Title

Inferring the cell type-specific functional roles of nucleic acid-binding proteins using 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

The regulatory target genes and functional roles of nucleic acid-binding proteins (NABPs), including DNA- and RNA-binding proteins, can vary in a cell- or tissue-specific manner. However, experimentally identifying NABP-binding sites and functionally characterizing these proteins remain technically challenging. Deep learning methods have been developed to predict gene expression levels across diverse cell and tissue types using DNA-binding profiles derived from ChIP-seq experiments. In this study, I predicted the regulatory targets and inferred the functional roles of NABPs using gene co-expression data. Gene co-expression correlations, derived from transcriptomic profiles across multiple cell and tissue types, were used to represent potential interactions between NABPs and their associated genes. These interactions were integrated into a deep learning framework by partially replacing the model's original input features, enabling expression-level prediction based on co-expression networks. The functional predictions of NABP target genes were consistent with their known biological roles. Moreover, several additional functional annotations not currently present in gene function databases were supported by evidence from previously published experimental studies. These results highlight the potential of combining gene co-expression data with deep learning to uncover both known and previously uncharacterized functional roles of NABPs in a cell type–specific context.

Keywords

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

Introduction

DNA-binding proteins, including transcription factors, play diverse biological roles and regulate gene expression at multiple levels. These functions differ across the cellular and tissue context [1-5] The expression of transcription factors is known to be cell- and tissue-specific, and regulatory networks have been investigated based on experimental data such as gene expression profiling, ChIP-seq, DNase-seq, and expression quantitative trait loci (eQTL) analysis [6-13]. However, elucidating the functional and regulatory roles of these proteins in a cell- and tissue-specific manner remains a significant challenge. The binding sites of nucleic acid-binding proteins (NABPs), including DNA-binding proteins (DBPs) and RNA-binding proteins (RBPs), can be experimentally identified using techniques such as ChIP-seq and ChIA-PET. These assays require high-quality antibodies against the target proteins, which are not always available. For certain NABPs, identifying binding sequences through experimental methods is impractical due to constraints in time, cost, and labor. Moreover, standard approaches to characterizing the functional roles of NABPs—such as transcriptomics, DNase-seq, and eQTL analyses—typically depend on prior knowledge of their nucleic acid-binding sequences. Thus, direct interactions involving proteins with unknown binding motifs remain difficult to identify. Massively parallel reporter assays (MPRAs) have been employed to measure the regulatory activity of hundreds of thousands of sequences [14]. In addition, large language model (LLM)-based deep learning architectures, such as *Enformer*, can predict genomic sequences that are critical for gene expression regulation [15]. However, it remains challenging to determine which NABPs actually interact with these predicted regulatory sequences.

Several recent studies have leveraged deep learning models to address these challenges and improve gene expression prediction from genomic features. Deep learning approaches have enhanced the accuracy of gene expression prediction based on the binding sites of NABPs, such as transcription factors, and other regulatory factors including microRNAs (miRNAs) [16]. Transcription factors bind to promoter regions near transcription start sites (TSSs) as well as to distal enhancers located upstream or downstream of their target genes. One such method predicts gene expression levels across 53 human tissues by integrating experimentally determined binding sites of NABPs in promoter regions and predicted miRNA-binding sites in transcribed regions. For example, this method achieved a correlation coefficient of 0.82 between predicted and actual gene expression levels. In a previous study, I modified this approach by incorporating binding site data for DNA-binding proteins located within 1 Mbp of TSSs in open chromatin regions—many of which likely correspond to enhancers—while excluding RNA-binding proteins and miRNAs. This modification reduced the correlation coefficient of 0.70 [17].

Analysis of gene co-expression across diverse cell and tissue types has provided insights into regulatory interactions among genes, including those encoding NABPs. Gene expression profiles from RNA-seq experiments and the results of co-expression analyses are publicly available in a variety of databases [18-22]. Numerous studies have utilized gene co-expression network analyses to infer functional relationships between genes [18,22-26]. These correlations may reflect both direct and indirect molecular interactions. Some of the regulatory molecules involved may be difficult to detect experimentally and challenging to predict using existing computational methods. Importantly, these co-expression patterns may reflect regulatory mechanisms that are still not fully captured by current experimental or computational methods.

To evaluate whether gene co-expression can be used to predict functionally meaningful regulatory targets of NABPs, I conducted the following analyses. In this study, I investigated the functional annotations enriched in the predicted target genes of NABPs using gene co-expression data. Specifically, I used gene co-expression profiles to infer interactions between these proteins and other genes. To test their functional significance, I replaced the DNA-binding sites of approximately 300 proteins that had shown low contribution scores─reflecting limited influence on gene expression prediction—in my previous deep learning model with binding sites from other NABPs.

The deep learning model was trained to predict gene expression levels and to assign importance scores to specific interactions between binding sites and genes that contributed to the prediction. As the regulatory targets of a transcription factor often consist of functionally related genes [27], as a result, functional enrichment can be detected among the predicted target genes, even if the prediction is not precise enough to identify direct regulatory relationships. I examined the overlap between the predicted target genes of NABPs and binding sites identified through experimental assays such as ChIP-seq and eCLIP. I found that predicted targets associated with experimentally validated binding sites for a given protein showed higher absolute contribution scores than those linked to other proteins. I then conducted functional enrichment analyses for the predicted target genes of ten NABPs. The enriched functional annotations corresponded to known functions documented in the UniProt database. Additionally, I identified enriched annotations that were not present in UniProt but had been previously described in published studies. Finally, I employed a ChatGPT-based method to infer the functional roles of the predicted target genes. The inferred functions and associated literature

116 retrieved by this method were consistent with the results of the functional enrichment
117 analyses.

118 These analyses establish a computational framework for identifying the functional
119 roles and regulatory targets of NABPs without relying on experimental binding data.
120 They also enable the discovery of indirect regulatory mechanisms that are difficult to
121 detect using conventional methods, thereby reducing experimental costs and facilitating
122 the identification of novel biological functions.

123

124 **Results**

125 **Relationship between gene expression levels and contribution scores**

126 To examine the effect of gene co-expression interactions on gene expression
127 prediction, the input data for 302 out of 1,310 DNA-binding proteins with low
128 contribution scores (DeepLIFT scores) were replaced with co-expression interaction data
129 (**Fig. 1**). These data included mRNAs encoding DNA-binding proteins and other mRNAs
130 not encoding NABPs. The substituted DNA-binding proteins were randomly selected
131 from gene symbols listed in the ENPD (Eukaryotic Nucleic acid Binding Protein)
132 database and other sources used in DNA-binding motif analysis [28]. Notably, in human
133 foreskin fibroblast (HFF) cells, the correlation coefficient between actual and predicted
134 gene expression levels increased from 0.70 to 0.80 when co-expression data were used
135 (**Fig. 2a, b**). Subsequently, for a separate analysis, 325 DNA-binding proteins with low
136 contribution scores were replaced with co-expression interaction data involving genes
137 encoding RNA-binding proteins and other mRNAs. The RNA-binding proteins were also
138 randomly selected from gene symbols in the ENPD database. This substitution further
139 improved the prediction accuracy, with the correlation coefficient increasing from 0.70

