. Ge, T.L. Jensen, L.H. Matherly, J.W. Taub, Synergistic regulation of human cystathionine-β-synthase-1b promoter by transcription factors NF-YA isoforms and Sp1, Biochim. Biophys. Acta Gene Struct. Expr. 1579 (2) (2002) 73–80, https://doi.org/10.1016/S0167-4781(02)00509-2.

[23] L. Cappabianca, A.R. Farina, L. Di Marcotullio, P. Infante, D. De Simone, M. Sebastiano, A.R. Mackay, Discovery, characterization and potential roles of a novel NF-YAx splice variant in human neuroblastoma, J. Exp. Clin. Cancer Res. 38 (1) (2019) 482, https://doi.org/10.1186/s13046-019-1481-8.

[24] G. Gusmaroli, C. Tonelli, R. Mantovani, Regulation of the CCAAT-binding NF-Y subunits in *Arabidopsis thaliana*, Gene 264 (2) (2001) 173–185, https://doi.org/10.1016/S0378-1119(01)00323-7.

[25] R. Hooft van Huijsduijnen, X.Y. Li, D. Black, H. Matthes, C. Benoist, D. Mathis, Co-evolution from yeast to mouse: cDNA cloning of the two NF-Y (CP-1/CBF) subunits, EMBO J. 9 (10) (1990) 3119–3127.

[26] S.N. Maity, T. Vuorio, B. de Crombrugghe, The B subunit of a rat heteromeric CCAAT-binding transcription factor shows a striking sequence identity with the yeast Hap2 transcription factor, PNAS 87 (14) (1990) 5378–5382, https://doi.org/10.1073/pnas.87.14.5378.

[27] K. Zemzoumi, E. Serra, R. Mantovani, J. Trolet, A. Capron, C. Dissous, Cloning of *Schistosoma mansoni* transcription factor NF-YA subunit: phylogenic conservation of the HAP-2 homology domain, Mol. Biochem. Parasitol. 77 (2) (1996) 161–172, https://doi.org/10.1016/0166-6851(96)02590-X.

[28] Q. Li, M. Herrler, N. Landsberger, N. Kaludov, V.V. Ogryzko, Y. Nakatani, A. P. Wolffe, Xenopus NF-Y pre-sets chromatin to potentiate p300 and acetylation-responsive transcription from the *Xenopus* hsp70 promoter in vivo, EMBO J. 17 (21) (1998) 6300–6315, https://doi.org/10.1093/emboj/17.21.6300.

[29] J. Inoue, Y. Sato, R. Sinclair, K. Tsukamoto, M. Nishida, Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling, PNAS 112 (48) (2015) 14918–14923, https://doi.org/10.1073/pnas.1507669112.

[30] A. Meyer, Y.V. de Peer, From 2R to 3R: evidence for a fish-specific genome duplication (FSGD), BioEssays 27 (9) (2005) 937–945, https://doi.org/10.1002/bies.20293.

[31] M. Nakatani, M. Miya, K. Mabuchi, K. Saitoh, M. Nishida, Evolutionary history of Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaean origin and Mesozoic radiation, BMC Evol. Biol. 11 (1) (2011) 177, https://doi.org/10.1186/1471-2148-11-177.

[32] P. Xu, J. Xu, G. Liu, L. Chen, Z. Zhou, W. Peng, Y. Jiang, Z. Zhao, Z. Jia, Y. Sun, et al., The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*, Nat. Commun. 10 (1) (2019) 4625, https://doi.org/10.1038/s41467-019-12644-1.

[33] S. Lien, B.F. Koop, S.R. Sandve, J.R. Miller, M.P. Kent, T. Nome, T.R. Hvidsten, J. S. Leong, D.R. Minkley, A. Zimin, et al., The Atlantic salmon genome provides insights into rediploidization, Nature 533 (7602) (2016) 200–205, https://doi.org/10.1038/nature17164.

[34] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA11Edited by F. E. Cohen, J. Mol. Biol. 268 (1) (1997) 78–94, https://doi.org/10.1006/jmbi.1997.0951.

[35] K.J. Hoff, M. Stanke, WebAUGUSTUS–a web service for training AUGUSTUS and predicting genes in eukaryotes, Nucleic Acids Res. 41 (Web Server issue) (2013) W123–W128, https://doi.org/10.1093/nar/gkt418.

[36] A.M. Heimberg, R. Cowper-Sal Lari, M. Sémon, P.C.J. Donoghue, K.J. Peterson, microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate, PNAS 107 (45) (2010) 19379–19383, https://doi.org/10.1073/pnas.1010350107.

