Title

Cell type–specific functions of nucleic acid-binding proteins revealed by deep learning on co-expression networks

Naoki Osato [1*]

[1] School of Life Science and Technology, Institute of Science Tokyo, Tokyo, Japan.

E-mail: naokiosato11@gmail.com

*Corresponding author.

Abstract

Nucleic acid-binding proteins (NABPs) exhibit cell type–specific regulatory functions, yet their target genes and roles remain incompletely defined due to limitations in current experimental methods. Here, we present a deep learning approach that integrates gene co-expression correlations to predict NABP regulatory targets and infer their functions across diverse cellular contexts, without requiring binding motif information. By replacing low-informative input features in a transcriptome-based prediction model with co-expression-derived interactions, we achieved improved accuracy in gene expression prediction. Functional enrichment analysis revealed biologically coherent annotations aligned with known NABP roles. In parallel, we employed a ChatGPT-assisted interpretation framework that provided functional predictions based on semantic patterns among gene sets—an approach that may offer interpretability even when gene sets are modest in size. Predictions supported by both statistical enrichment and ChatGPT-based inference were less likely to reflect spurious associations, reinforcing their biological relevance. Together, this integrative framework combines deep learning, co-expression networks, and generative AI to uncover both established and previously uncharacterized NABP functions in a cell type–specific manner.

Keywords

Computational biology, Deep learning, Gene expression, Gene regulation, Protein function prediction

Introduction

Understanding the cell type–specific regulatory functions of nucleic acid-binding proteins (NABPs), including both DNA-binding proteins (DBPs) and RNA-binding proteins (RBPs), remains a fundamental challenge in molecular biology [1-10]. Although experimental assays such as ChIP-seq and ChIA-PET can map DBP binding sites, they require high-quality antibodies and are not easily scalable across large sets of proteins or conditions. The challenge is even greater for RBPs, many of which lack well-defined binding motifs due to their reliance on intrinsically disordered regions (IDRs) for RNA recognition, rather than sequence-specific contacts [11-13]. Low-affinity DBP binding, which plays key roles in enhancer function and paralog specificity, is also difficult to interpret, especially in the context of phase-separated transcriptional condensates [14]. Conventional methods like eQTL mapping or DNase-seq require prior knowledge of binding motifs and struggle to resolve such indirect or context-dependent interactions [15-17]. While massively parallel reporter assays (MPRAs) and large language model–based architectures such as Enformer have improved our ability to relate sequence to function [18,19], they do not reveal which NABPs mediate the predicted regulatory effects. Thus, new strategies are needed to infer NABP–target relationships and their functional roles, especially in cases where binding motifs are unknown or direct binding evidence is lacking.

Deep learning has substantially improved the prediction of gene expression by modeling regulatory inputs from NABPs, including transcription factors and microRNAs (miRNA) [20]. For example, models incorporating ChIP-seq–derived NABP binding sites in promoter regions and predicted miRNA-binding sites in transcripts have achieved high correlation ($r = 0.82$) between predicted and observed expression levels. In a previous

70 study, we extended this approach by incorporating binding site data for DNA-binding
71 proteins within 1 Mb of transcription start sites (TSSs) in open chromatin, excluding
72 miRNAs and RBPs. To facilitate accurate computation of importance scores (also known
73 as contribution scores) and to streamline model interpretability, we unified the input
74 representation by excluding a previous input layer for RNA-binding proteins and
75 miRNAs [21]. While this modification slightly reduced predictive accuracy (r = 0.70), it
76 enabled broader coverage of distal regulatory elements and laid the foundation for
77 downstream functional analysis based on importance scores. These scores are later
78 leveraged to identify regulatory targets and infer NABP functions.

79 Gene co-expression analysis, a well-established method since the 2000s, offers a
80 complementary perspective by identifying genes with coordinated expression across
81 biological contexts [22-30]. These patterns, derived from transcriptomic datasets across
82 diverse cell types, can capture both direct and indirect regulatory associations—including
83 those involving NABPs with unknown or hard-to-characterize binding features. In this
84 study, we integrated gene co-expression data into a deep learning framework to assess its
85 utility for predicting gene expression and identifying context-specific regulatory targets
86 of NABPs. This combination provides a flexible means to infer functional relationships
87 even when direct binding data are lacking.

88 To test the effectiveness of this approach, we substituted the binding site features
89 of ~300 NABPs that contributed little to previous deep learning predictions with gene co-
90 expression–based interactions. The modified model successfully predicted gene
91 expression levels while assigning DeepLIFT-based importance scores to NABP–gene
92 interactions. Many predicted targets aligned with ChIP-seq or eCLIP-validated binding
93 partners and showed higher contribution scores than background interactions. Functional

enrichment analyses of ten representative NABPs revealed consistent correspondence with known annotations in UniProt [31], and several additional functions were supported by independent publications. To enhance interpretability, we also applied a ChatGPT-based method to infer gene functions and retrieve supporting literature. The results were consistent with enrichment-based findings and highlighted biologically meaningful associations such as circadian regulation (AKAP8) and glycolysis (PKM).

This study provides a computational strategy to uncover the functional roles and regulatory targets of NABPs without requiring experimentally defined binding sites. By integrating gene co-expression profiles with deep learning, the framework captures both direct and indirect interactions often missed by standard methods. This approach enables the systematic and scalable identification of NABP functions, offering a new avenue for exploring transcriptional regulation in a cell type–specific context.

Results

Gene co-expression features improve expression prediction accuracy

Replacing low-contribution DBP inputs with gene co-expression data led to a substantial increase in model performance (**Fig. 1**). In human foreskin fibroblasts (HFFs), substituting 302 out of 1,310 DBPs with co-expression profiles improved the correlation between predicted and actual gene expression from 0.70 to 0.80 (**Fig. 2a, b**). These co-expression inputs included both NABP-encoding and non-NABP mRNAs, drawn from the ENPD (Eukaryotic Nucleic acid Binding Protein) database and related motif studies [32]. A subsequent experiment replaced 325 low-contribution DBPs with co-expression interactions involving RBPs, resulting in a further improvement in prediction

accuracy (r = 0.81; **Fig. 2c**). Collectively, these findings demonstrate that integrating co-expression interactions into the model enhances its ability to capture gene regulatory dynamics beyond direct binding site information.

Contribution scores calculated from co-expression–based NABP–gene interactions showed broader variability and notably higher mean and median values than those of the original feature set (**Fig. 2d**, **Supplementary Table S1**). Correlation analysis further revealed consistent relationships between these contribution scores and $\log_2$-transformed transcript expression levels across cell types (**Supplementary Fig. S2**). Example plots demonstrated that the model adjusted its predictions in a cell type–specific manner, with contribution scores reflecting the strength of inferred NABP regulatory influence **(Fig. 2e**).

Experimental binding data support model-inferred regulatory targets

Experimentally validated binding sites align with deep learning–inferred regulatory targets. Genes with ChIP-seq or eCLIP peaks for a given NABP exhibited disproportionately high or low importance ranks, indicating that experimentally confirmed binding sites correspond to stronger regulatory influence in gene expression prediction models. This relationship supports the biological validity of DeepLIFT-based contribution scores, which quantify the importance of specific NABPs in driving gene expression changes.

To evaluate this association, we analyzed ChIP-seq data for 32 DBPs in HepG2 cells and eCLIP data for 29 RBPs in K562 cells—two cell lines with the most comprehensive publicly available NABP binding site datasets from ENCODE and other studies. A gene was considered associated with an NABP if a ChIP-seq peak was located

142 within ±5 kb of its transcription start site (TSS) or if an eCLIP peak fell within its
143 transcribed region. As illustrated in **Fig. 3a** and **Supplementary Fig. S1**, NABPs with
144 experimentally mapped binding sites consistently showed greater absolute contribution
145 scores across their predicted target genes, reinforcing their regulatory relevance and
146 aligning with prior interpretations of DeepLIFT scores [20,33].

147 Model-derived importance scores reliably predict DBP binding sites, as evidenced
148 by significant overlap with ChIP-seq peaks. To evaluate this correspondence, we
149 analyzed ChIP-seq data for 32 DNA-binding proteins (DBPs) and quantified how
150 frequently high-scoring gene regions—defined by DeepLIFT contribution scores—
151 coincided with experimental binding evidence. Overlap ratios were calculated as the
152 proportion of high-contribution regions that intersected with ChIP-seq peaks, considering
153 both promoter-proximal (±5 kb) and distal (±100 kb) regions around transcription start
154 sites (TSSs). To account for the functional heterogeneity and abundance of ChIP-seq
155 peaks, we applied a precision-at-K approach, focusing on the top-ranked or top-scoring
156 gene regions most likely to reflect regulatory relevance. Four thresholds were used to
157 define high-contribution regions: two based on contribution score rank (top/bottom 20 or
158 50) and two based on absolute score magnitude (≥ 0.01 or ≤ -0.01; ≥ 0.001 or ≤ -0.001).
159 A control analysis using randomly selected gene regions of matched size was performed
160 in parallel. In all tested conditions, the overlap between ChIP-seq peaks and high-
161 contribution regions exceeded that of the random control, supporting the model's ability
162 to prioritize functionally relevant regulatory sites (**Fig. 3b**).

163

164 Deep learning uncovers functionally enriched NABP target genes 165 beyond co-expression networks

166 Model-predicted regulatory targets of nucleic acid-binding proteins (NABPs)
167 display consistent functional enrichment patterns, highlighting the biological specificity
168 of the deep learning approach (**Table 1, 2** and **Supplementary Note**). To evaluate the
169 functional coherence of these predicted targets, we conducted Gene Ontology (GO)
170 enrichment analysis on ten representative NABPs [34,35]. The selected gene sets
171 demonstrated statistically significant enrichment for functionally related GO terms, in
172 contrast to randomly sampled gene sets of equivalent size, which lacked such enrichment
173 (**Fig. 4**). Importantly, comparison with all co-expressed genes (Pearson's $z > 2$) revealed
174 that the model-selected subsets contained enriched annotations not detectable in the
175 broader co-expression profiles. This suggests that the deep learning model filters co-
176 expression signals to identify a biologically focused set of target genes. Functional
177 enrichment was confirmed using the PANTHER Overrepresentation Test across multiple
178 annotation resources, including GO Slim, Reactome, and PANTHER Pathways (**Table**
179 **1**) [36]. To complement these results, we applied a ChatGPT-based gene set interpretation
180 method (**Table 2** and **Supplementary Note**) [37], which inferred functions and retrieved
181 literature consistent with the GO-based findings and UniProt annotations [31]. Notably,
182 AKAP8 was linked not only to its known role in the mitotic cell cycle (UniProt) but also
183 to circadian clock regulation, supported by literature associating AKAP8 with molecular
184 clock genes and melatonin biosynthesis (**Fig. 5** and **6**; **Supplementary Table S4** and
185 **Supplementary Note**) [38]. For this analysis, AKAP8's top 50 predicted targets were
186 defined by ranking all genes according to their model-derived NABP contribution scores,
187 and selecting those for which AKAP8 received the highest score in K562 cells. These
188 results demonstrate that integrating co-expression–guided deep learning with large

language model–based interpretation yields robust insights into both canonical and previously unrecognized biological functions of NABPs.

Deep learning–based refinement of NABP–target predictions enhances the detection of biologically specific functional annotations. To evaluate how training influences functional specificity, we compared GO term enrichment between co-expressed gene sets and predicted regulatory targets for ten NABPs. Although the co-expressed gene sets were larger and exhibited broader enrichment, several functionally informative annotations were uniquely or more significantly enriched among the deep learning–derived target genes (**Table 3**, **Supplementary Table S4** and **S5**). For instance, circadian rhythm–related functions linked to AKAP8 and kinase- or glycolysis-related annotations associated with PKM emerged only after model training (**Table 3** and **Supplementary Table S4**) [39-42]. For PKM, we defined 332 predicted targets by ranking all genes according to NABP contribution scores and selecting those for which PKM ranked within the top two contributors. These results suggest that the model captures biologically meaningful signals beyond mere expression correlations. Importantly, both PKM and AKAP8 are RNA-binding proteins whose functional predictions were derived from a deep learning model trained on gene expression data from K562 cells—a human erythroleukemia cell line widely used in leukemia research. This contextual alignment lends biological credibility to the inferred functions. ALDOA, a glycolysis-associated partner of PKM, and AKAP8, linked to circadian gene networks, have both been implicated in leukemia metabolism and clock regulation, underscoring the relevance of these functions to K562 biology [38,43-46]. Furthermore, the ChatGPT-based functional inference retrieved consistent biological themes—“circadian clock” for AKAP8 and

212 “glycolysis” for PKM—reinforcing the deep learning–based predictions through
213 independent, literature-derived validation (**Fig. 5** and **6**) [37].