140 to 0.81 (**Fig. 2c**). These findings suggest that gene co-expression interactions significantly
141 enhance the prediction of gene expression levels in deep learning models.
142 The contribution scores of input features selected based on gene expression
143 correlations were distributed over a wide range; however, both the mean and median
144 values were higher than those of the original input features (**Fig. 2d and Supplementary**
145 **Table S1**). I examined the distribution of Spearman's correlation coefficients between
146 the contribution scores and gene expression levels ($\log_2$ TPMs; transcripts per million)
147 (**Supplementary Fig. S2**), and plotted examples showing the relationship between the
148 contribution scores of NABPs and the expression levels of their predicted regulatory
149 target genes across different cell types (**Fig. 2e**). These results indicate that, based on co-
150 expression-derived interactions between genes and NABPs, the deep learning model
151 adjusted its gene expression predictions in a cell type–specific manner.
152

153 ## Overlap between high-scoring genes and experimental targets

154 To assess whether experimentally determined binding sites are associated with the
155 regulatory target genes of nucleic acid-binding proteins (NABPs) identified based on
156 contribution scores, I analyzed ChIP-seq and eCLIP data. A gene was considered
157 associated with a protein if a ChIP-seq peak for a DNA-binding protein was located
158 within ±5 kb of its transcription start site (TSS), or if an eCLIP peak for an RNA-binding
159 protein was located within its transcribed region. ChIP-seq data for 32 DNA-binding
160 proteins in HepG2 cells and eCLIP data for 29 RNA-binding proteins in K562 cells were
161 used, as these cancer-derived cell lines provide the most comprehensive NABP binding
162 site data available from public resources, including the ENCODE portal and published
163 studies. As shown in **Fig. 3a** and **Supplementary Fig. S1**, genes with ChIP-seq or eCLIP

164 peaks for a given NABP tended to exhibit either high or low ranks in contribution scores,
165 indicating that these NABPs had relatively larger absolute contribution scores compared
166 to other NABPs. These patterns suggest that the presence of experimentally identified
167 binding sites corresponds to a stronger regulatory influence—either activation or
168 repression—on gene expression, consistent with the interpretation of contribution scores
169 provided by the DeepLIFT framework and previous studies [16 29].

170 To assess whether gene regions with high contribution scores overlap with
171 experimentally determined binding sites, I analyzed ChIP-seq data for 32 DNA-binding
172 proteins. I calculated the overlap ratio as the number of gene regions with both a high
173 contribution score and a ChIP-seq binding site, divided by the total number of gene
174 regions with a high contribution score. Two types of genomic regions were examined:
175 those within ±5 kb of the transcription start site (TSS), and those within ±100 kb of the
176 TSS. High contribution scores were defined based on four threshold criteria: two rank-
177 based thresholds (top or bottom 20; top or bottom 50), and two score-based thresholds (≥
178 0.01 or ≤ -0.01; ≥ 0.001 or ≤ -0.001). As a control, the same analysis was performed using
179 randomly selected gene regions, with the number of regions matched to those containing
180 a ChIP-seq peak. Cumulative overlap ratios were calculated across a range of overlap
181 thresholds and visualized as a plot (**Fig. 3b**). The overlap ratios for gene regions with
182 high contribution scores and ChIP-seq peaks were consistently higher than those for
183 randomly selected gene regions across all four threshold criteria.

184

185 **Functional prediction of target genes of NABPs**

186 Transcriptional target genes of transcription factors often share functionally
187 related annotations. Functional enrichment analysis of gene annotations is useful for

predicting the function of a transcription factor and verifying that predicted target genes are not selected at random. To assess whether the regulatory target genes of NABPs—predicted by deep learning based on gene expression correlations—include functionally related genes, I performed functional enrichment analysis using Gene Ontology (GO) annotations for the target genes of ten such proteins [30,31]. These predicted regulatory targets exhibited significant enrichment for functionally related GO terms. In contrast, when the same number of genes was randomly selected from the entire gene set, GO term enrichment was negligible, indicating that the predicted target genes are functionally coherent rather than randomly distributed (**Fig. 4**). For a more detailed analysis of functional enrichment, I performed statistical overrepresentation tests using the gene list analysis tool provided by PANTHER (https://pantherdb.org/) [32]. This tool incorporates multiple functional annotation databases, including PANTHER Pathways, Protein Class, GO Slim, GO, and Reactome Pathways. The enriched functional annotations of the predicted regulatory target genes for each NABP were compared with those of the same protein retrieved from the UniProt database (e.g., https://www.uniprot.org/uniprotkb/Q9Y2L8/entry) [33]. In addition, I used a ChatGPT-based method to infer the functional roles of the predicted regulatory target genes. This method automatically retrieved literature in which the titles or abstracts included keywords related to the predicted functions [34]. Ten representative NABPs are shown in **Table 1, Supplementary Table S4** and **Supplementary Note**. The predicted functional annotations and ChatGPT-inferred roles for these proteins correspond well with those in the UniProt database and previously published studies. Notably, among the ten examples, AKAP8 is annotated in the UniProt database (https://www.uniprot.org/uniprotkb/O43823/entry) with cell cycle-related functions,

212 such as “GO:0000278 mitotic cell cycle” and “GO:0044839 cell cycle G2/M phase
213 transition.” In the present study, AKAP8 was also implicated in circadian clock regulation.
214 A search of the PubMed database identified two studies supporting the involvement of
215 AKAP8 in circadian rhythms. One study states that “genes such as NFIL3, AKAP8, and
216 SALL3, which are highly correlated with known molecular clock genes (CRY1), can be
217 identified as candidate circadian genes,” while another reports that “AKAP5 and AKAP8
218 are both involved in the biosynthesis of the circadian hormone melatonin, which is
219 markedly disrupted in individuals with ASDs.” [35,36]. The ChatGPT-based method
220 automatically retrieved three publications associated with the “circadian clock” based on
221 the predicted regulatory target genes of AKAP8 (**Fig. 5**) [34]. These findings demonstrate
222 that the methods and analyses employed in this study successfully identified both known
223 and potentially novel functional annotations associated with NABPs.

224 During training, the deep learning model adjusted the contribution (i.e., neural
225 network weights) of NABPs to the prediction of their regulatory target gene expression.
226 To evaluate the effect of learning on functional enrichment, I analyzed all co-expressed
227 genes (Pearson’s z-score > 2) for ten NABPs using PANTHER DB. I then compared the
228 enriched functional annotations of these co-expressed genes with those of the regulatory
229 target genes identified by the deep learning model (**Table 2** and **Supplementary Table**
230 **S5**). Although the number of enriched annotations was higher for the co-expressed gene
231 sets due to their larger size, certain annotations were uniquely or more significantly
232 enriched in the regulatory target genes identified after training. Notably, circadian
233 rhythm–related functional annotations, such as those associated with AKAP8, were
234 detected only after training (**Table 2**; yellow highlights in the "S2_AKAP8" tab of
235 **Supplementary Table S4**). Similarly, for PKM, kinase-related functional annotations

236 emerged only after training (**Table 2** and the "S2_PKM" tab in **Supplementary Table**
237 **S4**). PKM interacts with ALDOA, as shown in the STRING database for both mouse
238 (https://string-db.org/network/10090.ENSMUSP00000034834) and human [37,38].
239 Previous studies support the association between the functional annotation "aldolase
240 (PC00044)" and PKM in glycolysis. For example, one study reported that "Aldolase A
241 (ALDOA) is an enzyme that plays an important role in glycolysis and gluconeogenesis,
242 which is closely related to tumor metabolism." [39,40] These functional annotations were
243 enriched only after deep learning, suggesting their potential involvement in biological
244 processes specific to K562 cells, where the regulatory target genes of the RNA-binding
245 protein were identified. Circadian clock functions and PKM have both been linked to
246 glycolytic regulation in human acute myeloid leukemia (AML) cells [41-44]. The ChatGPT-
247 based method also inferred functional associations related to "circadian clock function"
248 for AKAP8 and "glycolysis" for PKM (**Fig. 5**) [34]. These findings suggest that combining
249 gene co-expression data with deep learning–based gene expression prediction can
250 uncover biologically meaningful functional associations of NABPs in a cell type– or
251 tissue-specific context.

