UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

Dipartimento di Bioscienze
PhD course in Molecular and Cellular Biology
XXXVI cycle

The mesenchymal identity of NF-YA long isoform: from vertebrate evolution to cancer

Scientific Supervisor:
Dott. Diletta Dolfini

PhD Student:
Alberto Gallo
Matriculation n° R12998
Academic Year 2022-2023

Abstract (English)

NF-Y is a pioneer, highly conserved transcription factor (TF) composed of the subunits NF-YB/NF-YC, containing the histone-fold domain (HFD), and the regulatory NF-YA. NF-YA recognizes the CCAAT box on the DNA, a motif widely enriched in the promoter of key genes involved in cell proliferation, metabolism, and apoptosis. Due to the functions it controls, unitedly to its ability to increase local chromatin accessibility, the NF-Y complex is a cornerstone of many cancer types development.

NF-YA undergoes alternative splicing (AS), mainly involving exon-3. NF-YAl includes exon-3 and is linked to differentiation, while NF-YAs lacks it and sustains embryonic stemness. In epithelial cancers, imbalances in NF-YA isoforms expression correlate with diverse expression patterns. For instance, breast cancer Claudin^{low} invasive subtype has an increased NF-YAl expression, which relates to a mesenchymal identity and poorer survival outcomes compared to the epithelial luminal subtypes, where NF-YAs prevails.

During my PhD, I initially conducted a comprehensive study of NF-YA isoforms expression in stomach cancer, which unveiled a link between poor prognoses and high NF-YAl/NF-YAs ratio (NF-YA$_{Ratio}$) levels and described a 79 samples Claudin^{low} node. I then proposed a 158-genes signature with prognostic value, gathering genes co-expressed with NF-YAl in breast and gastric cancers. Integrating scRNA-seq deconvolution, I identified specific cell populations with mesenchymal identity associated with both the Claudin^{low} subtype and high NF-YA$_{Ratio}$ samples.

I performed a phylogenetic analysis that uncovered distinct patterns in NF-YA splicing across vertebrates, with NF-YAl being primarily expressed in differentiated/ mesoderm-associated samples. I additionally highlighted a unique isoform in avian species, NF-YAg, and a conserved retrogene, NF-YAr, in selected groups of mammals. Overall, my work significantly contributed to understanding NF-YA isoforms role in cancer, carrying potential diagnostic and therapeutic implications, while shedding light on their evolutionary dynamics and the link between NF-YAl and mesodermal fate determination.

Abstract (Italiano)

NF-Y è un fattore di trascrizione pioniere altamente conservato, composto dalle subunità NF-YB/NF-YC, contenenti lo "histone fold domain" (HFD), e dal regolatore NF-YA. NF-YA riconosce il CCAAT box sul DNA, un motivo diffuso nei promotori di importanti geni coinvolti nella proliferazione cellulare, nel metabolismo e nell'apoptosi. A causa delle funzioni che controlla, unitamente alla sua capacità di incrementare l'accessibilità della cromatina, NF-Y rappresenta un elemento essenziale nello sviluppo di molti tipi di cancro.

NF-YA è soggetto a splicing alternativo, concernente soprattutto l'esone-3. NF-YAl include l'esone-3 ed è legato alla differenziazione, mentre NF-YAs, che ne è privo, sostiene la staminalità nell'embrione. Nei tumori epiteliali, gli squilibri nell'espressione delle isoforme di NF-YA correlano con diversi pattern di espressione. Nel sottotipo invasivo di cancro al seno Claudin^{low}, ad esempio, NF-YAl è ampiamente espresso, in correlazione con un'identità mesenchimale e una prognosi peggiore rispetto ai sottotipi luminali, in cui prevale NF-YAs.

Durante il mio dottorato ho condotto uno studio sulle isoforme di NF-YA nel cancro dello stomaco, rivelando una connessione tra prognosi sfavorevole e valori elevati del rapporto NF-YAl/NF-YAs (NF-YA$_{Ratio}$) e descrivendo un cluster Claudin^{low} di 79 campioni. Ho proposto una signature con valore prognostico, composta da geni coespressi con NF-YAl nei tumori al seno e dello stomaco. Integrando la deconvoluzione tramite scRNA-seq, ho identificato specifiche popolazioni cellulari con identità mesenchimale associate sia al sottotipo Claudin^{low} che a campioni con elevato NF-YA$_{Ratio}$.

Tramite un'analisi filogenetica nei vertebrati ho scoperto pattern distinti nello splicing di NF-YA, con NF-YAl prevalentemente espresso in campioni differenziati/associati al mesoderma. Ho evidenziato un'isoforma esclusiva negli uccelli, NF-YAg, e un retrogene, NF-YAr, conservato in gruppi selezionati di mammiferi. Il mio lavoro ha contribuito ad approfondire il ruolo delle isoforme di NF-YA nel cancro, facendo luce sulle loro dinamiche evolutive e sull'associazione tra NF-YAl e la differenziazione mesodermica.

Contents

List of publications produced during the PhD **7**

1 Introduction **8**
1.1 Transcription Factors . . . 8
1.1.1 Protein and cis-regulatory elements conservation . . . 8
1.1.2 Interplay . . . 10
1.1.3 Binding motifs definition and design . . . 11
1.2 Alternative Splicing . . . 13
1.2.1 Spliceosome, RNA-binding proteins . . . 13
1.2.2 Impact on cancer . . . 15
1.2.3 Therapeutic instruments and targets . . . 16
1.3 Cancer and Epithelial to Mesenchymal Transition . . . 18
1.3.1 The Claudin^{low} subtype . . . 19
1.4 The CCAAT box . . . 22
1.4.1 Motif specificities . . . 22
1.4.2 CCAAT-binding TFs . . . 23
1.5 The NF-Y complex . . . 25
1.5.1 Trimer structure and DNA interaction . . . 26
1.5.2 Action, partners . . . 26
1.5.3 Alternative splicing isoforms . . . 28
1.5.4 Target genes . . . 29
1.5.5 Role in cancer . . . 32

2 Aim of the project **34**

3 Results **36**
3.1 NF-YA isoforms balance in the Claudin^{low} subtype . . . 36
3.1.1 NF-Y Overexpression and isoforms balance in stomach adenocarcinoma . . . 36
3.1.2 NF-YAl expression correlates with EMT in BRCA and STAD Claudin^{low} tumors . . . 51
3.1.3 Single cell RNA-seq analyses of breast and gastric Claudin^{low} cell fraction . . . 67
3.1.4 Identification of Claudin^{low}-specific splicing program . . . 70

3.2 Phylogeny of NF-YA locus 72
3.2.1 NF-YA across vertebrates' evolution 72
3.2.2 NF-YA isoforms with alternative splicing of exon-5 in *Aves* 91
3.2.3 Retrotransposon-mediated NF-YA gene duplications with potential regulatory functions in selected groups of mammals 104

4 Conclusions & Future perspectives 142

5 Materials & Methods 146
5.1 Section 3.1.4 146
5.1.1 scRNA-seq gene expression data acquisition and processing 146
5.1.2 Per-cell expression data visualization 146
5.1.3 CNV prediction and cancer cell status assignment 146
5.2 Section 3.1.5 147
5.2.1 Retrieval and handling of STAD exon-level expression data 147
5.2.2 Significant Claudin^{low}-specific splicing event detection using satuRn 147
5.2.3 Functional analysis and exon-level expression visualization 148

6 References 149

List of Figures

1.1 The essential domains and functions of transcription factors 9
1.2 Pioneer TFs promote chromatin accessibility 11
1.3 Graphical interpretation of transcription factor binding site motifs 12
1.4 The constitutive and alternative splicing of pre-mRNAs 14
1.5 Regulation of epithelial- and mesenchymal-specific AS events 15
1.6 Alternative splicing in cancer hallmark genes & ASO-mediated therapy . . 17
1.7 EMT is a key step for cancer invasiveness 19
1.8 Claudin^{low} classification as a breast cancer subtype 21
1.9 CCAAT box sequence logo 23
1.10 NF-Y acts as a transcription initiation platform for ESC master TFs in enhancer regions 27
1.11 The NF-Y complex: functional domains, NF-YA isoforms 29
1.12 NF-YAl expression correlates with worst prognoses and with the Claudin^{low} phenotype in BRCA 33

3.1 Claudin^{low}-directed scRNA-seq analyses in BRCA and STAD 68
3.2 Preliminary analysis of Claudin^{low}-specific splicing events 71
3.3 NF-YA exon-3 expression patterns across vertebrates 73
3.4 NF-YAx expression patterns across vertebrates 74

4.1 NF-YA isoforms: impact on cancer & phylogeny 145

List of publications produced during the PhD

- Gallo A., Ronzio M., Bezzecchi E., Mantovani R., Dolfini D. **NF-Y subunits overexpression in gastric adenocarcinomas (STAD)**. Sci Rep 2021;11(1):23764.
- Bernardini A., Gallo A., Gnesutta N., Dolfini D., Mantovani R. **Phylogeny of NF-YA trans-activation splicing isoforms in vertebrate evolution**. Genomics 2022;114(4):110390.
- Gallo A.*, Londero M.*, Cattaneo C., Ghilardi A., Ronzio M., Del Giacco L., Mantovani R., Dolfini D. **NF-YAl drives EMT in Claudin^{low} tumours**. Cell Death Dis. 2023 Jan 28;14(1):65.
- Gallo A., Dolfini D., Bernardini A., Gnesutta N., Mantovani R. **NF-YA isoforms with alternative splicing of exon-5 in *Aves***. Genomics. 2023 Sep;115(5):110694.
- Gallo A.*, Bernardini A.*, Polettini S., Dolfini D., Gnesutta N. and Mantovani R. **Retrotransposon-mediated NF-YA gene duplications with potential regulatory functions in selected groups of mammals** (*Submitted*).

*These Authors contributed equally to the work.

1 Introduction

1.1. Transcription Factors

Transcription initiates gene expression by generating a primary RNA transcript from the DNA sequence of a gene. Serving as a pivotal first step, transcription is followed by post-transcriptional events, including RNA splicing and translation, culminating in the synthesis of a functional protein in the case of coding regions. Transcription also represents the prime target for gene regulation: via selecting of which genes are transcribed into RNA, different proteins are expressed in different tissues. Transcription and its regulation are both reliant on a specific class of proteins, transcription factors (TF)[1].

Traditionally, the term "transcription factor" broadly referred to any protein involved in transcription and capable of modifying gene expression levels. However, the contemporary use of the term implies a protein with the ability to (1) bind DNA in a sequence-specific manner through the DNA-binding domain (DBD), and (2) regulate transcription by its effector domain(s), while not joining the initiation complex[2]. Effector domains can either recruit RNA polymerase and/or other accessory factors that then facilitate specific stages of the transcription process, i.e. general transcription factors (GTF)[3], allow the activity of chromatin modifying enzymes, or interact with partner TFs[4] (**Figure 1.1**). Some transcription factors are even devoid of a functional effector, and simply act sterically by preventing other protein to contact DNA[5]. Typically, TFs contact the DNA at two well-defined locations: promoters, regions where RNA polymerase binds to initiate transcription from the downstream DNA sequence, and enhancers, distal regions that can increase the expression rate of genes by spatial proximity[6].

TFs are central players in cell-fate determination, being also able to de-differentiate adult cells back into induced pluripotent stem cells[7]. Moreover, approximately one-third of human developmental disorders have been ascribed to the dysfunction of these class of proteins[8], and transcription factors are overrepresented among oncogenes[9]. Designing a therapy approach that targets TFs can influence many facets of cancer progression, like blocking invasion capability, proliferation and self-renewal, evasion of immune system, and more[10]. Indeed, small molecules targeting either protein-protein interactions involving TFs[11] or the chromatin rearrangements induced[1] and/or exploited by TFs[12] have been tested with encouraging results.

1.1.1. Transcription Factors: protein and cis-regulatory elements conservation

In *Homo sapiens*, transcription factors make up approximately 8% of total genes[13], representing an essential and extensive component of the "genetic toolkit" that orchestrates development through precise spatial and temporal deployment of regulatory networks. Together with signalling proteins, TFs are among the most conserved protein sequences in the *Animalia* kingdom, with many families dating back to the last common ancestor of bilaterians, a timeframe that predates the Cambrian period[14].

[1] Pioneer TFs, see section **1.1.2**

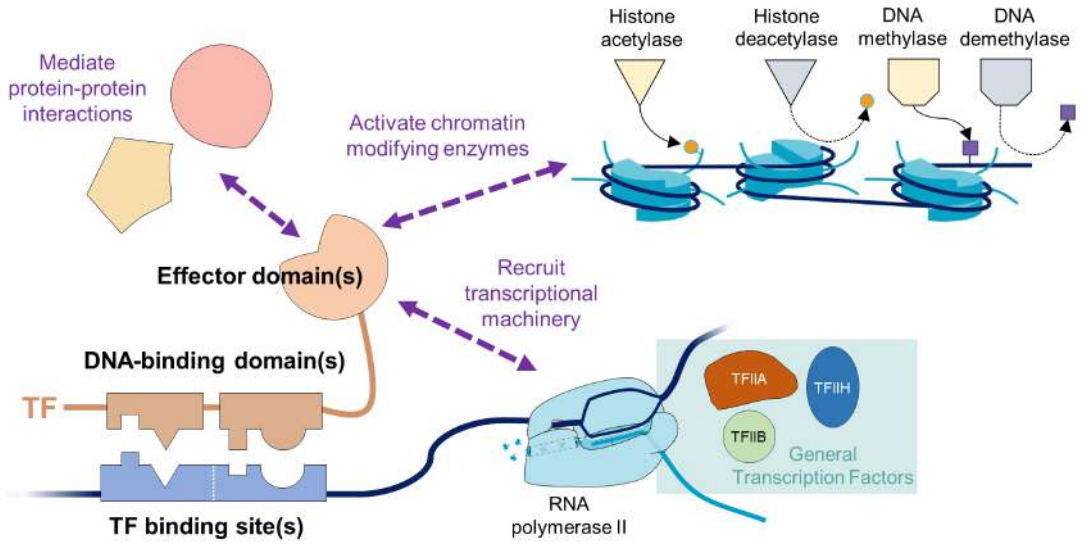


Figure 1.1: The essential domains and functions of transcription factors.
Scheme representing the minimal functional elements of transcription factors (TF). One or more DNA-binding domains recognize the corresponding binding site on the DNA, while the effector domain(s) enable TF transcriptional regulations. Nucleosomes graphic representation are from Franklin et al.*[15].*

It is the case of homeobox, *Sox*, *T-box*, and *Fox* orthologs found in the genomes of sponges and cnidarians, which may constitute a fundamental component of the core regulatory network at the basis of both animal development and multicellularity[16].

Within individual TFs, the DNA-binding domain is typically the most conserved portion of the polypeptide sequence. Changes in the DBD can alter the sequence specificity of the TF, and have a severe impact on the phenotype[17], and its conservation often exceeds the one of the *cis*-regulatory elements they recognize[18]. Intriguingly, divergent regulatory regions can even be bound by TFs and the transcriptional machinery across distant species and carry out comparable functions[2]. Generally, the increase of complexity of an organism is a function of the expansion of its *cis*-regulatory elements repertoire. In this context, TFs are integral part of an intricate system that recognizes several distal enhancers and core promoters and coordinates the higher-order spatial organization of genetic elements[20].

In light of these observations, the presence of a consistent genetic toolkit might suggest that TFs implicated in crucial processes are close to immutable, and that evolution primarily operates on the distribution of the DNA elements they bind within gene regulatory regions, rather than fixing alterations in the sequences contacted by highly conserved DBDs residues. Indeed, a study published in 2015 explored the binding specificities of 242 transcription factors in both *Drosophila melanogaster* and human, revealing that the majority of DNA sequences recognized by human TFs closely matched those targeted by fruit fly orthologs. Species-specific *cis*-regulatory elements were specifically bound by TFs governing processes exclusive to one of the two species, like wing development in

[2]A notable example of this is represented by enhancers driving the expression of human receptor tyrosine kinase (RET). Out of the 13 RET enhancers identified in humans, 11 can exercise the same regulations when introduced into zebrafish (*Danio rerio*), despite the sequence dissimilarity between the two species[19].

Drosophila and mucus-producing goblet cells differentiation in human[21]. Further, the main objection to ascribing a significant role to protein-mediated evolution of regulatory circuitry lies in pleiotropy[22], which describe the cases in which key mutations can impact multiple, unrelated phenotypic traits simultaneously[23].

Yet, a body of evidence demonstrates that *in trans* regulatory elements are susceptible to evolutionary pressures too: in fact, some adaptive and molecular mechanisms can counteract the negative pleiotropic effects induced by the mutation of a pivotal transcription factor. Among these, we find sub- and/or neo-functionalization following gene duplication[24], tissue-specific expression[25], alternative splicing[26], domain shuffling[27], and the gain and loss of protein–protein interaction motifs[28]. This data ultimately confirm how in actuality both regulatory DNA elements and transcription factors evolve "hand-in-hand" to effect phenotypic evolution[29].

1.1.2. Transcription Factors: interplay

Transcription factors generally collaborate to achieve both the requisite specificity in DNA binding and effector function. In cooperative binding, TFs aid each other in engaging DNA. In fact, numerous transcription factors are unable to independently bind DNA as single monomeric proteins, so they incorporate distinct protein–protein interaction domains that facilitate the formation of functional complexes, like dimers, trimers, or tetramers. These complexes can manifest as either homomeric, when formed by multiple copies of the same protein, or heteromeric assemblies, originated by the association of different proteins.

The interactions at the protein level enhance the complex's overall binding capacity. This boost, observed when multiple subunits or domains within the complex interact with the target DNA molecule, is termed avidity, in opposition to affinity, which in this context refers to the strength of the binding interaction between an individual TF and its binding site[30]. Moreover, the physical association of proteins can bring to the recognition of adjacent motifs, often palindromic if not, less commonly, direct repeats[31].

A crucial determinant of an active promoter is the presence of a transcription start site (TSS) within a region depleted of nucleosomes, the fundamental structural units of chromatin, which is also flanked by two well-positioned nucleosomes[32]. Another form of collaboration between TFs is the influence they may exert on chromatin state, which leads to a further layer of synergistic regulations. Pioneer transcription factors can in fact directly change the local chromatin conformation upon binding, making the TSS of genes as well as other DNA elements accessible for other TFs (**Figure 1.2**).

Besides this active form of enabling gene regulations, making the way to other binding partners, pioneer factors can also act in a more passive manner, when their mere presence allows for a speedier regulation by keeping chromatin unwound until the final, rate-limiting factor appearance, e.g. in response to a developmental signal[33].

This is substantiated by multiple observations reporting that most promoters and enhancers, especially those associated with tissue-specific expression, necessitate the binding of a combination of transcription factors for their functional activity[34, 35]. In order to access closed chromatin, pioneer TFs must possess the ability to engage nucleosomes. To achieve this, some of them have even evolved domains that mimic the structure of both core (H2A-H2B) and linker (H5) histones[36, 37].

1.1.3. Transcription Factors: binding motifs definition and design

The idiosyncrasies of sequence-specific transcription factors DNA-binding are often condensed into "motifs", which serve as models representing related and concise sequences favoured by each transcription factor. Establishing a DNA-binding motif is usually the initial phase in a comprehensive investigation of a transcription factor's function, since the identification of potential binding sites (referred to as TFBS) serves as a gateway to further in-depth analyses, and motifs can be employed to survey longer sequences like promoters and enhancers for this exact purpose[13]. Motifs are commonly visualized in the form of a sequence logo, where the height of each letter (nucleotide) is proportionate to its frequency, and the letters are arranged in descending order[38].

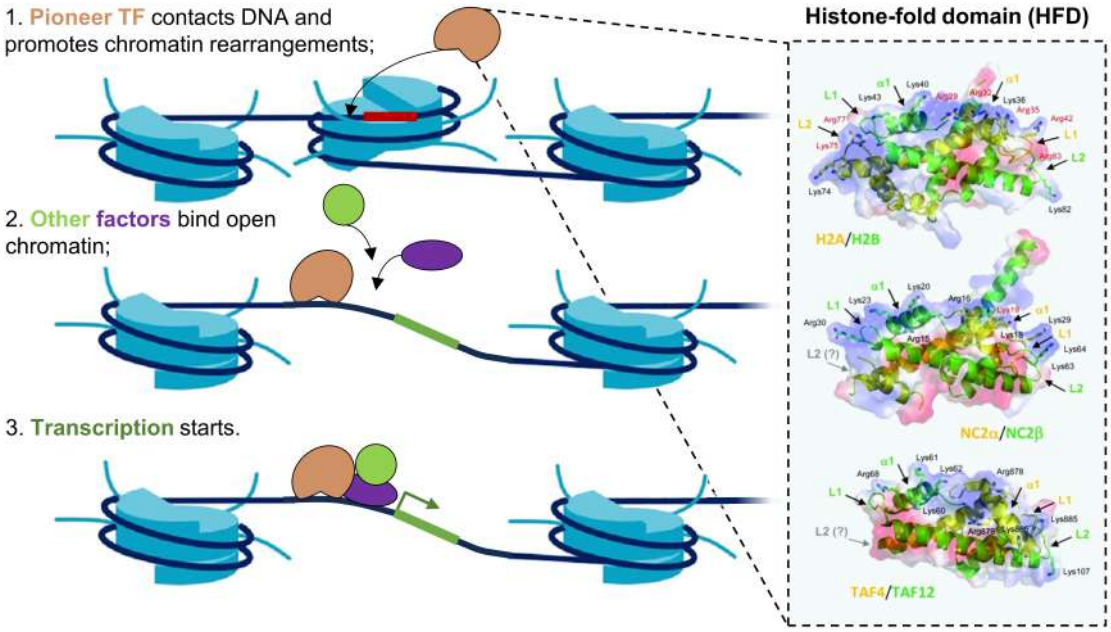


Figure 1.2: Pioneer TFs promote chromatin accessibility.
Simplified three-steps depiction of pioneer transcription factors action. Firstly, the pioneer TFs engages condensed DNA and stimulates the activity of chromatin-modifying enzymes; other TFs are then recruited to the open chromatin region, allowing transcription initiation. ***Right****: H2A/H2B crystal structure is shown, together with pioneer TFs proteins (NC2α/NFC2β and TAF4/TAF12) that mimick their structure through an histone-fold domain (HFD). Nucleosomes graphic representation are from Franklin et al.[15], while HFD structures were adapted from [37].*

The logo corresponds to an underlying table known as a "position weight matrix" (PWM), where the relative preference of the TF for each nucleotide in the binding site is indicated: within it, the four bases are assigned a score at each TFBS position. The computation of these scores for each base of a given candidate target sequence provides an estimate of the transcription factor's predicted relative affinity for that sequence[39] (**Figure 1.3**).

PWMs have been constructed from experimentally derived binding sites reported in the literature, as well as determined experimentally by assays such as bacterial-1-hybrid

system, protein-binding-microarrays, SELEX[3], MITOMI[4], or ChIP-chip / ChIP-seq[5][40].

For instance, in 2013 Jolma et al. obtained 830 human and murine TF binding profiles from a nonredundant set of 239 distinct PWMs using high-throughput SELEX and ChIP sequencing[41].

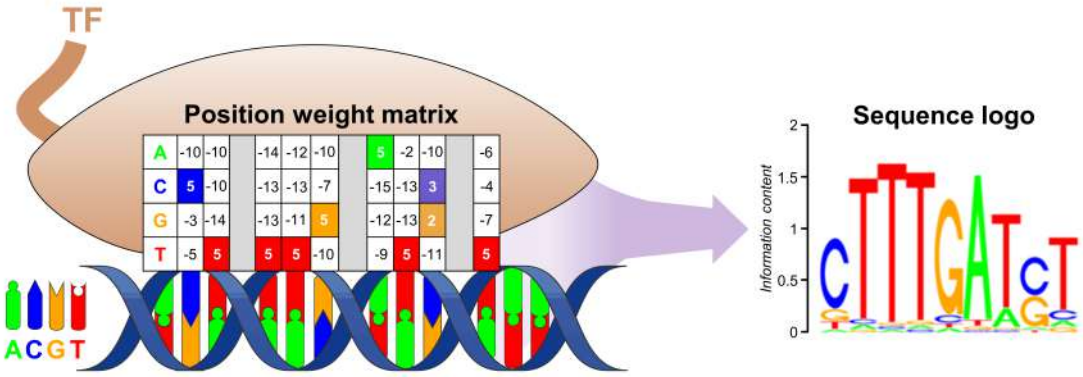


Figure 1.3: Graphical interpretation of transcription factor binding site motifs. ***Left Panel****: position weight matrix (PWM) model representing a hypothetical transcription factor binding site (TFBS) motif.* ***Right Panel****: the same motif is represented through a sequence logo, originally from Ambrosini* et al.*[42]. Parts of the figure were drawn by using pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).*

[3]"Systematic evolution of ligands by exponential enrichment".

[4]"Mechanically Induced Trapping of Molecular Interactions".

[5]"Chromatin ImmunoPrecipitation", combined with DNA microarray or DNA sequencing.

1.2. Alternative Splicing

TRANSCRIPTION occurs within the cellular nucleus, producing precursor mRNA (pre-mRNA) molecules, primarily consisting of two distinct regions originating from the corresponding gene: introns and exons. During RNA splicing introns are excised from the transcripts and exons, the coding portion of the gene, are retained. RNA splicing occurs concurrently with transcription, and evidence supporting the requirement for interaction between RNA polymerase II and pre-mRNA splicing have been documented[43, 44]. This essential step of gene expression contributes to the generation of mature transcripts, also known as messenger RNA (mRNA)[6].

Most multi-exon human genes undergo alternative splicing (AS), through which exons can be joined in different combinations. Consequently, a single gene can yield several distinct mature mRNAs, drastically increasing the protein diversity within the cell[48]. The members of a set of protein originated from the same gene, chiefly through AS, are called isoforms. The extensive generation of multiple isoforms encoded by key gene loci has been acknowledged as a crucial mechanism in cell differentiation[49], organismal development[50], and disease[51].

1.2.1. Alternative Splicing: spliceosome, RNA-binding proteins

RNA splicing is performed by the spliceosome, a massive and dynamic nuclear complex comprised of both RNA molecules and proteins. Core components of its machinery are five small nuclear ribonucleoproteins (snRNPs), produced by the association between small nuclear RNA molecules (snRNAs) and specific proteins, such as Sm and Lsm (Like-Sm). The spliceosome goes through continuous rearrangements during its assembly and activation, transitioning through several intermediate stages and ensuring that reactive groups in the pre-mRNA are accurately aligned for the catalytic splicing reactions to take place[52].

The complex detects three fundamental nucleotide sequences within the pre-mRNA: (1) the 5' and 3' splice sites (5'SS and 3'SS), which determine intron/exon boundaries, or splice junctions; (2) the branch point site (BPS), a short, conserved intronic sequence that connects, through the 2'-hydroxyl group of an adenosine, to the 5' splice site. This forms a lariat-shaped intermediate that is eventually resolved when adjacent exons are joined[53]; (3) the polypyrimidine tract, a stretch of consecutive pyrimidine nucleotides (C/U) located upstream of the 3' splice site, that stabilizes the interaction between the spliceosome and the pre-mRNA during the assembly[54] (**Figure 1.4A**).

Two spliceosomal complexes have been described, mainly distinguished by a subset of RNA components and the splice site sequences they target. Specifically, the U2-type (major spliceosome) preferentially recognizes GU-AG splice sites and is responsible for the removal of over 99% of introns[55], while the U12-type (minor spliceosome) can recognize both AU-AC and GC-AG sites, termed non canonical splice sites. These events collectively contribute to only about 1% of intronic excisions and are notably prevalent in the central nervous system[56, 57].

The spliceosome is often assisted by a plethora of regulatory splicing factors, a group of RNA-binding proteins (RBPs) that precisely modulate its activity depending on the

[6]This maturation process also involves two post-transcriptional modifications, namely polyadenylation at the 3' (poly-A) end and 5' end capping (5' cap)[45, 46], occurring at the transcript's untranslated regions (UTRs). UTR sequences, poly-A, and 5'cap control important aspects of the transcript fate, including initiation and rate of translation, mRNA stability, and localization[47].

cellular environment. These factors recognize and bind two categories of *cis*-regulatory elements on pre-mRNA: exonic or intronic splicing enhancer sequences (ESE or ISE) promote the inclusion of an exon in the final transcript, while intronic splicing silencer (ESS or ISS) repress it[58] (**Figure 1.4B-C**). Among RBPs families, two of the most studied are the serine/arginine-rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs)[59, 60]. SR proteins interact with splicing enhancers to facilitate splicing from adjacent splice sites, while hnRNPs typically impede splicing from neighboring splice sites by binding splicing silencers sequences[61]. The ratio between members of these two families in the nucleus is crucial in alternative splicing regulation, because SR family proteins act in opposition to hnRNPs proteins activity in splice site selection. Thus, they control AS in a concentration-dependent manner[62].

Other splicing factors exhibit some degrees of specificity, either being involved in distinct cellular processes or exclusively expressed within particular tissues. For instance, the NOVA1/2 and PTBP2 (Polypyrimidine Tract Binding Protein 2) factors are neuron-specific[63], whereas the CELF (CUGBP1, Elav-like factor) family regulations are heavily implicated in heart and skeletal muscle development[64].

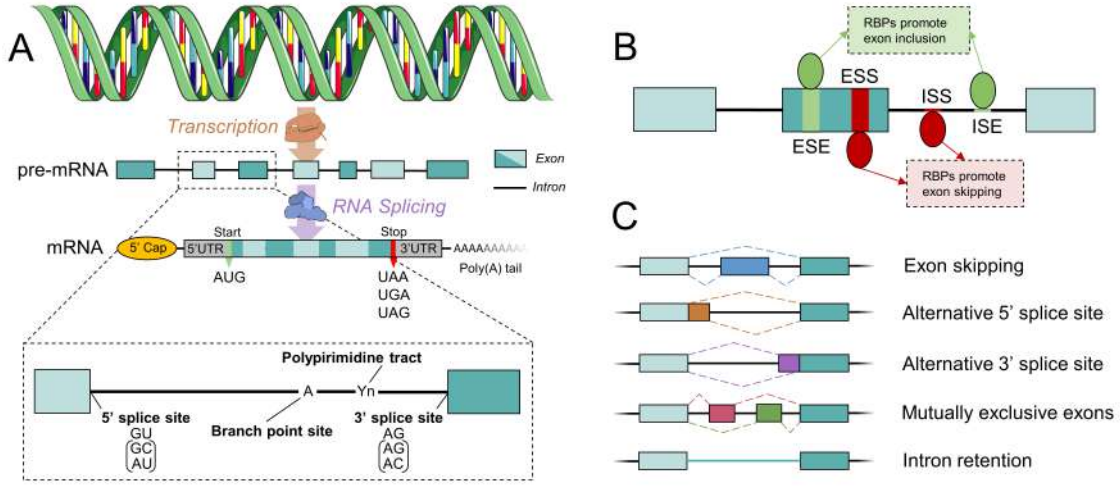


Figure 1.4: The constitutive and alternative splicing of pre-mRNAs.
***A**: After transcription, constitutive RNA splicing yields mature mRNA by excising introns from precursor mRNA (pre-mRNA). Aside from the coding sequence (CDS), final transcripts feature two untranslated regions at their 5' and 3' ends. A guanine cap binds the 5' UTR (yellow), while a poly-adenine chain (poly-A) is attached to the 3' UTR. RNA splicing requires three intronic regions: the branch point site (BPS), the polypirimidine tract, and the 5' and 3' splice sites. GU-AG represent the most common splice junctions (SJs), while noncanonical SJs are enclosed in parentheses. **B**: In alternative splicing, the spliceosome complex is usually joined by tissue-specific* trans*-acting regulatory splicing factors, which bind intronic and exonic splicing silencer (ISS/ESS) and enhancers (ISE/ESE), exercising opposing influences on the inclusion rate of exons undergoing alternative splicing. **C**: Alternative splicing encompasses various patterns, like exon skipping (also referred to as "cassette exons"), alternative usage of 5' and 3' splice sites, mutually exclusive exons, and intron retention. RBP = RNA-binding protein; parts of the figure were drawn by using pictures from Servier Medical Art.*

An alternative splicing regulatory model was proposed in 2016 for epithelial to mesenchymal transition (EMT), the process wherein epithelial cells undergo transdifferentiation into mesenchymal cells, a phenomenon crucial during embryo development, from gastrulation to the formation of mesoderm-derived tissues and organs, as well as for the origin of metastases[7][65]. In the model, two groups of splicing factors are predicted to carry out antagonistic AS regulations in epithelial and mesenchymal cells: on one hand ESRPs (Epithelial Splicing Regulatory Proteins) engage with downstream introns in epithelial cells to promote epithelial cell-specific exons inclusion, and with upstream introns to induce exon skipping. RBM47 (RNA Binding Motif Protein 47) either binds targets overlapping with the ones of ESRPs or contributes in some other way to maintain the epithelial splicing program. In mesenchymal cells, where ESRPs and/or RBM47 expression is substantially decreased or absent, RBFOX2 (RNA Binding Fox-1 Homolog 2) and QKI (Quaking) come into play by binding to downstream introns and fostering mesenchymal cell-specific exon inclusion. RBFOX2 may also binds to upstream introns, promoting in this case mesenchymal cell-specific exon skipping[66] (**Figure 1.5**).

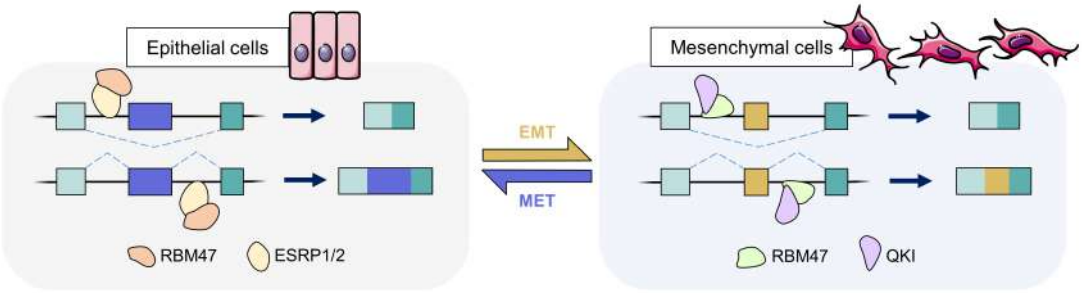


Figure 1.5: Regulation of epithelial- and mesenchymal-specific AS events.
EMT-associated AS regulation model presented by Yang et al.*[66]. In epithelial cells ESRPs, assisted by RBM47, promote tissue-specific exon inclusion by binding upstream and exon skipping by binding downstream. In mesenchymal cells, reduced ESRPs and RBM47, along with increased RBFOX2 and QKI, leads to similar mesenchymal-specific events. MET = mesenchymal to epithelial transition (see section **1.3**); parts of the figure were drawn by using pictures from Servier Medical Art.*

1.2.2. Alternative Splicing: impact on cancer

Anomalies in alternative splicing stand out as one of the prevalent molecular characteristic of most tumor types, either with splicing factors recurrent mutations and altered expression patterns, or in the form of cancer-associated isoforms that play crucial roles in cancer cell transformation, growth and invasion[51]. The frequent somatic mutations of genes such as *SF3B1*, *SRSF2*, *U2AF1*, and *ZRSR2* are collectively termed "spliceosomal mutations", and represent the most well-known alterations within the splicing machinery. They are especially diffused in hematological malignancies[67–69], where are also the are the most common genetic alterations, but also reported in solid tumors like uveal melanoma[70] and lung adenocarcinoma[71]. Spliceosomal mutations tend to be mutually exclusive, since their cumulative impact on alternative splicing and hematopoiesis results in apoptosis (synthetic lethality)[72].

Instead, splicing factors in solid tumors often display copy number or expression level alterations, but are hardly mutated[73]. Since SRs and hnRNPs operate in a

[7]EMT will be better detailed in section **1.3**.

concentration-dependent manner, unbalances in their expression can lead on its own to dysregulated AS patterns. Indeed, increased expression of the SR protein SRSF1 in breast, lung, colon, and bladder is consistently linked to tumorigenesis[74]. SRSF1 is located on chromosome 17q23, a region that is regularly amplified in breast cancer[75]. Additionally, the upregulation of this spicing factor is often coupled with MYC transcription factor overexpression[76]. SRSF1 dysregulation cascades into the production of protein isoforms that reduces cell death[77], augments proliferation[78], and affects resistance to DNA damage[79], ultimately contributing to cellular transformation.

As for tissue-specific RBPs, deregulations of ESRPs, RBFOX2, QKI and others elicit different effects depending on the context: ESRP1 functions as a tumor suppressor in EMT-driven metastasis formation[80] but exhibits oncogenic activity in specific cancer types[81–83]. In some cases QKI downregulation correlates with poorer prognoses, as it influences the alternative splicing of NUMB, to originate an isoform that reduces cell proliferation[84]; at the same time, it cooperates with RBFOX2 to promote mesenchymal-specific splicing events[66] and fuses to MYB to promote tumorigenesis in pediatric angiocentric glioma[85]. Finally, RBFOX2 overexpression facilitates cancer cells invasion in glioblastoma[86], breast[87], pancreatic[88] and laryngeal cancer[89], all while maintaining in non-pathological conditions a central role in the regulation of other RBPs expression, by modulating nonsense mediated decay (NMD) coupling with AS[90].

A comprehensive analysis of alternative splicing in human tumors revealed that, on average, cancers show up to a 30% increase in isoform variety when compared to normal tissue, with an average of 900 novel exon-exon junctions per cohort[91]. This inflation in spicing events repertoire can occur without the influence of mutations in the encoding gene or in splicing factors, but simply as inherent, tumor-exclusive differences in baseline splicing profiles.

Despite the specificity observed in tumor types or even subtypes, numerous dysregulated isoforms are recurrently shared across various cancers, suggesting the possibility of common splicing regulatory networks that transcend tissue-specific boundaries. Many of these cancer-specific isoforms are products of cancer hallmark genes (reviewed in Bradley & Anczuków[51] and **Figure 1.6A**). Deranged splicing patterns, like intron retention, widely affects tumor transcriptomes as a result of an interplay between *in-cis* (pre-RNA weaker splice sites, shorter lengths, and higher GC content) and *in-trans* (deregulated RBPs) mechanisms[92]. Many intron retention events, particularly affecting tumor-suppressor genes, also lead to the incorporation of a premature termination codon (PTC) in the final transcript, eventually triggering nonsense mediated decay response[93].

1.2.3. Alternative Splicing: therapeutic instruments and targets

The broad spectrum of alternative splicing defects shown in the previous section translates into several different strategies that can be devised for cancer therapy. The most generic approach is to limit AS all together, by targeting core spliceosome components like SF3B1[94] and the complex formed by the three snRNP U4/U5/U6[95], involved in the selection of 3'SS/BPS and assembly of both major and minor spliceosomes, respectively.

Dysregulated splicing factors can also be modulated, either directly inhibiting them[96] or hindering the activity of upstream protein that enable their functioning, like methyltransferases[97] or kinases[98]. Nonetheless, these therapeutical avenues impact splicing progression in both healthy and malignant cells, possibly leading to dangerous side effects. Thus, targeting cancer-specific products of deregulated AS may results in dedicated treatments with a more favorable side effect profile.

Using small molecules to bind cancer-relevant mRNA molecules has proven to be a challenging task, primarily due to the limited understanding of the mechanism of action of the hypothetical drug compound[99]. Designed splice-switching antisense oligonucleotides (ASOs), on the other hand, present delivery issues, but are otherwise suitable for clinical use. ASOs bind a reverse complementary sequence within the target transcript, effectively obstructing access to the spliceosome machinery and/or to AS regulatory factors. They can be customized to selectively target both splice sites, a splicing enhancer sequence (ESE /ISE) or a splicing silencer sequence (ESS/ISS), blocking splicing factor-mediated activation or repression, and inducing exon skipping or inclusion.

Preferred target sequences of splice-switching ASOs treatment are exonic splicing enhancers and the splice sites[100, 101]. ASOs can also cover a cryptic splice site caused by a mutation, reinstating the wild-type splice site[102, 103]. Lastly, ASOs usage can stimulate the inclusion of a poison exon, so-called due to the inclusion of a PTC in its sequence, into the mRNA of an oncogene[104] (**Figure 1.6B**).

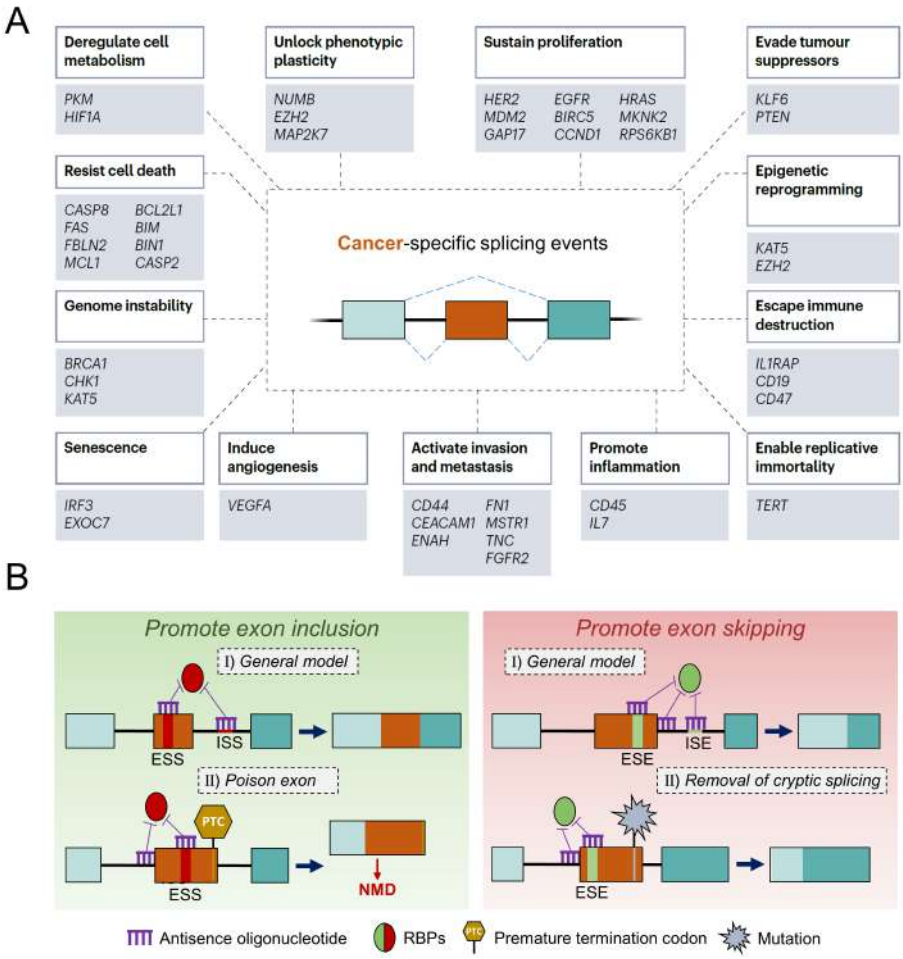


Figure 1.6: Alternative splicing in cancer hallmark genes & ASO-mediated therapy. ***A****. Examples of genes that encode cancer-specific isoforms and are implicated in the critical processes defined as cancer hallmarks; adapted from [51].* ***B****. Possible applications of splice switching antisense oligonucleotides:* ***Top Panels*** *represent the general strategy, in which ASOs are designed to target* cis*-regulatory regions and/or splice sites, preventing the binding of RBPs.* ***Bottom Panels*** *depict specific scenarios, like the inclusion of a poison exon in an oncogene's transcript to prevent its translation (****Left****) or the skipping of an exon carrying a deleterious mutation (****Right****); NMD = nonsense-mediated decay.*

1.3. Cancer and Epithelial to Mesenchymal Transition

EPITHELIAL to mesenchymal transition is a complex process where epithelial cells lose their apical-basolateral polarity as a consequence of cell junctions disassembly, including tight junctions (TJs), adherens junctions (AJs), gap junctions (GJs), and desmosomes. Particularly, TJs are dismantled trough reduced expression of claudins (*CLDNs*) and occludin (*OCLN*) and after depleting cell-cell contacts of zonula occludens 1 (*ZO1*); AJs are destabilized by the degradation of epithelial cadherin proteins (E-cadherin, *CDH1*); Desmosomes are mainly disrupted via desmoplakin (*DSP*) repression; finally, gap junction integrity is compromised due to decreased connexin levels (*CX* genes)[105]. The EMT switch is mediated by key TFs, including *SNAI1/2* (also referred to as SNAIL), the zinc-finger E-box-binding *ZEB1/2*, and basic helix-loop-helix transcription factors, such as *TWIST1*[106].

Via EMT, cells become motile and invasive following many intricate cytoskeleton reorganizations. These capabilities are paramount during gastrulation, when EMT drives pluripotent epithelial epiblast cells ingression, as they move from the surface layer into the underlying tissue to form the primary mesoderm. Yet, EMT is a reversible process, and its counterpart mesenchymal to epithelial transition (MET) is likewise crucial in embryogenesis and organogenesis[107] (**Figure 1.7A**).

Beyond development, EMT is implicated in cancer metastasis and the formation of cancer stem cells (CSCs), typically associated with poor responses to cancer therapies. In fact, introducing SNAIL or TWIST into mammary epithelial cells induces a mesenchymal population exhibiting a phenotype closely reminiscent of the one of epithelial stem cells, with the involvement of WNT and Notch signalling pathways. CSCs can initiate new tumors, as demonstrated by the *in vitro* ability of forming >30-fold more mammospheres than control cells, a feature shared with epithelial mammary stem cells[8]. Moreover, differentiated cancer cells and CSCs can transform one into the other as a product of EMT reversibility, carrying over any oncogenic mutations acquired. This depicts a scenario in which EMT promotes extravasation of CSCs/TICs from the primary tumor site, dissemination, and ultimately expansion into secondary tumors upon MET[109] (**Figure 1.7B**).

In recent years, the paradigm of EMT-focused research has shifted from a fixed, binary series of switches between epithelial to mesenchymal markers expression to a more plastic continuity of intermediate phenotypes within the E/M spectrum. Cells activating only a portion of EMT transcriptional programme, not fully committing to a complete mesenchymal transdifferentiation and expressing flexible mixtures of E/M markers, sustain a so-called partial EMT (pEMT)[110]. Partial EMT is especially impactful in pathological conditions: circulating cancer stem cells that show hybrid E/M phenotypes exit blood more efficiently, and an effective metastasis formation is now thought to involve the concomitant retention of epithelial morphology and aquisition of mesenchymal characteristics in CSCs. Therefore, cells undergoing partial EMT can even migrate collectively and not as individual tumor initiating cells, leveraging epithelial traits for a stable mutual adhesion and mesenchymal features for extracellular matrix attachment[111].

A comparative single-cell RNA-seq (scRNA-seq) study examined 12 distinct EMT time courses in lung, prostate, breast, and ovarian cell lines treated with the EMT-promoting factors TGFB1, EGF, and TNF. With over 100k cells from 960 samples, the study revealed the highly context-specific nature of partial EMT expression patterns: on average, the

[8]For this reason, they are sometimes termed Tumor Initiating Cells, TICs[108].

overlap in differentially expressed genes between pairwise comparisons of time courses was just over 20%, with cell line variability playing a more significant role compared to the differences introduced by the three different EMT-inducing stimuli[112].

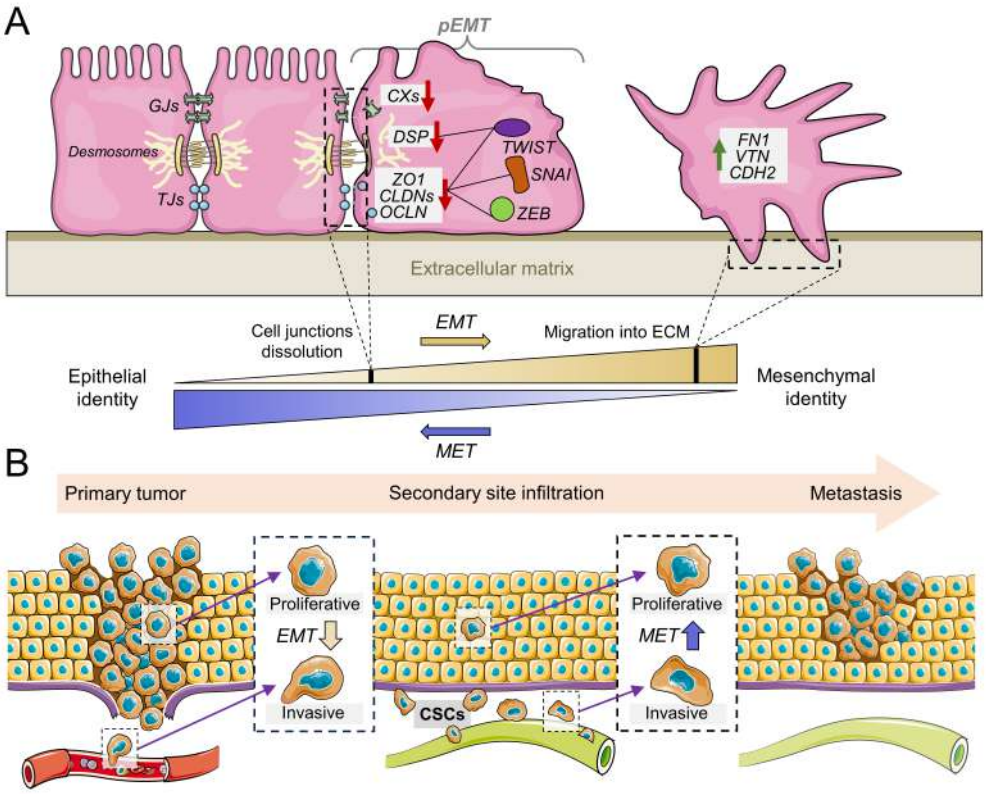


Figure 1.7: EMT is a key step for cancer invasiveness.
A. *Epithelial to mesenchymal and mesenchymal to epithelial transitions (EMT and MET respectively) are two reversible processes that drive the transdifferentiation of epithelial cells into mesenchymal cells, and vice versa. In EMT, the activation of master TFs such as ZEB, TWIST1 and SNAILs (encoded by the* SNAI1/2 *genes) determines the breakdown of cell-cell junction, extensive cytoskeleton reorganization, and upregulation of mesenchymal markers. Post EMT, partially transdifferentiated cells may invade the extracellular matrix. The label pEMT (partial EMT) designates the multifaceted phenotypic intermediates between the two extremes, common and advantageous during the formation of metastases.* ***B***. *In cancer, EMT serves as a pivotal strategy for metastatic dissemination. EMT cancer cells might also express stem cells markers, becoming cancer stem cells (CSCs), which invade the blood flow, extravasate and reach secondary sites. Here, they can alter their expression profile once more via MET, reacquiring an epithelial identity and a proliferative profile to initiate tumor growth. Parts of the figure were drawn by using pictures from Servier Medical Art.*

1.3.1. Cancer and EMT: the Claudin^{low} subtype

Breast cancer (BRCA) intrinsic subtypes were first determined via hierarchical clustering of the genes that exhibited significant alterations among different breast cancer samples[113]. This gene expression signature, consisting of 50 genes, was aptly named PAM50 (Prediction Analysis of Microarray 50), and generally describes four types: Basal-like, HER2-enriched, Luminal A and Luminal B[114].

The initial classification included also a Normal-like group, which showed high expression of marker genes for adipose tissue and other nonepithelial cell types[115]. Luminal subtypes make up the majority of categorized BRCA samples, and despite having each its specificities, e.g. Luminal B higher affinity for proliferation signatures, they share expression patterns associated to the epithelial cells lining the inner surfaces of mammary glands and ducts. As the name suggests, HER2-Enriched tumors are characterized by overexpression of the human epidermal growth factor receptor 2 (*HER2*) gene.

Basal-like cancers are characterized by the expression of genes found in basal epithelial cells, and are often triple-negative (TN), meaning they lack expression of estrogen receptor (ER), progesterone receptor (PR), and HER2[116]. Since treatment strategies for BRCA commonly target these receptors, basal-like tumors represent one of the most aggressive breast cancer groups. However, Basal-like and TN denominations should not be used interchangeably, as their definitions do not always overlap[117].

The Claudin^{low} subtype did not emerge in the original BRCA classification. Rather, it was proposed years after, as part of an analysis of conserved expression pattern in human and murine mammary tumors[118]. Claudin^{low} tumors share a low expression of cell-cell adhesion genes (especially *CLDN3/4/7*, hence the name), elevated expression of EMT- and CSCs-associated genes[119], pronounced immune and stromal cell infiltration[120], but are otherwise heterogenous in terms of mutational burdens and copy number aberrations[121]. These tumors are often triple-negative, a characteristic which, in conjunction with the highly invasive profile, determines markedly poor prognoses.

The first Claudin^{low} tumors predicting algorithm relied again on gene expression hierarchical clustering. It partitioned 9 cell lines out of 52, favouring those over primary tumors to circumvent the variability introduced by the tumor microenvironment[118]. This predictor has been in some cases paired to PAM50, overwriting with a Claudin^{low} label the previous subtype classifications[122].

A paper published in 2020 challenged Claudin^{low} status as a possible fifth (or sixth) BRCA intrinsic subtype, proposing instead a continuum model of "claudin-lowness", based on its inherent genetic heterogeneity. A compromise between the two paradigms could be to consider Claudin^{low} as a secondary, distinctive phenotype which may or may not be included in tumors belonging to the well-established intrinsic subtypes[123] (**Figure 1.8**).

Nevertheless, to understand the Claudin^{low} phenotype role in cancer, especially its speculated impact on worst survival, a deeper understanding of the cancer microenvironment, particularly the distinctive features of CSCs/TICs, is essential. To this end, comprehensive scRNA-seq analysis could help elucidate the issue. A scRNA-seq atlas of BRCA from 26 primary tumors, for instance, described a population of cells annotated as cancer associated fibroblast (CAFs), but that also expressed stem cells markers like *ALDH1A1*, *KLF4* and *LEPR118*[124].

Aside from breast cancer, Claudin^{low} signatures were proposed also for bladder[125] and gastric cancer[126], the latter also in correlation to decreased survival rates, possibly pointing towards shared underlying mechanism in the formation of metastases across epithelial cancers.

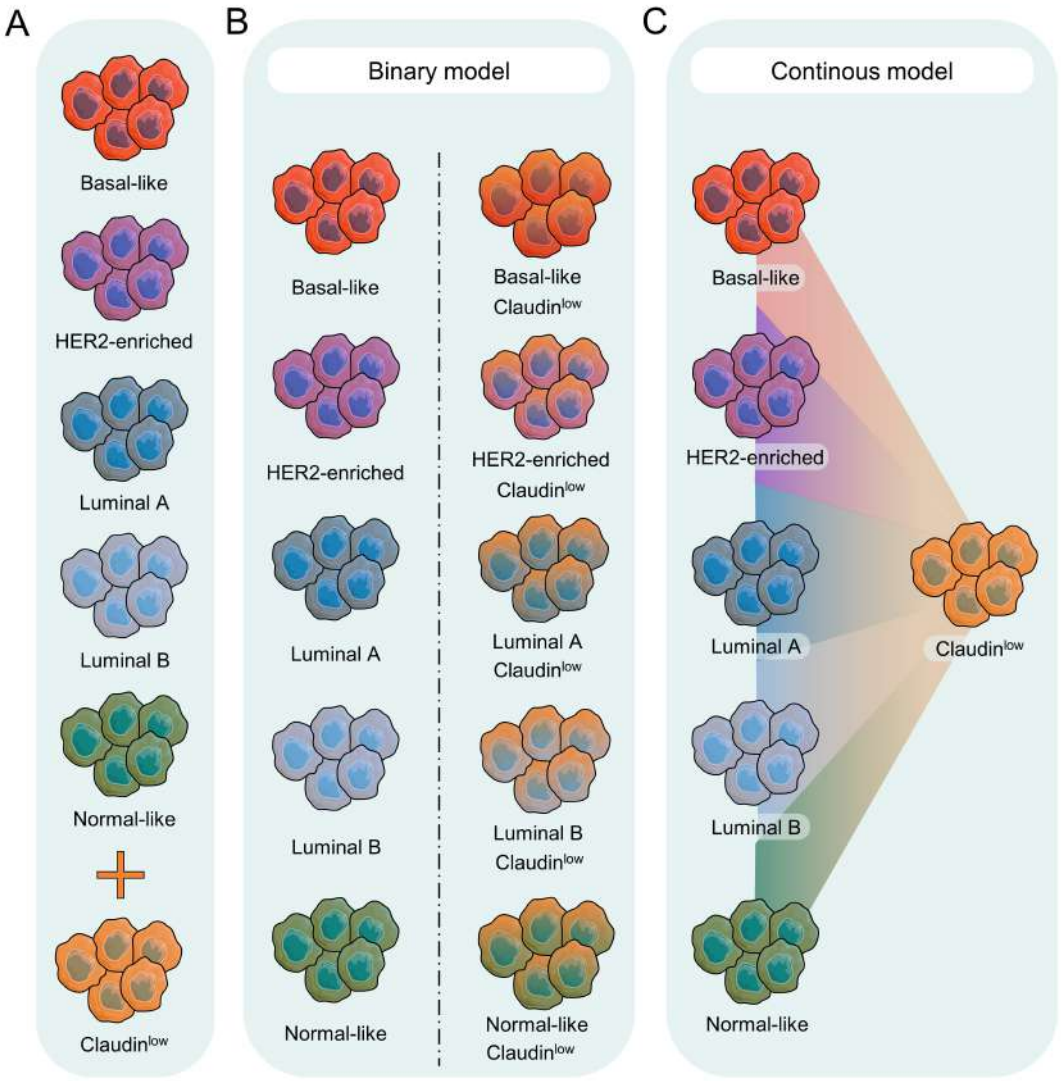


Figure 1.8: Claudin^{low} classification as a breast cancer subtype.
A. *Claudin^{low} is usually included as the sixth molecular subtype of breast cancer, alongside Basal-like, Luminal, HER2-enriched and Normal-like tumors.* ***B***. *Fougner* et al.*[123] proposed two alternative models for Claudin^{low} tumors classification: in the first one, Claudin^{low} is considered as a secondary phenotypic trait to add to the standard PAM50 subtype array.* ***C***. *The second model envisions a degree of "claudin-lowness" to assign to each sample, in a continuous scale between preexisting subtypes and "pure" Claudin^{low} cancers.*

1.4. The CCAAT box

The sequences of *cis*-regulatory elements like promoters and enhancers contain a constellation of short, conserved elements that mediate TFs activity, mainly transcriptional activation. Among these motifs we find the CCAAT box, one of the most enriched DNA elements found in the promoter regions of eukaryotic genes.

A statistical analysis conducted in 2004 encompassing over 4700 human putative promoter regions revealed that 60% (2839) of them harbored CCAAT boxes. This motif ranked as the second most frequent in the study, surpassed only by the GC box, with approximately 89% (4197) of total promoters containing it[127]. Six months later, another investigation focused on the distribution of all possible DNA 8-mers within a pool of approximately 13000 human promoters discovered that the ones exhibiting the most significant clustering around the transcription start site (TSS) could be categorized into nine groups of related sequences, with the CCAAT box being one of these[128].

The first observation of a consensus sequence including the CCAAT particle dates back to 1983, when a striking homology (17/18 identical bases) was denoted between the human and mouse sequences at positions -69 to -52 from the TSS of the Eα gene promoter, a member of the major histocompatibility complex (MHC) class II genes. The CCAAT box within this block was oriented in the 3'-5' direction, as ATTGG[129]. Later, this stretch of homology within MHC class-II gene promoter was reported in other vertebrates like rabbit[130] and mole-rat[131], and named Y box. The consensus sequence for this element was defined as CTGATTGG(T/C)(T/C)[9].

The Y box motif played a critical role in the activity of the Eα promoter, confirmed by deletion and replacement experiments across multiple immune cell types[132]. Moreover, Y box is essential for proper transcription initiation, as validated through *in vivo* experiments with transgenic mice[133]. Concomitantly, the CCAAT box was reported as a key functional element of promoters and enhancers in diverse conditions, regulating various genes, thus suggesting functional equivalence between Y and CCAAT boxes, and confirming that an intact CCAAT is often required for transcription to start[134–136].

1.4.1. The CCAAT box: motif specificities

The CCAAT box is usually present as a singular element between -60 and -100 relative to the TSS. Instances of multiple sites occurring within the same regulatory element are not common, but when this happens they are never closer than 27 bases one to another[137]. A thorough inspection of 178 CCAAT-containing promoters, mainly retrieved from vertebrate models, determined that the inverse motif ATTGG was slightly more frequent (60%) than the counterpart[138]. Only 57% of total promoters contained both CCAAT and TATA motifs, another prevalent motif crucial for transcriptional activation, contrasting with a previous estimate indicating that nearly 80% of eukaryotic promoters included a TATA box[139]. Furthermore, the presence of a TATA box shifted CCAAT sites, irrespective of their orientation, to more proximal positions[138].

This analysis was expanded in 2009, with the definition of two related position specific frequency matrix[10], termed p-CCAAT and g-CCAAT. The p-CCAAT matrix, from "promoter", was generated through a literature survey identifying promoters with CCAAT elements, resulting in 328 matches (294 human, 19 mouse, and 15 rat). Conversely, the

[9]Bases within parentheses indicate equal probability of occurrence at that position.

[10]PSFM, which directly shows observed frequencies of nucleotides at each position, as opposed to PWM in which log-odds scores are calculated, considering both observed and background frequencies.

g-CCAAT matrix, from "genomic", was derived from ChIP-chip data deliberately excluding RefSeq promoter regions to focus solely on distal locations of CCAAT motifs[140] (**Figure 1.9**).

Nucleotide frequencies in the regions flanking the CCAAT pentanucleotide were very similar between the two matrices, converging to the new proposed consensus sequence SRRCCAATSRSNVNSS[11] [140]. Additionally, promoter CCAAT sites exhibiting the highest affinity for the novel PSFMs were predominantly located within the -80 region, regardless of orientation, consistent with the previously noted positional bias. Though, in opposition to prior observations, CCAAT-TATA boxes co-localization within promoter regions appeared sparse, as signalled by a low number of promoters with high enrichment scores for both motifs. Instead, a clear correlation was reported between CCAAT and CpG island, DNA spans completely lacking methylation and rich in GC content, frequently overlapping promoter regions[141].

The CCAAT box has also been included as a cardinal motif within a general model that proposed a more "individualistic" approach for the *cis*-regulatory elements role in the core promoter (−150, +50 from TSS). In fact, these cardinal motifs recruit TFs that impart several distinctive cofactor signatures, leading to a potential subclassification of human promoters based on the presence of these key DNA elements, antithetical to the picture of a purely collaborative association of TFBS motif to mediate gene expression regulation[142].

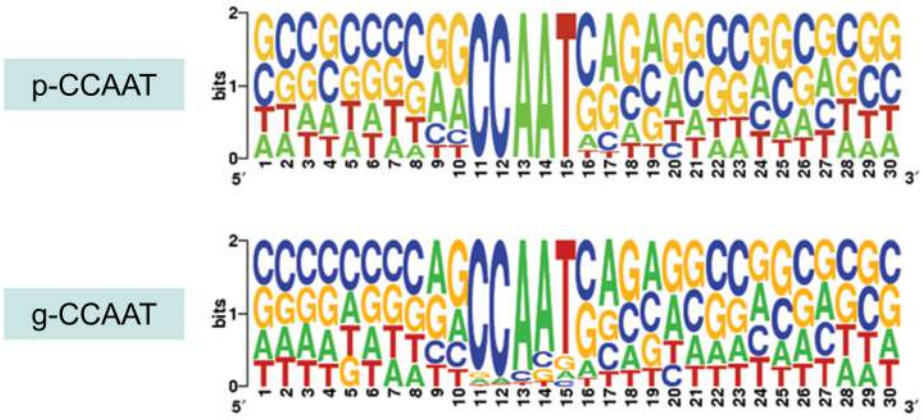


Figure 1.9: CCAAT box sequence logo.
The two CCAAT box sequence logos proposed by Dolfini et al. *in 2009[140]. The p-CCAAT, from "promoter", was the result of a literature survey focused on promoter regions, while the g-CCAAT, from "genomic", was obtained from ChIP-chip data of CCAAT motifs embedded within enhancers.*

1.4.2. The CCAAT box: CCAAT-binding TFs

Over the course of the decades since the discovery of the CCAAT box role in gene regulation, and by extension of the CCAAT containing Y-box, the two transcription factors YB-1 (Y Box-binding protein 1) and NF-Y (Nuclear Factor Y) have been suggested as the potential effectors that interact with the motif in promoter regions.

[11] Where S = G/C; R = G/ A; N = any base; V =G/C/A.

YB-1 was firstly identified as a CCAAT-binding factor by studying the promoter of DRa, the human homolog of mouse Eα class II protein[143]. YB-1 includes a cold-shock domain (CDS), a DNA-binding sequence conserved in eukaryotes and even prokaryotes[144]. However, conflicting observations arose regarding its specificity, as it was reported binding to sequences totally unrelated to CCAAT or Y boxes, and also to single-stranded and abasic[12] DNA[145].

These observations shifted YB-1 centered research from its controversial role in transcriptional initiation to its suspected involvement in mRNA post-transcriptional regulations[146]. *YB1* is a well-acknowledged oncogene, with a role in transformation and metastatic growth[147]. It localizes preferentially in the nucleus of cancer cells, in association with an increased *MDR1*[13] transcription. This genes encodes an efflux pump that cancer cells can use to evade cytotoxic compounds, leading to antitumor regimens resistance[148]. This suggests a potential mechanism wherein YB-1's transactivation activity may induce the overexpression of *MDR1*, by binding to the CCAAT box present in its promoter.

Nevertheless, this hypothesis has been challenged by two observations: the first relates to the multivocal YB-1 DNA binding specificities, as already discussed. On the other hand, this TF has a consistent RNA-binding activity, with both *in vitro* experimental confirmation and a ChIP-seq-derived consensus motif that mirrors the Kozak sequence, a mRNA element that includes the first codon and is recognized by the translation pre-initiation complex[149]. For these reasons, it has been postulated that YB-1 associates with newly synthesized mRNA rather than gene promoters, and due to proximity gets crosslinked to DNA during ChIP experiments. This assumption is sustained by the observation that YB-1-chromatin interactions are abolished when ribonucleases are used[150].

Being the focus of this thesis, the upcoming chapter will be entirely dedicated to exploring the second documented CCAAT-binding factor, NF-Y.

[12] A term that describes a site that has neither a purine nor a pyrimidine base.

[13] MultiDrug Resistance protein 1, also known as ABCB1, ATP Binding Cassette Subfamily B Member 1.

1.5. The NF-Y complex

The NF-Y complex[14] was discovered as well in mouse MHC class II Eα promoter[151]. In opposition to YB-1, NF-Y has consistently shown a strong specificity for the CCAAT box in multiple *in vitro* and *in vivo* analysis: in EMSAs[15], even single-nucleotide substitutions within the CCAAT pentanucleotide, and also in some of the flanking nucleotides, significantly reduce the binding affinity of the examined DNA sequence for NF-Y up to two orders of magnitude[152]. As for *in vivo* investigations, ChIP-chip/seq as well as inactivation experiments with dominant negative vectors[153] or small interfering RNAs (siRNA)[154] unequivocally demonstrated that NF-Y binds the CCAAT box and drive transcriptional initiation at target genes promoters.

NF-Y is a ubiquitous heterotrimer consisting of the three subunits NF-YA, NF-YB and NF-YC. NF-YA recognize the CCAAT box on the DNA through its DBD, while NF-YB and NF-YC feature an histone fold domain (HFD), which interacts with the DNA in a similar fashion to the core histones H2A/H2B[37]. The complex assembles within nuclei to carry out transcriptional regulation: NF-YA can inherently enter the compartment, while HFD access is coordinated[155]. All three proteins sequence and function are remarkably conserved in eukaryotes, to the point that the yeast *S. cerevisiae* HAP and *H. sapiens* NF-Y complexes bind to the same targets, establish the same interaction with the DNA, share the same responses to CCAAT box mutations, and are generally replaceable one with the other[156].

The HAP complex is composed of the four subunits HAP2/3/4/5: complexes missing HAP4 lack transcription activation activity[157], a function that in higher eukaryotes has been separated and included into the glutamine-rich transcription activation domains (TAD) of NF-YA (corresponding to yeast HAP2) and NF-YC (HAP5), which are redundant within a functional trimer[158]. Moreover, all NF-Y subunits are all essential for an effective binding to the CCAAT box: in both *Drosophila melanogaster* and Zebrafish, inactivation experiments of NF-YA and NF-YB, respectively, led to embryo lethality or severe developmental defects[159, 160]. In *Mus musculus*, NF-YA KO causes early embryonic death as a consequence of impaired S phase progression and apoptosis[161], whereas small hairpin RNA (shRNA) interference targeting NF-YB or NF-YC results in G2/M exit delay[162].

In animals and fungi each NF-Y subunit is encoded by a single gene, an uncommon feature among transcription factors. For instance, in *H. sapiens* only 6 other DNA-binding domain families out of a total of 77 have a single member like NF-Y does, according to the Human Transcription Factors database[13]. In plants, all NF-Y subunits underwent an extensive structural and functional diversification process, which generated three gene families with many members each, ranging from 8 to 39 depending on the species[163]. This heterogeneity is specifically attributed to gene duplications, events that show a higher incidence in plant genomes[164]. Moreover, the systematic analyses of this wide pool of NF-Y subunits pointed out many instances of functional specialization, with NF-Y proteins dedicated to key processes like embryo development, in which a whole group of NF-YB paralogs have been described and termed LEAFY COTYLEDON 1 (*LEC1*)[165], flowering time control[166], ER-stress[167], drought stress/resistance[168], and nodule de-

[14]Previously known also as CCAAT binding factor (CBF), CCAAT protein 1 (CP1), or HAP (Heme Activator Protein) in *Saccharomyces cerevisiae*.

[15]Molecular assays that highlight shifts in electrophoretic mobility of a DNA molecule when bound by a TF, compared to its free state.

velopment[169]. Broadly, plant NF-Y proteins can be categorized into two distinct classes: one with more general, ubiquitous functions similar to the animal counterpart, and the other with clear-cut roles in distinct pathways, in association with specific sets of partner TFs[163].

1.5.1. The NF-Y complex: trimer structure and DNA interaction

The acquisition of CCAAT-bound NF-Y complex crystal structure has revealed valuable insights on its protein-DNA interactions, as well as on the sequence-specificity conferred by the NF-YA subunit.

Just like core histones lack specificity to allow dynamic but stable interactions with DNA irrespective of the sequence, the HFD subunits NF-YB/NF-YC plausibly approach the DNA and bind aspecifically: NF-YA then recognize the CCAAT box by firstly associating to the dimer through its positively charged A1 helix, which interacts with a composite crevice formed by mostly negative residues near the NF-YC αC helix. Subsequently, the A2 helix contacts the pentapeptide, entering the DNA minor groove, while the A1A2 linker further stabilizes the interaction by simultaneously guaranteeing flexibility to the NF-YA chain and optimizing electrostatic interactions within the trimer. NF-YA A2 helix is closely followed by the conserved motif GxGGRF, which can also infiltrate the DNA minor groove. All these fundamental elements compose the NF-YA core domain, initially identified in *S.cerevisae* HAP2.

The protein-DNA interactions within the minor groove are a crucial aspect of the NF-Y DNA binding strategy, as they bend DNA double-helix between the CCAAT motif C and A sites and might facilitate or even encourage the binding of other TFs to the adjacent major grooves[170]. DNA minor groove binding is not a prerogative of the NF-Y complex: in fact, it has been previously reported in yeast TATA-binding protein (TBP)[171] and in sex-determining region Y protein (SRY), which serves as the initiator TF of male sexual differentiation in therian mammals[172]. However, both these factors exhibit a markedly distinct binding mechanism compared to NF-YA, as they are characterized by a broader interaction surface and they engage with the DNA through electrostatic and/or hydrophobic interactions. Instead, NF-YA's A2 helix contacts the CCAAT box with much less strength, adopting exclusively H-bonds and one Wan der Waals interaction generated by the phenylalanine of the GxGGRF motif. Furthermore, due to the flexibility of the first four residues of this motif, in the absence of DNA the C-term of NF-YA lack a secondary structure. Yet, in the presence of DNA, the region anchors the complex to the CCAAT, interacting with both strands. This results in a decreased entropy upon binding, a feature prominently associated with TF targeting the DNA major groove[173].

1.5.2. The NF-Y complex: action, partners

NF-YA distinctive binding of the CCAAT box, in conjunction with the high DNA affinity of the HFD and its ability to dislocate nucleosomes *in vitro*[174], as well as to associate with histone acetyltransferases (HATs) and deacetylases (HDACs)[175], has established the NF-Y complex as a vital pioneer transcription factor. In mouse embryonic stem cells (ESCs), NF-Y both regulate key cell-cycle genes binding their promoter proximal region and colocalizes with the master TFs *Oct4*, *Sox2*, *Nanog*, and *Prdm14* at active enhancers, hence detaining a precise role in stem cells self-renewal[176]. These factors exhibit heightened levels of occupancy, which drop after NF-Y depletion, concomitantly with a decreased chromatin accessibility[177] (**Figure 1.10**).

NF-YA inactivation results also in the accumulation of extra nucleosomes over the TSS region, in tandem with transcription starting at unusual locations, in a phenomenon known as ectopic transcription initiation[178]. As hinted by its interaction with both HATs and HDACs, NF-Y-mediated regulations can be divided in two groups: a major subset is correlated with an increased expression, marked by H3K9-14ac and H3K4me3 histone modifications, and a minor one linked to genes transcriptional silencing, accompanied by negative histone signatures like H4K20me3 and H3K27me3[179].