214 In neuronal cells, however, AKAP8 exhibited minimal predicted regulatory activity,
215 in contrast to its circadian-linked role in cancer. The model identified very few AKAP8
216 targets in the neuronal context, which limited downstream functional annotation by both
217 PANTHER and ChatGPT-based interpretation. This was based on a contribution score–
218 based ranking across all genes, where only four genes had AKAP8 ranked as the top
219 contributor, and two genes had AKAP8 ranked among the six lowest-scoring NABPs.
220 These findings suggest that AKAP8 may lack broad transcriptional influence in neurons,
221 reinforcing its potential role as a context-dependent regulator.

222 PKM exhibits distinct regulatory profiles in cancer and neuronal cells, reflecting
223 divergent metabolic adaptations. In K562 cancer cells, PKM is associated with a
224 Warburg-like metabolic reprogramming, as indicated by the significant enrichment of
225 aldolase activity among its predicted targets, notably including ALDOC and TALDO1
226 (**Supplementary Table S4**) [47,48]. TALDO1, a key enzyme in the non-oxidative branch of
227 the pentose phosphate pathway, contributes to nucleotide biosynthesis and redox balance,
228 supporting the proliferative and survival needs of cancer cells. This metabolic shift aligns
229 with the activation of the hypoxia-inducible factor (HIF) pathway, involving HIF1A,
230 VEGFA, BNIP3, and DDIT4, which collectively enhance glycolysis, angiogenesis, and
231 adaptation to hypoxic conditions (**Supplementary Note**) [49,50]. In contrast, in neuronal
232 cells, ALDOC—but not TALDO1 or ALDOA—was among the predicted PKM-
233 regulated genes, and aldolase activity was not significantly enriched (**Supplementary**
234 **Table S4**) [51]. For this analysis, 326 predicted targets were selected based on NABP
235 contribution scores, where PKM ranked among the top two contributors per gene. This

236 suggests that neuronal PKM (predominantly PKM1) maintains a stable glycolytic flux to
237 support basal ATP production, rather than engaging in metabolic reprogramming [52,53].
238 Furthermore, although several HIF-responsive genes (e.g., EGLN1, BNIP3, MIR210HG)
239 were among the predicted targets, HIF1A itself was not, indicating a context-dependent
240 modulation of the HIF axis in neurons. Collectively, these findings suggest that PKM is
241 repurposed across cell types, integrating core metabolic programs with distinct
242 transcriptional partners tailored to tissue-specific functional demands.

243 Finally, the integration of deep learning-based co-expression modeling with AI-
244 assisted literature mining enabled functional predictions even for less-characterized
245 regulators such as AKAP8. This framework facilitated both hypothesis generation and
246 validation across cell types, and highlights the utility of scalable, interpretable tools for
247 deciphering context-specific gene regulatory logic.

248

249

250 Discussion

251 These findings address a long-standing challenge in molecular biology: the
252 systematic functional characterization of nucleic acid-binding proteins (NABPs) across
253 diverse cellular contexts. While traditional methods rely heavily on experimental binding
254 data and known motifs, our approach leverages gene co-expression and machine learning
255 to circumvent these limitations. By capturing both direct and indirect regulatory signals,
256 it expands the scope of functional prediction for NABPs that lack conventional binding
257 site information.

258 Gene expression correlations alone can serve as a powerful basis for uncovering
259 the regulatory roles of nucleic acid-binding proteins (NABPs), even in the absence of

experimental binding data. Leveraging transcriptomic profiles from diverse public RNA-seq datasets, our deep learning framework inferred regulatory targets of NABPs based solely on co-expression patterns. Unlike binding site assays such as ChIP-seq or eCLIP—which are technically demanding and limited by the availability of high-quality antibodies—this computational approach is scalable and applicable across NABPs with poorly characterized binding motifs. Functional enrichment analysis of the predicted target genes revealed coherent biological roles, many of which aligned with existing annotations in public databases. Additionally, several novel functional associations—absent from UniProt or GO but supported by published literature—were uncovered. These findings were further corroborated using a ChatGPT-assisted literature mining method, which retrieved consistent biological themes and references. Together, these results underscore the biological relevance of NABP–target predictions based on gene expression correlations and demonstrate the value of integrating deep learning with AI-assisted interpretation. This strategy offers a scalable path toward elucidating NABP functions and discovering new regulatory mechanisms across diverse cellular contexts.

In cancer cells, AKAP8 appears to serve as a molecular integrator of transcriptional regulation under environmental stress. Our analyses, combined with recent reports, suggest that AKAP8 may support hypoxic adaptation through its interaction with DDX5, a multifunctional RNA helicase known to promote HIF-1 transcriptional activity (**Fig. 6**) [54]. DDX5 enhances the recruitment of HIF-1α/β heterodimers to hypoxia-responsive elements (HREs), thus facilitating the expression of genes involved in glycolysis, angiogenesis, and cell survival under low oxygen conditions [55]. Notably, AKAP8 has been shown to physically interact with DDX5 in lung carcinoma cells, forming a chromatin-associated complex that modulates R-loop stability and transcriptional

284 dynamics [56]. Given that DDX5 levels decrease under hypoxia, yet their re-expression
285 restores aberrant R-loop accumulation, AKAP8 may function as a stabilizing co-factor
286 that compensates for transcriptional stress in hypoxic tumor microenvironments. In
287 addition to hypoxia-associated functions, AKAP8 is implicated in circadian rhythm
288 regulation. In our study, AKAP8 was identified as a potential regulator of circadian genes
289 in K562 cancer cells, and ChatGPT-based inference retrieved functional annotations such
290 as "circadian clock" and "rhythm." Prior studies have linked the disruption of circadian
291 timing systems to tumor progression [57], suggesting that AKAP8 may contribute to
292 temporal dysregulation in cancer. Thus, AKAP8 may play dual roles by coupling
293 chromatin-associated RNA processing to both metabolic and temporal regulatory circuits
294 in malignancies. Taken together, these findings position AKAP8 as a context-dependent
295 regulatory hub that coordinates hypoxia response and circadian control through
296 interaction with DDX5 and possibly other nuclear partners. Targeting such
297 multifunctional scaffolds may offer new strategies for disrupting tumor cell plasticity and
298 transcriptional resilience in the face of environmental stress.

299 A key feature distinguishing cancer cells from non-proliferative cells such as
300 neurons is their ability to reprogram core metabolic pathways to support growth under
301 hypoxic and nutrient-limited conditions. Our analysis supports the presence of a PKM2–
302 HIF-1−ALDOC/TALDO1 regulatory axis in cancer cells (**Fig. 6**) [58-61], which coordinates
303 glycolytic and pentose phosphate pathway gene expression to enable metabolic flexibility.
304 Notably, although ALDOA is a well-established glycolytic enzyme and a known PKM
305 interactor in human and mouse (STRING database), it was not among the PKM co-
306 expressed genes in our K562 transcriptomic data (Pearson's z-score < 2), whereas
307 ALDOC—typically regarded as a brain-specific isoform—emerged as a strongly co-

308 expressed and PKM-regulated gene. This suggests that ALDOC may substitute or
309 complement ALDOA in hematopoietic contexts. In agreement with this, transcriptomic
310 profiles from the Human Protein Atlas indicate that both ALDOA and ALDOC are
311 expressed in K562 cells, raising the possibility that PKM2 may preferentially regulate
312 ALDOC under hypoxia, while ALDOA may operate independently or through non-PKM-
313 mediated control. Functionally, ALDOC contributes to anaerobic glycolysis and has been
314 implicated in chemoresistance and tumor cell survival, while TALDO1, a key enzyme in
315 the non-oxidative branch of the pentose phosphate pathway, facilitates carbon flux
316 recycling and supports NADPH and nucleotide biosynthesis—functions essential for
317 redox balance and proliferation under stress. These roles align with their upregulation in
318 various cancer models under hypoxic conditions. In contrast, neuronal cells
319 predominantly express the PKM1 isoform, which sustains a high-efficiency glycolytic
320 flux coupled to mitochondrial oxidative phosphorylation. In our neuronal cell model,
321 ALDOC—but not ALDOA or TALDO1—was predicted as a PKM-regulated gene, and
322 no significant enrichment for aldolase activity was detected. This underscores the
323 metabolic priorities of neurons, which emphasize consistent ATP supply over
324 biosynthetic plasticity, and exhibit minimal reliance on the HIF-1 pathway. The absence
325 of HIF1A among PKM targets in neurons further supports the notion of a transcriptionally
326 uncoupled hypoxic response in these cells. Together, these findings reveal how PKM
327 isoform usage and gene network context shape distinct metabolic wiring across cell types.
328 In cancer, PKM2 engages in metabolic reprogramming through the selective regulation
329 of genes like ALDOC and TALDO1 to support hypoxia-adaptive and proliferative states.
330 In neurons, PKM1 maintains homeostatic energy production under physiological
331 conditions. The selective engagement of ALDOC in both contexts—but through

divergent regulatory logic—highlights a potential metabolic node of interest, while the lack of ALDOA co-expression with PKM suggests a context-dependent separation of biochemical and transcriptional control. These insights may inform future strategies to exploit tumor-selective metabolic dependencies.

Contribution scores derived from gene expression correlations capture both direct and indirect regulatory interactions, offering insights beyond what ChIP-seq alone can provide. While ChIP-seq detects thousands of DNA-binding events per protein, many of these sites do not contribute directly to gene expression regulation and may reflect multifunctional roles of transcription factors [62,63]. Consequently, ChIP-seq–derived binding maps and DeepLIFT-based predictions show only partial overlap, underscoring both methodological differences and the inherent complexity of regulatory networks. In this study, gene co-expression analysis revealed fewer target genes than ChIP-seq but identified subsets that are likely to be more functionally relevant. Our previous work on insulator-associated factors linked such proteins to enhancer–promoter insulation and chromatin organization, including phase separation and loop formation [21]. That study incorporated eQTL data to anchor binding sites to expression outcomes. Here, by contrast, we focused on NABPs with unknown motifs and excluded motif-based or chromatin accessibility data such as DNase-seq or ATAC-seq, to assess the capacity of gene expression correlations to reveal regulatory roles. Notably, the distribution of contribution scores for co-expression–based inputs included both high and low values, implying that gene expression patterns reflect not only direct binding but also indirect regulation mediated through broader network effects. This highlights the value of contribution scores as indicators of both proximal and distal regulatory influences, even when binding motifs are unknown or experimental binding data are unavailable.

Functional roles of NABPs can be inferred even from small gene sets, provided that the target genes exhibit coherent functional patterns. In this study, ten NABPs showed functional enrichments consistent with known annotations, supporting the validity of the deep learning–based predictions. These enrichments were identified using Fisher's exact test with false discovery rate (FDR) correction via the PANTHER database, which typically requires at least 50 genes for robust statistical significance. For proteins with fewer than 50 predicted targets, enrichment was harder to detect, although functional coherence was still apparent in some cases. Relaxing the ranking thresholds increased the number of target genes and enhanced statistical power. Alternatively, more stringent thresholds could be retained if the contribution scores were strong, offering flexibility depending on analysis goals. Additional features—such as gene expression levels or co-expression correlations—could further refine enrichment analyses, particularly for borderline cases. Notably, enrichment analysis of both high- and low-ranking targets revealed candidate novel functions for specific NABPs, suggesting that genes with large-magnitude contribution scores are biologically informative. Moreover, a complementary ChatGPT-based approach demonstrated utility in interpreting small gene sets. Unlike standard enrichment tools, this method can infer biological roles even when gene counts fall below conventional significance thresholds, provided that functionally related genes are present. This capacity helps overcome statistical limitations and broadens the interpretability of regulatory predictions in a cell type–specific context.

