

Multiple Retrotransposon-mediated NF-YA Gene Duplication Events Recurred in Diverse Groups of Mammals at Different Ancestry Levels

Alberto Gallo ^{1,†}, Andrea Bernardini ^{1,†}, Sofia Poletti ¹, Diletta Dolfini ¹, Nerina Gnesutta ¹, Roberto Mantovani ^{1,*}

¹Dipartimento di Bioscienze, Università degli Studi di Milano, Via Celoria 26, 20133 Milano, Italy

[†]The authors equally contributed.

*Corresponding author: E-mail: mantor@unimi.it.

Accepted: April 10, 2025

Abstract

NF-Y is a transcription factor trimer 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. The subunits are extremely conserved in all eukaryotes and in mammals they are typically encoded by single copy genes. We describe here the presence of a second NF-YA termed NF-YAr for retrogene in diverse groups of mammals (*Cetacea*, *Ruminantia*, *Ursidae*, *Sciuridae*, hippopotamus, and greater horseshoe bat). NF-YAr retrogenes are located on different chromosomes with respect to the parental gene; they are compact and intronless, or with few annotated introns. Phylogenetic and synteny analyses indicate multiple independent retrotransposition events in the different orders. Analysis of RNA-seq data of *Bos taurus* suggests expression confined to spermatozoa. Conservation of translation initiation signals around predicted start codons, and of 5'UTR sequences, are consistent with protein expression, suggesting that NF-YAr is a translated, retroposed NF-YA. 3D-informed structural considerations of the predicted protein sequences point at deleterious changes for CCAAT-binding and, potentially, for trimer formation. These findings indicate that multiple independent NF-YA retrotransposition events were fixed in selected orders of mammals, generating a second NF-YA with a strict tissue distribution.

Key words: transcription factors, intrinsically disordered protein, retrogenes, mammals.

Significance

Normally, the flux of information is from DNA to RNA; mRNAs are sometimes transformed back into DNA and inserted in chromosomes into locations different from the original one (retrogenes). Retrogenes are often unexpressed, although in *Mammalia* functional ones have been associated to sperm cells. In this study we outline the conservation of a retrocopy of NF-YA, the regulatory subunit of the essential CCAAT-binding transcription factor complex, termed NF-YAr. This retrogene originated independently in selected mammalian orders, is possibly expressed in bull sperm cells, with similarities to its parental gene that may offer some insight on its function. The discovery of NF-YAr adds depth to NF-Y subunits phylogeny and may suggest novel tissue-specific regulatory mechanisms for transcription factors activity.

© The Author(s) 2025. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

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 (Jolma et al. 2013; Lambert et al. 2018; Wingender et al. 2018). 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 (Dolfini et al. 2009). The two other subunits NF-YB/NF-YC share the Histone Fold Domain (HFD) with core histones, notably H2B/H2A (Gnesutta et al. 2013). NF-YA provides the complex with sequence-specificity through a 56 amino acids domain present in all eukaryotes, named HAP2 after the yeast homolog (Nardone et al. 2017; Hortschansky et al. 2017). 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 nonsequence-specific DNA contacts by HFD subunits (Huber et al. 2012; Nardini et al. 2013; Chaves-Sanjuan et al. 2021). 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 (Nardone et al. 2017). In humans, the NF-YA gene is composed of 10 exons, located on the short arm of chromosome 6; the protein coding sequence (CDS) starts within exon-2 and ends within exon-10 (supplementary fig. S1, Supplementary Material online). NF-YA is involved in two major alternative splicing (AS) in mammals, comprising (NF-YA1) or lacking (NF-YAs) exon-3 sequences (Li et al. 1992). A third isoform NF-YAx was reported in human neuroblastomas, missing both exon-3 and exon-5 (Cappabianca et al. 2019). Another AS event is present uniquely in birds, skipping exon-5 but not exon-3, producing NF-YAg (Gallo et al. 2023). 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 increases 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 Casola and Betr n 2017; Ciomborowska-Basheer et al. 2021). In mammalian species, the retrocopy remains in most cases an unexpressed retropseudogene as it is devoid of the regulatory regions of the parental gene; these nonexpressed 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 (subfunctionalization). Depending on the expression patterns of the retrogene, the new entity can populate new “territories,” or be further restricted. It has been reported that a high number of retrogenes are expressed mostly or exclusively in male germ cells (Carelli et al. 2016).

We recently reconstructed the phylogenetic history of NF-YA in the evolution of deuterostomes (Bernardini et al. 2022; Gallo et al. 2023). Gene duplications are essentially found only in fishes (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.ccb.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 duplications in selected groups of mammals.

Results

Identification of a Second NF-YA Gene in Ruminants, Cetaceans, Bears, Squirrel, Hippopotamus and a Bat

During our phylogenetic studies on NF-YA in deuterostomes (Bernardini et al. 2022), we noticed an additional sequence in goat (*Capra hircus*), with a predicted amino acid sequence similar to canonical NF-YA, but also containing oddities, possibly due to errors in the sequencing and/or annotations. 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*), bison (*Bison bison*), cow (*Bos taurus*), sheep (*Ovis aries*), and Yarkand deer (*Cervus hanglu yarkandensis*). The two latter

species were excluded because of incomplete (and somewhat patchy) conservation with NF-YA. In pronghorn (*Antilocapra americana*), two additional sequences were found.

Next, we surveyed cetaceans, members of the *Cetruminantia* clade, such as common bottlenose 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 (*Hippopotamus amphibius*) also harbors 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 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* carnivores American black bear (*Ursus americanus*), Asiatic black bear (*Ursus thibetanus*) and giant panda (*Ailuropoda melanoleuca*), but not in polar bear (*Ursus maritimus*). We found a match also in squirrel and related species (*Sciuridae*): thirteen-lined ground squirrel (*Ictidomys tridecemlineatus*), 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 (see supplementary fig. S2, Supplementary Material online); note that we could not find NF-YAr in another *Sciuridae* member, the Arctic ground squirrel. Finally, we found a NF-YAr 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, Alpine marmot 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 correspond to glutamine codons (CAG or CAA) in the other sequences, which end up

being TAA or TAG (supplementary table S1, Supplementary Material online); on the other hand, the DBD stop codons have a coherent logic, discussed below. The translated CDS of NF-YAr is predicted to be slightly shorter, but otherwise showing clear homology in overall length and amino acid sequence. The three Blocks previously identified within the TADs of extant deuterostomes (Bernardini et al. 2022) contain a majority of conserved residues. As control, alignment of the canonical NF-YA genes in all these species confirms the almost perfect conservation, as expected (supplementary fig. S3, Supplementary Material online).

The search for related protein sequences across all main mammalian lineages retrieved a match only for a small stretch within the Q-rich TAD of the protein in almost all species (supplementary fig. S4, Supplementary Material online), but ORF conservation is lost. Repeating the search at DNA level (BLASTN) revealed that the genomic location of these sequences matched that of a human retropseudogene (annotated as NF-YAP1). This pseudogene shows signs of homology with NF-YA (supplementary fig. S5a, Supplementary Material online), but the unit is full of disruptive changes and insertions and the resulting short ORF lacks a coherent biological logic (supplementary fig. S5b, Supplementary Material online). Most importantly, searches for expression in numerous human RNA-seq databases yielded negative results (not shown), thus qualifying it as an actual pseudogene. Sequences corresponding to NF-YAP1 and sharing the same genomic context were retrieved in all placental mammals (except for *Ruminantia*, *Muroidea*, and hippopotamus, likely due to secondary losses), suggesting an ancient common retrotransposition event at the base of the eutherian lineage, followed by pseudogenization. We conclude that NF-YAP1 pseudogene is an element ancestral to all eutherian mammals, but its coding potential and its similarity with the parental NF-YA underwent massive disruptive changes. Instead, NF-YAr retrogenes are present in distinct and restricted mammalian lineages, and their potential functionality warrants further analyses.

Phylogenetic Reconstruction of NF-YAr

The observations above for NF-YAr could be explained by a single retrotransposition in the common ancestor of mammals, which was fixed in some species, but lost during evolution of the others. The alternative hypothesis is that multiple retrotransposition events might have occurred in the different ancestors of each taxon. To clarify this point, we constructed a phylogenetic tree using NF-YAr and NF-YA cDNAs alignments, devoid of gaps, with chicken NF-YA gene as the outgroup (Fig. 2). NF-YAr of ruminants, cetaceans, carnivores, hippopotamus, rodents, and bat are more similar to their respective parental NF-YA than to each other, with robust bootstrap confidence levels, suggesting a separate retrotransposition event in each clade. In

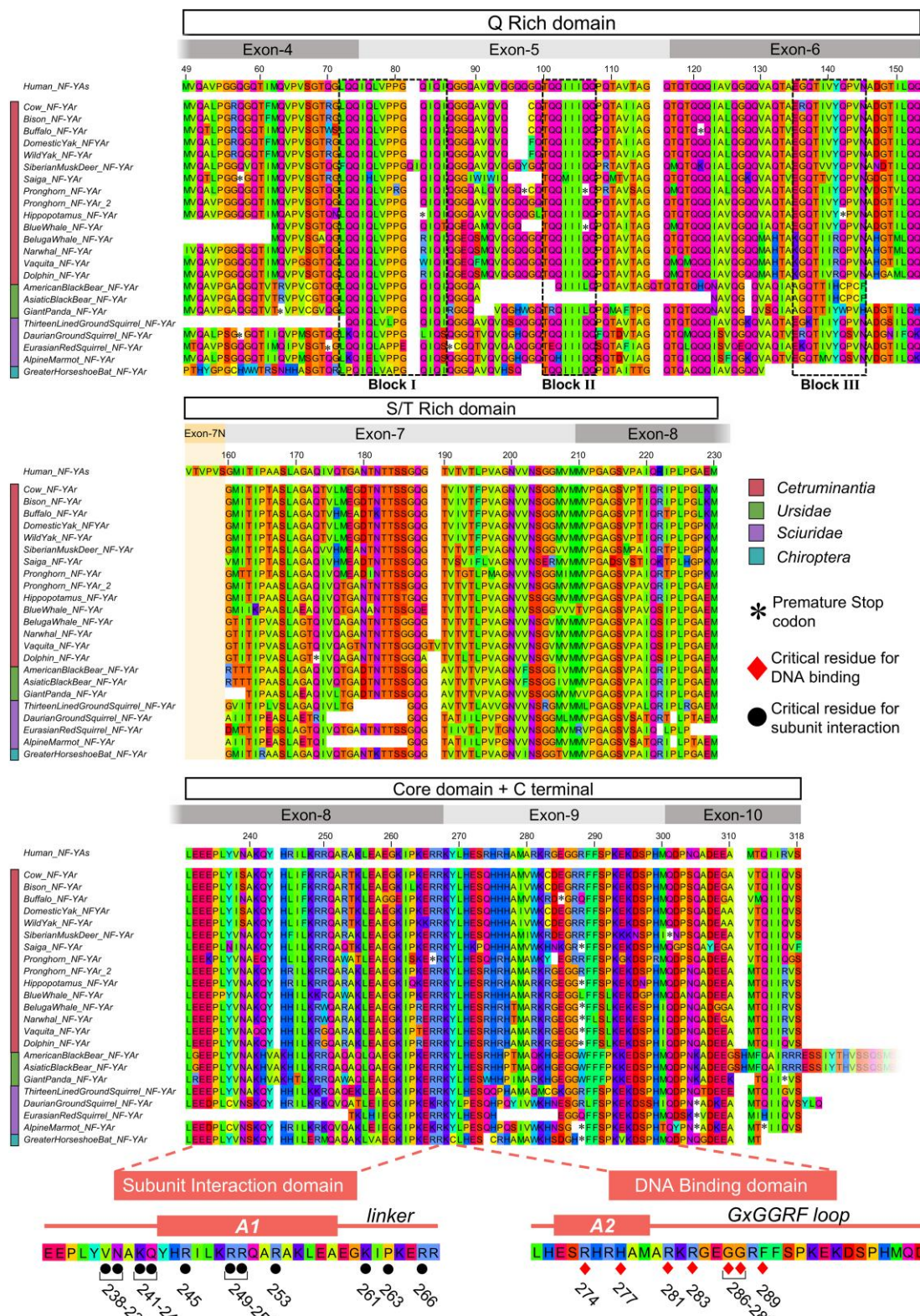


Fig. 1. MSA of three stretches of NF-YA amino acid sequence: from top to bottom, Q-rich domain, S/T rich domain and core domain + C-terminal. Regions encoded by each exon are indicated by gray boxes. Human NF-YAs was selected as reference for residues numbering. Premature Stop codons are signaled within the alignment by an asterisk, and the shaded area highlights 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.