Although the mechanisms by which NF-Y impacts gene expression via the TAD regions of NF-YA and NF-YC are not fully understood, all these lines of evidence concur on a preeminent role in recruiting chromatin modifying enzymes and/or the pre-initiation complex (PIC) to target genes promoters[180]. Given the nature of its activity as transcription initiation platform, the NF-Y complex can cooperate with an extensive array of partner factors. Sp1 (specific protein 1), a zinc-finger TF, binds to GC boxes and has been frequently associated with NF-Y in the literature, evidenced by both co-expression[181] and ChIP-on-chip experiments[182][16]. Sp1 may not be the sole GC box-binding zinc-finger TF that collaborates with NF-Y in gene regulation, given the identical binding specificity within this class, which also includes KLF (Kruppel-Like Factor) proteins[183].

E2Fs constitute another NF-Y partner class of TFs, jointly regulating cell cycle and growth driving genes[184]. A direct link between NF-YA and E2F1 has been established through an experiment on apoptosis regulation, where the programmed cell death of mouse fibroblast upon overexpression of NF-YA was negated by E2F1$-/-$ double mutant[185]. Many additional TFs have been shown to interact either with NF-YA or the HFD, some even being targeted by the complex through CCAAT elements in their promoters. Among these, we have the CCAAT/enhancer-binding protein (C/EBP) α/ζ, SMAD2/3 involved in TGF-β Signaling Transduction, E box proteins like USF1/2 and MYC, and the ER stress-associated ATF6[186–188].

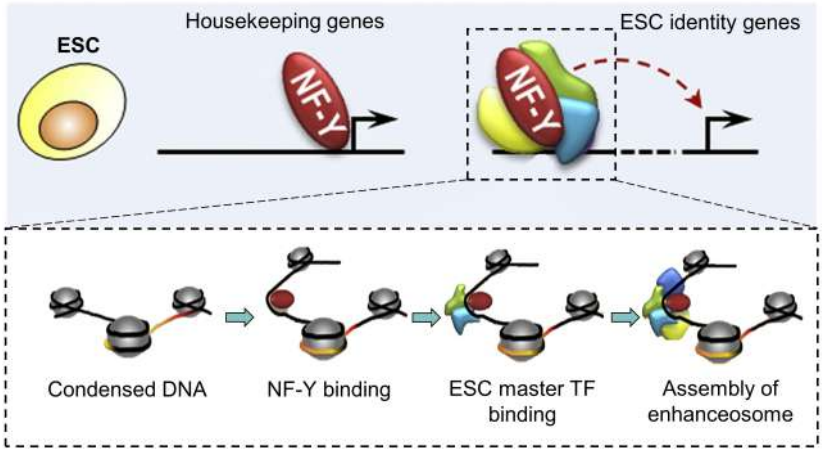


Figure 1.10: NF-Y acts as a transcription initiation platform for ESC master TFs in enhancer regions. *In embryonic stem cells (ESCs) the NF-Y complex promotes housekeeping genes expression by binding to proximal promoters region, while simultaneously increasing chromatin accessibility at distal regions, enabling the activity of ESCs self-renewal master TFs like* Oct4, Sox2, and Nanog. *Adapted from Oldfield* et al.*[177].*

[16]These analyses also indicate a positional bias for GC boxes, with two peaks at 15 and 27 nucleotides downstream of CCAAT[140].

In 2020, our laboratory conducted a comprehensive survey of the NF-Y regulome by analysing multiple ChIP-seq experiments retrieved from the ENCODE consortium[189]. This investigation revealed 116 factors that co-associated with NF-Y, which were assigned to four groups based on CCAAT enrichment in the corresponding ChIP-seq peaks and the percentage of peaks shared with NF-Y. Many classes of TFs were represented within these groups, with single predominant members. Some of them, such as the subunits of the repressive complexes NuRD and DREAM, were associated with NF-Y for the first time. The analysis also highlighted a novel partnership between NF-Y and RBPs, with the co-association to RBM25 and HNRNPLL[190].

1.5.3. The NF-Y complex: alternative splicing isoforms

NF-Y subunits are regulated by several post-transcriptional and post-translational mechanisms, including amino acids modifications like acetylation[191], phosphorylation[192], and Cysteine oxidation[193], and most notably in this context, alternative splicing.

In mammals, both NF-YA and NF-YC undergo AS. All currently reported NF-YA isoforms retain the HAP2 core domain, thus being able to interact with HFD subunits and bind the CCAAT box. The first variant of NF-YA, known as NF-YAs (short), lacks the 28/29 amino acids of exon-3, a segment of the N-terminal TAD domain. On the other hand, the isoform that includes exon-3 is referred to as NF-YAl (long). Two additional minor NF-YA splicing variants are associated with exon-3, namely L2 and L5, in which alternative splice sites selection results in exon-3 being one amino acid shorter at the N-term or C-term, respectively. In exon-7N, a segment from the 5' end of exon-7 is excited, generating a protein devoid of a valine-rich hexapeptide (VTVPVS), again located within the TAD[194] (**Figure 1.11**). Finally, an isoform skipping exon-3 and exon-5, as well as with the exon-7N variant, was characterized in primary human neuroblastoma and mouse embryo by RT-PCR[195].

NF-YC transcription activation domain is at the C-term of the protein, as opposed to NF-YA. Here, multiple splicing variants arise from different conformations of exon-8 and exon-9, albeit some of the variability is also conferred by two distinct promoter signals. The most expressed NF-YC isoforms are 37kD and 50kD in weight, and are mutually exclusive[196].

Since both NF-YA and NF-YC AS events involve the TAD domain, they could entail a different activation potential, although the precise nature of these differences remains purely speculative. What is known for certain is that NF-Y subunits' isoforms might be implicated in different types of regulations. In hematopoietic stem cells (HSCs), the master regulator HOXB4 (Homeobox B4) is subject to NF-Y-directed regulation[197].

When NF-YAs is overexpressed, it leads to the upregulation of HOXB4 and several other paralogous genes. Additionally, it enhances Notch and Wnt signaling and activates the telomerase enzyme, all of which are crucial pathways for the self-renewal of HSCs. In physiological conditions, the concentration of NF-YAs rapidly decreases during hematopoietic differentiation, leading to a significant loss (>50-fold) in HOXB4 promoter occupancy during terminal myelopoiesis[198].

NF-Y role as pioneer transcription factor extends to the differentiation of stem cells. It collaborates with JNK (Jun N-terminal kinases), a class of mitogen-activated protein kinases that phosphorylate and modify the activity of numerous nuclear proteins involved in processes such as response to UV irradiation, apoptosis, and notably differentiation, as their absence bring to mid-gestation embryonic lethality[199]. In an experiment on neuronal differentiation, where mouse ESCs committed to neuronal progenitors, a dominant

negative NF-YA mutant was found to abolish JNK DNA-binding. Given the crucial role of JNKs in ESCs differentiation, this suggests that the NF-Y/JNK connection might play a vital role in stem-to-differentiated cells transition[200].

One year later, NF-Y contribution to stemness maintenance and differentiation has been codified according to differential isoforms expression. In mouse ESCs, NF-YAs levels are initially high but decline as cells differentiate into embryoid bodies, where NF-YAl becomes the predominant isoform. Similarly, NF-YC isoforms exhibit a switch upon differentiation, with the 37 KDa isoform being prevalent in ESCs and the 50 KDa isoform dominating in embryoid bodies[201]. NF-YAl function in differentiation was reinforced in an experiment using mouse C2C12 myoblasts, where NF-YA exon-3 editing resulted in viable cells unable to form myotubes, a crucial step in the construction of muscle fibers. Moreover, this set a precedent for the involvement of NF-YAl in determining mesenchymal fate[202].

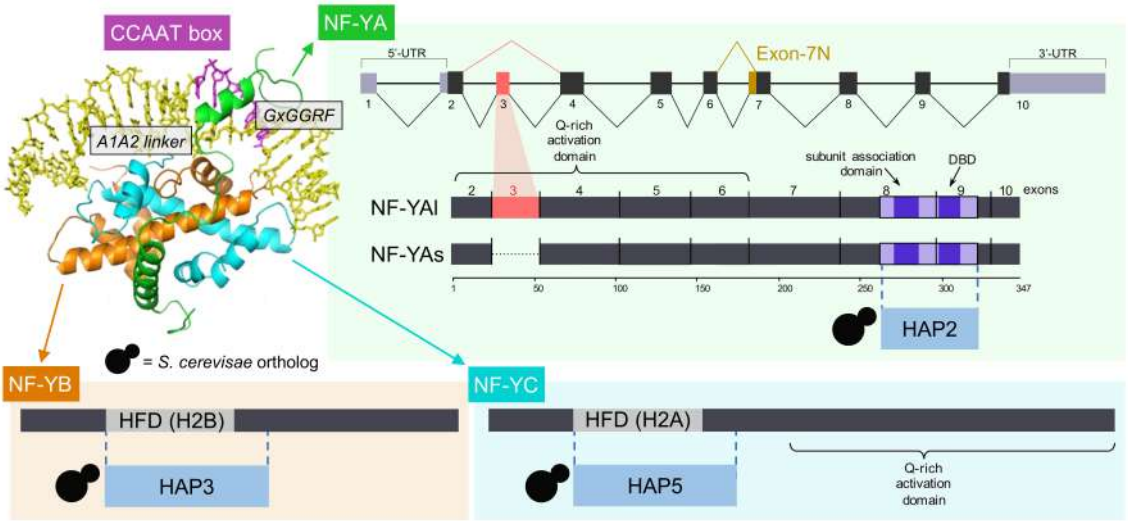


Figure 1.11: The NF-Y complex: functional domains, NF-YA isoforms.
Top Left Panel *shows the crystal structure of the NF-Y trimer bound to a CCAAT-box containing DNA region[170]. Grey boxes indicate the locations of NF-YA A1A2 linker and GxGGRF motif. The functional domains of NF-YA, NF-YB and NF-YC are presented, together with the yeast HAP orthologs. In the specific case of NF-YA, the main alternative splicing variants are also depicted.*

1.5.4. The NF-Y complex: target genes

Being a decisive and widely occurring motif in eukaryotes, The CCAAT box is found in the promoters of genes involved in a diverse range of cellular processes. Cell cycle is probably the most studied among these, given its intrinsic association with cancer development.

Although the CCAAT box is enriched in all five cell cycle phases gene promoter regions, it is especially linked to the G2/M phase regulation, where the cell undergoes the final preparations before cell division[203]. NF-Y regulates the expression of *CDC2* (Cell Division Cycle 2), Cyclin B1 (encoded by the *CCNB1* gene), and Cyclin B2 (*CCNB2* gene), the master regulators of this phase, as they form the maturation-promoting factor, essential for mitosis initiation and control[204, 205]; *PLK1* (Polo-like Kinase 1), that facilitates centrosomal maturation and the bipolar spindle formation[206]; *AURKB* (AURora

Kinase B), encoding a crucial protein for chromosomes alignment and segregation during mitosis and cytokinesis, via the localization of microtubules near kinetochores[207].

Outside G2/M, the CCAAT box is found in Cyclin A promoter (*CCNA1/2*), which is expressed from the S phase until late G2, when it is replaced by Cyclin B[208], as well as of three key genes involved in DNA replication: *PCNA* (Proliferating Cell Nuclear Antigen), *MYC*, and *MCM4* (Minichromosome Maintenance Complex Component 4)[209].

Another cellular process that prominently features CCAAT-containing promoters is apoptosis. Within the proapoptotic landscape, including all those genes which function is intertwined with programmed cell death, NF-Y regulates the expression of many key TFs: amidst them are *CHOP* (C/EBP Homologous Protein) and *XBP1* (X-Box Binding Protein 1), masters regulators of unfolded protein response, HSFs (Heat Shock Factors), that triggers the production of heat shock proteins in response to cellular stress, *HIF1α* (Hypoxia-inducible factor 1α), component of the heterodimeric TF HIF1, which orchestrates the cell reaction to hypoxia, as well as of *TP53/TP63* (Tumor Protein 53/63), two tumor suppressors heavily implicated in genome stability maintenance that are able to induce apoptosis[210].

In 2016, an intricate map of metabolic genes regulated by NF-Y was revealed through shRNA interference-mediated NF-YA inactivation and ChIP-seq experiments. Most of the genes that constitute this map are involved in lipids, sugars, amino acids, and nucleotides metabolisms. NF-Y regulations tended to increase the expression of anabolic genes bearing a CCAAT motif in their promoter, while repressing the ones related to catabolism[211]. Specifically, direct evidence of NF-Y binding were documented for *SREBP1*, one of the master regulator of cholesterol and fatty acids biosynthetic pathways[212], as well as for several enzymes comprising them, such as *HMGCS1*[213], *HMGCR*[214], *SQLE*[215], *DHCR7*[216], *ACACA*[217], *FASN*[218], *SCD1*[219].

Glycolytic gene promoters are densely populated by CCAAT motifs, which mediates the transcriptional control of *PFKFB2/3/4*, *PGK1/2*, *GAPDH*, *PKM*, *LDHA/B/C*[211], *HK2*[220], and *ALDOB*[221]. NF-Y exhibits a high enrichment in the Serine/Glycine pathway[17]. SOCG is required for both *de novo* production of the two amino acids, and for feeding into additional metabolisms essential for cell growth, like the folate cycle[222]. NF-YA removal decreases mRNA levels of key enzymes involved in Serine synthesis like *PHGDH*, *PSAT1*, and *PSPH*, the first of which was also reported to be bound by NF-YA *in vivo*[223]. Further, NF-Y regulates the SOCG mitochondrial branch enzymes *MTHFDL1*, *MTHFD2* and *DHFR*.

NF-Y seemed to interact with the promoter of most genes in the Glutamine pathway, with direct confirmation of the of CCAAT box genetic importance in the case of *GLS* (Glutaminase)[224], but the absence of *GLUD1* (Glutamate Dehydrogenase 1) regulation do not sustain anaplerosis, the replenishment of intermediates in the citric acid cycle (TCA). Therefore, NF-Y may be primarily involved in driving anaerobic energy production pathways. Supporting this hypothesis, NFY inactivation results in reduced expression of *ME1/2* (Malic Enzyme 1/2), which are responsible for pyruvate production from malate and exit from the TCA cycle[211]. Finally, NF-Y regulates the genes in polyamines and purines biosynthesis: in the former it exerts a direct control of *AMD1* expression and an indirect one of *ODC1*, in the latter it governs two tumor suppressor genes, namely *MTAP* and *AK2*, as well as the enzyme *RRM2*, implicated in the synthesis of deoxyribonucleotides and crucial for DNA replication[211].

[17] Also referred to as Serine, One Carbon, Glycine, SOCG pathway.

Cellular Process	Gene	Reference
Cell cycle, G2/M phase	*CDC2*	Zhu *et al.* 2004[204]
	CCNB1	
	CCNB2	Wasner *et al.* 2003[205]
	PLK	Uchiumi *et al.* 1997[206]
	AURKB	Kimura *et al.* 2004[207]
Cell cycle, G2 phase	*CCNA1/2*	Chae *et al.* 2011[208]
Cell cycle, S phase	*PCNA*	Benatti *et al.* 2016[209]
	MYC	
	MCM4	
Apoptosis	*CHOP*	Gatta *et al.* 2011[210]
	XBP1	
	HIF1α	
	TP53	
	TP63	
Lipid metabolism	*SREBP1*	Amemiya-Kudo *et al.* 2000[212]
	HMGCS1	Jackson *et al.* 1995[213]
	HMGCR	Howe *et al.* 2017[214]
	SQLE	Inoue *et al.* 1998[215]
	DHCR7	Prabhu *et al.* 2014[216]
	ACACA	Shi *et al.* 2012[217]
	FASN	Xiong *et al.* 2000[218]
	SCD1	Mauvoisin *et al.* 2007[219]
Glycolysis	*PFKFB2/3/4*	Benatti *et al.* 2015[211]
	PGK1/2	
	GAPDH	
	PKM	
	LDHA/B/C	
	HK2	Lee *et al.* 2003[220]
	ALDOB	Tsutsumi *et al.* 1989[221]
Serine/Glycine pathway	*PHGDH*	Jun *et al.* 2008[223]
	PSAT1	Benatti *et al.* 2015[211]
	PSPH	
	MTHFDL1	
	MTHFD2	
	DHFR	
Glutamine pathway	*GLS*	Pérez-Gomez *et al.* 2003[224]
Polyamines/purines biosynthesis	*AMD1*	Benatti *et al.* 2015[211]
	ODC1	
	MTAP	
	AK2	
	RRM2	

Table 1: *Summary table listing the NF-Y target genes mentioned in the previous section.*

1.5.5. The NF-Y complex: role in cancer

Considering that cell cycle, apoptosis, and metabolism are among the most dysregulated macroprocesses in cancer, it should come as no surprise the NF-Y complex plays a well-established role in cancer initiation and growth. For starters, the CCAAT box is among the most widely enriched motif in the promoters of genes that are deregulated in cancer, as reported by an unbiased *de novo* motif discovery performed on the Oncomine microarray database[225], as well as by analyses on RNA-seq data[226].

Unlike many other genes crucial for cancer biology, NF-Y subunits are rarely mutated. According to the COSMIC (Catalog Of Somatic Mutations In Cancer)[227] database, the highest rate of somatic mutations for the components of the trimer is barely over 3%. As a point of reference, *TP53*, the most frequently mutated gene in cancer, reaches rates over 90% in some of the tissues included in the catalog. Hence, the most likely mechanism by which NF-Y contributes to cancer progression is through alterations in the expression levels of its subunits compared to normal tissues.

In recent years, numerous studies have investigated the overexpression of NF-Y genes, starting from two works that linked a worst outcome in endometrial cancer patients to elevated NF-YA expression, specifically NF-YAs isoform[228, 229]. Subsequent studies reported NF-YA upregulation in stomach adenocarcinoma[230] and triple negative breast cancer[231]. In addition, NF-YC was found to be deregulated and linked to metastasis formation in both gliomas[232] and colon adenocarcinoma[233].

Building on these bases, our laboratory conducted a systematic analysis of NF-Y subunits and NF-YA isoforms in several epithelial cancer cohorts. In these, NF-YA was generally upregulated, and perturbations in the ratio of expression between its two main isoforms were consistently linked to distinct transcriptional programmes. In breast cancer (BRCA) a low NF-YAl/NF-YAs ratio (NF-YA_{Ratio}) were associated with a proliferative profile, marked by the overrepresentation of cell cycle-related gene signatures, whereas tumors with a high NF-YA_{Ratio} were characterized by the sustained expression of mesenchymal identity markers. $Claudin^{low}$ subtype cancers in particular were more prone to metastases and correlated to a worst prognosis[234] (**Figure 1.12**).

Similar findings were observed for lung squamous cell carcinoma (LUSC) [235], lung adenocarcinoma (LUAD)[236], and head and neck squamous cells carcinoma (HNSCC). In the latter, cell type deconvolution of bulk RNA-seq using a scRNA-seq reference revealed a direct correspondence between NF-YAl expression and the proportion of CAFs and partial EMT (p-EMT) cells, endowed with metastatic potential[237]. Liver hepatocellular carcinoma (HCC) represented the exception to the trend, since all NF-Y subunits were overexpressed, but no switch of NF-YA isoforms and no apparent role of NF-YAl in cancer invasiveness were reported[238]. Nevertheless, in all these cases NF-YA expression and/or the relative expression levels of NF-YA isoforms held some form of clinical impact.

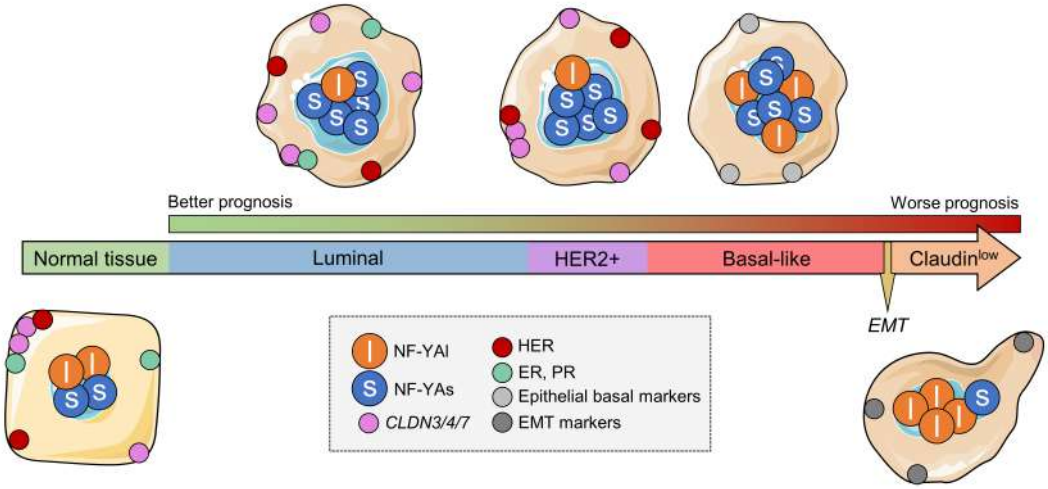


Figure 1.12: NF-YAl expression correlates with worst prognoses and with the Claudin^{low} phenotype in BRCA. *In this scheme, breast cancer molecular subtypes are arranged according to average prognosis, from left to right. Normal epithelial tissue is also included as reference. The partitioning of NF-YA long and short isoforms is represented by orange and blue circles in the nucleus of the cells representing each subtype. HER, ER and PR receptors are depicted in red and light green, while* CLDN3/4/7*, basal markers, and EMT markers expression is denoted by pink, light grey, and dark grey circles, respectively. Adapted from Dolfini* et al.*[234]; parts of this figure were drawn by using pictures from Servier Medical Art.*

2 Aim of the project

My PhD proceeded along two mostly parallel avenues ultimately converging under a common theme: the link between a heightened expression ratio of NF-YA main isoforms, specifically NF-YAl/NF-YAs (NF-YA_{Ratio}), and mesenchymal differentiation. The first investigation I performed had the purpose of expanding NF-Y-directed analyses of epithelial cancer cohorts to gastric cancer, leveraging the existing $Claudin^{low}$ signature as a starting point.

Gastric cancer, ranking fifth in cancer incidence and fourth in cancer-related deaths, primarily manifests as adenocarcinomas (STAD)[239]. The current gold standard for early-stage diagnosis relies on endoscopic examinations and forceps biopsies, allowing timely intervention to prevent cancer recurrence or spread. However, the heterogeneity of gastric tumors poses challenges in accurate diagnosis and predicting aggressiveness and treatment outcomes. To address this, molecular classification has become essential alongside the Lauren histological classification[240]. Particularly, in 2014 The Cancer Genome Atlas (TCGA) identified four gastric cancer subtypes using six molecular platforms[241], while a later investigation by the Asian Cancer Research Group (ACRG) introduced a second molecular classification, with a partial overlap with the previous one[242].

Using RNA-seq data from the extensive primary tumors collection of the TCGA repository, as well from available gastric cancer cell lines, I planned to inspect the expression patterns of NF-Y complex components, with a precise focus on the balance between the two main isoforms of NF-YA, in the different subtypes from established classifications. To ensure a comprehensive examination, I also proposed to expand the TCGA and ACRG categorizations to incorporate unallocated samples and a well-defined $Claudin^{low}$ subset.

Then, I intended to conduct survival analyses based on expression data, perform differential expression analysis followed by functional analysis, possibly validating any observed trends in NF-Y's direct involvement in cancer development and invasion, building upon the insights of previous investigations on the subject. Primarily, I was interested in the characteristic transcriptional patterns of different NF-YA_{Ratio} levels, since the analyses of other epithelial cancer cohorts, such as BRCA, revealed that low values were generally linked to a more proliferative profile, and tumors with high NF-YA_{Ratio} exhibited mesenchymal identity markers.

Further, to identify shared features within the $Claudin^{low}$ subtype across different cohorts, I aimed to identify genes consistently co-expressed with NF-YAl in both BRCA and STAD and elucidate the molecular mechanisms responsible for NF-YA isoform switching, including via the exploration of the involved splicing factors. Lastly, I intended to employ publicly accessible single-cell RNA-seq experiments to deconvolute bulk RNA-seq samples cell type proportions, to better characterize sample groups defined by different NF-YA_{Ratio} intervals.

The parallel project I was involved in focused on deciphering the phylogeny of NF-YA alternative splicing isoforms, chiefly directed at unveiling the identity of the "original" isoform between NF-YAl and NF-YAs. To achieve this, I planned to explore a wide collection of expression data, spanning from the lancelet (*Branchiostoma lanceolatum*), a representative of the *Chephalochordata* subphylum adopted as model organism for vertebrate evolution, to mammals. My strategy involved computing transcription levels for both isoforms and exons annotated within the NF-YA locus, with the aim of obtaining a precise overview of the transcripts expressed in tissues originating from distinct germ

layers: specifically, brain (ectoderm), liver (endoderm), and muscle (mesoderm). Additionally, this approach was devised to address the direct association between the prevalent expression of NF-YAl, reflected by exon-3 coverage, and mesoderm-related tissues.

3 Results

3.1. NF-YA isoforms balance in the Claudin^{low} subtype

3.1.1. NF-Y Overexpression and isoforms balance in stomach adenocarcinoma

During my initial NF-Y-focused analysis in epithelial tumors, I primarily examined the TCGA STAD RNA-seq dataset, which comprised 450 samples[1], and observed a significant increase in the expression of all NF-Y subunits in tumors samples when compared to normal adjacent tissue. At the isoform level, NF-YAs exhibited a substantial increase, unlike NF-YAl: in turn, NF-YAl/NF-YAs (NF-YA$_{\text{Ratio}}$) decreased in tumor tissues. However, in progression-free survival analysis (PFS) across all gastric cancers considered, samples with elevated NF-YA$_{\text{Ratio}}$ showed a significant association with a poorer prognosis.

The two currently available molecular classifications for STAD were initially proposed by TCGA[241] and ACRG[242], but did not cover the entire collection of TCGA gastric cancer samples. I extended these classifications using the DeepCC learning tool[243], employing already assigned samples as training set. In both, NF-YAs expression predominated in three out of four subtypes, while NF-YAl levels were elevated in the TCGA subtype GS (Genomically Stable) and in ACRG EMT tumors. In the TCGA classification, the genes upregulated in each of the four subtypes were enriched in several cell cycle-related Gene Ontology (GO) terms, with the CCAAT box being the most represented transcription factor binding site (TFBS) in their promoter regions.

Lastly, I isolated a new Claudin^{low} conservative node of 79 samples using a previously proposed gene signature[126], mostly within the partially overlapping GS and EMT subtypes. This novel subtype gathered STAD samples with the highest NF-YAl expression values, the lowest NF-YAs, and consequently the top NF-YA$_{\text{Ratio}}$ levels. Functional analysis of Claudin^{low} samples revealed many GO terms associated with both cell cycle, under direct transcriptional control of NF-Y, and mesenchyme development, where instead the CCAAT motif did not emerge among the most enriched motifs.

These results broadly recapitulated what observed in BRCA Claudin^{low} cell lines[234], underlying the similarities of the subtype across two distinct cancer cohorts, which we deemed worthy of further investigation.

[1] Specifically: 415 primary tumor samples + 35 normal tissue samples.

scientific reports

Check for updates

OPEN

NF-Y subunits overexpression in gastric adenocarcinomas (STAD)

Alberto Gallo, Mirko Ronzio, Eugenia Bezzecchi, Roberto Mantovani & Diletta Dolfini✉

NF-Y is a pioneer transcription factor—TF—formed by the Histone-like NF-YB/NF-YC subunits and the regulatory NF-YA. It binds to the CCAAT box, an element enriched in promoters of genes overexpressed in many types of cancer. NF-YA is present in two major isoforms—NF-YAs and NF-YAl—due to alternative splicing, overexpressed in epithelial tumors. Here we analyzed NF-Y expression in stomach adenocarcinomas (STAD). We completed the partitioning of all TCGA tumor samples (450) according to molecular subtypes proposed by TCGA and ACRG, using the deep learning tool DeepCC. We analyzed differentially expressed genes—DEG—for enriched pathways and TFs binding sites in promoters. CCAAT is the predominant element only in the core group of genes upregulated in all subtypes, with cell-cycle gene signatures. NF-Y subunits are overexpressed, particularly NF-YA. NF-YAs is predominant in CIN, MSI and EBV TCGA subtypes, NF-YAl is higher in GS and in the ACRG EMT subtypes. Moreover, NF-YAl^{high} tumors correlate with a discrete Claudin^{low} cohort. Elevated NF-YB levels are protective in MSS;TP53^{+} patients, whereas high NF-YAl/NF-YAs ratios correlate with worse prognosis. We conclude that NF-Y isoforms are associated to clinically relevant features of gastric cancer.

Abbreviations

TCGA	The Cancer Genome Atlas
ACRG	Asian Cancer Research Group
NF-YAl	Nuclear factor Y subunit A isoform long
NF-YAs	Nuclear factor Y subunit A isoform short
NF-YB	Nuclear factor Y subunit B
NF-YC	Nuclear factor Y subunit C
E2F	E2 factor
TF	Transcription factor
TFBS	Transcription factors binding sites
FDR	False discovery rate
HFD	Histone fold domain
STAD	Stomach adenocarcinoma
BRCA	Breast carcinoma
LUSC	Lung squamous cells carcinoma
LUAD	Lung adenocarcinoma
HCC	Hepatocellular carcinoma
HNSCC	Head and neck squamous cells carcinoma
CIN	Chromosome instability
EBV	Epstein-Barr virus
GS	Genomically stable
MSI	MicroSatellite instability
EMT	Endothelial to mesenchymal transition
MSS	MicroSatellite stable
TP53	Tumor protein 53
TIC	Tumor-initiating cells
DEG	Differentially expressed genes
PFI	Progression free interval

Gastroesophageal tumors are among the most widespread cancers worldwide[1]. Stomach adenocarcinomas—STAD—share a survival outcome of patients that, despite many efforts, remains poor. The Lauren histological

Dipartimento di Bioscienze, Università degli Studi di Milano, Via Celoria 26, 20133 Milan, Italy. ✉email: diletta.dolfini@unimi.it

classification divides gastric cancers into intestinal (IT), diffuse (DF) and mixed (MX)[2,3]. Further microarrays profilings studies have since classified tumors according to molecular subtypes[4–7]. More recently, TCGA has proposed a classification based on genetic mutations, chromosomal alterations, epigenetic features and RNA-seq expression data that included four subtypes: EBV (EBV-infected), MSI (MicroSatellite Instability), GS (Genomically Stable) and CIN (Chromosomal Instability)[8]. In parallel, the ACRG (Asian Cancer Research Group) proposed another classification, originally based on independent microarray profilings, also consisting of four subtypes: EMT (Epithelial to Mesenchymal Transition), MSS;TP53$^-$ (MicroSatellite Stable, inactive tumor protein 53), MSS;TP53$^+$ and MSI[9,10]. The two classifications are partially overlapping (Reviewed in Refs.[11–13]).

In general, cellular transformation causes—and in some cases is caused by—changes in mRNA production patterns. The first step in this process is the binding of sequence-specific transcription factors—TFs—to DNA elements in promoters and enhancers, entailing recruitment of chromatin modifying Cofactors[14]. Changes in the structure or expression of TFs can cause permanent changes that lead to transformation. The identification of TFBSs—transcription factor binding sites—in promoters of genes overexpressed in cancer led to the identification of the CCAAT box as one of the most widely enriched[15]. CCAAT is typically crucial for high-level expression of genes[16]. This box is recognized by NF-Y, a heterotrimer formed by the histone fold domain—HFD—dimer NF-YB/NF-YC and the sequence-specific NF-YA[17]. NF-YA has two alternatively spliced isoforms—NF-YAs and NF-YAl—differing in 28/29 amino acids coded by exon 3[18]. NF-YC is also present in multiple isoforms, resulting from alternative splicing at the C-terminal of the protein[19]. In both subunits, this involves the glutamine-rich trans-activation domains (TADs), while the subunits-interaction and DNA-binding domains are common to all isoforms.

NF-Y subunits are rarely mutated in tumors, yet the NF-Y regulome—ChIP-seq and functional analysis—point to cell-cycle and metabolic pathways being positively affected[20]: specifically, rate-limiting, cancer-promoting genes of different anabolic routes—amino acids, lipids, nucleotides—are activated[21].

Reports on the expression of NF-Y subunits in tumors emerged recently. In ovarian[22,23], breast[24,25], lung[26,27], liver[28] and head and neck squamous cell carcinomas (HNSCC)[29], overexpression of NF-YA was reported. As for gastric cancer, two studies provide evidence for a specific function of NF-YA: microarray-based differentially expressed genes (DEG) of gastric cancer identified NF-YA as a key TF, specifically in the DF subtype, with prognostic significance[30]; NF-YA inactivation has a more profound growth suppressive effect in a DF than in a IT cell line. Another study analyzing TCGA data found high expression of NF-YA, including of the protein in STAD specimens[31]; this correlated with Cyclin E, a gene often amplified and overexpressed in STAD datasets[32,33]. These two studies did not report on the relative levels of the two major NF-YA subunits, which are clinically important in breast, lung and HNSCC cancers[25–27], nor of the HFD subunits, which might be relevant in light on our recent finding on their overexpression in liver Hepatocarcinomas and HNSCC[28,29]. We report here on the analysis of STAD RNA-seq data present in TCGA, as further classified according to TCGA and ACRG. We confirm NF-YA global overexpression, extend this finding to HFD subunits, and investigate the isoforms of NF-YA.

Results

NF-Y subunits are overexpressed in STAD. Inspection of NF-Y subunits expression of the TCGA datasets (http://firebrowse.org) suggested that expression of NF-YA is globally increased in epithelial tumors[25]. We downloaded the available STAD RNA-seq dataset[8] and analyzed NF-Y subunits: NF-YA is robustly increased in STAD (p value: 10^{-14}). NF-YB and NF-YC are also increased (p values: $10^{-07/08}$) (Fig. 1a). We then analyzed the levels of NF-YA isoforms: Fig. 1b shows that the levels of the "short" NF-YAs increase in tumors (p value 10^{-15}), unlike NF-YAl. In conclusion, we confirm a generalized overexpression of NF-Y subunits, especially NF-YA, in STAD.

The predominance of NF-YAs prompted us to verify the relative expression in gastric cancer cell lines. For this, we interrogated two repositories: the Broad Institute CCLE—Cancer Cell Lines Encyclopedia (https://portals.broadinstitute.org/ccle/about) and a recently described set of gastric cancer lines[34]; overall, we analyzed 50 cell lines, with a partial overlap of lines common to the two datasets. We downloaded RNA-seq data, mapped reads and analyzed NF-Y subunits levels. The results are shown in Fig. S1: the overall levels of NF-YA mRNA expression are variable with the majority, but not all, cell lines expressing primarily NF-YAs (Fig. S1a). The levels of the two HFD subunits, particularly NF-YB, are comparably less variable among the cell lines (Fig. S1b,c). We conclude that NF-Y subunits are overexpressed in STAD, particularly NF-YA, whose predominant isoform is NF-YAs, in gastric tumors and cell lines.

Expression of NF-Y isoforms in STAD subtypes. According to several genetic, epigenetic and functional parameters, TCGA classified STAD in four subtypes[8]. Since overexpression of NF-Y subunits could be limited to one -or more- of the subtypes, we investigated the levels of the three subunits in the four cohorts. Currently, RNA-seq data on 415 tumors are available, of which 387 were categorized by TCGA. We first classified all tumors for which there are RNA-seq data, employing the DeepCC machine learning tool[35], with a training set represented by those already classified by TCGA: the relative proportions are indeed essentially maintained (Fig. 2a). Figure 2b (Left Panels) shows that the relative increase of NF-YA is similar in CIN, EBV and MSI (p values of $10^{-12/15}$ relative to normal samples), but in GS, the levels are lower. NF-YB and NF-YC are increased at comparable levels in all subtypes.

As for the isoforms, the data are shown in Fig. 2b (Right Panels): NF-YAs is increased in MSI, EBV and CIN (p values $10^{-14/16}$ with respect to normal samples), less in GS. NF-YAl, instead, shows a significant increase in GS. As a consequence of these changes, the NF-YAl/NF-YAs ratio is substantially increased in GS with respect to the other subtypes. In summary, overexpression of NF-YAs is generally widespread, but there is a distinctly higher NF-YAl/NF-YAs ratio in GS tumors.

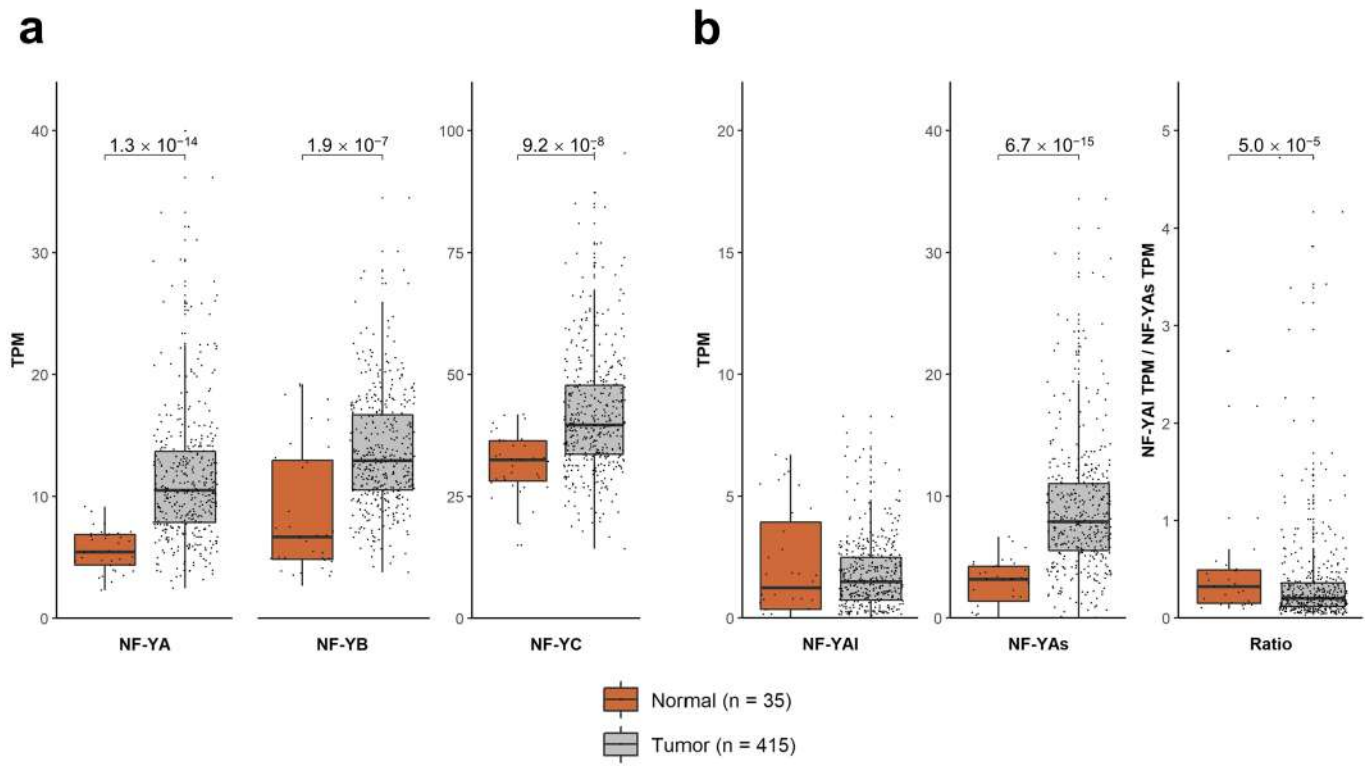


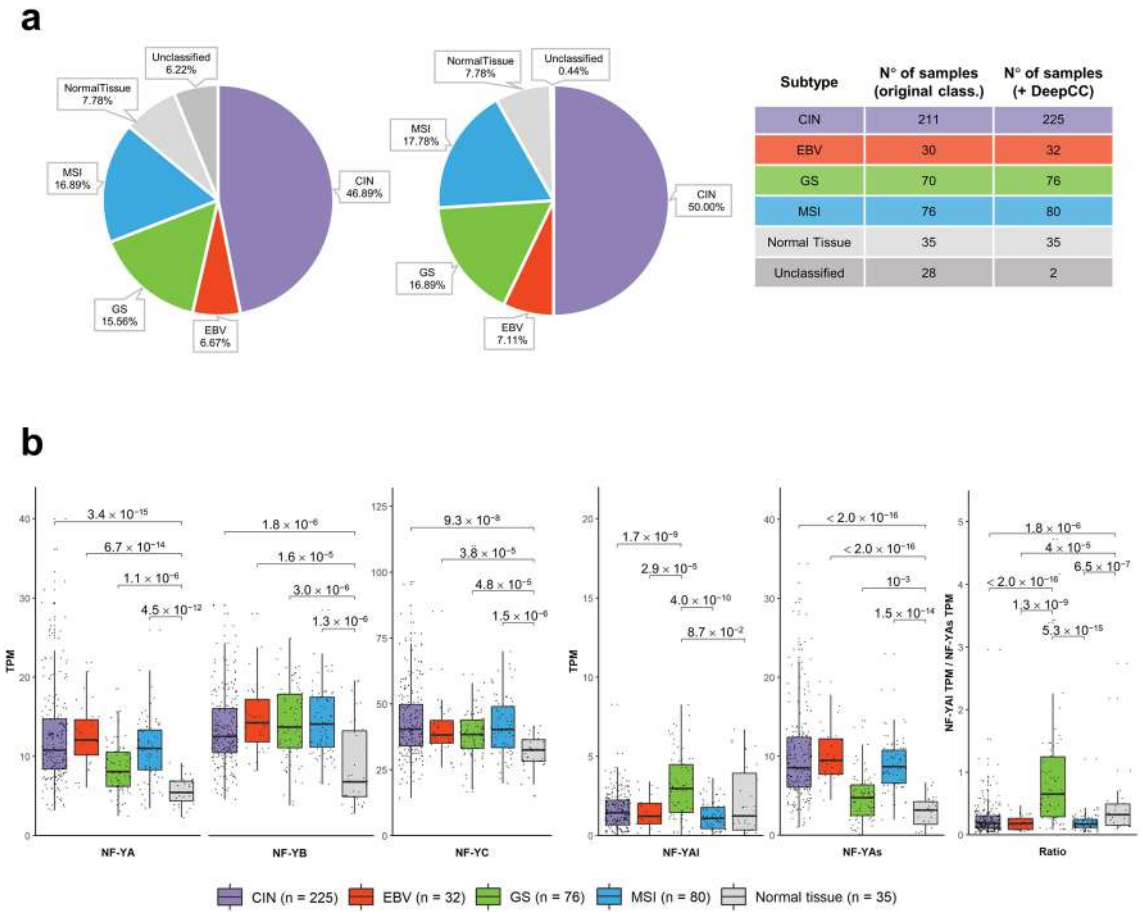
Figure 1. NF-Y subunits are overexpressed in STAD. (**a**) Box plots of expression levels of the three NF-Y subunits at gene level in the TCGA-STAD dataset, measured in TPMs. (**b**) Same as (**a**) with the analysis of the NF-YA, -NF-YAs, NF-YAl-isoforms levels as well as the NF-YAl/NF-YAs ratio. p values are calculated using the Wilcoxon rank-sum test.

STAD differentially expressed genes—DEG—have CCAAT in promoters. To gain insight on the gene expression programs altered in STAD, we compared RNA-seq data of STAD tumors to those of the respective normal samples, using a $|\log_2 FC| > 0.5$, FDR < 0.01 threshold. The lists of DEG are in Supplementary Table S1. We analyzed the promoters (−450 to +50 from the TSS) of overexpressed genes with the Pscan software, which pinpoints enriched TFs matrices[36]. The NF-Y matrix is absent, and E2Fs and SP/KLFs are at the top of the list of upregulated genes (Fig. S2a, Left Panel). As for downregulated genes (Fig. S2a, Right Panel), CCAAT is absent, and Zn Fingers TFs are enriched. Thereafter, we used KOBAS to identify Gene Ontology terms in DEG: in upregulated genes, nuclear terms—*nucleolus*, *nuclear chromatin*, *cell division*, *DNA replication*—predominate; different terms are also present in downregulated genes (Fig. S2b).

With the same thresholds, we then performed analysis of RNA-seq of the individual TCGA subtypes. Venn diagrams of the overlaps are shown in Fig. 3a and the lists of genes are in Supplementary Table S2. As for subtype-specific TFBS, distinct matrices are enriched in the four subtypes (Fig. S3a): SP1/2 in CIN, ETS-family in EBV, Zn fingers TFs in GS and MSI (EGR1/2/3, Sp2/4). We analyzed Gene Ontology terms of DEG: Fig. S3b shows specific gene signatures for individual subtypes: in CIN, *cellular protein metabolism, spermatogenesis*; in EBV, *viral process*, *T cell signaling*; in GS, *extracellular matrix*, *cell adhesion*; in MSI, *nucleolus*. Analysis of the common set of 898 genes upregulated in all subtypes have NF-Y at the top of the enriched matrices, and features described in global DEG, such as *extracellular matrix*, *cell division*, *DNA replication*, with the addition of *extracellular matrix* terms (Fig. 3b). Overall, we conclude that CCAAT is the primary site only in promoters of commonly upregulated genes, but it is absent in those specific to each TCGA subtype.

Clinical outcome of NF-Y overexpression in STAD according to the TCGA subtypes. We stratified the progression free interval—PFI—of STAD patients according to High, Intermediate, Low levels of NF-Y subunits expression. In addition, we considered the ratios of NF-YAl/NF-YAs, because this parameter was more informative than the overall levels of the two isoforms to predict patient outcomes in breast, lung and HNSCC cancers[25–27,29]. No correlation is scored according to the different levels of NF-YA and of the HFD subunits (Fig. S4), nor to the ones of NF-YAl and NF-YAs isoforms (Fig. 4a, Upper Panels). As for the NF-YAl/NF-YAs ratios, instead, we did find a robust correlation with worse prognosis (p value 0.0099) (Fig. 4a, Lower Panel). We then focused on PFIs of NF-YA ratios stratified according to the single subtypes: a correlation with poor prognosis was scored in CIN and EBV (Fig. 4b), but not in GS and MSI (Fig. S5). In summary, a higher NF-YAl/NF-YAs ratio does have relevant clinical implication in STAD, globally and in specific TCGA subtypes.

Expression of NF-Y according to the ACRG classification. A second STAD molecular classification was proposed by ACRG. This was originally based on profiling analysis, and thereafter applied to the TCGA



Subtype	N° of samples (original class.)	N° of samples (+ DeepCC)
CIN	211	225
EBV	30	32
GS	70	76
MSI	76	80
Normal Tissue	35	35
Unclassified	28	2

Figure 2. NF-Y levels in four subtypes of STAD (TCGA classification). (**a**) Classification of all TCGA samples for which there are RNA-seq data in the four subtypes. In the left pie the currently available samples (Unclassified 6.22%), in the right one our complete classification. The numbers for each subtype are in the scheme to the right. (**b**) Box plots of expression levels of the three NF-Y subunits at gene level (Left Panels), as well as of NF-YAl, NF-YAs and the NF-YAl/NF-YAs ratio (Right Panels), in the four subtypes of the TCGA STAD dataset as well as in normal tissues, measured in TPMs. p values are calculated using the Wilcoxon rank-sum test.

RNA-seq database on a partial set of 204 samples[9]. As above, we first used DeepCC and the training set to classify all TCGA tumors in the four ACRG subclasses: unclassified samples are reduced from 211 to 16 (Fig. S6a). The proportion of the four classes are relatively well maintained, with EMT being the most abundant (122 samples). A direct comparison between the TCGA and ACRG classifications is shown in Fig. 5a: most GS samples are found in EMT, which also harbors a sizeable number of CIN; MSI samples are largely shared, while EBV are partitioned among the four subclasses. With the extended ACRG dataset on hand, we evaluated the levels of NF-Y subunits and isoforms: Fig. 5b (Left Panels) shows similar levels of NF-YA and NF-YC, lower levels of NF-YB in MSS;TP53^{-} and MSS;TP53^{+}. Figure 5b (Right Panels) shows higher levels of NF-YAl, and lower of NF-YAs, in EMT samples, leading to an increased ratio of these isoforms. The presence of CIN samples in all ACRG subtypes, particularly EMT, led us to analyze NF-Y expression of CIN within ACRG subclasses: globally, the levels are similar (Fig. 5c, Left Panels), with those within the EMT group having distinctly higher levels of NF-YAl, lower NF-YAs and, by consequence, higher ratios (Fig. 5c, Right Panels). Note that analysis of STAD cell lines shows that most EMT lines, classified as such by Lee et al. [34], indeed express the lowest levels of NF-YAs and highest of NF-YAl (Fig. S1a). We conclude that the EMT subclass of ACRG includes GS, as well as a portion of tumors catalogued as CIN, having a high ratio between NF-YAl and NF-YAs.

Clinical outcome of NF-Y expression according to the ACRG subtypes. Next, we evaluated the clinical outcome of patients according to the ACRG classification. Stratification according to NF-YAl/NF-YAs ratios indicate no clinical relevance in MSI, MSS;TP53^{-} and MSS;TP53^{+}, but worst prognosis with high and intermediate levels in EMT (Fig. 6a). This is in agreement with the CIN data (Fig. 4b) and with the notion of a cluster of CIN tumors with high NF-YAl/NF-YAs ratios being inserted in the EMT subtype of ACRG (Fig. 5c): this could be responsible for the correlation seen in EMT, but not in GS. To substantiate this point, we calculated the distribution of the NF-YAl/NF-YAs ratios in GS and EMT: Fig. 6b shows that GS has a flatter distribution, with more samples with very high ratios (35% are ≥ 1), whereas EMT has fewer samples with high ratios (25% are ≥ 1), but a larger population with ratios between 0.2 and 0.5. Thus, EMT is in part fed by the CIN samples

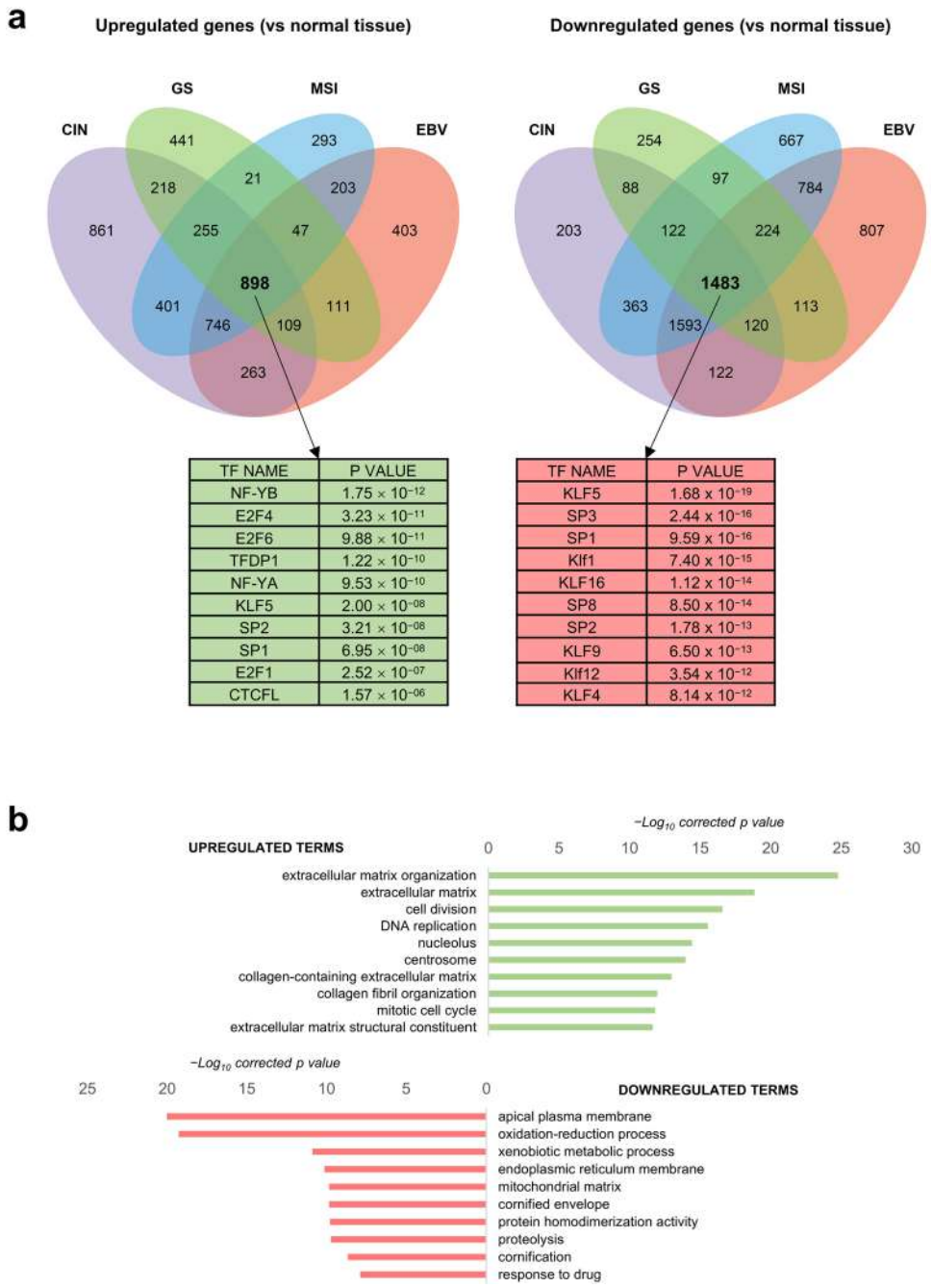


Figure 3. DEG in STAD core. (**a**) Upper Panels, Venn diagrams of DEG in the four TCGA subtypes. Left Diagram, upregulated genes, Right Diagram, downregulated genes. In the lower Panels, we show the Pscan analysis of Transcription Factors Core. (**b**) KOBAS analysis of upregulated and downregulated GO terms in the core set of commonly deregulated genes in the four STAD subtypes.

that show high ratios (Fig S6b). Note that EBV and MSI have essentially no samples above a 0.35 ratio. Thereafter, we stratified EMT samples according to low and intermediate/high ratios: the curve of the latter significantly correlates to a worst outcome (p value 0.012) (Fig. 6c, Left Panel). In addition, we reasoned that the overall levels of NF-YAs might also be impactful: stratification according to NF-YAs levels indeed indicates a protective effect of this isoform (Fig. 6c, Right Panel). Finally, analysis on the levels of HFD subunits in ACRG subtypes yielded negative results (Fig. S7), except for NF-YB, whose high levels are protective in MSS;TP53^{+} (Fig. 6d). Altogether, these data reinforce the role of the relative levels of the two NF-YA isoforms in the outcome of EMT, as well as pointing at a novel role of NF-YB in the MSS;TP53^{+} subtype.

NF-YAl is predominant in Claudin^{low} STAD tumors. We previously reported on association of high NF-YAl levels in a subclass of BRCA showing low levels of Claudin 3/4/7 expression[25], a cluster associated with

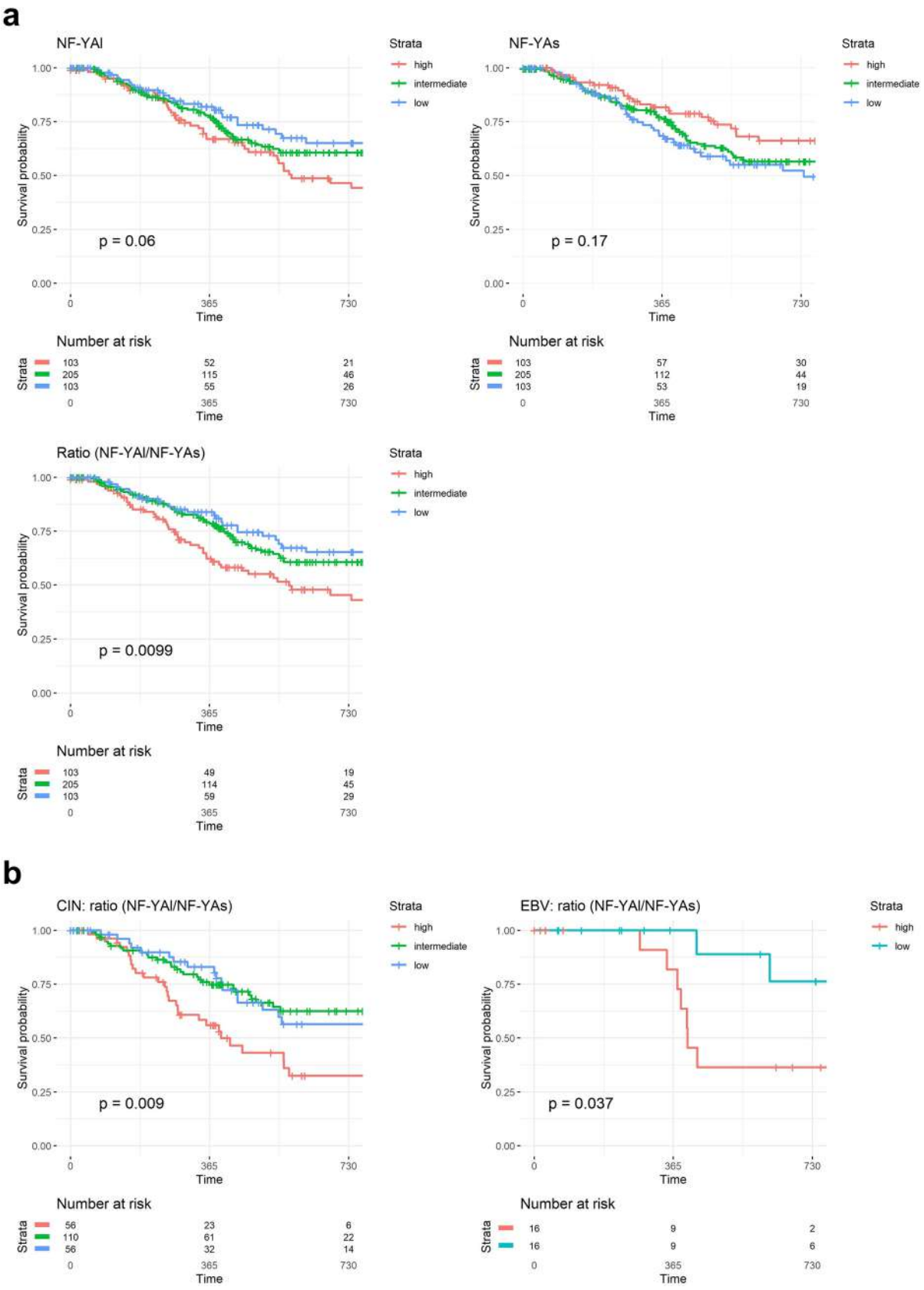


Figure 4. Clinical outcome of NF-Y expression according to the TCGA classification. (**a**) Upper Panels, Progression-Free-Interval-PFI-curves of survival probability of STAD tumors with stratification according to quartiles of NF-YA long and short isoforms (Intermediate, High and Low). PFI curves of NF-YAl/NF-YAs ratios are shown in the Lower Panel. We stratified all available tumors in the three groups according to the TPM levels: Low = first quartile, high = fourth quartile and intermediate for values included in the two middle quartiles. p values are calculated using the log-rank test. (**b**) PFI curves of the NF-YAl/NF-YAs ratios in the CIN (Left Panel) and EBV tumors (Right Panel). In this latter case, we stratified samples in two bins, High (third and fourth quartiles) and Low (first and second quartiles).

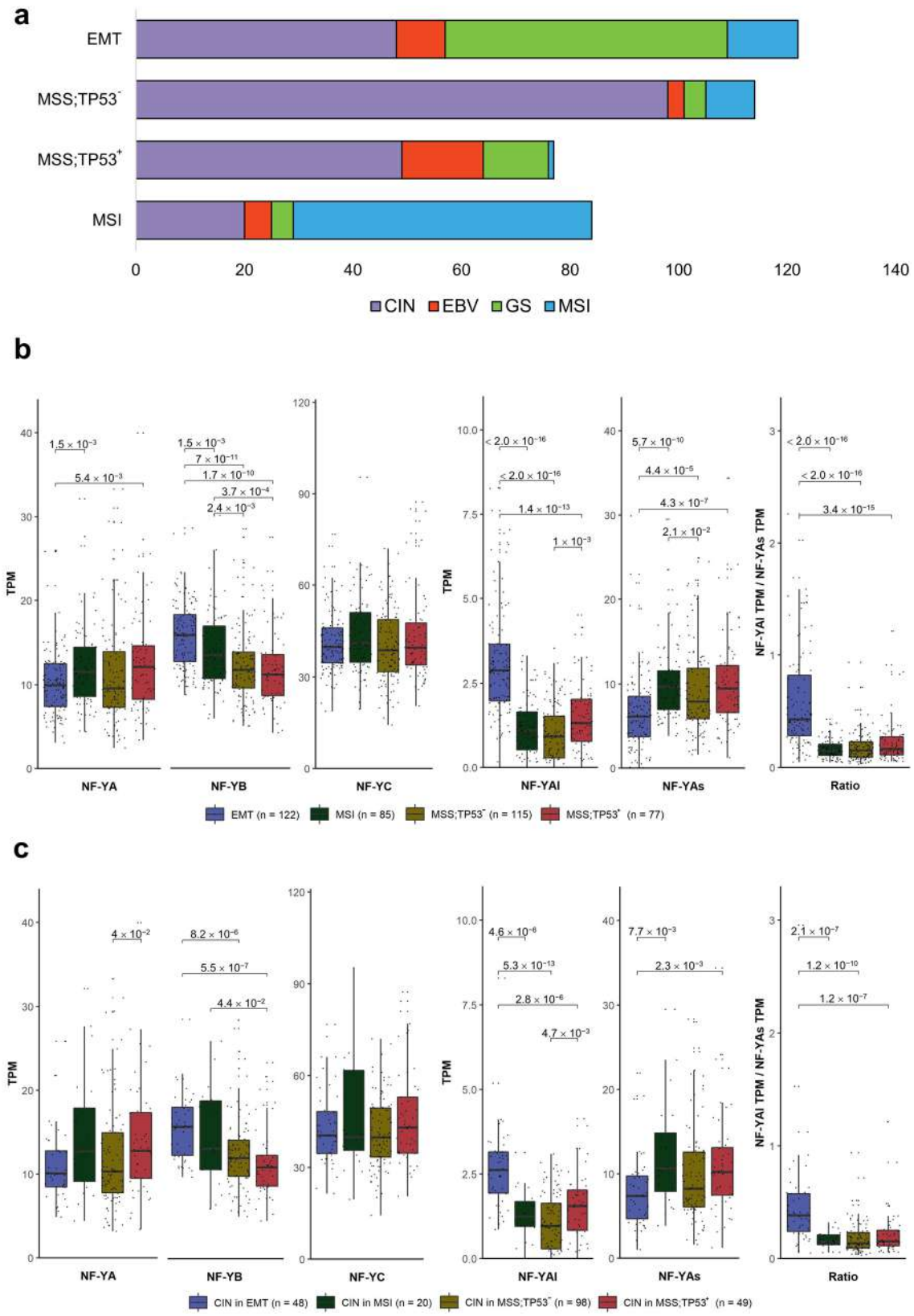


Figure 5. NF-Y expression in the ACRG subtypes. (**a**) Partitioning of TCGA subtypes in the four ACRG subtypes, according to the classification of TCGA RNA-seq dataset. (**b**) Box plot analysis of NF-Y subunits (Left Panel) and NF-YA isoforms (Right Panel) in the four ACRG subtypes. (**c**) Same as (**b**), except that expression was computed in the TCGA CIN samples partitioned in the ACRG subtypes. p values are calculated using the Wilcoxon rank-sum test.

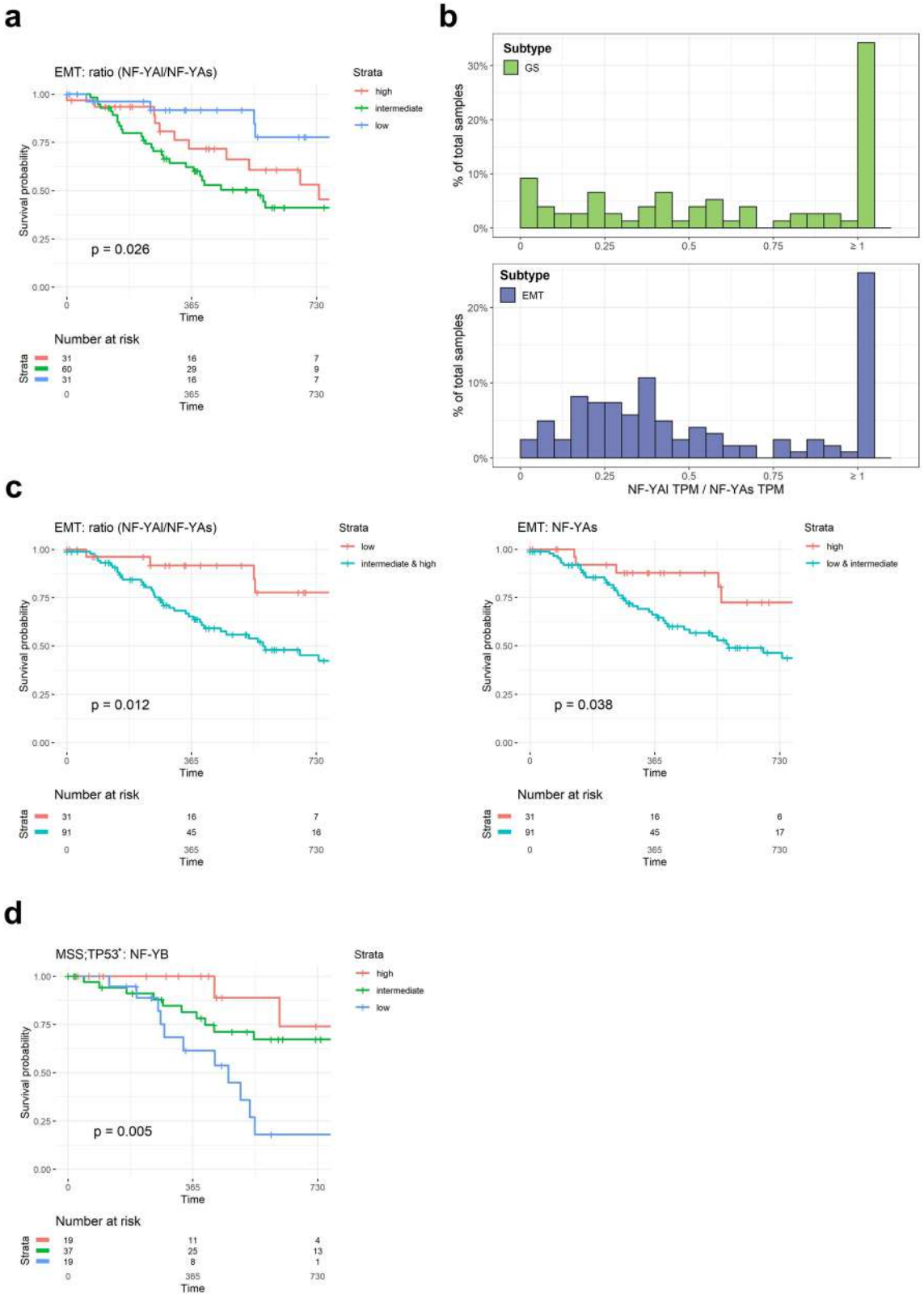


Figure 6. Clinical outcome of NF-Y expression according to the ACRG classification. (**a**) Progression-Free-Interval curves of survival probability of EMT tumors with stratification according to quartiles of NF-YAl/NF-YAs ratio (Intermediate, High and Low). We stratified all available tumors in the three groups according to the ratio levels: Low = first quartile, High = fourth quartile, and Intermediate for values included in the two middle quartiles. p values are calculated using the log-rank test. (**b**) Distribution of NF-YAl/NF-YAs ratios in the TCGA GS (Upper Panel) and ACRG EMT (Lower Panel) samples. The width of each bin was set to 0.05, and ratios equal to or greater than one were included in the last bin. (**c**) PFI curves in EMT tumors of NF-YAl/NF-YAs (Left Panel) and NF-YAs (Right Panel) stratified in Low and Intermediate/High values or High and Intermediate/Low, respectively. (**d**) Same as (**a**), except that the levels of NF-YB expression were correlated to prognosis in MSS;TP53^{+} tumors.

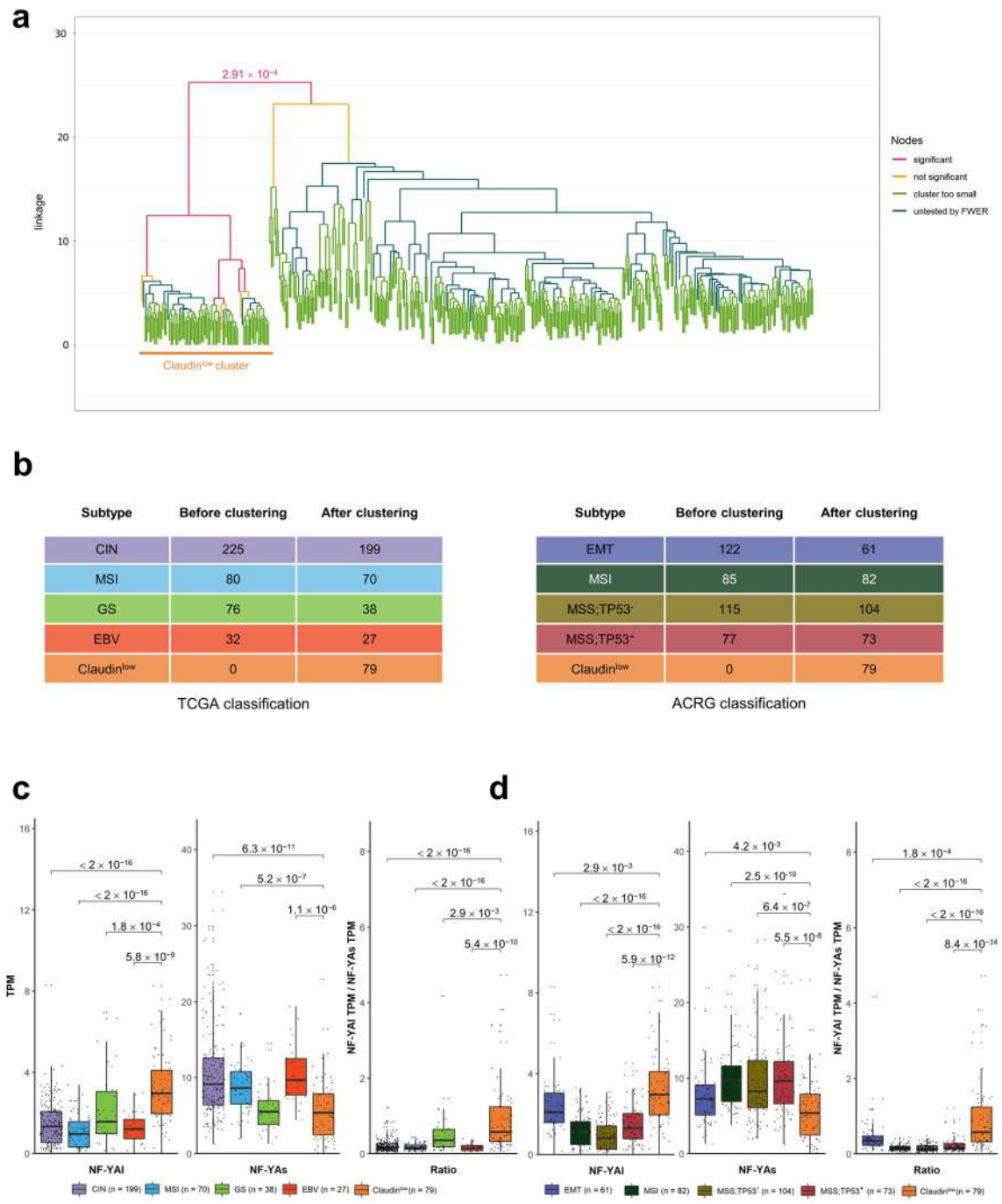


Figure 7. NF-YAl is predominant in Claudin^{low} STAD tumors. (**a**) Hierarchical clustering of all TCGA tumor samples according to the 24 genes signature described in Ref. [37]. (**b**) Sample repartition in the TCGA (Left Panel) and ACRG (Right Panel) classifications, including the new Claudin^{low} group identified in (**a**), before and after the clustering procedure. (**c**) Box plots of expression levels of NF-YAl (Left Panel), NF-YAs (Central Panel), and NF-YAl/NF-YAs ratio (Right Panel) in the TCGA classification, including the Claudin^{low} group identified in (**a**). (**d**) Same as (**c**), except that the NF-YA isoforms expression levels are calculated for the ACRG classification subtypes. p values are calculated using the Wilcoxon rank-sum test.

EMT features and poor prognosis. By analyzing TCGA STAD data, Nishijima et al. identified a specific group of tumors—46 samples—based on three features: epithelial to mesenchymal transition (EMT), tumor-initiating cells (TIC) and a Claudin^{low} phenotype[37]; this group was separated from CIN and GS (TCGA classification) and EMT (ACRG classification). Importantly, these Authors derived a 24-strong gene signature predictive of this subclass: we used it to conduct a hierarchical clustering of the entire TCGA dataset; Fig. 7a shows the dendrogram with the identification of 79 samples with these gene expression features; this cohort is clearly separated by the other tumors based on a strong statistical bias (p value: 2.91×10^{-4}). We first checked how this signature fea-

tures each subtypes: Fig. S8 shows below zero median Z scores of CIN, EBV and MSI (TCGA), MSI, MSS;TP53$^-$ and MSS;TP53$^+$ (ACRG); instead, good concordance is scored within the GS and EMT groups. Because of the presence of low levels of epithelial Claudins, we will refer to this group as Claudin^{low}. Next, we positioned this group within the other TCGA and ACRG subtypes (Fig. 7b): most tumors of the Claudin^{low} cluster are from the GS and CIN (TCGA) and EMT (ACRG) subtypes. In essence, the Claudin^{low} group could be classified as new within TCGA, while being essentially a subclass of the EMT ACRG subtype. Overall, these data confirm the existence of the subgroup proposed by Nishijima et al., further expanding it to 79 TCGA samples, with robust statistical significance.