Strategic use of co-expression data enhances functional inference, while indiscriminate replacement can diminish biological relevance. The selective integration of gene co-expression data improved target prediction and functional coherence. In contrast, replacing all input features with co-expression-derived interactions led to a

380 decrease in enriched annotations, and known functional roles were no longer recovered.
381 Excessive replacement of input features with gene co-expression data diminishes the
382 biological relevance of model predictions. When all NABP–target inputs were substituted
383 with co-expression-based interactions, the number of enriched functional annotations
384 declined, and previously identified functions from the UniProt database were no longer
385 recovered. This suggests that indiscriminate replacement may cause the model to capture
386 spurious or biologically irrelevant associations. A similar trend was observed in a prior
387 control analysis, where substituting true regulatory targets with randomly selected genes
388 led to a gradual loss of enrichment as randomness increased. Based on these observations,
389 the normalized count of enriched functional terms—adjusted by the number of predicted
390 targets—may serve as a useful proxy for evaluating the accuracy and biological validity
391 of NABP–target predictions [64]. These observations also raise intriguing questions about
392 how the deep learning model prioritizes informative input features during training. Future
393 work exploring the model's internal representation dynamics—such as how it
394 distinguishes signal from noise when presented with diverse regulatory interactions—
395 may provide valuable insights into the mechanisms of transcriptional regulation and
396 model interpretability. Investigating how the model converges on biologically
397 meaningful NABP–target relationships during training—possibly through attention
398 mechanisms or layer-wise relevance propagation—could deepen our understanding of
399 both gene regulation and model interpretability.

400 Cell type specificity and cooperative binding dynamics may underlie missing
401 regulatory signals for certain NABPs. While some NABPs showed no high- or low-
402 ranking target genes in HepG2 or K562 cells, this absence does not rule out regulatory
403 relevance. These proteins may exhibit cell type–specific regulatory activity detectable

404 only in other cellular contexts, where functional enrichment analysis could reveal distinct
405 biological roles. Moreover, many transcription factors and DNA-binding proteins act in
406 concert, forming multi-protein complexes that bind cooperatively to shared promoters [65-
407 67]. In such cases, individual components may not always yield strong contribution scores
408 due to distributed regulatory influence across the complex. This highlights the potential
409 of co-localization analysis as a complementary strategy to uncover cooperative regulatory
410 mechanisms and to disentangle the functional interplay of NABPs within transcriptional
411 networks.

412 This study establishes a scalable computational framework for uncovering the
413 functional roles of nucleic acid-binding proteins (NABPs) using gene expression
414 correlations and deep learning. By leveraging co-expression–based interactions and
415 contribution scoring, the model successfully identified both canonical and novel
416 regulatory associations across cell types—even in the absence of experimentally defined
417 binding motifs. The results highlight how expression-driven inference, complemented by
418 AI-assisted functional interpretation, can extend our understanding of NABP biology
419 beyond conventional binding site–centered approaches. Although the present analysis
420 focused on a representative subset of proteins and cell types, the framework is readily
421 adaptable to broader transcriptomic datasets and larger NABP cohorts. As such, it offers
422 a path forward for systematic functional annotation of NABPs, particularly those with
423 poorly characterized regulatory roles. Together, these findings lay the foundation for
424 future studies to decode complex gene regulatory architectures and integrate
425 transcriptomic and AI-driven predictions into functional genomics research at scale.
426
427

428 # Methods

429 ## Predictive modeling of gene expression across diverse cell types using
430 deep learning and NABP binding information

431 To predict gene expression across multiple human cell types, we adapted a
432 previously established deep learning framework that integrates DNA-binding site data for
433 transcription factors and other regulatory proteins. In this implementation, regulatory
434 elements located both proximally and distally (up to ±1 Mb) from transcription start sites
435 (TSSs) were included as input features (**Fig. 1**) [20,21]. To minimize the effects of cellular
436 heterogeneity in bulk tissue, RNA-seq datasets were selected at the individual cell type
437 level. Publicly available RNA-seq data were retrieved for four functionally distinct cell
438 types: human foreskin fibroblasts (HFFs), mammary epithelial cells (HMECs), neural
439 progenitor cells (NPCs), and either HepG2 hepatoma cells or K562 leukemia cells,
440 depending on the class of regulatory protein analyzed [68,69]. Gene expression values
441 (FPKM) were computed using the RSeQC toolkit based on GENCODE v19 annotations,
442 and subsequently $\log_2$-transformed after adding a small offset to avoid undefined values.
443 Following the approach described by Tasaki et al. [20,21], expression fold changes were
444 calculated relative to the median expression level across the four cell types. For each
445 model iteration, three common cell types (HFFs, HMECs, and NPCs) were used, and
446 either HepG2 (for DNA-binding protein analysis) or K562 (for RNA-binding protein
447 analysis) was included to match the availability of ChIP-seq and eCLIP validation
448 datasets. These cell types were chosen to represent diverse biological lineages and
449 regulatory contexts.

450

451 ## Extended regulatory features and DeepLIFT-compatible model design

452 To capture distal regulatory signals beyond promoter regions, we integrated DNA-
453 binding site data from putative enhancers into a deep learning framework for gene
454 expression prediction. Building upon our earlier work [20,21], this model extends the
455 architecture originally proposed by Tasaki et al. (2020) [20,21], which focused on promoter-
456 proximal regions and included RNA-binding protein and miRNA inputs. In our previous
457 study, aiming to identify insulator-associated DNA-binding proteins, we intentionally
458 excluded RNA-binding proteins and miRNAs to focus on DNA-centric regulatory
459 elements. Instead, we combined DNA-binding sites from both promoter (−2 kb to +1 kb)
460 and enhancer-like regions (±1 Mb from the transcription start site [TSS]) into a single
461 input matrix, enabling DeepLIFT analysis under the high-accuracy "genomics-default"
462 mode.

463 Binding site coordinates were obtained from the GTRD database (v20.06) [70], and
464 regulatory associations were inferred using cis-eQTLs from GTEx v8. A ChIP-seq peak
465 overlapping a 100-bp window centered on an expression-associated SNP was considered
466 linked to the corresponding transcript. Trans-eQTLs were excluded, and coordinates were
467 converted to the hg19 assembly using UCSC LiftOver [71]. To limit feature dimensionality
468 while maximizing genomic coverage, we computed peak lengths for each DNA-binding
469 protein (DBP) in 100-bp bins across ±1 Mb around the TSS, then selected the maximum
470 peak value across each 10-kb window, yielding up to 200 enhancer bins. These were
471 concatenated with 30 promoter-region bins to form a 230-bin unified input layer. This
472 single-layer structure was necessary because DeepLIFT's genomics-default
473 configuration ("RevealCancel rule on dense layers and Rescale rule on conv layers") does
474 not support models with multiple input layers. Concatenating all DBP-related features
475 into one matrix thus preserved compatibility while enhancing interpretability.

476 Model training followed the protocol described by Tasaki et al. and reused in our
477 prior work [20,21]. Specifically, 2,705 genes from chromosome 1 were used for testing,
478 while 22,251 genes from other chromosomes were randomly assigned to training and
479 2,472 to validation. Models were implemented in TensorFlow 2.6.0 with training on an
480 NVIDIA RTX 3080 GPU, typically completing within 2–4 hours. Feature importance
481 was assessed using DeepLIFT, which assigned contribution scores to input features
482 reflecting their impact on gene expression prediction.
483

484 Predicting NABP regulatory targets using gene co-expression–informed 485 substitutions

486 To uncover regulatory targets of NABPs without relying on direct binding site
487 assays, we implemented a co-expression–driven modification of the deep learning input.
488 Specifically, DNA-binding proteins with low contribution scores in the original model
489 were replaced with NABPs whose regulatory interactions were inferred from gene co-
490 expression data. Of the 1,310 DBPs initially included, 302 were substituted with other
491 DBPs, and in a separate RNA-binding protein analysis, 325 were replaced with RNA-
492 binding proteins (RBPs) identified from the ENPD database [32]. Putative NABP–target
493 relationships were established using co-expression profiles provided by COXPRESdb [22],
494 where a Pearson's *z*-score > 2 between a NABP-encoding gene and a non-NABP gene
495 indicated a strong co-expression link. These co-expression pairs served as proxies for
496 regulatory interactions and were used to generate new input features for the model. To
497 integrate these proxies into the deep learning framework, hypothetical NABP binding
498 sites were assigned at the transcription start site (TSS) and 2 kb upstream of each co-
499 expressed gene. The model was then retrained using this modified input, and DeepLIFT

500 was applied to calculate importance scores, evaluating the influence of each NABP–gene
501 interaction on predicted expression levels. This approach enables the functional
502 characterization of NABPs with unknown or uncharacterized binding motifs by
503 leveraging gene co-expression as a proxy for regulatory targeting.
504

505 Prioritization of NABP regulatory targets based on importance score 506 rankings

507 To identify the primary regulatory targets of each NABP, genes were ranked
508 based on the importance scores assigned to NABP binding sites by the deep learning
509 model. Genes with binding sites consistently ranked in the top five or bottom five across
510 all NABPs were selected as putative targets, reflecting high positive or negative
511 regulatory influence. These targets represent genes most strongly associated with a given
512 NABP's regulatory activity. To ensure sufficient statistical power for downstream
513 enrichment analysis, rank thresholds were adjusted for each NABP depending on the
514 number of high-confidence predictions. While a cutoff of five was sufficient in most cases,
515 thresholds were relaxed to include up to eight genes for certain proteins such as RPLP0,
516 which initially yielded a smaller target set.

517 Functional enrichment and AI-based interpretation confirm the biological
518 relevance of NABP regulatory targets. To evaluate the biological coherence of predicted
519 NABP regulatory targets, we performed functional enrichment analysis using the
520 Statistical Overrepresentation Test from the PANTHER database alongside custom in-
521 house scripts [36]. Gene Ontology (GO) terms from biological process, molecular function,
522 and cellular component categories were analyzed for enrichment across three gene sets:
523 (1) primary regulatory target genes identified by the model, (2) all co-expressed genes

524 (Pearson's z-score > 2), and (3) randomly selected genes of matched size. Enrichment
525 was assessed using Fisher's exact test with Benjamini–Hochberg false discovery rate
526 (FDR) correction, and FDR < 0.05 was considered statistically significant. For this
527 analysis, human gene symbols were mapped to Ensembl Gene IDs using the HGNC
528 database (https://www.genenames.org/), and GO annotations were obtained from the
529 Gene Ontology resource [34,35]. Target gene IDs served as the test set, while the full
530 expression dataset used in the deep learning model formed the reference background. To
531 complement the statistical analysis, we applied a ChatGPT-based functional inference
532 method [37]. This approach used curated prompts—detailed in *Extended Data Tables 1 and*
533 *5*—to predict functional roles based on gene sets, and retrieved relevant literature using
534 a modified version of the "4.Reference search and validation.ipynb" script [37]. This script
535 is publicly available on Code Ocean and GitHub
536 (https://doi.org/10.24433/CO.7045777.v1 ;
537 https://github.com/idekerlab/llm_evaluation_for_gene_set_interpretation). The
538 ChatGPT-based results provided consistent and interpretable biological themes aligned
539 with the enrichment findings.
540

541 Validation of predicted NABP targets using experimentally determined
542 binding

543 To assess the biological relevance of the predicted regulatory targets, we
544 compared deep learning–inferred NABP–gene associations with experimentally
545 validated binding data. Specifically, we evaluated the spatial overlap between predicted
546 target genes and ChIP-seq peaks for 32 DNA-binding proteins (within ±5 kb of
547 transcription start sites [TSSs]) as well as eCLIP peaks for 29 RNA-binding proteins