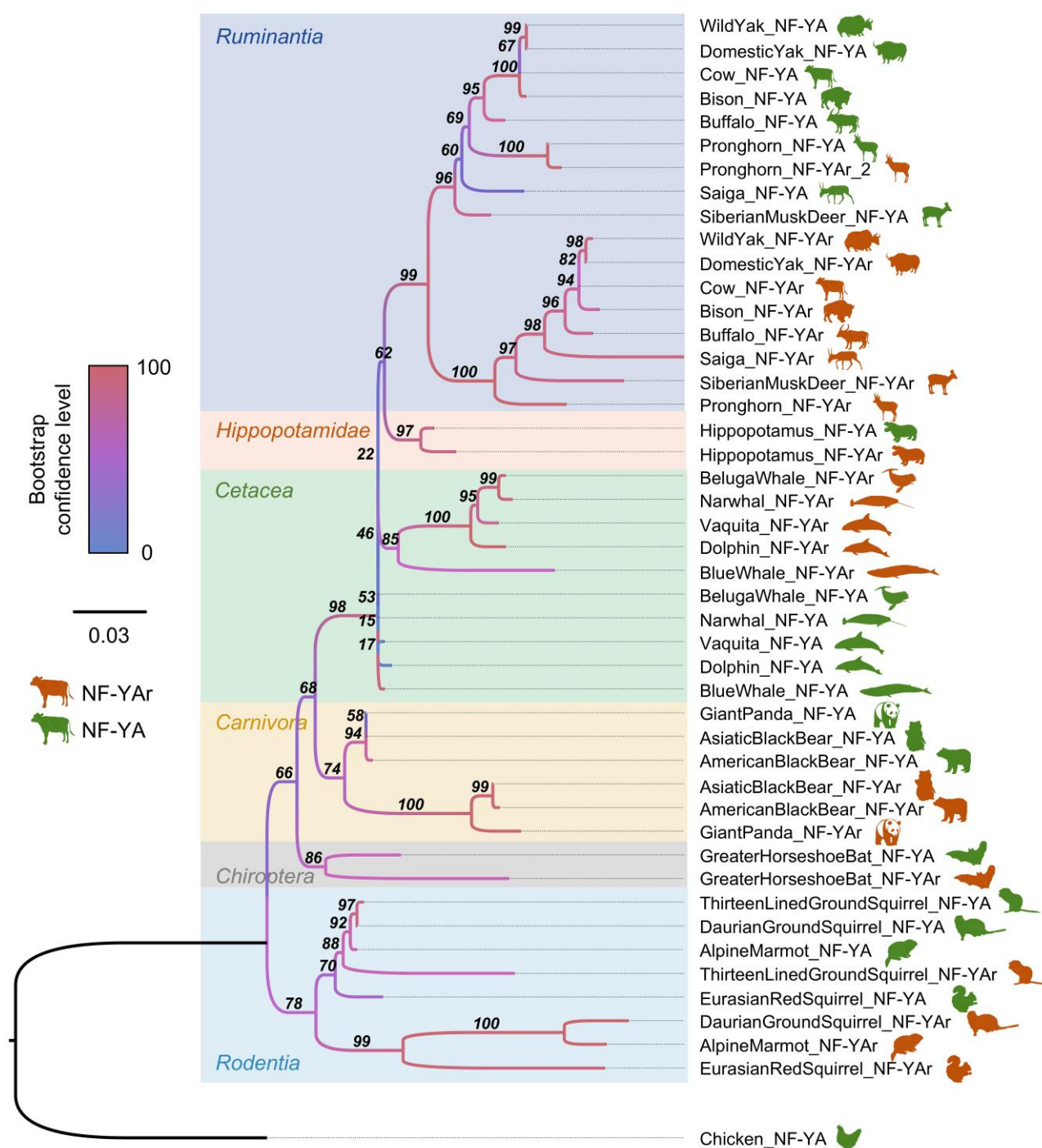


Fig. 2. The gene tree of NF-YA and NF-YAr includes the mammalian species from Fig. 1, plus *Gallus gallus* serving as an outgroup. Bootstrap confidence levels are represented by a color gradient and indicated above each branch. The six taxa associated with each candidate retrotransposition event are highlighted in the background of the tree. NF-YAr and NF-YA leaves are marked by orange and green silhouettes, respectively. Graphical representation of the tree was achieved with the FigTree software; tree branch length = number of nucleotide substitutions per site.

ruminants, we notice that the two NF-YAr of pronghorn have different locations on the tree: NF-YAr clusters with all other retrogenes of *Ruminantia*, an indication of an early retrotransposition in the common ancestor; NF-YAr_2,

instead, segregates with NF-YA, an indication that its retrotransposition is more recent and happened a second time in this particular species. In parallel, we constructed a tree based on the amino acid sequence of the N-terminal

Q-rich region, which is the less conserved and potentially more informative domain: [supplementary fig. S6, Supplementary Material](#) online shows a very similar distribution, with the exception of the thirteen-lined squirrel, which is closer to greater horseshoe bat in the protein tree, whereas it is grouped with the other squirrels in the cDNA tree. The conclusion of this analysis is that NF-YAr sequences can be divided into six groups. The difference related to the thirteen-lined squirrel positioning will be further explained by the synteny analysis below.

NF-YAr Genomic Structure

Retro(pseudo) genes are typically intronless, or possess two exons, and are located on chromosomal locations different from that of the parental gene. 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. 3. In most cetuminants, 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 (see [supplementary fig. S7, Supplementary Material](#) online). In pronghorn, the two NF-YAr genes are located on distinct scaffolds, confirming an independent origin. These data support the notion that NF-YAr was not generated by a local gene duplication event, but rather by events of retrotransposition of the processed mRNA and insertions in different chromosomal locations.

Synteny Analysis Confirms Independent Origins of NF-YAr in Different Orders

Analysis of synteny allows tracing the path of duplication events, and the subsequent post-duplication rearrangements (Liu et al. 2018) with detection of conserved genomic segments (blocks) across species, each featuring an analogous gene order derived from a common ancestor. To further explore whether NF-YAr arose from an ancient duplication of the NF-YA locus, or more recent independent events, we leveraged the Syntenet pipeline, which constructs whole-genome synteny networks by integrating pairwise intra/interspecies genomic associations (Almeida-Silva et al. 2023).

First, we selected 18 different species from the 22 depicted in Fig. 1, according to the availability of Ensembl annotations and proteomes, adding polar bear (*Ursus maritimus*) despite the absence of a complete homologous sequence to the related *Ursidae*. Figure 4a depicts four distinct synteny clusters of NF-YAr orthologs: one gathering ruminants; the second including all cetacean species (note that hippopotamus was included neither in cetaceans nor in ruminants); the third is populated by *Ursidae*; the fourth contains Alpine marmot and Eurasian red squirrel. Thirteen-lined

ground squirrel and greater horseshoe bat were initially clustered together in a group spanning between *Rodentia* and *Chiroptera*. However, this clustering was deemed unreliable due to two factors: (i) the significant distance (~2 million bp) between NF-YAr and the other genes of its syntenic block in the greater horseshoe bat, and (ii) the presence of several unrelated loci within this region that are not genomically associated with the thirteen-lined ground squirrel. As expected, NF-YA orthologs are part of a sixth separate cluster, retrieved in all mammals (not shown). American bison, black bear, and Daurian ground squirrel were not grouped into any of the NF-YAr clusters upon network inference: this could be attributed to poor assembly and/or annotation quality around the NF-YAr locus (unnamed genes, short and sparse scaffolds).

A close inspection of NF-YA neighboring genes ([supplementary fig. S8, Supplementary Material](#) online) and of the multiple NF-YAr synteny blocks (Fig. 4b) shows that none of the annotated genes are shared across blocks, barring a single, ancestral duplication of the NF-YA block, while hinting at least five different, taxon-specific insertion points of NF-YAr. In summary, these data lend further support to the hypothesis of independent origins of NF-YAr in different mammalian orders.

Evolutionary Dynamics of NF-YAr Domains

To understand the evolutionary dynamics of genes, a useful measure is the calculation of the rate of nonsynonymous versus synonymous substitutions per site (dN/dS). Leveraging the codeML tool from the PAML package, first we calculated ratios (ω) for the parental NF-YA sequences, selecting the species included in the NF-YAr syntenic blocks presented in Fig. 4. We used four models: (i) the free-ratios branch model, which assumes heterogeneous evolutionary rates among the different branches of the associated gene tree and homogeneous selective pressure across codons (Yang and Nielsen 2002); (ii) the neutral model M0, which infers a single ω value throughout tree branches and protein positions; (iii) the site model M1, underlying evolutionary relaxation at specific positions; (iv) M2, which allows for site-directed positive selection (Yang et al. 2000). [supplementary fig. S9a and b, Supplementary Material](#) online show data consistent with extreme constraints in NF-YA protein variability. Note that the apparent high evolutionary rate in yak species is dependent on the absence of synonymous substitutions in these sequences, causing a division by 0 (detailed in Álvarez-Carretero et al. 2023). We gathered NF-YAr results for the four syntenic groups analyzed separately, in Fig. 5 for ruminants and cetaceans (more numerous and reliable) and [supplementary fig. S9, Supplementary Material](#) online for bears and squirrels. Figure 5a shows the log likelihood values (lnL) for M1, M2, and M0 either with or without ω set to 1, and the pairwise comparison between the models