Next, we evaluated the expression of NF-YA isoforms and their relative ratio including the Claudin^{low} group. Figure 7c,d show the results according to the TCGA and ACRG subtypes, respectively: NF-YAl is mostly present in the Claudin^{low} class, with far lower levels in the remaining samples of the ACRG EMT subtype. On the contrary, NF-YAs is lowest in Claudin^{low}, and higher in all other ACRG and TCGA subtypes, with the exception of GS. As a consequence, the NF-YAl/NF-YAs ratio is significantly increased (lowest p values: 10^{-16}) mostly in the Claudin^{low} group. These data indicate that NF-YAl is mostly associated to a discrete number of STAD samples with EMT and Claudin^{low} features.

To verify the overlap between the Claudin^{low} and NF-YAl^{high} (and NF-YAs^{low}) subsets, we stratified the clinical outcome of Claudin^{low} tumors according to NF-YA isoforms expression (High, Intermediate, Low): no further worsening of prognosis in PFI curves is scored according to the different levels of NF-YA isoforms (Fig. S9, Upper Panels), nor NF-YAl/NF-YAs ratio (Fig. S9, Lower Panel). We conclude that there is a large overlap between the subset classified as Claudin^{low} and NF-YAl^{high} tumors.

CCAAT box is enriched in upregulated pathways of Claudin^{low} samples. To further investigate the Claudin^{low} cluster, we compared pathways in Claudin^{low} and EMT versus normal samples. The analysis of DEG in EMT shows absence of CCAAT in promoters (Fig. S10a). Across EMT upregulated pathways, we did find mesenchymal terms such as *extracellular matrix*, *heart development*, *mesenchyme development* (Fig. S10b). Within the TF motifs enriched in the promoters of genes of each single category, we observed significant enrichment of the NF-Y motif in *cell-cycle* terms, as expected, and in *mesenchyme development* and *pattern specification process*. In downregulated pathways, we observed different *metabolism* terms, also expected (Fig. S10c). The same analysis performed on Claudin^{low} samples did not yield NF-Y motifs as enriched in deregulated genes, but rather MAZ, E2F6 and KLFs motifs (Fig. 8a); these TFs were confirmed by analyzing ChIP-seq data from the ChIP-Atlas database[38] (Supplementary Table S3). Among upregulated pathways we found *extracellular matrix* and *mesenchyme development* terms (*heart development*, *skeletal system*, and *pattern specification process*). As above, the CCAAT box was enriched in terms related to mesenchyme (Fig. 8b). Various metabolic processes populated the downregulated pathways (fatty acid and lipid metabolic process), expectedly regulated by NF-Y and with CCAAT motifs (Fig. S11).

Discussion

Because of its histone-like structure[17], positioning within promoters[16], synergistic connections with many other TFs and interactions with coactivators, NF-Y is believed to play a pioneering role in "opening" promoter structures and correct positioning of RNA Pol II[39]. Specifically, NF-Y is important for genes required for cell proliferation[20]. We describe here an investigation on NF-Y subunits levels in gastric cancer. We report the presence of CCAAT in commonly overexpressed genes and overexpression of NF-YA isoforms, as well as a prognostic value of their relative levels. We also report on overexpression of the HFD subunits, and clinical significance of NF-YB.

CCAAT boxes have been routinely found in promoters of genes overexpressed in cancer, first in large microarrays profiling[15] and more recently in RNA-seq datasets. Our analysis of TCGA identified CCAAT in overexpressed genes, typically with E2Fs sites, in line with the pro-growth role of these TFs. Specifically, two schemes are starting to emerge. In the first, CCAAT is enriched globally, and indeed at the top of the TFBS list, when all upregulated genes are computed: it is the case of lung tumors[26,27]; in the second, the enrichment is found either in specific subtypes—iCluster 3 in HCC[28]—or only in DEG shared by all subtypes, as in BRCA[25] and STAD, as shown here. In global STAD DEG, TFBS in promoters of upregulated genes contain the familiar E2Fs motifs, along with Zn Finger TFBS (SPs/KLFs), but CCAAT is absent. As in BRCA, however, it comes out first when considering the core group of upregulated genes shared in all STAD subtypes. We also find that CCAAT is absent in promoters of genes downregulated in STAD, as for all other types of cancer examined so far. This further reinstates that this element is not a "general" signal enriched in promoters per se, but rather a core logo driving expression of genes associated to growth, not necessarily related to transcriptional features that are cancer- or subtype-specific.

The HFD subunits are overexpressed in STAD, unlike in lung and breast tumors. We recently reported a similar scenario in HCC, in which high levels of these subunits correlate with worst prognosis in a specific subtype, iCluster1. In STAD, global or subtype-specific PFI curves are globally superimposable based on NF-YB or NF-YC expression, with one notable exception: the MSS;TP53$^+$ ACRG subtype, in which high NF-YB levels correlate with a better prognosis. As for HCC, the fraction of p53 wt tumors in STAD is much higher—51%—than in other epithelial cancers (lung for example), in which the vast majority are p53 mutated, rendering comparisons with wt p53 samples essentially impossible. Note that the protective role of NF-YB in STAD is opposite to what we reported in HCC iCluster1 tumors, generally associated to wt p53 status: although direct NF-Y/p53 interactions have been reported in several studies[20], the reasons for association of NF-YB levels to such genetic background is unclear. Nevertheless, a role of HFD subunits in cancer progression is starting to emerge; in this respect,

a

Claudin^{low} – Normal tissue

UPREGULATED		DOWNREGULATED	
TF NAME	P VALUE	TF NAME	P VALUE
MAZ	1.63×10^{-29}	KLF5	3.86×10^{-60}
ZNF148	1.01×10^{-28}	KLF15	2.50×10^{-59}
KLF15	1.15×10^{-27}	SP3	1.46×10^{-55}
E2F6	1.23×10^{-27}	KLF2	9.30×10^{-55}
TFDP1	6.63×10^{-26}	SP1	3.01×10^{-54}
SP9	9.04×10^{-26}	KLF6	5.61×10^{-53}
KLF5	1.90×10^{-25}	KLF11	1.27×10^{-52}
SP3	4.36×10^{-23}	SP9	1.54×10^{-50}
ZBTB14	2.27×10^{-22}	KLF16	5.07×10^{-50}
EGR1	7.17×10^{-22}	KLF10	2.89×10^{-49}

b

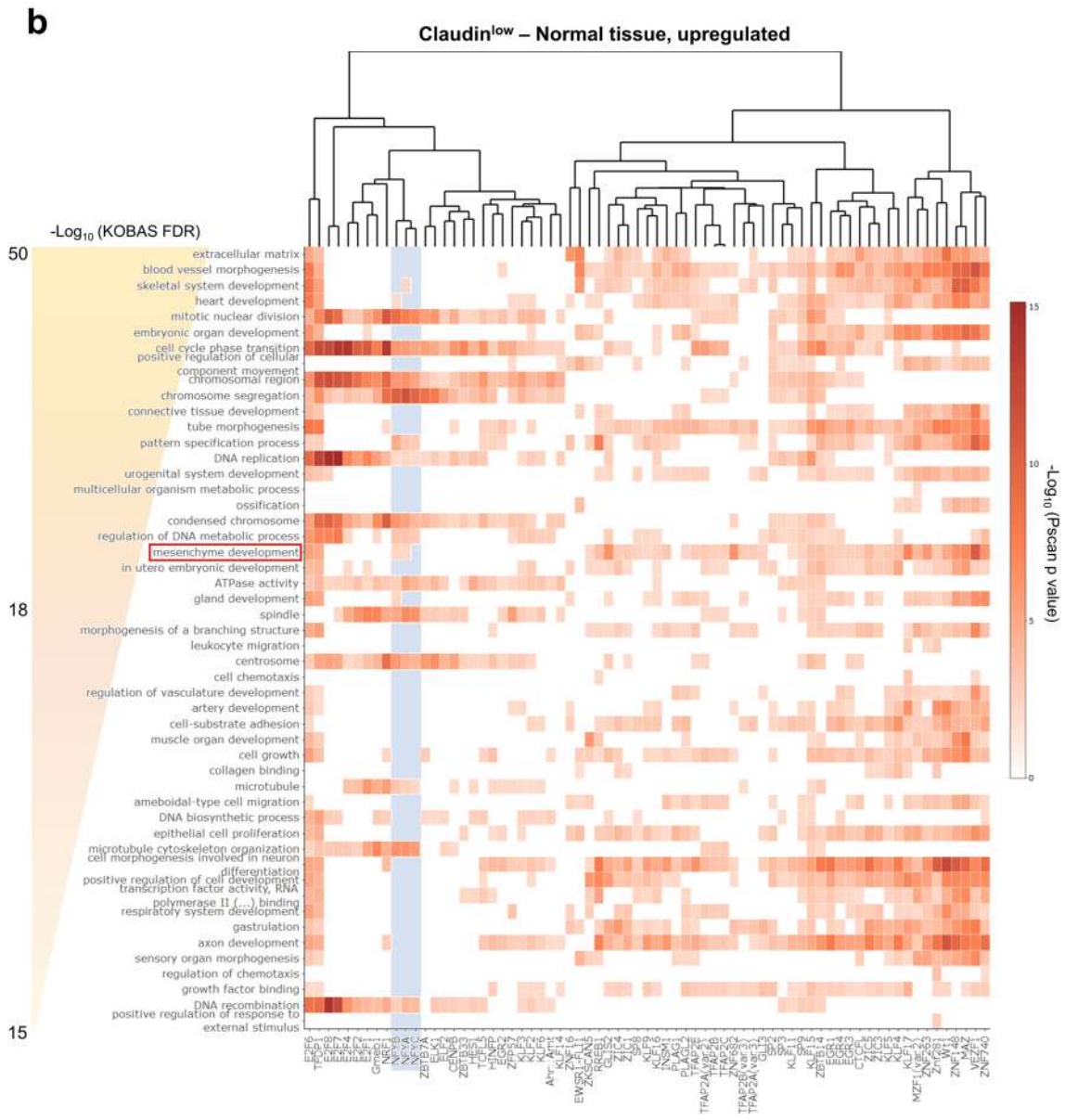


Figure 8. DEG in Claudin^{low} subtype. (**a**) List of enriched TFBS in promoters of upregulated and downregulated genes in the Claudin^{low} subtype, compared to normal tissue, as predicted by Pscan. (**b**) Heatmap showing the p values, as computed by Pscan, associated to the enrichment in multiple TFBS motifs (columns) of the top 50 most represented GO terms (rows) among upregulated genes in Claudin^{low} samples, as emerged from the KOBAS analysis. The light blue area highlights NF-Y subunits binding site motifs (CCAAT box). Only terms with less than 500 background genes and TFBS with a significant enrichment in more than 10 terms ($p < 0.01$) were included in the heatmap.

measurement of protein levels in tumors deserve a close look in the future: in BRCA cell lines, for example, the NF-YB protein seems to be more variable than one could anticipate from mRNA levels[25].

Overexpression of NF-YA mRNA is as obvious in STAD as in the tumors previously analyzed. Note that analysis of 22 cancer specimens confirms that higher expression is also found at the protein level[30]. In the same study, high levels of NF-YA and Cyclin E in TCGA STAD samples were associated to worsening of patients' prognosis: yet, we do not find here a prognostic value of global levels of NF-YA. In another study, NF-YA high expression correlated with prognosis in a separate set of tumor samples analyzed by microarray profilings[31], but only in the Diffuse (DF), not in the Intestinal (IT) subtype (Lauren classification). We add a novel and relevant twist, in that isoform ratios—rather than global levels—are clinically important within subclasses of STAD.

The two major NF-YA splicing isoforms differ in the Gln-rich trans-activation domain (TAD): NF-YAl has 28/29 extra amino acids coded by exon 3, predicted to impart different activation potential, as reported in mESCs and myoblasts[40,41]. In addition, a shorter isoform—NF-YAx—lacking sequences of exon-3 and exon-5 was recently found overexpressed in Neuroblastomas[42]. As in the other epithelial cancers, we find that NF-YAs predominates, but higher expression of NF-YAl, alone or coupled to lower levels of NF-YAs, is clinically relevant. The TCGA GS subtype is enriched in DF samples[8], which is indeed in line with the data reported by Cao et al.[30]. GS tumors are characterized by earlier onset and expression of "cell adhesion" signatures. The NF-YAl/NF-Ys ratio is shifted in GS and the same pattern is observed stratifying tumors according to the ACRG classification: higher NF-YAl/NF-YAs ratios are found in EMT tumors. The relatedness of these subtypes in the two classifications was commented before[11–13]: indeed analysis of GO terms and pathways of DEG in these subtypes are in agreement with a mesenchymal phenotype. The ACRG EMT has 48 samples catalogued as CIN by TCGA: interestingly, the PFI of CIN patients indicates a worst prognosis following the NF-YAl/NF-YAs ratios.

Our comparative analysis of the whole set of TCGA tumors suggest clinical relevance for NF-YB and NF-YA isoforms in subgroups of the ACRG classification. Specifically, NF-YA-wise, the ACRG EMT group is more revealing than the TCGA GS, most likely because of the inclusion of CIN tumors with EMT-like profilings. While in the EMT group the role of NF-YA ratios is clinically visible, in the TCGA GS it is not. One possible explanation is the lower dispersion of ratios and lower number of samples in this latter group, making comparison of quartiles difficult. Incidentally, this also allowed to score a protective role of NF-YAs, completely missed by adhering to the TCGA classification. Another feature emerging in the ACRG classification is the protective role of high NF-YB levels, as discussed above. These differences might reflect the fact that RNA profilings are the basis of ACRG, while TCGA factored in other genetic and epigenetic features of STAD.

The parallel of the present data with what we found in breast carcinoma is noteworthy. NF-YAs is also predominant in BRCA, except in the Claudin^{low} subset of Basal-like tumors, that have higher levels of NF-YAl. This is associated to a shift in DEG in these tumors, from signatures dominated by proliferative terms in NF-YAs^{high} tumors, toward activation of EMT signatures. In turn, this is clinically associated to an aggressive, metastatic, drug-resistant behavior. As in BRCA, the NF-YAl/NF-YAs ratio is clinically informative in STAD, but in this case the protective role of NF-YAs^{high} in the EMT subtype is novel. Nishijima et al. showed that overall survival curves and Hazard ratios of the 46 Claudin^{low} patients are indeed worse with respect to other subtypes, dramatically so within the ACRG-classified patients. This suggests that the Claudin^{low} partitioning is particularly significant with ACRG. We extended this group to 79 TCGA tumors by using the signature described: our results confirm and extend the scenario proposed by these Authors, particularly within the ACRG classification, which better partitions the protective role of NF-YAs from the detrimental role of NF-YAl in the Claudin^{low} group. Furthermore, it appears manifest the overlap of tumors with Claudin^{low} and NF-YAl^{high} features.

In general, these data invite further analysis in epithelial cancers to identify (i) Claudin^{low} signatures in other types of epithelial cancers, and (ii) a threshold of NF-YA isoforms ratios, rather than overall levels, possibly responsible for shifting DEG away from proliferative, cell cycle genes toward mesenchymal ones.

Materials and methods

RNA-seq datasets. As of December 2020, there were RNA-seq data on 415 STAD primary tumors in TCGA and 35 non-tumor tissues. We downloaded the corresponding RSEM scaled count data from the http://firebrowse.org/ web page. The last published classification of STAD samples in the four molecular subtypes made by TCGA referred to 387 of the 415 tumors, and we retrieved it from the https://www.cbioportal.org/ web page[43,44]; a different classification was proposed by ACRG on 204 TCGA tumors[9]. All the experiments involving human data in these public datasets adhered to relevant ethical guidelines. The DeepCC tool[35] was used to classify RNA-seq dataset of all tumors in TCGA, according to the TCGA and ACRG classification, using as a training set the tumors already classified by TCGA and ACRG, respectively.

We retrieved the FASTQ files associated to the 37 CCLE stomach cell lines (accession code: PRJNA523380)[45], as well as the 29 cell lines collected by Lee et al. (accession code: PRJNA327709)[34], using the SRA Explorer website (https://sra-explorer.info/). From the FASTQ files, we calculated mRNA expression with RSEM-1.3.3.

Gene expression analysis. Differential gene expression analysis of RNA-seq data was performed using R package *DESeq2*[46]. The Tumor versus Normal expression fold change (FC) denotes upregulation or downregulation according to the FC value. Log_2FC, and the corresponding false discovery rate (FDR), were reported by the R package. FDR < 0.01 and $|log_2FC| > 0.5$ were set as inclusion criteria for DEG selection in tumor/subtype versus normal samples.

Gene ontology, pathway enrichment and transcription factor binding site analysis. We used KOBAS 3.0 (http://kobas.cbi.pku.edu.cn/anno_iden.php) for pathway enrichment analysis using the ENTREZ gene IDs. The TFBS and de novo motif analyses were performed using the Pscan software[36], while ChIP-seq

experiments enrichment analyses were conducted with ChIP-Atlas[38]. To obtain TFBS enrichment heatmaps, input genes collections of the top GO terms from KOBAS analysis, sorted by FDR, were analyzed individually with Pscan. Only GO terms with less than 500 background genes were included, and TFBS motif enriched (Pscan p value < 0.01) in less than 10 terms were filtered out.

Analysis of clinical data. We retrieved clinical data related to the TCGA STAD samples and progression free interval—PFI—time records of patients, respectively, from the https://www.cbioportal.org/ and the http://xena.ucsc.edu/ web pages[43,44,47]. We stratified all the tumors for which PFI records were available according to NF-Y subunits expression at gene level, NF-YA isoforms expression, and NF-YAl/NF-YAs ratio, into three groups (Low = first quartile, Intermediate = second and third quartiles, High = fourth quartile). Survival analysis was performed according to the Kaplan–Meier analysis and log-rank test[48].

Hierarchical clustering and Z scores computation. TCGA samples RSEM scaled count data were converted into TPM, log_2-transformed, and median centered; we then performed a hierarchical clustering of the samples with the R package *SigClust2* (version 1.2.4) with "average" linkage and "euclidean" metric options, while the alpha parameter was set to 0.05. Daughter nodes were tested if significance was achieved at the corresponding parent node, according to the built-in FWER controlling procedure. We obtained Z scores from log_2-transformed expression data for each gene of the Claudin^{low} signature, and a median Z score for each sample was computed across the genes of the signature.

Statistical analysis. Analyses were performed in the R programming environment (version 4.0.3), with the *ggplot2*, *ggpubr*, *survival*, *survminer, tidyverse* packages. Single comparisons between two groups were performed with the Wilcoxon rank-sum test.

Received: 5 July 2021; Accepted: 22 November 2021
Published online: 09 December 2021

References

1. Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2012. *CA Cancer J. Clin.* **62**, 10–29 (2012).
2. Laurén, P. The two histological main types of gastric carcinoma: Diffuse and so-called intestinal-type carcinoma. *Acta Pathol. Microbiol. Scand.* **64**, 31–49 (1965).
3. Hartgrink, H. H., Jansen, E. P. M., van Grieken, N. C. T. & van de Velde, C. J. H. Gastric cancer. *Lancet* **374**, 477–490 (2009).
4. Kim, B. *et al.* Expression profiling and subtype-specific expression of stomach cancer. *Cancer Res.* **63**, 8248–8255 (2003).
5. Jinawath, N. *et al.* Comparison of gene-expression profiles between diffuse- and intestinal-type gastric cancers using a genome-wide cDNA microarray. *Oncogene* **23**, 6830–6844 (2004).
6. Lee, Y.-S. *et al.* Genomic profile analysis of diffuse-type gastric cancers. *Genome Biol.* **15**, R55 (2014).
7. Tanabe, S., Aoyagi, K., Yokozaki, H. & Sasaki, H. Gene expression signatures for identifying diffuse-type gastric cancer associated with epithelial-mesenchymal transition. *Int. J. Oncol.* **44**, 1955–1970 (2014).
8. Bass, A. J. *et al.* Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
9. Cristescu, R. *et al.* Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat. Med.* **21**, 449–456 (2015).
10. Yu, Y. A new molecular classification of gastric cancer proposed by Asian Cancer Research Group (ACRG). *Transl. Gastrointest. Cancer* **5**, 557–557 (2016).
11. Chia, N.-Y. & Tan, P. Molecular classification of gastric cancer. *Ann. Oncol.* **27**, 763–769 (2016).
12. Min, L. *et al.* Integrated analysis identifies molecular signatures and specific prognostic factors for different gastric cancer subtypes. *Transl. Oncol.* **10**, 99–107 (2017).
13. Battaglin, F., Naseem, M., Puccini, A. & Lenz, H.-J. Molecular biomarkers in gastro-esophageal cancer: Recent developments, current trends and future directions. *Cancer Cell Int.* **18**, 99–99 (2018).
14. Levine, M., Cattoglio, C. & Tjian, R. Looping back to leap forward: Transcription enters a new era. *Cell* **157**, 13–25 (2014).
15. Goodarzi, H., Elemento, O. & Tavazoie, S. Revealing global regulatory perturbations across human cancers. *Mol. Cell* **36**, 900–911 (2009).
16. Dolfini, D., Zambelli, F., Pavesi, G. & Mantovani, R. A perspective of promoter architecture from the CCAAT box. *Cell Cycle* **8**, 4127–4137 (2009).
17. Nardini, M. *et al.* Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. *Cell* **152**, 132–143 (2013).
18. Li, X. Y., Hooft van Huijsduijnen, R., Mantovani, R., Benoist, C. & Mathis, D. Intron-exon organization of the NF-Y genes. Tissue-specific splicing modifies an activation domain. *J. Biol. Chem.* **267**, 8984–8990 (1992).
19. Ceribelli, M., Benatti, P., Imbriano, C. & Mantovani, R. NF-YC complexity is generated by dual promoters and alternative splicing. *J. Biol. Chem.* **284**, 34189–34200 (2009).
20. Gurtner, A., Manni, I. & Piaggio, G. NF-Y in cancer: Impact on cell transformation of a gene essential for proliferation. *Biochim. Biophys. Acta* **1860**, 604–616 (2017).
21. Benatti, P. *et al.* NF-Y activates genes of metabolic pathways altered in cancer cells. *Oncotarget* **7**, 1633–1650 (2016).
22. Mamat, S. *et al.* Transcriptional regulation of aldehyde dehydrogenase 1A1 gene by alternative spliced forms of nuclear factor Y in tumorigenic population of endometrial adenocarcinoma. *Genes Cancer* **2**, 979–984 (2011).
23. Cicchillitti, L. *et al.* Prognostic role of NF-YA splicing isoforms and Lamin A status in low grade endometrial cancer. *Oncotarget* **8**, 7935–7945 (2017).
24. Yang, C., Zhao, X., Cui, N. & Liang, Y. Cadherins associate with distinct stem cell-related transcription factors to coordinate the maintenance of stemness in triple-negative breast cancer. *Stem Cells Int.* **2017**, 5091541–5091541 (2017).
25. Dolfini, D., Andrioletti, V. & Mantovani, R. Overexpression and alternative splicing of NF-YA in breast cancer. *Sci. Rep.* **9**, 12955 (2019).
26. Bezzecchi, E. *et al.* NF-YA overexpression in lung cancer: LUAD. *Genes* **11**, 198 (2020).
27. Bezzecchi, E., Ronzio, M., Dolfini, D. & Mantovani, R. NF-YA Overexpression in lung cancer: LUSC. *Genes (Basel)* **10**, 937 (2019).

28. Bezzecchi, E., Ronzio, M., Mantovani, R. & Dolfini, D. NF-Y overexpression in liver hepatocellular carcinoma (HCC). *Int. J. Mol. Sci.* **21**, 9157 (2020).
29. Bezzecchi, E. *et al.* NF-Y Subunits Overexpression in HNSCC. Cancers (Basel). **13**(12), 3019 (2021).
30. Cao, B. *et al.* Gene regulatory network construction identified NFYA as a diffuse subtype-specific prognostic factor in gastric cancer. *Int. J. Oncol.* **53**, 1857–1868 (2018).
31. Bie, L.-Y. *et al.* Analysis of cyclin E co-expression genes reveals nuclear transcription factor Y subunit alpha is an oncogene in gastric cancer. *Chronic Dis. Transl. Med.* **5**, 44–52 (2018).
32. Alsina, M. *et al.* Cyclin E amplification/overexpression is associated with poor prognosis in gastric cancer. *Ann. Oncol.* **26**, 438–439 (2015).
33. Ooi, A. *et al.* Gene amplification of CCNE1, CCND1, and CDK6 in gastric cancers detected by multiplex ligation-dependent probe amplification and fluorescence in situ hybridization. *Hum. Pathol.* **61**, 58–67 (2017).
34. Lee, J. *et al.* Selective cytotoxicity of the NAMPT inhibitor FK866 toward gastric cancer cells with markers of the epithelial-mesenchymal transition, due to loss of NAPRT. *Gastroenterology* **155**, 799-814.e13 (2018).
35. Gao, F. *et al.* DeepCC: A novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* **8**, 44–44 (2019).
36. Zambelli, F., Pesole, G. & Pavesi, G. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.* **37**, W247–W252 (2009).
37. Nishijima, T. F. *et al.* Molecular and clinical characterization of a claudin-low subtype of gastric cancer. *JCO Precis. Oncol.* https://doi.org/10.1200/PO.17.00047 (2017).
38. Oki, S. *et al.* ChIP-Atlas: A data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.* **19**, e46255 (2018).
39. Oldfield, A. J. *et al.* NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region. *Nat. Commun.* **10**, 3072–3072 (2019).
40. Dolfini, D., Minuzzo, M., Pavesi, G. & Mantovani, R. The short isoform of NF-YA belongs to the embryonic stem cell transcription factor circuitry. *Stem Cells* **30**, 2450–2459 (2012).
41. Libetti, D. *et al.* The switch from NF-YAl to NF-YAs isoform impairs myotubes formation. *Cells* **9**, 789 (2020).
42. Cappabianca, L. *et al.* Discovery, characterization and potential roles of a novel NF-YAx splice variant in human neuroblastoma. *J. Exp. Clin. Cancer Res.* https://doi.org/10.1186/s13046-019-1481-8 (2019).
43. Cerami, E. *et al.* The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
44. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* **6**, 11 (2013).
45. Ghandi, M. *et al.* Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019).
46. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550–550 (2014).
47. Goldman, M. J. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38**, 675–678 (2020).
48. Therneau, T. *A Package for Survival Analysis in R*, 95.

Acknowledgements
We thank P. Gandellini and N. Gnesutta for comments and critical reading of the manuscript. The authors acknowledge support from the University of Milan through the APC initiative.

Author contributions
D.D. designed the experiments. A.G., E.B. and M.R. performed and analyzed the experiments. R.M. and D.D. wrote the manuscript.

Funding
This work was supported by Ministero della Salute GR-2013-02355625 to DD.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-03027-y.

Correspondence and requests for materials should be addressed to D.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.1.2. NF-YAl expression correlates with EMT in BRCA and STAD Claudin^{low} tumors

Until now, our laboratory has investigated the balance between NF-YA isoforms in the Claudin^{low} subpopulations of breast[234] and gastric cancer. Two experimental approaches could provide a clearer understanding of the specific role of NF-YAl prevalence in this subtype: a direct confirmation of its activity via *in vivo* knockout experiments, and an inquiry on the shared characteristics of the two Claudin^{low} subsets.

Clones of two Claudin^{low} BRCA lines, SUM159PT and BT549, were ablated of NF-YA exon-3, forcing the expression of NF-YAs: these cells showed both diminished migratory capabilities in cell culture and decreased ability to generate metastases within Zebrafish embryo after xenografting. I analyzed the deregulated pathways of exon-3 edited clones and identified common downregulated terms such as *epithelial to mesenchymal transition*, *inflammatory signalling* and *hypoxia*. Conversely, *mitotic spindle* figured among the most upregulated hallmarks for both cell lines. This analysis confirmed the central role of NF-YAl in the acquisition of the mesenchymal identity of Claudin^{low} cell lines.

Then, I relabelled Normal-like and unclassified BRCA sample included in a previous PAM50-based classification[121], which included Claudin^{low}, again using the DeepCC tool. I conducted co-expression analysis of STAD and BRCA TCGA cancer samples the WGCNA tool[244], replacing gene-level NF-YA expression with NF-YAl and NF-YAs at isoform-level as well as including NF-YA$_{\text{Ratio}}$ as a regular gene. I defined modules of genes with comparable expression patterns across tumors.

NF-YAl and NF-YA$_{\text{Ratio}}$ grouped together in the same co-expressed gene module for both BRCA and STAD, and the overlap of the two modules resulted in a 158 genes-strong proposed signature. This gene set was enriched in GO terms related to muscle and cell differentiation, and PFS analysis using the median Z score for the signature yielded similar results to those obtained with NF-YA$_{\text{Ratio}}$ values. Thus, the BRCA/STAD NF-YA$_{\text{Ratio}}$ signature could serve as a clue for the mesenchymal-associated expression patterns linked to high NF-YAl expression, also bearing potential clinical implication.

Exploiting published scRNA-seq experiments for breast[124] and gastric[245] cancer, together with cell-type specific signatures, I deconvoluted the cell population composition of TCGA BRCA and STAD tumour samples using the SCDC package[246]. Bulk RNA-seq samples were firstly divided into equally sized bins according to NF-YA$_{\text{Ratio}}$: both datasets showed a slight CAFs increase in samples with high NF-YA$_{\text{Ratio}}$ values, while the predicted proportion of cancer cell remained constant. Within the subtype framework, samples with high NF-YA$_{\text{Ratio}}$ generally exhibited a more expanded population of EMT cancer cells in both breast and gastric cancers.

Searching for the RBPs potentially responsible for NF-YA alternative splicing, I studied the expression of genes included in RBPDB (RNA binding proteins database)[247] in 21 TCGA epithelial cancer cohorts. QKI and RBFOX2 were the factors that overall correlated best with high NF-YA$_{\text{Ratio}}$ levels, whereas ESRP1/2 and RBM47 were the top ranked anticorrelated RBPs. Substantiating this initial survey, I checked NF-YA isoforms expression in available RNA-seq experiments in which these RBPs were functionally inactivated. NF-YAs expression increased upon KO/KD of MBNL1/2, QKI and RBFOX2, whereas NF-YAl was upregulated in RBM47 and ESRP1-2 inactivation experiments.

CDDpress

www.nature.com/cddis

ARTICLE OPEN

Check for updates

NF-YAl drives EMT in Claudin^{low} tumours

Michela Londero[1,2], Alberto Gallo[1,2], Camilla Cattaneo[1], Anna Ghilardi[1], Mirko Ronzio[1], Luca Del Giacco[1], Roberto Mantovani[1] and Diletta Dolfini[1]✉

NF-Y is a trimeric transcription factor whose binding site -the CCAAT box- is enriched in cancer-promoting genes. The regulatory subunit, the sequence-specificity conferring NF-YA, comes in two major isoforms, NF-YA long (NF-YAl) and short (NF-YAs). Extensive expression analysis in epithelial cancers determined two features: widespread overexpression and changes in NF-YAl/NF-YAs ratios (NF-YAr) in tumours with EMT features. We performed wet and in silico experiments to explore the role of the isoforms in breast -BRCA- and gastric -STAD- cancers. We generated clones of two Claudin^{low} BRCA lines SUM159PT and BT549 ablated of exon-3, thus shifting expression from NF-YAl to NF-YAs. Edited clones show normal growth but reduced migratory capacities in vitro and ability to metastatize in vivo. Using TCGA, including upon deconvolution of scRNA-seq data, we formalize the clinical importance of high NF-YAr, associated to EMT genes and cell populations. We derive a novel, prognostic 158 genes signature common to BRCA and STAD Claudin^{low} tumours. Finally, we identify splicing factors associated to high NF-YAr, validating RBFOX2 as promoting expression of NF-YAl. These data bring three relevant results: (i) the definition and clinical implications of NF-YAr and the 158 genes signature in Claudin^{low} tumours; (ii) genetic evidence of 28 amino acids in NF-YAl with EMT-promoting capacity; (iii) the definition of selected splicing factors associated to NF-YA isoforms.

Cell Death and Disease (2023)14:65 ; https://doi.org/10.1038/s41419-023-05591-9

INTRODUCTION

Breast cancer (BRCA), the most common malignancy worldwide in women, is curable in 70% of cases, especially when identified at an early, non-metastatic stage [1]. It has been molecularly classified into different subtypes based on histopathological features and gene expression patterns. The use of Prediction Analysis of Microarray 50 -PAM50- was developed to identify four subtypes -Luminal A, Luminal B, HER2 + and Basal-like- exhibiting specific patterns of gene expression [2]. In parallel, the use of immunohistochemical (IHC) techniques concurs in the identification of the molecular subtypes. Unfortunately, neither PAM50 nor IHC-based techniques are absolutely accurate for clinical diagnosis, with numerous samples escaping from categorization. Recently, a discrete Claudin^{low} subtype was identified by bioinformatic means [3], characterized by low expression of critical cell adhesion molecules, including Claudin 3, 4 and 7, Occludin and E-cadherin, previously noticed in selected BRCA samples [4]. In general, commonalities are shared between Claudin^{low} and Basal-like tumours, among which mesenchymal and stem cell features [3].

Regulation of gene expression is at the heart of all developmental processes, including deviation from physiological patterns, such as cellular transformation and formation of tumours. The initial stage of gene expression is driven by the binding of sequence-specific Transcription Factors -TFs- to discrete promoter and enhancer elements. By analysing the structure of promoters driving expression of genes specifically overexpressed in cancer, many studies reported the overrepresentation of a specific element, the CCAAT box [5–10].

The TF regulating CCAAT box is the trimeric NF-Y, proven to be crucial for expression of many genes. It consists of three subunits, the Histone Fold Domain -HFD- NF-YB/NF-YC and the sequence-specific NF-YA. Several studies reported the overexpression of NF-Y subunits in different human cancers [11–18]. Typically, this came from analysing circuitries of TFs mediating transformation and NF-Y subunits were queried for increased expression. Most reports pointed to NF-YA as the overexpressed subunit in cancer, including our systematic analysis in TCGA and other datasets of breast, lung, liver, head and neck, prostate and stomach carcinomas [19–25].

A further level of complexity is brought by alternative splicing -AS- whose alterations have been shown to play a crucial role in the development of different types of tumours [26]. AS is governed by sets of proteins that guide either the inclusion of exons in the mature mRNA or their excision from the primary RNA transcript [27]. The NF-YA and NF-YC subunits are involved in AS: specifically, two major isoforms of NF-YA exist, NF-YAs (NF-YA short) and NF-YAl (NF-YA long), the latter comprising 28/29 amino acids within the large Q-rich Trans-Activation Domain (TAD), coded by exon-3. Both share the parts required for heterotrimerization and CCAAT-binding. We found that higher levels of NF-YAl correlates to a mesenchymal phenotype in BRCA, LUAD, HNSCC and STAD and a high ratio between NF-YAl and NF-YAs is predictive of a poor clinical outcome [19–21, 23, 25]. In particular, BRCA tumours and cell lines with a Claudin^{low} phenotype express high NF-YAl levels. Recently, it has been shown that these tumours are separate from the BRCA Basal-like subtype,

[1]Dipartimento di Bioscienze, Università degli Studi di Milano, Via Celoria 26, 20133 Milano, Italy. [2]These authors contributed equally: Michela Londero, Alberto Gallo.
✉email: diletta.dolfini@unimi.it

Received: 3 August 2022 Revised: 11 January 2023 Accepted: 13 January 2023
Published online: 28 January 2023

representing a fifth individual molecular subtype [28]. In general, Claudin^{low} tumours are not restricted to BRCA, since similar subtypes have been classified in Bladder (BLCA) and Stomach Adenocarcinomas, based on gene expression signatures resembling Claudin^{low} tumours [29, 30].

We performed wet and in silico experiments to explore the role of the NF-YA isoforms, initially in BRCA, then extending our findings to gastric cancers. We formalized the importance of the NF-YAr concept and derived a Claudin^{low} signature common to BRCA and STAD. Finally, we explored the role of splicing factors mediating expression of the NF-YAl isoform.

RESULTS

Deletion of NF-YA exon-3 in Claudin^{low} BRCA lines

Claudin^{low} breast cancer cell lines mostly express NF-YAl [19]. To understand the role of the isoform in breast cancer, we genetically ablated exon-3, coding for the 28 extra amino acids of NF-YAl, in two different Claudin^{low} cell lines, SUM159PT and BT549. We used the same strategy employed in murine cells [31], based on four guides flanking exon-3 and single strand-cutting Cas9-nickase (Cas9n), thus minimizing off-target effects (Supplementary Fig. S1A). After transfections, individual clones were isolated, expanded and genomic DNA screened by PCR, as shown in Supplementary Fig. S1B. We selected two clones with correct ablation in homozygosity for both cell lines and three random control clones without deletion (Supplementary Fig. S1C). The PCRs of clones #266 and #321 of SUM159PT, #12 and #242 of BT549, do not show the expected bands for the A amplicon, present only in parental cells and control clones, and for the B amplicon, shorter in edited clones respect to the controls (Supplementary Fig. S1C). Sequencing confirms deletion of exon-3, with somewhat different ends in the four YAl-KO clones (Supplementary Fig. S1D). We then checked isoform-specific mRNA expression: we did not score signals of NF-YA long transcript in edited clones (Supplementary Fig. S1E). Finally, analysis of protein levels by western blot confirmed exclusive expression of the NF-YAs in edited cells at comparable levels. As controls, NF-YB and NF-YC protein levels were assessed and found unchanged (Supplementary Fig. S1F).

NF-YAl sustains in vitro migration capacities of breast Claudin^{low} cells

SUM159PT and BT549 YAl-KO clones are stable upon repeated cycles of freezing/thawing and their morphology looks apparently similar to the parental cells and control clones (Supplementary Fig. S2A). We assessed the area occupied by cells and reported no difference in size (Supplementary Fig. S2B), nor in growth curves (Supplementary Fig. S2C). We also checked clonal expansion by clone formation assay: compared to controls, the number of colonies formed in YAl-KO cells was not statistically different. (Supplementary Fig. S2D).

We then employed the spheroid formation assay, a three-dimensional system, as a measure of cell-cell contacts and extracellular matrix formation. YAl-KO clones of SUM159PT and BT549 formed aggregates, whose morphology look looser and less compact than spheres formed by controls, and they did show many single, detached cells in the plates (Fig. 1A). This phenomenon was confirmed by a disaggregation-replating procedure, as shown in Fig. 1B. The looser aspect of YAl-KO clones suggested us to investigate their motility and invasion abilities, typical of Claudin^{low} cells. Transwell assays showed that YAl-KO clones significantly lost the capacity to migrate (Fig. 1C), and wound healing migration assays confirmed that YAl-KO clones fill the wound more slowly than controls (Fig. 1D). Altogether, these data indicate that expression of NF-YAs in Claudin^{low} cells guarantees basal growth features, but NF-YAl is involved in supporting invasion and migration. Of note, these features were shared by independent clones of two different Claudin^{low} cell lines.

Ablation of NF-YA exon-3 impairs in vivo migration of breast Claudin^{low} cells

Claudin^{low} breast cancer cells are highly invasive in vitro and metastasize in experimental and murine models in vivo [32]. To test the in vivo invasion/metastasis potential of the YAl-KO clones, we employed zebrafish, a widely used biosystem which allows to monitor these features. SUM159PT and BT549 YAl-KO clones and control cells were microinjected into the perivitelline space of 48 h post-fertilization (hpf) embryos and analysed 24 h post-injection (hpi), using an inverted fluorescent microscope. The spreading of fluorescently labelled extravasated cancer cells could be seen throughout the body of the fish, but predominantly in the area of the caudal haematopoietic tissue (CHT), located in the tail region (Fig. 2A). As expected, cells from SUM159PT and BT549 control clones clearly metastasize at 24 hpi, and the extent of metastasis was readily apparent, with large clusters of cells. The same results were obtained at 48 hpi (data not shown). Instead, SUM159PT YAl-KO clones showed less invasion potential and the BT549 ones hardly any (Fig. 2B). Moreover, they rarely showed extravasated cells in the body of the fish (Fig. 2). These correlations were consistent within the cohorts of fish used for each individual experiment. These data agree with in vitro invasion assays, further suggesting that NF-YAl is specifically involved in the metastatic process.

NF-YAl regulates distinct pathways, including EMT

To identify the deregulated pathways involved in the altered migratory capacities of YAl-KO clones, we performed RNA-seq of SUM159PT and BT549 YAl-KO clones and wt controls, defining differentially expressed genes, DEGs (Fig. 3A; listed in Supplementary Table S1). We observed that the genes collectively up- or down-regulated were relatively few in BT549 and more numerous in SUM159PT. This discrepancy is most likely due to the respective identities and to specific features of individual clones (Fig. 3B). Yet, extrapolation of the deregulated pathways led to the identification of common terms such as *EMT process*, *inflammatory signalling* (IFN), and *hypoxia* as collectively downregulated in YAl-KO clones of both cell lines (Fig. 3C). *EMT process* is expected, as its downregulation is consistent with the loss of migratory capacity of the YAl-KO clones. Intriguingly, *inflammatory* and *interferon signalling* genes, typically considered as tumour suppressive, are also downregulated, in keeping with such genes found over-expressed in breast Claudin^{low} tumours and correlating with worst prognosis [33, 34]. Overall, these terms suggest a less aggressive phenotype of the YAl-KO clones. Concerning upregulated terms, we found *p53* and *Estrogen Response pathways*, which seem counterintuitive, since Claudin^{low} belongs to the "triple negative" tumours, featuring loss of ERalpha expression and mutation of p53. However, the switch to NF-YAs could impact on the expression of genes previously classified as ER-responsive lacking EREs, hence independent from ER activities, characterized by the presence of E2F1, NF-Y and NRF1 motifs in their promoters [35]. As for p53, BT549 cells carry R249S, a "conformational" mutation *a-la* R175H: such mutants mostly behave like those impacting on DNA-binding (R248 and R273), that is, as Dominant Negatives, rather than Loss-of Function: for example, R249S fails to repress, and rather activates certain targets [36]: the differential regulation in the context of the two isoforms is an interesting issue worthy of further investigation.

A NF-YAr signature characterizes Claudin^{low} breast cancers

We previously reported on altered splicing of NF-YA isoforms in BRCA of TCGA and independent GEO datasets [19]. Traditionally, four subtypes have been described for BRCA, but more recently, the TCGA dataset has been re-classified into 5 distinct subtypes

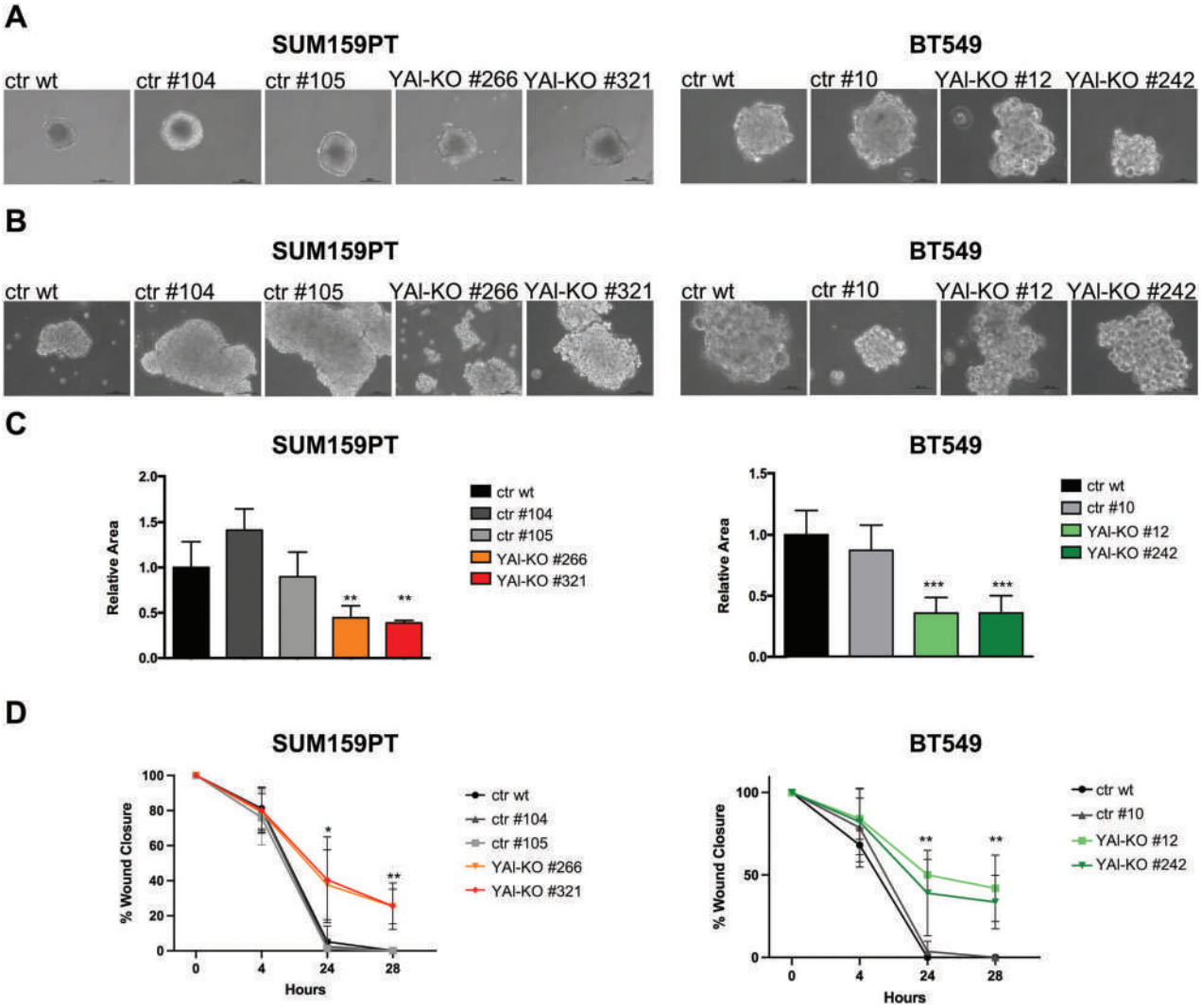


Fig. 1 In vitro characterization of NF-YAl-KO edited clones and controls. A Transmitted light images of SUM159PT and BT549 spheroids formation assay performed in complete medium. **B** Transmitted light images of SUM159PT and BT549 disaggregated and replated spheroids grown in complete medium. **C** Evaluation of invasion ability with transwell assay of SUM159PT and BT549; each bar represents mean value and error bars the SD of at least two independent experiments performed (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). **D** Migration evaluation ability through wound healing assay of SUM159PT and BT549; each point represents mean value and error bars the SD of three independent experiments performed (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

using the PAM50 signature [28]. Before proceeding with further analysis, we used the DeepCC machine learning method to classify all samples, including 28 Normal-like and 109 previously termed as Unclassified, using as training set the already classified tumours [22]. We checked the distribution of the new categorization (Supplementary Fig. S3A): only a small number of samples ($n = 12$) remained unclassified, with the majority being in the Luminal subtype (Supplementary Fig. S3B). Expression of NF-YA isoforms and their ratio (NF-YAr = NF-YAl/NF-YAs) was previously assessed according to the four subtypes classification [19]: we repeated this analysis with the complete classification, as in Fig. S3. This confirmed that the Claudin^{low} subtype shows highest expression of NF-YAl and lowest of NF-YAs, making NF-YAr higher in this subset (Fig. 4A). As expected, Basal-like tumours also show a similar trend.

Next, we extracted NF-YAr-associated gene signatures from RNA-seq datasets of all TCGA specimens. We replaced gene-level NF-YA expression with NF-YAl and NF-YAs at isoform-level, and we treated NF-YAr as a "normal" gene. Exploiting WGCNA (Weighted Gene Co-expression Network Analysis) and considering only tumour samples, 8 gene clusters emerged (Fig. 4B). NF-YAr and high NF-YAl expression grouped together in Module 7, suggesting that the ratio is influenced prevalently by the amount of NF-YAl (Fig. 4B). Module 7 is composed of 512 genes (Supplementary Table S2): plotting their expression according to the subtypes, we observed expression mostly in Claudin^{low} and in Basal-like (Fig.4C, Left Panel). Supplementary Fig. S4 shows a heatmap of gene expression according to the 8 modules and divided according to subtypes. Analysing the NF-YAr-driven 7th Module in BRCA cell lines expression data from CCLE, a significant correlation with Claudin^{low} and, to a lesser extent, Basal-like also emerged (Fig. 4C, Right Panel). Functional analysis of the NF-YAr signature revealed categories belonging to differentiation, specifically of tissues of mesenchymal origin; terms associated to cell adhesion and extracellular matrix emerged as the most deregulated GO categories (Fig. 4D). We also investigated the enrichment in MSigDB Hallmarks: inflammatory pathways and epithelial to mesenchymal transition were characterized as the most enriched sets (Fig. 4E). These data are in line both with the genetic and gene expression data of edited Claudin^{low} cell line shown above.

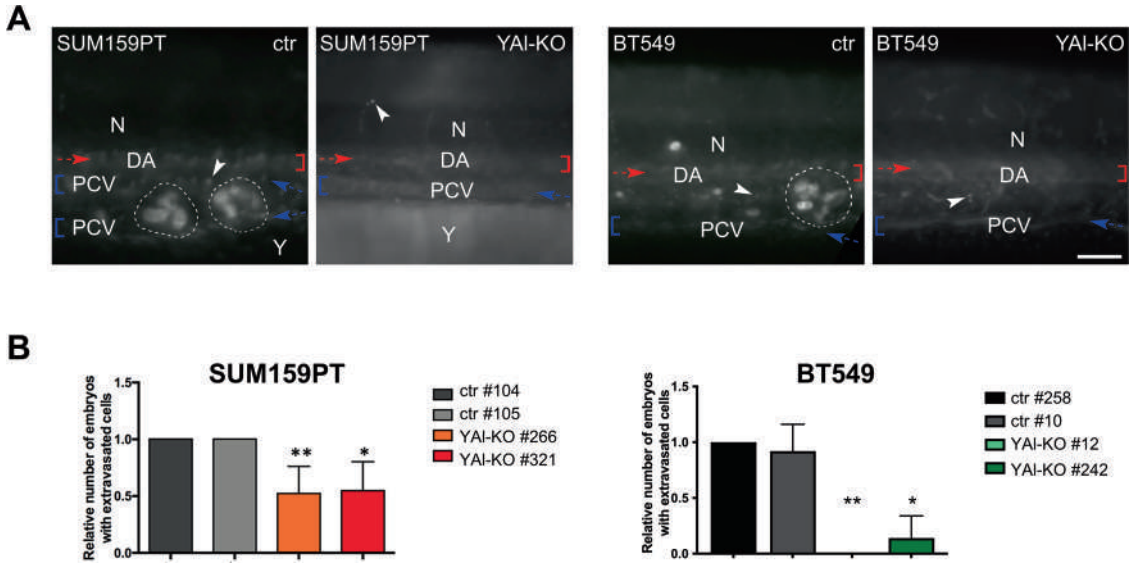


Fig. 2 In vivo characterization of NF-YAI-KO edited clones and controls. A SUM159PT and BT549 cancer ctr and YAI-KO were fluorescently labelled with Hoechst 33342 solution and directly injected into the Perivitelline Space (PVS) of 48 h post-fertilization (hpf) zebrafish larvae. Total number (*n*) of circulating-tumour-cells larvae analysed: SUM159PT ctr, $n = 98$; SUM159PT YAI-KO, $n = 112$; BT549 ctr, $n = 105$; BT549 YAI-KO, $n = 68$. All images are taken 24 h post-injection (hpi) and show the caudal haematopoietic tissue (CHT) tail region. The red and blue brackets indicate the DA and PCV, respectively, while the red and blue arrows show the blood flow in the two vessels. The white-dotted perimeters indicate the metastases (in the first image on the left of the panel the PCV is split in 2 branches because of the metastatic clusters). The arrowheads point out single circulating tumour cells. N notochord, DA dorsal aorta, PCV posterior cardinal vein, Y yolk. Scale bar: 100 µm. Anterior left, dorsal top. **B** Evaluation of migration ability through 48 hpf zebrafish embryos injection. Each bar represents mean value % of embryos that 24 hpi showed metastasis; error bars represent the SD of at least two independent experiments performed (*$p < 0.05$, **$p < 0.01$).

A NF-YAr 158 genes signature common to Claudin^{low} BRCA and STAD tumours

We recently reported on higher levels of NF-YAl in Claudin^{low} samples of gastric cancer [25]. We thus started to investigate TCGA STAD samples to define a NF-YAr signature that better marks Claudin^{low} tumours, by integrating co-expression network analysis. With the same pipeline described above, we extracted the genes clustering with NF-YAl and NF-YAr expression in gastric samples. Both fell in the same cluster in gastric cancer with 2508 genes. To restrict the NF-YAr signature, we extrapolated the genes commonly correlating with NF-YAr in both breast and gastric cancer, retrieving 158 genes (Fig. 5A and listed in Supplementary Table S2). We tested this signature across the different subtypes: Fig. 5B shows it mainly characterizes Claudin^{low} TCGA tumours of BRCA (Supplementary Fig. S3) and in STAD, the latter derived by using subtypes according to the ACRG classification, added of Claudin^{low} [25]. Functional characterization of the 158 genes indicates involvement in mesodermal related pathways and Gene Ontology indicates differentiation and muscle related categories (Fig. 5C); MSigDB Hallmark gene sets enrichment analysis further confirmed the GO results, highlighting pathways related to EMT and Myogenesis (Fig. 5D).

In conclusion, we identify a 158 genes signature defining Claudin^{low} tumours common to BRCA and STAD, correlating with high NF-YAr.

Prognostic value of NF-YAr and the 158 gene signature

We assessed the prognostic value of NF-YAr in the TCGA BRCA and STAD cohorts, independently from the subtype classification of samples. First, through the Cutoff Finder web tool, we found that the optimal NF-YAr value for dichotomization of Progression Free Survival -PFS- of BRCA samples is 0.86. The NF-YAr-directed stratification of BRCA samples shows an unbalanced distribution, yet survival analysis confirms that partition of samples with this cut-off predicts the survival outcome (Fig. 6A). Using the same approach, the cut-off in STAD was set to 0.27, predicting a threshold above which STAD patients show an unfavourable outcome (Fig. 6B). The same exercise was repeated with the 158 Claudin^{low} signature, yielding similar results in BRCA (Fig. 6C) and STAD (Fig. 6D). In summary, despite a numeric disproportion of Claudin^{low} samples –STAD ≫ BRCA- both NF-YAr and the 158 signature are predictive of a worst prognosis.

High NF-YAr correlates with EMT cancer cells in scRNA-seq BRCA and STAD

Bulk RNA-seq experiments have the limit of measuring gene expression of cell types within a tumour, including cancer cells. To discriminate correlations between NF-YAr and specific cell types, we exploited single cells experiments of breast [37] and gastric cancers [38]; we used the respective signatures to deconvolute the cell population composition of TCGA breast and gastric samples. In Fig. 7A, we ranked samples of the two cohorts, divided in deciles, according to NF-YAr, and we plotted average cell populations as predicted by the SCDC deconvolution tool. Both datasets show a slight increase in cancer-associated fibroblasts (CAFs) in samples with high NF-YAr values. Focusing on the predicted cancer cell population, we predict the composition of these cells in terms of cancer cells subpopulations: we observed a very robust correlation between the proportion of EMT cancer cells and NF-YAr (Fig. 7A, right panel). In addition, the BRCA Luminal population anticorrelated with NF-YAr values. Recently, the Claudin^{low} subtype emerged as separated, found primarily in Basal-like tumours and less in other subtypes [28]. We tested our NF-YAr predictor also inside the standard subtype framework: in Fig. 7B, we observed that within each subtype, the subset of samples with high NF-YAr correlated with a more expanded population of EMT cancer cells in breast and gastric, except for Basal-like. In summary, deconvolution analysis according to scRNA-seq data supports the hypothesis that high NF-YAr is predictive of cancer cells with an EMT phenotype.

RBFOX2 promotes NF-YA exon-3 inclusion in Claudin^{low} tumours

The above data showing that altered splicing of NF-YA isoforms has an impact on BRCA and STAD tumours begs the question as to

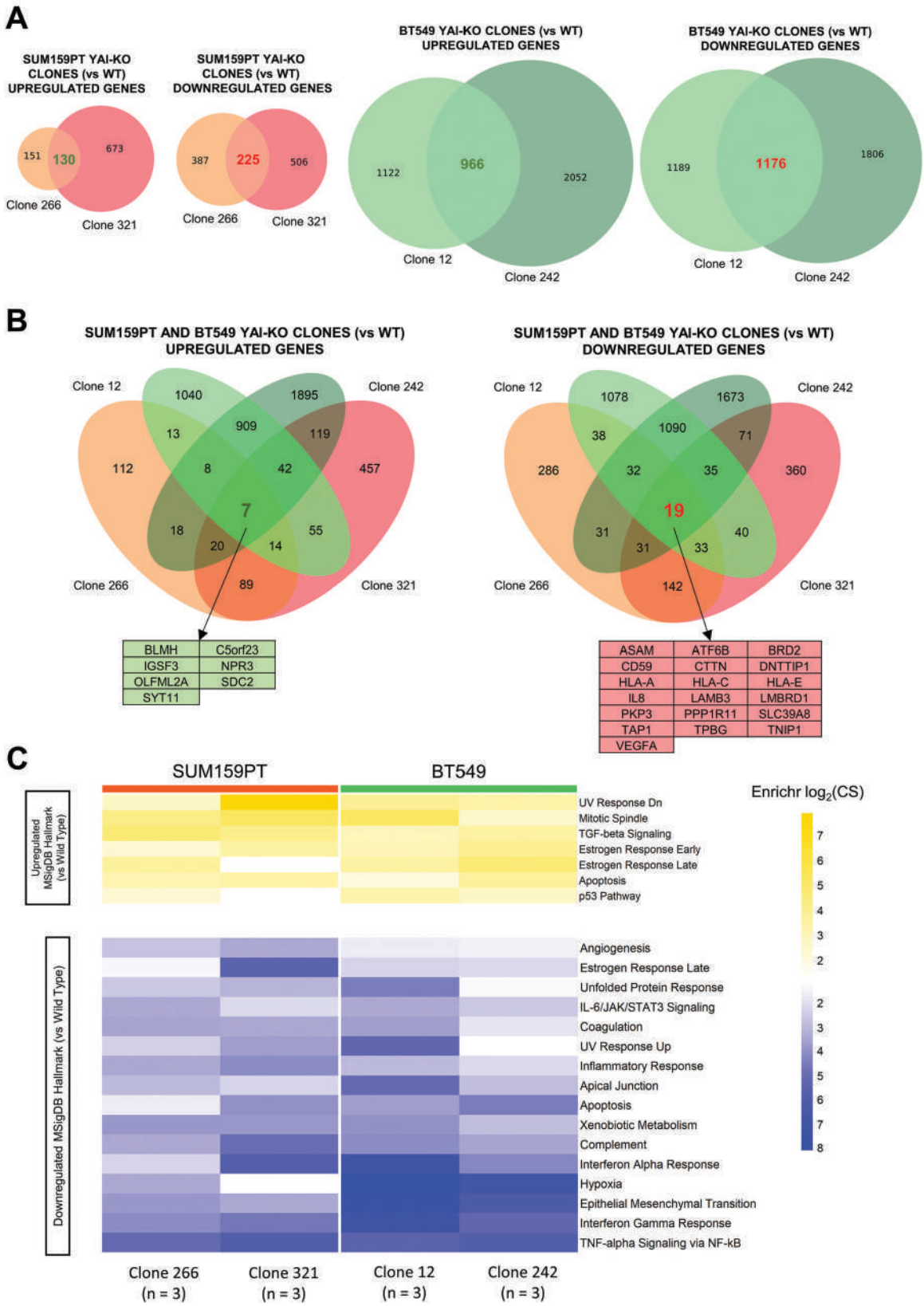
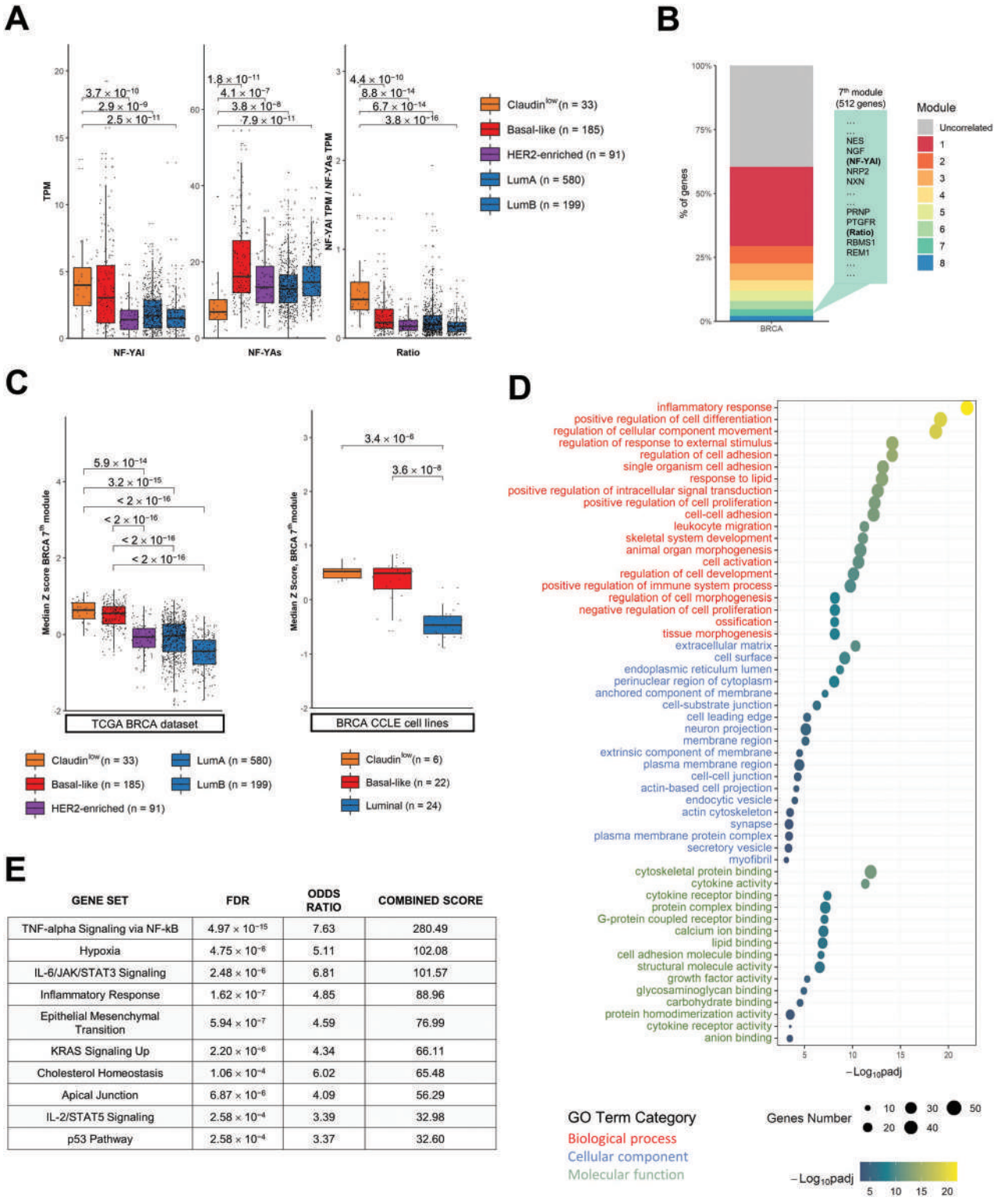


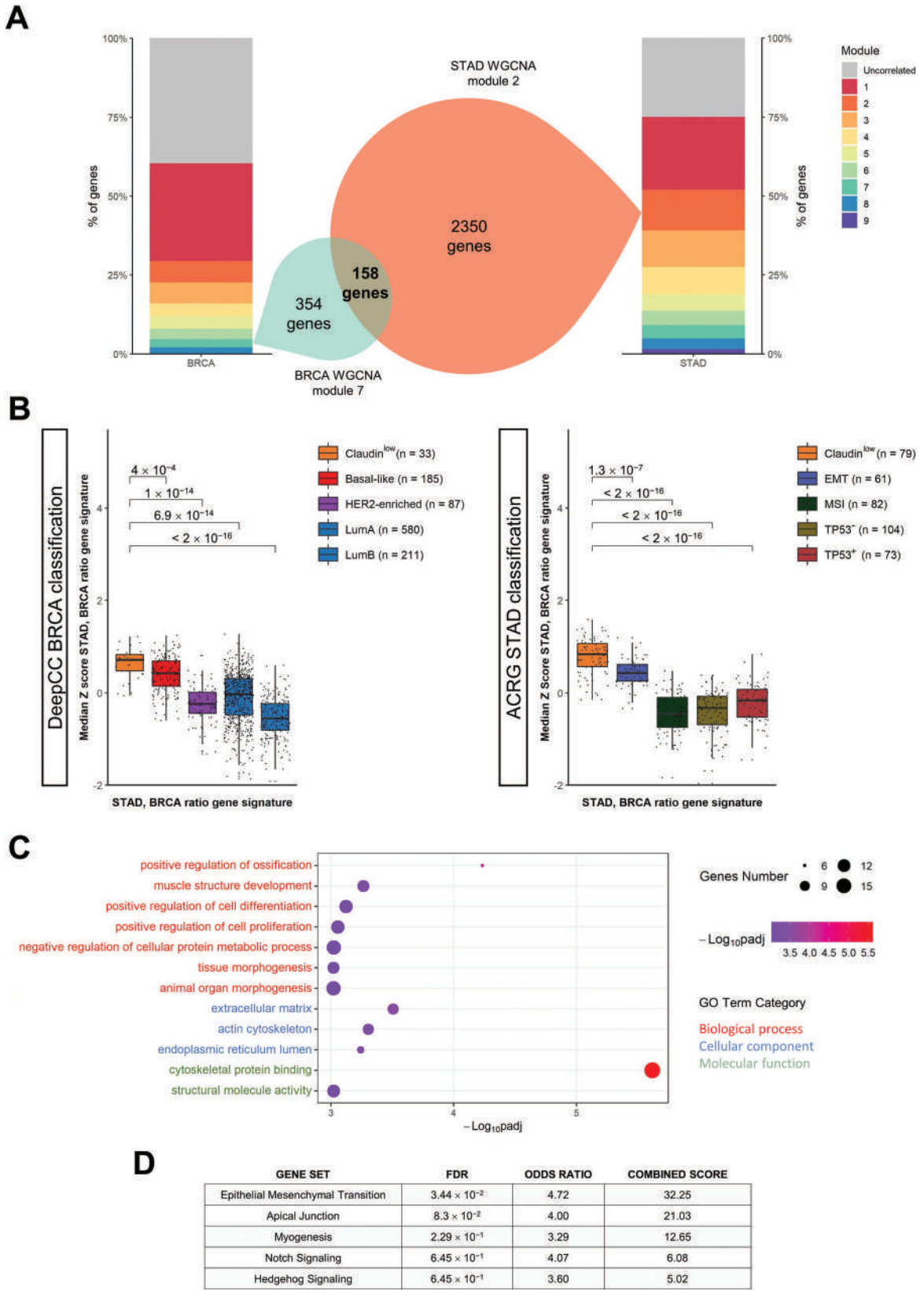
Fig. 3 SUM159PT and BT549 NF-YAI-KO clones differential expression and functional analysis. A Venn diagrams showing the overlap of differentially expressed genes (DEGs) in SUM159PT (Left Panel) and BT549 (Right Panel) YAI-KO clones, compared to the respective wt cell-lines. **B** Venn diagrams representing the number of shared DEGs between all four YAI-KO clones, compared to wt cell lines. In the Left Panel we included upregulated genes, in the Right Panel downregulated ones. Tables at the bottom detail the genes up- or downregulated in all four YAI-KO clones. **C** Heatmap depicting enriched MSigDB Hallmark gene sets within the DEGs (vs wt cell lines) of each deleted clone. Upregulated and downregulated gene sets are displayed in yellow and blue, respectively. Intensity of the colour = -$\log_2$(Enrichr Combined Score). Only gene sets with -$\log_2$(CS) > 1 in all four DEG lists were included in the analysis.



E

GENE SET	FDR	ODDS RATIO	COMBINED SCORE
TNF-alpha Signaling via NF-kB	4.97×10^{-15}	7.63	280.49
Hypoxia	4.75×10^{-6}	5.11	102.08
IL-6/JAK/STAT3 Signaling	2.48×10^{-6}	6.81	101.57
Inflammatory Response	1.62×10^{-7}	4.85	88.96
Epithelial Mesenchymal Transition	5.94×10^{-7}	4.59	76.99
KRAS Signaling Up	2.20×10^{-6}	4.34	66.11
Cholesterol Homeostasis	1.06×10^{-4}	6.02	65.48
Apical Junction	6.87×10^{-6}	4.09	56.29
IL-2/STAT5 Signaling	2.58×10^{-4}	3.39	32.98
p53 Pathway	2.58×10^{-4}	3.37	32.60

Fig. 4 BRCA NF-YAI/NF-YAr WGCNA gene module characterization. A Expression levels of the NF-YA isoforms and NF-YAr in the TCGA-BRCA dataset, measured in TPMs. Samples are divided according to the new DeepCC classification. **B** graphical representation of the 8 gene modules determined by weighted gene co-expression network analysis, with the associated proportion of total genes. The module depicted in grey gathers uncorrelated genes. **C** Box plots of the median Z scores calculated for each sample across genes of the BRCA WGCNA 7th gene module, in the TCGA-BRCA dataset (Left Panel) with samples divided according to the DeepCC classification, and in CCLE (Right Panel), where cell lines are separated following the classifications from Prat et al. [65] and Dai et al. [66]. **D** KOBAS-generated list of GO terms enriched in BRCA WGCNA 7th module. Terms are ranked according to -log$_{10}$(q value) and divided within the three main GO categories. Dot sizes represent the genes shared by the module and each GO term. **E** Most enriched MSigDB Hallmark gene sets in BRCA WGCNA 7th module, as calculated by the Enrichr website. p values in (A) and (C) box plots are calculated using the Wilcoxon rank-sum test.



D

GENE SET	FDR	ODDS RATIO	COMBINED SCORE
Epithelial Mesenchymal Transition	3.44×10^{-2}	4.72	32.25
Apical Junction	8.3×10^{-2}	4.00	21.03
Myogenesis	2.29×10^{-1}	3.29	12.65
Notch Signaling	6.45×10^{-1}	4.07	6.08
Hedgehog Signaling	6.45×10^{-1}	3.60	5.02

which splicing factor is responsible for this behaviour. First, we interrogated the TCGA repository by analysing all the epithelial cancers cohorts: we divided tumour samples for each type based on quartiles of NF-YAr, defining two groups: high NF-YAr samples associated to the fourth quartile, and low NF-YAr group with all other samples. Then, we compared expression of the RNA binding protein (RBPs) genes included in the RBP database [39] in the two groups. Fig. 8A shows the RBPs that significantly correlated (on the

Fig. 5 BRCA-STAD NF-YAr Signature definition and characterization. A Overlap between BRCA WGCNA 7th module and STAD 2nd module, including NF-YAl and NF-YAr. The stacked bar plots at the borders represent the percent of total genes belonging to each module in the two cohorts, and the grey module includes uncorrelated genes. **B** Box plots of the median Z scores calculated for each sample across the genes of BRCA-STAD NF-YAr gene signature, in the TCGA-BRCA dataset (Left Panel) and in TCGA-STAD dataset (Right Panel). Samples are distributed according to the respective DeepCC classifications. p values are calculated using the Wilcoxon rank-sum test. **C** GO terms enriched in BRCA-STAD NF-YAr gene signature, as computed by KOBAS. Terms are ranked according to -$\log_{10}$(q value) and divided within the three main GO categories. Dot sizes represent the number of genes shared by the module and each GO term. **D** Most enriched MSigDB Hallmark gene sets in BRCA-STAD NF-YAr gene signature, as calculated by Enrichr.