252 Together, these results demonstrate that deep learning models trained on gene co-
253 expression data can identify biologically relevant regulatory targets and functional
254 annotations of NABPs. Notably, the use of a ChatGPT-based method enabled the
255 identification of literature support for both established and potentially novel functions,
256 such as circadian clock regulation for AKAP8 and glycolytic involvement for PKM.
257 These findings highlight the potential of integrating deep learning with AI-assisted
258 literature analysis to uncover functional roles of regulatory proteins in a cell type–specific
259 context.

Discussion

While previous models relied on experimentally determined binding sites, this study demonstrates that gene expression correlations alone can provide valuable insights into nucleic acid-binding protein (NABP) functions. Gene expression correlations were derived from RNA-seq data obtained from public databases encompassing various cell and tissue types. Compared to RNA-seq, experimental identification of nucleic acid-binding sites using assays such as ChIP-seq and eCLIP remains more time-consuming, costly, and labor-intensive, particularly for certain NABPs where binding site determination is technically challenging. This highlights the need for alternative computational approaches, such as the one proposed here. The deep learning analysis based on gene expression correlations successfully predicted regulatory target genes of NABPs. Subsequent functional enrichment analyses, along with ChatGPT-based interpretation, revealed potential functional roles and associations of these NABPs. Many predicted functional annotations aligned with known annotations reported in public databases, while additional annotations—absent from these databases—were supported by evidence from previously published studies. Overall, functionally related annotations were frequently observed among the predicted regulatory targets, highlighting the biological relevance of the predictions. Both functional enrichment analysis and the ChatGPT-based method consistently identified known and potentially novel functional associations of NABPs, underscoring the utility of this integrative approach. This approach can facilitate the discovery of novel regulatory mechanisms across diverse biological contexts.

When comparing DNA-binding sites predicted based on contribution scores for gene expression prediction with those identified by ChIP-seq assays, the two sets of binding sites do not completely overlap. Although ChIP-seq can identify tens of thousands of DNA-binding sites for a given protein, not all of these sites are necessarily involved in direct gene expression regulation. Additionally, some transcription factors are known to participate in multiple cellular functions [45,46]. This discrepancy reflects both methodological differences and the biological complexity of gene regulation. The number of co-expressed genes (Pearson's z-score > 2) for each NABP was generally less than 1,500, which is smaller than the number of genes associated with ChIP-seq binding sites for the same proteins. However, most statistically significant co-expressed genes are likely to be functionally relevant to gene regulation. In our previous analysis of insulator-associated DNA-binding proteins, 97 transcription factors and DNA-binding proteins were identified as potentially involved in insulator functions that modulate enhancer–promoter interactions [17]. Some of these proteins may also play critical roles in chromatin loop formation, transcriptional condensates, and phase separation—mechanisms that regulate gene expression differently from direct transcription factor binding. In that deep learning analysis, we incorporated eQTL data to link DNA-binding sites with changes in gene expression levels. However, in this study, I did not use DNA-binding motif sequences or open chromatin regions identified by DNase-seq or ATAC-seq, as the objective was to investigate DNA-binding proteins with unknown motifs. While our previous work focused on insulator-associated functions, this study broadens the scope to NABPs with unknown binding motifs. Furthermore, associations between eSNPs and eGenes in eQTL data may reflect indirect regulatory interactions. The contribution scores of DNA-binding sites selected based on gene expression correlations exhibited both high

307 and low values, suggesting that indirect interactions between DNA-binding proteins and
308 other genes influence these scores. This highlights the potential of contribution scores to
309 capture not only direct regulatory effects but also complex indirect interactions within
310 gene regulatory networks.

311 In the functional enrichment analysis of regulatory target genes, ten NABPs
312 exhibited functional annotations consistent with those previously reported for the
313 respective proteins. This suggests that the number of predicted target genes was sufficient
314 to detect statistically significant functional associations. Statistical enrichment was
315 assessed using Fisher's exact test with false discovery rate (FDR) correction in
316 PANTHER DB, which generally requires more than 100 regulatory target genes to
317 achieve statistical significance. For some proteins, however, fewer than 100 predicted
318 target genes were identified, limiting the ability to detect enriched functions. Relaxing the
319 ranking threshold when selecting target genes tends to increase the number of genes
320 included in the analysis, which in turn can improve the statistical power for detecting
321 functionally enriched annotations. This approach is particularly useful when aiming to
322 ensure a sufficient gene set size for enrichment analysis, although even more stringent
323 thresholds can be effective when the predicted regulatory relationships are strong. A
324 potential next step is to improve the detection of functional enrichment in smaller gene
325 sets by incorporating additional features, such as gene expression levels and gene co-
326 expression correlations. Furthermore, functional enrichment analysis using high- and
327 low-ranking regulatory target genes revealed novel functions potentially associated with
328 specific NABPs. These findings indicate that regulatory target genes with high absolute
329 contribution scores can facilitate the identification of NABP functional roles in a cell
330 type–specific context. Additionally, unlike traditional functional enrichment analysis, the

ChatGPT-based approach holds promise for predicting functional roles even in smaller gene sets, provided that multiple functionally related genes are present. This capability may enhance the identification of biological functions for NABPs with a limited number of predicted target genes. This approach is particularly advantageous for small gene sets, as ChatGPT-based predictions can infer functional roles when multiple functionally related genes are present, overcoming limitations of conventional enrichment analyses.

When all input data in the deep learning analysis were replaced with interactions between NABPs and other genes based solely on gene expression correlations, the number of functionally enriched annotations among the predicted regulatory target genes decreased, and known functional annotations from the UniProt database were no longer detected. This suggests that a high proportion of input data replacement may lead the model to learn error-prone or biologically irrelevant interactions between NABPs and their target genes. In a previous study, replacing the regulatory target genes of a DNA-binding protein with randomly selected genes similarly resulted in a reduction of enriched functional annotations as the number of random genes increased. Therefore, the number of enriched functional annotations, when normalized to the total number of predicted target genes, can serve as an indicator of the accuracy and biological relevance of the predicted regulatory interactions [47].

Some nucleic acid-binding proteins (NABPs) did not exhibit high- or low-ranking target genes in HepG2 cells; however, these proteins may display such rankings in other cell types, enabling functional enrichment analysis to reveal their roles in distinct cellular contexts. Additionally, certain transcription factors and DNA-binding proteins are known to form complexes that cooperatively bind to the same promoters to regulate gene expression [48-50]. In these cases, not all proteins within the complex necessarily

355 exhibit consistently high or low ranks across target genes. Co-localization analysis of
356 NABPs could therefore provide further insights into the functional roles of these protein
357 complexes and their cooperative regulatory mechanisms. Such cooperative interactions
358 may obscure individual protein contributions, making co-localization analysis a valuable
359 tool for disentangling complex regulatory networks.