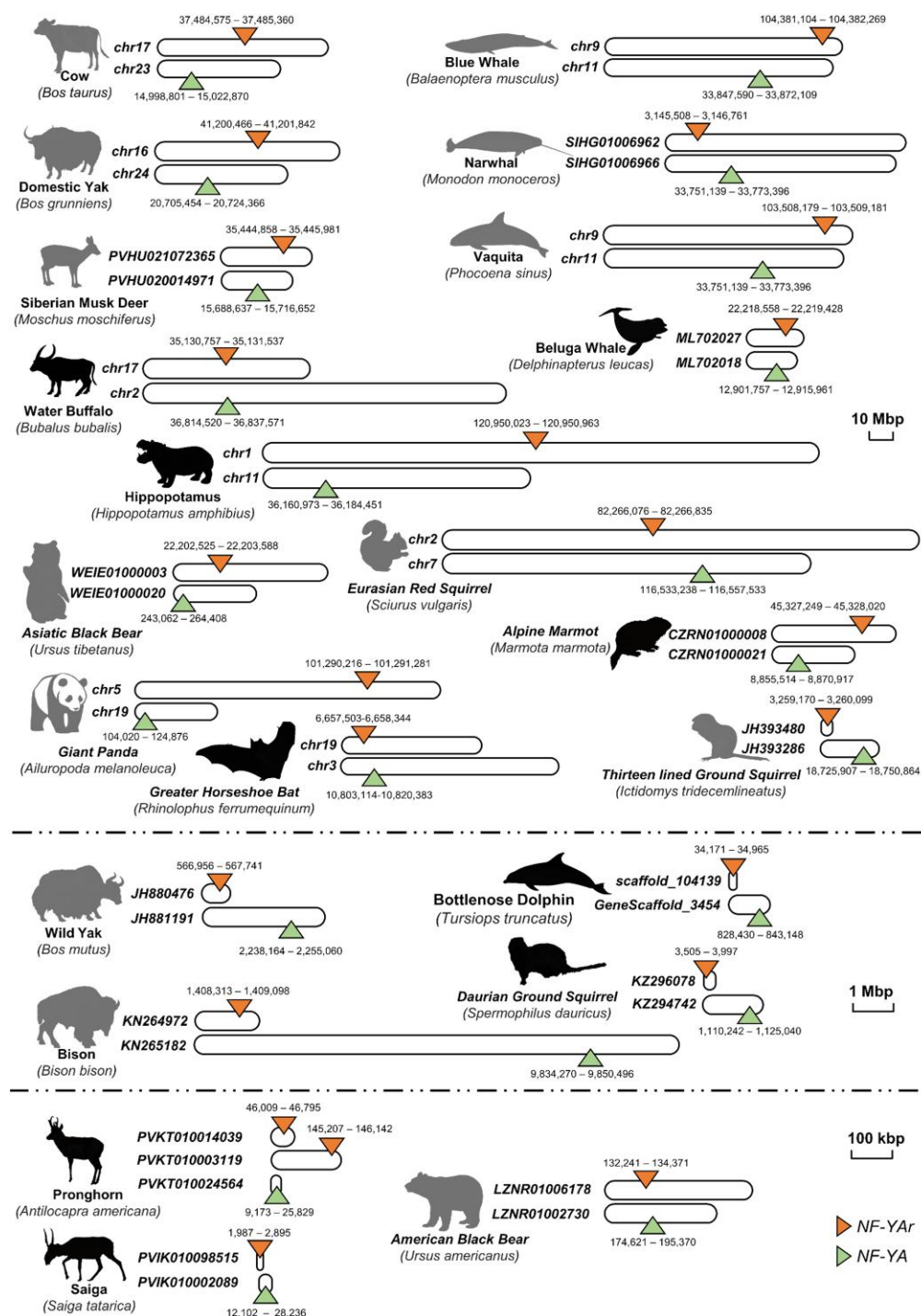


Fig. 3. Schematic representation of the chromosomes of 22 mammalian species, together with the coordinates of NF-YA (green) and NF-YAr (orange). The genome of animals represented by a gray silhouette includes an annotation for NF-YAr in Ensembl, as listed in [supplementary table S2, Supplementary Material online](#), while the ones with black silhouettes do not. Species are arranged in three main groups, depending on the scale used for chromosome length.

via likelihood ratio tests (LRT). Based on the LRT results, we concluded that the M1 model is the best-fit for ruminants, M0 for cetaceans. Figure 5b shows the free-ratios branch model, confirming higher evolutionary rates in ruminants

compared with cetaceans. Figure 5c shows the fraction of residues significant for the M1 model in the two groups, while Fig. 5d shows the partitioning of average ω scores for the M1 and M2 models of the four domains of the

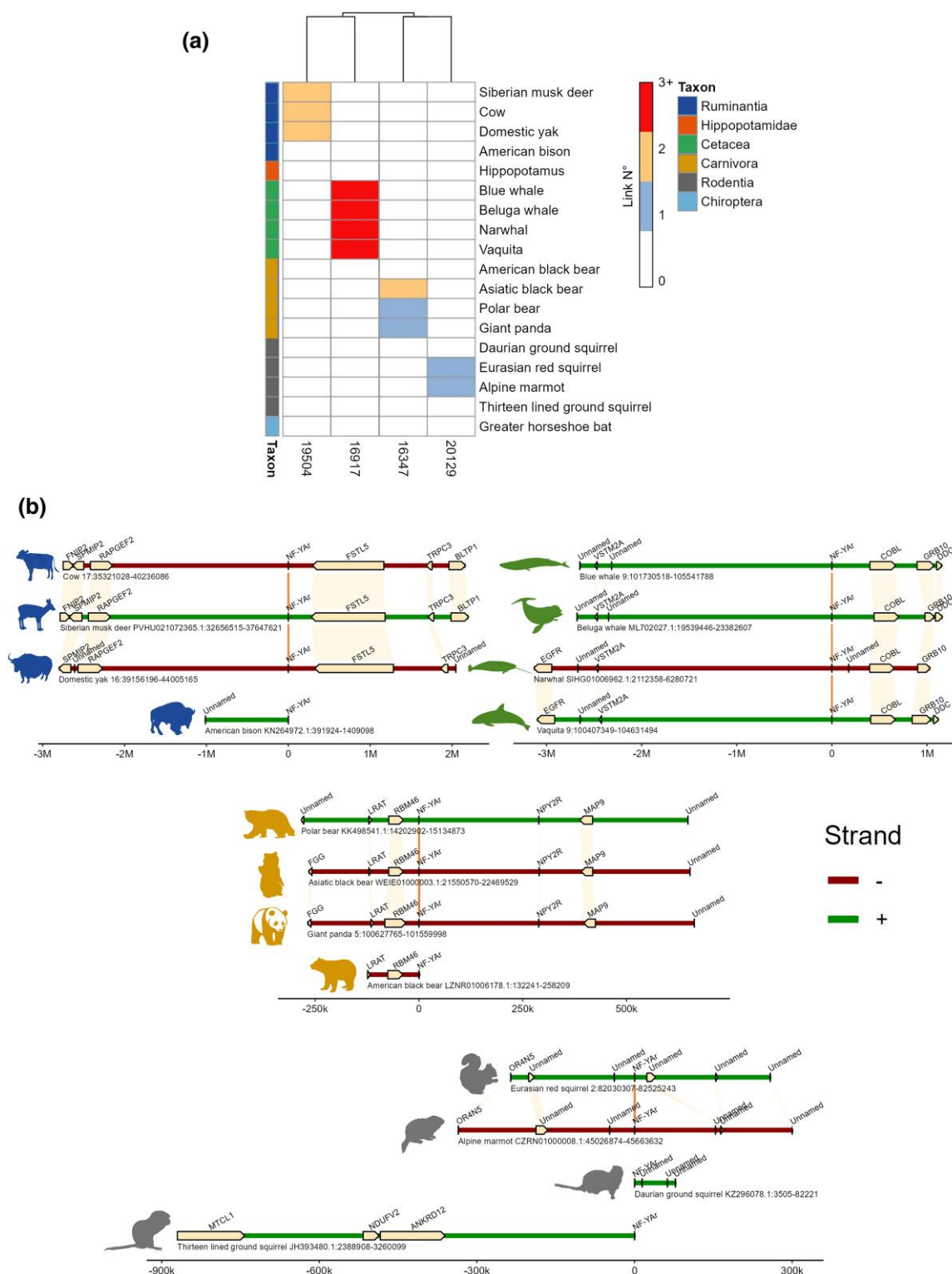


Fig. 4. a) Heatmap depicting the results of Syntenet phylogenomic profiling of NF-YAr loci from several species included in Fig. 1 MSA, with the addition of polar bear (*Ursus maritimus*). The color represents the number of predicted orthologs (links) for each protein sequence, as inferred during the Syntenet pipeline. b) Diagrams illustrating gene order and content surrounding NF-YAr across the species included within each cluster. Up to three genes are depicted upstream and downstream each NF-YAr locus; the "Unnamed" tag indicates genes without a specific denomination (i.e. gene name) besides Ensembl "gene_id" in any of the orthologs considered. Genes that were linked in the Syntenet synteny network are connected by color-coded shades, orange for NF-YAr and yellow for the others, whereas plots coordinates are referred to NF-YAr starting codon position. This illustration was produced using the gggenomes R package.