548 (within the transcribed region from TSS to transcription end site [TES]). Gene structure
549 annotations were derived from the GENCODE v19 release [72], and binding site data were
550 obtained from the ENCODE portal and prior publications (**Supplementary Table S6**) [73].
551 This overlap analysis provides orthogonal validation for the predicted targets and
552 supports the utility of gene co-expression–based modeling for inferring NABP–target
553 relationships.
554

555 ## Statistical analysis

556 Functional enrichment was assessed using Fisher's exact test, with multiple testing
557 correction applied via the Benjamini–Hochberg false discovery rate (FDR) procedure.
558 Statistical significance was defined as FDR < 0.05. Distributions of DeepLIFT
559 contribution scores for nucleic acid-binding sites derived from gene co-expression data
560 and ChIP-seq data were compared using the Mann–Whitney U test (also known as the
561 Wilcoxon rank-sum test). A *P*-value < 0.05 was considered statistically significant.
562

563 ## Data availability

564 RNA-seq data for human foreskin fibroblast (HFF) cells were obtained from the Gene
565 Expression Omnibus (GEO) under accession number GSM941744. RNA-seq data for
566 human mammary epithelial cells (HMEC) were retrieved from ENCODE via the UCSC
567 Genome Browser (http://genome.ucsc.edu/encode/) under the file name
568 *wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam*. RNA-seq data for human neural
569 progenitor cells (NPC) were obtained from GEO under accession number GSM915326.
570 RNA-seq data for HepG2 and K562 cells were obtained from the ENCODE portal under
571 accession numbers ENCFF670LIE and ENCFF584WML, respectively. mRNA co-

expression data were downloaded from COXPRESdb (https://zenodo.org/record/6861444). The list of nucleic acid-binding proteins was obtained from the ENPD (Eukaryotic Nucleic acid Binding Protein) database (http://qinlab.sls.cuhk.edu.hk/ENPD/). ChIP-seq data for transcription factors were downloaded from the GTRD database (v20.06 release; https://gtrd.biouml.org). eQTL data (GTEx_Analysis_v8_eQTL.tar) were obtained from the GTEx portal (v8 release; https://gtexportal.org). eCLIP data for RNA-binding proteins were retrieved from the ENCODE portal (ENCSR456FVU) and from Supplementary Data 2 of Van Nostrand et al., *Nature* (2020), https://doi.org/10.1038/s41586-020-2077-3.

Code availability

The deep learning model used in this study is available at https://github.com/reposit2/coexgene.

Figures

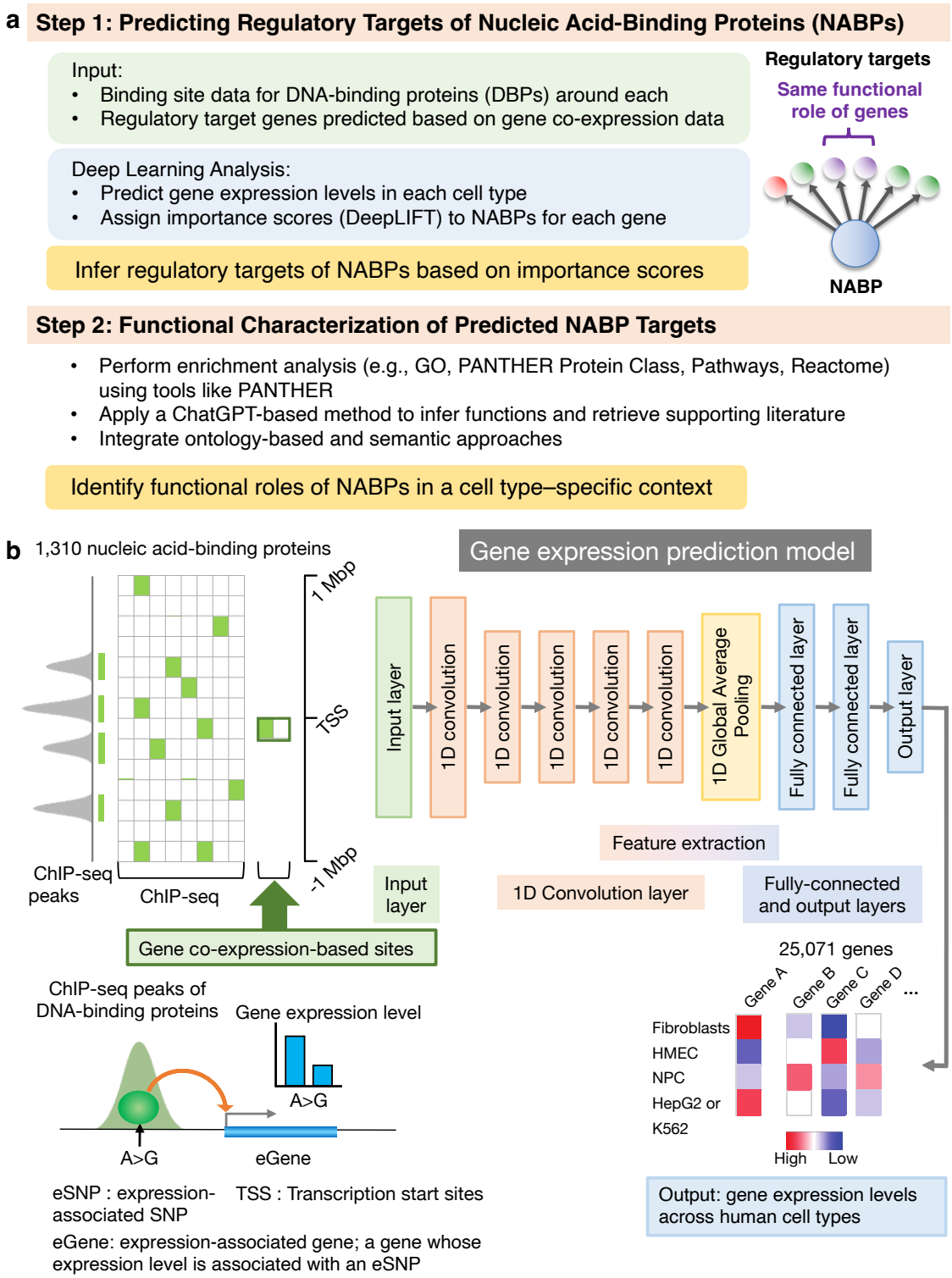


Figure 1. Overview of the deep learning framework for predicting and functionally characterizing NABP regulatory targets.

a. This diagram outlines the overall strategy used to infer the functional roles of nucleic acid-binding proteins (NABPs) from gene expression data. Gene expression profiles from various human cell types were integrated with NABP–gene co-expression relationships derived from COXPRESdb. These inferred interactions were substituted for low-contribution DNA-binding proteins in a deep learning model originally trained on ChIP-seq–based binding site data. The modified model was trained to predict gene expression levels and identify regulatory target genes using DeepLIFT-derived importance scores. Functional annotations of the predicted target genes were assessed using both conventional enrichment analysis and a ChatGPT-based method for literature-guided interpretation.

b. Gene expression levels in human cells were predicted using the genomic locations of DNA-binding sites at promoters and distal elements.

During model training, a DNA-binding protein bound to a distal region overlapping an expression-associated SNP (eSNP) was linked to a potentially regulated gene (eGene) based on eQTL data. For each gene–protein pair, ChIP-seq peak locations were binned (green boxes) and formatted into a matrix as input for deep learning.

In co-expression–based models, binding sites were located near transcription start sites (TSSs). The deep learning model consisted of 1D convolutional layers, global average pooling, and fully connected layers. It predicted gene expression levels across 25,071 genes and multiple human cell types.

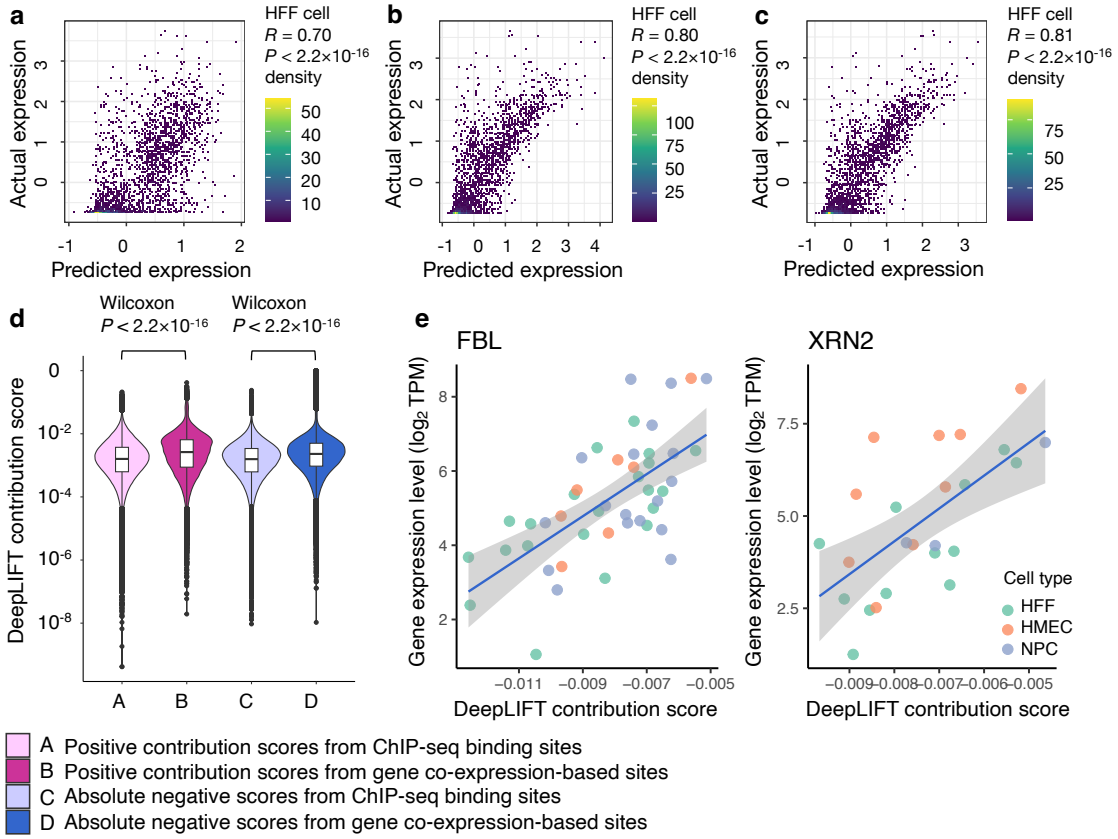


Figure 2. Relationship between gene expression levels and contribution scores. Gene expression levels ($\log_2$ TPM) predicted by a deep learning model were compared with actual expression levels in HFF cells using Spearman's rank correlation.