360 In conclusion, the deep learning model, parameter settings, and analytical
361 framework developed in this study provide a robust foundation for predicting the
362 functional roles of nucleic acid-binding proteins (NABPs) based on gene expression
363 correlations. Although this analysis focused on a limited number of proteins and cell types,
364 the approach demonstrated strong potential for identifying both known and previously
365 uncharacterized functional annotations. These findings contribute to advancing our
366 understanding of the biological and molecular functions of NABPs and their regulatory
367 target genes, particularly in cases where functional roles have not yet been fully
368 elucidated. These findings not only enhance our understanding of NABP functions but
369 also lay the groundwork for future studies aimed at elucidating complex gene regulatory
370 networks across diverse biological contexts.

371

372

373 # Methods

374 ## Deep learning analysis for predicting gene expression levels

375 I modified a previously developed deep learning approach to predict gene
376 expression levels across different cell and tissue types using DNA-binding site
377 information for transcription factors and other regulatory proteins. In this study, DNA-
378 binding sites located in regions far upstream and downstream of transcription start sites

379 (TSSs) were included in the input data [16,17] (**Fig. 1**). To reduce the effects of cellular
380 heterogeneity within tissues, gene expression data were collected at the cell type level.
381 RNA-seq data from various human cell types were obtained from public databases,
382 including human foreskin fibroblasts (HFFs; GSM941744), human mammary epithelial
383 cells (HMECs; ENCODE at UCSC Genome Browser,
384 wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam) [51,52], neural progenitor cells
385 (NPCs; GSM915326 and ENCFF529SIO), HepG2 hepatoma cells (ENCFF670LIE), and
386 K562 leukemia cells (ENCFF584WML). Gene expression levels (FPKM values) were
387 calculated using the RSeQC tool (http://rseqc.sourceforge.net/) based on GENCODE v19
388 gene annotations from the ENCODE portal (https://www.encodeproject.org/). FPKM
389 values were transformed to a $\log_2$ scale with the addition of 0.25 to avoid negative infinity.
390 Log_2-fold changes were calculated relative to the median $\log_2$-FPKM across the four cell
391 types, following the transcriptome data processing method described by Tasaki et al.
392 (2020) [16,17]. For the deep learning analysis, gene expression levels were predicted across
393 four cell types per model iteration. Three cell types (HFFs, HMECs, and NPCs) were
394 consistently included, while either HepG2 or K562 cells were incorporated depending on
395 the type of regulatory protein analyzed. Specifically, HepG2 cells were used for analyses
396 involving DNA-binding proteins to enable comparison with ChIP-seq data, whereas
397 K562 cells were used for RNA-binding protein analyses in conjunction with eCLIP data.
398 Regulatory target genes were selected based on gene co-expression data, representing
399 interactions between NABPs and their associated genes.

400 DNA-binding sites overlapping open chromatin regions in putative enhancers
401 located outside promoter regions were included as input features in the deep learning
402 model. To identify DNA-binding sites of DNA-binding proteins (DBPs) in promoter and

403 enhancer regions, ChIP-seq data for DBPs were obtained from the GTRD database
404 (release v20.06; https://gtrd.biouml.org) [53]. DBP binding sites regulating the expression
405 of each transcript were selected based on expression quantitative trait loci (eQTL) data
406 (GTEx_Analysis_v8_eQTL.tar) from the GTEx portal (release v8; https://gtexportal.org).
407 A DNA-binding site located within a 100-bp window centered on a single nucleotide
408 polymorphism (SNP) of an eQTL was considered to regulate the corresponding transcript.
409 Only cis-eQTLs were used; trans-eQTLs, where the SNP and transcript are located on
410 different chromosomes, were excluded. The genomic coordinates of the eQTLs were
411 converted to the hg19 human genome assembly using the UCSC LiftOver tool
412 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover) [54].
413 The original deep learning model utilized DNA-binding site data from promoter
414 regions as input features. Specifically, the input was defined as the total length of ChIP-
415 seq peaks for each DNA-binding protein (DBP) within 100-base-pair bins spanning from
416 2 kilobases (kb) upstream to 1 kb downstream of the transcription start site (TSS) for each
417 gene or transcript (**Fig. 1**). To incorporate DNA-binding sites located in more distal
418 upstream and downstream regions, and to reduce the total number of input bins, I first
419 calculated the peak length of each DBP within 100-bp bins across a ±1 megabase (Mb)
420 region surrounding the TSS. Subsequently, the maximum peak length within each
421 window of 100 consecutive bins was selected. This procedure yielded up to 200 input
422 values per transcript, representing the longest ChIP-seq peaks within the extended regions.
423 To define the input region for each gene, upstream and downstream boundaries were set
424 at the midpoint between the transcription start site (TSS) of the target gene and the TSS
425 of its nearest neighboring gene. This approach ensured that regulatory regions were
426 appropriately assigned without overlapping adjacent genes, resulting in a maximum of

427 200 input values per transcript, depending on the intergenic distances. The original model
428 was modified accordingly to accommodate these additional features.
429 The modified deep learning model was trained using the same test, training, and
430 validation datasets as described in the original method [16,17]. Specifically, all 2,705 genes
431 located on chromosome 1 were used as the test set; 22,251 genes were randomly selected
432 from other chromosomes for training, and 2,472 genes were allocated for validation. To
433 interpret the model, DeepLIFT [29] was employed to calculate contribution scores for DNA-
434 binding sites, indicating their importance in gene expression prediction. Input data
435 representing promoter and putative enhancer regions were concatenated into a single
436 input matrix comprising 230 bins: 30 bins for promoter regions and 200 bins for putative
437 enhancers, with each bin representing the length of ChIP-seq peaks for 1,310 DNA-
438 binding proteins (DBPs). This concatenation was necessary because the DeepLIFT
439 configuration used ('RevealCancel rule on dense layers and Rescale rule on conv layers',
440 recommended for genomics) does not support models with multiple input layers. Model
441 training and inference were performed on a system running Ubuntu 20.04 with Python
442 3.8.10, TensorFlow 2.6.0, the NVIDIA CUDA Toolkit 11.4.2, a GeForce RTX 3080 GPU
443 (10 GB memory), and an Intel Core i7-10700 CPU with 128 GB RAM. Original scripts
444 were updated to support newer software versions. The training procedure followed the
445 original method, with training times ranging from two to four hours.
446

447 **Deep learning analysis based on interactions between NABPs and genes**
448 **derived from gene co-expression data**

449 To predict the regulatory target genes of nucleic acid-binding proteins (NABPs)
450 without relying on protein-binding assay data such as ChIP-seq or eCLIP, I modified the

451 deep learning input by substituting low-contribution DNA-binding proteins with
452 alternative proteins. Among the 1,310 DNA-binding proteins originally used in the model,
453 302 proteins with low contribution scores for gene expression prediction were replaced
454 with other DNA-binding proteins. For RNA-binding protein analysis, 325 DNA-binding
455 proteins were substituted with RNA-binding proteins. Gene symbols for RNA-binding
456 proteins were obtained from the ENPD (Eukaryotic Nucleic acid Binding Protein)
457 database [28]. Interactions between the substituted NABPs and their putative target genes
458 were inferred based on gene co-expression pairs across different cell and tissue types, as
459 provided by COXPRESdb [18].