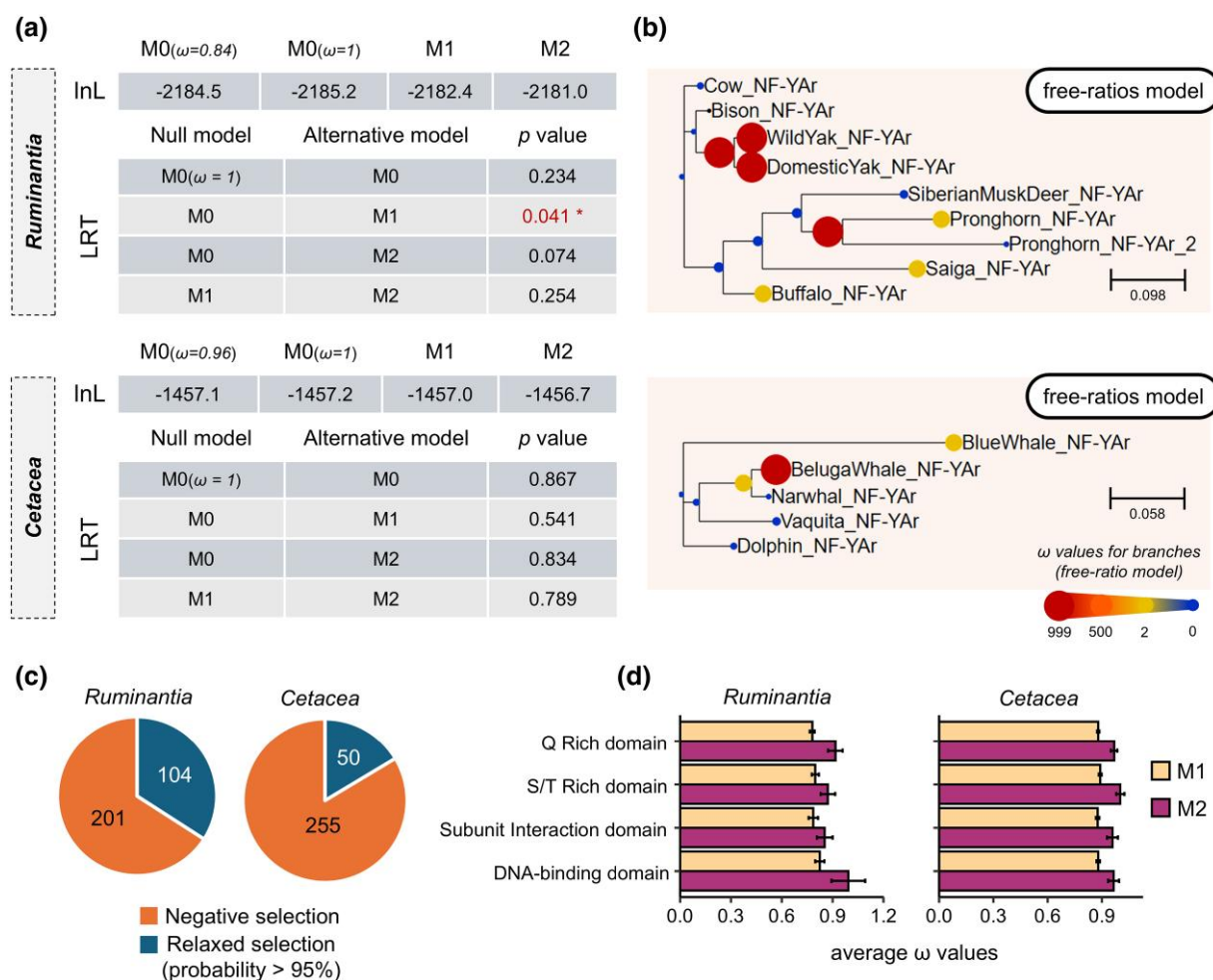


Fig. 5. a) Identification of the best evolutionary model for ruminants and cetaceans NF-YAr sequences. Top panel: log likelihood values (lnL) for M0, M0 with ω fixed to 1, M1 and M2 evolutionary models. Bottom panel: Pairwise comparisons of the models using the LRT. Significant P values are indicated by asterisks; ** = $P < 0.01$, *** = $P < 0.001$. b) Phylogenetic tree depicting node-specific dN/dS ratios (ω) of ruminants and cetaceans NF-YAr retrogenes according to the free-ratios branch model. High ω values are associated to larger dots. Phylogenetic tree branch length = number of nucleotide substitutions per codon. c) Fraction of residues with a significant relaxation in evolutionary pressure (M1 model) in *Ruminantia* (Left) and *Cetacea* (Right). d) Average ω scores for different NF-YAr domains. Error bars = standard error of the mean.

protein: Q-rich TAD, ST-rich, Subunit Interaction, DBD (see [supplementary fig. S1, Supplementary Material](#) online). In ruminants, the subunits interaction region has an average ω score comparable with the Q-rich and ST-rich domains, while in the DBD the average score is higher. Six residues in the DBD (highlighted in green in [Fig. 1](#)) are significant for evolutionary relaxation under the M1 model (probability > 95%), suggesting a local drive for neutral or nonsynonymous changes; this might be functionally relevant, as it will be discussed below. As for bears and squirrels, the latter are similar cetaceans, whereas the former show as best-fit M2, but the presence of only three sequences makes the interpretation of the results difficult ([supplementary fig. S10, Supplementary Material](#) online).

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 also 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 in the majority of the species of the ACC triplet before the ATG (-3/-1), and G at +4, fits with the consensus ([Fig. 6a](#), right panel); in many, the CC preceding at position -5/-4 is also consistent with an optimal Kozak. The exceptions are sequences of beluga whale

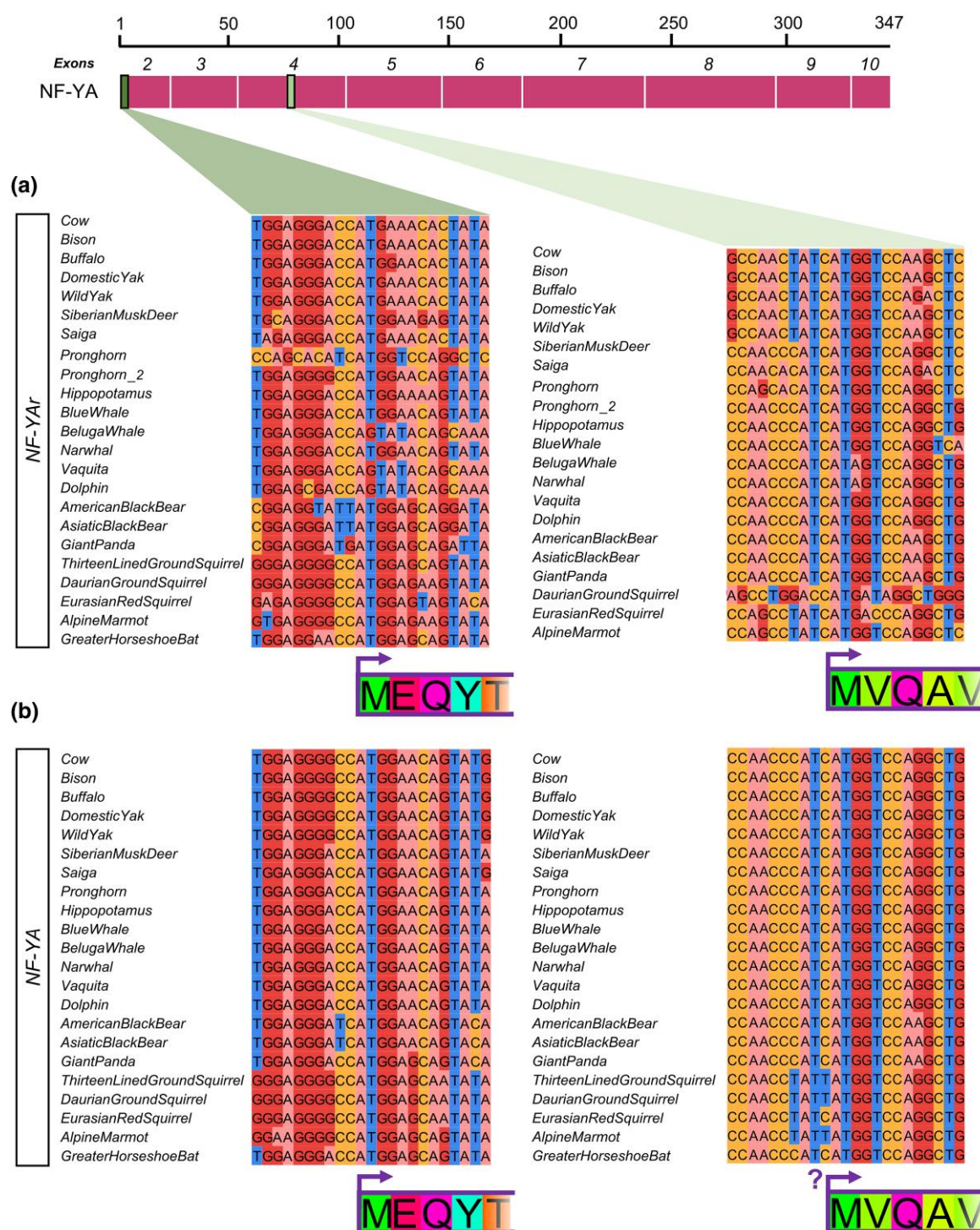


Fig. 6. a) Top: Schematic representation of NF-YA mature mRNA. The position of two ATG codons are marked, Met1 and Met49. Bottom: MSA of NF-YA sequences from selected mammalian species containing bases spanning from –10 to +10 from the two potential start codons. Bottom boxes: first five translated amino acids of NF-YA protein. b). Same as a), except the alignments for the regions around NF-YA Met1 and Met49 are shown. The alignment was exported from the software Jalview.

and narwhal, not because of the surroundings, but for the presence of ATA instead of ATG. On the other hand, several nucleotides deviate from a reasonable Kozak at Met63

(supplementary fig. S11, Supplementary Material online). supplementary fig. S12, Supplementary Material online shows conservation in several species of NF-YA 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, show a good Kozak consensus in most of the species considered (Fig. 6a, left panel). Several 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 (supplementary fig. S12, Supplementary Material online). As reference, Fig. 6b depicts an excellent Kozak consensus at both Met1 and Met49 in the parental NF-YA.

Another relevant observation is that none of the species have sequences corresponding to exon-3 (supplementary fig. S11, Supplementary Material online), which is alternatively spliced out in NF-YA to form the short NF-YAs isoform: this is a strong indication that NF-YAs, not NF-YAI, was the isoform originally retrotranscribed and transposed. We conclude that the presence of conserved Kozak sequences around predicted ATGs 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. Figure 7 (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 on the presence, in all animals except cetaceans and squirrels, of two conserved CCAAT boxes some 20 bp 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 (Oldfield et al. 2019).

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 (Casola and Betrán 2017; Ciomborowska-Basheer et al. 2021, and see RetrogeneDB2, Rosikiewicz et al. 2017). 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 (supplementary fig. S13a, Supplementary

Material online); on the other hand, NF-YAr RNA was found only in spermatozoa (Fig. 8, lower panel), but absent in the other tissues considered including testis (supplementary fig. S13a, Supplementary Material online right panel and not shown); expression of the parental NF-YA is ubiquitously recorded, as expected (supplementary fig. S13a, Supplementary Material online, left panel). Remarkably, we could not detect expression of the parental NF-YA in bovine spermatozoa, as measured by reads corresponding to exons (Fig. 8, upper panel), which is consistent with the reported very low level of NF-YA in zygotes (Halstead et al. 2020). Note that RNAs of NF-YC and, to a lesser degree, NF-YB, were scored (supplementary fig. S14, Supplementary Material online). 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 (supplementary fig. S13b, Supplementary Material online), in keeping with initiation of Zygotic Genome Activation (ZGA) as previously shown in these animals (Halstead et al. 2020). We conclude that, as in the case of many other retrogenes, NF-YAr shows expression in spermatozoa, where it appears to be the only NF-YA present.

Structural Considerations on the NF-YAr Protein Sequences

MSA of NF-YAr protein sequences (Fig. 1) shows features relative to the different domains that deserve punctual comments. Note that the numbering of amino acids refers to NF-YAs, as in Nardini et al. 2013.