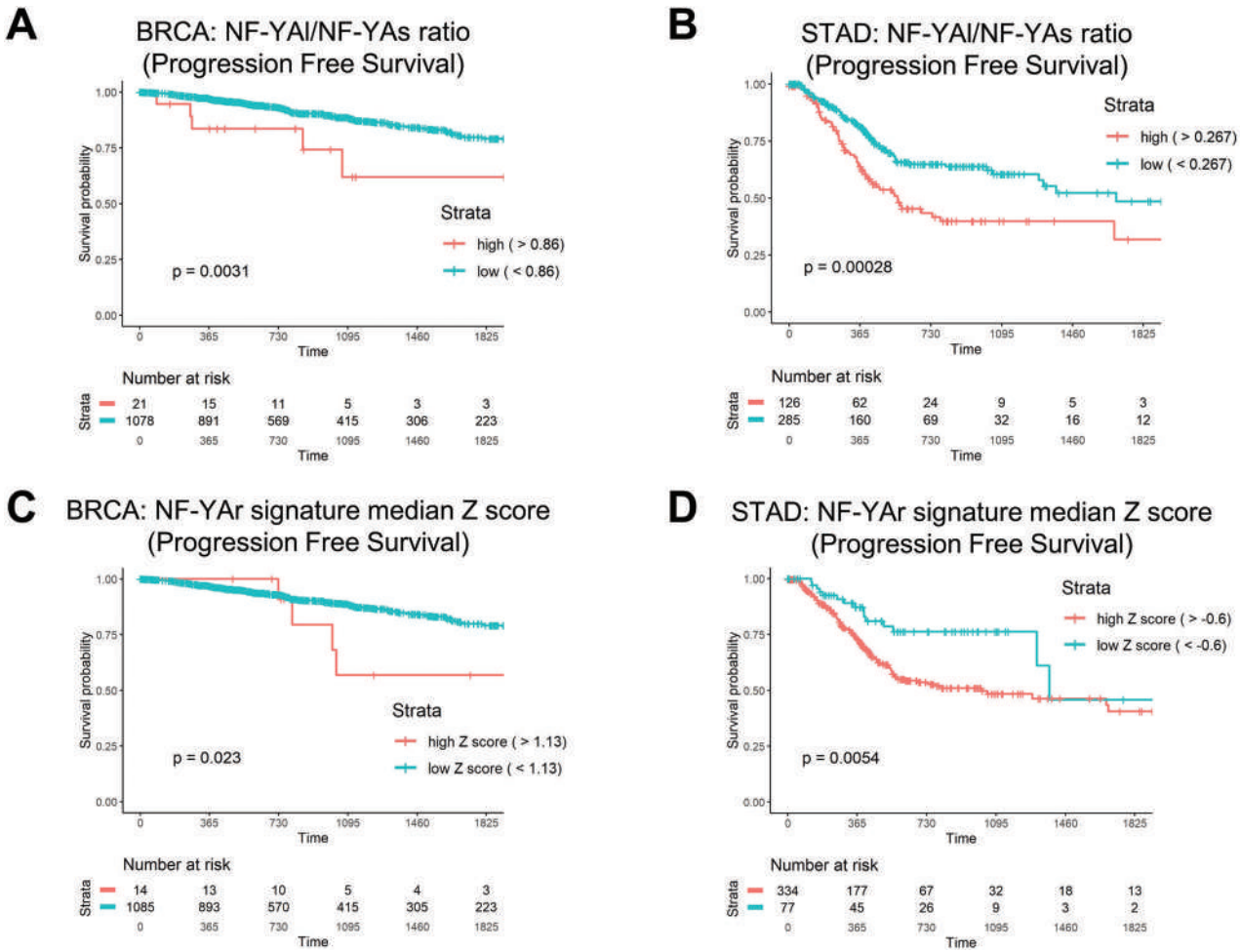


Fig. 6 Clinical analysis of NF-YAr in TCGA BRCA and STAD cohorts. Progression-Free-Interval curves of survival probability, with stratification according to the Cutoff Finder-determined threshold for NF-YAr (0.86) in TCGA BRCA patients (**A**), for NF-YAr in STAD (**B**) threshold = 0.27, for the NF-YAr signature median Z score in BRCA (**C**) threshold = 1.13, and for the NF-YAr signature median Z score in STAD (**D**) threshold = −0.60. *p* values are calculated using the log-rank test.

left) and anticorrelated (on the right) in expression with high levels of NF-YAr. The classification was performed ordering RBPs according to the number of cancer cohorts in which we found a significant correlation, trying to predict the candidate RBPs responsible for NF-YA isoforms switching. On the top of the correlating list, we found QKI and RBFOX2, while RBM47 and ESRP1 resulted as the most anticorrelating (Fig. 8A). In parallel, we searched for RBPs motifs in the NF-YA exon-3 flanking introns through the oRNAment tool [40]: we confirmed the presence of direct sites of RBFOX2 and QKI in NF-YA transcripts as indicated in Fig. 8A. We checked the expression of the top 20 correlating and anticorrelating RBPs across BRCA and STAD subtypes: the heatmaps of Supplementary Fig. S5 shows high expression of most correlating RBPs in Claudin^{low} samples, and low in LuminalA/B. The reverse is observed for anticorrelating RBPs. Finally, we analysed CCLE BRCA lines: high NF-YAr-related RBPs -MBNL1/1, QKI and RBFOX2- are expressed predominantly in Claudin^{low} lines; CELF2, instead, appears to be expressed almost exclusively in specific Basal-likes. Instead, the anticorrelating RBPs ESRP1 and RBM47 are specifically lowly expressed in Claudin^{low} lines and show high expression in Luminal cells (Supplementary Fig. S6): in such lines, we previously detected almost exclusive expression of NF-YAs [19].

To substantiate these findings, we analysed available RNA-seq experiments in which correlated and anticorrelated RBPs were functionally inactivated, or overexpressed, for expression of NF-YA isoforms. Upon alteration of levels of these RBPs, we scored a significant impact on isoforms expression, namely increase of NF-YAs in MBNL1/2, QKI and -markedly- RBFOX2 inactivation, and increase of NF-YAl in RBM47 and ESRP1-2 inactivation (Supplementary Fig. S7). To further validate our predictions, we over-expressed RBFOX2 in Luminal breast cells, chosen after analysis of

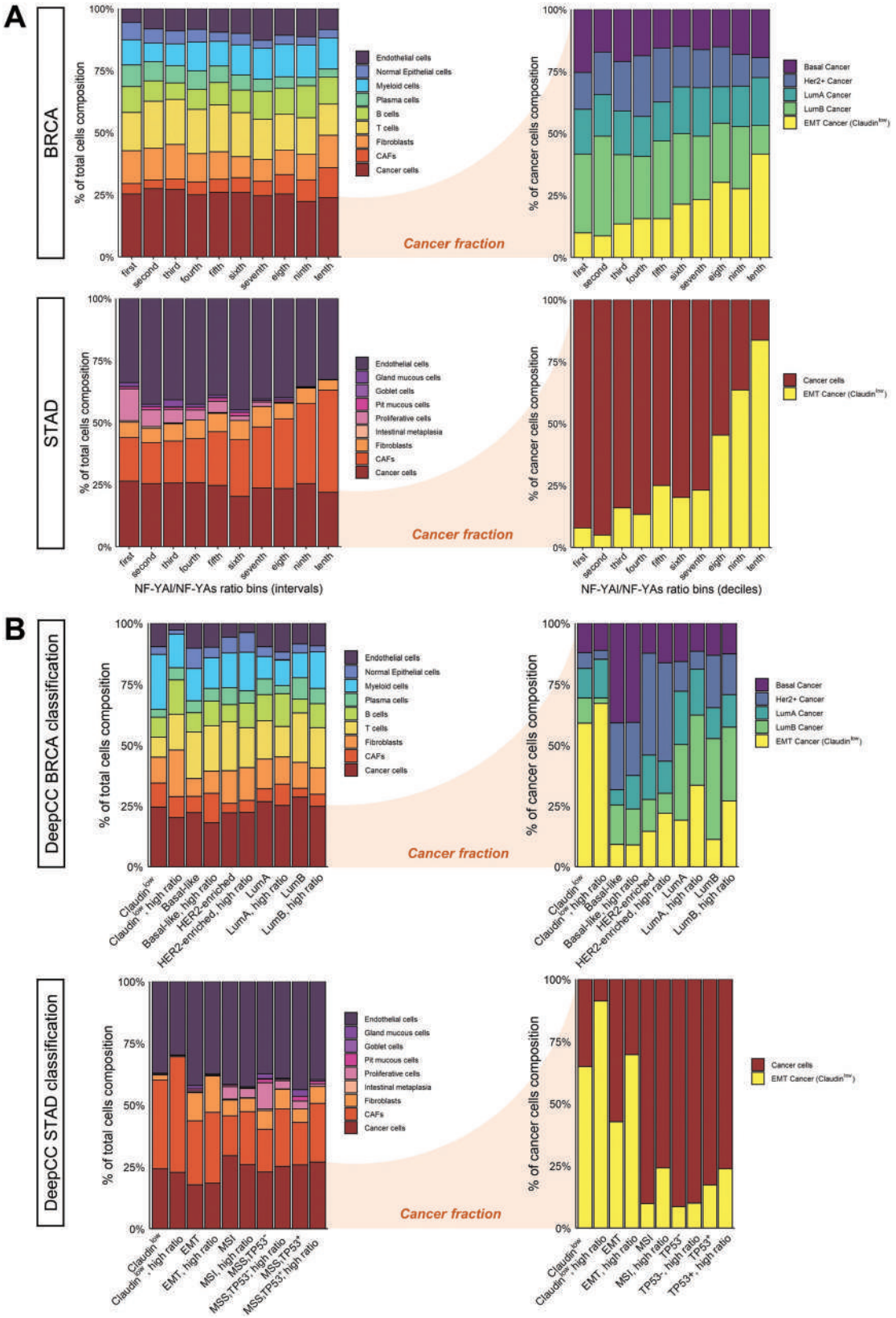


Fig. 7 Cell type deconvolution of TCGA BRCA and STAD samples. A Proportion of cells annotated as each cell type, as predicted by the SDCD deconvolution package, for TCGA BRCA and STAD samples: in the Left Panel all cell types constituting the tumour microenvironment are considered, whereas on the Right only cancer cells are evaluated. TCGA samples are divided according to NF-YAr deciles, and the average cell type proportions among samples comprising each decile are plotted. **B** Same as (A), except that TCGA BRCA and STAD samples are partitioned according to their respective DeepCC classifications.

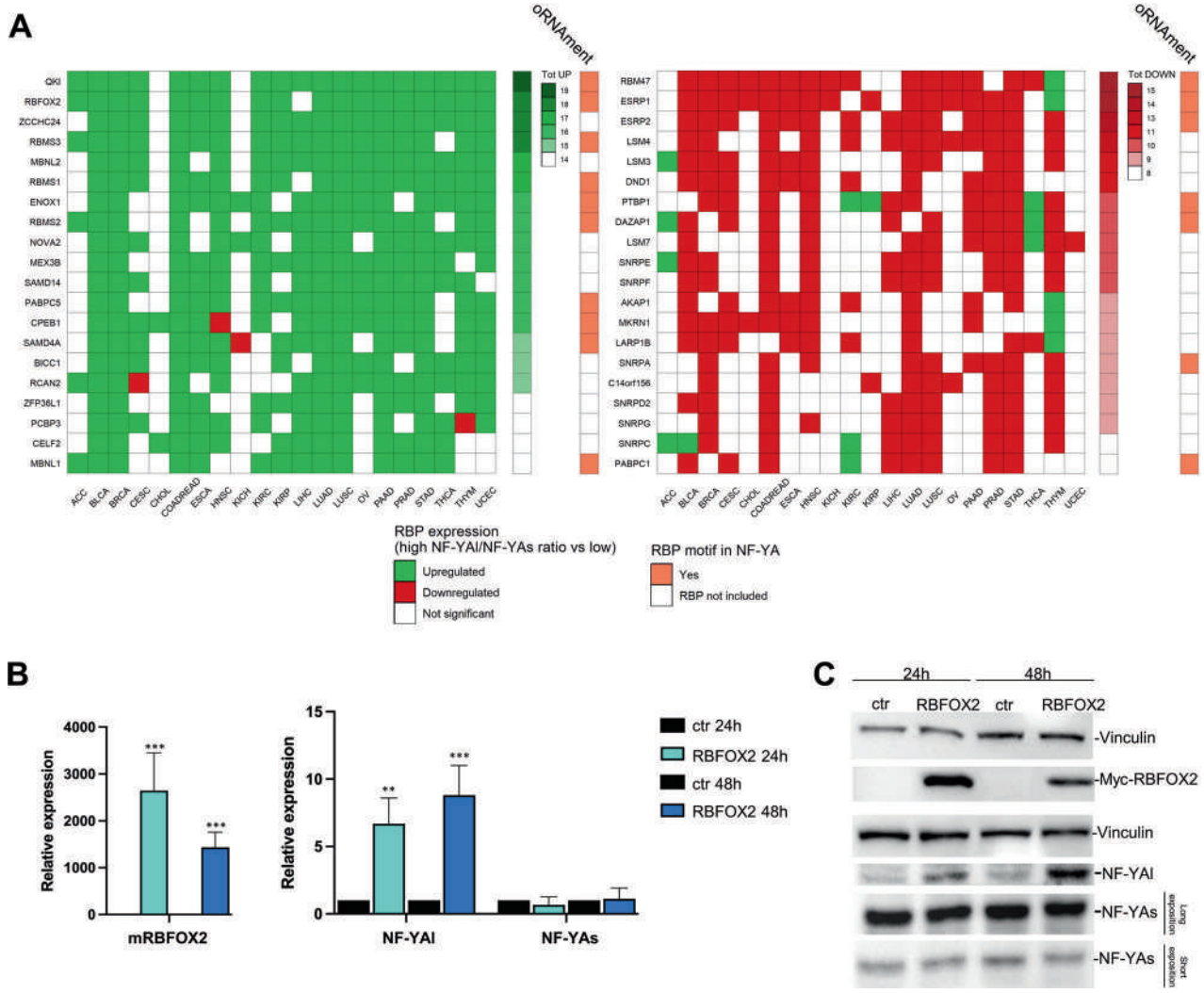


Fig. 8 RBFOX2 correlation with NF-YA exon-3 inclusion. A Heatmaps depicting variation of RBPs expression between high NF-YAr samples (fourth quartile) and all others, in 21 TCGA epithelial cancer cohorts. Green squares represent significant overexpression ($p < 0.05$) in high NF-YAr samples, as tested by Wilcoxon rank-sum test, and red squares are associated to lower expression values. Left Panel: RBPs are ranked according to the number of cohorts in which they are significantly overexpressed ("Tot UP" column). Right Panel: RBPs are ranked according to the number of cohorts in which they have a significantly lower expression ("Tot DOWN" column). Only the top 20 RBPs are shown in both panels. The oRNAment column illustrates the presence of each RBP binding site motif within the NF-YA locus. **B** Gene expression analysis of mouse RBFOX2, NF-YA long and NF-YA short levels in T47D luminal cells transfected with plasmid carrying mRBFOX2 gene or with an empty plasmid. Each bar represents mean value and error bars the SD of at least two independent experiments performed (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). **C** Western Blot analysis of MYC-RBFOX2, NF-YA short and NF-YA long in T47D luminal cells transfected with plasmid carrying mRBFOX2 gene or with the empty control.

the RBPs expression patterns across CCLE breast cell lines, confirming the same trend observed for tumour subtypes (Supplementary Fig. S6). We overexpressed RBFOX2 in the luminal cell line T47D, with prevalent expression of NF-YAs: as shown in Fig. 8B we observe maximum expression of RBFOX2 at 24 h, and after 48 h we observed an increase of the NF-YAl transcript. We confirmed this also at the protein level, with the peak of NF-YAl expression at 48 h post-transfection (Fig. 8C). In conclusion, bioinformatic and wet biology experiments suggest that two RBPs -RBFOX2 and QKI- are involved in controlling the levels of NF-YA isoforms, specifically increasing NF-YAl.

DISCUSSION

Our work sheds light on the importance of the individual NF-YA isoforms for the biology of BRCA and STAD. The role of NF-YAl was studied in vitro and in vivo by gene editing; bioinformatic analysis, including deconvolution of scRNA-seq data, led to the identification of a NF-YAr signature shared by Claudin^{low} BRCA and STAD. Finally, we provide evidence of a role of selected AS proteins in influencing NF-YA isoforms levels.

NF-YA isoforms

The ubiquitous presence of NF-Y-binding activity in all cell lines, of CCAAT sites in promoters of disparate genes and the results from YAl-KO experiments signalling an essential role of NF-YA in early embryogenesis, all contributed to the conclusion that NF-Y, including NF-YA, is not relevant for cellular transformation. This view changed when it was realised that NF-YA is absent in some post-mitotic cells, it is associated to cell proliferation in vivo and that the targeted genes are enriched in growth-promoting categories, namely cell-cycle progression and metabolism. Not surprisingly, functional removal of NF-YA -by siRNA or shRNA- leads to impairment -or lack- of cell growth [41]; conversely,

overexpression promotes different aspects of cell proliferation and tumorigenesis [24, 42–45]. Overexpression of TF genes can have widespread direct -and indirect- transcriptional consequences, and this is most likely true for NF-YA, whose protein expression levels are tightly controlled [46]. Indeed, stable, complete ablation of NF-YA in cell lines has not been reported so far, being likely lethal for growing cells.

We used genetic means to ablate exon-3, generating Claudin^{low} cells lines expressing physiological levels of NF-YAs instead of NF-YAl. Basic features -morphology, growth rates- of the YAl-KO clones are apparently unaffected, at least in the conditions tested. We reported similar findings with mouse myoblasts C2C12, substituting NF-YAl with NF-YAs by genome editing [31]. The two Claudin^{low} lines used here are different in morphology, growth rates and expression profiles, but they share commonalities, such as high migratory capacities, including to metastasize in vivo: in both lines, these features are substantially diminished -or lost- in YAl-KO clones expressing NF-YAs.

Over the last few years, several groups reported that NF-YA is overexpressed in cancers, specifically epithelial ones [19] and that the roles of the two major isoforms might be fundamentally different. In reports on endometrial carcinoma (EC), two studies reported a role of NF-YAs in aggressive ALDHA1^{+} tumours [47] and a direct correlation with EMT markers in high-grade EC [18]. A recent report delineates a complex interplay between the two isoforms in prostate cancer (PCa): on the one hand, NF-YAs levels are higher in LumB, the subtype with the worst clinical outcome, and overexpression increases spheroids and xenografts tumours [24]. However, a significant increase in the NF-YAl/NF-YAs ratio distinguishes circulating tumour cells from cells within metastasis: this is consistent with increased pro-migratory functions observed in NF-YAl-overexpressing PCa lines. This paints a scenario in which NF-YAr is modulated according to different stages of tumour progression: first NF-YAs (increased proliferation), then NF-YAl (detachment and migration to distant sites), finally NF-YAs again within metastasis (attachment/homing and growth?). In BRCA and STAD, instead, we report that locally aggressive Claudin^{low} tumours, irrespective of their metastatic status, have already high NF-YAr, which is a clinically impactful concept. Finally, an additional NF-YA isoform -NF-YAx, devoid of both exon-3 and exon-5- is expressed in neuroblastoma (NB) [48]. Overexpression of NF-YAx activates key genes -Nestin, SOX2, Nanog- that lead to selection of NB cancer stem cells. Note that all isoforms, including NF-YAx, share the conserved HAP2 domain responsible for heterotrimerization and efficient and selective DNA-binding, thus all CCAAT box-containing genes could be targeted and activated. These findings are intriguing, since this shorter version lacks large parts of the TAD, including stretches we recently identified as extremely conserved in evolution [49], one of which proven to be functionally important [50]. In fact, NF-YAx loses the capacity to interact with Sp1/3, well known NF-Y partners in activation, thus it is expected to have negative effects on expression of some genes. We have been unable to detect relevant levels of NF-YAx in the epithelial tumours we analysed, so this isoform might be very specific for neuronal malignancies.

NF-YAr

Our bioinformatic analyses expand previous concepts and provide new insights into the association between the ratio of NF-YA isoforms and epithelial tumours, particularly Claudin^{low}. First, we confirm previous reports suggesting that Claudin^{low} tumours are a discrete subtype in BRCA and STAD, based on profiling analysis of the respective TCGA samples. Their number in BRCA is lower -some 3%- than in STAD, yet upon categorization of DEGs, they share common pathways and GO terms. Second, we previously reported on NF-YA overexpression in BRCA and STAD, and on high levels of NF-YAl in breast and gastric Claudin^{low} cell lines and tumours: we now define and quantify this concept of discrete values of NF-YAr. In turn, these are relevant to predict the clinical outcome of these tumours. Note that there is a difference in the values in the two tumours, but this might have to do with the higher number of Claudin^{low} tumours in STAD, generating more robust statistics. Third, using NF-YAr as feature, we derived a 158 genes signature common to BRCA and STAD. Note that previous Claudin^{low}-based signatures, derived separately in BRCA and STAD, were substantially different. Thus, our data go along with the establishment of a unifying set of genes whose expression is commonly deregulated in these types of tumours, independently from the origin. This signature is enriched in EMT and mesenchymal terms in general, which is to be expected based on the previous and present analysis and on the genetic experiments and RNA-seq data presented in Figs. 1–3. In turn, these data directly impinge on the specific role of NF-YAl -and not NF-YAs- in these classes of genes. Fourth, by analysing scRNA-seq data to deconvolute BRCA and STAD cellular compositions present in TCGA samples, we found that there is a parallel increase of NF-YAr and the Cancer Associated Fibroblasts -CAF- population, but not that of normal fibroblasts, whose overall numbers are not affected. This is relevant, since CAFs are well known to play an important role in tumour expansion [51]. As for the population of cancer cells, which make up some 25% of both STAD and BRCA samples, the higher the NF-YAr, the more numerous are the cancer cells labelled as EMT. This effect is particularly striking in STAD samples. Even partitioning for tumour subtypes and high or low NF-YAr, the EMT phenotype follows well high ratios, further confirming that these features are intrinsically associated. We conclude that a substantial change in cellular behaviour is brought by exon-3 28 amino acids within the TAD. These results bring at least two avenues of future investigations: verification of the Claudin^{low} signature identified here in other epithelial tumours, in primis BLCA, and identification of the molecular mechanisms empowering these 28 amino acids with such specific transcriptional effects.

RBPs in cancer

AS is taking a centre-stage in studies of expression profiling of tumours [52], and many studies increasingly focus on the roles of RNA-Binding Proteins involved in AS. A global change in expression of specific isoforms of hundreds of genes is routinely detected in tumours and, indeed, this goes along with changes in expression of individual RBP genes. We asked the obvious question, that is, which RBP affects differential splicing of the two NF-YA isoforms. By interrogating acknowledged members of the RBP database for correlation with high NF-YAr in TCGA tumour samples, some factors emerge; other RBPs, predictably, anti-correlate. Reassuringly, the top hits did show the presence of the respective sites within the exon-3 RNA splicing areas. We validated changes of NF-YA isoforms by analysing RNA-seq experiments previously performed after RBFOX2, CELF2, MBNL1-2 and QKI knock-down: all hits behaved as expected, namely inactivation of these genes entails a decrease of NF-YAl. RBFOX2 was further tested upon overexpression in a luminal breast line predominantly expressing NF-YAs, and indeed we detect an increase of NF-YAl at the mRNA and protein level. In fact, RBFOX2 is involved in alternative splicing in mesenchymal tissues and during the epithelial-mesenchymal transition process, which is important for cancer cell metastasis [53].

As to the RBPs that anti-correlate with high NF-YAr, thus promoting eviction of exon-3 and production of NF-YAs, RBM47 and ESRP1-2 are validated by RNA-seq data of knocked-down cells. Two further results are consistent with our data. Experiments on ESRP1/2 functional KO by shRNA interference reported altered NF-YAr ratios, with a substantial increase of NF-YAl and decrease of NF-YAs, which was prevalent in the cell line used [54]. In the context of cellular reprogramming of MEFs -expressing

predominantly NF-YAl- to iPS -predominantly NF-YAs- Cieply et al. showed that inactivation of RBM47 leads to a shift of the NF-YA protein isoforms from NF-YAs to NF-YAl [55]. Epithelial specific alternative splicing is regulated by RBM47 that is also found inactive in some breast cancers and whose low expression in patients correlates with poor clinical outcome [56].

On the other hand, we were surprised not to find U2AF2 in the RBPs list emerging from our analysis. Conditional inactivation of U2AF1 in haematopoietic mouse cells led to failure of haematopoiesis due a defect in stem cell renewal, ultimately fatal [57]. NF-YA levels are decreased, along with other relevant TFs, such as PBX1, Meis1 and Runx2. Most importantly, U2AF1 inactivation by shRNA led to splicing alterations and decrease in NF-YA levels. Conversely, OE of NF-YA compensate for U2AF1-KO. These data are consistent with previous work showing a similar phenotype in mice in which NF-YA was conditionally KO in haematopoietic stem cells, mostly expressing NF-YAs [58]. The lack of U2AF1 in our lists might be due to tissue-specific effects since most datasets we analysed are from epithelial tumours. In general, it is possible that no single AS factor has a dominant role on NF-YA splicing, which might be coordinately controlled by several factors, partially in a tissue-preferred way.

In conclusion, we are left with a relatively small group of RBPs that are worth further genetic work: genome editing could shed light on the collective AS circuit in the cells we used. In addition, our data awaits the actual biochemical prove that these RBPs bind sequences in NF-YA exon-3 boundaries, in vitro and in vivo.

MATERIALS AND METHODS

Cell culture and transfection

Human breast cancer cells SUM159PT (ATCC) were cultured in DMEM/F12 (1:1) supplemented with 10% FBS (EuroClone), insulin 5 µg/ ml (Sigma-Aldrich), L-glutamine 1 mM 100 µg/ ml (EuroClone), Penicillin and Streptomycin 100 µg/ ml (EuroClone), Hydrocortisone 5 µg/ ml and HEPES 25 mM (EuroClone). BT549 (ATCC HTB-^{122TM}) were grown in RPMI 1640 supplemented with 10% FBS, insulin 5 µg/ ml, L-glutamine 1 mM 100 µg/ ml, Penicillin and Streptomycin 100 µg/ ml. T47D cells (ATCC HTB-^{122TM}) were cultured in DMEM/F12 (1:1) supplemented with 10% FBS, L-glutamine 1 mM 100 µg/ ml, Penicillin and Streptomycin 100 µg/ ml. All cell lines resulted negative for mycoplasma test. 4×10^5 T47D cells were transfected with 4 µg of pIRESneo plasmids carrying mouse Myc-RBFOX2 or empty control with Lipofectamine$_{2000}$.

Genome editing

To target NF-YA exon-3, we used the same strategy employed in murine muscle cells [31], with four gRNAs (Supplementary Table S3) designed to target the introns flanking the human exon-3. 10^6 SUM159PT and BT549 cells were transfected with 3 µg of the two gRNAs/nCas9 plasmids by electroporation (Nucleofector® 2b Device). A total of 420 individuals clones for each line were isolated 72 h post-transfection and expanded. Clones were then screened for exon-3 deletion by genomic analysis by semi-quantitative PCR using the amplicons indicated in Supplementary Fig. S1B. Two homozygously deleted clones were identified for each line and thereafter sequenced to verify the deletion ends.

RNA extraction, real-time PCR and RNA-seq

RNA samples were extracted with the Tri-Reagent protocol (Sigma-Aldrich). RNA samples were extracted, retrotranscribed and analysed as in murine cells [31]. RNA samples for RNA-seq were extracted from three independent biological replicates, treated with DNAse and checked for their quality by RNA ScreenTape Assay with TapeStation System. For BT549, mRNA was enriched using oligo(dT) beads, cDNA synthesis done using random hexamers and reverse transcriptase; a custom second-strand synthesis buffer (Illumina) was added with dNTPs, RNase H and Polymerase I to generate the second strand by nick-translation. After a round of purification, terminal repair, A-tailing, ligation of sequencing adapters, size selection and PCR enrichment, the cDNA libraries were sequenced. For SUM159PT, total RNA was depleted of ribosomal RNA and the RNA-seq libraries prepared with the Illumina TruSeq Stranded Total RNA kit following the manufacturer's protocol. Amplified libraries were checked on a bioanalyzer 2100 and quantified with picogreen reagent. Libraries with distinct TruSeq adapter UDIndexes were multiplexed and, after cluster generation on FlowCell, were sequenced for 50 bases in paired-end mode with a Novaseq 6000 sequencer (40×10^6 reads coverage).

Protein extraction and western blot analysis

For Whole Cell Extracts -WCE- preparation, SUM159PT, BT549 and T47D cells were extracted with RIPA buffer [31]. Primary antibodies used were anti-NF-YA G-2 (sc-17753 Santa Cruz), anti-NF-YB (home-made), anti-NF-YC (home-made), anti-Vinculin (05-386 Sigma-Aldrich), anti-Myc (hybridomas 9E10). Secondary antibodies were peroxidase-conjugated anti-mouse (A4416 Sigma-Aldrich) and anti-rabbit (A6154 Sigma-Aldrich). Western blot experiments were performed on three independent biological replicates. Full and uncropped western blots are presented in Supplementary Fig. S8.

Cellular assays

For size analysis, cells were fixed with 2% paraformaldehyde for 10', washed with PBS, permeabilized by 0.1% TritonX-100 for 20', blocked with 1% BSA for 20', stained with Rhodamine Phalloidin 1:500 for 45', washed three times and stained with Hoechst 33342 1:2000 for 5'. Pictures were taken at 40x magnification with Leica CTR6000 Fluorescent Microscope; areas of randomly chosen cells were collected and analysed with ImageJ 1.53.

Proliferation assays were performed by plating 2500 cells (SUM159PT) or 5000 (BT549) in 96-wells plates and counting every 24 h for 3 days, using the Hoechst 1:2000 with automated cell counting by High Content Screening. The number of cells were normalized on 4 h post-plating count.

For spheroids formation, 10^5 SUM159PT and BT549 cells were plated in 10 cm non-coated plates; the medium was changed every 4 days and the data acquired after 14 days of growth. We collected images and analysed the shape and aggregation state.

For clonogenic assays in 2D Culture, 10^3 SUM159PT and BT549 cells were plated in 10 cm dishes, medium changed after 7 days; after 14 days, plates were washed with PBS and fixed/stained with a Crystal Violet solution (Crystal violet 0.0005%, Formaldehyde 1%, Methanol 0.01%, in PBS) for 20' at room temperature. We counted colonies with more than 50 cells.

For wound healing assay, cells were cultured in a 24-well culture plate for 24 h to 90%–95% confluency. Wound line was created by scratching the plate with a 10 µl micropipette tip. Cells were washed with PBS and the average width of the gaps calculated from the image taken with a microscope. Invasion assay was performed Corning strain with 0.45 µm pore size membranes. Cells were resuspended in 100 µl of media without FBS and insulin and seeded at a density of 10^4 cells per well. Membranes bottom sides were put in contact with 600 µl of complete media. After 24 h of culture, invasive cells were fixed and stained with crystal violet 0.2% solution, pictures taken at optical microscope (LEICA ICC50 W, 4×) and % of membrane area occupied by invasive cells was counted with the ImageJ software (LAS V4.9).

All biological data were obtained from at least three independent biological replicates, except for the BT549 transwell assay which were done in duplicate. Multiple comparisons were performed using the one-way ANOVA test.

Zebrafish experiments

Zebrafish larvae of the AB strain were obtained through natural spawning of wild type adult fish. Our facility strictly complies with the relevant European (EU Directive, 2010/63/EU for animal experiments) and Italian (Legislative Decree No. 26/2014) laws, rules, and regulations, confirmed by the authorization issued by the municipality of Milan (PG 384983/2013). The procedures were carried out in accordance with the relevant guidelines and regulations.

48 h post-fertilization (hpf) anesthetized zebrafish larvae were microinjected in a fully randomized order with fluorescently labelled tumour cells as previously reported [59 and Reference therein]. Depending on the cells' size, ~600 to ~1200 cells were injected directly into the perivitelline space. Briefly, SUM159PT and BT549, both controls and YAl-KO cells, were labelled with Hoechst 33342 solution (ThermoFisher) and resuspended in complete medium at a final concentration of 300 cells/nl, and 2 to 4 nl/larva were inoculated in a complete blind manner. Injected live larvae were immediately observed under a fluorescent microscope to ensure the presence of labelled tumour cells. At 24 h post-injection (hpi) the larvae were anesthetized, individually placed on a microscope slide, and the number of extravasated cells was counted using an inverted

fluorescent microscope. At least two independent biological replicates were performed. The number (n) of larvae for each experiment is indicated in the Figure Legends.

RNA-seq datasets and molecular classifications

We downloaded the RSEM scaled count data for TCGA BRCA and STAD cohorts from the http://firebrowse.org/ web page. As of June 2022, we found RNA-seq data on 1093 primary tumours, 7 metastatic and 35 non-tumour tissues for BRCA, 415 primary tumours and 35 normal adjacent tissue samples for STAD. We retrieved the FASTQ files associated to 51 BRCA cell lines from CCLE, and to six SUM159 cell line samples from Yang et al. [60]. For RBPs analyses, we acquired expression data from KD, OE and KO experiments [54, 61–64]. All FASTQ files were downloaded using the SRA Explorer website (https://sra-explorer.info/), and the accession codes are in Supplementary Table S4. From the FASTQ files, we calculated mRNA expression with RSEM 1.3.1.

For BRCA samples, the molecular classification was retrieved from Fougner et al. [28], and used as the training set for the DeepCC tool to classify samples previously unclassified or classified as Normal-like. As for STAD, we used the classification detailed in a previous work [25]. Finally, the molecular classification for all the analysed BRCA cell lines was derived from Prat et al. [65] and Dai et al. [66].

Weighted gene co-expression network analysis

BRCA and STAD gene expression, as $\log_2(TPM + 1)$, were used as input for weighted gene co-expression network analysis, employing the WGCNA R package (version 1.70-3) [67]. NF-YA gene-level expression was replaced with NF-YAl and NF-YAs isoforms expression, together with the NF-YAl/NF-YAs ratio (NF-YAr). The soft-thresholding powers for BRCA and STAD were set to 4 and 7, respectively, as suggested by the function *pickSoftThreshold*. The parameters for the *blockwiseModules* function responsible for building the networks and find gene modules were: *networkType* = " signed", *maxBlockSize* = 30000, *minModuleSize* = 30, *reassignThreshold* = 0, and *mergeCutHeight* = 0.25.

Gene expression analysis

Differential gene expression analysis of RNA-seq data was performed using the R package DESeq2 (version 1.30.1) [68]. We used expression fold change (FC) to denote upregulation or downregulation in BT549 and SUM159 YAl-KO clones versus parental cell lines. Log_2FC, and the corresponding false discovery rate (FDR), were reported by the R package. FDR < 0.01 were set as inclusion criteria for DEG selection.

Gene ontology and gene sets enrichment analyses

We used KOBAS 3.0 (http://kobas.cbi.pku.edu.cn/anno_iden.php) for Gene Ontology (GO) terms enrichment analysis using the ENTREZ gene IDs as input and filtering out terms with FDR ≥ 0.01. MSigDB Hallmark gene sets enrichment analyses were conducted with the Enrichr website (https://maayanlab.cloud/Enrichr). For heatmap of pathways, only gene sets with a Combined Score >1 in all four experiments were included.

Z scores computation for signatures evaluation

We obtained Z scores from $\log_2$-transformed expression data (TPM) for each gene of the BRCA 7th gene module and BRCA-STAD ratio signatures. Then, a median Z score computed across all genes of the signatures was associated to each sample.

Analysis of clinical data

We retrieved TCGA BRCA and STAD patients Progression Free Interval -PFI- time records from the http://xena.ucsc.edu/ web page. We employed the Cutoff Finder tool [69] to find the optimal threshold for dichotomization of tumour samples based on the NF-YAr levels and PFI data. NF-YAr values >1 were set to 1, and survival analysis was performed according to the Kaplan–Meier analysis and a two-sided log-rank test.

scRNA-seq cell type annotation and deconvolution

scRNA-seq experiments conducted on breast and gastric cancers [37, 38] were first annotated with the R package scTyper (version 0.1.0) [70] choosing the Nearest Template Prediction -NTP- as cell typing method. We used the same markers for the cell types included in the original papers, when present. For BRCA cancer fractions, Basal, Luminal (A and B), and HER2-enriched tumours markers were included, plus an EMT signature from Taube et al. [71]. Likewise, we added an EMT signature in STAD cell typing process. Supplementary Table S5 contains the complete list of markers selected for the scTyper analysis.

We used the package SCDC (version 0.0.0.9000) [72] to lead a bulk RNA-seq composition deconvolution and a predicted proportion for each cell type was associated to TCGA BRCA and STAD tumour samples. These samples were then divided according to NF-YAr deciles or molecular subtypes, and cell types average proportions were computed for these so-defined groups. Within each subtype, samples with high NF-YAr (fourth quartile) were considered separately from all others.

Statistical analysis

All computational analyses were performed in the R programming environment (version 4.0.3), with the ggplot2, ggpubr, pheatmap, rstatix and tidyverse packages installed. Single comparisons between two groups were performed with the Wilcoxon rank-sum test (two-sided) or, in case of triplicates of two conditions, the Student *t*-test (two-sided).

DATA AVAILABILITY

Rna-seq datasets generated and analysed during this study are included in this published article and its Supplementary Information files. The raw data are available at Gene Expression Omnibus (GEO) as GSE208088.

REFERENCES

1. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, et al. Breast cancer. Nat Rev Dis Prim. 2019;5:1–31.
2. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.
3. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. 2010;12:R68.
4. Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol. 2007;8:R76.
5. Wang Y, Liu Z-P. Identifying biomarkers for breast cancer by gene regulatory network rewiring. BMC Bioinforma. 2022;22:308.
6. Islam MR, Abdulrazak LF, Alam MK, Paul BK, Ahmed K, Bui FM, et al. Identification of potential key genes and molecular mechanisms of medulloblastoma based on integrated bioinformatics approach. Biomed Res Int. 2022;2022:1776082.
7. Motalebzadeh J, Eskandari E. Transcription factors linked to the molecular signatures in the development of hepatocellular carcinoma on a cirrhotic background. Med Oncol. 2021;38:121.
8. Rocha D, García IA, González Montoro A, Llera A, Prato L, Girotti MR, et al. Pan-cancer molecular patterns and biological implications associated with a tumor-specific molecular signature. Cells. 2020;10:E45.
9. Meier T, Timm M, Montani M, Wilkens L. Gene networks and transcriptional regulators associated with liver cancer development and progression. BMC Med Genomics. 2021;14:41.
10. Hossain SMM, Halsana AA, Khatun L, Ray S, Mukhopadhyay A. Discovering key transcriptomic regulators in pancreatic ductal adenocarcinoma using dirichlet process gaussian mixture model. Sci Rep. 2021;11:7853.
11. Kallergi G, Tsintari V, Sfakianakis S, Bei E, Lagoudaki E, Koutsopoulos A, et al. The prognostic value of JUNB-positive CTCs in metastatic breast cancer: from bioinformatics to phenotypic characterization. Breast Cancer Res. 2019;21:86.
12. Takegoshi K, Honda M, Okada H, Takabatake R, Matsuzawa-Nagata N, Campbell JS, et al. Branched-chain amino acids prevent hepatic fibrosis and development of hepatocellular carcinoma in a non-alcoholic steatohepatitis mouse model. Oncotarget. 2017;8:18191–205.
13. Bie L-Y, Li D, Mu Y, Wang S, Chen B-B, Lyu H-F, et al. Analysis of cyclin E co-expression genes reveals nuclear transcription factor Y subunit alpha is an oncogene in gastric cancer. Chronic Dis Transl Med. 2018;5:44–52.
14. Cao B, Zhao Y, Zhang Z, Li H, Xing J, Guo S, et al. Gene regulatory network construction identified NFYA as a diffuse subtype-specific prognostic factor in gastric cancer. Int J Oncol. 2018;53:1857–68.
15. Pan Z, Li L, Fang Q, Qian Y, Zhang Y, Zhu J, et al. Integrated bioinformatics analysis of master regulators in anaplastic thyroid carcinoma. BioMed Res Int. 2019;2019:e9734576.
16. Cui H, Zhang M, Wang Y, Wang Y. NF-YC in glioma cell proliferation and tumor growth and its role as an independent predictor of patient survival. Neurosci Lett. 2016;631:40–9.

17. Kottorou AE, Antonacopoulou AG, Dimitrakopoulos F-ID, Tsamandas AC, Scopa CD, Petsas T, et al. Altered expression of NFY-C and RORA in colorectal adenocarcinomas. Acta Histochem. 2012;114:553–61.
18. Cicchillitti L, Corrado G, Carosi M, Dabrowska ME, Loria R, Falcioni R, et al. Prognostic role of NF-YA splicing isoforms and Lamin A status in low grade endometrial cancer. Oncotarget. 2017;8:7935–45.
19. Dolfini D, Andrioletti V, Mantovani R. Overexpression and alternative splicing of NF-YA in breast cancer. Sci Rep. 2019;9:12955–12955.
20. Bezzecchi E, Ronzio M, Dolfini D, Mantovani R. NF-YA Overexpression in Lung Cancer: LUSC. Genes (Basel). 2019;10:937.
21. Bezzecchi E, Ronzio M, Semeghini V, Andrioletti V, Mantovani R, Dolfini D. NF-YA Overexpression in Lung Cancer: LUAD. Genes. 2020;11:198.
22. Bezzecchi E, Ronzio M, Mantovani R, Dolfini D. NF-Y Overexpression in Liver Hepatocellular Carcinoma (HCC). Int J Mol Sci. 2020;21:9157.
23. Bezzecchi E, Bernardini A, Ronzio M, Miccolo C, Chiocca S, Dolfini D, et al. NF-Y Subunits Overexpression in HNSCC. Cancers. 2021;13:3019.
24. Belluti S, Semeghini V, Rigillo G, Ronzio M, Benati D, Torricelli F, et al. Alternative splicing of NF-YA promotes prostate cancer aggressiveness and represents a new molecular marker for clinical stratification of patients. J Exp Clin Cancer Res. 2021;40:362.
25. Gallo A, Ronzio M, Bezzecchi E, Mantovani R, Dolfini D. NF-Y subunits overexpression in gastric adenocarcinomas (STAD). Sci Rep. 2021;11:23764.
26. Zhang Y, Qian J, Gu C, Yang Y. Alternative splicing and cancer: a systematic review. Sig Transduct Target Ther. 2021;6:1–14.
27. Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer. 2016;16:413–30.
28. Fougner C, Bergholtz H, Norum JH, Sørlie T. Re-definition of claudin-low as a breast cancer phenotype. Nat Commun. 2020;11:1787.
29. Kardos J, Chai S, Mose LE, Selitsky SR, Krishnan B, Saito R, et al. Claudin-low bladder tumors are immune infiltrated and actively immune suppressed. JCI Insight. 2016;1:e85902.
30. Nishijima TF, Kardos J, Chai S, Smith CC, Bortone DS, Selitsky SR, et al. Molecular and clinical characterization of a claudin-low subtype of gastric cancer. JCO Precision Oncology 2017;1:1–10.
31. Libetti D, Bernardini A, Sertic S, Messina G, Dolfini D, Mantovani R. The Switch from NF-YAl to NF-YAs isoform impairs myotubes formation. Cells. 2020;9:789.
32. Flanagan L, Van Weelden K, Ammerman C, Ethier SP, Welsh J. SUM-159PT cells: a novel estrogen independent human breast cancer model system. Breast Cancer Res Treat. 1999;58:193–204.
33. Broad RV, Jones SJ, Teske MC, Wastall LM, Hanby AM, Thorne JL, et al. Inhibition of interferon-signalling halts cancer-associated fibroblast-dependent protection of breast cancer cells from chemotherapy. Br J Cancer. 2021;124:1110–20.
34. Fougner C, Bergholtz H, Kuiper R, Norum JH, Sørlie T. Claudin-low-like mouse mammary tumors show distinct transcriptomic patterns uncoupled from genomic drivers. Breast Cancer Res. 2019;21:85.
35. Scafoglio C, Ambrosino C, Cicatiello L, Altucci L, Ardovino M, Bontempo P, et al. Comparative gene expression profiling reveals partially overlapping but distinct genomic actions of different antiestrogens in human breast cancer cells. J Cell Biochem. 2006;98:1163–84.
36. Rogha M, Berjis N, Lajevardi SM, Alamdaran M, Hashemi SM. Identification of R249 Mutation in P53 gene in tumoral tissue of tongue cancer. Int J Prev Med. 2019;10:129.
37. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, et al. A single-cell and spatially resolved atlas of human breast cancers. Nat Genet. 2021;53:1334–47.
38. Kim J, Park C, Kim KH, Kim EH, Kim H, Woo JK, et al. Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage diversity and intratumoral heterogeneity. npj Precis Onc. 2022;6:1–11.
39. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. RBPDB: a database of RNA-binding specificities. Nucleic Acids Res. 2011;39:D301–308.
40. Benoit Bouvrette LP, Bovaird S, Blanchette M, Lécuyer E. oRNAment: a database of putative RNA binding protein target sites in the transcriptomes of model species. Nucleic Acids Res. 2020;48:D166–73.
41. Benatti P, Chiaramonte ML, Lorenzo M, Hartley JA, Hochhauser D, Gnesutta N, et al. NF-Y activates genes of metabolic pathways altered in cancer cells. Oncotarget. 2016;7:1633–50.
42. Poluri RTK, Paquette V, Allain ÉP, Lafront C, Joly-Beauparlant C, Weidmann C, et al. KLF5 and NFYA factors as novel regulators of prostate cancer cell metabolism. Endocr-Relat Cancer. 2021;28:257–71.
43. Li Y, Xiao X, Chen H, Chen Z, Hu K, Yin D. Transcription factor NFYA promotes G1/S cell cycle transition and cell proliferation by transactivating cyclin D1 and CDK4 in clear cell renal cell carcinoma. Am J Cancer Res. 2020;10:2446–63.
44. Dolfini D, Minuzzo M, Sertic S, Mantovani R. NF-YA overexpression protects from glutamine deprivation. Biochim Biophys Acta Mol Cell Res. 2020;1867:118571.
45. Yang W, Feng Q, Ma H, Lei D, Zheng P. NF-YA promotes the cell proliferation and tumorigenic properties by transcriptional activation of SOX2 in cervical cancer. J Cell Mol Med. 2020;24:12464–75.
46. Manni I, Caretti G, Artuso S, Gurtner A, Emiliozzi V, Sacchi A, et al. Posttranslational Regulation of NF-YA modulates NF-Y transcriptional activity. Mol Biol Cell. 2008;19:5203–13.
47. Mamat S, Ikeda J-I, Tian T, Wang Y, Luo W, Aozasa K, et al. Transcriptional regulation of aldehyde dehydrogenase 1A1 gene by alternative spliced forms of nuclear factor Y in tumorigenic population of endometrial adenocarcinoma. Genes Cancer. 2011;2:979–84.
48. Cappabianca L, Farina AR, Di Marcotullio L, Infante P, De Simone D, Sebastiano M, et al. Discovery, characterization and potential roles of a novel NF-YAx splice variant in human neuroblastoma. J Exp Clin Cancer Res. 2019;38:482.
49. Bernardini A, Gallo A, Gnesutta N, Dolfini D, Mantovani R. Phylogeny of NF-YA trans-activation splicing isoforms in vertebrate evolution. Genomics. 2022;114:110390.
50. Silvio di A, Imbriano C, Mantovani R. Dissection of the NF-Y transcriptional activation potential. Nucleic Acids Res. 1999;27:2578–84.
51. Mhaidly R, Mechta-Grigoriou F. Role of cancer-associated fibroblast subpopulations in immune infiltration, as a new means of treatment in cancer. Immunol Rev. 2021;302:259–72.
52. El Marabti E, Younis I. The cancer spliceome: reprograming of alternative splicing in cancer. Front Mol Biosci. 2018;5:80.
53. Venables JP, Brosseau J-P, Gadea G, Klinck R, Prinos P, Beaulieu J-F, et al. RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. Mol Cell Biol. 2013;33:396–405.
54. Yang Y, Park JW, Bebee TW, Warzecha CC, Guo Y, Shang X, et al. Determination of a comprehensive alternative splicing regulatory network and combinatorial regulation by key factors during the epithelial-to-mesenchymal transition. Mol Cell Biol. 2016;36:1704–19.
55. Cieply B, Park JW, Nakauka-Ddamba A, Bebee TW, Guo Y, Shang X, et al. Multiphasic and dynamic changes in alternative splicing during induction of pluripotency are coordinated by numerous RNA-binding proteins. Cell Rep. 2016;15:247–55.
56. Vanharanta S, Marney CB, Shu W, Valiente M, Zou Y, Mele A, et al. Loss of the multifunctional RNA-binding protein RBM47 as a source of selectable metastatic traits in breast cancer. eLife. 2014;3:e02734.
57. Dutta A, Yang Y, Le BT, Zhang Y, Abdel-Wahab O, Zang C, et al. U2af1 is required for survival and function of hematopoietic stem/progenitor cells. Leukemia. 2021;35:2382–98.
58. Domashenko AD, Danet-Desnoyers G, Aron A, Carroll MP, Emerson SG. TAT-mediated transduction of NF-Ya peptide induces the ex vivo proliferation and engraftment potential of human hematopoietic progenitor cells. Blood. 2010;116:2676–83.
59. Rebelo de Almeida C, Mendes RV, Pezzarossa A, Gago J, Carvalho C, Alves A, et al. Zebrafish xenografts as a fast screening platform for bevacizumab cancer therapy. Commun Biol. 2020;3:1–13.
60. Yang Y, Lu H, Chen C, Lyu Y, Cole RN, Semenza GL. HIF-1 interacts with TRIM28 and DNA-PK to release paused RNA polymerase II and activate target gene transcription in response to hypoxia. Nat Commun. 2022;13:316.
61. Mérien A, Tahraoui-Bories J, Cailleret M, Dupont J-B, Leteur C, Polentes J, et al. CRISPR gene editing in pluripotent stem cells reveals the function of MBNL proteins during human in vitro myogenesis. Hum Mol Genet. 2021;31:41–56.
62. Piqué L, Martinez de Paz A, Piñeyro D, Martínez-Cardús A, Castro de Moura M, Llinàs-Arias P, et al. Epigenetic inactivation of the splicing RNA-binding protein CELF2 in human breast cancer. Oncogene. 2019;38:7106–12.
63. Chen X, Liu Y, Xu C, Ba L, Liu Z, Li X, et al. QKI is a critical pre-mRNA alternative splicing regulator of cardiac myofibrillogenesis and contractile function. Nat Commun. 2021;12:89.
64. Zhou D, Couture S, Scott MS, Abou Elela S. RBFOX2 alters splicing outcome in distinct binding modes with multiple protein partners. Nucleic Acids Res. 2021;49:8370–83.
65. Prat A, Karginova O, Parker JS, Fan C, He X, Bixby L, et al. Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes. Breast Cancer Res Treat. 2013;142:237–55.
66. Dai X, Cheng H, Bai Z, Li J. Breast cancer cell line classification and its relevance with breast tumor subtyping. J Cancer. 2017;8:3131–41.
67. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinforma. 2008;9:559.
68. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550–550.
69. Budczies J, Klauschen F, Sinn BV, Győrffy B, Schmitt WD, Darb-Esfahani S, et al. Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. PLoS One. 2012;7:e51862.
70. Choi J-H, In Kim H, Woo HG. scTyper: a comprehensive pipeline for the cell typing analysis of single-cell RNA-seq data. BMC Bioinforma. 2020;21:342.
71. Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J, et al. Core epithelial-to-mesenchymal transition interactome gene-expression signature is

associated with claudin-low and metaplastic breast cancer subtypes. Proc Natl Acad Sci. 2010;107:15449–54.
72. Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Brief Bioinform. 2021;22:416–27.

ACKNOWLEDGEMENTS

The Authors thank Prof. Gabellini at Division of Genetics and Cell Biology, IRCCS Ospedale San Raffaele for providing the myc-mRBFOX2 and control plasmids, Nadia Zaffaroni at Fondazione IRCSS Istituto Nazionale dei Tumori for providing the ATCC breast cancer cell lines. The Authors would like to acknowledge the Genomic Unit at Istituto Europeo di Oncologia for NGS experiments. The authors acknowledge the support of the APC central fund of the University of Milan. This work was supported by PNRR M4C2-Investimento 1.4 -CN00000041-PNRR_CN3RNA_SPOKE2 to R.M. and Ministero della Salute GR-2013-02355625 to D.D.

AUTHOR CONTRIBUTIONS

ML, AG, MR, CC, and AG contributed to the production of data. LDG and DD designed the study and assisted in report preparation. RM and DD supervised the study and wrote the manuscript. All Authors have read and approved the final manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41419-023-05591-9.

Correspondence and requests for materials should be addressed to Diletta Dolfini.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.1.3. Single cell RNA-seq analyses of breast and gastric Claudin^{low} cell fraction

As a corollary analysis of STAD and BRCA sample cell type deconvolution, I further characterized cells associated with the Claudin^{low} subtype. Previously I isolated via marker-based cell typing a cluster of cells predicted as Claudin^{low} in both tumor types, starting from the two large-scale scRNA-seq experiments[2] I mentioned in the former section. These cells were clearly distinct from epithelial cancer cells clusters in dimensionality reduction plots, showed minimal to negligible expression of epithelial tumor markers (*EPCAM*, *KRT8/18/19*) and cancer-associated fibroblasts markers (*FAP*, *FN1*). Intriguingly, they also displayed a robust correlation with the 158-gene NF-YA$_{\text{Ratio}}$ signature.

Taking into account the intrinsic heterogeneity of gene copy number variation (CNV) rates in Claudin^{low} tumors, I proceeded to estimate the proportion of cancer cells within within these identified clusters. This estimation was accomplished using the inferCNV tool[3], which operates at the single-cell level to identify chromosomal amplifications and deletions, serving as a proxy for tumor cells' deranged control of genome stability. inferCNV takes as input the expression profiles of single cell clusters and employs a statistical framework to detect genomic aberrations signaling CNVs through the use of a reference baseline. In this case, I used normal epithelial cells as negative control, and found comparable percentages between BRCA and STAD predicted Claudin^{low} cells and the corresponding epithelial cancer clusters.

In addition, this outcome substantiated the assumption that Claudin^{low} clusters incorporate a proportion of highly invasive cancer cells, and is not solely composed of stromal/immune cells composing the tumor microenvironment (**Figure 3.1**).

[2] >50k cells for BRCA, and ∼30k for STAD.

[3] https://github.com/broadinstitute/inferCNV.

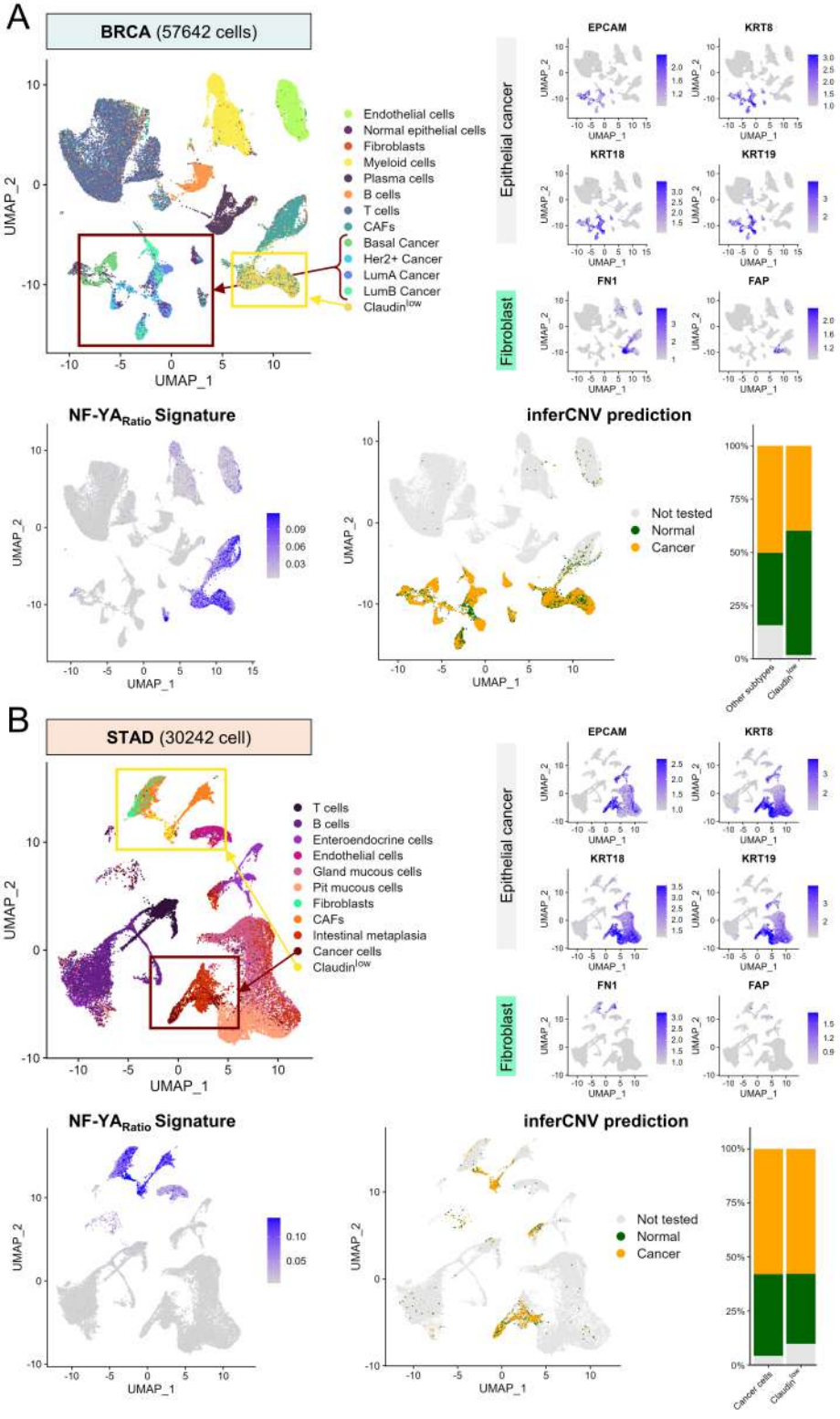


Figure 3.1: Claudin^{low}-directed scRNA-seq analyses in BRCA and STAD.
Figure caption is on the following page.

Figure 3.1: Claudin^{low}-directed scRNA-seq analyses in BRCA and STAD.
A. ***Top Left Panel****: UMAP plot showing the results of the marker-based cell typing procedure I conducted employing the BRCA scRNA-seq expression atlas data[124]. Cells portrayed in dark yellow were labelled as Claudin^{low} based on the corresponding BRCA signature[119].* ***Top Right Panel****: normalized expression of the epithelial markers* EPCAM, KRT8, KRT18, and KRT19*, as well as the fibroblast markers* FN1 and FAP*.* ***Bottom Left Panel****: UMAP representation of per-cell correlation scores to the NF-YA$_{Ratio}$ signature, described in section* ***3.1.2****.* ***Bottom Right Panel****: Cancer cell assignment based on inferCNV results, adopting normal epithelial cells as the reference cluster.* ***B****. Same as* ***A****, but this time the analysis was conducted on a STAD scRNA-seq dataset[245]. For cell typing, I employed the Claudin^{low} signature proposed by Nishijima* et al.*[126], whereas gland/pit mucous cells were selected as inferCNV reference group.*

3.1.4. Identification of Claudin^{low}-specific splicing program

IDENTIFYING specific splicing events unique to the Claudin^{low} subtype of gastric cancer holds potential for streamlined diagnostics and the development of targeted therapeutic approaches. This is particularly significant in this cancer cohort, where limited molecular tools are available for accurate differential diagnoses. Thus, we identified a preliminary collection of Claudin^{low} specific events by examining the TCGA STAD RNA-seq data using computational tools designed to determine differential transcripts or exons usage (DTU and DEU, respectively) between distinct conditions, such as subtypes.

Using differential transcript expression analysis alone may not be sufficient to get a precise picture of changes in transcript usage, as it overlooks the gene of origin. Hence, an approach like this would expend statistical power in identifying many predominant transcripts of differentially expressed genes. To address this limitation, we examined the relative proportions of isoforms originating from genes, allowing us to identify differentially used transcripts even when total gene expression remains stable.

DTUrtle was elected as our DTU calling tool[248], and applied to STAD samples by comparing the recently identified Claudin^{low} tumor samples with all other molecular subtypes from the ACRG classification (EMT, MSI, MSS;TP53^{-}, and MSS;TP53^{+}) in a pairwise manner. The intersection of the results produced an unfiltered collection of 1203 differentially used isoforms originating from 740 genes, setting the overall false discovery rate threshold to 0.05.

I curated the exon differential usage analysis, where I instead employed satuRn, a method suitable for both isoform-level and exon-level approaches [249]. Comparing once again Claudin^{low} to the other subtypes, the results overlap consisted in over 9000 differentially expressed exons from 3803 distinct genes, with the same statistical threshold as the DTU analysis. Then, out of 478 unique genes identified as common between exon- and isoform-level analyses, I selectively retained only the instances where isoforms with increased usage in Claudin^{low} included differentially used exons.

This integration yielded 363 genes, with functional enrichment highlighting the MSigDB Hallmarks *Myogenesis*, *Epithelial to Mesenchymal Transition*, and *Mitotic Spindle*. Notably, this analysis revealed Claudin^{low}-specific overexpression of isoforms encoded by six genes involved in regulating critical cellular functions, which are considered cancer hallmarks. In particular, I identified distinct splicing patterns in four of the seven genes known to produce isoforms that promote invasiveness and metastasis, namely *ENAH, FN1, TNC*, and *FGFR2*[51] (**Figure 3.2**).

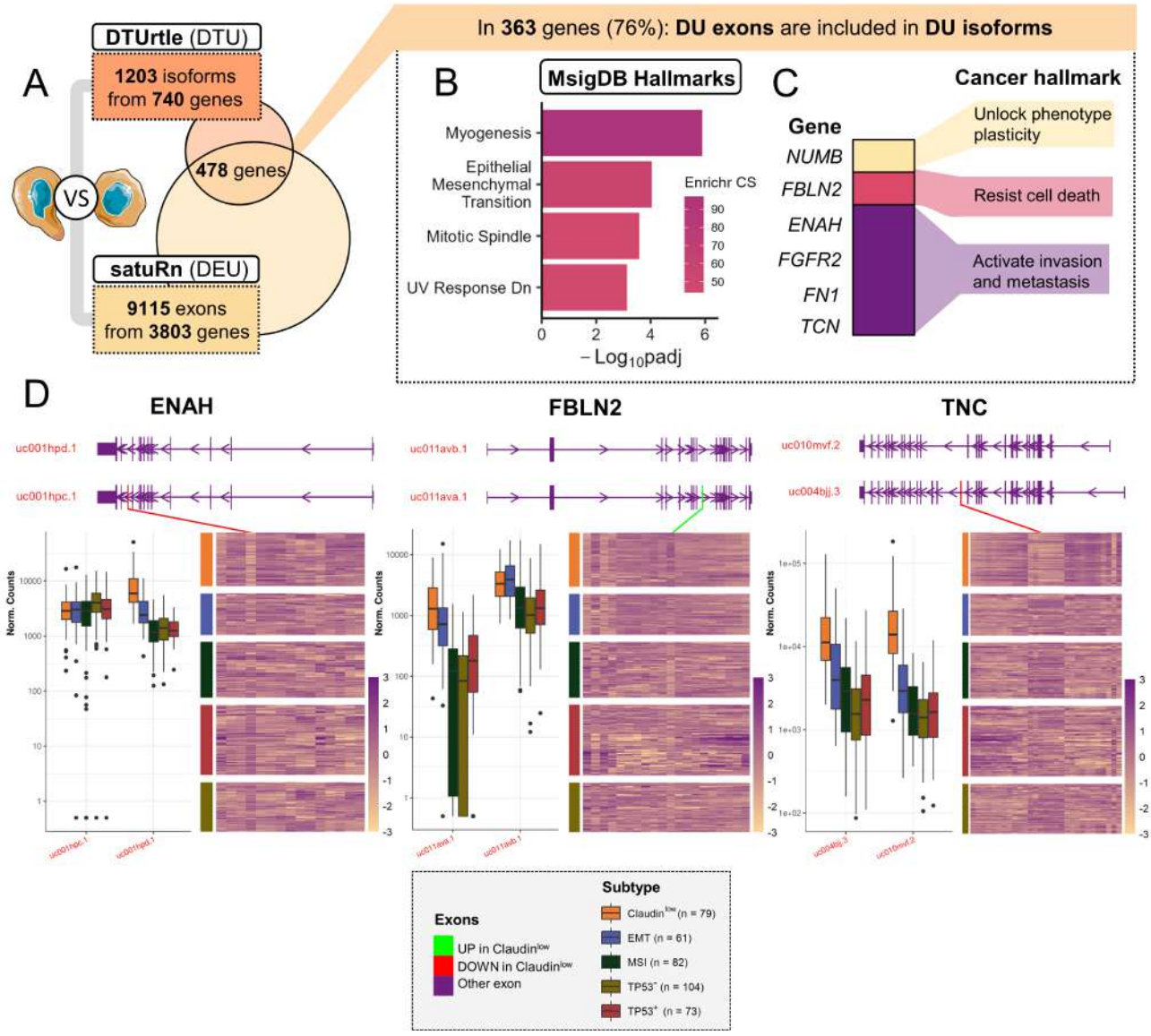


Figure 3.2: Preliminary analysis of Claudin^{low}-specific splicing events.
A. *Schematic representation of DTUrtle and satuRn analyses.* ***B***. *Enrichment in MSigDB Hallmarks of common genes where differentially used exons (satuRn) where included in differentially used isoforms (DTUrtle).* ***C***. *Genes found within preliminary data whose isoforms have been associated to cancer hallmarks.* ***D***. *Transcripts structure (**Top panels**), isoform-level (**Bottom Left**), and exon-level expression (**Bottom Right**) of significant splicing events in* ENAH, FBLN2 *and* TNC. *Parts of the figure were drawn by using pictures from Servier Medical Art.*

3.2. Phylogeny of NF-YA locus

3.2.1. NF-YA across vertebrates' evolution

As part of a broader phylogenic analysis focused on NF-YA intron-exon structure, I quantified exon-level transcription at the corresponding locus in *Petromyzon marinus* (lamprey), *Danio rerio* (zebrafish), *Xenopus laevis* (frog), and *Mus musculus* (mouse). This evaluation spanned multiple embryonic stages and brain, liver, and muscle adult tissues.

Generally, exon-3 inclusion in the final mRNA was significantly decreased during development compared to the mean expression of all other exons in lamprey, zebrafish (where *nfya* has a paralog, *nfyal*) and frog. In contrast, mouse embryo data depicted increased exon-3 expression at the zygote stage, blastulation, and in mesoderm upon gastrulation (E7.5 stage). In adult tissues, exon-3 levels largely mirrored the ones of all others, representing sustained NF-YAl expression, except for zebrafish and frog liver samples. This confirmed the assumption that NF-YAs predominates during embryogenesis, while NF-YAl splicing is linked to differentiation, as well as to mesoderm fate determination (**Figure 3.3**).

Furthermore, in the same species selection, exon seven N-term consistently underwent alternative splicing, with its shorter version frequently associating with NF-YAs in embryonic samples. Additionally, the mean expression of all exons minus exon-3 and exon-5, in a conformation resembling NF-YAx[4], was identical to the one representing NF-YAs, implying that NF-YAx is not produced within the sample pool considered for the analysis (**Figure 3.4**).

[4]An isoform described in 2019 in human neuroblastoma and mouse embryo[195].

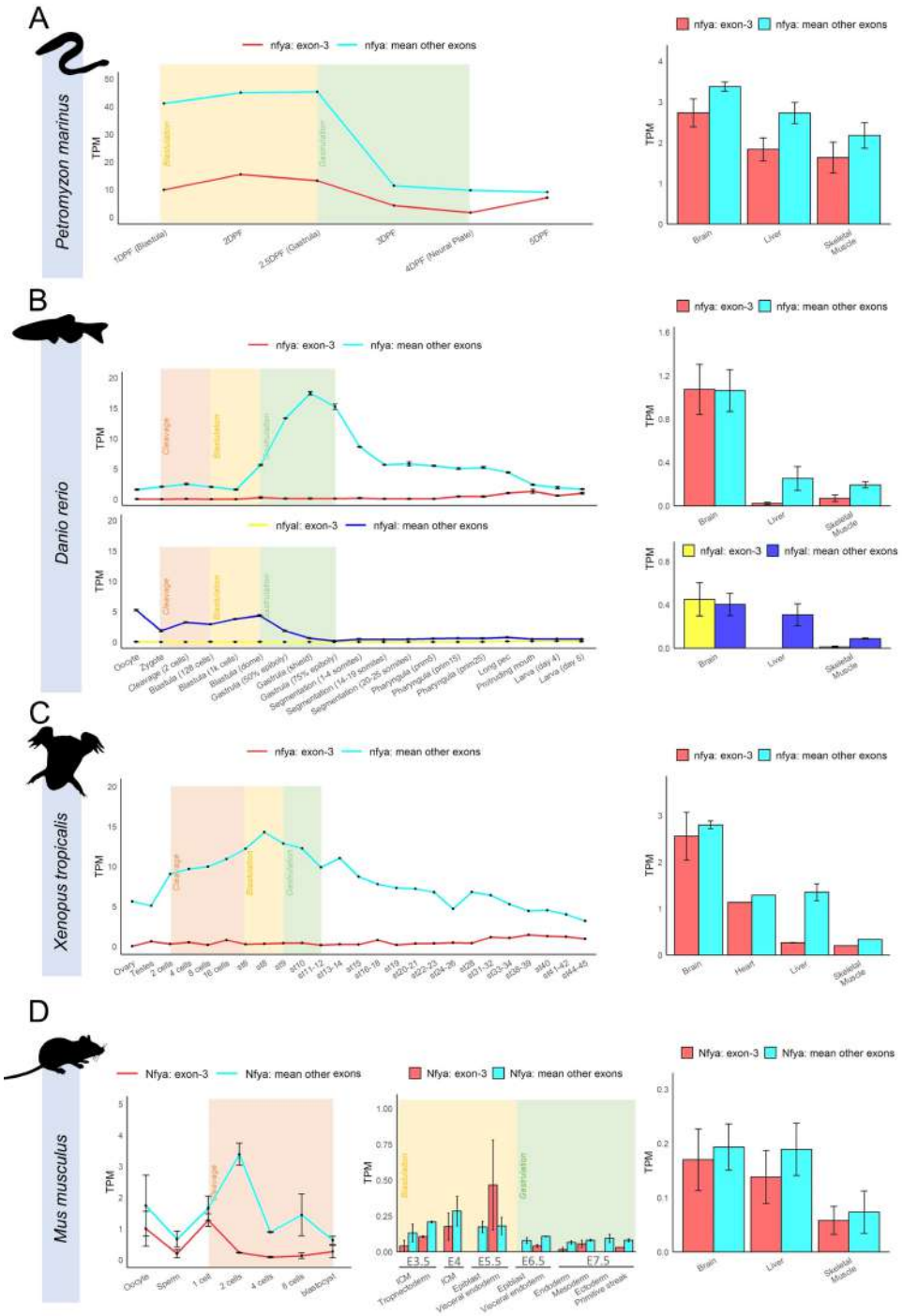


Figure 3.3: NF-YA exon-3 expression patterns across vertebrates.
A. *Expression of the exon-3 of sea lamprey (*Petromyzon marinus*) NFYA ortholog compared to the expression mean of all other exons of the gene, as measured in six embryonic stages (**Left Panel**) as well as in the adult brain, liver, and skeletal muscle (**Right Panel**).* ***B***. *Expression of exon-3 of zebrafish (*Danio rerio*)* nfya *(**Top panels**) and* nfyal *(**Bottom panels**) genes compared to the expression mean of all other exons of the gene, as measured in multiple developmental timeframes, spanning from oocyte to larva (**Left**), as well as in the adult brain, liver, and skeletal muscle (**Right**).* ***C***. *Same as* ***B****, except the expression of* Xenopus tropicalis nfya *exon-3 is depicted.* ***D***. *Expression of* Mus musculus Nfya *exon-3 compared to the expression mean of all other exons of the gene, as measured in early development (**Left Panel**), in the embryonic tissues of five distinct stages (**Central Panel**), as well as in the adult brain, liver, and skeletal muscle (**Right Panel**). Error bars indicate the standard error of the mean, and expression levels are represented by Transcripts Per Million (TPM).*

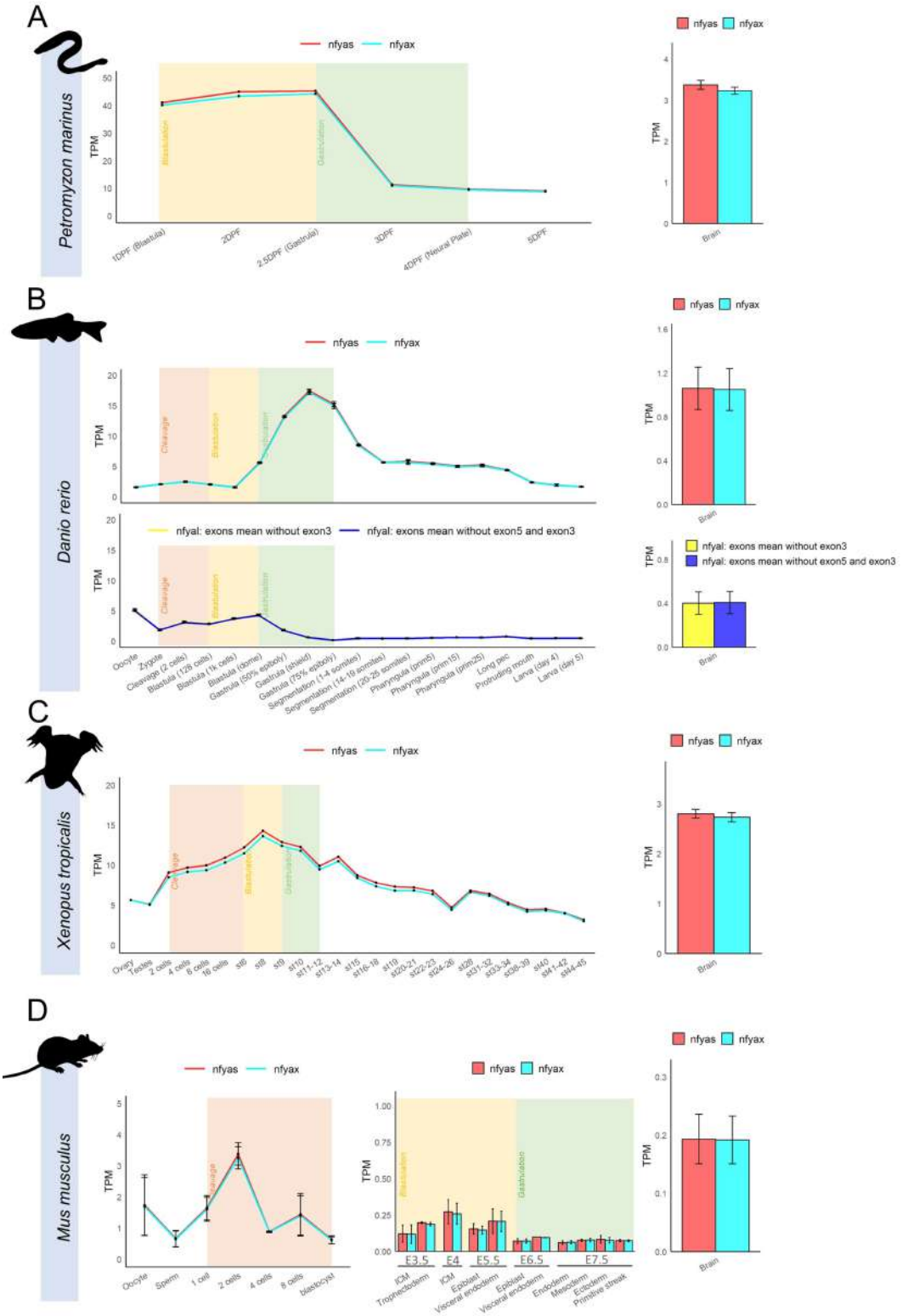


Figure 3.4: NF-YAx expression patterns across vertebrates.
Expression of two alternative splicing configurations representing NF-YAs (expression mean of all exons except exon-3) and NF-YAx (expression mean of all exons except exon-3 and exon-5), as measured in the adult brain and in the same embryonic stages of ***Figure 3.3****. Data are from* ***A*** *sea lamprey (*Petromyzon marinus*);* ***B*** *zebrafish (*Danio rerio*), both paralogs nfya and nfyal are shown;* ***C*** Xenopus tropicalis*;* ***D*** Mus musculus*. Error bars indicate the standard error of the mean, and expression levels are represented by Transcripts Per Million (TPM).*

Following the revision process, we decided to extend the expression analysis to 17 vertebrate species picked from all vertebrate classes, this time only considering the adult tissues. The final picture (**Figure 7** in the following paper) depicted the mapped read coverage at three NF-YA regions: exon-3, exon-7N, and exon-7C, a previously undocumented six amino acids stretch prone to skipping in teleosts (bony fishes). NF-YAl was expressed in the brain of all species considered, whereas in muscle NF-YAs was the prevalent isoform in *Ambystoma mexicanum* (axolotl), as indicated by the absence of exon-3 coverage. In the liver, exon-3 reads were detected solely in *Felis catus* (cat) and *Sus scrofa* (pig), aligning with NF-YAs being the dominant isoform.