460 Gene co-expression refers to the similarity in expression profiles between pairs
461 of genes. Functionally related genes, as well as genes under shared regulatory control, are
462 typically co-expressed across different cell and tissue types. When a gene encoding a
463 nucleic acid-binding protein (NABP) exhibits expression patterns similar to those of other
464 genes, these genes are likely to be regulated by the NABP and functionally associated
465 with it. Co-expression interactions with a Pearson's z-score > 2 between genes encoding
466 NABPs and other genes (excluding NABPs) were used to replace low-contribution
467 features in the input data. For modeling purposes, the binding sites of NABPs were
468 hypothetically assigned to regions 2 kb upstream of the transcription start sites (TSSs)
469 and at the TSSs of their co-expressed genes. Using this modified input, deep learning
470 analysis combined with the DeepLIFT tool was applied to calculate contribution scores
471 for NABP binding sites in predicting gene expression levels.

472

473 **Functional prediction of target genes of NABPs**

474 For each nucleic acid-binding protein (NABP), I evaluated the rank of its
475 contribution scores by comparing the contribution scores of its binding sites to those of
476 other NABPs for each gene. Based on consistent tendencies toward high or low ranks
477 (i.e., high positive or low negative contribution scores) across the predicted regulatory
478 target genes, genes in which the NABP binding site ranked within the top five or bottom
479 five were selected as primary regulatory targets. To detect functional enrichment, the rank
480 threshold was adjusted individually for each NABP. In most cases, the threshold was
481 limited to the top five or bottom five ranks; however, for RPLP0, the threshold was
482 extended to the top eight to increase the number of selected genes.

483 Functional enrichment analysis of the primary regulatory target genes for each
484 nucleic acid-binding protein (NABP) was performed using in-house scripts and the
485 Statistical Overrepresentation Test provided by PANTHER (https://pantherdb.org/) [32]. To
486 evaluate the specificity of enriched Gene Ontology (GO) annotations, I compared the
487 number of significantly enriched GO terms across three gene sets: (1) the primary target
488 genes of each NABP, (2) all co-expressed genes with a Pearson's z-score > 2, and (3)
489 randomly selected genes matched in number to the co-expressed genes. GO annotations
490 from the three main categories—biological process, molecular function, and cellular
491 component—were combined for this analysis. The GO ontology file
492 (https://purl.obolibrary.org/obo/go.obo) and the human gene annotation file
493 (https://current.geneontology.org/annotations/goa_human.gaf.gz) were downloaded
494 from the Gene Ontology resource (https://geneontology.org/docs/download-ontology/)
495 [30,31]. Human gene symbols were converted to Ensembl Gene IDs using the HUGO Gene
496 Nomenclature Committee (HGNC) database (https://www.genenames.org/). For
497 PANTHER analysis, Ensembl Gene IDs of the primary target genes were used as the test

498 set, while Ensembl Gene IDs from the full gene expression dataset used in the deep
499 learning analysis served as the reference set. Functional enrichment was assessed using
500 Fisher's exact test, with multiple testing correction applied via the Benjamini–Hochberg
501 false discovery rate (FDR), considering FDR < 0.05 as statistically significant.

502 A ChatGPT-based method was used to infer the functional roles of the predicted
503 target genes of nucleic acid-binding proteins (NABPs) [34]. Functional roles were predicted
504 using the "Engineered prompt for gene set analysis" and the "Engineered prompt for
505 identifying genes supporting a proposed name," as described in Extended Data Tables 1
506 and 5, respectively [34]. Supporting literature linking the predicted functional roles to the
507 target genes was obtained by modifying and applying the Python script "4.Reference
508 search and validation.ipynb," available in the code repository on the Code Ocean platform
509 (Published Version 1.0; https://doi.org/10.24433/CO.7045777.v1) and GitHub
510 (https://github.com/idekerlab/llm_evaluation_for_gene_set_interpretation).

511

512 **Overlap between high-scoring genes and experimental targets**

513 To investigate the overlap between experimentally identified binding sites and
514 the predicted regulatory target genes of nucleic acid-binding proteins (NABPs) derived
515 from deep learning on gene co-expression data, DNA-binding sites within ±5 kb of
516 transcription start sites (TSSs) were selected based on ChIP-seq data, and RNA-binding
517 sites between transcription start and end sites (TESs) were selected based on eCLIP data.
518 Gene annotations were obtained from the human GENCODE v19 release [55]. ChIP-seq
519 data for 32 DNA-binding proteins and eCLIP data for 29 RNA-binding proteins were
520 collected from the ENCODE portal and a published study (**Supplementary Table S6**) [56].

521

522 **Statistical analysis**

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

529

530 **Data availability**

531 RNA-seq data for human foreskin fibroblast (HFF) cells were obtained from the Gene
532 Expression Omnibus (GEO) under accession number GSM941744. RNA-seq data for
533 human mammary epithelial cells (HMEC) were retrieved from ENCODE via the UCSC
534 Genome Browser (http://genome.ucsc.edu/encode/) under the file name
535 *wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam*. RNA-seq data for human neural
536 progenitor cells (NPC) were obtained from GEO under accession number GSM915326.
537 RNA-seq data for HepG2 and K562 cells were obtained from the ENCODE portal under
538 accession numbers ENCFF670LIE and ENCFF584WML, respectively. mRNA co-
539 expression data were downloaded from COXPRESdb
540 (https://zenodo.org/record/6861444). The list of DNA/RNA-binding proteins was
541 obtained from the ENPD (Eukaryotic Nucleic acid Binding Protein) database
542 (http://qinlab.sls.cuhk.edu.hk/ENPD/). ChIP-seq data for transcription factors were
543 downloaded from the GTRD database (v20.06 release; https://gtrd.biouml.org). eQTL
544 data (GTEx_Analysis_v8_eQTL.tar) were obtained from the GTEx portal (v8 release;
545 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