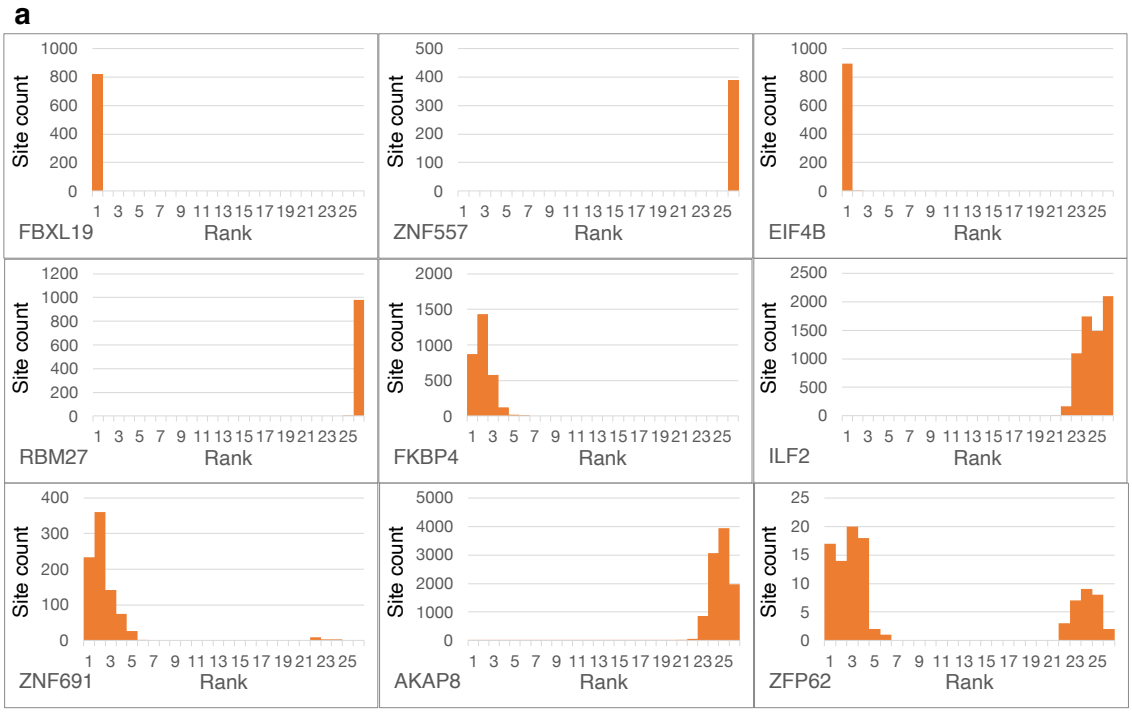
a. Input included DNA-binding sites of 1,310 proteins obtained from ChIP-seq data.

b. Among these, 302 proteins with low contribution scores were replaced with an equal number of DNA-binding proteins selected based on gene co-expression.

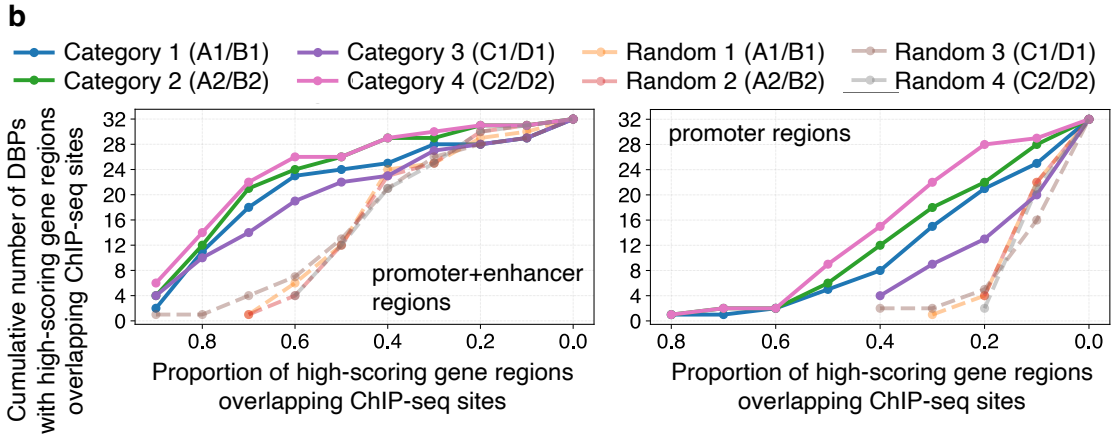
c. Similarly, 325 DNA-binding proteins were replaced with RNA-binding proteins selected via co-expression data.

d. DeepLIFT contribution scores were significantly higher for binding sites derived from gene co-expression data compared to those from ChIP-seq data (Wilcoxon signed-rank test, $P < 2.2 \times 10^{-16}$). A and B show distributions of positive contribution scores, while C and D show absolute values of negative contribution scores.

628 **e.** Representative examples of the correlation between a regulator's DeepLIFT
629 contribution score and the expression levels of its predicted target genes. Linear
630 regression lines with 95% confidence intervals are shown.
631
632



633



634

635 **Figure 3. Comparison of predicted regulatory target genes with experimental ChIP-**
636 **seq and eCLIP binding data.** This figure illustrates how predicted gene-level regulatory

contributions derived from deep learning align with experimentally identified nucleic acid-binding protein targets across genomic regions.

a. Distribution of contribution scores across ChIP-seq–identified binding sites. The x-axis indicates the rank of contribution scores assigned to nucleic acid-binding sites used to predict gene expression levels. Scores were ranked from 1 (highest positive contribution, left) to 1,310 (lowest negative contribution, right) and binned into 50-rank intervals. Rank 1 corresponds to ranks 1–50, rank 2 to ranks 51–100, and rank 26 to ranks 1,251–1,310. nucleic acid-binding sites were categorized as being located in the promoters or transcribed regions of genes experimentally bound by the nucleic acid-binding protein shown in each graph. These sites showed higher or lower contribution scores than those across all ChIP-seq binding sites. The y-axis indicates the number of sites within each 50-rank bin. Positive contribution scores are indicative of gene activation, whereas negative scores reflect gene repression.

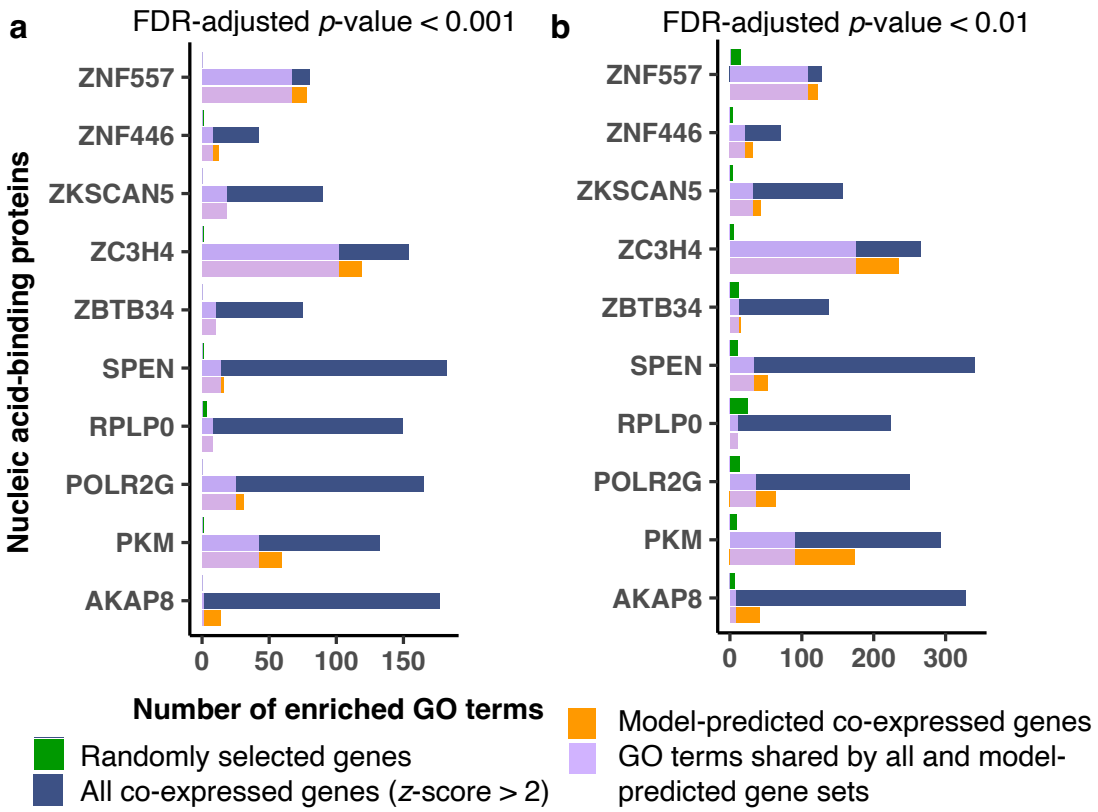
b. Proportion of high-contribution gene regions overlapping ChIP-seq binding sites in promoter+enhancer and promoter regions, relative to all high-contribution regions. Labels indicating "promoter+enhancer regions" and "promoter regions" are shown above the corresponding panels. The proportion represents the number of high-contribution promoter+enhancer regions (within ±100 kb of the transcription start site [TSS]; left panel) and promoter regions (within ±5 kb of the TSS; right panel) that overlap with ChIP-seq binding sites, divided by the total number of high-contribution regions in each category. To define high- and low-contribution regions, multiple thresholds were applied to the 1,310 DNA-binding proteins analyzed: **Rank-based thresholds** — (A1 and B1): rank ≥ 20 or ≤ 1,290; (A2 and B2): rank ≥ 50 or ≤ 1,260. **Score-based thresholds**

660 — (C1 and D1): contribution score ≥ 0.01 or ≤ -0.01; (C2 and D2): contribution score ≥

661 0.001 or ≤ -0.001 (Supplementary Note and Table S2 and S3).

662

663



664

665 **Figure 4. Overrepresented functional annotation terms among target genes of**

666 **nucleic acid-binding proteins.**

667 **a.** and **b.** show results filtered at FDR-adjusted *p*-value < 0.001 and < 0.01, respectively.

668 Functional enrichment was performed on three gene sets: model-predicted co-expressed

669 genes, all co-expressed genes with a Pearson correlation *z*-score > 2, and randomly

670 selected genes of equal size. Gene Ontology (GO) terms from biological process,

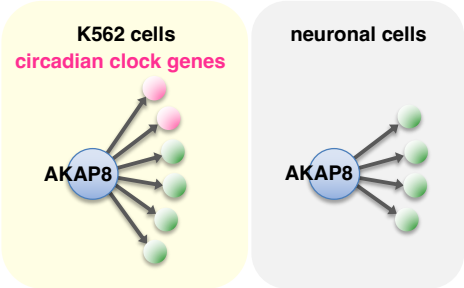
671 molecular function, and cellular component categories were combined. Fisher's exact test

672 *p*-values were FDR-adjusted.

AKAP8 (K562 cells)

1. Circadian Control of Metabolic Regulation

This system includes multiple components of the circadian clock (*PER2, PER3, CRY2, NR1D1*), which regulate daily oscillations of gene expression in metabolic tissues. These transcriptional regulators control the rhythmic expression of key metabolic enzymes and transporters, thereby orchestrating the timing of energy storage, mobilization, and hormonal signaling across the light-dark cycle. Their co-presence with nutrient transporters and signaling proteins suggests functional integration in time-dependent metabolic homeostasis.



PKM (K562 cells)

1. Hypoxia Sensing and Adaptive Response

This system prominently features components regulated by or regulating **hypoxia-inducible factors (HIFs).** *HIF1A*, *EGLN1*, *EGLN3*, *MIR210HG*, *VEGFA*, *BNIP3*, and *DDIT4* are core mediators of the transcriptional and metabolic response to low oxygen conditions. *EGLNs* hydroxylate HIF1A under normoxia, while *HIF1A* activates genes that enhance glycolytic flux, angiogenesis (*VEGFA*), and survival during hypoxic stress (*BNIP3*, *MIR210HG*). *DDIT4* connects hypoxia to mTOR inhibition, reducing energy expenditure. These factors establish a functional axis adapting metabolism and signaling to oxygen availability.

2. Metabolic Reprogramming Toward Glycolysis

Glycolytic genes including *PFKP*, *PFKL*, *ENO1*, *ENO2*, *ENO3*, *ALDOC*, *LDHA*, and *HK2* are key effectors of the hypoxia-induced metabolic shift. HIF1A upregulates these enzymes, favoring glycolysis and lactate production over mitochondrial respiration. This shift is supported by regulatory components like *PDK1*, *PDK3*, and *SLC16A3* (lactate transporter), which further suppress pyruvate entry into mitochondria and facilitate acid-base balance under anaerobic metabolism.

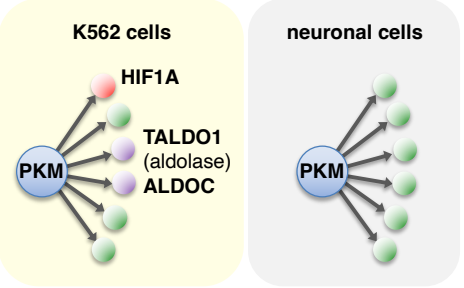


Figure 5. Functional mapping of nucleic acid-binding proteins reveals context-dependent regulatory roles in cancer and neuronal cells.