These findings largely reiterated what I described in the original figure, and agreed with the established knowledge about NF-YA main isoforms expression in human and mice. Exon-7N was generally included in the mature mRNA in both brain and muscle tissues, albeit presenting a reduced expression in *Alligator sinensis* (Chinese alligator) and frog, while in liver was prevalent in many species. Axolotl was still an exception to the rule, since no evidence of the alternative splicing site at exon-7 N-term was detected. Of note, no direct correlation between the expression of exon-3 and exon-7N was detectable in the tissues considered. Exon-7C in fish species exhibited a variable expression pattern across the three tissues, with *Gadus morua* (Atlantic cod) completely missing this variant.

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.

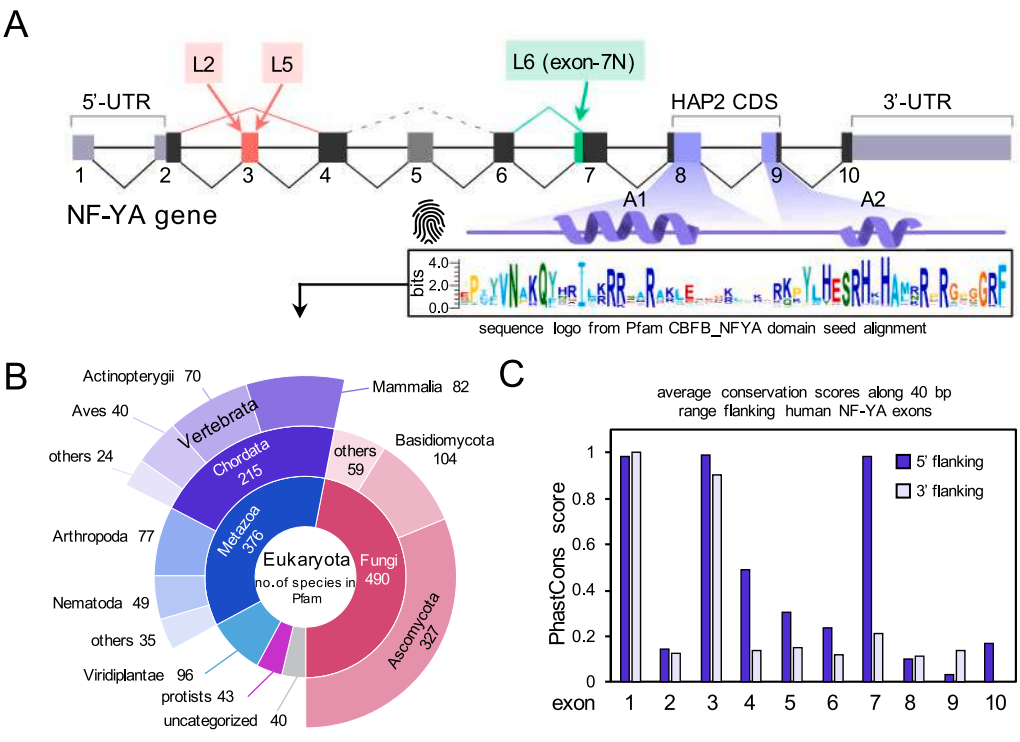


Fig. 1. *NF-YA* genes in all major eukaryotic taxa.
A. Scheme of the human *NF-YA* gene structure. Known alternative splicing events are indicated. The conserved HAP2 domain is highlighted in violet and the corresponding protein seed alignment from distantly related species in Pfam database is represented 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

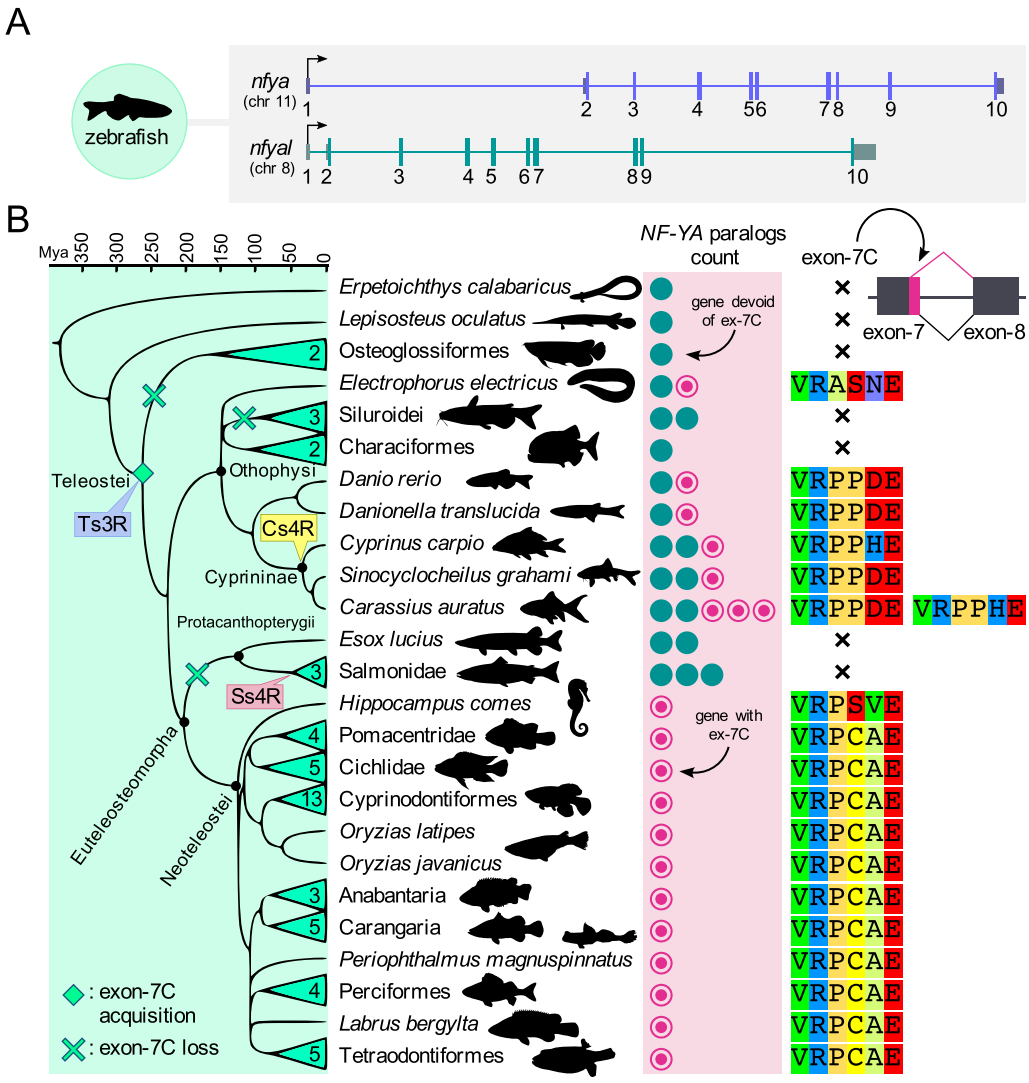


Fig. 2. *NF-YA* paralogs and exon-7C in teleost fishes.
A. Gene structure of the two *NF-YA* paralogs found in zebrafish. B. Phylogenetic time-tree of 63 species of ray-finned fishes (left panel) checked for the number of *NF-YA* paralogous genes (central red panel) and the exon-7C extension sequence (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-

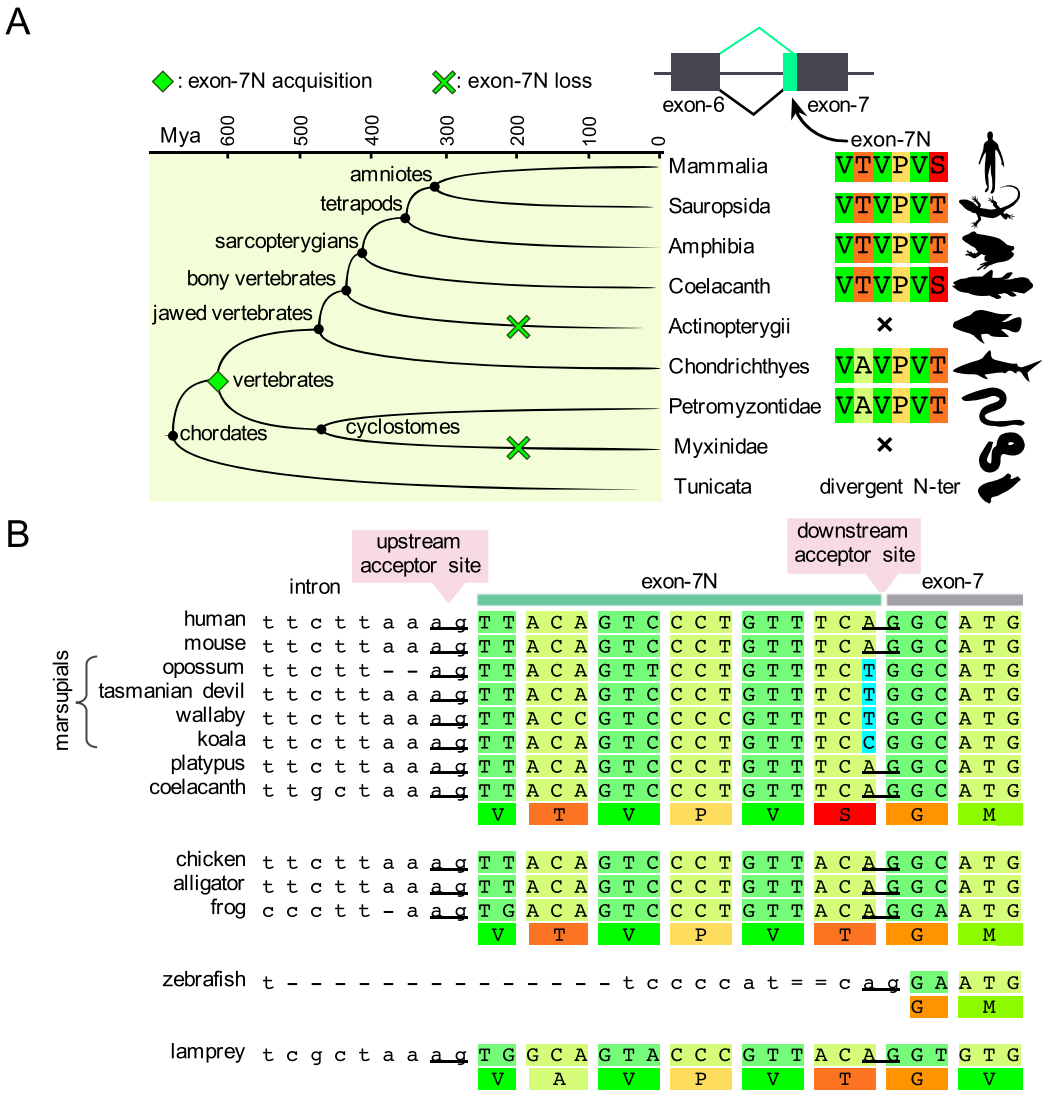


Fig. 3. Exon-7N conservation and phylogenetic distribution in vertebrates.
A. Hypothetical phylogeny of exon-7N among the main groups of vertebrates. For each group, exon-7N protein sequence is reported. The time-tree depicts the estimated time divergence among the animal groups considered. Tunicates 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

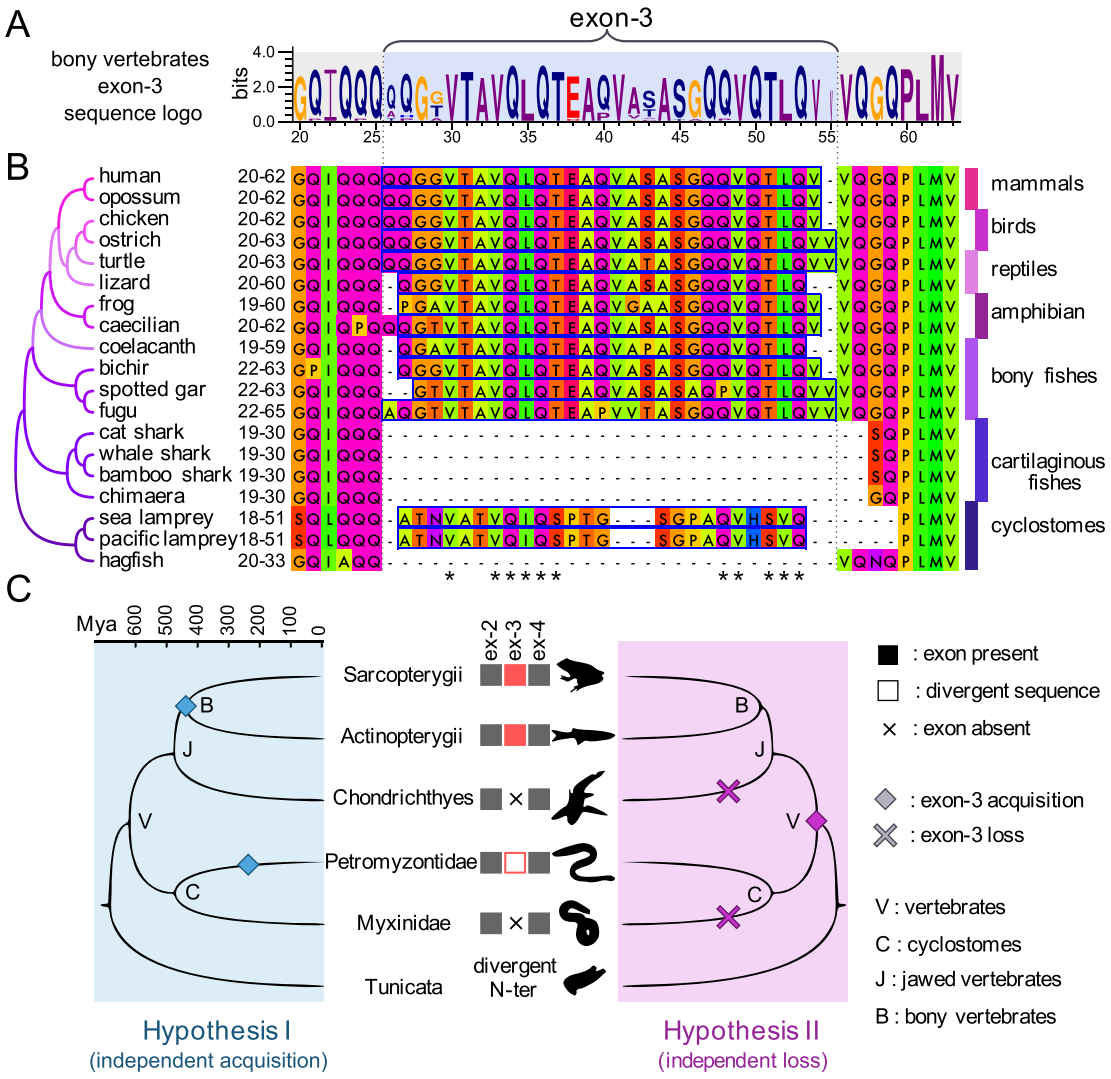


Fig. 4. Exon-3 conservation and phylogenetic distribution in vertebrates.
A. Sequence logo of NF-YA exon-3 and adjacent regions derived from a MSA of 15 species representative of the bony vertebrate groups possessing canonical exon-3. Species belonging to the following groups were chosen: mammals, sauropsids, 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*

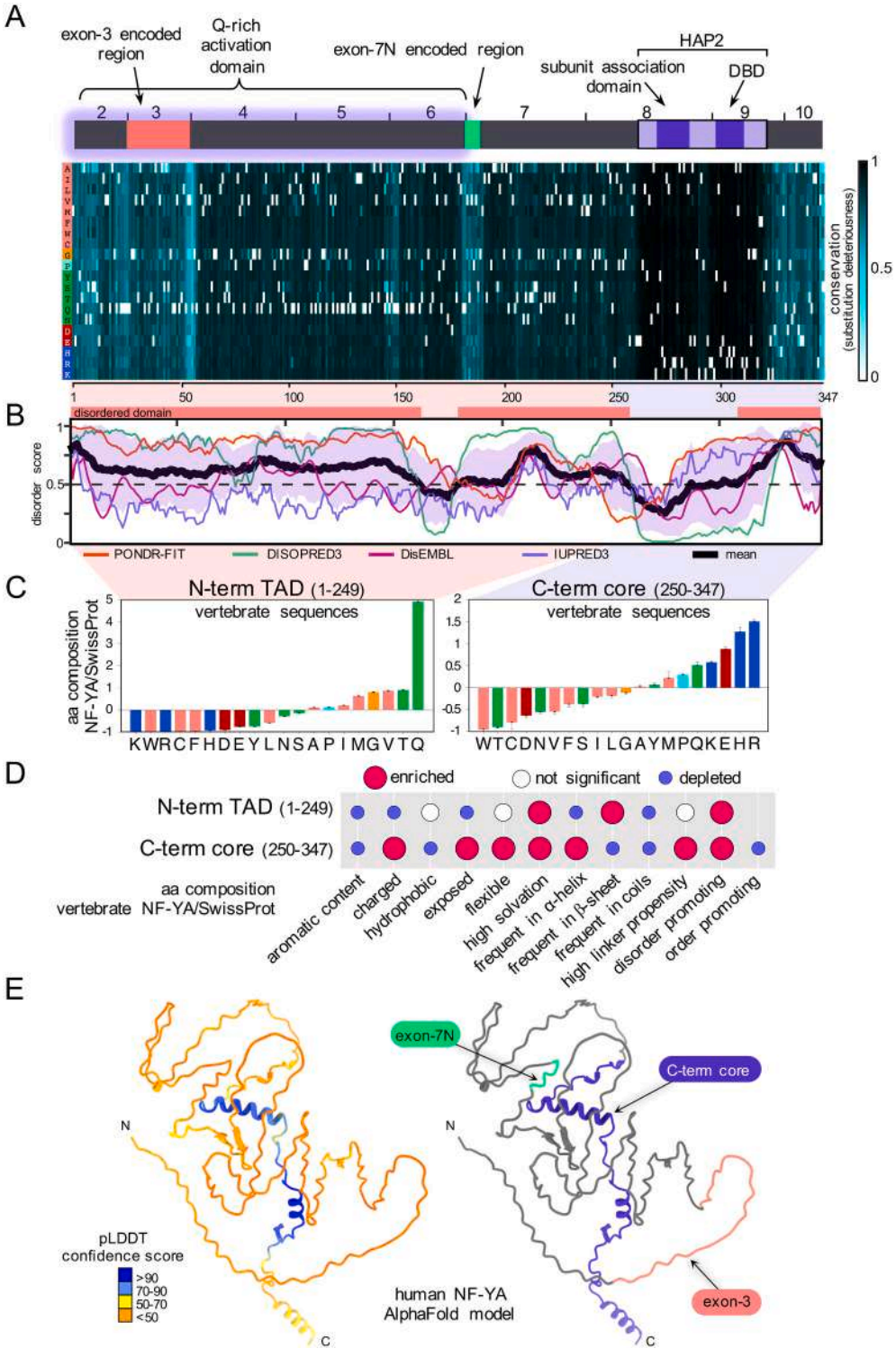


Fig. 5. NF-YA domains conservation, disorder and sequence composition.
A. NF-YA protein annotation and conservation. Exon boundaries, regions affected by AS and functional domains are indicated. The heatmap shows the deleteriousness scores computed for every possible substitution at each position along 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

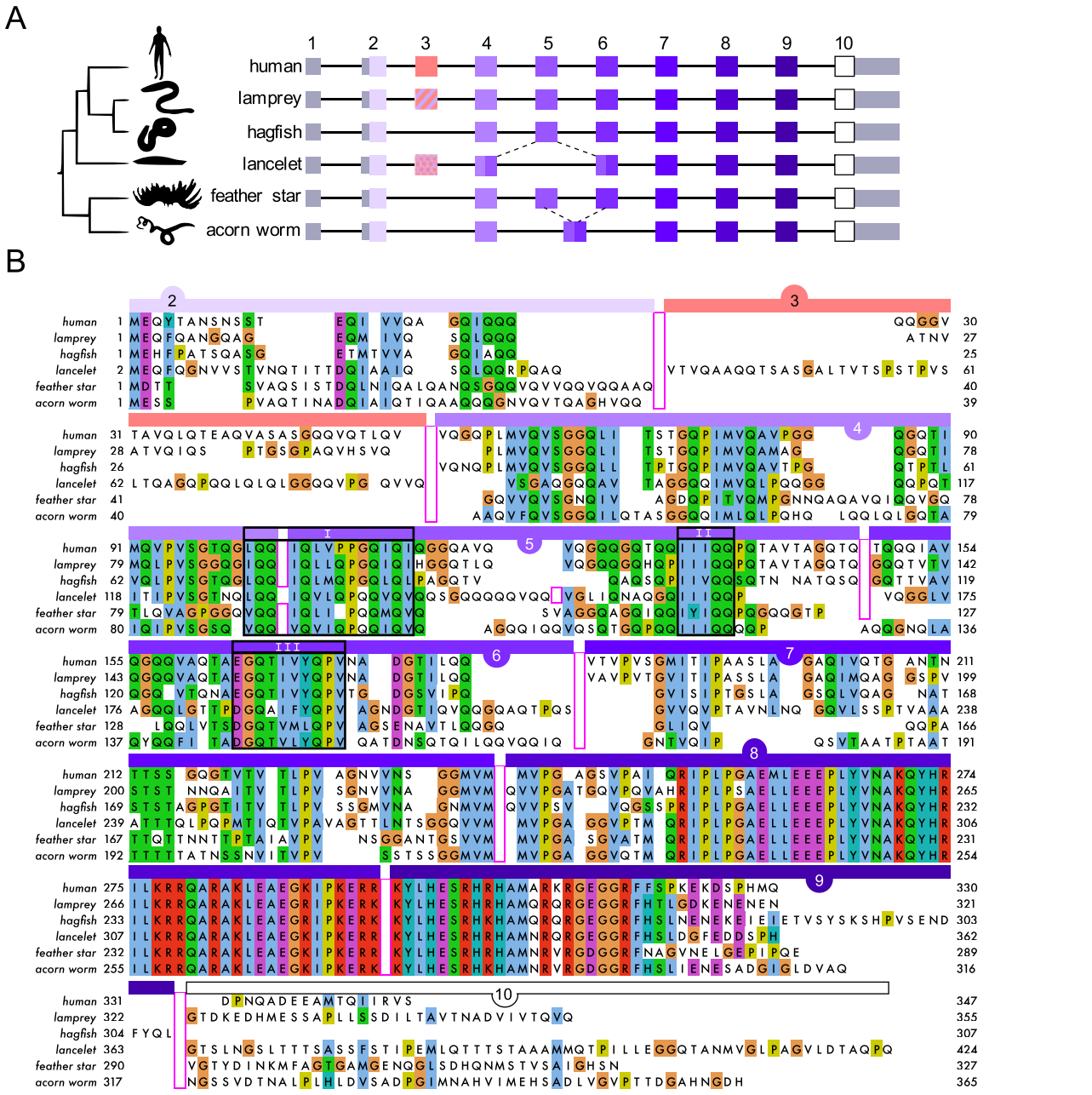


Fig. 6. NF-YA gene structure and conservation blocks in other deuterostomes.
A. Intron-exon structure of *NF-YA* gene orthologs in different species. Exon numbering is referred to human gene. Exons for which a recognizable amino acid signature could be traced among the considered species have the same 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.

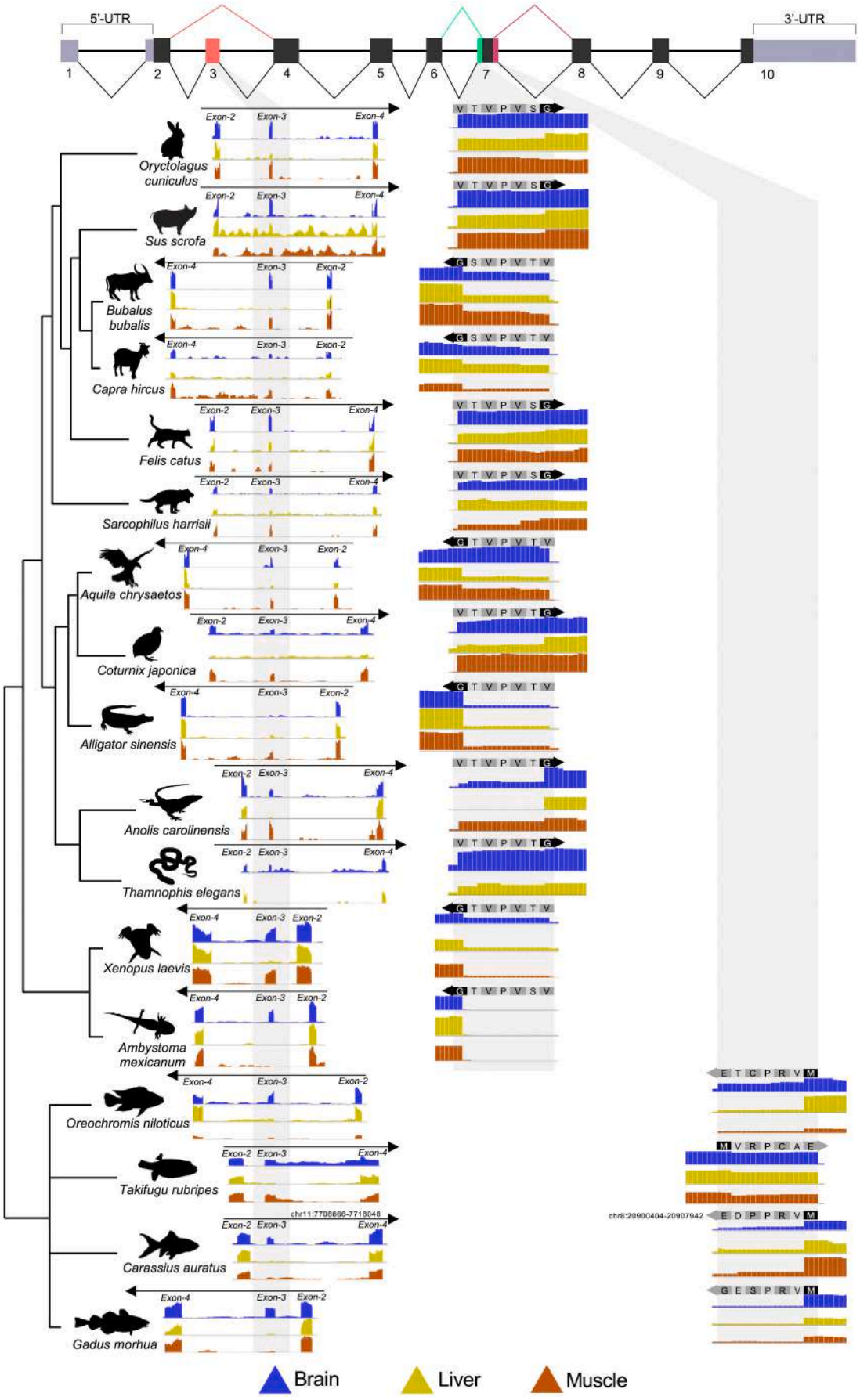
Another important point is that AS of NF-YA starts with vertebrates, since lancelet lacks exon-3 splicing, as well as variability at exon-7C or exon-7N. There is ample evidence of AS in amphioxus, including in TF genes, as documented for *Pax2/5/8* [52,53] and *Hif1a* [54]: this makes the lack of AS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

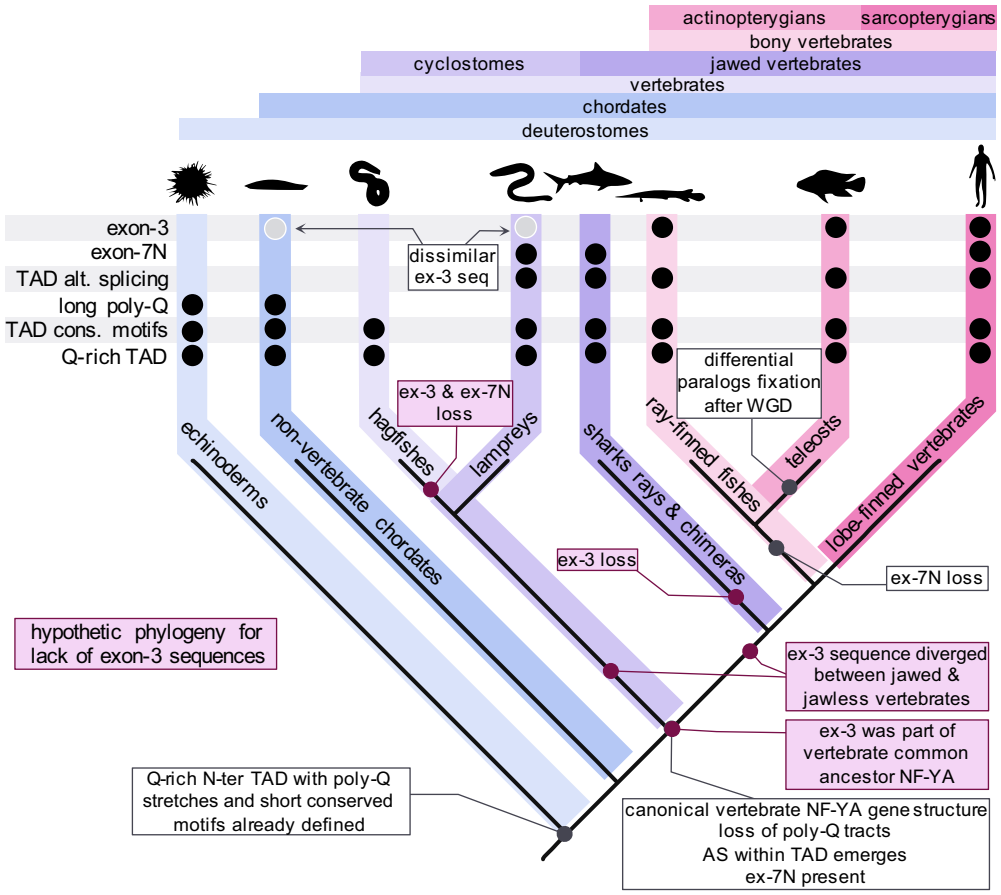


Fig. 8. Summary of NF-YA evolution in vertebrates.
The major events characterizing NF-YA transactivation domain (TAD) evolution in vertebrates are summarized. Lancelet and echinoderms are used as outgroups to infer ancestral patterns. The top panel summarizes the presence of different protein features 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.

3.2.2. NF-YA isoforms with alternative splicing of exon-5 in Aves

The multiple protein sequence alignment (MSA) designed as part of the phylogenetic assessment of NF-YA splicing unveiled that a few entries skipped exon-5, with or without the concomitant exclusion of exon-3 in the final transcripts. These sequences belonged to several *Sauropsida* species, a clade gathering birds and reptiles.

Intrigued by this observation, I collected all currently available NF-YA protein sequences[5] for this vertebrate group, generating another MSA which depicted several bird species missing the exon-5 encoded region, another portion of the transcription activation domain of the protein (TAD). This led to the identification of a new isoform, which was termed NF-YAg. Additionally, some entries lacked exon-3, indicating the expression of NF-YAx. Compared to the main isoforms, NF-YAg and NF-YAx appeared to be infrequent variants, with NF-YAl being the prevalent isoform in most *Sauropsida* species with a limited sequence count.

Further, I studied NF-YA intron-exon boundaries conservation in 77 *Sauropsida* species, assessing the average negative selection probability for the 40 base pairs up- and downstream of each exonic region. Exon-3 and Exon-5 edges in *Aves* (birds) and *Archosauria* (alligators, crocodiles, turtles, and tortoises) exhibited higher average conservation scores compared to other exons. In opposition, in *Squamata* (lizard and snakes) only exon-3 showed above-average conservation levels. Focusing on distal intronic regions, I reported a correlation between two *Aves*-specific conserved sequences located in the introns surrounding exon-5 and the binding site motifs of various RNA-binding proteins involved in exon skipping, such as MBNL1, SF1, HNRNPA1, and members of the SRSF family.

Lastly, through exon-level analysis of RNA-seq data I found that NF-YAg and NF-YAx expression was widespread in *Aves* adult liver and muscle tissues, but not in *Reptilia*. The two rare isoforms prevailed over NF-YAl and NF-YAs during the early stages of embryogenesis in chicken (*Gallus gallus*) and were also detected in full-length long-read libraries from the same species, spanning from embryo day 1 to day 7. Overall, the data suggested that both NF-YAg and NF-YAx are consistently included in the *Aves* parental gene alternative splicing repertoire.

[5]InterPro database entry: IPR001289

Genomics 115 (2023) 110694

Contents lists available at ScienceDirect

Genomics

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

NF-YA isoforms with alternative splicing of exon-5 in *Aves*

A. Gallo, D. Dolfini, A. Bernardini [1], N. Gnesutta, R. Mantovani [*]

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

ARTICLE INFO

Keywords:
Transcription factors
Alternative splicing
Transactivation domain
Intrinsically disordered protein
Birds

ABSTRACT

NF-YA, the regulatory subunit of the trimeric CCAAT-binding transcription factor NF-Y, is present in vertebrates in two major alternative spliced isoforms: NF-YAl and NF-YAs, differing for the presence of exon-3. NF-YAx, a third isoform without exon-3/-5, was reported only in human neuronal cells and tumors. These events affect the Trans-Activation Domain. We provide here evidence for the expression of NF-YAx and for the existence of a new isoform, NF-YAg, skipping only exon-5. These isoforms are abundant in *Aves*, but not in reptiles, and are the prevalent transcripts in the initial phases of embryo development in chicken. Finally, we analyzed NF-YAg and NF-YAx amino acid sequence using AlphaFold: absence of exon-5 denotes a global reduction of β-stranded elements, while removal of the disordered exon-3 sequence has limited effects on TAD architecture.

These data identify an expanded program of NF-YA isoforms within the TAD in *Aves*, implying a role during early development.

1. Introduction

The binding of sequence-specific Transcription Factors -TFs- to specific DNA sequences located in promoters and enhancers drives initiation of transcription, whose regulation is fundamental in all developmental processes. Structurally, TFs are minimally composed of two domains: a DNA-binding domain -DBD- interacting with DNA, and a Transcription Activation Domain -TAD- involved in interactions with coactivators and/or General Transcription Factors [1–3].

NF-Y is a TF that binds to the CCAAT box, a well characterized, asymmetric sequence -RRCCAATC/GA/G- present in a relatively precise location within promoters [4]. It is a complex formed by three different subunits -NF-YA/NF-YB/NF-YC- all required for sequence-specific binding to DNA. Each subunit has evolutionarily conserved parts present in all eukaryotes, whose structural features are well understood in yeast, mammals, and plants [5–7]. NF-YB/NF-YC have a Histone Fold Domain -HFD- resembling core histones, and as such, they contact DNA non-specifically. NF-YA provides CCAAT sequence-specificity through a short -56 aa- domain named HAP2 after the yeast homologue: it is composed of the A1 α-helix, responsible for heterotrimerization with the HFD dimer, and the A2 α-helix followed by the GXGGRF motif, mediating DNA recognition [8,9]. With the exceptions of plants, which underwent large duplications of NF-Y genes and diversification into NF-YA and CCT sub-families, other kingdoms do not show substantial expansion of the three genes, a rare feature among TFs. Evolutionary studies of TFs face difficulties due to the expansion and diversification of members of most families, making the identification of orthologues often difficult. In this respect, NF-YA, considered as the regulatory subunit of the trimer, appears to be informative, precisely because of its uniqueness: based on the Human Transcription Factors database (http://humantfs.ccbr.utoronto.ca/index.php), only 7/77 DBD families have, like NF-YA, a single member. The gene of NF-YA is composed of 10 exons, located in humans on the short arm of chromosome 6. The protein coding sequence starts within exon-2 and ends within exon-10.

Most protein coding genes are involved in alternative splicing (AS), whose importance is being progressively understood in various species. AS dramatically variegates the emporium of RNA and protein products: together with initiation/elongation of transcription, it is the crucial step that regulates all developmental processes and responses to external stimuli [10–12]. AS events in TFs are particularly relevant, since they can impact directly on transcriptional initiation of hundreds/thousands of genes, by generating isoforms with different, sometimes opposite functions. Within TFs, AS events are more often found in domains outside of the DBDs. As far as NF-YA, in addition to transcripts with 3′-untranslated regions (UTR) of different lengths, the coding part is subject to AS at the N-terminal, with mRNA species sharing the C-terminal

* Corresponding author.
E-mail address: 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 de la Santé et de la Recherche Médicale, U964, Illkirch, France; Université de Strasbourg, 67404, Illkirch, France.

https://doi.org/10.1016/j.ygeno.2023.110694
Received 20 March 2023; Received in revised form 21 July 2023; Accepted 31 July 2023
Available online 1 August 2023

HAP2 domain, hence CCAAT-binding capacity. Two major AS events have been described (Fig. S1). (i) The first is the skipping of exon-3, generating NF-YAs ("Short") isoform, as opposed to NF-YAl ("Long"), where exon-3 is maintained [13–15]. NF-YAs lacks a 28/29 aa stretch with glutamines and hydrophobics, similar in composition to the surrounding TAD domain. Mouse Embryonic Stem cells -mESCs- express NF-YAs but, following differentiation, NF-YAl arises [16,17]. Similarly, Hematopoietic Stem cells -HSCs- express NF-YAs, and NF-YAl increases after differentiation [18]. In muscle C2C12 cells, expressing exclusively NF-YAl, forcing expression of NF-YAs by ablation of exon-3 leads to loss of expression of "master" TFs required for muscle commitment and differentiation [19]. (ii) A second event involves the concurrent skipping of exon-3 and the 45 amino acids long exon-5, also part of the TAD: the resulting isoform -NF-YAx- was shown to have selective properties in gene regulation [20]. It was found in human neuroblastomas, in the head of mouse embryos and, interestingly, it was induced in neuronal cells by DNA-damaging agents. Finally, minor events arise from alternative acceptor/donor splice sites of exon-7, involving short segments -6 amino acids- at the 5′ or, in fishes, at the 3′ end of the exon, producing exon-7N and exon-7C-containing transcripts, respectively [13,14,21]. As to isoforms expression, we found that NF-YAs and NF-YAl, with 7N or 7C, but not NF-YAx, are variously present in different adult tissues of mammals, amphibians, fishes, reptiles, and birds. Further genetic evidence suggested that NF-YAs is the "ancestral" form of NF-YA: exon-3 sequences appear in chordates, are independently lost in hagfish and cartilaginous fishes, and AS events originate in vertebrates [21].

During the Bernardini et al. study, we did notice polypeptide sequences predicted to contain exon-3, but not exon-5, in a restricted number of animals. We were puzzled by this finding and decided to follow up more deeply on this observation to determine whether new isoform(s) of NF-YA exist, and, in case, in which species.

2. Results

2.1. A new NF-YA isoform in birds (and alligator)

A scheme of the NF-YA gene structure common to vertebrates is shown in the upper part of Fig. 1 and Fig. S1. Previously, we generated multiple sequence alignments -MSA- of NF-YA proteins from vertebrates retrieved from the Pfam database using as query the 56 amino acids HAP2 domain shared by all eukaryotes [21]. We did notice few sequences skipping exon-5, but not exon-3, among others, from chicken, turkey, and Chinese alligator, which we previously placed in the collective branch of reptiles and birds (*Sauropsida)* (See Fig. S1C in [21]). Similarly, we noticed 2 sequences potentially corresponding to NF-YAx from the bony fishes *Anabas testudineus* and *Labrus bergylta,* and one from *Vombatus ursinus*, a mammal (Fig. S1C and Fig. S1B, respectively, in [21]). Because of the paucity of these "outliers", and the possibility of erroneous annotations (See below), we initially did not pay attention as to the actual origin of exon-5-less sequences. Thanks to the ongoing efforts of the Vertebrate Genomes Project [22], the number of available annotated vertebrate genomes vastly increased since our report [21]. This prompted us to further investigate the presence and distribution of potential novel NF-YA isoforms based on our previous observations. Fig. 1 shows an updated MSA of currently available *Sauropsida* NF-YA sequences, based on the InterPro database IPR001289 (>8000 unique sequences). We included in the shown MSA only species with a minimum of 3 protein sequences: we obtained 112 NF-YA sequences from 30 species of *Aves* and 42 from 7 species of reptiles. Inspection shows that isoforms without exon-5, but which include exon-3, are annotated in one reptile species, Chinese alligator (*Alligator sinensis*) and 12 birds: chicken (*Gallus gallus*), mallard (*Anas plathyrinchos*), eastern spot-billed duck (*Anas zonorhyincha),* tufted duck (*Aythia fuligufa*), Muscovy duck (*Cairina moschata domestica*), blue-capped manakin (*Lepidothrix coronata*), wire-tailed manakin (*Pipra filicauda*), blue tit (*Cyanistes caeruleus*), Japanese quail (*Coturnix japonica*), New Caledonian crow (*Corvus moleduloides*), spoon-billed sandpiper (*Calidris pygmaea*) and small tree-finch (*Geospiza parvula*). Because of the recurrent retrieval in diverse species, we felt confident that this represents a new isoform and decided to name it NF-YAg, a combination of *Gallus gallus,* the species in which we first noticed it, and the name of the first Author of this manuscript. Importantly, several bird sequences lack both exon-5 and exon-3, hinting at NF-YAx being expressed and at relatively abundant levels. Collectively, the four isoforms retrieved from *Aves* are depicted in Fig. S1. The presence of one NF-YAg sequence in Chinese alligator potentially makes it not exclusive for *Aves*. As expected, visual inspection indicates that the amino acids variation within *Sauropsida* is marginal, at best. A similar exercise focused on amphibians yielded only NF-YAl and the exon-3-less NF-YAs, but not exon-5-less isoforms (Fig. S2).

We noticed a few *Sauropsida* sequences with apparent variations at their 3′ ends (Fig. 1). We believe that these variants are the results of errors of sequencing and/or reconstruction of the amino acids sequence, for two reasons: (i) the corresponding conserved stretches are visible exclusively within a single species, marked as diamond for *Anas zonorhyincha*, closed circle for *Amazona collaria*, asterisk for *Gallus gallus*, square for *Melopsittacus ondulatus,* triangle for *Alligator sinensis*. There are no conserved amino acids blocks shared across species, in striking contrast with the rest, almost immutable protein. (ii) Genomic analysis of intron/exon boundaries in chicken shows that a single nucleotide gap at the end of exon-9 leads to an extended putative ORF into intron-9, with a termination codon that would erroneously exclude the ubiquitously conserved region coded by exon-10 (Not shown). We therefore decided to ignore these apparent variations at the C-terminal end.

The complete list of *Aves* NF-YA sequences present in InterPro is shown in Fig. S3: there are a total of 272 hits from 183 *Aves*. By computing them (Fig. S4) NF-YAg and NF-YAx isoforms are found mostly in species in which multiple sequences are recovered; in those with 1/2 sequences (Not shown in Fig. 1), it is mostly NF-YAl that is found. Therefore, it is likely that the absence of NF-YAg -and of NF-YAx- in most of the bird species considered might be merely due to the paucity of annotated mRNA sequences.

2.2. Conservation of intronic sequences flanking exon-5 in Archosauria and Testudines

Alternative splicing is genetically controlled by regulatory motifs within intronic regions located upstream and downstream of the exons involved; this implies that higher nucleotide conservation is scored in proximity of AS junctions, and lower around constitutive exons. We previously used the PhastCons tool to verify this aspect across vertebrate species, finding that conservation is mostly limited to exon-3 sequences and to the 5′ end of exon-7 [21]. Note, however, that the reference genomic sequence we used in this exercise was the human one and that most of the species considered in this conservation analysis were not from birds or reptiles.

For this, we analyzed the sequences included into a basewise conservation experiment consisting of 77 species (phyloP77way), of which 62 are *Sauropsida,* using the chicken sequence as reference. *Sauropsida* are classified in at least three different orders, *Archosauria* (*Aves* and *Crocodilia*), *Testudines* and *Squamata* [23]. First, we analyzed the 52 *Aves* species: Fig. 2A shows high conservation scores in sequences flanking both exon-3, as expected, and exon-5, both at 5′ and 3′ ends, a result consistent with AS regulation of exon-5. The 5′, but not 3′ end, of exon-7 shows a high score, also expected from the generation of 7N isoforms. Note that exon-6 also scores higher than the average constitutive exons at the 5′ end: it remains to be seen whether there is heterogeneity at the N-terminal portion of this exon in birds. We then aligned the other 11 species available, separating *Testudines* (4 species) and *Crocodilia* (*Alligator sinensis* and *Alligator mississippiensis*) from *Squamata* (3 species): in the formers, conservation of both exon-3 and exon-5 boundaries is above average (Fig. 2B). In *Squamata*, only exon-3 clearly stands out as above average (Fig. 2C); the relatively high overall level of conservation in all

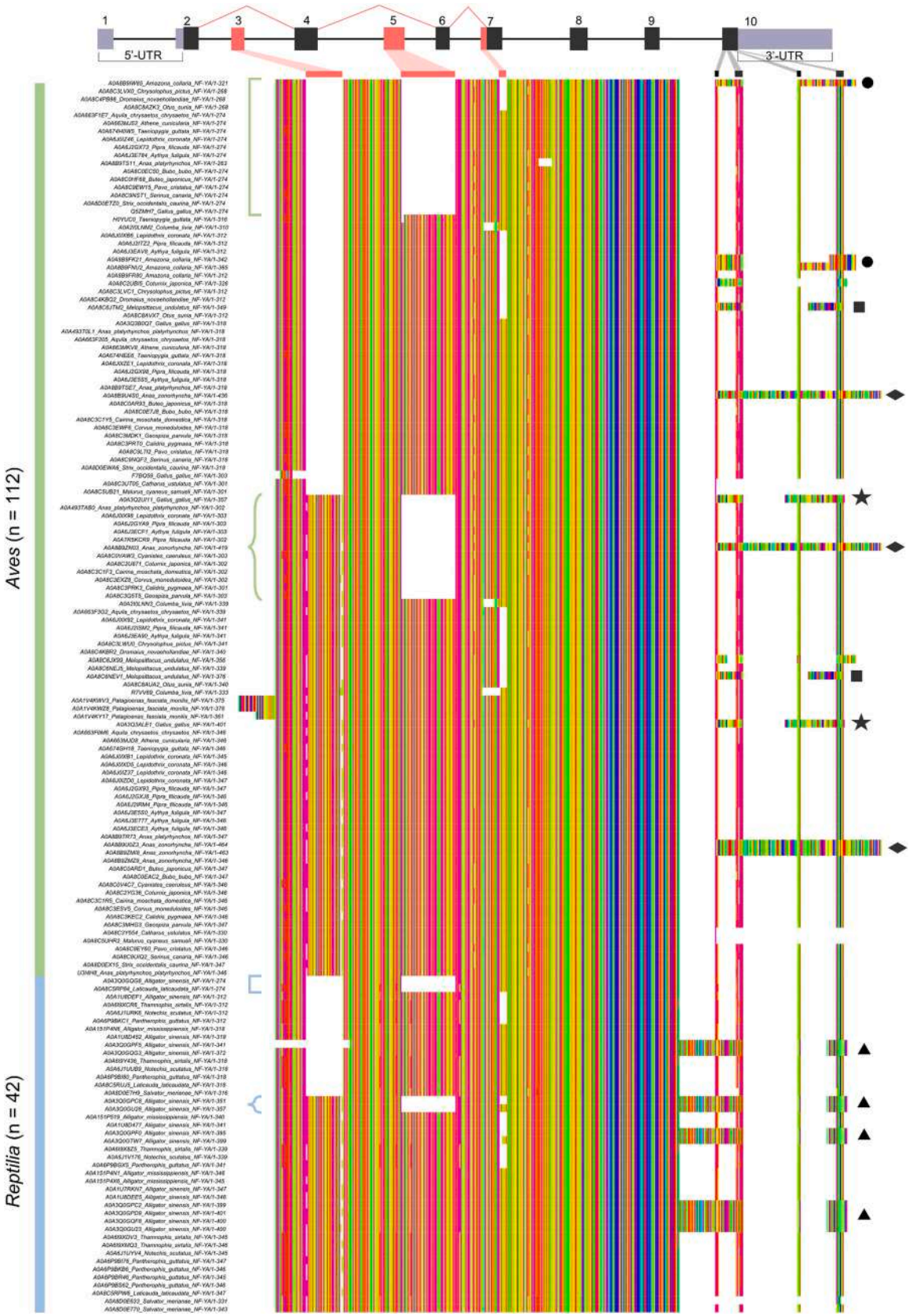


Fig. 1. Alignment of *Aves* and *Reptilia* NF-YA protein sequences.
Top: schematic representation of NF-YA intron-exon structure, with emphasis on the AS of exon-3, exon-5, and exon-7N, depicted in red. Bottom: multiple sequence alignment of NF-YA of *Aves* and *Reptilia* species, represented by at least three protein sequences within the InterPro accession IPR001289. Square and curly brackets signal NF-YAx and NF-YAg transcripts, respectively. Black symbols on the right indicate species-specific C-term variants. Alignments were exported from the software Jalview. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

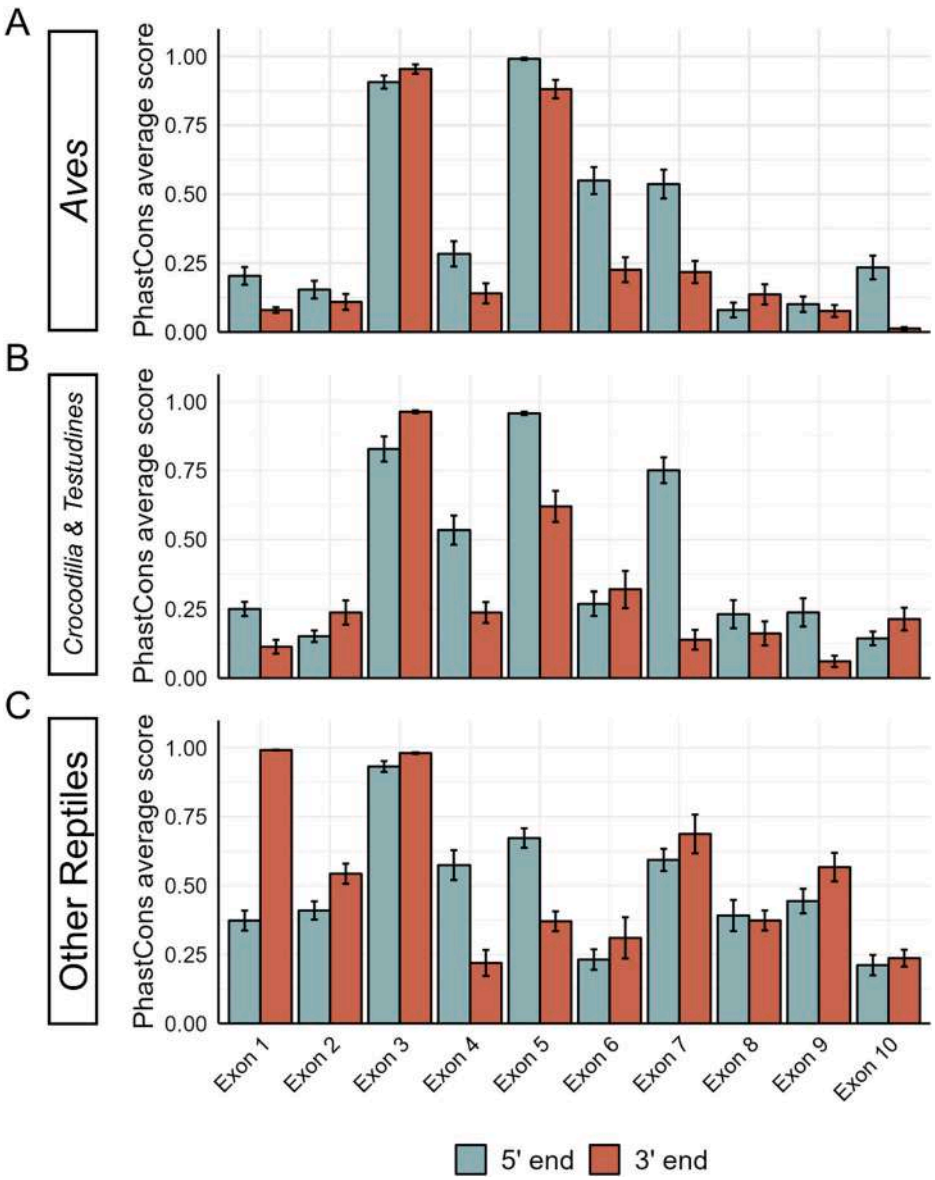


Fig. 2. Conservation of NF-YA exons boundaries in *Sauropsida*.
A. Mean conservation score, representing per-base pair probability of negative selection, across 40 bp upstream and downstream of each NF-YA exon, as calculated by PhastCons. The *Gallus gallus* sequence was selected as reference and compared to those of 52 bird species. B. Same as A, except that sequences from six *Crocodilia* and *Testudines* species were compared. C. Same as A and B, except that sequences from three *Squamata* species were compared. Error bars represent the standard error of the mean.

exons might be biased by the paucity and relatedness of the *Squamata* species available.

Finally, we looked for *cis* alternative splicing regulatory elements that might be present in the flanking regions (150 bp) of exon-5. In Fig. 3A, we confirm conservation of intronic sequences within the 15 *Aves* species included in this analysis, as represented by percent identity. The average conservation is slightly reduced in *Crocodilia* and *Testudines*, especially at distal sequences, and largely diminished in *Squamata* and 15 non-*Sauropsida* vertebrate species. The ATtRACT software finds *de novo* motifs within a DNA sequence and compares them to a collection of documented RNA binding protein (RBP) motifs [24]: the tool predicts three 10 bp long motifs within each extremity of NF-YA exon-5 boundaries. In Fig. 3B, we focused on the most distal motifs, highly conserved in *Aves*, but absent in the other species. We report a significant degree of association between the *Aves* specific distal motifs and sequences recognized by RBPs implicated in exon skipping, such as MBNL1, SF1, HNRNPA1 and the SRSF family.

Overall, these results provide genetic evidence as to the possible generation of AS of exon-5 in *Archosauria* and *Testudines*, but not in *Squamata,* and identify potential regulatory motifs.

2.3. Widespread expression of NF-YAg and NF-YAx is restricted to birds

As mentioned above, erroneously annotated sequences included in the InterPro protein family database, which is not completely curated, could impact on phylogenetic analysis, calling for a further verification as to individual isoform expression in RNA-seq datasets. This was particularly relevant for *Crocodilia*, since only one predicted NF-YAx and one NF-YAg sequence were retrieved -and only in Chinese alligator (Fig. 1)- and even more so in *Testudines*, from which we could not retrieve any sequence lacking exon-5 from the MSA analysis. To verify and quantify expression of NF-YAg and NF-YAx, we first used an exon-level strategy, inspecting all reads from raw NGS data mapped to each exon, thus avoiding possible confusion generated by annotations [21]. We then compared the levels of exon-5 and exon-3 reads to the average ones of the other constitutive exons, visualizing their expression with the IGV -Integrative Genomics Viewer- tool. Null/low exon-3 coverage signals a prevalence of either NF-YAs or NF-YAx; null/low exon-5 coverage indicates NF-YAg or NF-YAx; null/low exon-3 and exon-5 coverage suggests NF-YAx; finally, a similar coverage of both exons to the ones of constitutive exons reports a prevailing NF-YAl expression. We analyzed RNA-seq from species for which three adult tissues -brain,

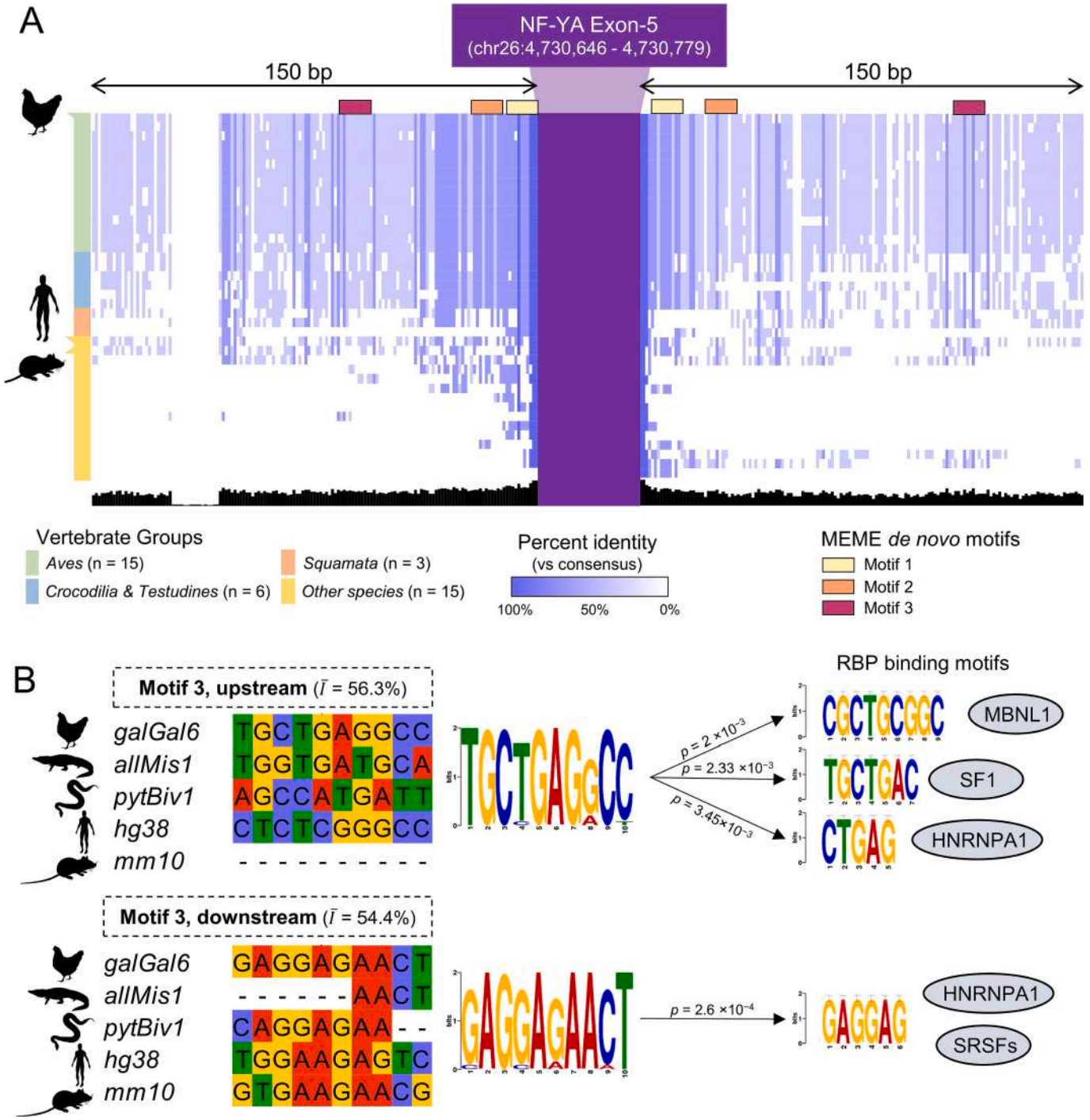


Fig. 3. Conservation of AS *cis* regulatory elements in the introns flanking NF-YA exon-5.
A. Conservation of intronic sequences 150 bp up- and downstream of chicken NF-YA exon-5. The intensity of the color represents the percentage of the bases in each position that agree with the consensus sequence, as calculated by Jalview. The species included in the analysis are part of the phyloP77way analysis. Bird species were randomly downsized from 53 to 15 (14 *plus G. gallus*). Three 10 bp motifs are included in both introns, as predicted through MEME *de novo* motif discovery in the ATtRACT tool pipeline, are indicated on top of the alignment. B. Left Panel: DNA sequence alignments including five representative species of most distal Aves-specific motifs. $\bar{I}$ = average percent identity for the motif. Right Panel: results of Tomtom comparison of predicted motifs and human RBP binding sites included in the ATtRACT database. Comparative *p* values are depicted on each arrow.

liver, muscle- are available: Japanese quail (*Coturnix japonica*), collared flycatcher (*Ficedula albicollis*), turkey (*Melagris gallopavo*), mallard (*Anas platyrhynchos),* golden eagle (*Aquila chrysaetos*), two turtles (*Malaclemys terrapin* and *Chrysemys picta*) and two alligators (*A. mississippiensis* and *A. sinensis*). For turtles, we could not find data from brain, and added one set of data from embryo; for *A. mississippiensis*, we could only find embryo data. The results are shown in Fig. 4. NF-YAl abundance -high exon-3 and exon-5 counts- is found in brain across all bird species, but not in liver or muscle. In liver, based on the absence of exon-3 and exon-5 mapped reads, NF-YAx prevails in flycatcher, NF-YAx and NF-YAg in turkey and duck, NF-YAs -low exon-3/high exon-5- in golden eagle. As for muscle, in all species exon-5 counts are either absent or lower than those of exon-3, signalling substantial levels of NF-YAg/NF-YAx.

Turtles and alligators show a very different pattern: in no tissue we

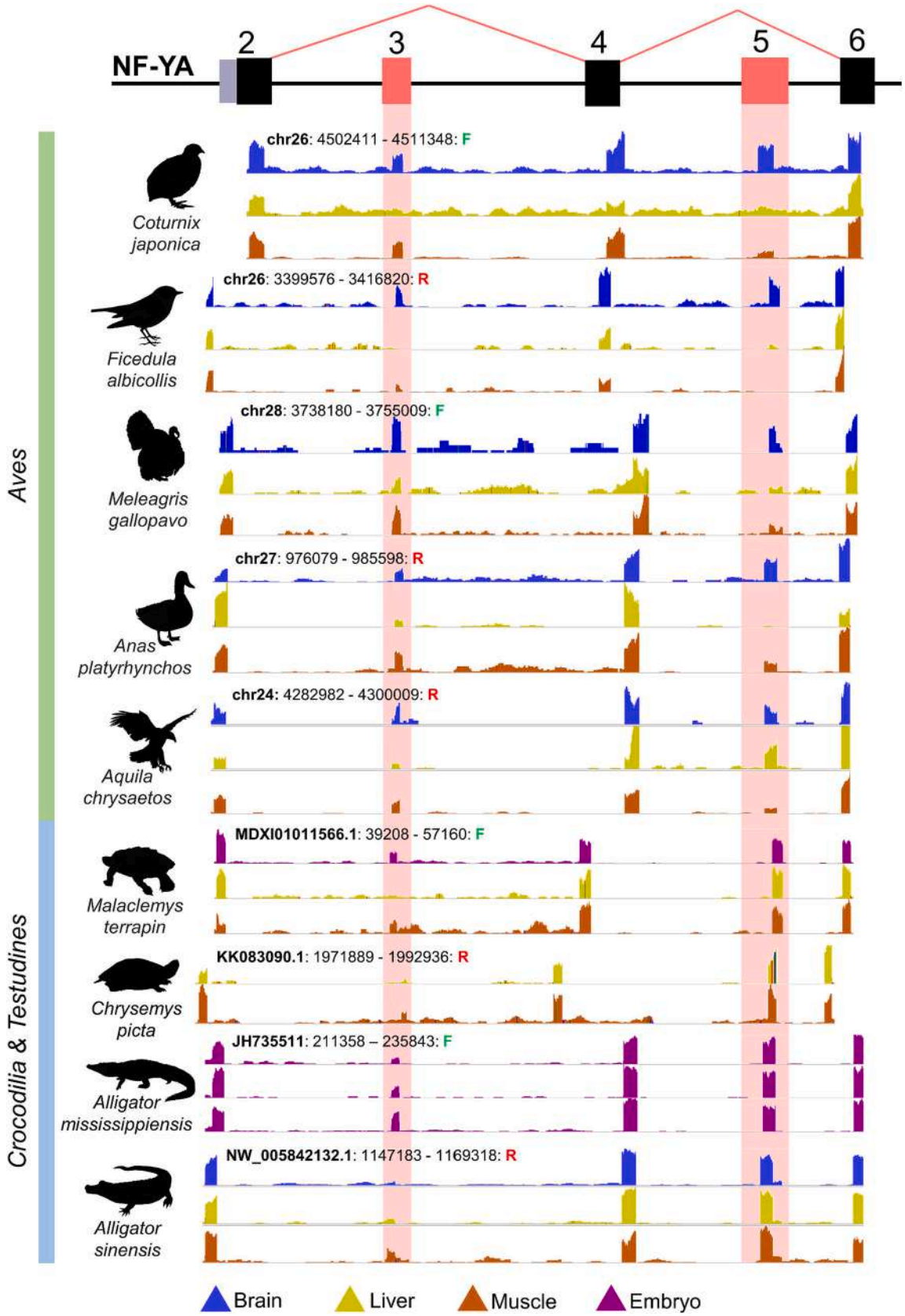


Fig. 4. Expression of NF-YA isoforms in different species of *Archosauria* and *Testudines*.
Expression of NF-YA exons-2 to exon-6 in five *Aves* and four reptile species from *Crocodilia* or *Testudines* orders, represented by mapped read coverage, in adult brain (blue track), liver (yellow), muscle (orange) or embryonic (purple) tissues. Samples from skeletal muscle were preferentially selected; when not available, samples from the heart were included in the analysis instead. Mapped reads were visualized with the IGV software. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

observe a sizeable difference in mapped reads between exon-5 and constitutive exons, indicating absence -or marginal levels- of NF-YAg expression. Exon-3 levels are lower, but still substantial in terrapin and *A. mississippiensis* embryos, suggesting expression of both NF-YAs and NF-YAl. Unlike birds, in *A. sinensis* brain there is exclusive expression of NF-YAs (no exon-3 counts); as for liver, we report a prevalence of NF-YAs in the reptile species, while muscle expresses both NF-YAs and NF-YAl in the three species considered. In conclusion, analysis of RNA-seq data support the existence of substantial levels of both NF-YAx and NF-YAg isoforms in selected tissues of birds, but not in the corresponding tissues of the turtles and alligators we examined.

2.4. NF-YAx and NF-YAg are predominant in the early phases of chicken development

NF-YAs is predominant in the totipotent mouse Embryonic Stem Cells (mESCs), while NF-YAl increases upon differentiation to Embryoid Bodies (EBs) [17,25]. The different phases of chicken early development have been intensely studied and classified: first cell division 0–5 h; 5–15 h EGK-I/VI (Eyal-Giladi Kochav classification); 15–25 h EGK-X. Several sets of bulk RNA-seq data are available: we analyzed them by two methods, first using the exon-level strategy mentioned above, calculating the associated Transcript Per Million -TPM- scores, second by making use of the RefSeq *Gallus gallus* NF-YA annotations. The results of exon-level analysis are shown in Fig. 5A, left Panel: in zygotes and EGK-I/III/VI, most transcripts have very low exon-5 counts, with presence of NF-YAx (no exon-3/-5) or NF-YAg (no exon-5). At EGK-VI, the major Zygotic Genome Activation -ZGA- occurs in chicken [26]: exon-3 TPMs are only slightly lower than the average of all other exons, suggesting high expression of NF-YAg (low levels of exon-5). At EGK-VIII and EGK-X, exon-3 and exon-5 counts increase and equal those of the other exons, signalling that NF-YAl prevails. Subsequently, we analyzed a different RNA-seq dataset derived from later stages of chicken embryo development, according to the HH classification [27], which only has one replicate. Fig. 5A, central Panel, shows low exon-3 counts, but higher of exon-5, suggesting the presence of NF-YAs, from HH4 until HH14, and further at HH36; from HH14 until HH32, the TPMs of exon-3 and exon-5 were identical, but lower than the line represented by the counts of other exons, indicating some expression of NF-YAx. Finally, we analyzed adult tissues: NF-YAl seems to predominate in brain and skeletal muscle, as TPMs of all exons equal that of exon-3 (Fig. 5A, right Panels); this result is in line with the data from other species in Fig. 4 and previous data in mouse [28–30]. However, some NF-YAg is expressed, since exon-5 counts are lower, as particularly evident in liver; this is also found in other bird species (Fig. 4).

Fig. 5B shows isoform-level expression from the same datasets, as calculated with the RSEM software: the zygote has high levels of both NF-YAx and NF-YAg, which are maintained throughout EGK-I and EGK-III/VI. Thereafter, at EGK-VIII and EGK-X, NF-YAx and NF-YAg expression drops, to the benefit of NF-YAl and, to a lower level, NF-YAs. Analysis of HH4–36 data (Fig. 5B central Panel) concurs with the exon-level one shown above, except for a more pronounced drop of NF-YAl at HH14. As for adult tissues, it is confirmed that NF-YAl predominates in brain, and even more so in muscle. In liver, it is mostly NF-YAx that is expressed (Fig. 5B, right Panel). Note that NF-YAg is expressed in all tissues, at levels higher than NF-YAs. Overall, these data concur that the NF-YAx and NF-YAg isoforms are highly expressed in the initial phases of development and further maintained in selected adult tissues.

Finally, to further substantiate these results, we considered transcripts derived from high quality “long reads” datasets, produced at Day 1 (corresponding to HH6/7 and EGK-X), Day 3 (HH20), Day 5 (HH27) and Day 7 (HH31) of chicken development. This system, bypassing the need of reconstructing annotations, or of computing single exon reads, is to be considered a definitive proof of the presence of various isoforms. We discarded partial mRNAs and only considered full-length transcripts. NF-YAg, along with NF-YAx, are present at Day 1, 3 and 5 (Fig. 5C). NF-YAl is absent at Day 1, present at Day 3 and 5, becoming apparently abundant at Day 7; NF-YAs is present throughout. Although it is not possible to draw quantitative conclusions from such modest numbers, the results are consistent with those of the HH RNA-seq, whose staging are more directly comparable.

In summary, we conclude that the NF-YAg isoform is indeed expressed in chicken and that, together with NF-YAx, it is a major isoform in the very early stages of embryogenesis.

2.5. NF-YA protein isoforms as per AlphaFold

The four NF-YA isoforms impact on the N-terminal TAD, predicted to be disordered, based on analysis of the primary amino acid sequence across species [21]. AlphaFold (AF) is a tool that provides AI powered informed predictions of the 3D structure of individual proteins -or parts of- based on the vast available knowledge of proteins and conserved domains structures [31]. We previously analyzed AF models of the human, bovine and zebrafish NF-YA, which render runs of β-stranded motifs across sequences from exon-4 to exon-7, within the poorly structured N-terminal domain [21]. We submitted the protein sequences of the chicken isoforms to this exercise: all individual subunits AF rendered models display the conserved core DBD A1 and A2 helical elements and a globally disordered N-terminal portion, hosting different arrays of β-stranded motifs. The A1/A2 part is predicted with high confidence, the N-terminal elements with low confidence scores (Fig. 6; see also Fig. S5A for AF confidence score color-code depictions). The TAD structural heterogeneity is further evident for the individual isoforms when the five different models provided by AF predictions are superimposed: for NF-YAg, three selected AF models overlay, aligned on the core domain, are shown in Fig. S5B.

Analysis of secondary structure elements displayed by N-terminal region of different isoforms -considering AF best model and lower ranking ones- shows recurrent β-stranded motifs described in NF-YAl as present in the different isoforms: a twisted antiparallel β-sheet in exon-4, spanning the exon-4/-5 boundary in NF-YAl and NF-YAs, the exon-6 hairpin, and a β-motif in exon-7 (Fig. 6A-D, Fig. S5B, and Fig. S8 in [21]).

NF-YAg N-terminal denotes even lower secondary structure content, and higher disorder: while all models display the exon-6 hairpin, only one model displays exon-7 β-motif; exon-4 β-stranded regions are arranged in a 4 stranded anti-parallel sheet in the best model only, otherwise they are limited to a β-hairpin or disordered (Fig. 6 and Fig. S5B). Exon-5 seems to provide some higher order degree to the N-terminal domain: NF-YAl and NF-YAs display a 3/4 stranded β-sheet -up to 7 strands in NF-YAl, always involving a strand encompassing the exon-4/-5 boundary sequence. NF-YAx confirms a more compact arrangement, with exon-4 β-strands hardly arranged in a twisted 4 stranded sheet (best model only), and with exon-6 hairpin often sided by the exon-7 motif.

Based on AF predictions, we infer that, by exclusion of exon-5, the conformational flexibility of the intrinsically disordered TAD could be further enhanced in NF-YAg, potentially providing a wider interaction platform, while concomitant exclusion of exon-3 in NF-YAx could limit the interactions range of the protein.