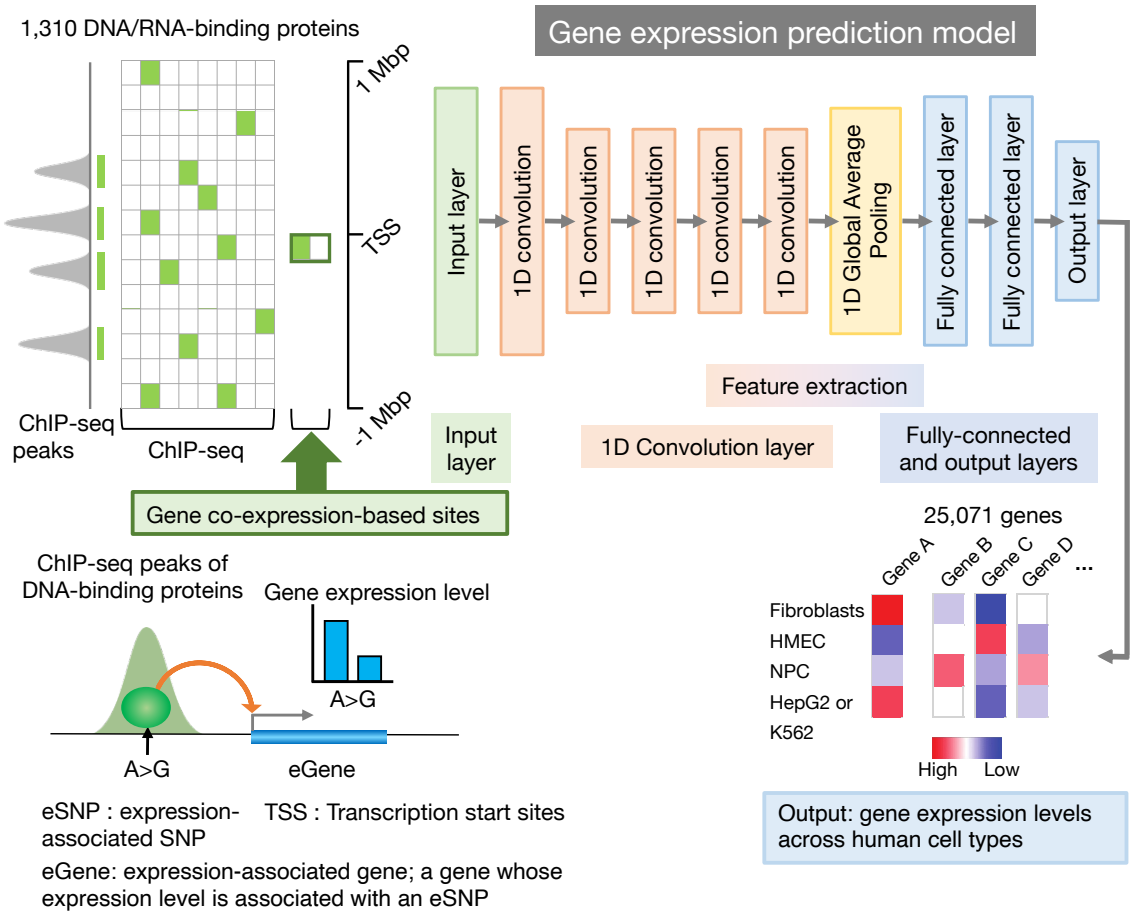
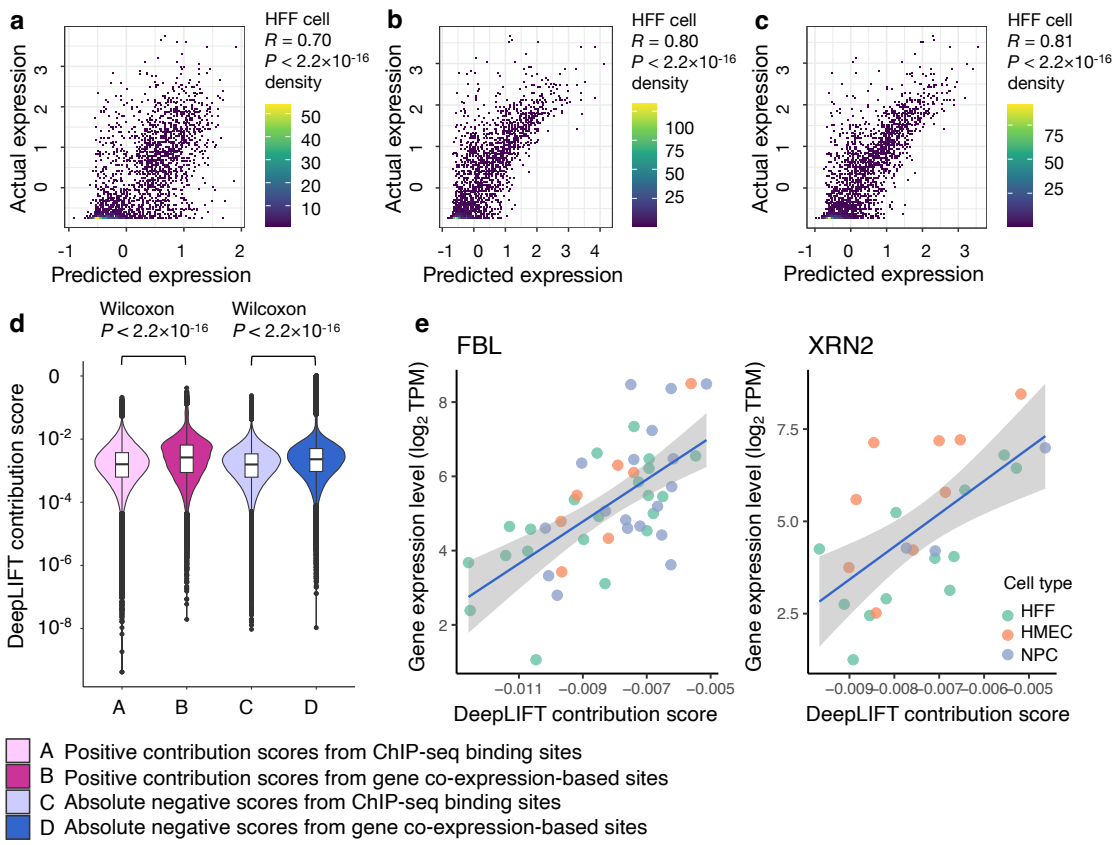


Figure 1. Overview of gene expression prediction.

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.



570 **Figure 2. Relationship between gene expression levels and contribution scores.**

571 Gene expression levels ($\log_2$ TPM) predicted by a deep learning model were compared

572 with actual expression levels in HFF cells using Spearman's rank correlation.

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

574 **b.** Among these, 302 proteins with low contribution scores were replaced with an equal

575 number of DNA-binding proteins selected based on gene co-expression.

576 **c.** Similarly, 325 DNA-binding proteins were replaced with RNA-binding proteins

577 selected via co-expression data.

578 **d.** DeepLIFT contribution scores were significantly higher for binding sites derived from

579 gene co-expression data compared to those from ChIP-seq data (Wilcoxon signed-rank

580 test, $P < 2.2 \times 10^{-16}$). A and B show distributions of positive contribution scores, while C

581 and D show absolute values of negative contribution scores.

582 **e.** Representative examples of the correlation between a regulator's DeepLIFT

583 contribution score and the expression levels of its predicted target genes. Linear

584 regression lines with 95% confidence intervals are shown.