A deep learning model trained on transcriptomic co-expression data was used to predict the regulatory influence of nucleic acid-binding proteins (NABPs) on gene expression. Contribution scores were computed to quantify the impact of each NABP across target genes, and top-ranking genes were designated as predicted regulatory targets. Functional interpretation was performed using a large language model–based inference method, which identified key biological pathways associated with each NABP. The resulting schematic highlights the distinct functional roles of AKAP8 and PKM in K562 leukemia cells—circadian regulation and metabolic reprogramming under hypoxia, respectively—compared with their more limited or divergent activities in neuronal cells.

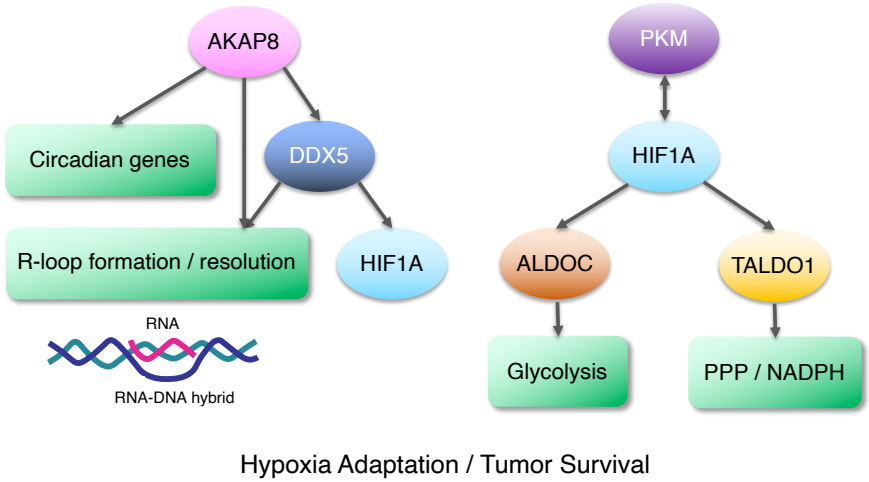


Figure 6. Integrated hypoxia-responsive regulatory network orchestrated by AKAP8 and PKM in cancer cells. This model illustrates how two nucleic acid-binding proteins (NABPs)—AKAP8 and PKM—regulate distinct but converging transcriptional ›programs to support tumor adaptation under hypoxic stress. AKAP8 acts through DDX5 to influence R-loop resolution [74], circadian gene expression, and stabilization of HIF1A transcriptional activity. In parallel, PKM modulates HIF1A-driven metabolic reprogramming, promoting glycolysis via ALDOC and redox homeostasis via TALDO1. Predicted regulatory links were derived from contribution scores computed by a deep learning framework and are supported by published functional studies in cancer models.

Tables

Table 1. Functional roles inferred for nucleic acid-binding proteins: integration of UniProt annotations, ChatGPT-based predictions, and confidence scores.

Functional annotations derived from UniProt entries were compared with significantly enriched biological processes in the model-predicted regulatory targets, identified via the PANTHER Overrepresentation Test (see Supplementary Table S4). Shared or overlapping functions between each nucleic acid-binding protein (NABP) and its predicted targets are shown. Functional roles inferred using a ChatGPT-based interpretation method are highlighted in yellow, with parentheses indicating confidence scores (proportion of annotated genes within the gene set). Confidence is categorized as high (0.87–1.00), medium (0.82–0.86), low (0.01–0.81), or none (0). For reference, see Hu M. et al., *Nat. Methods*, 2025 (https://doi.org/10.1038/s41592-024-02525-x) [37].

AKAP8
negative regulation of RNA biosynthetic process
Circadian clock system
Cry-Per complex
Circadian-regulated energy metabolism and insulin signaling (0.74)

PKM
kinase
glycolytic process
ATP metabolic process
Hypoxia-adaptive metabolic remodeling and cell migration (0.79)

POLR2G
RNA polymerase II transcription regulatory region sequence-specific DNA binding
general transcription factor
translational protein
Ribosome biogenesis and RNA maturation at the interface of translation and nuclear export (0.91)

RPLP0
large ribosomal subunit
ribonucleoprotein complex
cytosolic large ribosomal subunit
Mitochondrial-linked ribosome function and RNA-guided surveillance in translational homeostasis (0.84)

SPEN
mRNA binding
chromatin binding
transcription coregulator activity
Chromatin-associated transcriptional regulation and nuclear RNA export integration (0.83)

ZBTB34
C2H2 zinc finger transcription factor
DNA-binding transcription activator activity, RNA polymerase II-specific
RNA polymerase II transcription regulatory region sequence-specific DNA binding
Transcriptional regulation by C2H2 zinc finger proteins and signaling modulation (0.72)

ZC3H4
chromatin/chromatin-binding, or -regulatory protein
transcription regulator activity
metal ion binding
RNA-binding protein–mediated transcriptional integration and pre-mRNA processing (0.89)

ZKSCAN5
DNA-binding transcription factor activity
C2H2 zinc finger transcription factor
RNA polymerase II cis-regulatory region sequence-specific DNA binding
metal ion binding
RNA processing, ribonucleoprotein assembly, and transcription-coupled quality control (0.83)

ZNF446
C2H2 zinc finger transcription factor
extracellular space
Transcription-coupled RNA processing and mitochondrial RNA metabolism (0.81)

ZNF557
DNA-binding transcription activator activity, RNA polymerase II-specific
RNA polymerase II transcription regulatory region sequence-specific DNA binding
ion binding, metal ion binding
Transcriptional regulation and RNA metabolic coordination via zinc finger proteins and RNA-associated complexes (0.88)

Table 2. Functional characterization of predicted regulatory targets of nucleic acid-binding proteins using a ChatGPT-based inference framework.

717 Representative biological functions were inferred for each NABP by applying a large
718 language model–assisted gene set interpretation method. Bold labels summarize the
719 predominant functional theme for each NABP, while numbered subcategories denote
720 distinct functional roles associated with individual gene subsets.
721

AKAP8 (K562 cells)
Circadian-regulated energy metabolism and insulin signaling (0.74)
1. Circadian Control of Metabolic Regulation
2. Insulin Signaling and Glucose Transport
3. Fatty Acid and Lipid Homeostasis
4. Metabolite Transport and Mitochondrial Function
5. Signal Integration and Vesicle Trafficking
6. Developmental and Structural Modulators

PKM (K562 cells)
Hypoxia-adaptive metabolic remodeling and cell migration (0.79)
1. Hypoxia Sensing and Adaptive Response
2. Metabolic Reprogramming Toward Glycolysis
3. Extracellular Matrix Remodeling and Migration
4. Vesicular and Membrane Trafficking
5. Mitochondrial Function and ROS Homeostasis
6. Transcriptional and Epigenetic Regulation

PKM (neuronal cells)
Hypoxia-driven metabolic and structural reprogramming (0.87)
1.Hypoxia Sensing and Transcriptional Response
2.Glycolytic Shift and Mitochondrial Modulation
3.Cytoskeletal Remodeling and Motility Enhancement
4.Endocytosis, Vesicle Transport, and Membrane Dynamics
5.RNA Processing and Translation Control
6.Ubiquitination and Proteostasis
7.Stress Signaling and Nuclear Regulation

POLR2G (K562 cells)
Ribosome biogenesis and RNA maturation at the interface of translation and nuclear export (0.91)
1. Cytoplasmic and Nuclear Ribosomal Component Assembly
2. Ribosomal RNA Processing and Modification
3. Translational Initiation and Ribosome Recycling
4. Proteasome-Linked Ribosome Quality Control
5. Nuclear Export, mRNP Surveillance, and RNA Chaperoning

RPLP0 (K562 cells)
Mitochondrial-linked ribosome function and RNA-guided surveillance in translational homeostasis (0.84)
1. Ribosomal Structural Components and Translational Core
2. RNA Modification and Surveillance by Small Non-Coding RNAs
3. RNA Surveillance and Nuclear Export Quality Control
4. Mitochondrial and Redox-Linked Translational Interface
5. Membrane Trafficking and Endoplasmic Reticulum Integration
6. Zinc Finger Regulators and DNA Repair Coupling

SPEN (K562 cells)
Chromatin-associated transcriptional regulation and nuclear RNA export integration (0.83)
1. Transcriptional Regulation and Chromatin Remodeling
2. RNA Binding, Processing, and Export
3. Protein Quality Control and Ubiquitin-Mediated Regulation
4. Signal-Responsive Transcription Integration
5. RNA Surveillance and Stability
722 6. Nuclear Structural and Trafficking Components

ZBTB34 (HepG2 cells)
Transcriptional regulation by C2H2 zinc finger proteins and signaling modulation (0.72)
1. Transcriptional Regulation by Zinc Finger Proteins
2. Signaling and Transcriptional Crosstalk
3. Modulators of Chromatin and RNA Stability
4. Functional Integration and Cell Fate Control

ZC3H4 (HepG2 cells)
RNA-binding protein–mediated transcriptional integration and pre-mRNA processing (0.89)
1. RNA-Binding Protein Network and Pre-mRNA Processing
2. Transcription-Coupled Regulatory Scaffold
3. Spliceosome Assembly and Dynamic Coordination
4. mRNA Export and Quality Control
5. Epigenetic and Histone Regulatory Interface
6. RNA Modification and Stability Regulation
7. Nucleolar and Ribosome Biogenesis Linkage

ZKSCAN5 (HepG2 cells)
RNA processing, ribonucleoprotein assembly, and transcription-coupled quality control (0.83)
1. RNA Processing and mRNA Surveillance
2. Transcriptional Regulation and Chromatin Remodeling
3. Ribosome Biogenesis and Mitochondrial Translation
4. Protein Quality Control and Ubiquitin-Dependent Regulation
5. Nucleocytoplasmic Transport and Organelle Coordination
6. ZNF and ZBTB Family-Mediated Regulatory Modules

ZNF446 (HepG2 cells)
Transcription-coupled RNA processing and mitochondrial RNA metabolism (0.81)
1. Transcriptional Regulation and Initiation
2. Splicing and RNA Processing
3. Mitochondrial RNA Metabolism and Translation
4. RNA Transport and Nuclear Export
5. Ubiquitin-Proteasome and RNA-Protein Complex Quality Control
6. Chromatin-Linked Transcription Modulators

ZNF557 (HepG2 cells)
Transcriptional regulation and RNA metabolic coordination via zinc finger proteins and RNA-associated complexes (0.88)
1. Transcriptional Regulation by Zinc Finger Proteins
2. General Transcription Machinery and RNA Polymerase Assembly
3. RNA Splicing, Processing, and Surveillance
4. rRNA Processing and Ribosome Biogenesis
5. Epigenetic and Chromatin-Level Regulation
6. Coupling with RNA Modification and Translation Control

723

724

725 **Table 3. Functional enrichments among predicted regulatory target genes of nucleic**

726 **acid-binding proteins.**

727 Functional enrichment analysis was conducted using the PANTHER database. This table

728 lists annotations that were statistically enriched in regulatory target genes predicted by a

729 deep-learning model trained on gene co-expression data. These annotations were not

730 enriched in the full set of co-expressed genes (Pearson's z-score > 2) that included the

731 nucleic acid-binding protein itself. 'Reference' refers to the total number of human genes.

732 'Analyzed' denotes the number of regulatory target genes examined. 'Expected' is the

733 number of genes expected to have the annotation by chance. 'Over/Under' indicates

734 whether the annotation is overrepresented or underrepresented. 'Fold Enrichment' is the

735 observed-to-expected ratio. 'Raw P-value' is the result of Fisher's exact test. 'FDR'

736 indicates the Benjamini–Hochberg false discovery rate.