3. Discussion

This study sheds further light on the phylogenetic history of the splicing isoforms at the N-terminal TAD of NF-YA in vertebrates, notably birds and reptiles. We identify a new, “short” splicing isoform -NF-YAg- in which exon-skipping involves exclusively exon-5. In addition, NF-YAx, whose expression was previously not detected in adult tissues of several vertebrates, is easily scored in chicken embryos and adult liver of all birds analyzed. These data impact on the geometry of the Blocks of conserved amino acids previously identified within the TAD.

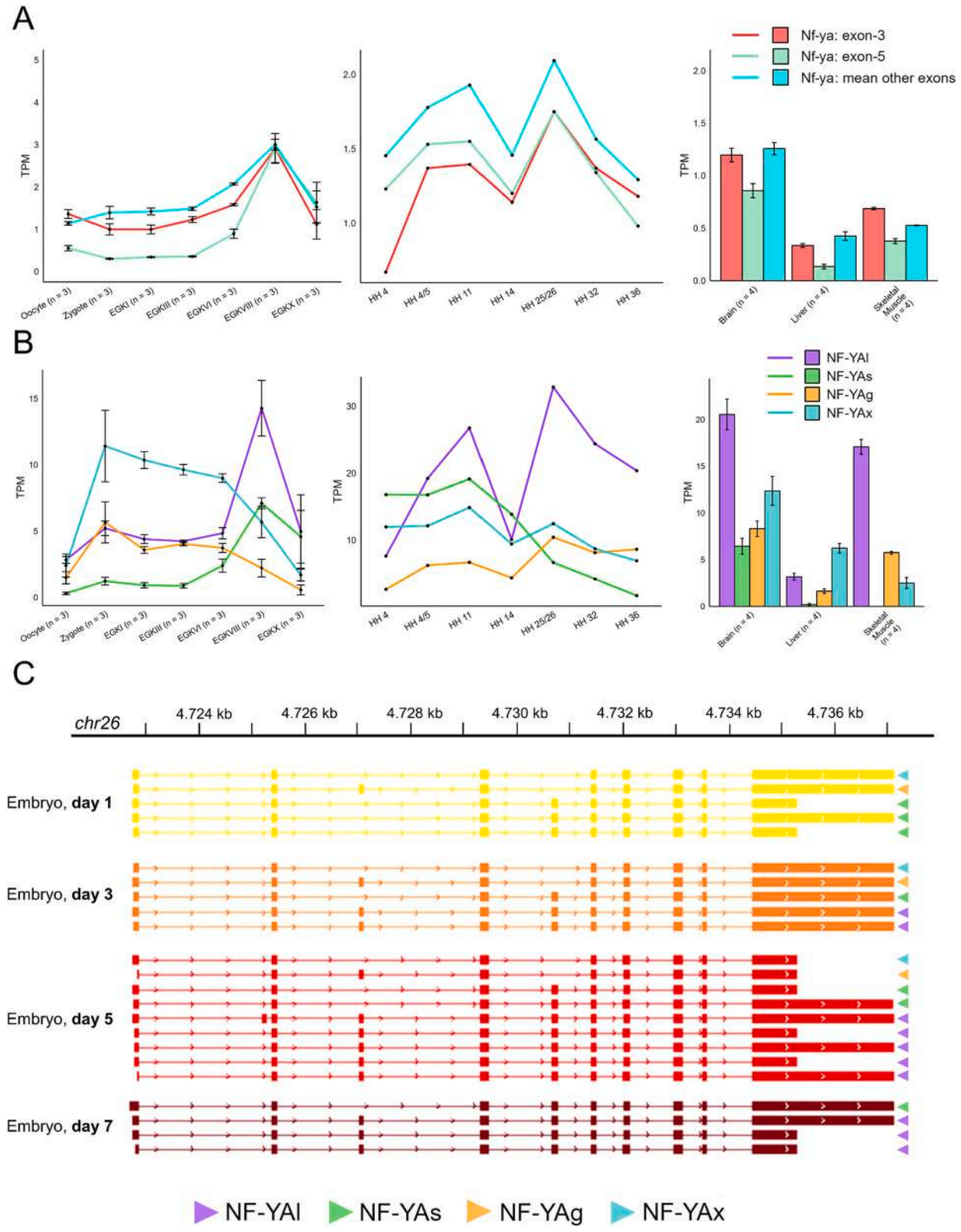


Fig. 5. Expression of NF-YA isoforms in chicken development.
A. Relative expression levels (in TPMs) of chicken NF-YA exon-3 and exon-5, compared to the expression mean of all other exons, as measured in five embryonic Eyal-Giladi and Kochav (EGK) stages, in addition of the oocyte and the zygote (Left Panel), as well as in seven Hamburger-Hamilton (HH) stages (Middle Panel), and adult brain, liver, and skeletal muscle (Right Panel). B. Same as A, but TPMs values relative to the isoforms NF-YAl, NF-YAs, NF-YAg, and NF-YAx expression are depicted. Where replicates were available: error bars represent the standard error of the mean. C. IGV visualization of collapsed transcript sets derived from chicken FLNC long reads mRNA-seq libraries at embryonic days 1, 3, 5, and 7.

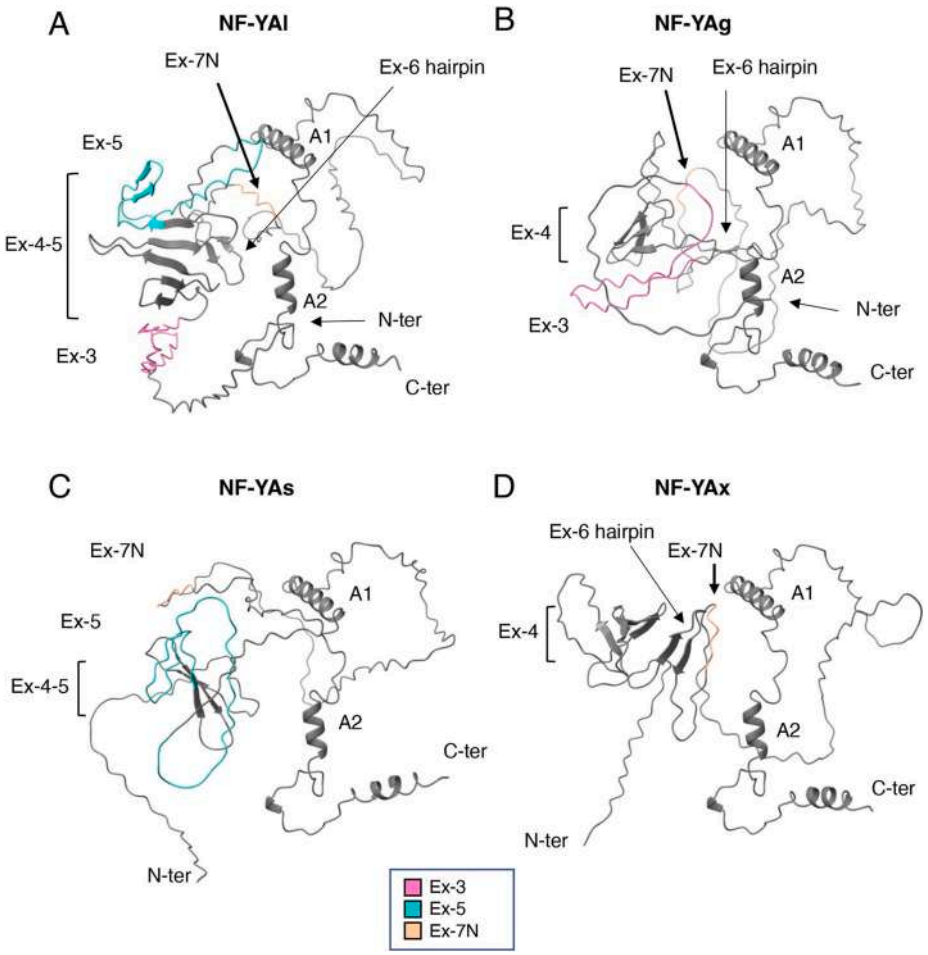


Fig. 6. AlphaFold structural predictions of chicken NF-YA protein isoforms.
AlphaFold predicted models [31] of isolated NF-YA protein isoforms (A-D) are shown in ribbon (gray color) with alternative spliced exon residues main-chain highlighted in pink (exon-3: Ex-3), light green (exon-5: Ex-5), and light orange (exon-7N: Ex-7N) color. Predicted secondary structure elements of the core domain α helices, exon-6 β hairpin, and exon-4 β sheet (spanning the exon-4/-5 boundary in YAl and YAs) are indicated. See also Fig. S5. AF models were represented with ChimeraX [58]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Alternative splicing of exon-5

Data produced in mouse and human over the last three decades indicated that the two major isoforms are NF-YAl and NF-YAs, the latter lacking exon-3. The newly described AS event brings the total of vertebrate isoforms to 4, not considering the -apparently- minor changes at the N-terminal of exon-7, lost in ray-finned fishes, and 7C, present only in Teleosts. Our previous phylogenetic analysis supported the idea that NF-YAs is the ancestral isoform of NF-YA, mainly because of the absence of exon-3 sequences in basal deuterostomes, its appearance in chordates, and further loss in sharks, rays, chimaeras, and in hagfish. The acquisition of AS of exon-3 in vertebrates led us to look for conservation of boundaries upstream and downstream of exon-3 in 99 vertebrate species, a minority of which (14) were birds. We were initially surprised not to find conservation around exon-5, considering the previous discovery of NF-YAx in human neuroblastomas [20]. This was rationalized by the undetectable levels of NF-YAx scored across 3 tissues of the 17 vertebrate species we analyzed [21]. Repeating the analysis placing chicken genomic sequences as reference now shows conservation around exon-5 in birds, alligators, and *Testudines* (Fig. 2), but not in reptiles (snakes, lizards *etc.*). Conservation of exon-5 boundaries is clearly above constitutive exons and approaching that of exon-3. We conclude that exon-5 boundaries are under positive selection in some, but not all, *Sauropsida* species. This suggests that divergence occurred before *Archosauria* radiation. Further refined analysis of exon-5 boundaries flanking sequences in *Aves* and other vertebrates identified blocks of sequences with various degree of conservation, some specific to *Aves* (Fig. 3). Some of these stretches show substantial resemblance to binding sites of known RBPs, which therefore are candidates to play a role in *Aves*-specific AS of exon-5.

Overall, these genetic data go along with the presence of transcripts of NF-YAg and NF-YAx, both lacking exon-5, easily scored in RNA-seq datasets of the 5 bird species we searched for. In alligators and turtles, RNA-seq data failed to support the existence of substantial amounts of these two isoforms. While many more tissues and developmental stages of turtles and crocodilians need to be analyzed before discounting expression of NF-YAx and NF-YAg, especially in embryogenesis when these isoforms are highly expressed in chicken, it is fair to conclude that overall levels are likely to be low or tissue/organ/developmental stage-specific, as in the other vertebrate species where conservation around exon-5 sequences was not scored.

This scenario is not unprecedented: a thorough evolutionary study on AS events in vertebrates found that most exons involved are species-specific, and that only a small group are following evolutionarily conserved tissue-specific expression patterns [32]. NF-YA would have both: on one hand, exon-3 inclusion is phylogenetically more ancient, based on conservation of intron-exon boundaries, and it follows a logic dictated by the tissue: in brain, for example, we find essentially NF-YAl, the only exception being Chinese alligator (Fig. 4 and Fig. 7 of [21]) [30]. On the other hand, AS of exon-5 is *Aves*-specific. An example of *Aves*-specific AS is represented by PTPB1 exon-9 [33]: in this case, the exon is always included in chicken and alternatively spliced in other vertebrates. The cited study by Barbosa-Morais and colleagues ascribed primarily to *cis*-acting elements around exons -rather than *trans*-acting splicing factors- the reason for species-specificity of AS. Therefore, it remains to be understood which of the elements identified in Fig. 3 conserved in birds, and absent in other vertebrates are responsible for exon-5 AS.

3.2. AS and the TAD of NF-YA

NF-YAx, and to a lesser degree NF-YAg, predominate during the major wave of ZGA of chicken embryogenesis (Fig. 5). In differentiation from mESCs to EBs NF-YAs is predominant in stem cells, NF-YAl kicks in at later stages and thereafter is highly expressed in the many adult tissues, particularly post-mitotic ones (brain, muscle). In *Aves*, NF-YAx and NF-YAg supplant NF-YAs during the early phases of development, and in some tissues (liver). What might be the consequences at the protein level of these findings, and how do they impact on NF-YAg -or NF-YAx-functionality? Within the limits of the modest variation of NF-YA, exon-3 is not one of the most conserved parts: it is absent in basal deuterostomes (echinoderms and hemichordates); it shows some variation within chordates, differences within vertebrates and it is even completely lost in *Chondrichthyes* and in hagfish [21]. It is short -28/29 amino acids- and its exclusion appears to be far from dramatic for overall TAD function: in fact, NF-YAs was shown to be fully transcriptionally competent in trans-activation assays [13,15,34–36], and mESCs require it for transcription of key stem TFs genes [16,17,25].

None of these observations apply to exon-5 sequences: it is present in all deuterostomes and, indeed, alignment of birds and reptile sequences show only marginal differences (Fig. 1 and Fig. S2). It harbors two of the three highly conserved amino acids Blocks -I and II- present in all deuterostomes. In particular, the QQIII stretch within Block II has been experimentally validated as important for full TAD function [37]. We have no data about the function of NF-YAg, but we can speculate based upon functional analysis of NF-YAx in human cells. (i) It serves as a Dominant Negative (DN), by associating to the NF-YB/NF-YC subunits, forming a trimer that binds DNA normally, but is crippled in activation. In the specific case of brain progenitors, an overexpressed NF-YAx competes with the coexpressed NF-YAl to inhibit BMI1 transcription and expansion of neuronal progenitors [20]. Mechanistically, NF-YAx is defective in binding to Sp1, another TF with a large Q-rich TAD found very often next to NF-Y in genome-wide locations studies [38–40]. A DN activity was somewhat to be expected, since an artificial construct without the entire N-terminal 160 amino acids was shown to act as a DN on, among others, growth promoting genes [41]. It remains to be seen whether the absence of exon-5 is sufficient to confer DN activities. (ii) NF-YAx also activates selected genes, such as KIF1Bb, or certain stem TFs genes (Nanog, SOX2). Lack of exon-3 and exon-5 sequences eliminates 73/74 amino acids, almost half of the TAD. As stretches within TADs were shown to act often in an additive way, one can imagine that NF-YAg might be somewhat intermediate between the fully active NF-YAs and the partially inactive/DN NF-YAx-

To understand function, we will have to eventually rationalize the three-dimensional conformations of these TADs. NF-YA was found to be one of the least structured of all proteins [42] and our previous analysis confirms that the 160 amino acids TAD is predicted to be intrinsically unstructured. IDRs (Intrinsically Disordered Regions) are typical hallmarks of TFs, generally present within their TADs [43,44]. AS provides a multiplicity of isoforms believed to mediate activation through association of modules with different, specific coactivators/repressors. For these reasons, conservation within IDRs should be significantly lower than that of structured domains such as DBDs. Instead, conservation in IDRs is very heterogeneous [[45] and References therein]: concerning NF-YA, it is far superior for Exon-4/-5/-6, with respect to exon-7 or exon-10. The AS events impacting on IDRs suggest evolutionary pressure to maintain multiple isoforms as a source of specific programs of gene expression, in this case, specific to birds. Bird-specific programs of AS are being increasingly revealed [46–48] and therefore NF-YAg should add to the list of genes identified.

In as much as the AlphaFold predictions shown in Fig. 6 are, for the time being, predictions, they suggest a regional organization of the TAD, pointing specifically at two important areas of strong evolutive pressure: the junction between exon-4 and exon-5 (Block I) and Block III of exon-6 (See Fig. S1). The second is maintained in all isoforms, corresponding rather precisely to a β-hairpin. The former is partly missed in NF-YAg and NF-YAx, leading to a decreased area and dynamic rearrangement of the β-sheets, presumably present in NF-YAl. Thus, missing exon-5 sequences would have a profound influence on the compactness of the TAD conformational *ensemble*.

In conclusion, the data presented here provide evidence for new AS splicing of this important TF, pointing at future avenues of investigation, notably the comprehension of the gene expression programs dependent upon NF-YAg -and NF-YAx- in systems expressing physiological levels of these subunits, such as chicken cells; analysis of the relationships between the DNA-encoded splicing signals mediating species-specific AS of exon-5 in birds and suspected RBPs acting on them; a systematic verification of NF-YAg/NF-YAx expression in cancer cells, which often show crucially altered AS events. Finally, the effort we made so far to elucidate NF-YA AS should be extended to NF-YC, whose multiple AS events also involve the TAD within the C-terminal part of the protein [36].

4. Materials and methods

4.1. Retrieval of NF-YA amphibians, reptiles, and birds orthologs and sequence filtering

Sauripsida and *Amphibia* NF-YA protein sequences were retrieved from the InterPro database (https://www.ebi.ac.uk/interpro/), under the accession number IPR001289. Sequences of all species belonging to these groups were downloaded in FASTA format. As for December 2022, the number of entries were 553 from 372 different species. Protein sequences not starting with Met and/or with a degenerate core domain were filtered out, while the remaining were aligned in Jalview [49,50] using Muscle [51] with default settings. The resulting MSA was manually edited and ordered according to taxonomy and isoform identity and consisted of 272 sequences from 183 species of birds, 57 sequences from 19 species of reptiles and 29 sequences from 6 species of amphibians, respectively. For the data presented in Fig. 1, we considered only species represented by three or more distinct proteins, including in the final alignment 112 *Aves* sequences from 30 species and 42 *Reptilia* proteins from 7 reptile species. All protein sequences used in the study are listed in File S1.

4.2. Exon flanking-sequences conservation

We evaluated the conservation of NF-YA exon-flanking sequences employing *Gallus gallus* as reference species and the tool PhastCons (version 1.3) [52]. We downloaded a multiple genomic alignment in MAF format from a 77 vertebrates -53 birds, 9 reptiles, and 15 other species- basewise conservation experiment included in the UCSC *galGal6* database (http://hgdownload.soe.ucsc.edu/goldenPath/galGal6/multiz77way/maf/). Species included in the *Aves*, *Crocodilia/Testudines* and *Squamata* groups −53, 6, and 3 respectively- were singled out from the original alignment, and average PhastCons scores for the 40 base pairs up- or downstream every NF-YA exon were calculated for each group.

4.3. Discovery of alternative splicing cis regulatory elements

The phyloP77way multiple genomic alignment for the 150 bp up- and downstream NF-YA exon-5 were converted from MAF to FASTA format using Galaxy MAF to FASTA tool [53], selecting one sequence per species as type of FASTA output. *Aves* species were randomly down-sampled from 53 to 15 (14 plus *G.gallus*) to guarantee a balanced representation with other vertebrate groups within the alignment. We then employed the ATtRACT web tool [24], using the alignment in FASTA format as input and the following parameters: “Zero or one per sequence” as the MEME model; 4 and 10 as minimum and maximum length of the motif, respectively; 5 as the Tomtom Evalue threshold. Finally, human RNA binding protein motifs were selected within the

complete ATtRACT dataset.

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

We retrieved the FASTQ files associated to each of the datasets considered for the analysis (Table S1), using the SRA Explorer website (https://sra-explorer.info/). We mapped the FASTQ files using *STAR* (version 2.7.8a) [54] and calculated mRNA expression at exon level with the software *featureCounts* (version 2.0.2) [55]. We obtained TPM values from raw counts mapped to each exon with an in-house script written in the Python programming language (version 3.7.1), while *Gallus gallus* transcript level expression analyses were conducted using RSEM (version 1.3.1). 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) [56]. We performed all the analyses downstream of the expression quantification in the R programming environment (version 4.0.3), with the *ggplot2*, *ggpubr*, *here*, *tidyverse* packages installed. We calculated the standard error of the mean whenever replicates were available, using the function *summarySE* from the R package *Rmisc* (version 1.5).

4.5. Long reads RNA-seq analyses

We downloaded full-length non-chimeric (FLNC) long reads collections in FASTA file format corresponding to *Gallus gallus* embryonic samples at days 1, 3, 5, and 7 of development from the NCBI BioProject under the accession number PRJNA488330 [57]. First, reads were mapped using minimap2 (version 2.24-r1122) employing the *splice* preset and the options *–secondary = no*, *-uf*, *-C5*. Resulting SAM files were sorted using samtools (version 1.10), and then processed with the script *collapse_isoforms_by_sam.py* from the Cupcake collection. Finally, all the collapsed, unique, full transcripts obtained were visualized with the IGV.

4.6. Protein structure modelling and analysis

Structural model predictions were generated using AlphaFold2 [31] with the AlphaFold Structure prediction Tool of the UCSF ChimeraX application [58] using standard settings. Chicken NF-YA isoforms input sequences were derived from FLNC long reads, correcting base pairs mismatches according to the chicken genomic sequences. AlphaFold generated models were inspected and depicted using UCSF ChimeraX [58].

Authors' contributions

AG, AB and NG performed the experiments; DD supervised the work; NG and RM wrote the manuscript.

Funding

This work was supported by institutional funding from Università degli Studi di Milano -PSR-Linea2 to DD.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2023.110694.

References

[1] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K.R. Nitta, P. Rastas, et al., DNA-binding specificities of human transcription factors, Cell. 152 (1–2) (2013) 327–339.
[2] S.A. Lambert, A. Jolma, L.F. Campitelli, P.K. Das, Y. Yin, M. Albu, et al., The human transcription factors, Cell. 172 (4) (2018) 650–665.
[3] 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.
[4] D. Dolfini, F. Zambelli, G. Pavesi, R. Mantovani, A perspective of promoter architecture from the CCAAT box, Cell Cycle 8 (24) (2009) 4127–4137.
[5] 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.
[6] M. Nardini, N. Gnesutta, G. Donati, R. Gatta, C. Forni, A. Fossati, et al., Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination, Cell. 152 (1) (2013) 132–143.
[7] A. Chaves-Sanjuan, N. Gnesutta, A. Gobbini, D. Martignago, A. Bernardini, F. Fornara, et al., Structural determinants for NF-Y subunit organization and NF-Y/DNA association in plants, Plant J. 105 (1) (2021) 49–61.
[8] V. Nardone, A. Chaves-Sanjuan, M. Nardini, Structural determinants for NF-Y/DNA interaction at the CCAAT box, Biochim Biophys Acta Gene Regul Mech. 1860 (5) (2017) 571–580.
[9] P. Hortschansky, H. Haas, E.M. Huber, M. Groll, A.A. Brakhage, The CCAAT-binding complex (CBC) in aspergillus species, Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 1860 (5) (2017) 560–570.
[10] J. Ule, B.J. Blencowe, Alternative splicing regulatory networks: functions, mechanisms, and evolution, Mol. Cell 76 (2) (2019) 329–345.
[11] F. Mantica, M. Irimia, The 3D-Evo space: evolution of gene expression and alternative splicing regulation, Annu. Rev. Genet. 56 (2022) 315–337.
[12] J.P. Verta, A. Jacobs, The role of alternative splicing in adaptation and evolution, Trends Ecol. Evol. 37 (4) (2022) 299–308.
[13] 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.
[14] K. Roder, S.S. Wolf, K.J. Larkin, M. Schweizer, Interaction between the two ubiquitously expressed transcription factors NF-Y and Sp1, Gene. 234 (1) (1999) 61–69.
[15] 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, Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression. 1579 (2) (2002) 73–80.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] L. Cappabianca, A.R. Farina, L. Di Marcotullio, P. Infante, D. De Simone, M. Sebastiano, et al., Discovery, characterization and potential roles of a novel NF-YAx splice variant in human neuroblastoma, J. Exp. Clin. Cancer Res. 38 (1) (2019) 482.
[21] A. Bernardini, A. Gallo, N. Gnesutta, D. Dolfini, R. Mantovani, Phylogeny of NF-YA trans-activation splicing isoforms in vertebrate evolution, Genomics. 114 (4) (2022), 110390.
[22] A. Rhie, S.A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, et al., Towards complete and error-free genome assemblies of all vertebrate species, Nature. 592 (7856) (2021) 737–746.
[23] M. Tollis, E.D. Hutchins, K. Kusumi, Reptile genomes open the frontier for comparative analysis of amniote development and regeneration, Int J Dev Biol. 58 (10–12) (2014) 863–871.
[24] G. Giudice, F. Sánchez-Cabo, C. Torroja, E. Lara-Pezzi, ATtRACT—a database of RNA-binding proteins and associated motifs, Database. (2016) baw035.
[25] A.J. Oldfield, P. Yang, A.E. Conway, S. Cinghu, J.M. Freudenberg, S. Yellaboina, et al., Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors, Mol. Cell 55 (5) (2014) 708–722.
[26] D. Rengaraj, Y.S. Hwang, H.C. Lee, J.Y. Han, Zygotic genome activation in the chicken: a comparative review, Cell. Mol. Life Sci. 77 (10) (2020) 1879–1891.
[27] V. Hamburger, H.L. Hamilton, A series of normal stages in the development of the chick embryo. 1951, Dev. Dyn. 195 (4) (1992 Dec) 231–272.
[28] A. Farina, I. Manni, G. Fontemaggi, M. Tiainen, C. Cenciarelli, M. Bellorini, et al., 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.
[29] A. Gurtner, I. Manni, P. Fuschi, R. Mantovani, F. Guadagni, A. Sacchi, et al., Requirement for Down-regulation of the CCAAT-binding activity of the NF-Y transcription factor during skeletal muscle differentiation, MBoC. 14 (7) (2003) 2706–2715.

[30] T. Yamanaka, H. Miyazaki, A. Tosaki, S.N. Maity, T. Shimogori, N. Hattori, et al., Gene expression profiling in neuronal cells identifies a different type of transcriptome modulated by NF-Y, Sci. Rep. 10 (1) (2020) 21714.
[31] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, et al., Highly accurate protein structure prediction with AlphaFold, Nature. 596 (7873) (2021) 583–589.
[32] N.L. Barbosa-Morais, M. Irimia, Q. Pan, H.Y. Xiong, S. Gueroussov, L.J. Lee, et al., The evolutionary landscape of alternative splicing in vertebrate species, Science. 338 (6114) (2012) 1587–1593.
[33] C.J. Wright, C.W.J. Smith, C.D. Jiggins, Alternative splicing as a source of phenotypic diversity, Nat Rev Genet. 23 (11) (2022) 697–710.
[34] 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.
[35] 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.
[36] 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.
[37] A. Silvio di, C. Imbriano, R. Mantovani, Dissection of the NF-Y transcriptional activation potential, Nucleic Acids Res. 27 (13) (1999) 2578–2584.
[38] 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.
[39] G. Suske, NF-Y and SP transcription factors — New insights in a long-standing liaison, Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1860 (5) (2017) 590–597.
[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] P.R. Romero, S. Zaidi, Y.Y. Fang, V.N. Uversky, P. Radivojac, C.J. Oldfield, et al., Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms, Proc. Natl. Acad. Sci. U. S. A. 103 (22) (2006) 8390–8395.
[45] 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.
[46] G. Wang, S. Yu, J. Liao, Identification and characterisation of alternative splice variants of hoxb9 and their correlation with melanogenesis in the black-boned chicken, Braz J Poult Sci. 21 (2019) eRBCA.
[47] X. Zhang, Q. Xiao, K. Tian, Y. Wang, X. Zhao, H. Yin, et al., Identification of three novel splicing variants and expression analysis of chicken GPR1 gene, Biomed. Res. Int. 2017 (2017) 1074054.
[48] G.T. Waites, I.R. Graham, P. Jackson, D.B. Millake, B. Patel, A.D. Blanchard, et al., Mutually exclusive splicing of calcium-binding domain exons in chick alpha-actinin, J. Biol. Chem. 267 (9) (1992) 6263–6271.
[49] 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.
[50] 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.
[51] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (5) (2004) 1792–1797.
[52] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, et al., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes, Genome Res. 15 (8) (2005) 1034–1050.
[53] D. Blankenberg, J. Taylor, A. Nekrutenko, Galaxy Team, Making whole genome multiple alignments usable for biologists, Bioinformatics. 27 (17) (2011) 2426–2428.
[54] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, et al., STAR: ultrafast universal RNA-seq aligner, Bioinformatics. 29 (1) (2013) 15–21.
[55] Y. Liao, G.K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics. 30 (7) (2014) 923–930.
[56] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, et al., Integrative genomics viewer, Nat. Biotechnol. 29 (1) (2011) 24–26.
[57] J. Ren, C. Sun, M. Clinton, N. Yang, Dynamic transcriptional landscape of the early Chick embryo, Front Cell Dev Biol. 7 (2019) 196.
[58] E.F. Pettersen, T.D. Goddard, C.C. Huang, E.C. Meng, G.S. Couch, T.I. Croll, et al., UCSF ChimeraX: structure visualization for researchers, educators, and developers, Protein Sci. 30 (1) (2021) 70–82.

3.2.3. Retrotransposon-mediated NF-YA gene duplications with potential regulatory functions in selected groups of mammals

Numerous mammalian species harbor in their genome an annotated processed retrogene that shows a high mRNA and protein sequence homology to NF-YA, called NF-YAr throughout this investigation. Most of this species are from the *Cetruminantia* clade, encompassing even-toed ungulates and cetaceans. I started retrieving candidate NF-YAr sequences in 22 mammalian species from the UCSC and Ensembl genome browsers, based on the integrity of the final protein and on significant homology to cow (*Bos taurus*) annotation. NF-YAr exhibited a high degree of similarity to the parental gene in all considered species, often lacked introns, resided on a different chromosome than the NF-YA locus, and frequently initiated from the NF-YAs M49 position. Moreover, exon-3 sequence was absent in all species, hinting that NF-YAs was the isoform originally retrotransposed and inserted into the genome.

The rate of nonsynonymous vs synonymous substitutions per codon (dN/dS, or ω) is often used to evaluate the degree of conservation of a gene, or of specific protein domains encoded by it[250]. Using this technique to analyze NF-YAr sequence alignments, I demonstrated that the domain responsible for the interaction with the HFD showed the lowest average ω score, indicating heightened evolutionary constraints within this region. Conversely, residues within the DNA-binding domain were associated with the highest ω values, thus to a greater sequence variability. This suggested that NF-YAr, if translated into protein, might be unable to recognize the CCAAT box on the DNA but could potentially bind to the other NF-Y subunits. Additionally, the three TAD regions highly conserved in deuterostomes and originally described in the first phylogenetic analysis of NF-YA protein (section **3.2.1**) were identified as conservation hotspots within the NF-YAr sequence.

I then reported favourable Kozak sequences at both M1 and M49 potential NF-YAr start sites, consistent with the hypothesis that that translation initiation of the retrogene could occur from either point. Examination of the cDNA region upstream M1, corresponding to *NF-YA* gene exon-2, revealed two closely spaced inverted CCAAT box motifs, possibly conveying a transcriptional promoting signal. After processing several *B.taurus* RNA-seq experiments, read counts mapped within NF-YAr locus boundaries were uniquely detected in 11 bull sperm samples, coinciding with the absence of NF-YA expression. Hence, NF-YAr expression might be restricted to spermatozoa, akin to other retrogenes[251].

Retrotransposon-mediated NF-YA gene duplications with potential regulatory functions in selected groups of mammals.

Gallo A.*, Bernardini A.*°, Polettini S., Dolfini D., Gnesutta N. and Mantovani R§.

Dipartimento di Bioscienze, Università degli Studi di Milano, Via Celoria 26, 20133, Milano Italy
* The Authors equally contributed.
§ Corresponding Author
Mail: mantor@unimi.it

°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.

ABSTRACT

Background

NF-Y is a transcription factor formed by Histone Fold subunits -NF-YB/NF-YC- and NF-YA, which confers sequence-specificity for the CCAAT box, an important *cis* regulatory element. In deuterostomes, NF-YA is typically a single copy gene.

Results

We describe here a second gene termed NF-YAr -for retrogene- in *Cetruminantia* and selected *Scrotifera*, *Ursidae*, squirrels and greater horseshoe bat. NF-YAr are located on different chromosomes with respect to the parental gene; they are compact, intronless, or with few introns. Analysis of RNA-seq data of *Bos taurus* indicates expression confined to male germ line cells. Conservation of Kozak signals around predicted ATGs, and of 5'UTR sequences, are consistent with protein expression, suggesting that NF-YAr is a translated, retroposed NF-YA. As to the CDSs, the transactivation domain is conserved, apparently lacking the N-terminal 77 amino acids in selected species. 3D-informed structural considerations of the predicted protein sequences point at substitutions/truncation deleterious for CCAAT-binding and, potentially, for trimer formation.

Conclusions

These findings indicate that a NF-YA retrotransposition event was fixed in selected mammals, generating a second NF-YA whose potential role is discussed.

Keywords: transcription factors; retrogene; transactivation domain; intrinsecally disordered region; NF-YA; mammals.

1. Introduction

Regulation of transcriptional initiation is key in development and physiology of all living organisms. The process involves the recognition of short DNA sequences in gene promoters and enhancers by Transcription Factors (TFs). The human genome devotes a considerable part of its protein coding capacity -some 8%- to the production of TFs, generally organized in the form of relatively few gene families, whose members are typically expanded, sometimes to considerable numbers [1–3]. Structurally, TFs have at least two domains, one conferring DNA-binding -DBD- responsible for sequence-specificity, and a Transcription Activation Domain -TAD- enhancing RNA production.

NF-YA is a subunit of the trimeric complex NF-Y, a TF that binds to an important element of regulatory regions, the CCAAT box [4]. The two other subunits -NF-YB/NF-YC- share the Histone Fold Domain -HFD- with core histones, notably H2B/H2A [5]. NF-YA provides the complex with sequence-specificity through a 56 amino acids domain present in all eukaryotes, named HAP2 after the yeast homologue [6,7]. The evolutionarily conserved parts of yeast, mammalian and plant NF-Y trimers in complex with the CCAAT box have been structurally characterized, detailing the principles for HFD heterodimerization, trimerization, sequence-specificity (NF-YA) and extended non-sequence-specific DNA contacts by HFD subunits [8–10]. The HAP2 domain of NF-YA is composed of two subdomains: the A1 α-helix mediates contacts with the HFD dimer, and a second α-helix A2, followed by the GXGGRF motif, form the DNA-recognition subdomain; A1 and A2 are connected by a flexible linker [6] . In humans, the NF-YA gene is composed of 10 exons, located on the short arm of chromosome 6; the protein coding sequence starts within exon-2 and ends within exon-10 (Fig. S1). NF-YA is involved in two major alternative splicing -AS- in mammals, comprising -NF-YAl- or lacking -NF-YAs- exon-3 sequences [11]. A third isoform -NF-YAx- was reported in human neuroblastomas, missing both exon-3 and exon-5 [12]. Another AS event is present uniquely in birds, skipping exon-5 but not exon-3, producing NF-YAg [13]. All AS isoforms have the HAP2 domain and are therefore capable of trimer formation and CCAAT binding.

Retrotransposition is a relevant force in driving the evolution of genomes, as it allows increasing the number of genes, potentially leading to new functions. Typically, an mRNA is retrotranscribed by the machineries of transposable elements and then “fixed” in the genome by insertion at locations of different chromosomes, (Reviewed by [14,15]). In mammalian species, the retrocopy remains in most cases unexpressed -retropseudogene- as it is devoid of the regulatory regions of the parental gene; these non-expressed units progressively accumulate mutations/alterations, ending up being functionally irrelevant. In some cases, the retrotransposed gene does express an RNA, also referred to as a transcribed processed retropseudogene, which might have various regulatory functions; more rarely, the expressed mRNA is translated into a protein. To be considered functional, a retrogene needs to possess an intact ORF -Open Reading Frame- with comparable length and sequence to the parental protein, as measured by nonsynonymous/synonymous conservation with respect to the parental gene, and be expressed in one or more tissues. The resulting protein

either conserves the same function of the parental one, provides a new function -neofunctionalization- or specializes into a partial function -sub-functionalization. Depending on the expression patterns of the retrogene, the new entity can populate new “territories”, or be further restricted. Intriguingly, it has been reported that a high number of retrogenes are expressed mostly -or exclusively- in male germ cells.

We recently reconstructed the phylogenetic history of NF-YA in the evolution of deuterostomes [13,16]. Gene duplications are essentially found only in fish (teleosts), due to well described whole genome duplication -WGD- events. In most other deuterostomes, a single NF-YA copy is present. In this context, NF-YA shares no homology with any other TF and therefore this subunit joins the few TF genes maintained unique across evolution (http://humantfs.ccbr.utoronto.ca/index.php). Following up on our previous studies, another set of results is presented here, describing the characterization of a retrotransposition-mediated NF-YA gene duplication in cetaceans and ruminants (*Cetruminantia*) and other -surprisingly selected- mammals.

2. Results

2.1 Identification of a second NF-YA gene in *Cetruminantia*, bears, rodents and a bat.

During our phylogenetic studies on NF-YA in deuterostomes, we noticed an additional sequence in goat (*Capra hircus*) potentially producing a shorter protein. The second goat NF-YA sequence is located on a different chromosome (Chr23 and Chr17), it is compact (0.85 *vs* 15.6 kb), it has fewer annotated exons (4 *vs* 10), whose boundaries do not match with the ones of the NF-YA gene, and it has very short (~30 bps) introns. The conclusion was that there is a pseudogene in goat. The pseudo-YA shows a predicted amino acidic sequence similar to canonical NF-YA. However, it also contains oddities, possibly due to errors in the sequencing and/or annotations. For this reason, we initially set aside this finding for further evaluation.

To ascertain whether the potential pseudo-YA was limited to goat, we analyzed the genomes of other ruminant species and indeed retrieved a second gene in all those examined, namely domestic yak (*Bos grunniens*), wild yak (*Bos mutus*), siberian musk deer (*Moschus moschiferus*), buffalo (*Bubalus bubalis*), saiga (*Saiga tatarica*), pronghorn (*Antilocapra americana*), bison (*Bison bison*), and cow (*Bos taurus*). A second sequence was also found in sheep (*Ovis aries*) and Yarkand deer (*Cervus hanglu yarkandensis*), but as for goat, they were excluded from further analysis because of incomplete -and somewhat patchy- conservation with NF-YA and with the other pseudo-YA (Not shown). Next, we surveyed cetaceans, members of the *Cetruminantia* clade: dolphin (*Tursiops truncatus*), vaquita (*Phocoena sinus*), beluga whale (*Delphinapterus leucas*), narwhal (*Monodon monoceros*), sperm whale (*Physeter catodon*), bowhead whale (*Balaena mysticetus*) and blue whale (*Balaenoptera musculus*): all have a second gene with resemblance to the canonical gene and to the pseudo-YA of ruminants. The related hippopotamus also harbours a second gene, indeed more similar to cetaceans than to ruminants, as expected from phylogenetic studies. The multiple sequence alignment (MSA) of the predicted protein sequences is shown in **Fig. 1**. Some of these genes are currently annotated as pseudogenes. The sperm and bowhead whale sequences of the pseudo-YA show the patchy inconsistencies mentioned above (Not shown) and were not included in the MSA. Because of the relatedness of the ORFs, and of additional data shown below, we will hereafter refer to these genes as NF-YAr (for retrogene).

We then searched other mammals for NF-YAr using TBLASTN with the cow protein sequence as input: we did not retrieve any sequence in the *Suidae* radiation within the *Artiodactyla* order (represented by pig, *Sus scrofa*), nor in horse (*Equus Ferus Caballus*) of *Perissodactyla*; we did find it in *Ursidae* carnivors American black bear (*Ursus americanus)*, Asiatic black bear (*Ursus thybetanus*), giant panda (*Ailuropoda melanoleuca*), but not in polar bear (*Ursus maritimus*). We found a match also in rodents, thirteen-lined ground squirrel, (*Ictidomys tridecemlineatum)*, eurasian red squirrel (*Sciurus vulgaris*), Daurian ground squirrel (*Spermophilus dauricus*) and alpine marmot (*Marmota marmota*). In these two latter species, a coherent ORF can be reconstructed only by combining two distinct reading frames, shifting at the same position in both species (**See Fig. S2**); note that we could not find NF-YAr in another *Sciuridae* member, the arctic ground squirrel. Finally, we found it in the *Chiroptera* greater horseshoe bat (*Rhinolophus*

ferrumequinum), but not in other mega- or micro-bats, nor in Egyptian rousette (Not shown). In some species, we noticed the presence of predicted Stop codons (Asterisks in **Fig. 1**), all located at the 5'-end of the CDS, but lacking a coherent conservation, unlike a Stop codon present in all cetaceans (except blue whale), hippopotamus, greater horseshoe bat and saiga, which are conserved in the same position within the DBD (corresponding to canonical exon-9 in **Fig. 1**). The 5'-end Stop codons might be the result of misreadings, since they all correspond to glutamine codons (CAG or CAA) in the other sequences, which end up being TAA or TAG; on the other hand, the DBD Stop codon has a coherent logic, discussed below. The translated CDS of NF-YAr is predicted to be shorter, but otherwise showing clear homology in length and amino acid sequence. As control, alignment of the canonical NF-YA genes in all these species confirms the almost perfect conservation, as expected by the high conservation of the NF-YA TAD across bony vertebrates (**Fig. S3**).

By searching for related DNA sequences across mammals, we retrieved a match only for a small stretch within the Q-rich TAD of the protein in almost all species (**Fig. S4**), but ORF conservation is lost. In humans, for example, an intronless pseudogene is indeed annotated as NF-YAP1 and the DNA sequence does show conservation (**Fig. S5A**), yet the unit is full of disruptive changes, insertions -see the AT-rich stretch at the 3' end- and gaps, making the resulting short ORF lacking a coherent biological logic (**Fig. S5B**). Most importantly, searches for NF-YAP1 expression in numerous human RNA-seq databases yielded negative results (Not shown). Collectively, NF-YAr is present in cetruminants and other selected species, but not in primates and most other mammals.

2.2 NF-YAr genomic structure.

The observations above could be explained by an ancient retrotransposition event in the common ancestor of mammals, which was fixed for protein readability in some species, but lost during evolution of the other. Retro(pseudo)genes are typically located on chromosomal locations different from that of the parental gene, intronless or carrying two exons. The genomic locations of NF-YAr are indeed on a different chromosome -or scaffolds- with respect to NF-YA in all species, as detailed in Fig. 2, including in humans. In most cetruminant species, the matched sequence belongs to a region annotated as a single-exon pseudogene. Exceptions are domestic yak, where NF-YAr is predicted to be a two exons gene, blue whale and thirteen-lined ground squirrel with three exons, and American black bear with four exons. These data indicate that NF-YAr was not generated by a local gene duplication event, but rather by an event of retrotransposition of the processed mRNA and insertion in a different chromosomal location.

To understand the evolutionary dynamics of genes, a useful measure is the calculation of the rate of nonsynonymous *vs* synonymous substitutions per site (dN/dS). We calculated ratios (ω) both for the parental NF-YA and NF-YAr, using two models: M2, which assumes positive selection [17] and SLR, which considers both positive and negative (or purifying) selection [18]. **Fig. 3A** shows the results of such analysis on a per-residue basis, **Fig. 3B** the partitioning of average ω scores on the five domains of the protein: Q-rich TAD, ST-rich, Subunits Interaction, DBD, C-term (See **Fig. S1**). As expected, the ratios are extremely low

for canonical NF-YA, <<0.1 throughout all domains, and indeed 0 in the HAP2 parts required for trimer formation and DNA-binding, according to SLR (**Fig. S6**), confirming the gene is under a strong purifying selection. Ratios are higher for NF-YAr, on average <1, but > 0.5, considered as a threshold level to indicate evolutionary constraints. These values do not change significantly when corrected for those of NF-YA, since the latter are extremely low. Interesting features emerge: (i) the three conserved Blocks previously identified within the TADs of extant deuterostomes [16] contain a majority of residues under selective constraints (< 0.5): 12/15 in Block I, 8/9 in Block II and 6/11 in Block III (**Fig. 3A**); (ii) with both models, the subunits interaction region has the lowest average score, and the DBD the highest (**Fig. 3B**); this dichotomy might be functionally relevant, as it will be discussed below.

2.3 Conservation of the 5'-end and Kozak sequences.

A key feature that indicates a functional CDS is the presence of Kozak sequences -CCACCATGG- around the Start codon. Many NF-YAr proteins are predicted to be shorter than the canonical NF-YA, with their Met1 corresponding to Met49 of the short isoform of parental NF-YA, thus lacking sequences of exon-2, exon-3 and part of exon-4. We notice another potential ATG just downstream -Met63- of NF-YAr Met49: we aligned the nucleotides around each of the three predicted start codons of NF-YAr. The presence of the ACC triplet before the ATG (-3/-1), and G at +4, in the majority of the species fits with the consensus (**Fig. 4A**, Right Panel); in many, the CC preceding at position -5/-4 is also consistent with an optimal Kozak. The exceptions are sequences of pronghorn, beluga whale and narwhal, not because of the surroundings, but for the presence of ATA instead of ATG. On the other hand, at Met63 several nucleotides deviate from a reasonable Kozak (**Fig. S7**). **Fig. S8** shows conservation in several species of NF-YAr sequences corresponding to canonical exon 2 and the initial part of exon 4, including the canonical ATG corresponding to Met1 of parental NF-YA. Alignment of NF-YAr sequences at the 5' end, corresponding to parental NF-YA Met1, shows that indeed pronghorn has an ATG corresponding to Met1, in line with the other species, all showing good Kozak consensus (**Fig. 4A**, Left Panel). Most of the cetaceans lack an ATG codon at position 1, replaced by AGT. In particular, narwhal and dolphin have a potentially extended ORF of 49 amino acids, whose composition -lack of glutamines, high content of charged residues- is very different from the rest of NF-YA N-terminal: this makes the effective presence of this stretch questionable (**Fig. S8**). As reference, **Fig. 4B** depicts an excellent Kozak consensus at both M1 and M49 in the parental NF-YA. Another relevant observation is that none of the species have sequences corresponding to exon-3, which is alternatively spliced in NF-YA to form the short NF-YAs isoform: this is a strong indication that NF-YAs, not NF-YAl, was the isoform originally retrotranscribed and transposed. We conclude that the presence of conserved nucleotides forming a realistic Kozak sequence around a predicted ATG -whether Met1 or Met49- supports the hypothesis that NF-YAr CDS can be translated.

Because of the extended conservation at the N-terminal shown above, we aligned genomic sequences of the various species upstream of 5' UTR sequences of NF-YAr. **Fig. 5** (Lower Panel) shows a high degree of conservation beyond upstream borders of what constitutes exon-2 of parental NF-YA: these could be the

regulatory regions of NF-YAr. In the 5'end region corresponding to the parental NF-YA exon-1, sequences are very similar in ruminants, bears and bat (Upper Panel). Incidentally, we remark the presence, in all animals except cetaceans and rodents, of two conserved CCAAT boxes some 20 bps apart, with a canonical sequence (Pu,Pu,CCAATCAG). It is possible that retrotransposition carried part of the regulatory region of the parental NF-YA, including CCAAT boxes, potentially exploiting it as a transcriptional promoting signal upon insertion [19].

2.4 Expression of NF-YAr

The majority of retrogenes in humans are not expressed and thus classified as retropseudogenes, as NF-YAP1; some 18% of them are transcribed, often in male germ cells ([14,15] and see RetrogeneDB2). To establish whether NF-YAr is a retropseudogene or a retrogene, expression was evaluated in available GEO datasets from different tissues (RNA-seq profiles) of *Bos Taurus*. As internal control, NF-YA is readily retrieved from RNA-seq datasets of liver, brain and muscle, consistent with its widespread expression (**Fig. S9A**); on the other hand, NF-YAr RNA was found only in spermatozoa (**Fig. 6**, Lower Panel), but absent in the other tissues considered including testis (**Fig. S9A** Right Panel and not shown); expression of the parental NF-YA is ubiquitously recorded, as expected (**Fig. S9A**, Left Panel). Intriguingly, we could not detect expression of the parental NF-YA in bovine spermatozoa, as measured by reads corresponding to exons (**Fig. 6**, Upper Panel). Note that RNAs of NF-YC and, to a lesser degree, NF-YB, were scored (**Fig. S10**). We were puzzled by this finding and interrogated RNA-seq datasets of early embryo stages of bovine development, from oocytes to blastocysts: NF-YA expression indeed becomes substantial at the 8-cell stage (**Fig. S9B**), in keeping with initiation of Zygotic Genome Activation -ZGA- as previously shown in these animals [20]. We conclude that, as in the case of many other retrogenes, NF-YAr shows exquisite expression in spermatozoa, where it appears to be the only NF-YA present.

2.5 Structural considerations on the NF-YAr protein sequences.

MSA of NF-YAr protein sequences (**Fig. 1**) shows notable features that deserve punctual comments.
(i) As mentioned above, starting at M49 some proteins are apparently devoid of the first 77 amino acids, with respect to the full-length NF-YAl, lacking exon-2, exon-3 and part of exon-4, while this N-terminal is present in other species. Note that this part of the TAD is the evolutionarily least conserved across deuterostomes [16]. The rest of the TAD -some 80 amino acids- is conserved, particularly the three Blocks previously identified in deuterostomes (depicted in **Fig. 3**). The notable differences: a 2 amino acids insertion in Siberian musk deer, gaps of 3 amino acids in ruminants and bat, larger ones in "exon-5" of *Ursidae*, and "exon-7" of squirrels. The common feature is the lack of the 6 amino acids corresponding to the N-terminal of exon-7, due to alternative splice donor sites [11], which points at a 7N-less mRNA as targeted by the retrotranspostion.
Hereafter the numbering of amino acids refers to mouse NF-YAs, as in Nardini et al. 2013 [8].

(ii) Within the A1 trimerization domain, residues making crucial contacts with NF-YB/NF-YC are conserved, others are variant (**Fig. 1**).

V238: conservative changes (to Ile) in ruminants should retain trimerization (involved in van der Waals -VdW- interactions with NF-YC).

N239 is an important residue, whose change to glycine or proline affects trimerization [21]; it is conserved in cetaceans, bears, squirrels and bat. This position is converted to serine in some ruminants, representing a potential non-conservative change that could impact VdW contacts with NF-YC αC.

K241 makes VdW interactions with NF-YB, it is conserved, except in vaquita (K241>Q).

Q242 is conserved in all, except a histidine in bears; it makes side chains interactions with NF-YB and NF-YC.

R245 makes ionic contacts with NF-YB α2 and is consistently different in NF-YAr, except for hippopotamus, squirrel and marmot: an R245L mutation, as found in ruminants, abolishes trimerization of the yeast HAP2 [21]. A similar effect could be observed with Phe or Cys substitutions. Less obvious to predict a negative effect of the histidine present in cetaceans, squirrels, bat and bears.

R249 is conserved, except a conservative change to Lys in blue whale and squirrels.

R250 makes ionic interactions with NF-YC D116, it is conserved in ruminants and bears, a conservative Lys in squirrel, but a Met, Gly or Trp in bat and marine species: it is relevant that R250G is detrimental for trimerization in HAP2 [21].

R253 establishes ionic interactions with NF-YB/NF-YC; it is conserved, but a Trp in blue whale and giant panda, and Gln in saiga, bears, squirrels and bat; in HAP2, changes to Ser, Gly or Lys do not affect trimer formation, whereas Ala does; due to the proximity of other hydrophobic residues, it is not easy to figure out what the consequences of a Trp might be.

K261 interacts *via* VdW with NF-YB (F50): it is conserved except in buffalo (Glu). It is a minor contact, and Ala mutation retains trimerization.

R266 interacts with NF-YB (E52 and E86): it is conserved.

As for the other changes found at L247, A254 and K264, they are solvent-exposed and conservative (or not-) changes to Phe, Thr/Glu, or Arg/Thr, respectively, should allow trimerization.

Overall, the prediction is that NF-YAr should be crippled -at least to some extent- in trimerization, mostly due to changes of N239 and R245 -in ruminants- R250 -in cetaceans- or R253, in bears, squirrels and bat. Note that most species display two or more relevant substitutions, the only exception being hippopotamus. Remarkably, the only NF-YB/NF-YC-contacting residue showing low dN/dS *ratios* – hence selective pressure- is Q242, unlike several solvent-exposed residues.

The sequence and length of the linker between A1 and A2 helices, which guides the trajectory of A2 toward the DNA, is rather conserved.

(iii) NF-YA sequence-specificity relies on 7 amino acids known to contact DNA directly (**Fig. 1**). These are all present in cetruminants, bear, squirrels and bat parental NF-YA, but only three are completely conserved in NF-YAr: H277, G286 and F289. The other four residues are variable among the different species: some

have Gln or Trp instead of R274; Trp or Gln instead of R281; Cys or His instead of R283; Arg instead of G287. A single amino acid change in some -R274G, R281I or A, R283A- abolishes DNA-binding, as experimentally tested in HAP2 [22]; G287 of the GXGGRF loop makes crucial main chain contacts with DNA bases, which is inconsistent with an Arg at this position being neutral for DNA binding [8]. Marine species all have "canonical" residues at DNA-contacting positions, but display Stop codons after Gly287: the following F289 is essential for DNA-binding [6]. Bears have a helix-crippling Pro, in addition to Gln at R281, as squirrels and marmot, having a Trp at R281. A Trp further substitutes R289 in bears. Thus, based on structural considerations, all NF-YAr display multiple changes in the principal (positively charged) residues involved in the delicate and precise contacts mediating minor-groove binding, indicating loss of (sequence-specific) DNA-binding potential.

(iv) The C-terminal of NF-YAr is conserved, in particular S291 and S297 (S320 and S326 in NF-YAl), residues modified by phosphorylation [23,24]. Conservation includes the surroundings of Ser297, which suggests that NF-YAr could be modified by the same kinases, among which CDK2 [23]. Surprisingly, C-term conservation is maintained in cetaceans and hippopotamus downstream of the aforementioned Stop codon in the DBD. We checked for the consistency of actual DNA sequences and indeed verified them in genomic sequences of all cetaceans.

In conclusion, the data lead us to infer that the changes present in NF-YAr affect NF-Y binding to DNA and possibly trimer formation with NF-YB/NF-YC.

2.6 NF-YAr structure as per AlphaFold.

The overall features of the retroprotein, based on structural and mutational analysis, can be further subjected to analysis by AlphaFold (AF), an AI tool providing predictions of the 3D structure based on knowledge of proteins and conserved domains structures [25]. We previously analyzed AF models of the human, bovine, zebrafish and chicken NF-YA, specifically predicting runs of potential β-stranded motifs from exon-4 to exon-7 within the intrinsically disordered TAD [13,16]. We analyzed NF-YAr to verify this aspect, as well as the A1 and A2 helical elements of the DBD, whose folding might be impacted by the changes mentioned above. For structural comparison, domestic yak protein sequences were selected, considering the complete and precise annotation of NF-YAr. Specifically, we compared the models with those generated for NF-YAs, given the absence of exon-3 sequences in NF-YAr.

The AF output provides five structural NF-YAs and NF-YAr models all displaying an ordered C-term (which hosts the DBD A1 and A2 helices) as *per* low Predicted Aligned Error -PAE- values (blue color in **Figure S11**). The N-term is generally disordered and can achieve some degree of folding in at least two subdomains, the most frequent involving the central portion of the protein corresponding to residues encoded by exon-6 of NF-YAs. High error values -red color- are predicted for the arrangement and packing of the N-terminal portion -relative to the C-terminal- indicating high degree of flexibility and can be thus viewed as independent domains.

Within the N-terminal, the models show different degrees of low-confidence structural organization, centered on two regions (**Figure S11**), namely aa 130-148 (NF-YAr) of exon-6 (including the conserved Block 3), always folded as a β-hairpin, and aa 60-78, which comprise exon-4/-5 boundary motif Block 1, as observed in previous analyses [13,16]. It is interesting to note that for NF-YAs models with higher structural order, secondary structure (β-) elements extend towards the C-terminus of exon-6 hairpin, including the previously described exon-7 β motif (aa 155-188). This is not observed in NF-YAr, where the ordered regions extend instead towards the N-ter of the protein. AF best NF-YAr model is shown in **Figure 7**, as compared to rank 4 model of NF-YAs, which was selected for its higher secondary structure content. As observed in **Figure 7** top Panels, NF-YAr N-terminal displays, with some degree of confidence, a wide twisted β-sheet, comprising 10 antiparallel strands, where both exon-4/5 and exon 6 elements are included. In NF-YAs, exon-6 and exon-7 motifs fold as a β-sandwich, with low confidence scores. We infer that changes within the N-terminal region, involving in particular hydrophobic residues, might contribute to the higher β-stranded structural content observed for NF-YAr N-terminal model.

Regarding the C-terminal portion, all models display the conserved A1 and A2 helical regions, predicted with high confidence scores (**Figure 7**, **Figure S11** and Not shown). This result suggests that changes in ruminant NF-YAr do not impede proper folding of the DBD region. To evaluate NF-YAr interaction potential for NF-YB/YC and DNA-binding activity, we compared the surface charge of the proteins, as shown in the bottom Panels of **Figure 7**. Despite the changes in the SI subdomain, we observe that helix A1 of NF-YAr retains an overall positive charge. In the A2 surface, instead, the essential positively-charged features of the DNA recognition subdomain are almost completely lost.

In summary, the intrinsic flexibility of the disordered TAD could be maintained in NF-YAr, as well as the overall configuration of the A1 and A2 helices; regarding protein/DNA interactions, the A1 might be still instrumental for HFD association, while the loss of positively charged residues and overall non-polar surface of the A2 helix would significantly cripple (sequence-specific) DNA binding.

3. Discussion

We report here a retrotransposon-mediated NF-YA gene duplication producing NF-YAr, expressed in male germ cells. The result of retrocopies production is typically a compact, processed retrogene, or a retropseudogene, depending on whether it is expressed or not. Given the uniqueness of NF-Y subunits in mammals, we were initially surprised to find a second gene in cetruminants and even more so when this observation was extended to other mammals. Teleostei fishes do show an increase in NF-YA copy numbers - up to five- also located on different chromosomes, but with the same intron/exon organization, as a consequence of WGD in these species [26,27]. The NF-YA retrotransposition event presumably took place in male germ cells of the primordial mammal, starting from the "short", 7N-less isoform, as shown by absence of exon-3 and of the six amino acids at the N-terminal of exon-7. It is unclear which isoform currently prevails in these cells, but NF-YAs is predominant in totipotent stem cells [28,29], as well as in stem cells of other mammalian tissues. The conservation of Kozak sequences around Met1, in species with exon-2 sequences, and around Met49, in all species, further suggest that the retrogene can be translated, although we have no direct evidence of that. Incidentally, the presence of Kozak sequences around Met49 also in the human -and rodents- sequence of parental NF-YA could suggest that this signal could be used under certain, yet undetermined, circumstances.

Most of the studies on retrocopies regard mouse and human genomes, with a minority of retrogenes carrying -or acquiring upon integration- functional elements leading to expression; an even tinier minority produces a protein. Thus, our findings are notable and, to the best of our knowledge, a *première* for *Cetruminantia* and for the other species considered. NF-YAr appears to be a non-expressed retropseudogene in most mammals, with changes accumulating, leading to abolition of the CDS. A most intriguing finding concerns the conservation of NF-YAr in selected species within the same order: in *Ursidae*, the absence of NF-YAr in Polar bear, which is more related to American and Asiatic bear than to giant panda, is notable. Obviously, this could be due to insufficient genomic data, but in bats, the CDS is absent in 5 of the 6 species for which a complete genome is available [30]. In rodents, it has been lost except in *Sciuridae*. This "patchy" evolution is not simple to explain. An alternative hypothesis to a single retrotransposition event in the mammalian ancestor could be that multiple events independently occurred in ancestors of cetruminants, carnivors, rodents and bat: additional genomic analysis of the landing location are required to shed light on this. Nevertheless, we notice that some of the species - whales, squirrels, possibly bat- are long-living animals, which have low incidence of cancer, compared to other mammals [31]: many retrogenes are known to impact on the fundamental aspects of cancer development [32]. Most studies focus on other rodents -blind mole rat- and microbat, thus less is known for the species analyzed here. NF-Y, in general, and NF-YA in particular, is increasingly implicated in cancer development: its levels are increased in epithelial tumors and overexpression is oncogenic *in vitro* and *in vivo*. NF-Yar could have a dominant-negative (DN) role on trimer function, and thus a tumor suppressive activity, but it is difficult to envisage a general role as long as expression is confined to male germ cells. However, many retrogenes become active in cancer cells and

further expression analysis in cancer specimens is required before one can rule out a more widespread role in such species.

The conspicuous conservation of the protein coding sequence of NF-YAr deserves some considerations. The positively charged A1 helix mediates formation of the trimer by contacting a negatively charged patch on the HFD heterodimer [6]. In turn, the trimer is required for formation of a stable complex with DNA, mediated by specific residues of A2 and the GXGGRF, also positively charged. Changes are present in the subunits-interacting A1, but even more in the A2/GXGGRF DBD parts present in all eukaryotes, in particular with regard to positive charges (**Fig. 7**). Mutations within the A2 and GXGGRF motif allow trimerization but they obliterate DNA-binding *in vitro* [33] and, when overexpressed in mammalian cells (Reviewed in [4]) or mice, *in vivo* [34,35]. These mutants act as DN by squelching the HFD dimer. The many changes in NF-YAr, including a dramatic truncation in marine mammals and hippopotamus, are consistent with lack of DNA-binding by NF-YAr: so, are these naturally occurring DN versions of NF-YA? The answer depends in part on the A1 and its capacity to mediate trimerization. Conservation of the A1 helix is visible, including in terms of positively charged surface, but important residues mediating trimerization are different in NF-YAr, and some of the introduced amino acids have been tested detrimental for trimer formation [21]. Is this sufficient to rule out a DN activity? Possibly not and different scenarios can be evoked. (i) We do not know whether analogous retrocopies of NF-YB and/or NF-YC exist in these species, which might have co-evolved to "adjust" to the changes of NF-YAr. (ii) Other evolutionarily conserved NF-YB/NF-YC-likes exist, potentially apt to trimerize with NF-YAr: the H2A/H2B subclass NC2a/b and Pole3/4. The former forms a complex with TBP and it is believed to work -negatively- on TATA boxes, the latter is part of DNA Pole involved in DNA replication [5]. At present, there is no data as to their association to NF-YA, but what about NF-YAr? (iii) It is even formally possible that an NF-YAr-based trimer formed with an HFD dimer other than NF-YB/NF-YC could bind to a sequence slightly different from CCAAT, through the changes in A2. In essence, rather than a DN for NF-Y, could NF-YAr redirect HFDs toward other DNA targets? This might be theoretically hypothesized in ruminants, although the overall loss of positive charges suggests otherwise, but the truncation of cetacean NF-YAr makes this hypothesis rather remote.

The second relevant issue is the conservation of the TAD, which is not inferior to that of the HAP2 domain, except for the N-terminal 49 amino acids apparently missing in some species. Conservation of the TAD, while the A2 loses a basic function of NF-Y, is an argument in favour of a Q-rich region maintaining ability to compete with NF-Y-interacting cofactors, at least some of them, away from CCAAT promoters. The apparent conservation of the sub-structures within the otherwise unstructured TAD, as predicted by Alpha-Fold is a further suggestion. In other words, playing a DN function by squelching coactivators, rather than CCAAT-binding. Finally, in addition to all these protein-based considerations, one should always consider that the NF-YAr RNA might have a role on its own, for example by sponging miRNAs or lncRNAs. Analysis of ruminants RNA-seq databases suggest that NF-YAr expression is not detectable in most tissues. Yet, bovine spermatozoa show substantial levels, unlike NF-YA. This is typical of expressed retrogenes [36]: how this is accomplished, as well as the evolutionary scope are subjects currently debated [14]. Assuming

that it is functional, what might be the specific role of NF-YAr in spermatozoa? Numerous studies in mice suggested an important role of NF-YA in spermatogonial stem cells -SSCs- and in the early phases of spermatocytes differentiation [37–39]: given the conservation of sequence, expression and splicings in vertebrates, it seems unlikely that the parental NF-YA would behave differently in the species in which NF-YAr is expressed. We can think of two possible answers. The first relates to a potential DN role during terminal differentiation of spermatozoa, by competition with parental NF-YA on activation of *cell-cycle* genes, one of the major targets within the NF-Y regulome [40]. The second is based on studies of early development of mammals. Following fertilization, zygotes undergo genome activation (ZGA, reviewed by [41]), whose timing is different in mammals: 2-cell stage in mice, and later -8-cell stage- in bovines and humans. Studies of chromatin opening by DNase I accessibility or ATAC-seq concur that the CCAAT box, among other elements, is enriched in units activated in the early wave of ZGA in several species [42,43] including bovines [20]. Lu et al. knocked-down expression of NF-YA in mouse zygotes, obtaining a substantial decrease of open *loci* of the early wave of transcriptional activation [42]. In turn, this fits with mounting evidence that NF-Y plays a pivotal "pioneering" role in establishing an open chromatin configuration [8,19,29]. Therefore, NF-YAr might delay ZGA in ruminants by competing with the lowly expressed NF-YA, which we confirm its activation at the onset of ZGA in bovines (**Fig S9**, [20]).
In summary, many phylogenetic and biochemical questions are posed by our observations: as to the formers, was the retrotransposition event single or reiterated in different mammalian ancestors? As to the latters, is NF-YAr expressed and can it form a trimer with HFDs? Can its TAD function as a squelching entity for cofactors required for NF-Y to function on CCAAT promoters? Or does it work as an actionable TAD on different targets?

4. Materials and methods.

4.1 Retrieval of NF-YAr sequences.

NF-YAr protein and *NF-YAr* DNA sequences were obtained by consulting Ensembl (release 110) [44] and UCSC genome browsers [45]. For bowhead whale, we retrieved the sequences from the website http://www.bowhead-whale.org/ [46]. We performed TBLASTN [47] or BLAT [48] searches against the genome assembly of 83 mammalian species, using as query the cow annotated (Ensembl gene: ENSBTAG00000016059) NF-YAr protein sequence. The complete list of the species considered is in Supplementary Table 1, together with *NF-YA* and *NF-YAr* gene locations. Protein and DNA sequences were edited in Jalview [49, 50], aligned using Muscle [51] with default settings, and are listed in Supplementary File 1 and Supplementary File 2, respectively. CCAAT box enrichment in the suspected promoter region of *NF-YAr* locus was calculated with the tool Simple Enrichment Analysis (SEA), from the MEME suite [52].

4.2 dN/dS ratio calculation.

Per-codon dN/dS ratios (ω) were computed, through the *ete3 evol* command, with the ETE toolkit (version 3) [53], underlying both M2 and SLR evolutionary models. The phylogenetic tree provided as input together with *Cetruminantia* NF-YA and NF-YAr protein sequences was originally obtained from NCBI Taxonomy Common tree tool (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi).

4.3 Mapping, and mRNA expression quantification.

We retrieved the FASTQ files associated to each of *Bos taurus* RNA-seq samples included in the expression analyses (Supplementary Table SX), using the SRA Explorer website (https://sra-explorer.info/). We mapped FASTQ files using STAR (version 2.7.8a) [54], and visualized mapped reads coverage by loading the BAM file corresponding to each sample into the software Integrative Genomic Viewer (IGV, version 2.10.2) [55]. Finally, we acquired bovine early embryogenesis NF-Y subunit expression data from the NCBI GEO DataSets accession GSE52415 [56]. We calculated the standard error of the mean for each developmental stage with the function *summarySE* from the R package Rmisc (version 1.5). The *ggplot2*, *ggpubr*, *here*, *tidyverse* packages were installed within the same R programming environment.

4.4 Protein Structure modelling and analysis

Protein structure models were generated using AlphaFold2 [25, 57] with the AlphaFold Structure Prediction tool within the UCSF ChimeraX application [58] using standard settings. Domestic yak (*Bos grunniens*) NF-YAs isoform input sequence was derived from the corresponding Ensembl annotation (ENSBGRT00000042807.1), which was then edited by removing the six amino acids (VTVPVS) of the 7N splicing isoform to allow for folding comparison with NF-YAr. NF-YAr sequence was obtained from the Ensembl transcript annotation ENSBGRT00000022504.1. AlphaFold computed models were inspected and depicted using UCSF ChimeraX [58].

5. Declarations.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Protein and DNA sequences of NF-YAr are included in Supplementary Files 1 and 2, respectively. Other data will be made available upon request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by institutional funding from the University of Milano PSR Linea 2 to NG and DD.

Authors' contribution

AB made the original observation and AG, AB, SP, NG performed the experiments; DD supervised the work; RM wrote the manuscript.

Acknowledgements

Not applicable.

References

1. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013 Jan 17;152(1–2):327–39.

2. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018 Feb 8;172(4):650–65.

3. Wingender E, Schoeps T, Haubrock M, Krull M, Dönitz J. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. Nucleic Acids Res. 2018 Jan 4;46(D1):D343–7.

4. Dolfini D, Zambelli F, Pavesi G, Mantovani R. A perspective of promoter architecture from the CCAAT box. Cell Cycle. 2009 Dec 15;8(24):4127–37.

5. Gnesutta N, Nardini M, Mantovani R. The H2A/H2B-like histone-fold domain proteins at the crossroad between chromatin and different DNA metabolisms. Transcription. 2013;4(3):114–9.

6. Nardone V, Chaves-Sanjuan A, Nardini M. Structural determinants for NF-Y/DNA interaction at the CCAAT box. Biochim Biophys Acta Gene Regul Mech. 2017 May;1860(5):571–80.

7. Hortschansky P, Haas H, Huber EM, Groll M, Brakhage AA. The CCAAT-binding complex (CBC) in Aspergillus species. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2017 May 1;1860(5):560–70.

8. Nardini M, Gnesutta N, Donati G, Gatta R, Forni C, Fossati A, et al. Sequence-Specific Transcription Factor NF-Y Displays Histone-like DNA Binding and H2B-like Ubiquitination. Cell. 2013 Jan 17;152(1):132–43.

9. Huber EM, Scharf DH, Hortschansky P, Groll M, Brakhage AA. DNA minor groove sensing and widening by the CCAAT-binding complex. Structure. 2012 Oct 10;20(10):1757–68.

10. Chaves-Sanjuan A, Gnesutta N, Gobbini A, Martignago D, Bernardini A, Fornara F, et al. Structural determinants for NF-Y subunit organization and NF-Y/DNA association in plants. Plant J. 2021 Jan;105(1):49–61.

11. Li XY, Hooft van Huijsduijnen R, Mantovani R, Benoist C, Mathis D. Intron-exon organization of the NF-Y genes. Tissue-specific splicing modifies an activation domain. J Biol Chem. 1992 May 5;267(13):8984–90.

12. Cappabianca L, Farina AR, Di Marcotullio L, Infante P, De Simone D, Sebastiano M, et al. Discovery, characterization and potential roles of a novel NF-YAx splice variant in human neuroblastoma. J Exp Clin Cancer Res [Internet]. 2019 Dec 5 [cited 2020 Nov 4];38. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6896337/

13. Gallo A, Dolfini D, Bernardini A, Gnesutta N, Mantovani R. NF-YA isoforms with alternative splicing of exon-5 in Aves. Genomics. 2023 Sep 1;115(5):110694.

14. Casola C, Betrán E. The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? Genome Biol Evol. 2017 Jun 1;9(6):1351–73.

15. Ciomborowska-Basheer J, Staszak K, Kubiak MR, Makałowska I. Not So Dead Genes-Retrocopies as Regulators of Their Disease-Related Progenitors and Hosts. Cells. 2021 Apr 15;10(4):912.

16. Bernardini A, Gallo A, Gnesutta N, Dolfini D, Mantovani R. Phylogeny of NF-YA trans-activation splicing isoforms in vertebrate evolution. Genomics. 2022 Jul 1;114(4):110390.

17. Yang Z, Nielsen R, Goldman N, Pedersen AMK. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics. 2000 May 1;155(1):431–49.

18. Massingham T, Goldman N. Detecting Amino Acid Sites Under Positive Selection and Purifying Selection. Genetics. 2005 Mar 1;169(3):1753–62.

19. Oldfield AJ, Henriques T, Kumar D, Burkholder AB, Cinghu S, Paulet D, et al. NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region. Nat Commun. 2019 Jul 11;10(1):3072–3072.

20. Halstead MM, Ma X, Zhou C, Schultz RM, Ross PJ. Chromatin remodeling in bovine embryos indicates species-specific regulation of genome activation. Nat Commun. 2020 Sep 17;11(1):4654.

21. Xing Y, Zhang S, Olesen JT, Rich A, Guarente L. Subunit interaction in the CCAAT-binding heteromeric complex is mediated by a very short alpha-helix in HAP2. Proc Natl Acad Sci U S A. 1994 Apr 12;91(8):3009–13.

22. Xing Y, Fikes JD, Guarente L. Mutations in yeast HAP2/HAP3 define a hybrid CCAAT box binding domain. EMBO J. 1993 Dec;12(12):4647–55.

23. Chae HD, Yun J, Bang YJ, Shin DY. Cdk2-dependent phosphorylation of the NF-Y transcription factor is essential for the expression of the cell cycle-regulatory genes and cell cycle G1/S and G2/M transitions. Oncogene. 2004 May 20;23(23):4084–8.

24. Bernardini A, Lorenzo M, Nardini M, Mantovani R, Gnesutta N. The phosphorylatable Ser320 of NF-YA is involved in DNA binding of the NF-Y trimer. FASEB J. 2019 Apr;33(4):4790–801.

25. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.

26. Meyer A, Van de Peer Y. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays. 2005 Sep;27(9):937–45.

27. Inoue J, Sato Y, Sinclair R, Tsukamoto K, Nishida M. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. Proc Natl Acad Sci U S A. 2015 Dec 1;112(48):14918–23.

28. Dolfini D, Minuzzo M, Pavesi G, Mantovani R. The Short Isoform of NF-YA Belongs to the Embryonic Stem Cell Transcription Factor Circuitry. STEM CELLS. 2012 Oct 22;30(11):2450–9.

29. Oldfield AJ, Yang P, Conway AE, Cinghu S, Freudenberg JM, Yellaboina S, et al. Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors. Mol Cell. 2014 Sep 4;55(5):708–22.

30. Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, et al. Six reference-quality genomes reveal evolution of bat adaptations. Nature. 2020 Jul;583(7817):578–84.