[37] T. Miyashita, M.I. Coates, R. Farrar, P. Larson, P.L. Manning, R.A. Wogelius, N. P. Edwards, J. Anné, U. Bergmann, A.R. Palmer, et al., Hagfish from the Cretaceous Tethys Sea and a reconciliation of the morphological–molecular conflict in early vertebrate phylogeny, PNAS. 116 (6) (2019) 2146–2151, https://doi.org/10.1073/pnas.1814794116.

[38] T.J. Near, Conflict and resolution between phylogenies inferred from molecular and phenotypic data sets for hagfish, lampreys, and gnathostomes, J. Exp. Zool. B Mol. Dev. Evol. 312B (7) (2009) 749–761, https://doi.org/10.1002/jez.b.21293.

[39] M. Sémon, M. Schubert, V. Laudet, Programmed genome rearrangements: in lampreys, all cells are not equal, Curr. Biol. 22 (16) (2012) R641–R643, https://doi.org/10.1016/j.cub.2012.06.022.

[40] J.J. Smith, N. Timoshevskaya, C. Ye, C. Holt, M.C. Keinath, H.J. Parker, M. E. Cook, J.E. Hess, S.R. Narum, F. Lamanna, et al., The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution, Nat. Genet. 50 (2) (2018) 270–277, https://doi.org/10.1038/s41588-017-0036-1.

[41] J.J. Smith, V.A. Timoshevskiy, C. Saraceno, Programmed DNA elimination in vertebrates, Annu. Rev. Anim. Biosci. 9 (1) (2021) 173–201, https://doi.org/10.1146/annurev-animal-061220-023220.

[42] N. Malhis, S.J.M. Jones, J. Gsponer, Improved measures for evolutionary conservation that exploit taxonomy distances, Nat. Commun. 10 (1) (2019) 1556, https://doi.org/10.1038/s41467-019-09583-2.

[43] E. Schad, L. Kalmar, P. Tompa, Exon-phase symmetry and intrinsic structural disorder promote modular evolution in the human genome, Nucleic Acids Res. 41 (8) (2013) 4409–4422, https://doi.org/10.1093/nar/gkt110.

[44] A.S. Garza, N. Ahmad, R. Kumar, Role of intrinsically disordered protein regions/domains in transcriptional regulation, Life Sci. 84 (7) (2009) 189–193, https://doi.org/10.1016/j.lfs.2008.12.002.

[45] J. Liu, N.B. Perumal, C.J. Oldfield, E.W. Su, V.N. Uversky, A.K. Dunker, Intrinsic disorder in transcription factors, Biochemistry. 45 (22) (2006) 6873–6888, https://doi.org/10.1021/bi0602718.

[46] Y. Minezaki, K. Homma, A.R. Kinjo, K. Nishikawa, Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation, J. Mol. Biol. 359 (4) (2006) 1137–1149, https://doi.org/10.1016/j.jmb.2006.04.016.

[47] C. Wang, V.N. Uversky, L. Kurgan, Disordered nucleiome: abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea, Proteomics. 16 (10) (2016) 1486–1498, https://doi.org/10.1002/pmic.201500177.

[48] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589, https://doi.org/10.1038/s41586-021-03819-2.

[49] X. Li, S. Dayal, W.H. Klein, C. Bhattacharya, S. Maity, Ectoderm gene activation in sea urchin embryos mediated by the CCAAT-binding factor, Differentiation 70 (2) (2002) 109–119, https://doi.org/10.1046/j.1432-0436.2002.700206.x.

[50] Z. Li, S.R. Kalsapudi, G. Childs, Isolation and characterization of cDNAs encoding the sea urchin (*Strongylocentrotus purpuratus*) homologue of the CCAAT binding protein NF-Y a subunit, Nucleic Acids Res. 21 (19) (1993) 4639, https://doi.org/10.1093/nar/21.19.4639.

[51] B. Rost, Twilight zone of protein sequence alignments, Protein Eng. 12 (2) (1999) 85–94, https://doi.org/10.1093/protein/12.2.85.

[52] S. Short, L.Z. Holland, The evolution of alternative splicing in the Pax Family: the view from the basal chordate Amphioxus, J. Mol. Evol. 66 (6) (2008) 605, https://doi.org/10.1007/s00239-008-9113-5.

[53] S. Short, Z. Kozmik, L.Z. Holland, The function and developmental expression of alternatively spliced isoforms of *Amphioxus* and *Xenopus laevis* Pax2/5/8 genes: revealing divergence at the invertebrate to vertebrate transition, J. Exp. Zool. B Mol. Dev. Evol. 318 (7) (2012) 555–571, https://doi.org/10.1002/jez.b.22460.