AKAP8							
PANTHER Pathways							
PANTHER Pathways	Reference (17130)	Analyzed (49)	Expected	Over/ Under	Fold Enrichment	raw P-value	FDR
Circadian clock system (P00015)	10	3	0.03	+	> 100	2.60E-06	4.16E-04
PANTHER GO-Slim Biological Process							
PANTHER GO-Slim Biological Process	Reference (17130)	Analyzed (49)	Expected	Over/ Under	Fold Enrichment	raw P-value	FDR
circadian regulation of gene expression (GO:0032922)	16	3	0.05	+	65.55	1.20E-05	1.27E-02
response to light stimulus (GO:0009416)	21	3	0.06	+	49.94	2.82E-05	1.99E-02
circadian rhythm (GO:0007623)	21	3	0.06	+	49.94	2.82E-05	1.49E-02
rhythmic process (GO:0048511)	24	3	0.07	+	43.7	4.27E-05	1.81E-02
GO cellular component complete							
GO cellular component complete	Reference (17130)	Analyzed (49)	Expected	Over/ Under	Fold Enrichment	raw P-value	FDR
Cry-Per complex (GO:1990512)	3	2	0.01	+	> 100	2.40E-05	4.77E-02
PKM							
PANTHER Protein Class							
PANTHER Protein Class	Reference (17130)	Analyzed (326)	Expected	Over/ Under	Fold Enrichment	raw P-value	FDR
aldolase (PC00044)	5	2	0.1	+	21.02	3.48E-03	5.68E-02
kinase (PC00137)	134	9	2.55	+	3.53	1.07E-03	2.62E-02
PANTHER GO-Slim Molecular Function							
PANTHER GO-Slim Molecular Function	Reference (17130)	Analyzed (326)	Expected	Over/ Under	Fold Enrichment	raw P-value	FDR
enzyme inhibitor activity (GO:0004857)	107	8	2.04	+	3.93	1.01E-03	4.68E-02
kinase activity (GO:0016301)	562	23	10.7	+	2.15	7.08E-04	3.57E-02
PANTHER GO-Slim Biological Process							
PANTHER GO-Slim Biological Process	Reference (17130)	Analyzed (326)	Expected	Over/ Under	Fold Enrichment	raw P-value	FDR
endoplasmic reticulum tubular network organization (GO:0071786)	12	3	0.23	+	13.14	1.32E-03	3.68E-02
regulation of cellular component biogenesis (GO:0044087)	139	10	2.65	+	3.78	3.31E-04	1.05E-02

Acknowledgements

Supercomputing resources were provided by the Human Genome Center, the Institute of Medical Science, the University of Tokyo, and the ROIS National Institute of Genetics.

Author contributions

N.O. designed and conducted the research, and wrote and revised the manuscript.

Supplementary data

Supplementary data are available at

Conflict of interests

None declared.

Funding

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI [16K00387], a start-up research grant from Osaka University, the KIOXIA Young Scientist Research Grant at Waseda University, and the Grant for Basic Science Research Projects from the Sumitomo Foundation. All grants were awarded to N.O.

760 **Supplementary Note and Tables**

761

762 **Gene co-expression–based predictions align with ChIP-seq–defined**
763 **DNA-binding targets**

764 High-ranking predicted binding regions frequently coincide with ChIP-seq peaks,
765 particularly in promoter–enhancer intervals. Deep learning–derived DNA-binding site
766 predictions based on gene co-expression data demonstrate substantial concordance with
767 ChIP-seq data, particularly when broader genomic regions encompassing promoters and
768 putative enhancers are considered. When overlap was assessed within ±100 kb of
769 transcription start sites (TSSs)—a window supported by prior enhancer–gene mapping
770 studies [19,75]—the mean and median overlap ratios reached 0.69 and 0.76, respectively
771 (Ratio A2 / B2; Supplementary Tables S2 and S3). In contrast, overlap was lower for
772 promoter regions alone (±5 kb), with corresponding mean and median values of 0.33 and
773 0.34. Some DNA-binding proteins, including FBXL19, POGK, SPEN, THAP7, ZNF142,
774 and ZNF800, showed notably high promoter-region overlap ratios (>0.5), indicating
775 potential direct regulatory activity. Nevertheless, not all ChIP-seq peaks necessarily
776 reflect functionally relevant transcriptional binding events. Consequently, model-inferred
777 high-scoring sites do not universally coincide with experimentally determined binding
778 regions, underscoring the complexity and context dependence of regulatory interactions.

779

780

781 **Context-specific predictions reveal known and emerging functional**
782 **roles of NABPs**

783 The deep learning–based predictions successfully recapitulate both established and
784 potentially novel biological functions of NABPs. For example, AKAP8—previously
785 annotated in UniProt with canonical cell cycle functions (e.g., "GO:0000278 mitotic cell
786 cycle" and "GO:0044839 cell cycle G2/M phase transition")—was also predicted to be
787 associated with circadian clock regulation in this study. Supporting literature retrieved
788 from PubMed indicates a potential role for AKAP8 in circadian rhythm control: one study
789 identified AKAP8 as highly co-expressed with known clock genes like *CRY1*, while
790 another reported its involvement in melatonin biosynthesis, a process disrupted in
791 individuals with autism spectrum disorders [76,77].

792 PKM (pyruvate kinase M) was functionally linked to glycolytic regulation,
793 consistent with literature-based protein–protein interaction data, including its
794 experimentally validated interaction with ALDOA in both human and mouse systems [39,40].
795 However, co-expression analysis in our dataset revealed a notable context-specific
796 pattern: ALDOA and ALDOB were not significantly co-expressed with PKM (Pearson's
797 z-score < 2), whereas ALDOC emerged as a strongly co-expressed gene prior to deep
798 learning model training. This result suggests that ALDOC may serve as a context-
799 dependent effector of PKM-regulated glycolysis, particularly in hematopoietic settings
800 such as K562 cells. Indeed, ALDOC was not only among the predicted regulatory targets
801 of PKM, but also appeared in gene sets annotated with hypoxia-responsive functions
802 involving HIF1A. Moreover, transcriptomic data from the Human Protein Atlas
803 confirmed that both ALDOA and ALDOC are expressed in K562 cells, reinforcing the
804 possibility that PKM may selectively regulate ALDOC in specific cellular contexts
805 despite the known biochemical role of ALDOA in glycolysis. These findings exemplify

806 how deep learning–guided predictions can refine our understanding of tissue-specific
807 metabolic wiring and highlight functional divergence within enzyme families.
808
809

810 Supplementary Table S1. Co-expression–based selection enhances the 811 importance scores of NABP binding sites.

812 Binding sites identified through gene co-expression exhibited higher absolute
813 DeepLIFT contribution scores compared to those selected solely from ChIP-seq data
814 without co-expression information. This trend was observed for both DNA-binding
815 proteins in HepG2 cells and RNA-binding proteins in K562 cells. For each group, the
816 table reports the mean and median of positive and negative contribution scores. These
817 findings suggest that gene co-expression–based selection prioritizes functionally
818 impactful regulatory interactions.

DNA-binding proteins (HepG2 cell)			
Co-expression–based		Non-co-expression–based	
	Plus score		Plus score
Mean	0.005673	Mean	0.003391
Median	0.003829	Median	0.001893
	Minus score		Minus score
Mean	-0.017000	Mean	-0.002502
Median	-0.002286	Median	-0.001583

RNA-binding proteins (K562 cell)			
Co-expression–based		Non-co-expression–based	
	Plus score		Plus score
Mean	0.004916	Mean	0.002962
Median	0.002422	Median	0.001689
	Minus score		Minus score
Mean	-0.011987	Mean	-0.002485
Median	-0.001521	Median	-0.001414

819
820
821

822 Supplementary Table S2. Enrichment of high-contribution promoter 823 regions with experimentally validated DBP binding sites. (See 824 Supplementary Excel file)

825 This table summarizes the proportion of promoter regions (±5 kb from transcription
826 start sites) that were both predicted to be functionally important by DeepLIFT

827 contribution scores and experimentally supported by ChIP-seq data for the same DNA-
828 binding proteins (DBPs). Overlap ratios were calculated across four thresholds—two
829 based on rank and two on absolute score—to assess the robustness of the predictions.
830 To classify high and low contributions, multiple thresholds were applied to the 1,310
831 DNA-binding proteins analyzed:
832 **Rank-based thresholds**:
833 (A1 and B1) proteins ranked in the top or bottom 20 (i.e., rank ≤ 20 or ≥ 1,290)
834 (A2 and B2) proteins ranked in the top or bottom 50 (i.e., rank ≤ 50 or ≥ 1,260)
835 **Score-based thresholds**:
836 (C1 and D1) contribution score ≥ 0.01 or ≤ –0.01
837 (C2 and D2) contribution score ≥ 0.001 or ≤ –0.001
838

839 **Supplementary Table S3. Enrichment of high-contribution promoter**
840 **and enhancer regions with ChIP-seq–validated DBP binding sites.**

841 This table presents the overlap ratios between predicted high-contribution regions
842 (defined within ±100 kb of transcription start sites, including distal enhancers) and ChIP-
843 seq binding sites for corresponding DBPs. High concordance between model-based
844 predictions and ChIP-seq peaks across multiple thresholds highlights the reliability of the
845 contribution scoring method in identifying functionally relevant regulatory sites.
846 To classify high and low contributions, the following thresholds were applied to
847 the 1,310 DNA-binding proteins analyzed:
848 **Rank-based thresholds**:
849 (A1 and B1) proteins ranked in the top or bottom 20 (i.e., rank ≤ 20 or ≥ 1,290)
850 (A2 and B2) proteins ranked in the top or bottom 50 (i.e., rank ≤ 50 or ≥ 1,260)

Score-based thresholds:

(C1 and D1) contribution score ≥ 0.01 or ≤ –0.01

(C2 and D2) contribution score ≥ 0.001 or ≤ –0.001

References

1 Belz, G. T. & Nutt, S. L. Transcriptional programming of the dendritic cell network. *Nature reviews. Immunology* **12**, 101-113, doi:10.1038/nri3149 (2012).

2 Miller, J. C. *et al.* Deciphering the transcriptional network of the dendritic cell lineage. *Nature immunology* **13**, 888-899, doi:10.1038/ni.2370 (2012).

3 Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744-752, doi:10.1016/j.cell.2010.01.044 (2010).

4 Gertz, J. *et al.* Distinct properties of cell-type-specific and shared transcription factor binding sites. *Molecular cell* **52**, 25-36, doi:10.1016/j.molcel.2013.08.037 (2013).

5 Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665, doi:10.1016/j.cell.2018.01.029 (2018).

6 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics* **10**, 252-263, doi:10.1038/nrg2538 (2009).

7 Pierson, E. *et al.* Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS Comput Biol* **11**, e1004220, doi:10.1371/journal.pcbi.1004220 (2015).

8 Sonawane, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. *Cell Rep* **21**, 1077-1088, doi:10.1016/j.celrep.2017.10.001 (2017).

9 Saha, A. *et al.* Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome research* **27**, 1843-1858, doi:10.1101/gr.216721.116 (2017).

10 Boyle, A. P. *et al.* Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453-456, doi:10.1038/nature13668 (2014).

11 Ray, D. *et al.* RNA-binding proteins that lack canonical RNA-binding domains are rarely sequence-specific. *Sci Rep* **13**, 5238, doi:10.1038/s41598-023-32245-9 (2023).

12 Zeke, A. *et al.* Deep structural insights into RNA-binding disordered protein regions. *Wiley Interdiscip Rev RNA* **13**, e1714, doi:10.1002/wrna.1714 (2022).