585

586

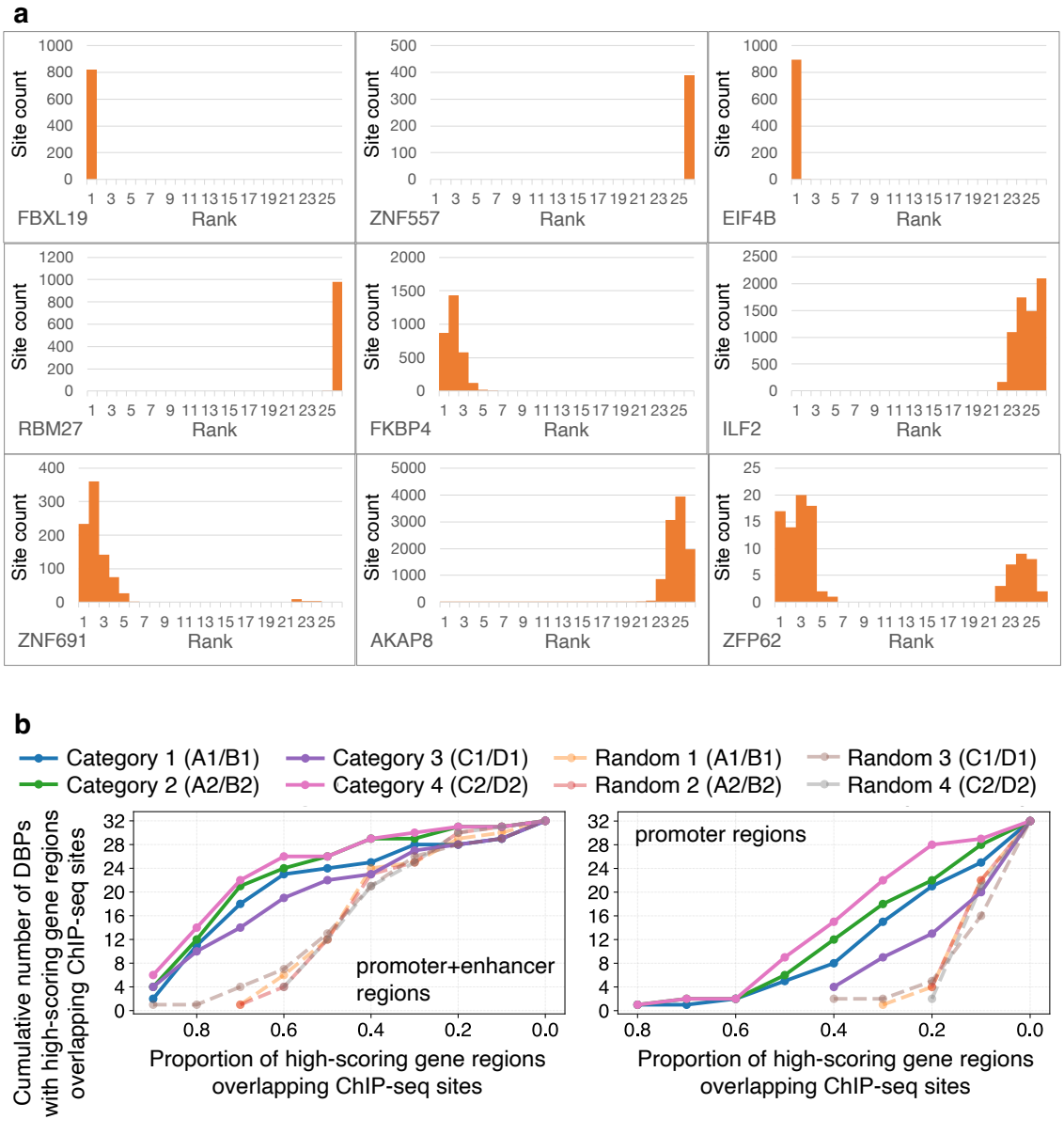


Figure 3. Comparison of predicted regulatory target genes with experimental ChIP-seq and eCLIP binding data. This figure illustrates how predicted gene-level regulatory contributions derived from deep learning align with experimentally identified DNA/RNA-binding protein targets across genomic regions.

a. Distribution of contribution scores across ChIP-seq–identified binding sites. The x-axis indicates the rank of contribution scores assigned to DNA/RNA-binding sites used to predict gene expression levels. Scores were ranked from 1 (highest positive

596 contribution, left) to 1,310 (lowest negative contribution, right) and binned into 50-rank
597 intervals. Rank 1 corresponds to ranks 1–50, rank 2 to ranks 51–100, and rank 26 to ranks
598 1,251–1,310. DNA/RNA-binding sites were categorized as being located in the promoters
599 or transcribed regions of genes experimentally bound by the DNA/RNA-binding protein
600 shown in each graph. These sites showed higher or lower contribution scores than those
601 across all ChIP-seq binding sites. The y-axis indicates the number of sites within each
602 50-rank bin. Positive contribution scores are indicative of gene activation, whereas
603 negative scores reflect gene repression.
604 **b. Proportion of high-contribution gene regions overlapping ChIP-seq binding sites**
605 **in promoter+enhancer and promoter regions, relative to all high-contribution**
606 **regions.** Labels indicating "promoter+enhancer regions" and "promoter regions" are
607 shown above the corresponding panels. The proportion represents the number of high-
608 contribution promoter+enhancer regions (within ±100 kb of the transcription start site
609 [TSS]; left panel) and promoter regions (within ±5 kb of the TSS; right panel) that overlap
610 with ChIP-seq binding sites, divided by the total number of high-contribution regions in
611 each category. To define high- and low-contribution regions, multiple thresholds were
612 applied to the 1,310 DNA-binding proteins analyzed: **Rank-based thresholds** — (A1
613 and B1): rank ≥ 20 or ≤ 1,290; (A2 and B2): rank ≥ 50 or ≤ 1,260. **Score-based thresholds**
614 — (C1 and D1): contribution score ≥ 0.01 or ≤ -0.01; (C2 and D2): contribution score ≥
615 0.001 or ≤ -0.001 (Supplementary Note and Table S2 and S3).
616
617

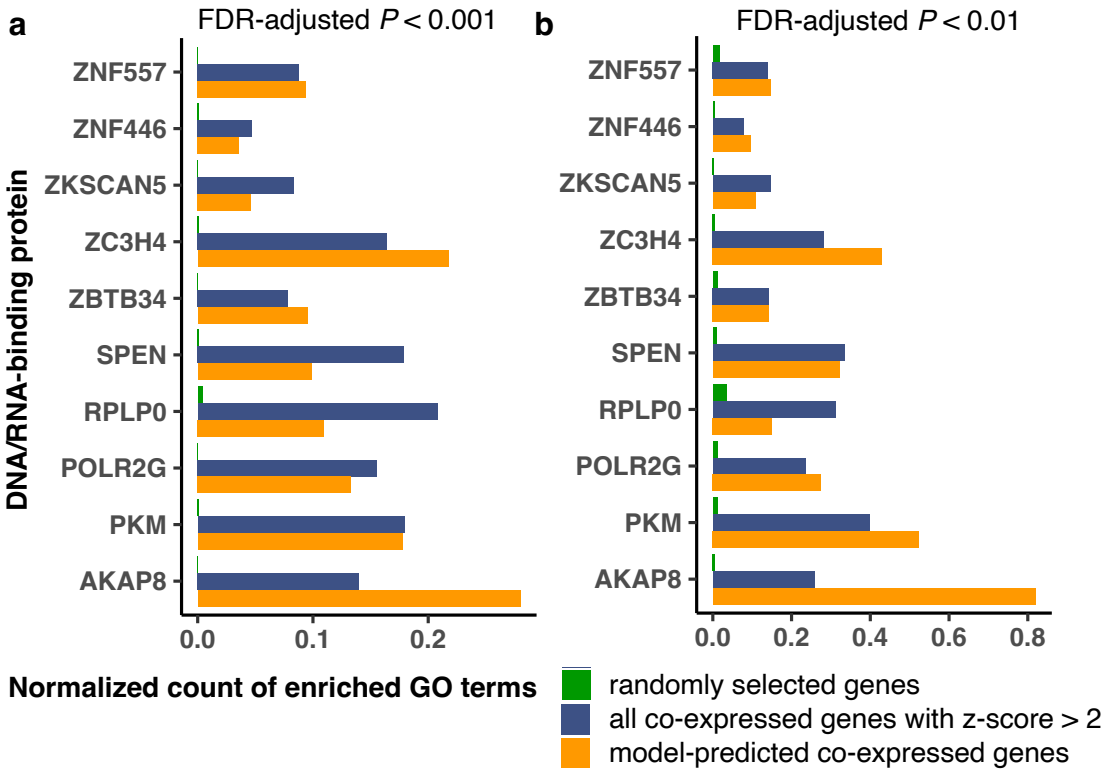


Figure 4. Overrepresented functional annotation terms among target genes of DNA/RNA-binding proteins.

a. and **b.** show results filtered at FDR-adjusted $P < 0.001$ and < 0.01, respectively. Functional enrichment was performed on three gene sets: model-predicted co-expressed genes, all co-expressed genes with a Pearson correlation z-score > 2, and randomly selected genes of equal size. Gene Ontology (GO) terms from biological process, molecular function, and cellular component categories were combined. Fisher's exact test *P*-values were FDR-adjusted.