- i) Starting at Met49, some proteins apparently lack the evolutionarily least conserved region across deuterostomes (Bernardini et al. 2022). The three blocks identified in deuterostomes (depicted in Fig. 1 and supplementary S1, Supplementary Material online) are generally conserved, with a 2 amino acids insertion in Siberian musk deer and extensive deletions in bears. The absence of the 6 amino acids corresponding to the N-terminal of exon-7, due to alternative splice donor sites (Li et al. 1992), points at a 7N-less mRNA as targeted by the retrotransposition.
- ii) Within the A1 trimerization domain, residues making crucial contacts with NF-YB/NF-YC are conserved (Fig. 1): V238 (conservative changes to Ile in ruminants); N239 (an important residue, it is conserved in cetaceans, bears, squirrels, and bats) is a serine in some ruminants, representing a potential nonconservative change; K241 (conserved, except in vaquita K241 > Q); Q242 is conserved in all, except a histidine in bears; R245 is different in NF-YAr, except for hippopotamus, squirrel, and marmot; R249 is conserved, except a conservative change to Lys in blue whales and squirrels; R250 is conserved in ruminants

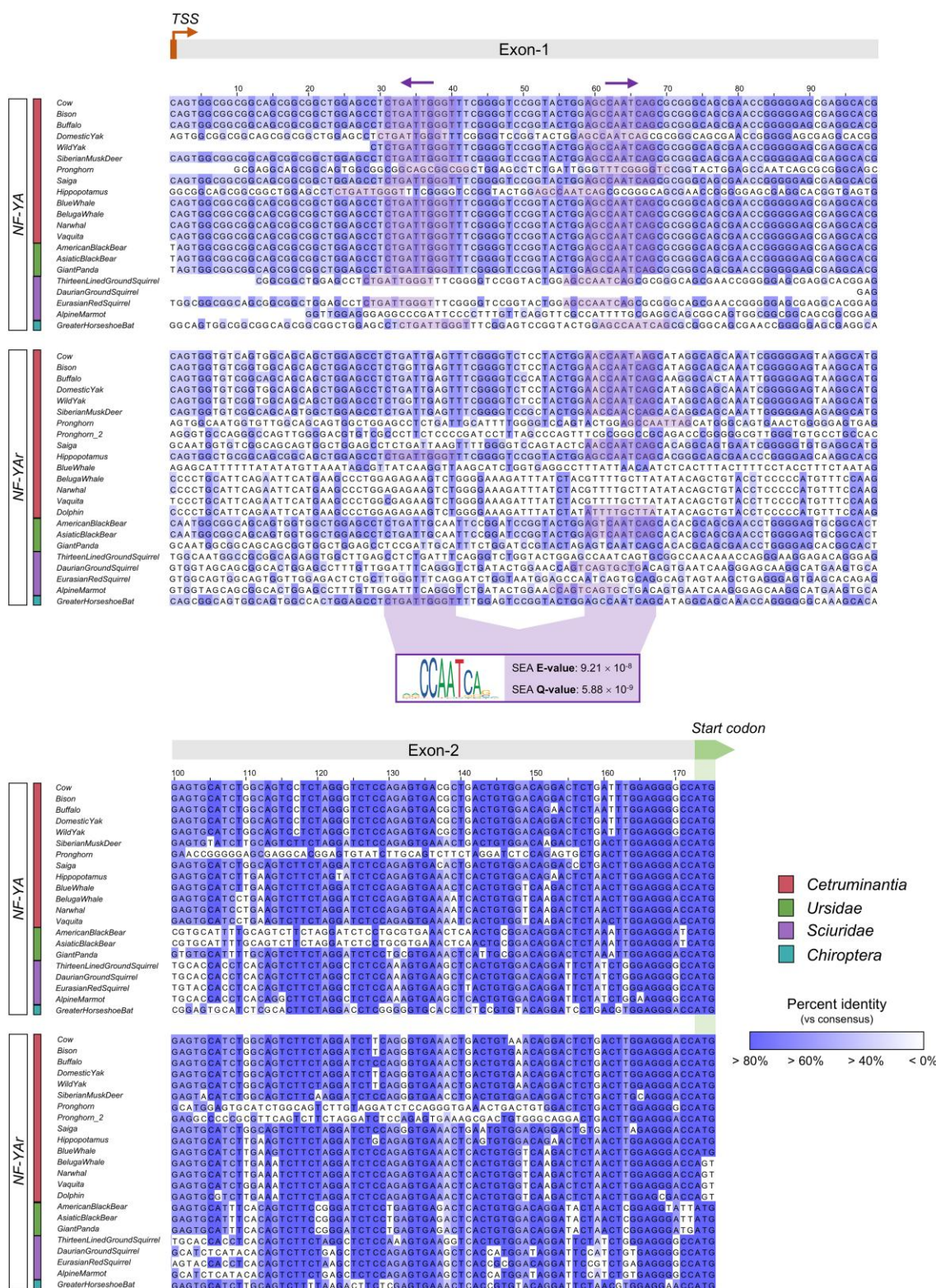


Fig. 7. MSA of NF-YA and NF-YAR cDNA sequences from selected mammalian species, spanning from start codon to 172 bp upstream of it, corresponding to the TSS in *bos Taurus*. NF-YA Exon-1 and Exon-2 are indicated above the sequences by gray boxes. For the species in which a NF-YA Exon-1 annotation was not available, homologous sequences obtained through BLAST/BLAT are included. The CCAAT box enrichment analysis performed by SEA is highlighted, as well as is the motifs within the alignment. Intensity of the color = per-nucleotide percent identity, when compared with the whole alignment consensus. The alignment was exported from the software Jalview.

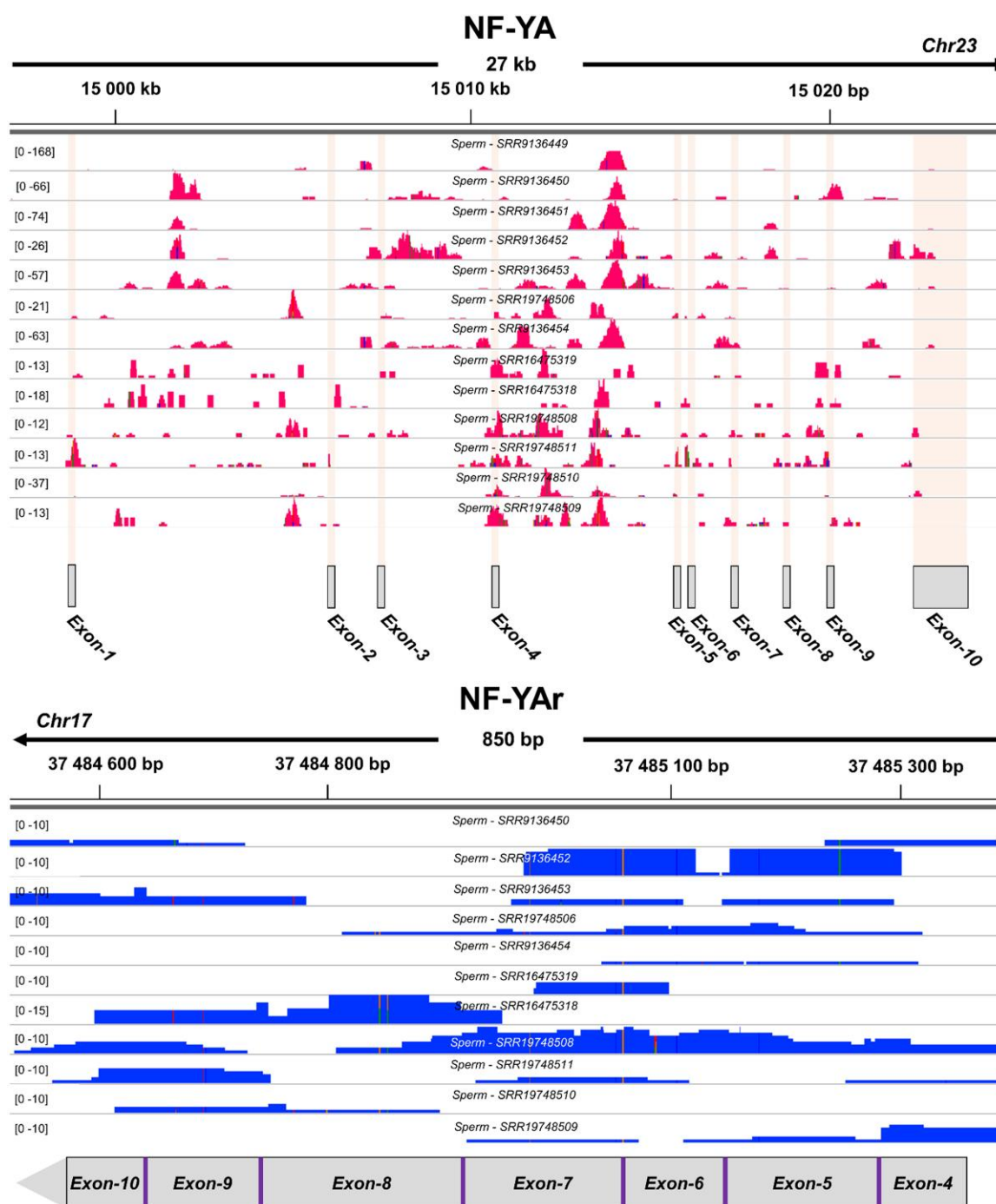


Fig. 8. Top: NF-YA expression from 13 *Bos Taurus* sperm RNA-Seq samples, represented by mapped read coverage. The regions corresponding to each exon are highlighted. Bottom: NF-YAr expression from 11 *Bos Taurus* sperm RNA-Seq samples. The gray boxes delineate regions within NF-YAr with a strong homology 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.

and bears, a conservative Lys in squirrel, but potentially negative Met, Gly, or Trp in bats and marine species. R253 is conserved, but a Trp in blue whale and giant panda, and Gln in saiga, bears, squirrels, and bats: it is not easy to figure out what the consequences of these

changes might be. K261 is conserved except in buffalo (Glu): a minor contact, and Ala mutation retains trimerization. R266 is conserved. The changes found at L247, A254, and K264, all solvent-exposed, to Phe, Thr/Glu, or Arg/Thr should allow trimerization. Overall, the

prediction is that NF-YAr trimerization efficiency should be diminished, at least to some extent, mostly due to changes of N239 and R245 in ruminants, R250 in cetaceans or R253 in bears, squirrels, and bats. Most species display two or more relevant substitutions, the only exception being hippopotamus.

- iii) The sequence and length of the linker between A1 and A2 helices, which guides the trajectory of A2 toward the DNA, is rather conserved.
- iv) NF-YA sequence-specificity relies on 7 amino acids known to contact DNA directly (Fig. 1), all present in cetuminants, bear, squirrels and bat parental NF-YA, but only three are conserved in NF-YAr: H277, G286, and F289. The other residues are variable: 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 (Xing et al. 1993); 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 (Nardini et al. 2013). Marine species all have “canonical” residues at DNA-contacting positions, but display Stop codons after Gly287: the following F289 is essential for DNA binding (Nardone et al. 2017). Bears have a helix-hindering Pro, in addition to Gln at R281, as squirrels and marmot, having a Trp at R281. A Trp further substitutes R289 in bears. 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.
- v) The C-terminal of NF-YAr is conserved, in particular S291 and S297 (S320 and S326 in NF-YAI), residues modified by phosphorylation (Chae et al. 2004; Bernardini et al. 2019). Conservation includes the surroundings of Ser297, which suggests that NF-YAr could be modified by the same kinases, among which is CDK2 (Chae et al. 2004). Surprisingly, C-term conservation is maintained in cetaceans and hippopotamus downstream of the aforementioned stop codon in the DBD.

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 (Jumper et al. 2021). We previously analyzed AF models of the human, bovine, zebrafish, and chicken NF-YA, specifically predicting runs of transient β -stranded motifs from exon-4 to exon-7 within the intrinsically disordered TAD (Gallo et al. 2023; Bernardini et al. 2022). 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 [supplementary fig. S15, Supplementary Material](#) online). The N-terminal domain 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 part relative to the C-terminal part indicating high degree of flexibility, and can be thus viewed as independent domains.