13 Zaharias, S. *et al.* Intrinsically disordered electronegative clusters improve stability and

884 binding specificity of RNA-binding proteins. *J Biol Chem* **297**, 100945,
885 doi:10.1016/j.jbc.2021.100945 (2021).
886 14 Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. & Mann, R. S. Low-Affinity Binding Sites
887 and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu Rev Cell Dev Biol*
888 **35**, 357-379, doi:10.1146/annurev-cellbio-100617-062719 (2019).
889 15 Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks.
890 *Cell* **150**, 1274-1286, doi:10.1016/j.cell.2012.04.040 (2012).
891 16 Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular
892 perturbations across complex diseases. *Nature methods* **13**, 366-370,
893 doi:10.1038/nmeth.3799 (2016).
894 17 Fagny, M. *et al.* Exploring regulation in tissues with eQTL networks. *Proc Natl Acad Sci*
895 *U S A* **114**, E7841-E7850, doi:10.1073/pnas.1707375114 (2017).
896 18 Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory elements.
897 *Nature*, doi:10.1038/s41586-024-08430-9 (2025).
898 19 Avsec, Z. *et al.* Effective gene expression prediction from sequence by integrating long-
899 range interactions. *Nature methods* **18**, 1196-1203, doi:10.1038/s41592-021-01252-x
900 (2021).
901 20 Tasaki, S., Gaiteri, C., Mostafavi, S. & Wang, Y. Deep learning decodes the principles of
902 differential gene expression. *Nat Mach Intell* **2**, 376-386, doi:10.1038/s42256-020-0201-
903 6 (2020).
904 21 Osato, N. & Hamada, M. Systematic discovery of directional regulatory motifs associated
905 with human insulator sites. *bioRxiv* (2025).
906 22 Obayashi, T., Kodate, S., Hibara, H., Kagaya, Y. & Kinoshita, K. COXPRESdb v8: an
907 animal gene coexpression database navigating from a global view to detailed investigations.
908 *Nucleic acids research* **51**, D80-D87, doi:10.1093/nar/gkac983 (2023).
909 23 Yang, S. *et al.* COEXPEDIA: exploring biomedical hypotheses via co-expressions
910 associated with medical subject headings (MeSH). *Nucleic acids research* **45**, D389-D396,
911 doi:10.1093/nar/gkw868 (2017).
912 24 Zhu, Q. *et al.* Targeted exploration and analysis of large cross-platform human
913 transcriptomic compendia. *Nature methods* **12**, 211-214, 213 p following 214,
914 doi:10.1038/nmeth.3249 (2015).
915 25 Miller, H. E. & Bishop, A. J. R. Correlation AnalyzeR: functional predictions from gene
916 co-expression correlations. *BMC Bioinformatics* **22**, 206, doi:10.1186/s12859-021-
917 04130-7 (2021).
918 26 Raina, P. *et al.* GeneFriends: gene co-expression databases and tools for humans and
919 model organisms. *Nucleic acids research* **51**, D145-D158, doi:10.1093/nar/gkac1031

(2023).

27 Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)* **302**, 249-255, doi:10.1126/science.1087447 (2003).

28 Lemoine, G. G., Scott-Boyer, M. P., Ambroise, B., Perin, O. & Droit, A. GWENA: gene co-expression networks analysis and extended modules characterization in a single Bioconductor package. *BMC Bioinformatics* **22**, 267, doi:10.1186/s12859-021-04179-4 (2021).

29 Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research* **38**, W214-220, doi:10.1093/nar/gkq537 (2010).

30 Wong, A. K., Krishnan, A. & Troyanskaya, O. G. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic acids research* **46**, W65-W70, doi:10.1093/nar/gky408 (2018).

31 UniProt, C. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic acids research* **53**, D609-D617, doi:10.1093/nar/gkae1010 (2025).

32 Tak Leung, R. W., Jiang, X., Chu, K. H. & Qin, J. ENPD - A Database of Eukaryotic Nucleic Acid Binding Proteins: Linking Gene Regulations to Proteins. *Nucleic acids research* **47**, D322-D329, doi:10.1093/nar/gky1112 (2019).

33 Shrikumar, A., Greenside, P. & Kundaje, A. in *Proceedings of the 34th International Conference on Machine Learning* Vol. 70 (eds Precup Doina & Teh Yee Whye) 3145--3153 (PMLR, Proceedings of Machine Learning Research, 2017).

34 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29, doi:10.1038/75556 (2000).

35 Gene Ontology, C. *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, doi:10.1093/genetics/iyad031 (2023).

36 Thomas, P. D. *et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci* **31**, 8-22, doi:10.1002/pro.4218 (2022).

37 Hu, M. *et al.* Evaluation of large language models for discovery of gene set function. *Nature methods* **22**, 82-91, doi:10.1038/s41592-024-02525-x (2025).

38 Farshadi, E., van der Horst, G. T. J. & Chaves, I. Molecular Links between the Circadian Clock and the Cell Cycle. *J Mol Biol* **432**, 3515-3524, doi:10.1016/j.jmb.2020.04.003 (2020).

39 Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research* **28**, 3442-3444, doi:10.1093/nar/28.18.3442 (2000).

40 Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research* **51**, D638-D646, doi:10.1093/nar/gkac1000 (2023).

41 Tian, W. *et al.* Bioinformatics analysis of the role of aldolase A in tumor prognosis and immunity. *Sci Rep* **12**, 11632, doi:10.1038/s41598-022-15866-4 (2022).

42 Zhang, F. *et al.* Elevated transcriptional levels of aldolase A (ALDOA) associates with cell cycle-related genes in patients with NSCLC and several solid tumors. *BioData Min* **10**, 6, doi:10.1186/s13040-016-0122-4 (2017).

43 Zlacka, J. & Zeman, M. Glycolysis under Circadian Control. *Int J Mol Sci* **22**, doi:10.3390/ijms222413666 (2021).

44 Yang, Y., Pu, J. & Yang, Y. Glycolysis and chemoresistance in acute myeloid leukemia. *Heliyon* **10**, e35721, doi:10.1016/j.heliyon.2024.e35721 (2024).

45 Puram, R. V. *et al.* Core Circadian Clock Genes Regulate Leukemia Stem Cells in AML. *Cell* **165**, 303-316, doi:10.1016/j.cell.2016.03.015 (2016).

46 Lee, Y. Roles of circadian clocks in cancer pathogenesis and treatment. *Exp Mol Med* **53**, 1529-1538, doi:10.1038/s12276-021-00681-0 (2021).

47 Liberti, M. V. & Locasale, J. W. The Warburg Effect: How Does it Benefit Cancer Cells? *Trends Biochem Sci* **41**, 211-218, doi:10.1016/j.tibs.2015.12.001 (2016).

48 Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science (New York, N.Y.)* **324**, 1029-1033, doi:10.1126/science.1160809 (2009).

49 Semenza, G. L. Hypoxia-inducible factors: mediators of cancer progression and targets for cancer therapy. *Trends Pharmacol Sci* **33**, 207-214, doi:10.1016/j.tips.2012.01.005 (2012).

50 Keith, B., Johnson, R. S. & Simon, M. C. HIF1alpha and HIF2alpha: sibling rivalry in hypoxic tumour growth and progression. *Nat Rev Cancer* **12**, 9-22, doi:10.1038/nrc3183 (2011).

51 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science (New York, N.Y.)* **347**, 1260419, doi:10.1126/science.1260419 (2015).

52 Israelsen, W. J. & Vander Heiden, M. G. Pyruvate kinase: Function, regulation and role in cancer. *Semin Cell Dev Biol* **43**, 43-51, doi:10.1016/j.semcdb.2015.08.004 (2015).

53 Bluemlein, K. *et al.* No evidence for a shift in pyruvate kinase PKM1 to PKM2 expression during tumorigenesis. *Oncotarget* **2**, 393-400, doi:10.18632/oncotarget.278 (2011).

54 Wang, X. & Liu, L. AKAP8 interacting with DDX5 to regulate R-loop balance involves in lung carcinoma cell growth. *bioRxiv*, doi:10.1101/2023.11.28.568985 (2024).

55 Shirai, Y., Suwa, T., Kobayashi, M., Koyasu, S. & Harada, H. DDX5 enhances HIF - 1

992 activity by promoting the interaction of HIF - 1α with HIF - 1β and recruiting the
993 resulting heterodimer to its target gene loci. *Biology of the Cell* **116**,
994 doi:10.1111/boc.202300077 (2023).
995 56 Leszczynska, K. B. *et al.* Hypoxia - mediated regulation of DDX5 through decreased
996 chromatin accessibility and post - translational targeting restricts R - loop accumulation.
997 *Molecular Oncology* **17**, 1173-1191, doi:10.1002/1878-0261.13431 (2023).
998 57 Farshadi, E., van der Horst, G. T. J. & Chaves, I. Molecular Links between the Circadian
999 Clock and the Cell Cycle. *Journal of Molecular Biology* **432**, 3515-3524,
1000 doi:10.1016/j.jmb.2020.04.003 (2020).
1001 58 Ran, J. *et al.* Hypoxia regulates glycolysis through the HIF-1α/BMAL1/ALDOC axis to
1002 reduce oxaliplatin sensitivity in colorectal cancer. *Journal of Cancer* **16**, 2503-2515,
1003 doi:10.7150/jca.108582 (2025).
1004 59 De Vitis, C. *et al.* ALDOC- and ENO2- driven glucose metabolism sustains 3D tumor
1005 spheroids growth regardless of nutrient environmental conditions: a multi-omics analysis.
1006 *Journal of Experimental & Clinical Cancer Research* **42**, doi:10.1186/s13046-023-02641-
1007 0 (2023).
1008 60 Minchenko, O. H., Garmash, I. A. & Minchenko, D. O. Inhibition of IRE1 modifies
1009 hypoxic regulation of G6PD, GPI, TKT, TALDO1, PGLS and RPIA genes expression in
1010 U87 glioma cells. *The Ukrainian Biochemical Journal* **89**, 38-49,
1011 doi:10.15407/ubj89.01.038 (2017).
1012 61 Obeidat, N. *et al.* Effects of cyclic acute and chronic hypoxia on the expression levels of
1013 metabolism related genes in a pancreatic cancer cell line. *Biomedical Reports* **17**,
1014 doi:10.3892/br.2022.1564 (2022).
1015 62 Ladomery, M. Multifunctional proteins suggest connections between transcriptional and
1016 post-transcriptional processes. *Bioessays* **19**, 903-909, doi:10.1002/bies.950191010
1017 (1997).
1018 63 Kim, S., Yu, N. K. & Kaang, B. K. CTCF as a multifunctional protein in genome regulation
1019 and gene expression. *Exp Mol Med* **47**, e166, doi:10.1038/emm.2015.33 (2015).
1020 64 Osato, N. Characteristics of functional enrichment and gene expression level of human
1021 putative transcriptional target genes. *BMC Genomics* **19**, 957, doi:10.1186/s12864-017-
1022 4339-5 (2018).
1023 65 Burley, S. K. & Kamada, K. Transcription factor complexes. *Curr Opin Struct Biol* **12**, 225-
1024 230, doi:10.1016/s0959-440x(02)00314-7 (2002).
1025 66 Goos, H. *et al.* Human transcription factor protein interaction networks. *Nat Commun* **13**,
1026 766, doi:10.1038/s41467-022-28341-5 (2022).
1027 67 Berenson, A. *et al.* Paired yeast one-hybrid assays to detect DNA-binding cooperativity

and antagonism across transcription factors. *Nat Commun* **14**, 6570, doi:10.1038/s41467-023-42445-6 (2023).

68 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research* **41**, D991-995, doi:10.1093/nar/gks1193 (2013).

69 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207-210, doi:10.1093/nar/30.1.207 (2002).

70 Kolmykov, S. *et al.* GTRD: an integrated view of transcription regulation. *Nucleic acids research* **49**, D104-D111, doi:10.1093/nar/gkaa1057 (2021).

71 Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic acids research* **34**, D590-598, doi:10.1093/nar/gkj144 (2006).

72 Mudge, J. M. *et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic acids research* **53**, D966-D975, doi:10.1093/nar/gkae1078 (2025).

73 Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711-719, doi:10.1038/s41586-020-2077-3 (2020).

74 Young Cheon, N., Kim, S. & Yil Lee, J. in *Gene Expression* Ch. Chapter 3, (IntechOpen, 2022).

75 Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 1516, doi:10.1016/j.cell.2019.02.027 (2019).

76 Tan, X. *et al.* Novel Insights into the Circadian Rhythms Based on Long Noncoding and Circular RNA Profiling. *Int J Mol Sci* **25**, doi:10.3390/ijms25021161 (2024).

77 Poelmans, G., Franke, B., Pauls, D. L., Glennon, J. C. & Buitelaar, J. K. AKAPs integrate genetic findings for autism spectrum disorders. *Transl Psychiatry* **3**, e270, doi:10.1038/tp.2013.48 (2013).