[54] S. Gao, L. Lu, Y. Bai, P. Zhang, W. Song, C. Duan, Structural and functional analysis of amphioxus HIFα reveals ancient features of the HIFα family, FASEB J. 28 (4) (2014) 1880–1890, https://doi.org/10.1096/fj.12-220152.

[55] M.J. McCoy, A.Z. Fire, Intron and gene size expansion during nervous system evolution, BMC Genomics 21 (1) (2020) 360, https://doi.org/10.1186/s12864-020-6760-4.

[56] K. Kawasaki, The SCPP gene family and the complexity of hard tissues in vertebrates, Cells Tissues Organs 194 (2–4) (2011) 108–112, https://doi.org/10.1159/000324225.

[57] Y. Lv, K. Kawasaki, J. Li, Y. Li, C. Bian, Y. Huang, X. You, Q. Shi, A genomic survey of SCPP family genes in fishes provides novel insights into the evolution of fish scales, Int. J. Mol. Sci. 18 (11) (2017) 2432, https://doi.org/10.3390/ijms18112432.

[58] B. Ryll, S. Sanchez, T. Haitina, P. Tafforeau, P.E. Ahlberg, The genome of *Callorhinchus* and the fossil record: a new perspective on SCPP gene evolution in gnathostomes, Evol. Dev. 16 (3) (2014) 123–124, https://doi.org/10.1111/ede.12071.

[59] B. Venkatesh, A.P. Lee, V. Ravi, A.K. Maurya, M.M. Lian, J.B. Swann, Y. Ohta, M. F. Flajnik, Y. Sutoh, M. Kasahara, et al., Elephant shark genome provides unique insights into gnathostome evolution, Nature 505 (7482) (2014) 174–179, https://doi.org/10.1038/nature12826.

[60] S. Chen, J. Gluhak-Heinrich, M. Martinez, T. Li, Y. Wu, H.-H. Chuang, L. Chen, J. Dong, I. Gay, M. MacDougall, Bone morphogenetic protein 2 mediates dentin sialophosphoprotein expression and odontoblast differentiation via NF-Y