Within the N-terminal domain, the models show different degrees of low-confidence structural organization, centered on two regions ([supplementary fig. S15, Supplementary Material](#) online), namely aa 130 to 148 (NF-YAr) of exon-6 (including the conserved Block 3), always folded as a β -hairpin, and aa 60 to 78, which comprise exon-4/-5 boundary motif Block 1, as observed in previous analyses (Gallo et al. 2023; Bernardini et al. 2022). It is interesting to note that for NF-YAs models with higher structural order, secondary structure (β -) elements extend toward the C-terminus of exon-6 hairpin, including the previously described exon-7 β motif (aa 155 to 188). This is not observed in NF-YAr, where the ordered regions extend instead toward the N-terminal part of the protein. AF best NF-YAr model is shown in Fig. 9, as compared to rank 4 model of NF-YAs, which was selected for its higher secondary structure content. As observed in Fig. 9 (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 domain, 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 (Fig. 9, [supplementary fig. S15, Supplementary Material](#) online 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 Fig. 9. 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 residues 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

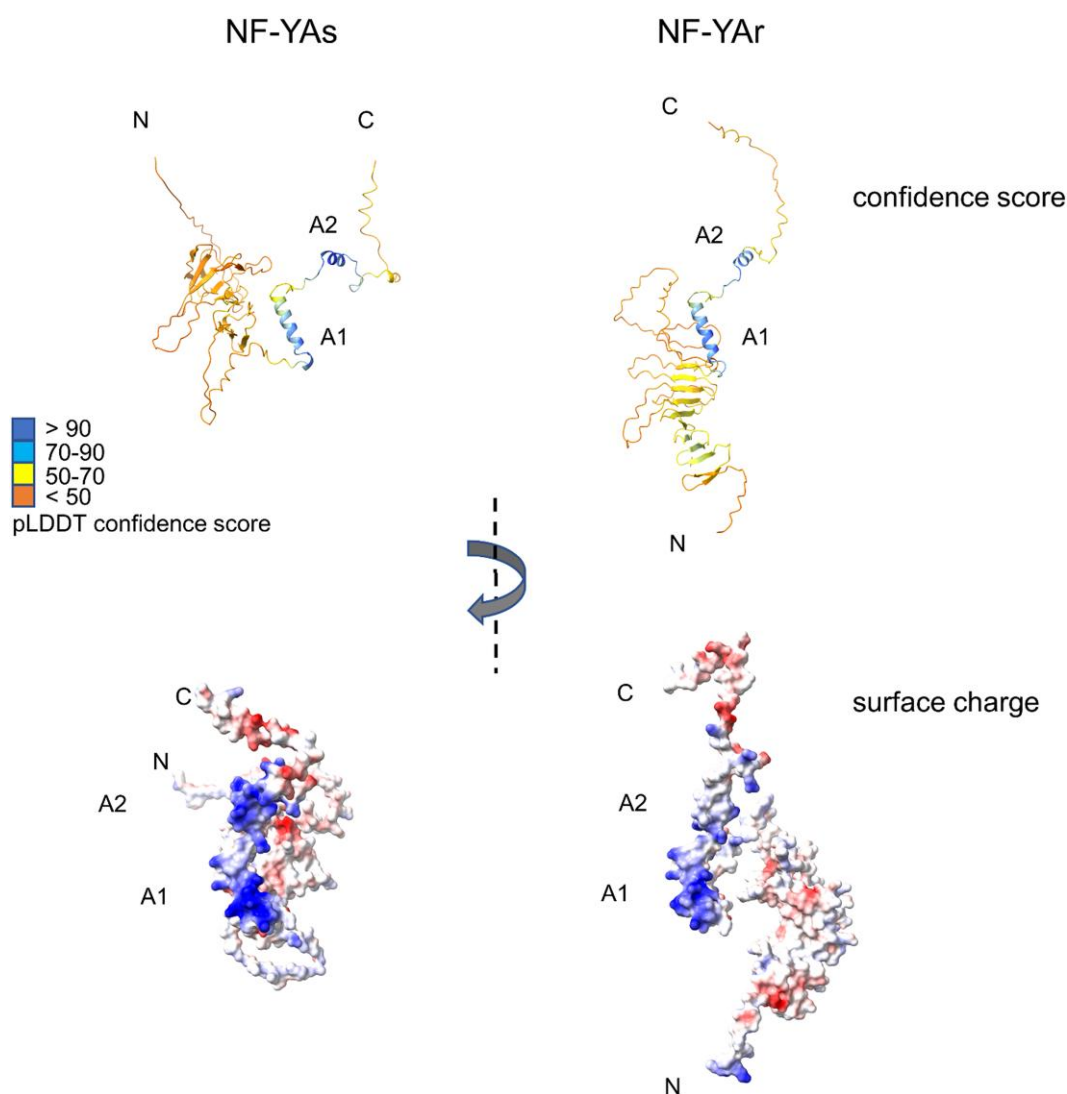


Fig. 9. Top: ribbon depiction of AF models obtained for domestic yak NF-YAs (left) and NF-YAr (right) with AF per-residue confidence score color palette. Bottom: the same models, rotated around the y axis, are shown below in surface charge coloring (blue: positive, red: negative), to highlight interaction 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 [supplementary fig. S11, Supplementary Material](#) online). Structural models images were obtained with ChimeraX.

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 nonpolar surface of the A2 helix would significantly impair (sequence-specific) DNA binding.

Discussion

We systematically searched for NF-YA retrogenes in mammals. Our findings are collectively novel for ruminants, cetaceans, hippopotamus, bears, squirrels, and greater horseshoe bats. We provide evidence that multiple NF-YA retrotransposition events happened independently in different taxa and were then fixed in selected mammals (Figs. 2 to 4).

Phylogenetic analyses and different syntenies in ruminants, *Cetacea* and *Ursidae* are consistent with four independent clade-specific duplication events. Moreover, presence of NF-YAr in hippopotamus, greater horseshoe bats and a second copy in pronghorn and thirteen-lined ground squirrel is clear evidence of four additional independent (potentially species-specific) retrotransposition events. By considering also the NF-YAP1 pseudogene found in placental mammals, the total number of independent NF-YA gene duplication events raises to nine (Fig. 10), hinting at a rather strong retrotransposition tendency of the parental NF-YA gene in mammals.

Most of the studies on retrocopies regard mouse and human genomes, with a minority of retrogenes carrying or

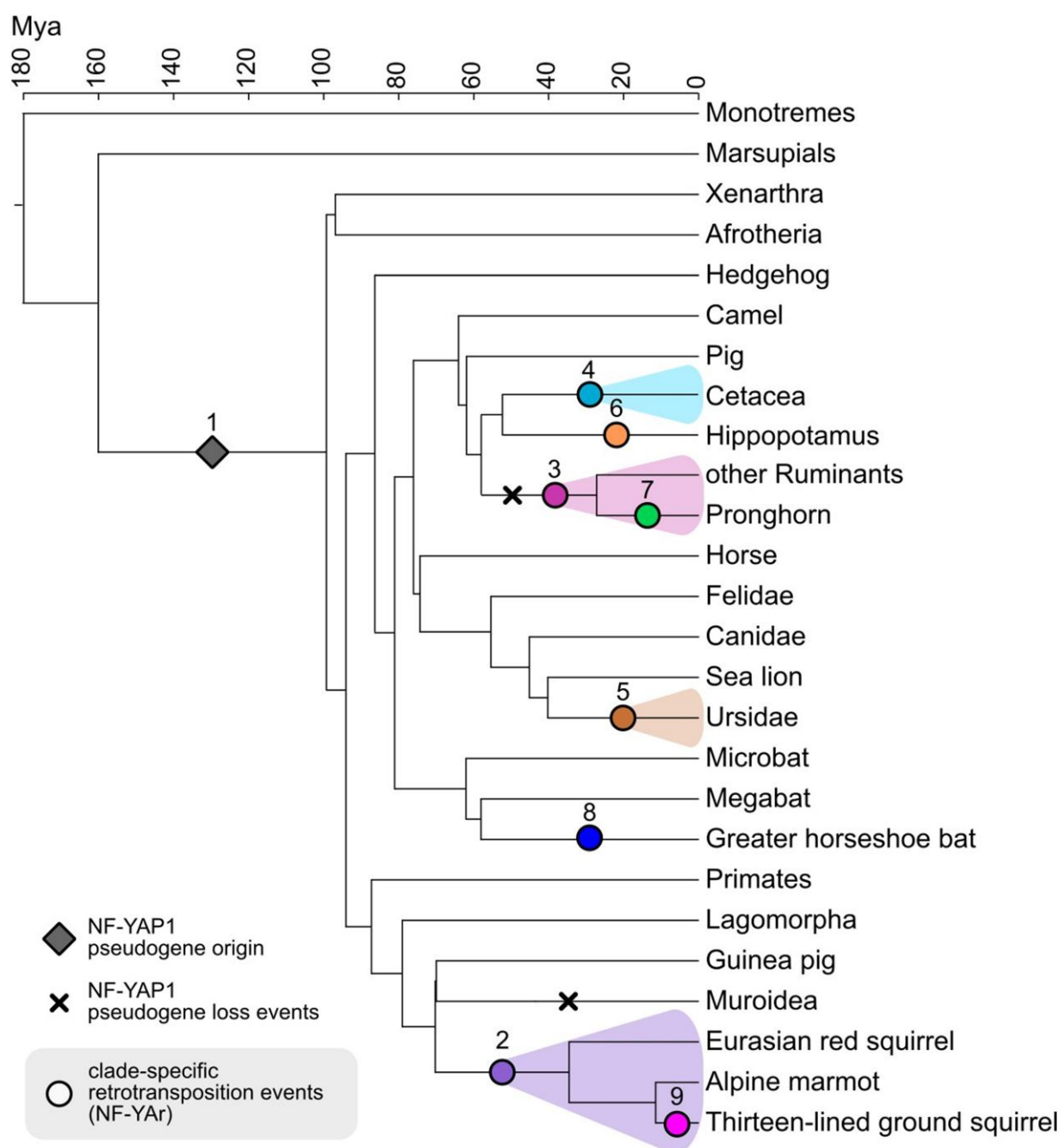


Fig. 10. Summary of the NF-YA retrotransposition events across mammals mapped in this study. The species tree was constructed using TimeTree (Kumar et al. 2022) and every independent retrotransposition event inferred in the study is indicated with a colored mark in the middle of the corresponding branch and numbered according to the predicted chronological order. Secondary losses of the NF-YAP1 pseudogene are indicated with a cross.

acquiring upon integration functional elements leading to expression; an even tinier minority produces a protein. NF-YAr appears to be a nonexpressed retropseudogene in most mammals, with changes accumulating, leading to abolition of the CDS. A relevant finding concerns the conservation of NF-YAr in selected species within the same order: in *Ursidae*, the absence of a secondary NF-YAr in Polar

bear, which are more related to American and Asiatic bear than to giant panda, is notable. In bats, the retrogene is absent in 5 of the 6 species for which a complete genome is available (Jebb et al. 2020). In rodents, it is present only in *Sciuridae*.