31. Seluanov A, Gladyshev VN, Vijg J, Gorbunova V. Mechanisms of cancer resistance in long-lived mammals. Nat Rev Cancer. 2018 Jul;18(7):433–41.

32. Staszak K, Makałowska I. Cancer, Retrogenes, and Evolution. Life (Basel). 2021 Jan 19;11(1):72.

33. Mantovani R, Li XY, Pessara U, Hooft van Huisjduijnen R, Benoist C, Mathis D. Dominant negative analogs of NF-YA. J Biol Chem. 1994 Aug 12;269(32):20340–6.

34. van Wageningen S, Nikoloski G, Vierwinden G, Knops R, van der Reijden BA, Jansen JH. The transcription factor nuclear factor Y regulates the proliferation of myeloid progenitor cells. Haematologica. 2008 Oct;93(10):1580–2.

35. Silvestre-Roig C, Fernández P, Esteban V, Pello ÓM, Indolfi C, Rodríguez C, et al. Inactivation of nuclear factor-Y inhibits vascular smooth muscle cell proliferation and neointima formation. Arterioscler Thromb Vasc Biol. 2013 May;33(5):1036–45.

36. Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res. 2016 Mar;26(3):301–14.

37. Guo J, Grow EJ, Yi C, Mlcochova H, Maher GJ, Lindskog C, et al. Chromatin and Single-Cell RNA-Seq Profiling Reveal Dynamic Signaling and Metabolic Transitions during Human Spermatogonial Stem Cell Development. Cell Stem Cell. 2017 Oct 5;21(4):533-546.e6.

38. Li J, Shen S, Chen J, Liu W, Li X, Zhu Q, et al. Accurate annotation of accessible chromatin in mouse and human primordial germ cells. Cell Res. 2018 Nov;28(11):1077–89.

39. Maezawa S, Yukawa M, Alavattam KG, Barski A, Namekawa SH. Dynamic reorganization of open chromatin underlies diverse transcriptomes during spermatogenesis. Nucleic Acids Res. 2018 Jan 25;46(2):593–608.

40. Ronzio M, Bernardini A, Pavesi G, Mantovani R, Dolfini D. On the NF-Y regulome as in ENCODE (2019). PLoS Comput Biol. 2020 Dec;16(12):e1008488.

41. Kobayashi W, Tachibana K. Awakening of the zygotic genome by pioneer transcription factors. Curr Opin Struct Biol. 2021 Dec;71:94–100.

42. Lu F, Liu Y, Inoue A, Suzuki T, Zhao K, Zhang Y. Establishing Chromatin Regulatory Landscape during Mouse Preimplantation Development. Cell. 2016 Jun 2;165(6):1375–88.

43. Liu L, Leng L, Liu C, Lu C, Yuan Y, Wu L, et al. An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. Nat Commun. 2019 Jan 21;10(1):364.

44. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic Acids Research. 2023 Jan 6;51(D1):D933–41.

45. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002 Jan 6;12(6):996–1006.

46. Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, et al. Insights into the evolution of longevity from the bowhead whale genome. Cell Rep. 2015 Jan 6;10(1):112–22.

47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.

48. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome Res. 2002 Jan 4;12(4):656–64.

49. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009 May 1;25(9):1189–91.

50. Troshin PV, Procter JB, Barton GJ. Java bioinformatics analysis web services for multiple sequence alignment--JABAWS:MSA. Bioinformatics. 2011 Jul 15;27(14):2001–2.

51. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

52. Bailey TL, Grant CE. SEA: Simple Enrichment Analysis of motifs [Internet]. bioRxiv; 2021 [cited 2023 Oct 27]. p. 2021.08.23.457422. Available from: https://www.biorxiv.org/content/10.1101/2021.08.23.457422v1

53. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution. 2016 Jun 1;33(6):1635–8.

54. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

55. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative Genomics Viewer. Nat Biotechnol. 2011 Jan;29(1):24–6.

56. Graf A, Krebs S, Zakhartchenko V, Schwalb B, Blum H, Wolf E. Fine mapping of genome activation in bovine embryos by RNA sequencing. Proc Natl Acad Sci U S A. 2014 Mar 18;111(11):4139–44.

57. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022 Jun;19(6):679-682. doi: 10.1038/s41592-022-01488-1. Epub 2022 May 30. PMID: 35637307; PMCID: PMC9184281.

58. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. 2021 Jan;30(1):70–82.

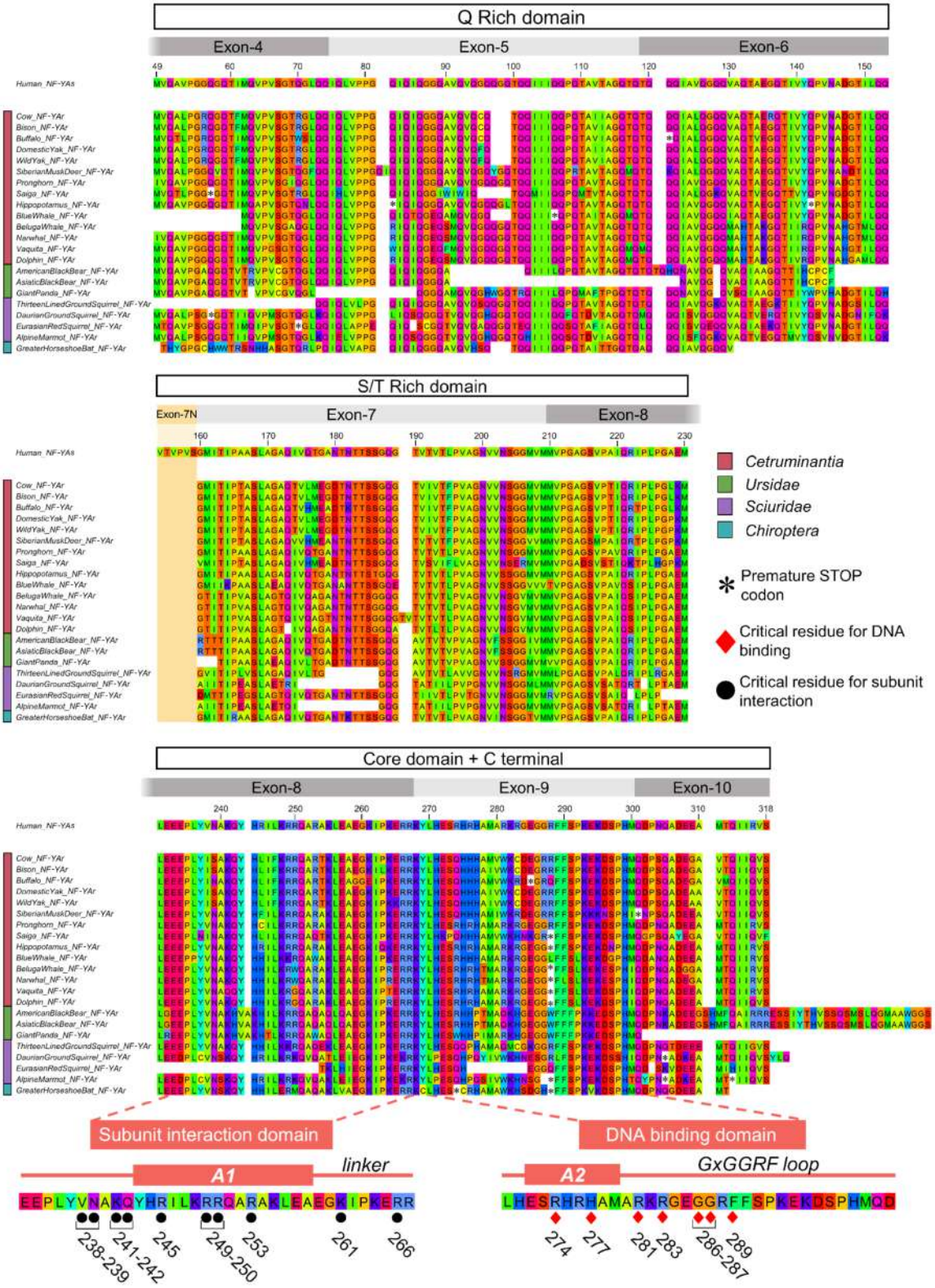


Figure 1. Alignment of NF-YAr protein sequences in selected mammalian species.
Multiple sequence alignment of three stretches of NF-YA aminoacidic sequence: from top to bottom, Q-rich domain, S/T rich domain and core domain + C terminal. Regions encoded by each exon are indicated by grey boxes. Human NF-YAs was selected as reference for residues numbering. Premature stop codons are signaled within the alignment by an asterisk, and the yellow area highlight the region encoded by Exon-7N. Bottom: close-up view of subunit interaction and DNA binding domains; the crucial residues and structures are indicated. The alignment was exported from the software Jalview.

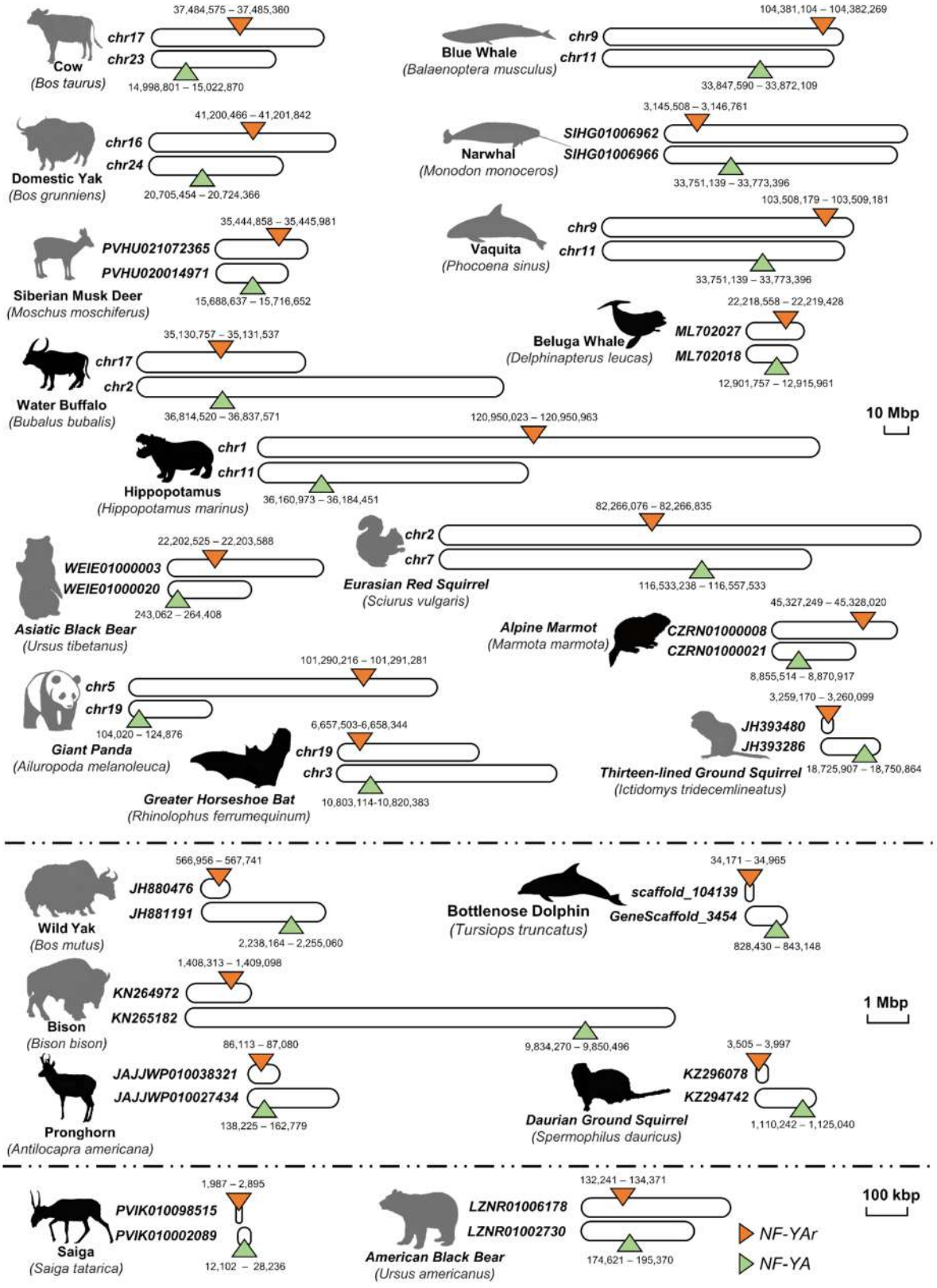


Figure 2. *NF-YA* and *NF-YAr* chromosomal localization in selected mammalian species. Schematic representation of chromosomes of 22 mammalian species, together with the coordinates of *NF-YA* (orange) and *NF-YAr* (green). The genome of animals represented by a grey silhouette includes an annotation for *NF-YAr* in Ensembl, as listed in Supplementary Table 1, while black silhouettes do not. Species are arranged in three main groups, depending on the scale used for chromosome length.

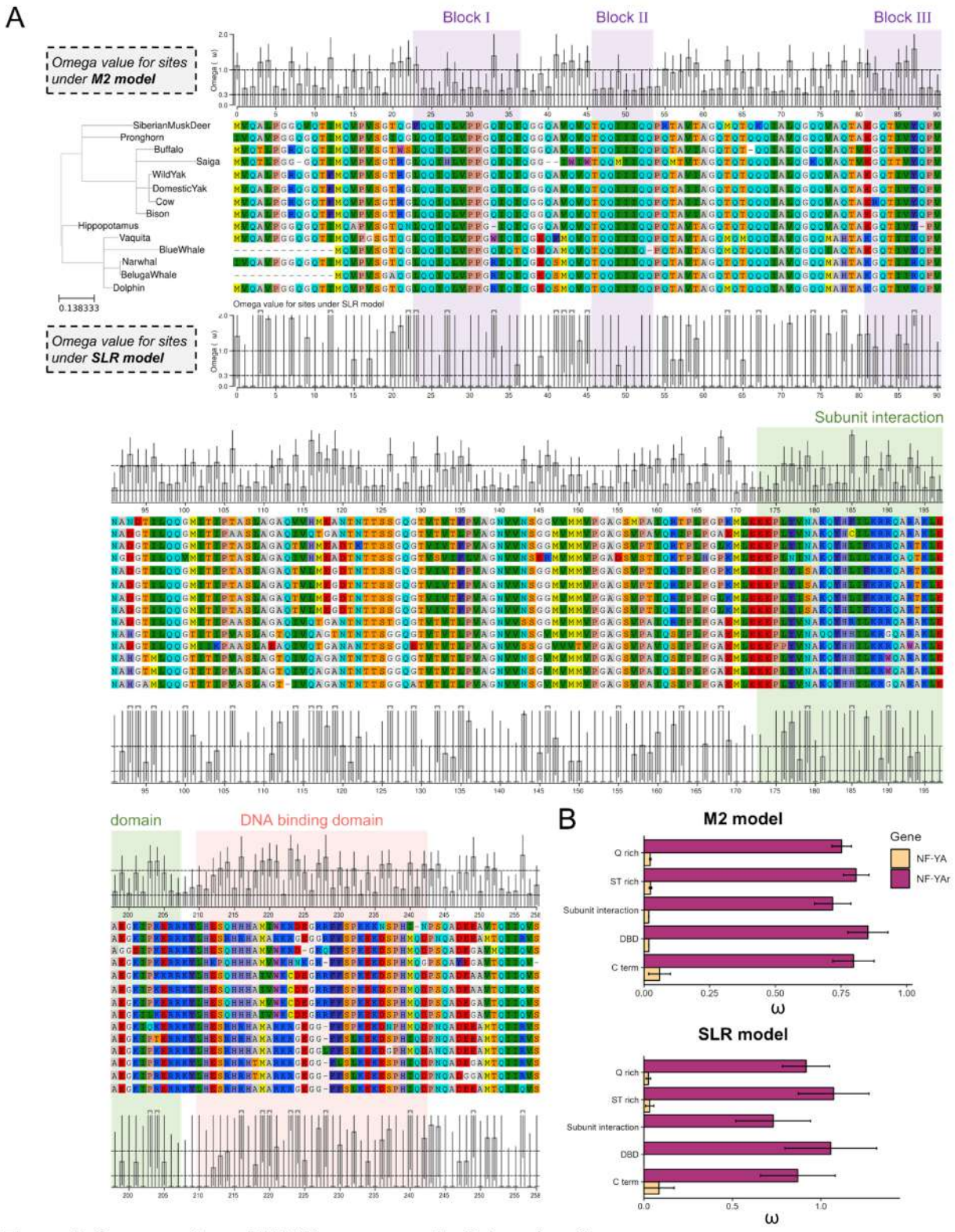


Figure 3. Conservation of NF-YAr sequence in *Cetruminantia*.
A. Per-codon dN/dS ratios (ω) of NF-YAr represented as barplots for *Cetruminantia* species. Above the alignment, computed ω values underly the M2 evolutionary model (positive selection), below ω values are determined under the SLR model (positive/purifying selection). TAD blocks conserved in deuterostomes are highlighted in purple. The alignment was generated by ETE toolkit. **B.** average ω score for different domains, in both NF-YA (complete alignment in Supplementary Figure S6) and NF-YAr.
Error bars = standard error of the mean; phylogenetic tree branch length = number of nucleotide substitutions per codon.

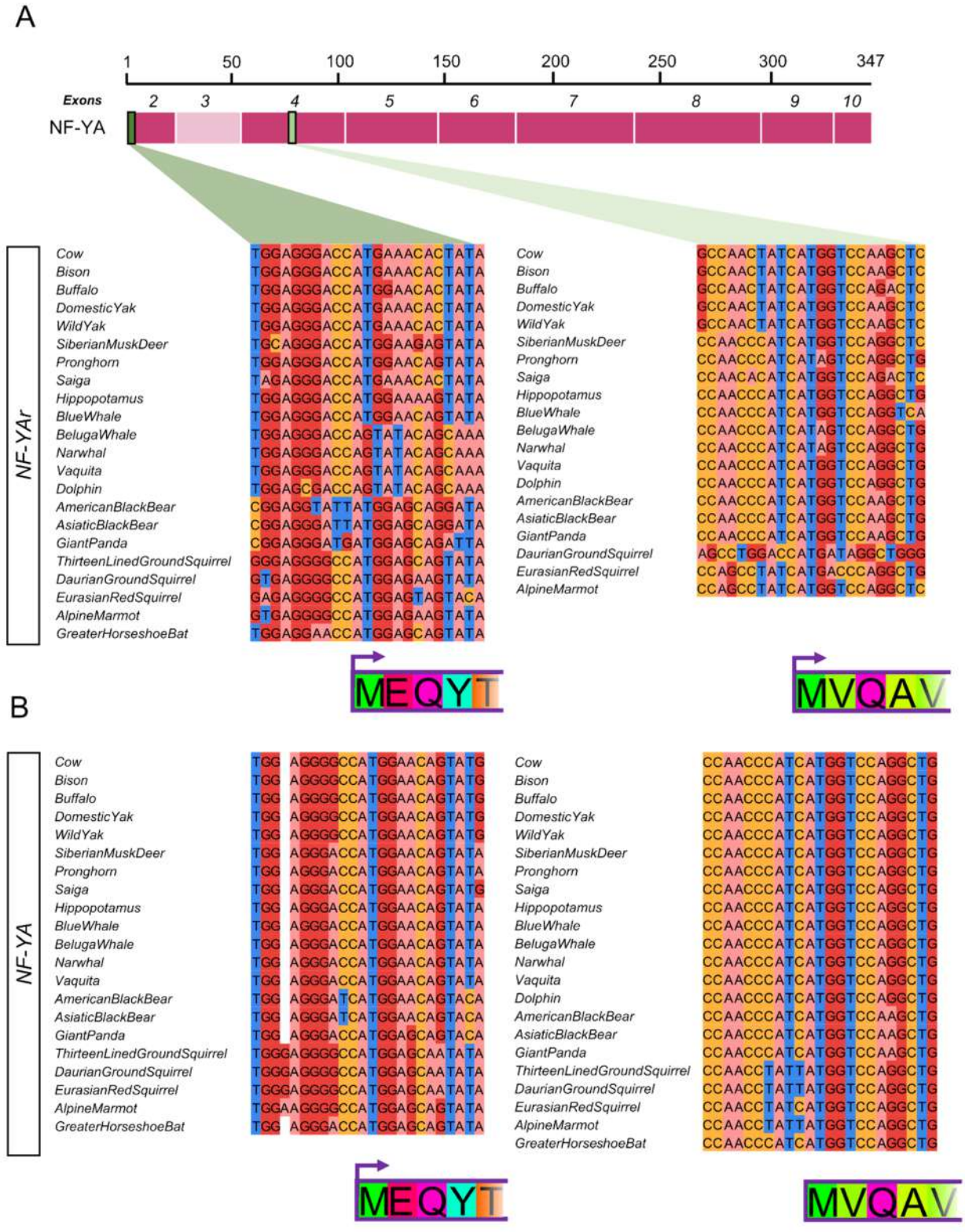


Figure 4. Conservation of *NF-YA* and *NF-YAr* Kozak sequence.
A. Top: Schematic representation of NF-YA mature mRNA. The position of two ATG codons are marked, M1 and M49. Bottom: MSA of *NF-YAr* sequences from selected mammalian species containing bases spanning from -10 to +10 from the two potential start codons. Purple boxes: first five translated amino acids of NF-YAr protein. **B.** Same as A, except the alignments for the regions around *NF-YA* M1 and M49 are shown. The alignment was exported from the software Jalview.

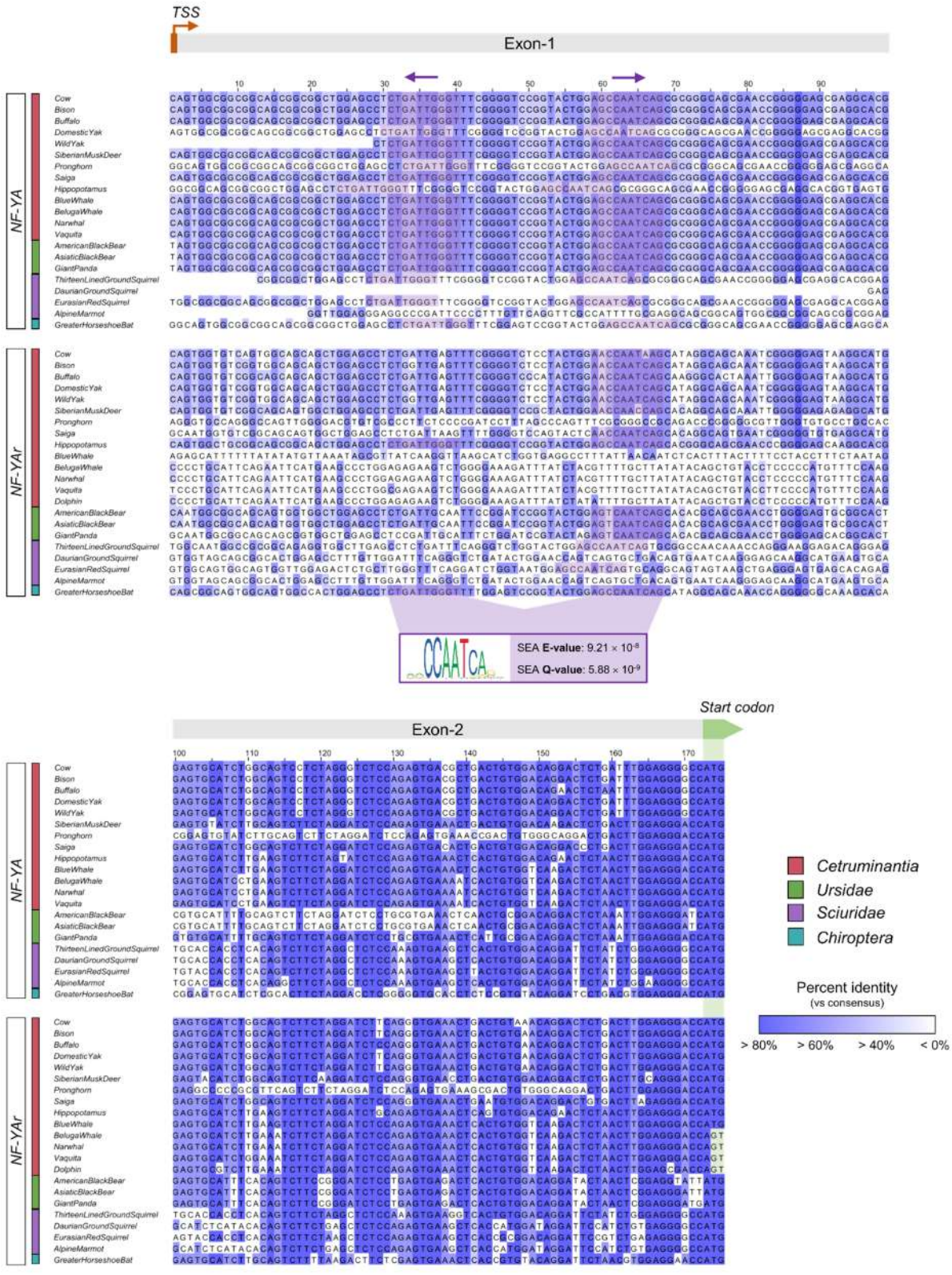


Figure 5. Conservation of *NF-YA* and *NF-YAr* 5' UTR.
MSA of *NF-YA* and *NF-YAr cDNA* sequences from selected mammalian species, spanning from start codon to 172 bps upstream of it, corresponding to the TSS in *Bos taurus*. *NF-YA* Exon-1 and Exon-2 are indicated above the sequences by gray boxes. For the species in which a NF-YA Exon-1 annotation was not available, homologous sequences obtained through BLAST/BLAT are included. The CCAAT box enrichment analysis performed by SEA is highlighted in purple, as well as is the motifs within the alignment. Intensity of the color = per-nucleotide percent identity, when compared to the whole alignment consensus. The alignment was exported from the software Jalview.

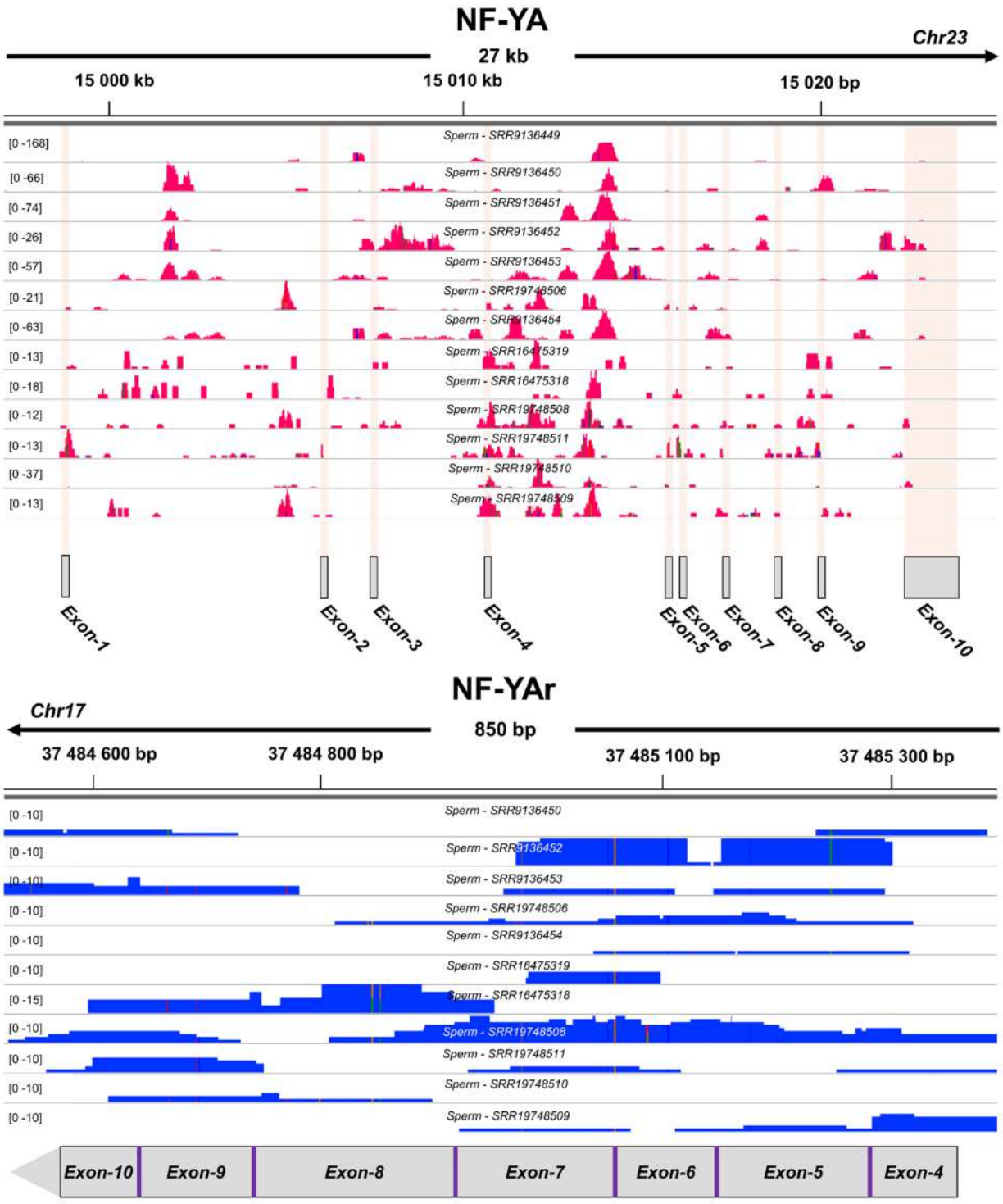


Figure 6. *NF-YAr* expression in cow sperm samples.
Top: *NF-YA* expression from 13 *Bos taurus* sperm RNA-seq samples, represented by mapped read coverage. The regions corresponding to each exon are highlighted in pink. Bottom: *NF-YAr* expression from 11 *Bos taurus* sperm RNA-seq samples. The gray boxes delineate regions within *NF-YAr* with a strong homology for the nucleotide sequences encoded by the specified *NF-YA* exons. On the left of each sample track, read number ranges are depicted as [min-max]. Mapped reads were visualized with the IGV software.

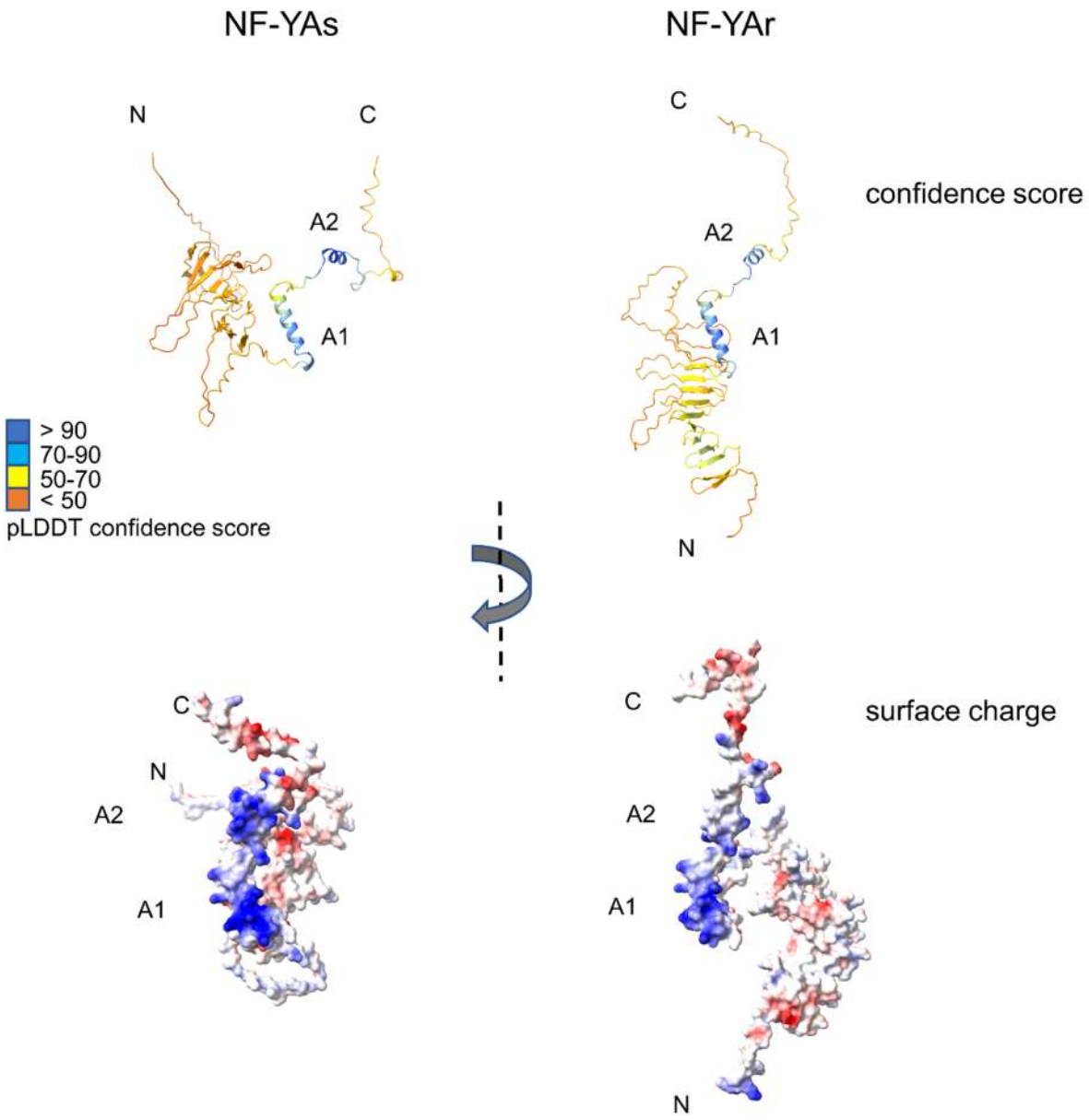


Figure 7. AlphaFold models of Domestic Yak NF-YA.
Top: ribbon depiction of AF models obtained for Domestic Yak NF-YAs (left) and NF-YAr (right) with AF per-residue confidence score (pLDDT) color palette. Bottom: the same models, rotated around the y-axis, are shown below in surface charge coloring (blue: positive, red: negative), to highlight interaction surfaces features. Secondary structure alpha-helical elements A1 and A2 of the DBD are indicated. AF rank 4 and rank 1 models are shown for NF-YAs and NF-YAr, respectively (see also Figure S11). Structural models images were obtained with ChimeraX.

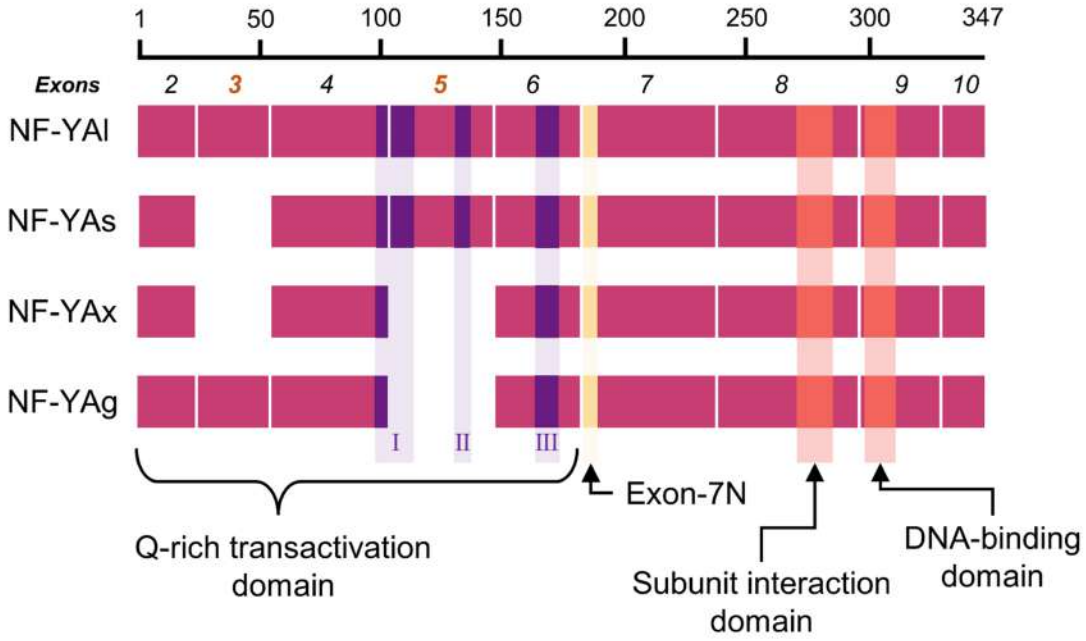


Figure S1. Scheme of NF-YA isoforms sequences.
Schematic representation of the four NF-YA AS variants currently reported: NF-YAl, NF-YAs, NF-YAx, NF-YAg. Exon-3 and exon-5 are highlighted in red; purple regions: blocks of strong evolutive pressure; yellow region: six amino acids encoded by the N-term of exon-7 (Exon-7N variant); orange regions: the two components of NF-YA core domain: the subunit interaction domain and the DNA-binding domain (DBD).

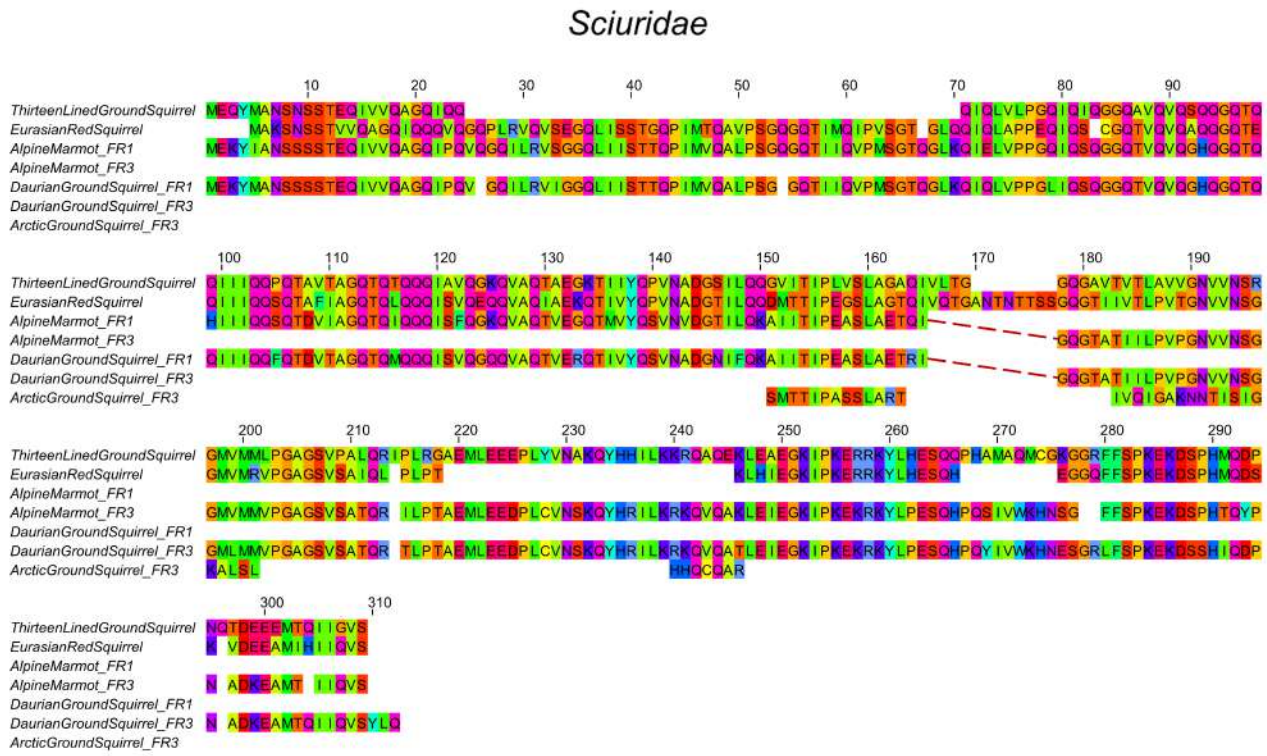


Figure S2. Alignment of NF-YAr protein sequences in five *Siuridae* species.
NF-YAr protein MSA of five squirrels: Thirteen-lined Ground Squirrel (*Ictidomys tridecemlineatus*), Eurasian Red Squirrel (*Sciurus vulgaris*), Alpine Marmot (*Marmota marmota marmota*), Daurian Ground Squirrel (*Spermophilus dauricus*), and Arctic Ground Squirrel (*Urocitellus parryii*). For the species in which a frame shift event was suspected, both frames involved are included in the alignment and connected by dashed lines. The alignment was exported from the software Jalview.

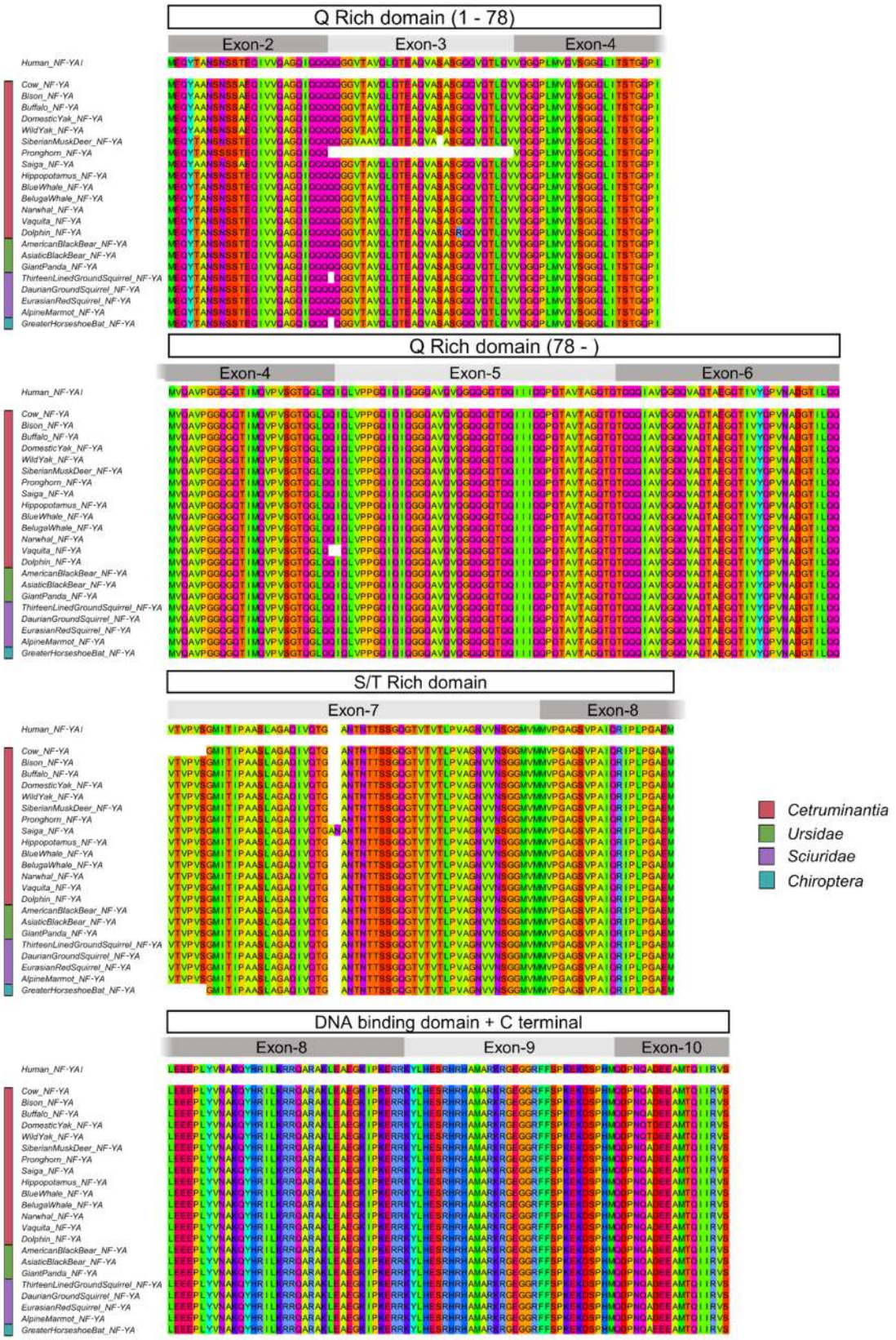


Figure S3. Alignment of NF-YA protein sequences in selected mammalian species.
Multiple protein sequence alignment of four NF-YA regions, as indicated by white labels. Regions encoded by each exon are indicated by grey boxes. Human NF-YAI was selected as reference. The alignment was exported from the software Jalview.

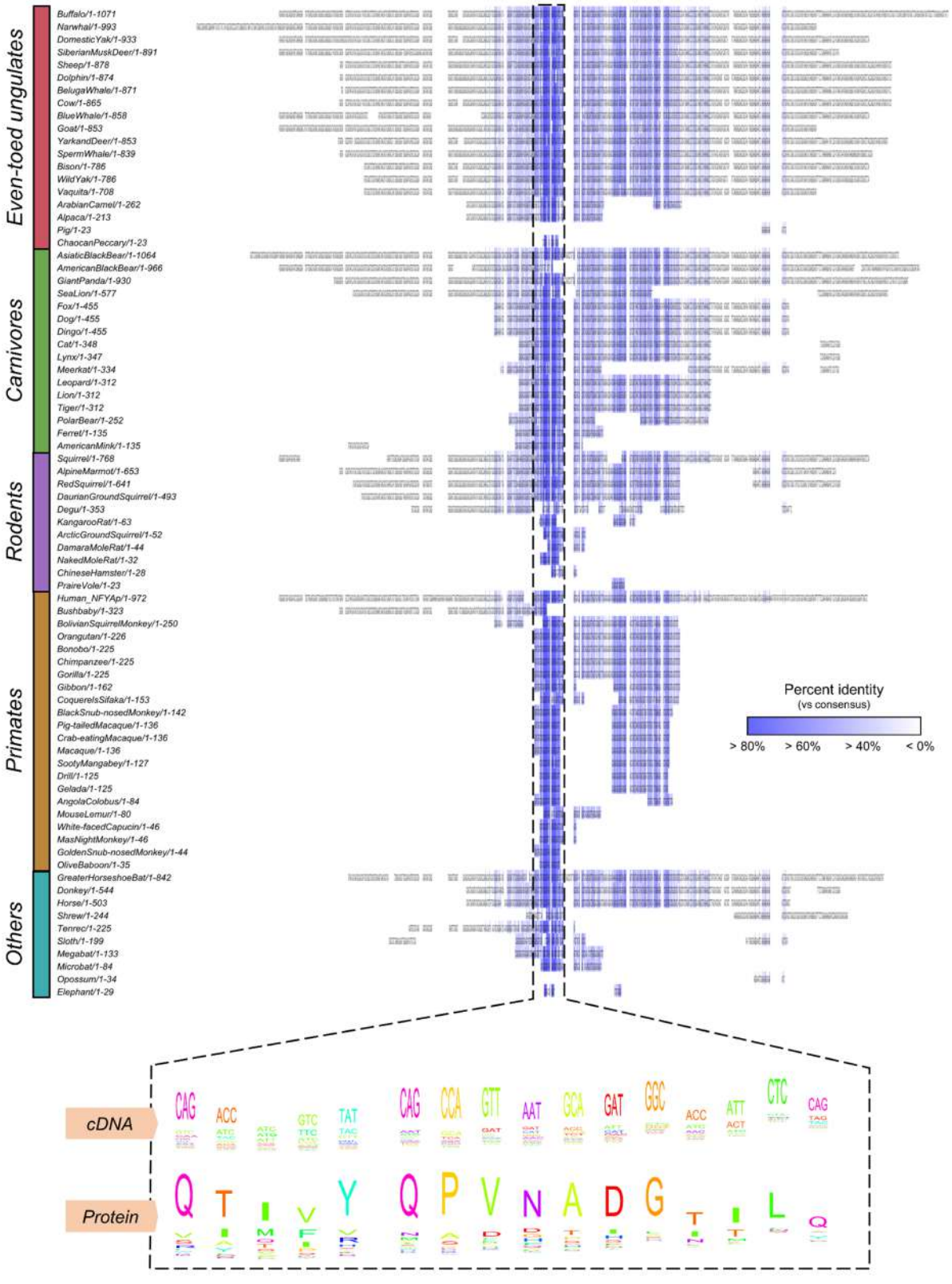


Figure S4. Alignment of *NF-YAr* orthologs in mammalians.
Multiple sequence alignment of DNA regions with a significant degree of homology to *NF-YAr*, procured through TBLASTN or BLAT analyses. The length of each sequence is indicated. Intensity of the color = per-nucleotide percent identity compared to the whole alignment consensus. Bottom: codon and protein sequence of the region with the highest identity (~ 80%). The alignment was exported from the software Jalview.

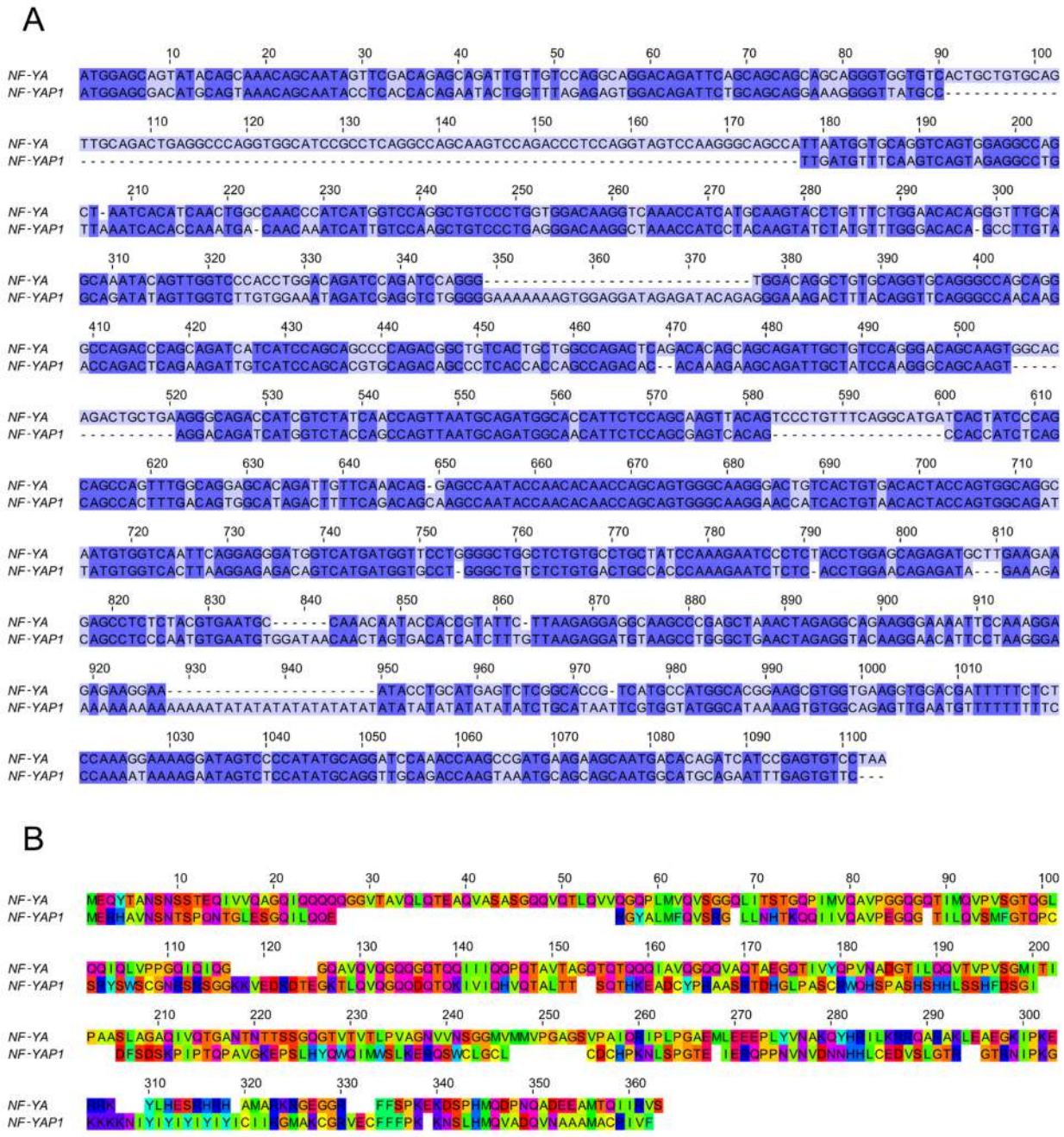


Figure S5. Alignment of human *NF-YA* and *NF-YAP1*.
A. MSA of H.*sapiens NF-YA* and *NF-YAP1* cDNA, as annotated in the UCSC genome browser. Intensity of the color = per-nucleotide percent identity compared to the alignment consensus. **B.** Same as A., except NF-YA and NF-YAP1 protein sequences are depicted. The alignment was exported from the software Jalview.

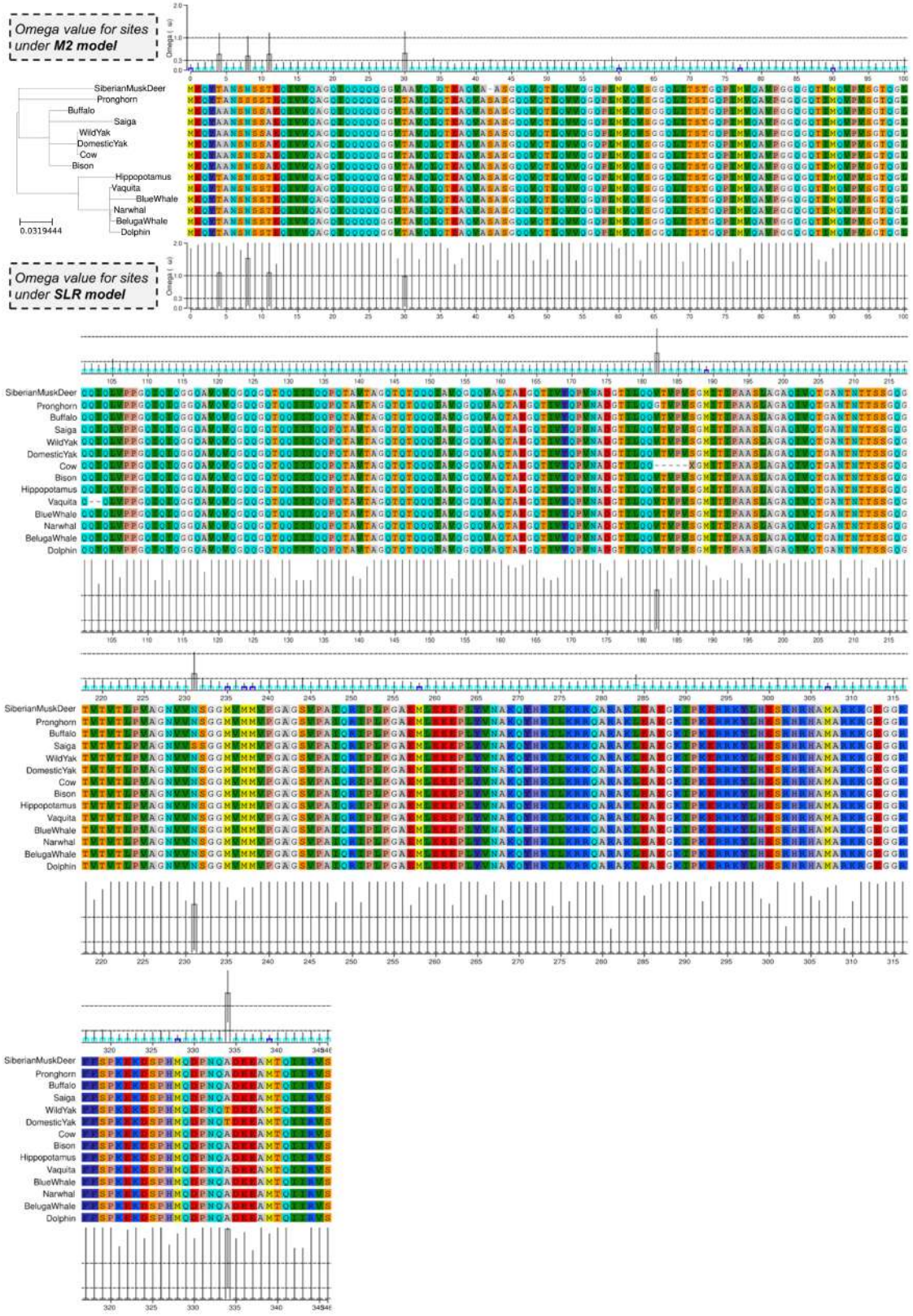


Figure S6. Conservation of NF-YA sequence in *Cetruminantia*.
A. Per-codon dN/dS ratios (ω) of NF-YA for the *Cetruminantia* species included in the analysis. Above the alignment, computed ω values underly the M2 evolutionary model (positive selection) are depicted as barplots, while those according to the SLR model (positive/purifying selection) are shown below. Error bars = standard error of the mean; branch length = number of nucleotide substitutions per codon. The alignment was generated by ETE toolkit.

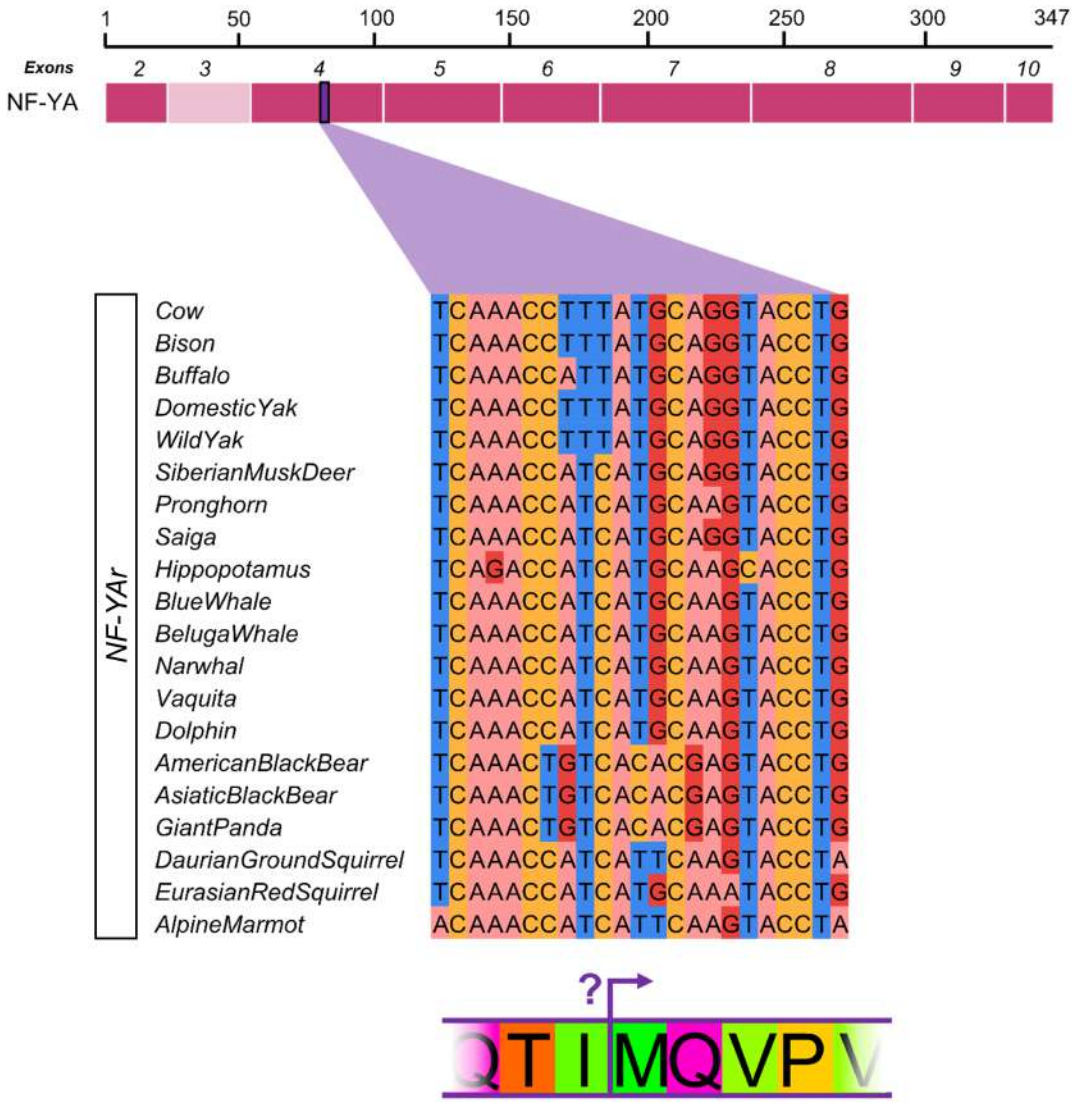


Figure S7. Conservation of *NF-YAr* M14 Kozak sequence.
Top: Schematic representation of NF-YA mature mRNA. The position of an ATG codon, NF-YAr M63, is marked. Bottom: MSA of *NF-YAr* sequences containing bases spanning from -10 to +10 from the selected ATG codon. Purple boxes: first five translated amino acids of NF-YA protein. The alignment was exported from the software Jalview.

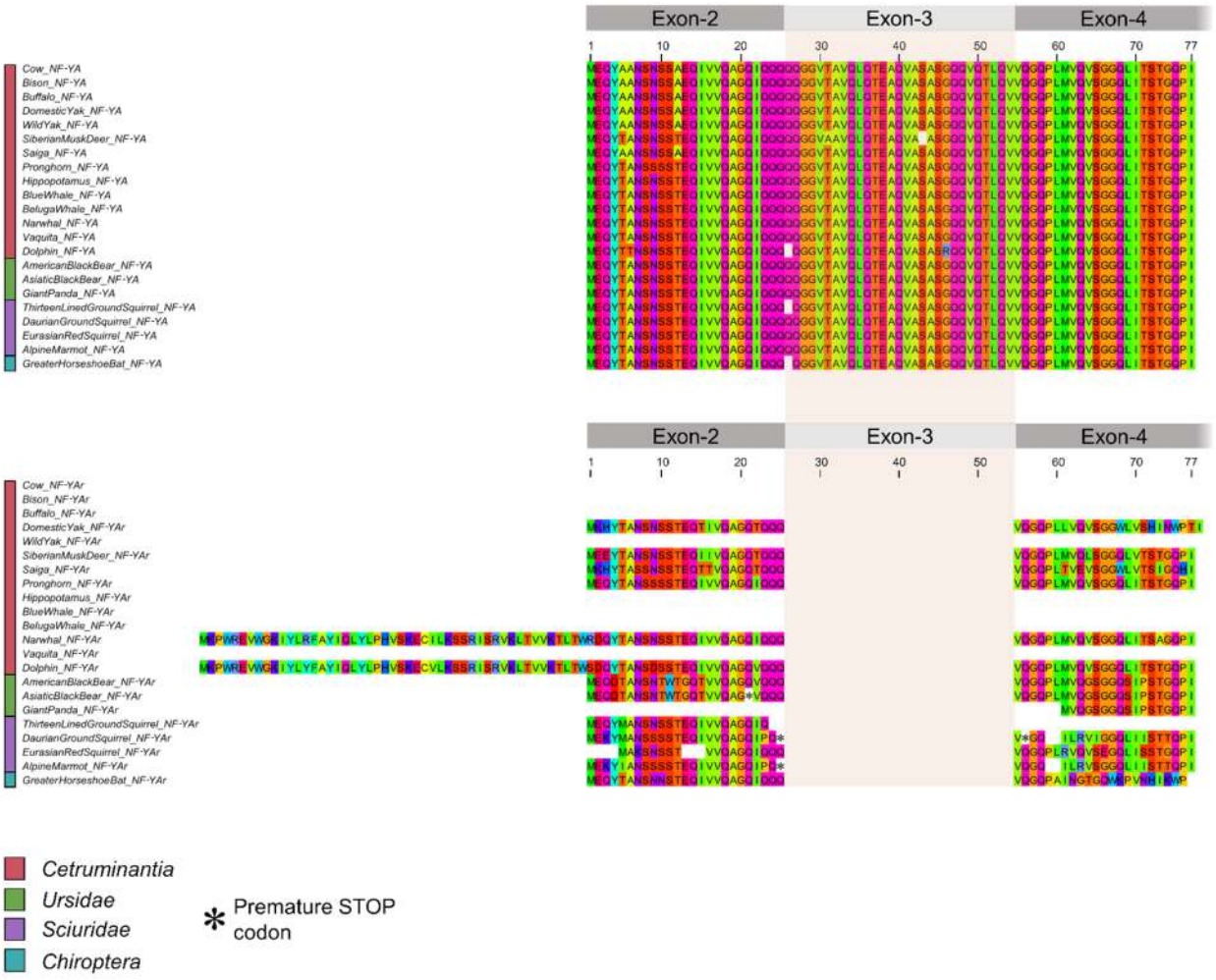


Figure S8. Alignment of NF-YA and NF-YAr N term.
Multiple sequence alignment of NF-YA (Top) and NF-YAr (Bottom) N terminus. Regions encoded by each exon are indicated by grey boxes. Premature stop codons are signaled within the alignment by an asterisk, and the orange area highlight the region encoded by Exon-3. The alignment was exported from the software Jalview.

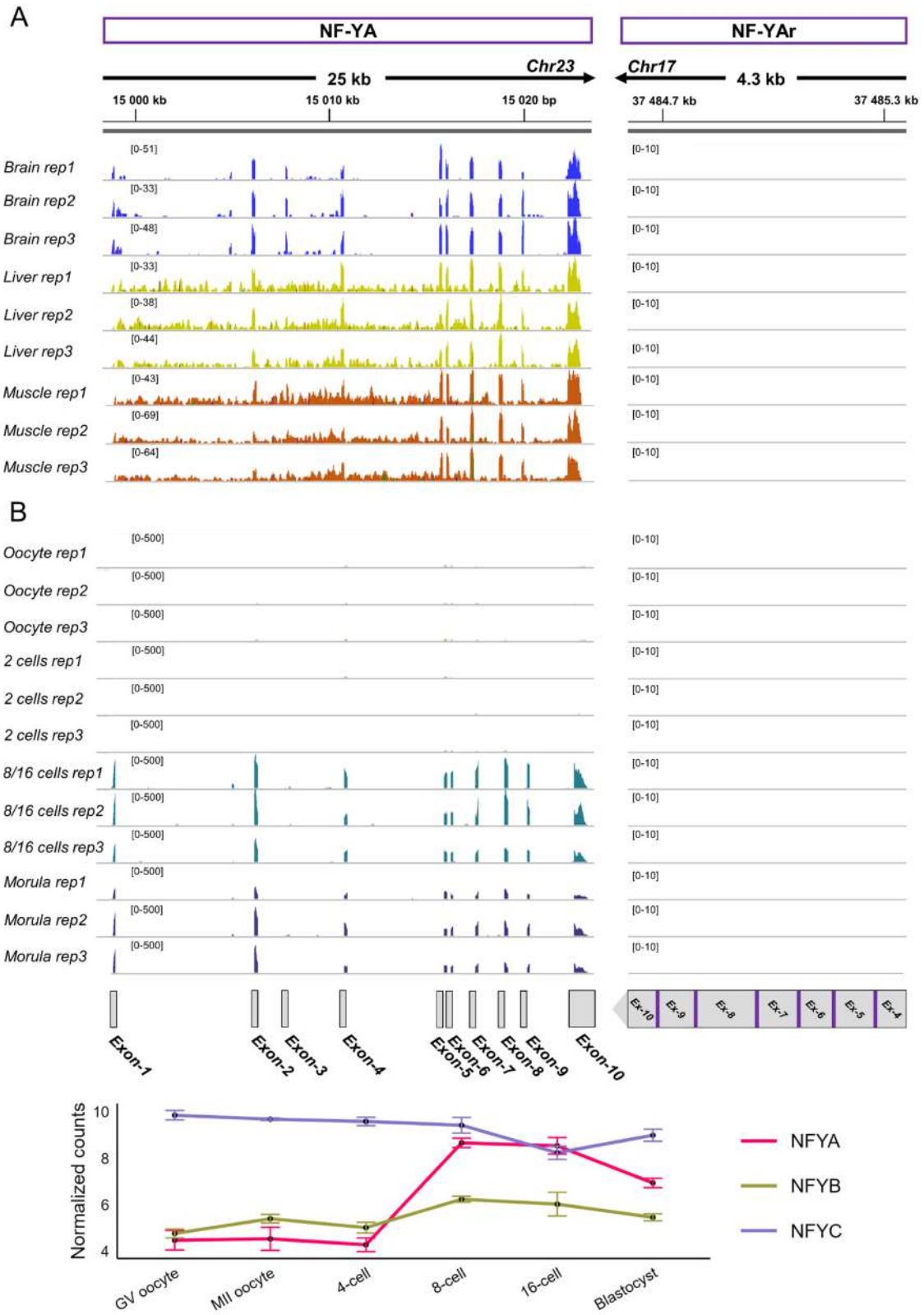


Figure S9. NF-Y expression in multiple cow RNA-seq samples.
A. RNA-seq analysis of NF-YA and NF-YAr expression from *Bos taurus* adult tissues (brain, liver, skeletal muscle), represented by mapped read coverage. **B.** *NF-YA* and *NF-YAr* expression from several *Bos taurus* embryo stages, represented by mapped read coverage (Top). Expression of NF-Y subunits by normalized counts from the GEO dataset GSE52415 (Bottom). Mapped reads were visualized with the IGV software, and the regions corresponding to each exon are underlined by grey boxes. On the left of each sample track, read number ranges are depicted as [min-max].

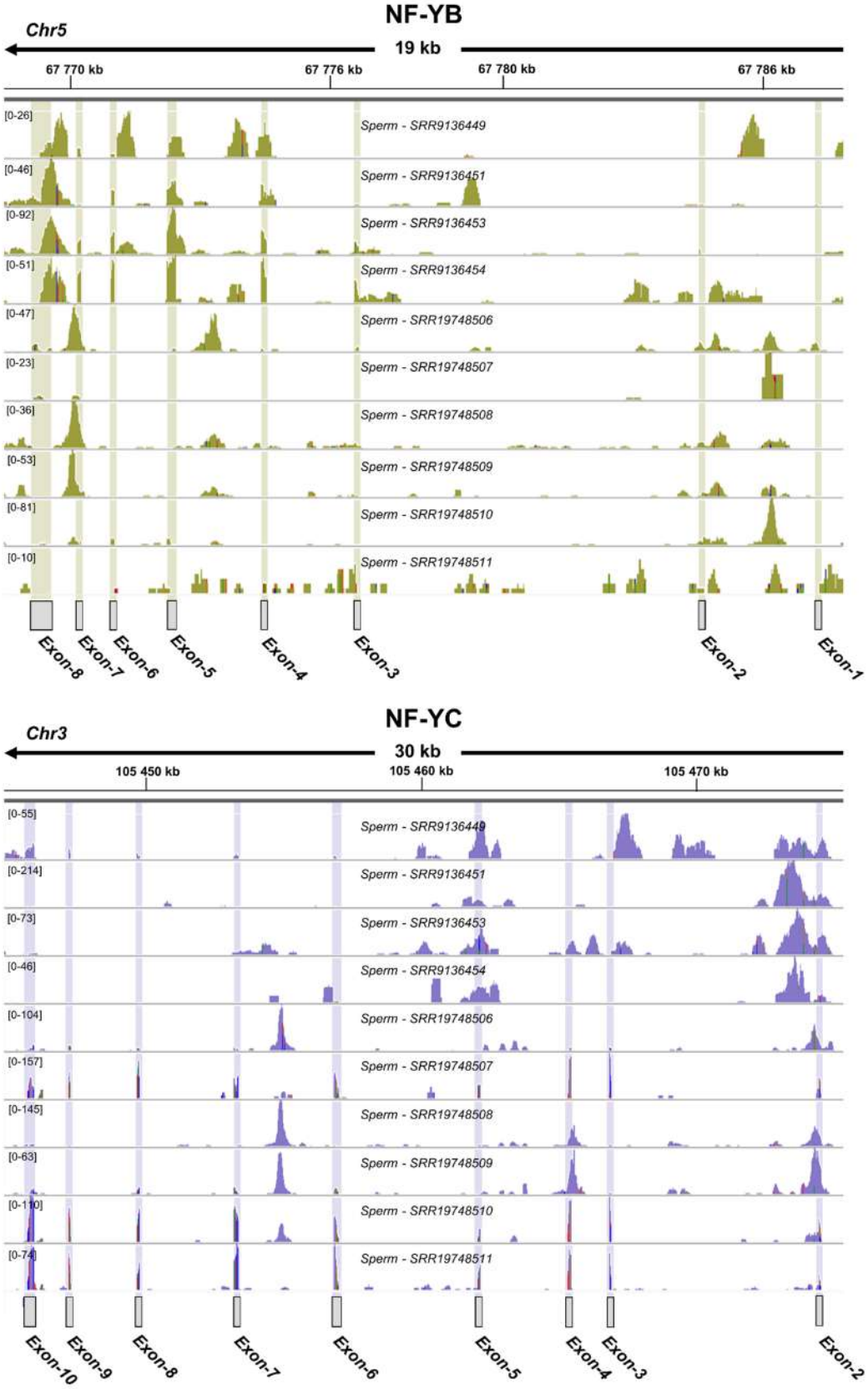


Figure S10. HFD NF-Y subunits expression in cow sperm RNA-seq samples.
A. *NF-YB* (Top) and *NF-YC* (Bottom) expression from *Bos taurus* sperm RNA-seq samples, represented by mapped read coverage. Mapped reads were visualized with the IGV software, and the regions corresponding to each exon are underlined by grey boxes. On the left of each sample track, read number ranges are depicted as [min-max].