signaling, J. Biol. Chem. 283 (28) (2008) 19359–19370, https://doi.org/10.1074/jbc.M709492200.
[61] S. Banerjee, S. Chakraborty, R.K. De, Deciphering the cause of evolutionary variance within intrinsically disordered regions in human proteins, J. Biomol. Struct. Dyn. 35 (2) (2017) 233–249, https://doi.org/10.1080/07391102.2016.1143877.
[62] C.N. Ravarani, T.Y. Erkina, G. De Baets, D.C. Dudman, A.M. Erkine, M.M. Babu, High-throughput discovery of functional disordered regions: investigation of transactivation domains, Mol. Syst. Biol. 14 (5) (2018), e8190, https://doi.org/10.15252/msb.20188190.
[63] M.V. Staller, A.S. Holehouse, D. Swain-Lenz, R.K. Das, R.V. Pappu, B.A. Cohen, A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain, Cell Syst. 6 (4) (2018), https://doi.org/10.1016/j.cels.2018.01.015, 444-455.e6.
[64] E. Bochkareva, L. Kaustov, A. Ayed, G.-S. Yi, Y. Lu, A. Pineda-Lucena, J.C.C. Liao, A.L. Okorokov, J. Milner, C.H. Arrowsmith, et al., Single-stranded DNA mimicry in the p53 transactivation domain interaction with replication protein A, PNAS 102 (43) (2005) 15412–15417, https://doi.org/10.1073/pnas.0504614102.
[65] J.B. Thoden, L.A. Ryan, R.J. Reece, H.M. Holden, The interaction between an acidic transcriptional activator and its inhibitor: the molecular basis of Gal4p recognition by Gal80p, J. Biol. Chem. 283 (44) (2008) 30266–30272, https://doi.org/10.1074/jbc.M805200200.
[66] M. Uesugi, G.L. Verdine, The α-helical FXXΦΦ motif in p53: TAF interaction and discrimination by MDM2, PNAS 96 (26) (1999) 14801–14806, https://doi.org/10.1073/pnas.96.26.14801.
[67] J.M. Wojciak, M.A. Martinez-Yamout, H.J. Dyson, P.E. Wright, Structural basis for recruitment of CBP/p300 coactivators by STAT1 and STAT2 transactivation domains, EMBO J. 28 (7) (2009) 948–958, https://doi.org/10.1038/emboj.2009.30.
[68] R. Mantovani, X.Y. Li, U. Pessara, R. Hooft van Huisjduijnen, C. Benoist, D. Mathis, Dominant negative analogs of NF-YA, J. Biol. Chem. 269 (32) (1994) 20340–20346, https://doi.org/10.1016/S0021-9258(17)31997-X.
[69] D. Dolfini, F. Zambelli, M. Pedrazzoli, R. Mantovani, G. Pavesi, A high definition look at the NF-Y regulome reveals genome-wide associations with selected transcription factors, Nucleic Acids Res. 44 (10) (2016) 4684–4702, https://doi.org/10.1093/nar/gkw096.
[70] M. Ronzio, A. Bernardini, G. Pavesi, R. Mantovani, D. Dolfini, On the NF-Y regulome as in ENCODE (2019), PLoS Comput. Biol. 16 (12) (2020), e1008488, https://doi.org/10.1371/journal.pcbi.1008488.
[71] G. Suske, NF-Y and SP transcription factors — new insights in a long-standing liaison, Biochim. Biophys. Acta Gene Regul. Mech. 1860 (5) (2017) 590–597, https://doi.org/10.1016/j.bbagrm.2016.08.011.
[72] P.R. Romero, S. Zaidi, Y.Y. Fang, V.N. Uversky, P. Radivojac, C.J. Oldfield, M. S. Cortese, M. Sickmeier, T. LeGall, Z. Obradovic, et al., Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms, PNAS 103 (22) (2006) 8390–8395, https://doi.org/10.1073/pnas.0507916103.
[73] D. Dolfini, M. Minuzzo, G. Pavesi, R. Mantovani, The short isoform of NF-YA belongs to the embryonic stem cell transcription factor circuitry, Stem Cells 30 (11) (2012) 2450–2459, https://doi.org/10.1002/stem.1232.
[74] M. Grskovic, C. Chaivorapol, A. Gaspar-Maia, H. Li, M. Ramalho-Santos, Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells, PLoS Genet. 3 (8) (2007), e145, https://doi.org/10.1371/journal.pgen.0030145.
[75] A.J. Oldfield, P. Yang, A.E. Conway, S. Cinghu, J.M. Freudenberg, S. Yellaboina, R. Jothi, Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors, Mol. Cell 55 (5) (2014) 708–722, https://doi.org/10.1016/j.molcel.2014.07.005.
[76] A.D. Domashenko, G. Danet-Desnoyers, A. Aron, M.P. Carroll, S.G. Emerson, TAT-mediated transduction of NF-Ya peptide induces the ex vivo proliferation and engraftment potential of human hematopoietic progenitor cells, Blood 116 (15) (2010) 2676–2683, https://doi.org/10.1182/blood-2010-03-273441.
[77] A. Farina, I. Manni, G. Fontemaggi, M. Tiainen, C. Cenciarelli, M. Bellorini, R. Mantovani, A. Sacchi, G. Piaggio, Down-regulation of cyclin B1 gene transcription in terminally differentiated skeletal muscle cells is associated with loss of functional CCAAT-binding NF-Y complex, Oncogene 18 (18) (1999) 2818–2827, https://doi.org/10.1038/sj.onc.1202472.
[78] D. Libetti, A. Bernardini, S. Sertic, G. Messina, D. Dolfini, R. Mantovani, The switch from NF-YAl to NF-YAs isoform impairs myotubes formation, Cells 9 (3) (2020) 789, https://doi.org/10.3390/cells9030789.
[79] E. Bezzecchi, A. Bernardini, M. Ronzio, C. Miccolo, S. Chiocca, D. Dolfini, R. Mantovani, NF-Y subunits overexpression in HNSCC, Cancers 13 (12) (2021) 3019, https://doi.org/10.3390/cancers13123019.