NF-YAr could have a regulatory role at the level of the transcript, for example by sponging miRNAs or lncRNAs,

as it is the case for other retrogenes (Poliseno et al. 2010; Troskie et al. 2021); yet, conservation of the predicted CDS and translation signals suggests that NF-YAr could be translated.

In contrast to this, the protein alignment of the NF-YAr Q-rich domain (Fig. 1) shows multiple nonconserved stop codons. However, many of these may result from sequencing errors present in some of the assemblies we included in the analysis. For example, in pronghorn, a Cys residue found in the Subunit interaction domain in one assembly (GCA_021018645) was corrected to a more “conventional” Leu in a different one (GCA_004027515). Additionally, most of these presumed stop codons are followed by the restoration of a continuous ORF, suggesting that these sequences might undergo transcript readthrough.

Evolutionary considerations based on substitutions/truncation found in the C-terminal domain of the predicted protein suggest loss of DNA binding, whereas changes in the Subunit interaction domain and in the N-terminal TAD leave open the possibility that NF-YAr acts negatively with respect to the basic functions of the parental gene.

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 ruminants 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 Whole Genome Duplications in these species (Meyer and Van de Peer 2005; Inoue et al. 2015). The NF-YA retrotransposition events presumably took place in male germ cells, originating from the “short,” 7N-less RNA isoform, as shown by absence of exon-3 and of the six amino acids at the N-terminal side of exon-7. NF-YAs isoform is predominant in totipotent stem cells (Dolfini et al. 2012; Oldfield et al. 2014), in stem cells of other mammalian tissues (Dolfini, Imbriano, and Mantovani 2025). 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 strong Kozak sequences around Met49 also in the human and rodents sequence of parental NF-YA could suggest that this signal might be used under certain, yet undetermined, circumstances.

The conservation of the protein CDS of NF-YAr deserves some considerations. The positively charged A1 helix mediates formation of the trimer by contacting a negatively charged surface groove on the HFD heterodimer (Nardone et al. 2017). The trimer is required for formation of a stable complex with DNA, mediated by specific residues of A2 and the GXGGRF motif. Changes are present

in the subunits-interacting A1, but even more in the A2/GXGGRF DNA-binding domain part, in particular with regard to positive charges (Fig. 1). Mutations in the A2 and GXGGRF motif allow trimerization but they obliterate DNA binding in vitro (Mantovani et al. 1994) and, when overexpressed in mammalian cells (Reviewed in Dolfini et al. 2009) or mice in vivo (van Wageningen et al. 2008; Silvestre-Roig et al. 2013). The many changes in NF-YAr are consistent with lack of DNA binding by NF-YAr: so, are these naturally occurring Dominant Negative 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, but important residues mediating trimerization are different in NF-YAr, and some amino acids have been shown to be detrimental for trimer formation (Xing et al. 1994). Is this sufficient to rule out a Dominant Negative activity? Possibly not and different scenarios can be evoked. (i) There might be retrocopies of NF-YB and/or NF-YC in these species, which might have co-evolved to “adjust” to the changes of NF-YAr. (ii) Other conserved NF-YB/NF-YC-like exist, potentially apt to trimerize with NF-YAr: NC2a/b and Pole3/4. The former forms a complex with TBP impacting -negatively- on TATA boxes, the latter is part of DNA Pol ϵ involved in DNA replication (Gnesutta et al. 2013). 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 could bind to a sequence slightly different from CCAAT, through the changes in A2. This might be theoretically hypothesized in ruminants, but the truncation of NF-YAr makes this hypothesis remote in cetaceans.

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, even with loss of the DNA-binding, argues in favor of maintaining the ability to compete with NF-Y-interacting cofactors. The short structural motifs within the otherwise unstructured TAD, as predicted by AlphaFold, is a further suggestion. In other words, playing a Dominant Negative function by “squenching” coactivators.

Analysis of bovine RNA-seq databases suggests that NF-YAr expression is selective for spermatozoa, unlike NF-YA. This is typical of expressed retrogenes (Carelli et al. 2016): how this is accomplished, as well as the evolutionary scope, are subjects currently debated (Casola and Betrán 2017). Numerous studies in mice indicated an important role of NF-YA in spermatogonial stem cells (SSCs) and in the early phases of spermatocytes differentiation (Guo et al. 2017; Li et al. 2018; Maezawa et al. 2018): it seems unlikely that the parental NF-YA would behave differently in cow. Assuming that NF-YAr is functional, what might be the specific role in spermatozoa? We can think of two possible answers. The first relates to a potential

Dominant Negative role during terminal differentiation of spermatozoa, by competition with parental NF-YA on activation of *cell-cycle* genes, one of the major targets in the NF-Y regulome (Ronzio et al. 2020). The second is based on studies of early development of mammals. 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 Zygotic Genome Activation ZGA in several species (Lu et al. 2016; Liu et al. 2019) including bovines (Halstead et al. 2020), which fits with mounting evidence that NF-Y plays a pivotal “pioneering” role in establishing an open chromatin configuration (Nardini et al. 2013; Oldfield et al. 2019; 2014). Therefore, NF-YAr might delay ZGA in ruminants by competing with the lowly expressed NF-YA, which we confirm is activated at the onset of ZGA in bovines (supplementary fig. S13, Supplementary Material online, Halstead et al. 2020).

Materials and Methods

Retrieval of NF-YAr Sequences

NF-YAr protein and DNA sequences were obtained by consulting Ensembl release 110 (Martin et al. 2023) and UCSC genome browsers (Kent et al. 2002). For bowhead whale, we retrieved the sequences from the website <http://www.bowhead-whale.org/> (Keane et al. 2015). We performed TBLASTN (Altschul et al. 1990) or BLAT (Kent 2002) 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 S2, Supplementary Material online, together with NF-YA and NF-YAr gene locations. Protein and DNA sequences were edited in Jalview (Waterhouse et al. 2009; Troshin et al. 2011), aligned using Muscle (Edgar 2004) with default settings, and are listed in supplementary tables S3 and S4, Supplementary Material online, respectively. CCAAT box enrichment in the suspected promoter region of NF-YAr locus was calculated with the tool Simple Enrichment Analysis, from the MEME suite (Bailey and Grant 2021).

Gene Tree Construction

NF-YAr and NF-YA cDNA sequences associated to the species included in Fig. 1 MSA, plus NF-YA from *Gallus gallus*, were aligned again using muscle, and gaps were removed with the tool trimAl (version 1.4, Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009), using the *-nogaps* option. A Newick format tree was then built using the gap-free alignment with IQ-TREE (version 1.6.12, Minh et al. 2020; Hoang et al. 2018), selecting GTR+I+G as models (*-m* option) and 1000 ultrafast bootstrap replicates (*-bb*

option). The tree was then imported FigTree (version 1.4.4, <http://tree.bio.ed.ac.uk/software/figtree/>), rerooted on the Chicken_NF-YA branch, and labeled according to the bootstrap confidence value of each branch.

Synten Analysis of NF-YA and NF-YAr Loci

We built a whole-genome synteny network using the R package Syntenet (version = 1.3.5, Almeida-Silva et al. 2023), collecting data from 18 different *Mammalia* species mostly derived from Fig. 1 MSA, with the addition of polar bear. Specifically, 9 *Artiodactyla*, 4 *Ursidae*, 4 *Rodentia*, and one *Chiroptera* species were selected.

The Syntenet pipeline requires two classes of input: gene annotations and amino acidic sequences collections, or proteomes. These species-specific paired input files were produced through an in house Python script (version 3.9). Original annotation files in GFF3 format and the collections of CDS in Fasta format were obtained from Ensembl (release 110). From the initial GFF3 files, only *gene* features were retained for final gene annotation files. Protein sequences from the UniProt database proteomes (release 2024_2, <https://www.uniprot.org/>) were preferentially adopted for preparing protein collections; when not available, translated CDS sequences were employed instead. NF-YA and NF-YAr synteny clusters were singled out from the complete network and depicted through Syntenet *plot_profiles* command, while gene order diagrams were generated using the gggenomes R package (version 1.0.0, <https://github.com/thackl/gggenomes>).

dN/dS Ratio Calculation

NF-YA and NF-YAr dN/dS ratios (ω) were computed through the *ete3 evol* command from the ETE toolkit, a python-based resources which can automate and streamline PAML codeml analyses (version 3, Huerta-Cepas, Serra, and Bork 2016), underlying multiple evolutionary models: the free-ratios branch model (*fb* option), the zero-ratio M0 model, and the site models M1 and M2. To test whether ω values were overall significantly different from 1, we also considered a variant M0 model where ω was fixed to 1 (*fix_omega* = 1 and *omega* = 1 in the codeml control file). The phylogenetic trees provided as input together with NF-YA and NF-YAr cDNA aligned sequences were generated using IQ-TREE, with the same parameters described above.

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 S5, Supplementary Material online), using the SRA Explorer website (<https://sra-explorer.info/>). We mapped FASTQ files using STAR (version 2.7.8a, Dobin et al. 2013), 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) (Robinson et al. 2011). Finally, we acquired bovine early embryogenesis NF-Y subunit expression data from the NCBI GEO DataSets accession GSE52415 (Graf et al. 2014). 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.

Protein Structure Modeling and Analysis

Protein structure models were generated using AlphaFold2 (Jumper et al. 2021; Mirdita et al. 2022) with the AlphaFold Structure Prediction tool within the UCSF ChimeraX application (Pettersen et al. 2021) 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 (Pettersen et al. 2021).

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

The authors acknowledge support from the University of Milan through the APC initiative.

Author Contributions

A.B. made the original observation and A.G., A.B., S.P., and N.G. performed the experiments; D.D. supervised the work; R.M. wrote the manuscript.

Funding

This work was supported by Ministero dell'Università e della Ricerca (MUR) - PRIN #20224TWKNJ to D.D.

Conflict of Interest

The authors declare that they have no competing interests.

Data Availability

Protein and DNA sequences of NF-YAr are included in supplementary tables S3 and S4, Supplementary Material

online, respectively. Other data will be made available upon request.