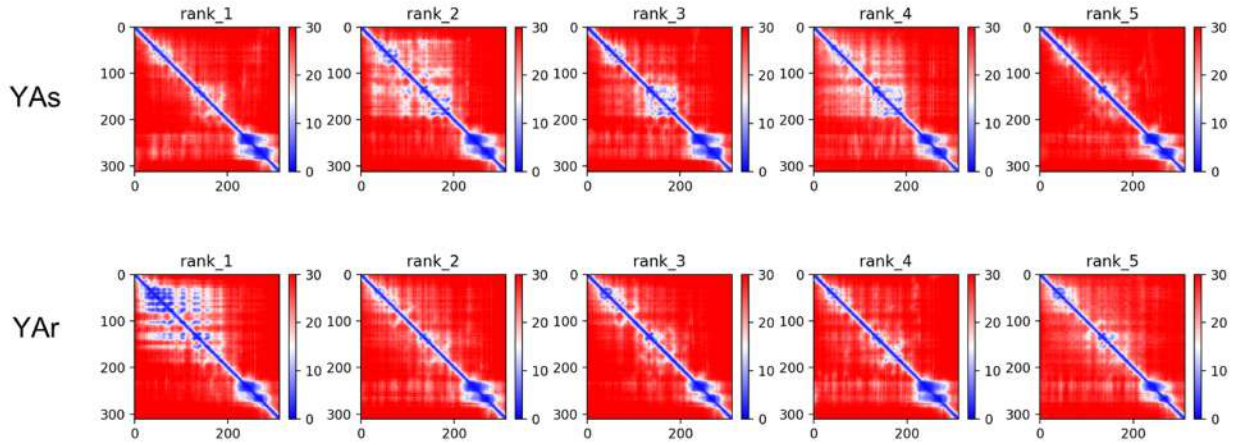


Figure S11. Domestic Yak NF-YA AF models PAE plots.
Predicted Aligned Errors (PAE) plots depict predicted errors for reciprocal distances of each residue along the polypeptide chain within each of the five AF computed 3D structure models. Higher confidence on the relative positioning within the model, or low PAE values, are colored in blue, while higher predicted error is colored in red. Domestic Yak NF-YA structures displayed in Figure 7 represent the AF rank 4 model for NF-YAs, and the rank 1 model (best model) for NF-YAr.

Conclusions & Future perspectives 4

In this Thesis I demonstrated the direct involvement of NF-YA long isoform, characterized by the inclusion of the 28/29 amino acids exon-3 in the mature transcript, in directing a mesenchymal differentiation within cancer metastatic dissemination and embryonic development and organogenesis, two apparently unrelated systems. In actuality, as the molecular basis of the process, EMT progresses in both scenarios along the same signaling pathways like TGF-β, Wnt, and Notch, and triggers the activity of the same master TFs, with SNAI and TWIST factors simultaneously repressing epithelial markers and activating mesenchymal genes, thus promoting cell phenotypic transformation[105].

It is still unclear how the integration of the exon-3 encoded region within the Q-rich TAD domain of NF-YA could impart such specific regulation program, but it certainly represents a remarkable testament to the impact of alternative splicing on genes functional repertoire. Given that the most plausible role for NF-Y in transcription regulation is of a pioneer complex that offers a nucleosome-free, CCAAT-containing platform within promoters and enhancers for other specialized TFs to operate on[177, 178], exon-3 encoded region could simply facilitate NF-Y interaction with TFs driving mesoderm differentiation. Our lab investigation on the NF-Y regulome did not show any co-association with SNAI/TWIST factors[190], but the CCAAT box is consistently included in *SNAI1/2/3* gene promoters at positions consistent with NF-Y regulation of expression (Not Shown). Due to technical limitations related to the production and the efficacy of dedicated antibodies, we still cannot perform NF-YA isoform-specific ChIP-seq experiments, which could definitely shine a light on the subject in the near future.

The exploration of NF-Y in stomach adenocarcinoma (STAD) revealed elevated NF-Y subunits expression levels, which, based on the previous analyses of our group, is a somewhat unusual occurrence. Typically, only the regulatory NF-YA is overexpressed, in line with the assumption that NF-YA concentration is the limiting factor for the trimer activity[180]. As for NF-YA isoforms, NF-YAs levels were highly enhanced compared to normal tissue, but only $\text{NFYA}_{\text{Ratio}}$ correlated with poor prognosis. A $\text{Claudin}^{\text{low}}$ subtype was also characterized, having high $\text{NF-YA}_{\text{Ratio}}$ scores and being associated with unique functional traits, with enrichment in EMT-related Gene Ontology terms. Notably, the main functions of NF-Y did not appear to be abolished within this $\text{Claudin}^{\text{low}}$ node, since many genes overexpressed in $\text{Claudin}^{\text{low}}$ when compared to normal tissues were still associated with cell-cycle related GO terms, although we also observed a diffused downregulation of terms linked to the regulation of lipid and sugar metabolisms. In the second project focused on the molecular determinants of $\text{Claudin}^{\text{low}}$ subtype, NF-YA exon-3 knockout in two different breast cancer $\text{Claudin}^{\text{low}}$ cell lines led to decreased migratory/metastatic capabilities, providing further *in vivo* and *in vitro* evidence of the intrinsic association between high $\text{NF-YA}_{\text{Ratio}}$ and tumor invasiveness (**Figure 4.1A**).

As a result of extensive *in silico* analyses, we then proposed a prognostic 158-gene signature associated with $\text{NF-YA}_{\text{Ratio}}$, again enriched in mesoderm related terms. This set of concurrently deregulated genes has a strikingly narrow overlap with the BRCA and STAD $\text{Claudin}^{\text{low}}$ signatures[119, 126]: this could serve as validation for the existence of distinct "flavours" of the $\text{Claudin}^{\text{low}}$ phenotype, as postulated by Fougner et al. in 2020[123]. In this sense, our findings delineate a $\text{Claudin}^{\text{low}}$-adjacent, but not $\text{Claudin}^{\text{low}}$-defining, cohesive group of genes that might experience shared dysregulation in epithelial tumors, irrespective of the tissue of origin.

Similarly to BRCA and STAD, NF-YAl and NF-YA$_{\text{Ratio}}$ were identified within the same co-expression module of genes in bladder cancer (BLCA), for which a Claudin$^{\text{low}}$ signature has already been advanced[125]. This urges the need for a comprehensive, NF-Y-directed investigation of BLCA, akin to those presented here.

Upon single cell RNA-seq deconvolution of breast and gastric cancers, which started from two sizeable, previously published scRNA-seq datasets[124, 245], high NF-YA$_{\text{Ratio}}$ correlated with an expanded EMT/Claudin$^{\text{low}}$ cancer cells and CAFs populations. The former composed a cluster that showed a clear distinction from epithelial and fibroblast clusters, and strongly correlated with the NF-YA$_{\text{Ratio}}$ signature. The latter could be consistent with the established role CAFs play within the tumor microenvironment, promoting angiogenesis, extracellular matrix rearrangements that favour invasion, and even drug resistance by inducing suppressive T lymphocytes and/or physically preventing immune infiltration[252]. Finally, the rate of cancer cells within the Claudin$^{\text{low}}$ group predicted by CNV evaluation appeared coherent with the one of epithelial cancer cells, in both cohorts. Authors of the original scRNA-seq atlas of breast cancer assigned most cells included in the Claudin$^{\text{low}}$ group I predicted to mesenchymal stem cells (MSCs), but included in the inferCNV call only cells of epithelial lineage[124]. For the future, it would be intriguing to elucidate the precise expression profile of these cells, possibly coinciding with the one of CSCs.

Our ongoing effort revolves around the recognition of Claudin$^{\text{low}}$-specific alternative splicing patterns, and originated with the integration of isoform and exon analyses resulted in a signature of AS events involving cancer hallmark genes. The end goal of this project is twofold: on one end I will pursue a reliable molecular predictive model for the Claudin$^{\text{low}}$ subtype, likely incorporating the activity of the EMT-associated RBPs that I recently linked to the splicing of NF-YA transcript. Additionally, this could also lead to a novel list of splicing exon targets for molecular therapeutic strategies, like newly designed ASOs. One potential caveat of this project is the reliance on short-reads RNA-seq databases, less precise than long read sequencing technologies for studying AS. Conversely, even employing a hybrid strategy, in which a *bona fide* set of transcripts generated by long-read sequencing is subsequently quantified by short-reads RNA-seq, may lack the depth and variability offered by repositories like TCGA.

In the phylogenetic analyses of NF-YA, we ultimately postulated that NF-YAs poses as the original alternative splicing isoform for the *Deuterostomia* group, based on conservation data from echinoderms and hemichordates, lacking the exon-3 encoded region of the TAD, and from the cephalochordate amphioxus, which instead harbors a divergent exon-3 sequence. Exon-3 inclusion is limited during development, emphasizing NF-YAs predominance in embryogenesis. NF-YAl instead prevailed in mouse mesoderm during gastrulation, and in adult muscle and brain of most vertebrate species considered. The exception represented by the axolotl, where no exon-3 expression is ever scored, could be somehow connected to its regenerating capabilities, for which the animal is studied as model organism[253].

Surprisingly, *Aves* displayed an unique NF-YA isoform lacking exon-5, called NF-YAg. Both NF-YAg and NF-YAx, missing both exon-3 and exon-5, were expressed in birds but absent in *Reptilia* samples, suggesting their consistent and close-to exclusive inclusion in the AS collection of the Aves class (**Figure 4.1B**). We speculated that NF-YAg can function similarly to NF-YAx[195], acting as a dominant negative (DN) for NF-YAl and NF-YAs, forming a trimer with NF-YB/NF-YC subunits that binds DNA but is impaired in transcription activation. NF-YAg may also activate only a limited

pool of genes, resembling an intermediate state between fully active NF-YAs and the partially inactive/DN NF-YAx. Nevertheless, the discovery of NF-YAg opens new avenues regarding its plausible implications in bird embryo development.

A similar DN activity could be imagined also for NF-YAr, the retrotransposed, possibly translated NF-YA pseudogene we recently described, conserved mostly between the boundaries of the *Cetruminantia* clade. Particularly, the higher degree of identity of subunits interaction domain residues, especially when compared to the DBD, portraits a slightly different inhibition mechanism, where NF-YAr sequesters the HFD and prevents proper trimer formation.

As NF-YC is equally subjected to alternative splicing, with an even higher isoform variability compared to what I described here for NF-YA, a similar systematic investigation on the NF-YC locus evolution should be accounted by a future phylogenetic project.

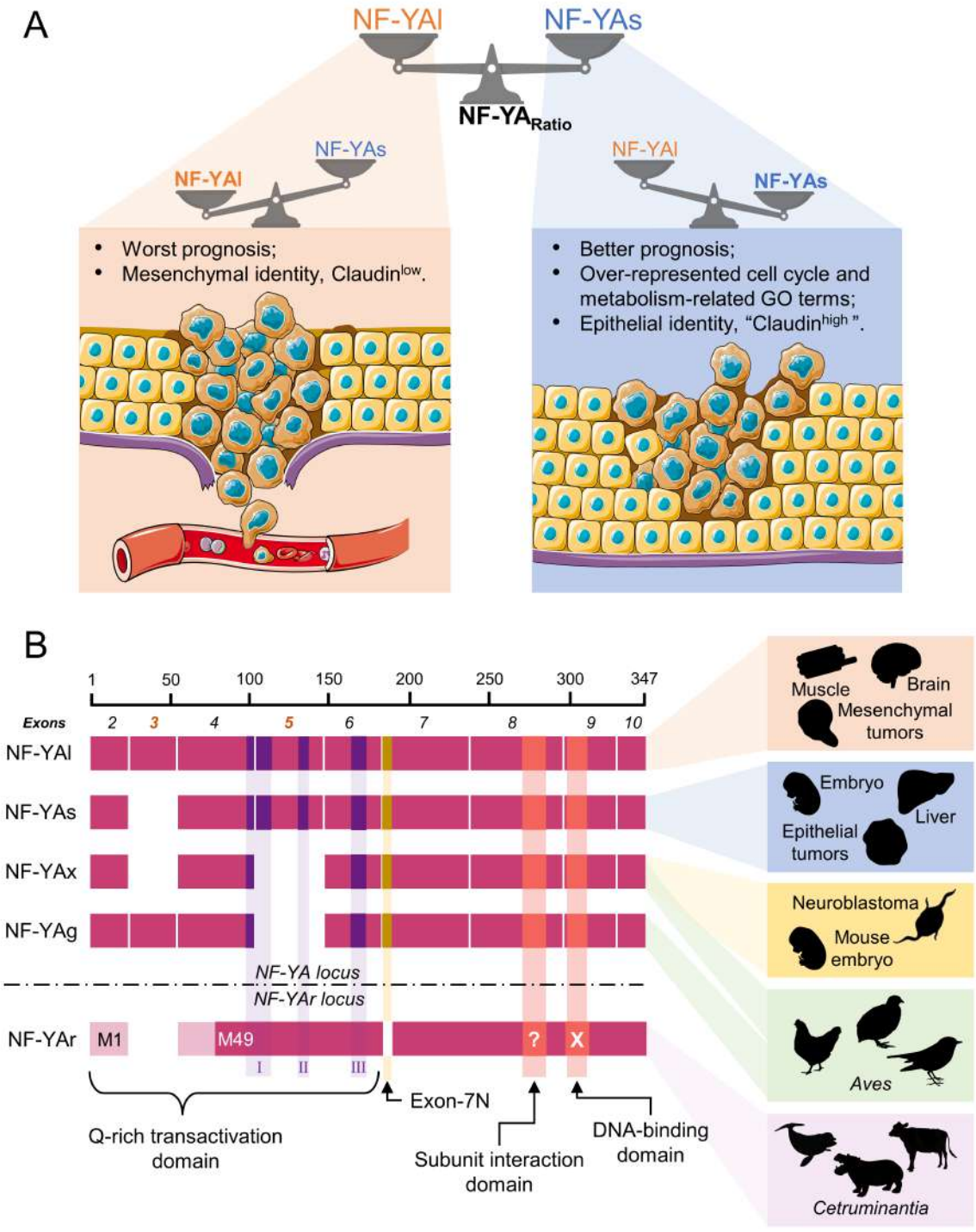


Figure 4.1: NF-YA isoforms: impact on cancer & phylogeny.
A. *The model presents NF-YA$_{Ratio}$ influence on epithelial cancers as a balance between the expression of the two main isoforms, in association with distinct molecular and phenotypical features.* ***B***. *All NF-YA alternative splicing variants discussed throughout the Thesis are included in this alignment, with their structural differences highlighted. On the* ***Right****, the silhouettes provide a summary of the tissues/organisms in which each isoform was reported or is chiefly expressed. Parts of the figure were drawn by using pictures from Servier Medical Art.*

5 Materials & Methods

Below, I outline the methodologies and workflows used to generate the currently unpublished or unsubmitted data presented in the Results chapter. For additional details on the remaining data discussed in this Thesis, please refer to the Materials & Methods sections of the attached papers.

5.1. Section 3.1.4

5.1.1. scRNA-seq gene expression data acquisition and processing

Raw counts from two single cell RNA-seq experiments in breast and gastric cancers were obtained from NCBI Gene Expression Omnibus (GEO) under accessions GSE176078[124] and GSE150290[245], respectively. The count matrices were filtered using Seurat (version 5.0.1)[254], retaining cells with 200 to 2500 detected genes, less than 10% of total counts mapped to mitochondrial genes, and genes expressed in at least 3 distinct cells. Data was then normalized, scaled, subjected to cell clustering using a `resolution` parameter of 0.5, and to the uniform manifold approximation and projection (UMAP) technique for dimensionality reduction. Cell typing was performed using scTyper (version 0.1.0)[255] with the Nearest Template Prediction (NTP) method, employing the markers from original papers, when available. Additionally, markers for Basal, Luminal (A and B), and HER2-enriched tumors, as well as Claudin^{low} signatures from Prat et al.[119] and Nishijima et al.[126], were included for breast and gastric cancer samples typing, respectively.

5.1.2. Per-cell expression data visualization

The $\log_2$-normalized expression levels of epithelial markers (*EPCAM*, *KRT8*, *KRT18*, and *KRT19*) and fibroblast markers (*FN1* and *FAP*) were visualized in UMAP plots using the Seurat `FeaturePlot` function. Cut-off points were set at the 90th and 10th quantiles. Correlation of single cells with the Claudin^{low} signature derived from BRCA and STAD WGCNA analysis (see section **3.1.2**) was calculated using the `AddModuleScore` function. These correlations were then visualized using `FeaturePlot` with the same cut-off values as used for the marker panels.

5.1.3. CNV prediction and cancer cell status assignment

I utilized the `inferCNV` R package (version 1.14.2)[1] to evaluate copy number variation in BRCA and STAD cells classified either as epithelial cancer or Claudin^{low} cancer cells. For BRCA, normal epithelial cells served as the reference, while gland/pit mucous cells were selected for STAD. This process was repeated for each sample from the two original experiments, excluding those lacking suspected malignant cells. The full command used is provided below:

[1] https://github.com/broadinstitute/inferCNV

```
infercnv::run(infercnv_obj,
  num_threads=8, out_dir=paste0("./Out_", sample_name),
  cutoff=0.1, window_length=101, max_centered_threshold=3,
  cluster_by_groups=F, plot_steps=F, denoise=T,
  write_expr_matrix = T, sd_amplifier=1.3,
  analysis_mode = "samples")
```

The normal/cancer cell determination procedure based on CNV values closely followed the methodology outlined by Wu et al.[124]: CN changes at each genomic locus were scaled between $-1:+1$, and the mean of their squares determined the genomic instability score ($\overline{CNA^2}$) for each cell. These were then correlated with the top 5% of cells with the highest scores in each tumor sample using the Kendall method, and this correlation (Cor) was used along with $\overline{CNA^2}$ for cancer cell determination.

Partitioning around medoids clustering was then performed using the `pamk` R function to find the optimal number of clusters ($2 \leq k \leq 4$), and validated through silhouette scores. To be classified as neoplastic, a cell $\overline{CNA^2}$ needed to be 2 standard deviations above the first partitioning around medoids cluster mean, while the Cor value had to exceed 1.5 standard deviations. In cases with only one cluster, the threshold was set at 1 standard deviation above the mean for both metrics. If neither of the two requirements were met, a cell was labelled as normal, in all other cases it was considered "unassigned".

5.2. Section 3.1.5

5.2.1. Retrieval and handling of STAD exon-level expression data

I downloaded exon-level raw count data for the TCGA STAD cohort from the web page http://firebrowse.org/, for a total of 415 primary tumour samples. In our previous reclassification procedure based on the ACRG molecular subtypes I assigned 399 of these samples to one of the four original groups, EMT, MSI, MSS;TP53$^-$, and MSS;TP53$^+$, and the hierarchical clustering I conducted using the corresponding 24-gene signature[126] for STAD isolated a 79-samples Claudin^{low} node (see section **3.1.1**). After filtering out unclassified samples, I converted interval labels (in the format chromosome:start-end:strand) to exons (transcript:exon number) using a custom Python script (version 3.8). This script took as input the count matrix and the hg19 GTF file adopted during the analysis that produced the exon-level expression data (this file was deposited as "unc_hg19.gtf" in the directory https://webshare.bioinf.unc.edu/public/mRNAseq_TCGA/rsem_ref). Finally, lines corresponding to monoexonic genes were removed from the complete count matrix.

5.2.2. Significant Claudin^{low}-specific splicing event detection using satuRn

My colleague and I conducted parallel analyses for examining differential exon usage (DEU) and differential transcript usage (DTU). We utilized the R packages satuRn (version 1.6.0)[249] for DEU, and DTUrtle (version 1.0.2)[248] for the DTU procedure. satuRn required raw exon-level counts as input, while DTUrtle used the estimated counts and scaled TPM values computed by the RSEM pipeline[256]. Default parameters were employed for both analyses, focusing on the same pairwise comparisons, contrasting Claudin^{low} samples with all other subtypes. Results were filtered based on a false discovery rate threshold of 0.05, and only values meeting this criterion across all four comparisons

were retained. We then integrated the findings from both analyses, identifying differentially used exons included in differentially used transcripts for further investigation.

5.2.3. Functional analysis and exon-level expression visualization

The MSigDB Hallmark enrichment analysis was carried out on the genes that included both DEU and DTU filtered results, using the Enrichr web resource[2]. Graphical representation of annotated transcript and coding sequences was obtained with the ggtranscript R package (version 0.99.9)[3]. For the boxplots of **Figure 3.2D**, estimated counts were normalized with the R package DESeq2 (version 1.38.3)[257], while for the heatmaps, created with the pheatmap package (version 1.0.12)[4], I divided exon-level raw counts by total per-gene counts to compute each exon's contribution to total expression at each locus. These contribution fractions were then scaled across all classified STAD samples.

[2]https://maayanlab.cloud/Enrichr
[3]https://github.com/dzhang32/ggtranscript
[4]https://github.com/raivokolde/pheatmap

6 References

[1] D. S. Latchman. “Transcription factors: an overview.” In: *International Journal of Experimental Pathology* 74.5 (Oct. 1993), pp. 417–422.

[2] Juan M. Vaquerizas et al. “A census of human transcription factors: function, expression and evolution”. In: *Nature Reviews. Genetics* 10.4 (Apr. 2009), pp. 252–263.

[3] Mary C. Thomas and Cheng-Ming Chiang. “The General Transcription Machinery and General Cofactors”. In: *Critical Reviews in Biochemistry and Molecular Biology* 41.3 (Jan. 1, 2006), pp. 105–178.

[4] Seth Frietze and Peggy J. Farnham. “Transcription Factor Effector Domains”. In: *A Handbook of Transcription Factors*. Ed. by Timothy R. Hughes. Subcellular Biochemistry. Dordrecht: Springer Netherlands, 2011, pp. 261–277.

[5] Ingrid E. Akerblom et al. “Negative Regulation by Glucocorticoids Through Interference with a cAMP Responsive Enhancer”. In: *Science* 241.4863 (July 15, 1988), pp. 350–353.

[6] Glenn A. Maston, Sara K. Evans, and Michael R. Green. “Transcriptional Regulatory Elements in the Human Genome”. In: *Annual Review of Genomics and Human Genetics* 7.1 (2006), pp. 29–59.

[7] Kazutoshi Takahashi and Shinya Yamanaka. “A decade of transcription factor-mediated reprogramming to pluripotency”. In: *Nature Reviews. Molecular Cell Biology* 17.3 (Mar. 2016), pp. 183–193.

[8] Ada Hamosh et al. “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders”. In: *Nucleic Acids Research* 33 (Database Issue Jan. 1, 2005), p. D514.

[9] Simon J. Furney et al. “Structural and functional properties of genes involved in human cancer”. In: *BMC genomics* 7 (Jan. 11, 2006), p. 3.

[10] John H. Bushweller. “Targeting transcription factors in cancer — from undruggable to reality”. In: *Nature Reviews Cancer* 19.11 (Nov. 2019), pp. 611–624.

[11] Christian Tovar et al. “MDM2 small-molecule antagonist RG7112 activates p53 signaling and regresses human tumors in preclinical cancer models”. In: *Cancer Research* 73.8 (Apr. 15, 2013), pp. 2587–2597.

[12] Yali Xu and Christopher R. Vakoc. “Targeting Cancer Cells with BET Bromodomain Inhibitors”. In: *Cold Spring Harbor Perspectives in Medicine* 7.7 (July 2017), a026674.

[13] Samuel A. Lambert et al. “The Human Transcription Factors”. In: *Cell* 172.4 (Feb. 8, 2018), pp. 650–665.

[14] Sean B. Carroll. “Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution”. In: *Cell* 134.1 (July 11, 2008), pp. 25–36.

[15] Kierra A. Franklin, Cara E. Shields, and Karmella A. Haynes. “Beyond the marks: reader-effectors as drivers of epigenetics and chromatin engineering”. In: *Trends in Biochemical Sciences* 47.5 (May 1, 2022), pp. 417–432.

[16] Claire Larroux et al. “Genesis and Expansion of Metazoan Transcription Factor Gene Classes”. In: *Molecular Biology and Evolution* 25.5 (May 1, 2008), pp. 980–996.

[17] Luis A. Barrera et al. “Survey of variation in human transcription factors reveals prevalent DNA binding changes”. In: *Science* 351.6280 (Mar. 25, 2016), pp. 1450–1454.

[18] Matthew T. Weirauch and Timothy R. Hughes. “Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same”. In: *Trends in Genetics* 26.2 (Feb. 2010), pp. 66–74.

[19] Shannon Fisher et al. “Conservation of RET regulatory function from human to zebrafish without sequence similarity”. In: *Science (New York, N.Y.)* 312.5771 (Apr. 14, 2006), pp. 276–279.

[20] Michael Levine, Claudia Cattoglio, and Robert Tjian. “Looping Back to Leap Forward: Transcription Enters a New Era”. In: *Cell* 157.1 (Mar. 27, 2014), pp. 13–25.

[21] Kazuhiro R Nitta et al. “Conservation of transcription factor binding specificities across 600 million years of bilateria evolution”. In: *eLife* 4 (Mar. 2015), e04837.

[22] Gregory A. Wray. “The evolutionary significance of cis-regulatory mutations”. In: *Nature Reviews Genetics* 8.3 (Mar. 2007), pp. 206–216.

[23] Frank W. Stearns. “One Hundred Years of Pleiotropy: A Retrospective”. In: *Genetics* 186.3 (Nov. 2010), pp. 767–773.

[24] Kristen Anne Panfilio and Michael Akam. “A comparison of Hox3 and Zen protein coding sequences in taxa that span the Hox3/zen divergence”. In: *Development Genes and Evolution* 217.4 (Apr. 2007), pp. 323–329.

[25] Jason Gertz et al. “Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner”. In: *Genome Research* 22.11 (Nov. 2012), pp. 2153–2162.

[26] Bahar Taneri et al. “Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific”. In: *Genome Biology* 5.10 (2004), R75.

[27] Thomas R. Bürglin. “The Hedgehog protein family”. In: *Genome Biology* 9.11 (Nov. 19, 2008), p. 241.

[28] U. Löhr, M. Yussa, and L. Pick. “Drosophila fushi tarazu. a gene on the border of homeotic function”. In: *Current biology: CB* 11.18 (Sept. 18, 2001), pp. 1403–1412.

[29] Vincent J. Lynch and Günter P. Wagner. “Resurrecting the role of transcription factor change in developmental evolution”. In: *Evolution; International Journal of Organic Evolution* 62.9 (Sept. 2008), pp. 2131–2154.

[30] Simon Erlendsson and Kaare Teilum. “Binding Revisited—Avidity in Cellular Function and Signaling”. In: *Frontiers in Molecular Biosciences* 7 (Jan. 14, 2021), p. 615565.

[31] Ekaterina Morgunova and Jussi Taipale. “Structural perspective of cooperative transcription factor binding”. In: *Current Opinion in Structural Biology*. Protein–nucleic acid interactions • Catalysis and regulation 47 (Dec. 1, 2017), pp. 1–8.

[32] Benjamin S. Scruggs et al. “Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin”. In: *Molecular Cell* 58.6 (June 18, 2015), pp. 1101–1112.

[33] Kenneth S. Zaret and Jason S. Carroll. “Pioneer transcription factors: establishing competence for gene expression”. In: *Genes & Development* 25.21 (Nov. 1, 2011), pp. 2227–2241.

[34] Laurie A. Boyer et al. “Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells”. In: *Cell* 122.6 (Sept. 23, 2005), pp. 947–956.

[35] Mike Levine. “Transcriptional enhancers in animal development and evolution”. In: *Current biology: CB* 20.17 (Sept. 14, 2010), R754–763.

[36] K. L. Clark et al. “Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5”. In: *Nature* 364.6436 (July 29, 1993), pp. 412–420.

[37] Nerina Gnesutta, Marco Nardini, and Roberto Mantovani. “The H2A/H2B-like histone-fold domain proteins at the crossroad between chromatin and different DNA metabolisms”. In: *Transcription* 4.3 (May 1, 2013), pp. 114–119.

[38] Thomas D. Schneider and R.Michael Stephens. “Sequence logos: a new way to display consensus sequences”. In: *Nucleic Acids Research* 18.20 (Oct. 25, 1990), pp. 6097–6100.

[39] Gary D. Stormo. “DNA binding sites: representation and discovery”. In: *Bioinformatics* 16.1 (Jan. 1, 2000), pp. 16–23.

[40] Stein Aerts. “Chapter five - Computational Strategies for the Genome-Wide Identification of cis-Regulatory Elements and Transcriptional Targets”. In: *Current Topics in Developmental Biology*. Ed. by Serge Plaza and François Payre. Vol. 98. Transcriptional Switches During Development. Academic Press, Jan. 1, 2012, pp. 121–145.

[41] Arttu Jolma et al. “DNA-Binding Specificities of Human Transcription Factors”. In: *Cell* 152.1 (Jan. 17, 2013), pp. 327–339.

[42] Giovanna Ambrosini et al. “Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study”. In: *Genome Biology* 21.1 (May 11, 2020), p. 114.

[43] Rita Das et al. “Functional coupling of RNAP II transcription to spliceosome assembly”. In: *Genes & Development* 20.9 (May 1, 2006), pp. 1100–1109.

[44] Martin J. Hicks et al. “Linking splicing to Pol II transcription stabilizes pre-mRNAs and influences splicing patterns”. In: *PLoS biology* 4.6 (June 2006), e147.

[45] Anand Ramanathan, G. Brett Robb, and Siu-Hong Chan. “mRNA capping: biological functions and applications”. In: *Nucleic Acids Research* 44.16 (Sept. 19, 2016), pp. 7511–7526.

[46] Xiangyue Wu and Gary Brewer. “The Regulation of mRNA Stability in Mammalian Cells: 2.0”. In: *Gene* 500.1 (May 25, 2012), pp. 10–21.

[47] Flavio Mignone et al. “Untranslated regions of mRNAs”. In: *Genome Biology* 3.3 (Feb. 28, 2002), reviews0004.1.

[48] Marina Reixachs-Solé and Eduardo Eyras. “Uncovering the impacts of alternative splicing on the proteome with current omics techniques”. In: *WIREs RNA* 13.4 (2022), e1707.

[49] John D. Blair et al. “Widespread Translational Remodeling during Human Neuronal Differentiation”. In: *Cell Reports* 21.7 (Nov. 14, 2017), pp. 2005–2016.

[50] Laura M. Agosto and Kristen W. Lynch. “Alternative pre-mRNA splicing switch controls hESC pluripotency and differentiation”. In: *Genes & Development* 32.17 (Sept. 1, 2018), pp. 1103–1104.

[51] Robert K. Bradley and Olga Anczuków. “RNA splicing dysregulation and the hallmarks of cancer”. In: *Nature Reviews Cancer* 23.3 (Mar. 2023), pp. 135–155.

[52] Cindy L. Will and Reinhard Lührmann. “Spliceosome Structure and Function”. In: *Cold Spring Harbor Perspectives in Biology* 3.7 (July 2011).

[53] Max E. Wilkinson, Clément Charenton, and Kiyoshi Nagai. “RNA Splicing by the Spliceosome”. In: *Annual Review of Biochemistry* 89.1 (2020), pp. 359–388.

[54] C J Coolidge, R J Seely, and J G Patton. “Functional analysis of the polypyrimidine tract in pre-mRNA splicing.” In: *Nucleic Acids Research* 25.4 (Feb. 15, 1997), pp. 888–896.

[55] M. Burset, I. A. Seledtsov, and V. V. Solovyev. “Analysis of canonical and non-canonical splice sites in mammalian genomes”. In: *Nucleic Acids Research* 28.21 (Nov. 1, 2000), pp. 4364–4375.

[56] Christopher R. Sibley, Lorea Blazquez, and Jernej Ule. “Lessons from non-canonical splicing”. In: *Nature Reviews Genetics* 17.7 (July 2016), pp. 407–421.

[57] Abhijit A. Patel and Joan A. Steitz. “Splicing double: insights from the second spliceosome”. In: *Nature Reviews Molecular Cell Biology* 4.12 (Dec. 2003), pp. 960–970.

[58] Zefeng Wang and Christopher B. Burge. “Splicing regulation: From a parts list of regulatory elements to an integrated splicing code”. In: *RNA* 14.5 (May 2008), pp. 802–813.

[59] Thomas Geuens, Delphine Bouhy, and Vincent Timmerman. “The hnRNP family: insights into their role in health and disease”. In: *Human Genetics* 135.8 (Aug. 2016), pp. 851–867.

[60] Jonathan M. Howard and Jeremy R. Sanford. “THE RNAissance Family: SR proteins as multifaceted regulators of gene expression”. In: *Wiley interdisciplinary reviews. RNA* 6.1 (Jan. 2015), pp. 93–110.

[61] Yeon Lee and Donald C. Rio. “Mechanisms and Regulation of Alternative Pre-mRNA Splicing”. In: *Annual review of biochemistry* 84 (2015), pp. 291–323.

[62] Jennifer C. Long and Javier F. Caceres. “The SR protein family of splicing factors: master regulators of gene expression”. In: *Biochemical Journal* 417.1 (Dec. 12, 2008), pp. 15–27.

[63] Robert B. Darnell. “RNA protein interaction in neurons”. In: *Annual Review of Neuroscience* 36 (July 8, 2013), pp. 243–270.

[64] Eric T. Wang et al. “Antagonistic regulation of mRNA expression and splicing by CELF and MBNL proteins”. In: *Genome Research* 25.6 (June 2015), pp. 858–871.

[65] Jormay Lim and Jean Paul Thiery. “Epithelial-mesenchymal transitions: insights from development”. In: *Development (Cambridge, England)* 139.19 (Oct. 2012), pp. 3471–3486.

[66] Yueqin Yang et al. “Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition”. In: *Molecular and Cellular Biology* 36.11 (June 1, 2016), pp. 1704–1719.

[67] E. Papaemmanuil et al. “Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts”. In: *The New England Journal of Medicine* 365.15 (Oct. 13, 2011), pp. 1384–1395.

[68] Kenichi Yoshida et al. “Frequent pathway mutations of splicing machinery in myelodysplasia”. In: *Nature* 478.7367 (Sept. 11, 2011), pp. 64–69.

[69] Lili Wang et al. “SF3B1 and other novel cancer genes in chronic lymphocytic leukemia”. In: *The New England Journal of Medicine* 365.26 (Dec. 29, 2011), pp. 2497–2506.

[70] J. William Harbour et al. “Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma”. In: *Nature Genetics* 45.2 (Feb. 2013), pp. 133–135.

[71] Marcin Imielinski et al. “Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing”. In: *Cell* 150.6 (Sept. 14, 2012), pp. 1107–1120.

[72] Stanley Chun-Wei Lee et al. “Synthetic Lethal and Convergent Biological Effects of Cancer-Associated Spliceosomal Gene Mutations”. In: *Cancer Cell* 34.2 (Aug. 13, 2018), 225–241.e8.

[73] Laura Urbanski, Nathan Leclair, and Olga Anczuków. “Alternative-splicing defects in cancer: splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics”. In: *Wiley interdisciplinary reviews. RNA* 9.4 (July 2018), e1476.

[74] Rotem Karni et al. “The gene encoding the splicing factor SF2/ASF is a proto-oncogene”. In: *Nature Structural & Molecular Biology* 14.3 (Mar. 2007), pp. 185–193.

[75] Colleen S. Sinclair et al. “The 17q23 amplicon and breast cancer”. In: *Breast Cancer Research and Treatment* 78.3 (Apr. 2003), pp. 313–322.

[76] Shipra Das et al. “Oncogenic splicing factor SRSF1 is a critical transcriptional target of MYC”. In: *Cell Reports* 1.2 (Feb. 23, 2012), pp. 110–117.

[77] Olga Anczuków et al. “The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation”. In: *Nature Structural & Molecular Biology* 19.2 (Feb. 2012), pp. 220–228.

[78] Claudia Ghigna et al. “Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene”. In: *Molecular Cell* 20.6 (Dec. 22, 2005), pp. 881–890.

[79] Linlin Chen et al. “SRSF1 Prevents DNA Damage and Promotes Tumorigenesis through Regulation of DBF4B Pre-mRNA Splicing”. In: *Cell Reports* 21.12 (Dec. 19, 2017), pp. 3406–3413.

[80] Claude C. Warzecha et al. “An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition”. In: *The EMBO journal* 29.19 (Oct. 6, 2010), pp. 3286–3300.

[81] Toshifumi Yae et al. “Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell”. In: *Nature Communications* 3 (June 6, 2012), p. 883.

[82] Morton Freytag et al. “Epithelial splicing regulatory protein 1 and 2 (ESRP1 and ESRP2) upregulation predicts poor prognosis in prostate cancer”. In: *BMC cancer* 20.1 (Dec. 18, 2020), p. 1220.

[83] Yesim Gökmen-Polar et al. “Splicing factor ESRP1 controls ER-positive breast cancer by altering metabolic pathways”. In: *EMBO reports* 20.2 (Feb. 2019), e46078.

[84] Feng-Yang Zong et al. “The RNA-Binding Protein QKI Suppresses Cancer-Associated Aberrant Splicing”. In: *PLOS Genetics* 10.4 (Apr. 10, 2014), e1004289.

[85] Pratiti Bandopadhayay et al. “MYB-QKI rearrangements in angiocentric glioma drive tumorigenicity through a tripartite mechanism”. In: *Nature Genetics* 48.3 (Mar. 2016), pp. 273–282.

[86] Shangbiao Li et al. “FBXO7 Confers Mesenchymal Properties and Chemoresistance in Glioblastoma by Controlling Rbfox2-Mediated Alternative Splicing”. In: *Advanced Science (Weinheim, Baden-Wurttemberg, Germany)* 10.33 (Nov. 2023), e2303561.

[87] Irina M. Shapiro et al. “An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype”. In: *PLoS genetics* 7.8 (Aug. 2011), e1002218.

[88] Amina Jbara et al. “RBFOX2 modulates a metastatic signature of alternative splicing in pancreatic cancer”. In: *Nature* 617.7959 (May 2023), pp. 147–153.

[89] Baocai Lu et al. “LncRNA ZFAS1 promotes laryngeal cancer progression through RBFOX2-mediated MENA alternative splicing”. In: *Environmental Toxicology* 38.3 (2023), pp. 522–533.

[90] Mohini Jangi et al. “Rbfox2 controls autoregulation in RNA-binding protein networks”. In: *Genes & Development* 28.6 (Mar. 15, 2014), pp. 637–651.

[91] André Kahles et al. “Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients”. In: *Cancer Cell* 34.2 (Aug. 13, 2018), 211–224.e6.

[92] Heidi Dvinge et al. “RNA splicing factors as oncoproteins and tumour suppressors”. In: *Nature Reviews Cancer* 16.7 (July 2016), pp. 413–430.

[93] Hyunchul Jung et al. “Intron retention is a widespread mechanism of tumor-suppressor inactivation”. In: *Nature Genetics* 47.11 (Nov. 2015), pp. 1242–1248.

[94] Anna Corrionero, Belén Miñana, and Juan Valcárcel. “Reduced fidelity of branch point recognition and alternative splicing induced by the anti-tumor drug spliceostatin A”. In: *Genes & Development* 25.5 (Mar. 1, 2011), pp. 445–459.

[95] Kristine O’Brien et al. “The biflavonoid isoginkgetin is a general inhibitor of Pre-mRNA splicing”. In: *The Journal of Biological Chemistry* 283.48 (Nov. 28, 2008), pp. 33147–33154.

[96] Ting Han et al. “Anticancer sulfonamides target splicing by inducing RBM39 degradation via recruitment to DCAF15”. In: *Science (New York, N.Y.)* 356.6336 (Apr. 28, 2017), eaal3755.

[97] Jia Yi Fong et al. “Therapeutic Targeting of RNA Splicing Catalysis through Inhibition of Protein Arginine Methylation”. In: *Cancer Cell* 36.2 (Aug. 12, 2019), 194–209.e9.

[98] Maki Sakuma, Kei Iida, and Masatoshi Hagiwara. “Deciphering targeting rules of splicing modulator compounds: case of TG003”. In: *BMC molecular biology* 16 (Sept. 24, 2015), p. 16.

[99] Cormac Sheridan. “First small-molecule drug targeting RNA gains momentum”. In: *Nature Biotechnology* 39.1 (Jan. 2021), pp. 6–8.

[100] Mallory A. Havens and Michelle L. Hastings. “Splice-switching antisense oligonucleotides as therapeutic drugs”. In: *Nucleic Acids Research* 44.14 (Aug. 19, 2016), pp. 6549–6563.

[101] Takenori Shimo, Rika Maruyama, and Toshifumi Yokota. “Designing Effective Antisense Oligonucleotides for Exon Skipping”. In: *Methods in Molecular Biology (Clifton, N.J.)* 1687 (2018), pp. 143–155.

[102] Alejandro Garanto et al. “In vitro and in vivo rescue of aberrant splicing in CEP290-associated LCA by antisense oligonucleotide delivery”. In: *Human Molecular Genetics* 25.12 (June 15, 2016), pp. 2552–2563.

[103] Jennifer J. Lentz et al. “Rescue of hearing and vestibular function by antisense oligonucleotides in a mouse model of human deafness”. In: *Nature Medicine* 19.3 (Mar. 2013), pp. 345–350.

[104] Yanan Sun et al. “Downregulation of SRSF3 by antisense oligonucleotides sensitizes oral squamous cell carcinoma and breast cancer cells to paclitaxel treatment”. In: *Cancer Chemotherapy and Pharmacology* 84.5 (Nov. 2019), pp. 1133–1143.

[105] Samy Lamouille, Jian Xu, and Rik Derynck. “Molecular mechanisms of epithelial–mesenchymal transition”. In: *Nature Reviews Molecular Cell Biology* 15.3 (Mar. 2014), pp. 178–196.

[106] Héctor Peinado, David Olmeda, and Amparo Cano. “Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype?” In: *Nature Reviews. Cancer* 7.6 (June 2007), pp. 415–428.

[107] M. Angela Nieto and Amparo Cano. “The epithelial-mesenchymal transition under control: global programs to regulate epithelial plasticity”. In: *Seminars in Cancer Biology* 22.5 (Oct. 2012), pp. 361–368.

[108] Sendurai A. Mani et al. “The epithelial-mesenchymal transition generates cells with properties of stem cells”. In: *Cell* 133.4 (May 16, 2008), pp. 704–715.

[109] Christina Scheel and Robert A. Weinberg. “Cancer stem cells and epithelial-mesenchymal transition: concepts and molecular links”. In: *Seminars in Cancer Biology* 22.5 (Oct. 2012), pp. 396–403.

[110] Jing Yang et al. “Guidelines and definitions for research on epithelial–mesenchymal transition”. In: *Nature Reviews. Molecular Cell Biology* 21.6 (2020), pp. 341–352.

[111] Masao Saitoh. “Involvement of partial EMT in cancer progression”. In: *The Journal of Biochemistry* 164.4 (Oct. 1, 2018), pp. 257–264.

[112] David P. Cook and Barbara C. Vanderhyden. “Context specificity of the EMT transcriptional response”. In: *Nature Communications* 11 (May 1, 2020), p. 2142.

[113] C. M. Perou et al. “Molecular portraits of human breast tumours”. In: *Nature* 406.6797 (Aug. 17, 2000), pp. 747–752.

[114] Brett Wallden et al. “Development and verification of the PAM50-based Prosigna breast cancer gene signature assay”. In: *BMC Medical Genomics* 8.1 (Aug. 22, 2015), p. 54.

[115] T. Sørlie et al. “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications”. In: *Proceedings of the National Academy of Sciences of the United States of America* 98.19 (Sept. 11, 2001), pp. 10869–10874.

[116] Aleix Prat and Charles M. Perou. “Deconstructing the molecular portraits of breast cancer”. In: *Molecular Oncology* 5.1 (Feb. 2011), pp. 5–23.

[117] Aleix Prat et al. “Molecular characterization of basal-like and non-basal-like triple-negative breast cancer”. In: *The Oncologist* 18.2 (2013), pp. 123–133.

[118] Jason I. Herschkowitz et al. “Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors”. In: *Genome Biology* 8.5 (2007), R76.

[119] Aleix Prat et al. “Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer”. In: *Breast cancer research: BCR* 12.5 (2010), R68.

[120] Kay Dias et al. “Claudin-Low Breast Cancer; Clinical & Pathological Characteristics”. In: *PloS One* 12.1 (2017), e0168669.

[121] Christian Fougner et al. “Claudin-low-like mouse mammary tumors show distinct transcriptomic patterns uncoupled from genomic drivers”. In: *Breast Cancer Research* 21.1 (July 31, 2019), p. 85.

[122] Renaud Sabatier et al. “Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization”. In: *Molecular Cancer* 13 (Oct. 2, 2014), p. 228.

[123] Christian Fougner et al. “Re-definition of claudin-low as a breast cancer phenotype”. In: *Nature Communications* 11.1 (Apr. 14, 2020), p. 1787.

[124] Sunny Z. Wu et al. “A single-cell and spatially resolved atlas of human breast cancers”. In: *Nature Genetics* 53.9 (Sept. 2021), pp. 1334–1347.

[125] Jordan Kardos et al. “Claudin-low bladder tumors are immune infiltrated and actively immune suppressed”. In: *JCI Insight* 1.3 (Mar. 2016), e85902.

[126] Tomohiro F. Nishijima et al. “Molecular and Clinical Characterization of a Claudin-Low Subtype of Gastric Cancer”. In: *JCO Precision Oncology* 1 (Nov. 2017), pp. 1–10.

[127] Leonardo Mariño-Ramírez et al. “Statistical analysis of over-represented words in human promoter sequences”. In: *Nucleic Acids Research* 32.3 (2004), pp. 949–958.

[128] Peter C. FitzGerald et al. “Clustering of DNA sequences in human promoters”. In: *Genome Research* 14.8 (Aug. 2004), pp. 1562–1574.

[129] D. J. Mathis et al. “The murine E alpha immune response gene”. In: *Cell* 32.3 (Mar. 1983), pp. 745–754.

[130] N Sittisombut. “Two distinct nuclear factors bind the conserved regulatory sequences of a rabbit major histocompatibility complex class II gene.” In: *Molecular and Cellular Biology* 8.5 (May 1988), pp. 2034–2041.

[131] R Schöpfer et al. “Evolutionary diversification of class II P loci in the Mhc of the mole-rat Spalax ehrenbergi.” In: *Molecular Biology and Evolution* 4.3 (May 1, 1987), pp. 287–299.

[132] C. Benoist and D. Mathis. “Regulation of major histocompatibility complex class-II genes: X, Y and other letters of the alphabet”. In: *Annual Review of Immunology* 8 (1990), pp. 681–715.

[133] A. Dorn et al. “Conserved major histocompatibility complex class II boxes–X and Y–are transcriptional control elements and specifically bind nuclear proteins”. In: *Proceedings of the National Academy of Sciences of the United States of America* 84.17 (Sept. 1987), pp. 6249–6253.

[134] R. M. Myers, K. Tilly, and T. Maniatis. “Fine structure genetic analysis of a beta-globin promoter”. In: *Science (New York, N.Y.)* 232.4750 (May 2, 1986), pp. 613–618.

[135] B. T. Greuel, L. Sealy, and J. E. Majors. “Transcriptional activity of the Rous sarcoma virus long terminal repeat correlates with binding of a factor to an upstream CCAAT box in vitro”. In: *Virology* 177.1 (July 1990), pp. 33–43.

[136] N. J. Zeleznik-Le, J. C. Azizkhan, and J. P. Ting. “Affinity-purified CCAAT-box-binding protein (YEBP) functionally regulates expression of a human class II major histocompatibility complex gene and the herpes simplex virus thymidine kinase gene”. In: *Proceedings of the National Academy of Sciences of the United States of America* 88.5 (Mar. 1, 1991), pp. 1873–1877.

[137] Roberto Mantovani. “The molecular biology of the CCAAT-binding factor NF-Y”. In: *Gene* 239.1 (Oct. 18, 1999), pp. 15–27.

[138] R Mantovani. “A survey of 178 NF-Y binding CCAAT boxes.” In: *Nucleic Acids Research* 26.5 (Mar. 1, 1998), pp. 1135–1143.

[139] P. Bucher. “Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences”. In: *Journal of Molecular Biology* 212.4 (Apr. 20, 1990), pp. 563–578.

[140] Diletta Dolfini et al. “A perspective of promoter architecture from the CCAAT box”. In: *Cell Cycle* 8.24 (Dec. 15, 2009), pp. 4127–4137.

[141] F. Antequera. “Structure, function and evolution of CpG island promoters”. In: *Cellular and molecular life sciences: CMLS* 60.8 (Aug. 2003), pp. 1647–1658.

[142] Christopher Benner et al. “Decoding a Signature-Based Model of Transcription Cofactor Recruitment Dictated by Cardinal Cis-Regulatory Elements in Proximal Promoter Regions”. In: *PLoS Genetics* 9.11 (Nov. 7, 2013), e1003906.

[143] D. K. Didier et al. “Characterization of the cDNA encoding a protein binding to the major histocompatibility complex class II Y box”. In: *Proceedings of the National Academy of Sciences of the United States of America* 85.19 (Oct. 1988), pp. 7322–7326.

[144] Marija Mihailovich et al. “Eukaryotic cold shock domain proteins: highly versatile regulators of gene expression”. In: *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 32.2 (Feb. 2010), pp. 109–118.

[145] S L Hasegawa et al. “DNA binding properties of YB-1 and dbpA: binding to double-stranded, single-stranded, and abasic site containing DNAs.” In: *Nucleic Acids Research* 19.18 (Sept. 25, 1991), pp. 4915–4920.

[146] Valentina Evdokimova, Lev P. Ovchinnikov, and Poul H. B. Sorensen. “Y-box binding protein 1: providing a new angle on translational regulation”. In: *Cell Cycle (Georgetown, Tex.)* 5.11 (June 2006), pp. 1143–1147.

[147] Valentina Evdokimova et al. “Reduced proliferation and enhanced migration: two sides of the same coin? Molecular mechanisms of metastatic progression by YB-1”. In: *Cell Cycle (Georgetown, Tex.)* 8.18 (Sept. 15, 2009), pp. 2901–2906.

[148] Kathleen W. Scotto. “Transcriptional regulation of ABC drug transporters”. In: *Oncogene* 22.47 (Oct. 2003), pp. 7496–7511.

[149] D. Dolfini and R. Mantovani. “YB-1 (YBX1) does not bind to Y/CCAAT boxes in vivo”. In: *Oncogene* 32.35 (Aug. 2013), pp. 4189–4190.

[150] Teresa Soop et al. “A p50-like Y-box protein with a putative translational role becomes associated with pre-mRNA concomitant with transcription”. In: *Journal of Cell Science* 116 (Pt 8 Apr. 15, 2003), pp. 1493–1503.

[151] R A Hooft van Huijsduijnen et al. “Properties of a CCAAT box-binding protein.” In: *Nucleic Acids Research* 15.18 (Sept. 25, 1987), pp. 7265–7282.

[152] François Tronche et al. “NFY or a related CCAAT binding factor can be replaced by other transcriptional activators for co-operation with HNF1 in driving the rat albumin promoter in vivo”. In: *Journal of Molecular Biology* 222.1 (Nov. 5, 1991), pp. 31–43.

[153] R. Mantovani et al. “Dominant negative analogs of NF-YA”. In: *The Journal of Biological Chemistry* 269.32 (Aug. 12, 1994), pp. 20340–20346.

[154] Paolo Benatti et al. “A balance between NF-Y and p53 governs the pro- and anti-apoptotic transcriptional response”. In: *Nucleic Acids Research* 36.5 (Mar. 2008), pp. 1415–1428.

[155] Mattia Frontini et al. “Cell cycle regulation of NF-YC nuclear localization”. In: *Cell Cycle (Georgetown, Tex.)* 3.2 (Feb. 2004), pp. 217–222.

[156] L. A. Chodosh et al. “A yeast and a human CCAAT-binding protein have heterologous subunits that are functionally interchangeable”. In: *Cell* 53.1 (Apr. 8, 1988), pp. 25–35.

[157] S. L. Forsburg and L. Guarente. “Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer”. In: *Genes & Development* 3.8 (Aug. 1989), pp. 1166–1178.

[158] Alberto Silvio di, Carol Imbriano, and Roberto Mantovani. “Dissection of the NF-Y transcriptional activation potential”. In: *Nucleic Acids Research* 27.13 (July 1, 1999), pp. 2578–2584.

[159] Yasuhide Yoshioka et al. “Complex interference in the eye developmental pathway by Drosophila NF-YA”. In: *Genesis (New York, N.Y.: 2000)* 45.1 (Jan. 2007), pp. 21–31.

[160] Yau-Hung Chen, Yung-Tsang Lin, and Gang-Hui Lee. "Novel and unexpected functions of zebrafish CCAAT box binding transcription factor (NF-Y) B subunit during cartilages development". In: *Bone* 44.5 (May 1, 2009), pp. 777–784.

[161] Anuradha Bhattacharya et al. "The B subunit of the CCAAT box binding transcription factor complex (CBF/NF-Y) is essential for early mouse development and cell proliferation". In: *Cancer Research* 63.23 (Dec. 1, 2003), pp. 8167–8172.

[162] Paolo Benatti et al. "Specific inhibition of NF-Y subunits triggers different cell proliferation defects". In: *Nucleic Acids Research* 39.13 (July 2011), pp. 5356–5368.

[163] Tom Laloum et al. "CCAAT-box binding transcription factors in plants: Y so many?" In: *Trends in Plant Science* 18.3 (Mar. 2013), pp. 157–166.

[164] Amy Lawton-Rauh. "Evolutionary dynamics of duplicated genes in plants". In: *Molecular Phylogenetics and Evolution* 29.3 (Dec. 2003), pp. 396–409.

[165] T. Lotan et al. "Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells". In: *Cell* 93.7 (June 26, 1998), pp. 1195–1205.

[166] Roderick W. Kumimoto et al. "The Nuclear Factor Y subunits NF-YB2 and NF-YB3 play additive roles in the promotion of flowering by inductive long-day photoperiods in Arabidopsis". In: *Planta* 228.5 (Oct. 2008), pp. 709–723.

[167] Jian-Xiang Liu and Stephen H. Howell. "bZIP28 and NF-Y transcription factors are activated by ER stress and assemble into a transcriptional complex to regulate stress response genes in Arabidopsis". In: *The Plant Cell* 22.3 (Mar. 2010), pp. 782–796.

[168] Wen-Xue Li et al. "The Arabidopsis NFYA5 Transcription Factor Is Regulated Transcriptionally and Posttranscriptionally to Promote Drought Resistance". In: *The Plant Cell* 20.8 (Aug. 2008), pp. 2238–2251.

[169] María Eugenia Zanetti et al. "A C subunit of the plant nuclear factor NF-Y required for rhizobial infection and nodule development affects partner selection in the common bean-Rhizobium etli symbiosis". In: *The Plant Cell* 22.12 (Dec. 2010), pp. 4142–4157.

[170] Marco Nardini et al. "Sequence-Specific Transcription Factor NF-Y Displays Histone-like DNA Binding and H2B-like Ubiquitination". In: *Cell* 152.1 (Jan. 17, 2013), pp. 132–143.

[171] Youngchang Kim et al. "Crystal structure of a yeast TBP/TATA-box complex". In: *Nature* 365.6446 (Oct. 1993), pp. 512–520.

[172] E. C. Murphy et al. "Structural basis for SRY-dependent 46-X,Y sex reversal: modulation of DNA bending by a naturally occurring point mutation". In: *Journal of Molecular Biology* 312.3 (Sept. 21, 2001), pp. 481–499.

[173] Peter L. Privalov et al. "What drives proteins into the major or minor grooves of DNA?" In: *Journal of Molecular Biology* 365.1 (Jan. 5, 2007), pp. 1–9.

[174] F. Coustry et al. "CBF/NF-Y functions both in nucleosomal disruption and transcription activation of the chromatin-assembled topoisomerase IIalpha promoter. Transcription activation by CBF/NF-Y in chromatin is dependent on the promoter structure". In: *The Journal of Biological Chemistry* 276.44 (Nov. 2, 2001), pp. 40621–40630.

[175] Shengkan Jin and Kathleen W. Scotto. "Transcriptional Regulation of the MDR1 Gene by Histone Acetyltransferase and Deacetylase Is Mediated by NF-Y". In: *Molecular and Cellular Biology* 18.7 (July 1998), pp. 4377–4384.

[176] Tianyun Zhao et al. "A role for polyamine regulators in ESC self-renewal". In: *Cell Cycle (Georgetown, Tex.)* 11.24 (Dec. 15, 2012), pp. 4517–4523.

[177] Andrew J. Oldfield et al. "Histone-Fold Domain Protein NF-Y Promotes Chromatin Accessibility for Cell Type-Specific Master Transcription Factors". In: *Molecular Cell* 55.5 (Sept. 4, 2014), pp. 708–722.

[178] Andrew J. Oldfield et al. "NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region". In: *Nature Communications* 10.1 (July 11, 2019), p. 3072.

[179] Michele Ceribelli et al. "The Histone-Like NF-Y Is a Bifunctional Transcription Factor". In: *Molecular and Cellular Biology* 28.6 (Mar. 2008), pp. 2047–2058.

[180] Diletta Dolfini, Raffaella Gatta, and Roberto Mantovani. "NF-Y and the transcriptional activation of CCAAT promoters". In: *Critical Reviews in Biochemistry and Molecular Biology* 47.1 (Feb. 1, 2012), pp. 29–49.

[181] K. Yamada et al. "Sp family members and nuclear factor-Y cooperatively stimulate transcription from the rat pyruvate kinase M gene distal promoter region via their direct interactions". In: *The Journal of Biological Chemistry* 275.24 (June 16, 2000), pp. 18129–18137.

[182] Brian D. Reed et al. "Genome-wide occupancy of SREBP1 and its partners NFY and SP1 reveals novel functional roles and combinatorial regulation of distinct classes of genes". In: *PLoS genetics* 4.7 (July 25, 2008), e1000133.

[183] Joanna Kaczynski, Tiffany Cook, and Raul Urrutia. "Sp1- and Krüppel-like transcription factors". In: *Genome Biology* 4.2 (2003), p. 206.

[184] Claire Attwooll, Eros Lazzerini Denchi, and Kristian Helin. "The E2F family: specific functions and overlapping interests". In: *The EMBO Journal* 23.24 (Dec. 8, 2004), pp. 4709–4716.

[185] Aymone Gurtner et al. "Transcription factor NF-Y induces apoptosis in cells expressing wild-type p53 through E2F1 upregulation and p53 activation". In: *Cancer Research* 70.23 (Dec. 1, 2010), pp. 9711–9720.

[186] Feifei Chen et al. "Repression of Smad2 and Smad3 transactivating activity by association with a novel splice variant of CCAAT-binding factor C subunit". In: *The Biochemical Journal* 364 (Pt 2 June 1, 2002), pp. 571–577.

[187] Timothy Ravasi et al. "An atlas of combinatorial transcriptional regulation in mouse and man". In: *Cell* 140.5 (Mar. 5, 2010), pp. 744–752.

[188] Yoshihiro Ito et al. "NF-Y and USF1 transcription factor binding to CCAAT-box and E-box elements activates the CP27 promoter". In: *Gene* 473.2 (Mar. 1, 2011), pp. 92–99.

[189] ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (Sept. 6, 2012), pp. 57–74.

[190] Mirko Ronzio et al. "On the NF-Y regulome as in ENCODE (2019)". In: *PLoS Computational Biology* 16.12 (Dec. 28, 2020), e1008488.

[191] Jianqi Yang et al. “A Novel Mechanism Involving Coordinated Regulation of Nuclear Levels and Acetylation of NF-YA and Bcl6 Activates RGS4 Transcription*”. In: *Journal of Biological Chemistry* 285.39 (Sept. 24, 2010), pp. 29760–29769.

[192] Isabella Manni et al. “Posttranslational Regulation of NF-YA Modulates NF-Y Transcriptional Activity”. In: *Molecular Biology of the Cell* 19.12 (Dec. 2008), pp. 5203–5213.

[193] Marcel Thön et al. “The CCAAT-binding complex coordinates the oxidative stress response in eukaryotes”. In: *Nucleic Acids Research* 38.4 (Mar. 2010), pp. 1098–1113.

[194] X. Y. Li et al. “Intron-exon organization of the NF-Y genes. Tissue-specific splicing modifies an activation domain.” In: *Journal of Biological Chemistry* 267.13 (May 5, 1992), pp. 8984–8990.

[195] Lucia Cappabianca et al. “Discovery, characterization and potential roles of a novel NF-YAx splice variant in human neuroblastoma”. In: *Journal of Experimental & Clinical Cancer Research* 38.1 (Dec. 5, 2019), p. 482.

[196] Michele Ceribelli et al. “NF-YC Complexity Is Generated by Dual Promoters and Alternative Splicing”. In: *The Journal of Biological Chemistry* 284.49 (Dec. 4, 2009), pp. 34189–34200.

[197] Jiang Zhu et al. “NF-Y cooperates with USF1/2 to induce the hematopoietic expression of HOXB4”. In: *Blood* 102.7 (Oct. 1, 2003), pp. 2420–2427.

[198] Jiang Zhu et al. “NF-Ya activates multiple hematopoietic stem cell (HSC) regulatory genes and promotes HSC self-renewal”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.33 (Aug. 16, 2005), pp. 11728–11733.

[199] Bernard Binétruy et al. “Concise review: regulation of embryonic stem cell lineage commitment by mitogen-activated protein kinases”. In: *Stem Cells (Dayton, Ohio)* 25.5 (May 2007), pp. 1090–1095.

[200] Vijay K. Tiwari et al. “A chromatin-modifying function of JNK during stem cell differentiation”. In: *Nature Genetics* 44.1 (Jan. 2012), pp. 94–100.

[201] Diletta Dolfini et al. “The Short Isoform of NF-YA Belongs to the Embryonic Stem Cell Transcription Factor Circuitry”. In: *Stem Cells* 30.11 (Nov. 1, 2012), pp. 2450–2459.

[202] Debora Libetti et al. “The Switch from NF-YAl to NF-YAs Isoform Impairs Myotubes Formation”. In: *Cells* 9.3 (Mar. 2020), p. 789.

[203] Chaim Linhart et al. “Deciphering Transcriptional Regulatory Elements That Encode Specific Cell-Cycle Phasing by Comparative Genomics Analysis”. In: *Cell Cycle* 4.12 (Dec. 1, 2005), pp. 1788–1797.

[204] Wencheng Zhu, Paloma H Giangrande, and Joseph R Nevins. “E2Fs link the control of G1/S and G2/M transcription”. In: *The EMBO Journal* 23.23 (Nov. 24, 2004), pp. 4615–4626.

[205] Mark Wasner et al. “Three CCAAT-boxes and a single cell cycle genes homology region (CHR) are the major regulating sites for transcription from the human cyclin B2 promoter”. In: *Gene* 312 (July 17, 2003), pp. 225–237.

[206] Takeshi Uchiumi, Dan L. Longo, and Douglas K. Ferris. "Cell Cycle Regulation of the Human Polo-like Kinase (PLK) Promoter*". In: *Journal of Biological Chemistry* 272.14 (Apr. 4, 1997), pp. 9166–9174.

[207] Masashi Kimura et al. "Cell cycle-dependent regulation of the human aurora B promoter". In: *Biochemical and Biophysical Research Communications* 316.3 (Apr. 9, 2004), pp. 930–936.

[208] Hee-Don Chae, Jungbin Kim Shin, and Deug Y. "NF-Y binds to both G1- and G2-specific cyclin promoters; a possible role in linking CDK2/Cyclin A to CDK1/Cyclin B". In: *BMB Reports* 44.8 (Aug. 31, 2011), pp. 553–557.

[209] Paolo Benatti et al. "Direct non transcriptional role of NF-Y in DNA replication". In: *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1863.4 (Apr. 1, 2016), pp. 673–685.

[210] R. Gatta, D. Dolfini, and R. Mantovani. "NF-Y joins E2Fs, p53 and other stress transcription factors at the apoptosis table". In: *Cell Death & Disease* 2.5 (May 1, 2011), e162–e162.

[211] Paolo Benatti et al. "NF-Y activates genes of metabolic pathways altered in cancer cells". In: *Oncotarget* 7.2 (Dec. 3, 2015), pp. 1633–1650.

[212] M. Amemiya-Kudo et al. "Promoter analysis of the mouse sterol regulatory element-binding protein-1c gene". In: *The Journal of Biological Chemistry* 275.40 (Oct. 6, 2000), pp. 31078–31085.

[213] S. M. Jackson et al. "NF-Y has a novel role in sterol-dependent transcription of two cholesterogenic genes". In: *The Journal of Biological Chemistry* 270.37 (Sept. 15, 1995), pp. 21445–21448.

[214] Vicky Howe et al. "New insights into cellular cholesterol acquisition: promoter analysis of human HMGCR and SQLE, two key control enzymes in cholesterol synthesis". In: *Biochimica Et Biophysica Acta. Molecular and Cell Biology of Lipids* 1862.7 (July 2017), pp. 647–657.

[215] J. Inoue, R. Sato, and M. Maeda. "Multiple DNA elements for sterol regulatory element-binding protein and NF-Y are responsible for sterol-regulated transcription of the genes for human 3-hydroxy-3-methylglutaryl coenzyme A synthase and squalene synthase". In: *Journal of Biochemistry* 123.6 (June 1998), pp. 1191–1198.

[216] Anika V. Prabhu, Laura J. Sharpe, and Andrew J. Brown. "The sterol-based transcriptional control of human 7-dehydrocholesterol reductase (DHCR7): Evidence of a cooperative regulatory program in cholesterol synthesis". In: *Biochimica Et Biophysica Acta* 1842.10 (Oct. 2014), pp. 1431–1439.

[217] Xuanming Shi, Cornelia C. Metges, and Hans-Martin Seyfert. "Interaction of C/EBP-beta and NF-Y factors constrains activity levels of the nutritionally controlled promoter IA expressing the acetyl-CoA carboxylase-alpha gene in cattle". In: *BMC Molecular Biology* 13.1 (June 27, 2012), p. 21.

[218] Shunbin Xiong, Subrahmanyam S. Chirala, and Salih J. Wakil. "Sterol regulation of human fatty acid synthase promoter I requires nuclear factor-Y- and Sp-1-binding sites". In: *Proceedings of the National Academy of Sciences of the United States of America* 97.8 (Apr. 11, 2000), pp. 3948–3953.

[219] Daniel Mauvoisin et al. “Role of the PI3-kinase/mTor pathway in the regulation of the stearoyl CoA desaturase (SCD1) gene expression by insulin in liver”. In: *Journal of Cell Communication and Signaling* 1.2 (Sept. 2007), pp. 113–125.

[220] Min Gyu Lee and Peter L. Pedersen. “Glucose metabolism in cancer: importance of transcription factor-DNA interactions within a short segment of the proximal region og the type II hexokinase promoter”. In: *The Journal of Biological Chemistry* 278.42 (Oct. 17, 2003), pp. 41047–41058.

[221] K Tsutsumi, K Ito, and K Ishikawa. “Developmental appearance of transcription factors that regulate liver-specific expression of the aldolase B gene.” In: *Molecular and Cellular Biology* 9.11 (Nov. 1989), pp. 4923–4931.

[222] Jason W. Locasale. “Serine, glycine and one-carbon units: cancer metabolism in full circle”. In: *Nature Reviews Cancer* 13.8 (Aug. 2013), pp. 572–583.

[223] Do Youn Jun et al. “Positive regulation of promoter activity of human 3-phosphoglycerate dehydrogenase (PHGDH) gene is mediated by transcription factors Sp1 and NF-Y”. In: *Gene* 414.1 (May 15, 2008), pp. 106–114.

[224] Cristina Pérez-Gómez et al. “Genomic organization and transcriptional analysis of the human l-glutaminase gene”. In: *The Biochemical Journal* 370 (Pt 3 Mar. 15, 2003), pp. 771–784.

[225] Hani Goodarzi, Olivier Elemento, and Saeed Tavazoie. “Revealing global regulatory perturbations across human cancers”. In: *Molecular cell* 36.5 (Dec. 11, 2009), pp. 900–911.

[226] Erik Andrews et al. “Contextual Refinement of Regulatory Targets Reveals Effects on Breast Cancer Prognosis of the Regulome”. In: *PLOS Computational Biology* 13.1 (Jan. 19, 2017), e1005340.

[227] Zbyslaw Sondka et al. “COSMIC: a curated database of somatic variants and clinical data for cancer”. In: *Nucleic Acids Research* (Nov. 1, 2023), gkad986.

[228] Suhana Mamat et al. “Transcriptional Regulation of Aldehyde Dehydrogenase 1A1 Gene by Alternative Spliced Forms of Nuclear Factor Y in Tumorigenic Population of Endometrial Adenocarcinoma”. In: *Genes & Cancer* 2.10 (Oct. 1, 2011), pp. 979–984.

[229] Lucia Cicchillitti et al. “Prognostic role of NF-YA splicing isoforms and Lamin A status in low grade endometrial cancer”. In: *Oncotarget* 8.5 (Jan. 31, 2017), pp. 7935–7945.

[230] Bin Cao et al. “Gene regulatory network construction identified NFYA as a diffuse subtype-specific prognostic factor in gastric cancer”. In: *International Journal of Oncology* 53.5 (Nov. 2018), pp. 1857–1868.

[231] Chuanwei Yang et al. “Cadherins Associate with Distinct Stem Cell-Related Transcription Factors to Coordinate the Maintenance of Stemness in Triple-Negative Breast Cancer”. In: *Stem Cells International* 2017 (Mar. 14, 2017), e5091541.

[232] Hua Cui et al. “NF-YC in glioma cell proliferation and tumor growth and its role as an independent predictor of patient survival”. In: *Neuroscience Letters* 631 (Sept. 19, 2016), pp. 40–49.

[233] Anastasia E. Kottorou et al. “Altered expression of NFY-C and RORA in colorectal adenocarcinomas”. In: *Acta Histochemica* 114.6 (Oct. 1, 2012), pp. 553–561.

[234] Diletta Dolfini, Valentina Andrioletti, and Roberto Mantovani. "Overexpression and alternative splicing of NF-YA in breast cancer". In: *Scientific Reports* 9.1 (Sept. 10, 2019), p. 12955.

[235] Eugenia Bezzecchi et al. "NF-YA Overexpression in Lung Cancer: LUSC". In: *Genes* 10.11 (Nov. 17, 2019), p. 937.

[236] Eugenia Bezzecchi et al. "NF-YA Overexpression in Lung Cancer: LUAD". In: *Genes* 11.2 (Feb. 2020), p. 198.

[237] Eugenia Bezzecchi et al. "NF-Y Subunits Overexpression in HNSCC". In: *Cancers* 13.12 (Jan. 2021), p. 3019.

[238] Eugenia Bezzecchi et al. "NF-Y Overexpression in Liver Hepatocellular Carcinoma (HCC)". In: *International Journal of Molecular Sciences* 21.23 (Dec. 1, 2020), E9157.

[239] Maria Alsina et al. "Current developments in gastric cancer: from molecular profiling to treatment strategy". In: *Nature Reviews Gastroenterology & Hepatology* 20.3 (Mar. 2023), pp. 155–170.

[240] Pekka Laurén. "The Two Histological Main Types of Gastric Carcinoma: Diffuse and so-Called Intestinal-Type Carcinoma". In: *Acta Pathologica Microbiologica Scandinavica* 64.1 (1965), pp. 31–49.

[241] Adam J. Bass et al. "Comprehensive molecular characterization of gastric adenocarcinoma". In: *Nature* 513.7517 (Sept. 2014), pp. 202–209.

[242] Razvan Cristescu et al. "Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes". In: *Nature Medicine* 21.5 (May 2015), pp. 449–456.

[243] Feng Gao et al. "DeepCC: a novel deep learning-based framework for cancer molecular subtype classification". In: *Oncogenesis* 8.9 (Aug. 16, 2019), pp. 44–44.

[244] Peter Langfelder and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9.1 (Dec. 29, 2008), p. 559.

[245] Jihyun Kim et al. "Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage diversity and intratumoral heterogeneity". In: *npj Precision Oncology* 6.1 (Jan. 27, 2022), pp. 1–11.

[246] Meichen Dong et al. "SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references". In: *Briefings in Bioinformatics* 22.1 (Jan. 18, 2021), pp. 416–427.

[247] Kate B. Cook et al. "RBPDB: a database of RNA-binding specificities". In: *Nucleic Acids Research* 39 (Database issue Jan. 2011), pp. D301–308.

[248] Tobias Tekath and Martin Dugas. "Differential transcript usage analysis of bulk and single-cell RNA-seq data with DTUrtle". In: *Bioinformatics* 37.21 (Nov. 1, 2021), pp. 3781–3787.

[249] Jeroen Gilis et al. *satuRn: Scalable Analysis of differential Transcript Usage for bulk and single-cell RNA-sequencing applications*. Jan. 16, 2021.

[250] Carolin Kosiol et al. "Patterns of Positive Selection in Six Mammalian Genomes". In: *PLOS Genetics* 4.8 (Aug. 1, 2008), e1000144.

[251] Joanna Ciomborowska-Basheer et al. “Not So Dead Genes-Retrocopies as Regulators of Their Disease-Related Progenitors and Hosts”. In: *Cells* 10.4 (Apr. 15, 2021), p. 912.

[252] Rana Mhaidly and Fatima Mechta-Grigoriou. “Role of cancer-associated fibroblast subpopulations in immune infiltration, as a new means of treatment in cancer”. In: *Immunological Reviews* 302.1 (July 2021), pp. 259–272.

[253] Catherine McCusker and David M. Gardiner. “The axolotl model for regeneration and aging research: a mini-review”. In: *Gerontology* 57.6 (2011), pp. 565–571.

[254] Andrew Butler et al. “Integrating single-cell transcriptomic data across different conditions, technologies, and species”. In: *Nature Biotechnology* 36.5 (June 2018), pp. 411–420.

[255] Ji-Hye Choi, Hye In Kim, and Hyun Goo Woo. “scTyper: a comprehensive pipeline for the cell typing analysis of single-cell RNA-seq data”. In: *BMC Bioinformatics* 21.1 (Aug. 4, 2020), p. 342.

[256] Bo Li and Colin N. Dewey. “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome”. In: *BMC bioinformatics* 12 (Aug. 4, 2011), p. 323.

[257] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12 (2014), pp. 550–550.