[80] E. Bezzecchi, M. Ronzio, D. Dolfini, R. Mantovani, NF-YA overexpression in lung cancer: LUSC, Genes 10 (11) (2019) 937, https://doi.org/10.3390/genes10110937.
[81] E. Bezzecchi, M. Ronzio, V. Semeghini, V. Andrioletti, R. Mantovani, D. Dolfini, NF-YA overexpression in lung cancer: LUAD, Genes 11 (2) (2020) 198, https://doi.org/10.3390/genes11020198.
[82] L. Cicchillitti, G. Corrado, M. Carosi, M.E. Dabrowska, R. Loria, R. Falcioni, G. Cutillo, G. Piaggio, E. Vizza, Prognostic role of NF-YA splicing isoforms and Lamin A status in low grade endometrial cancer, Oncotarget 8 (5) (2016) 7935–7945, https://doi.org/10.18632/oncotarget.13854.
[83] D. Dolfini, V. Andrioletti, R. Mantovani, Overexpression and alternative splicing of NF-YA in breast cancer, Sci. Rep. 9 (1) (2019) 12955, https://doi.org/10.1038/s41598-019-49297-5.
[84] S. Mamat, J. Ikeda, T. Tian, Y. Wang, W. Luo, K. Aozasa, E. Morii, Transcriptional regulation of aldehyde dehydrogenase 1A1 gene by alternative spliced forms of nuclear factor Y in tumorigenic population of endometrial adenocarcinoma, Genes Cancer 2 (10) (2011) 979–984, https://doi.org/10.1177/1947601911436009.
[85] A. Gallo, M. Ronzio, E. Bezzecchi, R. Mantovani, D. Dolfini, NF-Y subunits overexpression in gastric adenocarcinomas (STAD), Sci. Rep. 11 (1) (2021) 23764, https://doi.org/10.1038/s41598-021-03027-y.
[86] P.V. Troshin, J.B. Procter, G.J. Barton, Java bioinformatics analysis web services for multiple sequence alignment–JABAWS:MSA, Bioinformatics 27 (14) (2011) 2001–2002, https://doi.org/10.1093/bioinformatics/btr304.
[87] A.M. Waterhouse, J.B. Procter, D.M.A. Martin, M. Clamp, G.J. Barton, Jalview Version 2 – a multiple sequence alignment editor and analysis workbench, Bioinformatics 25 (9) (2009) 1189–1191, https://doi.org/10.1093/bioinformatics/btp033.
[88] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (5) (2004) 1792–1797, https://doi.org/10.1093/nar/gkh340.
[89] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (6) (2004) 1188–1190, https://doi.org/10.1101/gr.849004.
[90] S. Kumar, G. Stecher, M. Suleski, S.B. Hedges, TimeTree: a resource for timelines, timetrees, and divergence times, Mol. Biol. Evol. 34 (7) (2017) 1812–1819, https://doi.org/10.1093/molbev/msx116.
[91] I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation, Nucleic Acids Res. 49 (W1) (2021) W293–W296, https://doi.org/10.1093/nar/gkab301.
[92] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, et al., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes, Genome Res. 15 (8) (2005) 1034–1050, https://doi.org/10.1101/gr.3715005.
[93] P. Puigdevall, R. Castelo, GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor, Bioinformatics 34 (18) (2018) 3208–3210, https://doi.org/10.1093/bioinformatics/bty311.
[94] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (1) (2013) 15–21, https://doi.org/10.1093/bioinformatics/bts635.
[95] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative genomics viewer, Nat. Biotechnol. 29 (1) (2011) 24–26, https://doi.org/10.1038/nbt.1754.
[96] V. Vacic, V.N. Uversky, A.K. Dunker, S. Lonardi, Composition profiler: a tool for discovery and visualization of amino acid composition differences, BMC Bioinformatics 8 (1) (2007) 211, https://doi.org/10.1186/1471-2105-8-211.
[97] B. Xue, R.L. Dunbrack, R.W. Williams, A.K. Dunker, V.N. Uversky, PONDR-FIT: a meta-predictor of intrinsically disordered amino acids, Biochim. Biophys. Acta Proteins Proteomics 1804 (4) (2010) 996–1010, https://doi.org/10.1016/j.bbapap.2010.01.011.
[98] D.T. Jones, D. Cozzetto, DISOPRED3: precise disordered region predictions with annotated protein-binding activity, Bioinformatics 31 (6) (2015) 857–863, https://doi.org/10.1093/bioinformatics/btu744.
[99] R. Linding, L.J. Jensen, F. Diella, P. Bork, T.J. Gibson, R.B. Russell, Protein disorder prediction: implications for structural proteomics, Structure 11 (11) (2003) 1453–1459, https://doi.org/10.1016/j.str.2003.10.002.
[100] G. Erdős, M. Pajkos, Z. Dosztányi, IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation, Nucleic Acids Res. 49 (W1) (2021) W297–W303, https://doi.org/10.1093/nar/gkab408.
[101] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold - making protein folding accessible to all, bioRxiv (2022 Jan 1), https://doi.org/10.1101/2021.08.15.456425, 2021.08.15.456425.
[102] E.F. Pettersen, T.D. Goddard, C.C. Huang, E.C. Meng, G.S. Couch, T.I. Croll, J. H. Morris, T.E. Ferrin, UCSF ChimeraX: structure visualization for researchers, educators, and developers, Protein Sci. 30 (1) (2021) 70–82, https://doi.org/10.1002/pro.3943.