Literature Cited

- Almeida-Silva F, Zhao T, Ullrich KK, Schranz ME, Van de Peer Y. Syntenet: an R/bioconductor package for the inference and analysis of synteny networks'. *Bioinformatics*. 2023;39(1):btac806. <https://doi.org/10.1093/bioinformatics/btac806>.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- Álvarez-Carretero S, Kapli P, Yang Z. Beginner's guide on the use of PAML to detect positive selection. *Mol Biol Evol*. 2023;40(4):msad041. <https://doi.org/10.1093/molbev/msad041>.
- Bailey TL, Grant CE. 'SEA: Simple Enrichment Analysis of Motifs'. *bioRxiv* 457422. <https://doi.org/10.1101/2021.08.23.457422>, 24 August 2021, preprint: not peer reviewed.
- 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. <https://doi.org/10.1016/j.ygeno.2022.110390>.
- 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;33(4):4790–4801. <https://doi.org/10.1096/fj.201801989R>.
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses'. *Bioinformatics*. 2009;25(15):1972. <https://doi.org/10.1093/bioinformatics/btp348>.
- Cappabianca L, Farina AR, Di Marcotullio L, Infante P, De Simone D, Sebastiano M, Mackay AR. Discovery, characterization and potential roles of a novel NF-YAx splice variant in human neuroblastoma. *J Exp Clin Cancer Res*. 2019;38(1):482. <https://doi.org/10.1186/s13046-019-1481-8>.
- 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;26(3):301–314. <https://doi.org/10.1101/gr.198473.115>.
- 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;9(6):1351–1373. <https://doi.org/10.1093/gbe/evx081>.
- Chae H-D, Yun J, Bang Y-J, 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;23(23):4084–4088. <https://doi.org/10.1038/sj.onc.1207482>.
- Chaves-Sanjuan A, Gnesutta N, Gobbi A, Martignago D, Bernardini A, Fornara F, Mantovani R, Nardini M. Structural determinants for NF-Y subunit organization and NF-Y/DNA association in plants'. *Plant J*. 2021;105(1):49–61. <https://doi.org/10.1111/tpj.15038>.
- Ciomborowska-Basheer J, Staszak K, Kubiak MR, Makalowska I. Not so dead genes-retrocopies as regulators of their disease-related progenitors and hosts'. *Cells*. 2021;10(4):912. <https://doi.org/10.3390/cells10040912>.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics* (Oxford, England). 2013;29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
- Dolfini D, Imbriano C, Mantovani R. The role(s) of NF-Y in development and differentiation. *Cell Death Differ*. 2025;32(2):195–206. <https://doi.org/10.1038/s41418-024-01388-1>.

- 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;30(11):2450–2459. <https://doi.org/10.1002/stem.1232>.
- Dolfini D, Zambelli F, Pavesi G, Mantovani R. A perspective of promoter architecture from the CCAAT box. *Cell Cycle (Georgetown, Tex.)*. 2009;8(24):4127–4137. <https://doi.org/10.4161/cc.8.24.10240>.
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–1797. <https://doi.org/10.1093/nar/gkh340>.
- Gallo A, Dolfini D, Bernardini A, Gnesutta N, Mantovani R. NF-YA Isoforms with alternative splicing of exon-5 in aves'. *Genomics*. 2023;115(5):110694. <https://doi.org/10.1016/j.ygeno.2023.110694>.
- 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):114–119. <https://doi.org/10.4161/trns.25002>.
- 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;111(11):4139–4144. <https://doi.org/10.1073/pnas.1321569111>.
- Guo J, Grow EJ, Yi C, Mlcochova H, Maher GJ, Lindskog C, Murphy PJ, Wike CL, Carrell DT, Goriely A, 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;21(4):533–546.e6. <https://doi.org/10.1016/j.stem.2017.09.003>.
- 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;11(1):4654. <https://doi.org/10.1038/s41467-020-18508-3>.
- Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 2018;35(2):518–522. <https://doi.org/10.1093/molbev/msx281>.
- Hortschansky P, Haas H, Huber EM, Groll M, Brakhage AA. The CCAAT-binding Complex (CBC) in *Aspergillus* Species'. *Biochim Biophys Acta Gene Regul Mech*. 2017;1860(5):560–570. <https://doi.org/10.1016/j.bbagr.2016.11.008>.
- Huber EM, Daniel HS, Hortschansky P, Groll M, Brakhage AA. DNA Minor groove sensing and widening by the CCAAT-binding Complex'. *Structure*. 2012;20(10):1757–1768. <https://doi.org/10.1016/j.str.2012.07.012>.
- Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 2016;33(6):1635–1638. <https://doi.org/10.1093/molbev/msw046>.
- 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;112(48):14918–14923. <https://doi.org/10.1073/pnas.1507669112>.
- Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, Winkler S, Jermin LS, Skirmuntt EC, Katourakis A, et al. Six reference-quality genomes reveal evolution of bat adaptations'. *Nature*. 2020;583(7817):578–584. <https://doi.org/10.1038/s41586-020-2486-3>.
- Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. DNA-binding specificities of human transcription factors'. *Cell*. 2013;152(1-2):327–339. <https://doi.org/10.1016/j.cell.2012.12.009>.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, van Dam S, Brawand D, Marques PJ, et al. Insights into the evolution of longevity from the bowhead whale genome. *Cell Rep*. 2015;10(1):112–122. <https://doi.org/10.1016/j.celrep.2014.12.008>.
- Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res*. 2002;12(4):656–664. <https://doi.org/10.1101/gr.229202>.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996–1006. <https://doi.org/10.1101/gr.229102>.
- Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. TimeTree 5: an expanded resource for species divergence times'. *Mol Biol Evol*. 2022;39(8):msac174. <https://doi.org/10.1093/molbev/msac174>.
- Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. The human transcription factors'. *Cell*. 2018;172(4):650–665. <https://doi.org/10.1016/j.cell.2018.01.029>.
- Li J, Shen S, Chen J, Liu W, Li X, Zhu Q, Wang B, Chen X, Wu L, Wang M, et al. Accurate annotation of accessible chromatin in mouse and human primordial germ cells'. *Cell Res*. 2018;28(11):1077–1089. <https://doi.org/10.1038/s41422-018-0096-5>.
- 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;267(13):8984–8990. [https://doi.org/10.1016/S0021-9258\(19\)50377-5](https://doi.org/10.1016/S0021-9258(19)50377-5).
- Liu D, Hunt M, Tsai J. Inferring synteny between genome assemblies: a systematic evaluation. *BMC Bioinformatics*. 2018;19(1):26. <https://doi.org/10.1186/s12859-018-2026-4>.
- Liu L, Leng L, Liu C, Lu C, Yuan Y, Wu L, Gong F, Zhang S, Wei X, Wang M, et al. An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos'. *Nat Commun*. 2019;10(1):364. <https://doi.org/10.1038/s41467-018-08244-0>.
- Lu F, Liu Y, Inoue A, Suzuki T, Zhao K, Zhang Y. Establishing chromatin regulatory landscape during mouse preimplantation development. *Cell*. 2016;165(6):1375–1388. <https://doi.org/10.1016/j.cell.2016.05.050>.
- Maezawa S, Yukawa M, Alavattam KG, Barski A, Namekawa SH. Dynamic reorganization of open chromatin underlies diverse transcriptomes during spermatogenesis'. *Nucleic Acids Res*. 2018;46(2):593–608. <https://doi.org/10.1093/nar/gkx1052>.
- Mantovani R, Li XY, Pessara U, Hooft van Huijsduijnen R, Benoist C, Mathis D. Dominant negative analogs of NF-YA. *J Biol Chem*. 1994;269(32):20340–20346. [https://doi.org/10.1016/S0021-9258\(17\)31997-X](https://doi.org/10.1016/S0021-9258(17)31997-X).
- Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, Becker A, Bennett R, Berry A, Bhai J, et al. Ensembl 2023. *Nucleic Acids Res*. 2023;51(D1):D933–D941. <https://doi.org/10.1093/nar/gkac958>.
- Meyer A, Van de Peer Y. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*. 2005;27(9):937–945. <https://doi.org/10.1002/bies.20293>.
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;37(5):1530–1534. <https://doi.org/10.1093/molbev/msaa015>.
- Mirdita M, Schütze K, Moriaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods*. 2022;19(6):679–682. <https://doi.org/10.1038/s41592-022-01488-1>.
- Nardini M, Gnesutta N, Donati G, Gatta R, Forni C, Fossati A, Vonrhein C, Moras D, Romier C, Bolognesi M, et al. Sequence-specific

- transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. *Cell*. 2013;152(1-2):132–143. <https://doi.org/10.1016/j.cell.2012.11.047>.
- 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;1860(5):571–580. <https://doi.org/10.1016/j.bbagr.2016.09.006>.
- Oldfield AJ, Henriques T, Kumar D, Burkholder AB, Cinghu S, Paulet D, Bennett BD, Yang P, Scruggs BS, Lavender CA, et al. NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region. *Nat Commun*. 2019;10(1):3072–3072. <https://doi.org/10.1038/s41467-019-10905-7>.
- Oldfield AJ, Yang P, Conway AE, Cinghu S, Freudenberg JM, Yellaboina S, Jothi R. Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors. *Mol Cell*. 2014;55(5):708–722. <https://doi.org/10.1016/j.molcel.2014.07.005>.
- Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci*. 2021;30(1):70–82. <https://doi.org/10.1002/pro.3943>.
- Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*. 2010;465(7301):1033. <https://doi.org/10.1038/nature09144>.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24–26. <https://doi.org/10.1038/nbt.1754>.
- Ronzio M, Bernardini A, Pavesi G, Mantovani R, Dolfini D. On the NF-Y regulome as in ENCODE (2019). *PLoS Comput Biol*. 2020;16(12):e1008488. <https://doi.org/10.1371/journal.pcbi.1008488>.
- Rosikiewicz W, Kabza M, Kosinski JG, Ciomborowska-Basheer J, Kubiak MR, Makalowska I. RetrogeneDB—a database of plant and animal retrocopies. *Database*. 2017;2017:bax038. <https://doi.org/10.1093/database/bax038>.
- Silvestre-Roig C, Fernández P, Esteban V, Pello ÓM, Indolfi C, Rodríguez C, Rodríguez-Calvo R, López-Maderuelo MD, Bauriedel G, Hutter R, et al. Inactivation of nuclear factor-Y inhibits vascular smooth muscle cell proliferation and neointima formation. *Arterioscler Thromb Vasc Biol*. 2013;33(5):1036–1045. <https://doi.org/10.1161/ATVBAHA.112.300580>.
- Troshin PV, Procter JB, Barton GJ. Java bioinformatics analysis web services for multiple sequence alignment-JABAWS:MSA. *Bioinformatics*. 2011;27(14):2001–2002. <https://doi.org/10.1093/bioinformatics/btr304>.
- Troskie R-L, Faulkner GJ, Cheetham SW. Processed pseudogenes: a substrate for evolutionary innovation. *BioEssays*. 2021;43(11):2100186. <https://doi.org/10.1002/bies.202100186>.
- 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;93(10):1580–1582. <https://doi.org/10.3324/haematol.12425>.
- Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25(9):1189–1191. <https://doi.org/10.1093/bioinformatics/btp033>.
- Wingender E, Schoeps T, Haubrock M, Krull M, Dönitz J. TFCClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res*. 2018;46(D1):D343–D347. <https://doi.org/10.1093/nar/gkx987>.
- Xing Y, Fikes JD, Guarente L. Mutations in yeast HAP2/HAP3 define a hybrid CCAAT box binding domain. *EMBO J*. 1993;12(12):4647–4655. <https://doi.org/10.1002/j.1460-2075.1993.tb06153.x>.
- 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;91(8):3009–3013. <https://doi.org/10.1073/pnas.91.8.3009>.
- Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 2002;19(6):908–917. <https://doi.org/10.1093/oxfordjournals.molbev.a004148>.
- Yang Z, Nielsen R, Goldman N, Pedersen A-MK. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. 2000;155(1):431–449. <https://doi.org/10.1093/genetics/155.1.431>.

Associate editor: Esther Betran