AKAP8

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. [1][2][3]

[1] Miro, Caterina, Docimo, Annamaria, Barrea, Luigi, Verde, Ludovica, Cernea, Simona, Sojat, Antoan Stefan, Marina, Ljiljana V, Docimo, Giovanni, Colao, Annamaria, Dentice, Monica, Muscogiuri, Giovanna. "Time" for obesity-related cancer: The role of the circadian rhythm in cancer pathogenesis and treatment." Seminars in cancer biology, 2023, pp. 99-109, doi: https://doi.org/10.1016/j.semcancer.2023.03.003

[2] Liu, Zhihua, Chu, Guiyan. "Chronobiology in mammalian health." Molecular biology reports, 2013, pp. 2491-501, doi: https://doi.org/10.1007/s11033-012-2330-4

[3] Levine, Daniel C, Hong, Heekyung, Weidemann, Benjamin J, Ramsey, Kathryn M, Affinati, Alison H, Schmidt, Mark S, Cedernaes, Jonathan, Omura, Chiaki, Braun, Rosemary, Lee, Choogon, Brenner, Charles, Peek, Clara Bien, Bass, Joseph. "NAD^+ Controls Circadian Reprogramming through PER2 Nuclear Translocation to Counter Aging." Molecular cell, 2020, pp. 835-849.e7, doi: https://doi.org/10.1016/j.molcel.2020.04.010

PKM

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. [1][2][3]

[1] Massari, Francesco, Ciccarese, Chiara, Santoni, Matteo, Iacovelli, Roberto, Mazzucchelli, Roberta, Piva, Francesco, Scarpelli, Marina, Berardi, Rossana, Tortora, Giampaolo, Lopez-Beltran, Antonio, Cheng, Liang, Montironi, Rodolfo. "Metabolic phenotype of bladder cancer." Cancer treatment reviews, 2016, pp. 46-57, doi: https://doi.org/10.1016/j.ctrv.2016.03.005

[2] Cao, Lili, Wang, Min, Dong, Yujin, Xu, Bo, Chen, Ju, Ding, Yu, Qiu, Shusheng, Li, Liang, Karamfilova Zaharieva, Elena, Zhou, Xinwen, Xu, Yanbin. "Circular RNA circRNF20 promotes breast cancer tumorigenesis and Warburg effect through miR-487a/HIF-1α/HK2." Cell death & disease, 2020, pp. 145, doi: https://doi.org/10.1038/s41419-020-2336-0

[3] Pouysségur, J, Marchiq, I, Parks, S K, Durivault, J, Ždralević, M, Vucetic, M. "Warburg effect' controls tumor growth, bacterial, viral infections and immunity - Genetic deconstruction and therapeutic perspectives." Seminars in cancer biology, 2022, pp. 334-346, doi: https://doi.org/10.1016/j.semcancer.2022.07.004

Figure 5. Schematic overview of functional role prediction for DNA/RNA-binding proteins.

Gene–gene co-expression correlations derived from transcriptomic profiles were used as input to a deep learning model to predict gene expression levels. Contribution scores were

computed for each DNA/RNA-binding protein, and high-scoring genes were selected as predicted regulatory targets. Functional roles were inferred using a ChatGPT-based method by linking the predicted gene sets to relevant biological terms, and supporting publications were retrieved based on function-related keywords extracted from those terms.

Tables

Table 1. Functional annotations and inferred functional roles of predicted regulatory target genes of DNA/RNA-binding proteins.

Functional annotations for DNA/RNA-binding proteins were retrieved from the UniProt database. Significantly enriched annotations in the predicted regulatory target genes were identified using the PANTHER database (see Supplementary Table S4). Common or similar annotations shared between each DNA/RNA-binding protein and its predicted targets are shown. Functional roles inferred using a ChatGPT-based method are highlighted in yellow. Numbers in parentheses indicate confidence scores, which reflect the proportion of target genes involved in the corresponding biological process(es). Confidence scores range from 0.00 to 1.00: high (0.87–1.00), medium (0.82–0.86), low (0.01–0.81), and none (0). (Hu M. et al., *Nat. Methods*, 2025; https://doi.org/10.1038/s41592-024-02525-x).

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 enrichments among predicted regulatory target genes of DNA/RNA-binding proteins.

Functional enrichment analysis was conducted using the PANTHER database. This table lists annotations that were statistically enriched in regulatory target genes predicted by a deep-learning model trained on gene co-expression data. These annotations were not enriched in the full set of co-expressed genes (Pearson's z-score > 2) that included the DNA/RNA-binding protein itself. 'Reference' refers to the total number of human genes. 'Analyzed' denotes the number of regulatory target genes examined. 'Expected' is the number of genes expected to have the annotation by chance. 'Over/Under' indicates whether the annotation is overrepresented or underrepresented. 'Fold Enrichment' is the observed-to-expected ratio. 'Raw P-value' is the result of Fisher's exact test. 'FDR' 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.

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 Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274-1286, doi:10.1016/j.cell.2012.04.040 (2012).

12 Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature methods* **13**, 366-370, doi:10.1038/nmeth.3799 (2016).

13 Fagny, M. *et al.* Exploring regulation in tissues with eQTL networks. *Proc Natl Acad Sci U S A* **114**, E7841-E7850, doi:10.1073/pnas.1707375114 (2017).

14 Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory elements. *Nature*, doi:10.1038/s41586-024-08430-9 (2025).

15 Avsec, Z. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods* **18**, 1196-1203, doi:10.1038/s41592-021-01252-x (2021).

16 Tasaki, S., Gaiteri, C., Mostafavi, S. & Wang, Y. Deep learning decodes the principles of differential gene expression. *Nat Mach Intell* **2**, 376-386, doi:10.1038/s42256-020-0201-6 (2020).

17 Osato, N. & Hamada, M. Systematic discovery of directional regulatory motifs associated with human insulator sites. *bioRxiv* (2025).

18 Obayashi, T., Kodate, S., Hibara, H., Kagaya, Y. & Kinoshita, K. COXPRESdb v8: an animal gene coexpression database navigating from a global view to detailed investigations. *Nucleic acids research* **51**, D80-D87, doi:10.1093/nar/gkac983 (2023).

19 Yang, S. *et al.* COEXPEDIA: exploring biomedical hypotheses via co-expressions associated with medical subject headings (MeSH). *Nucleic acids research* **45**, D389-D396, doi:10.1093/nar/gkw868 (2017).

20 Zhu, Q. *et al.* Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nature methods* **12**, 211-214, 213 p following 214, doi:10.1038/nmeth.3249 (2015).

21 Miller, H. E. & Bishop, A. J. R. Correlation AnalyzeR: functional predictions from gene

co-expression correlations. *BMC Bioinformatics* **22**, 206, doi:10.1186/s12859-021-04130-7 (2021).

22 Raina, P. *et al.* GeneFriends: gene co-expression databases and tools for humans and model organisms. *Nucleic acids research* **51**, D145-D158, doi:10.1093/nar/gkac1031 (2023).

23 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).

24 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).

25 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).

26 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).

27 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

28 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).

29 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).

30 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).

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

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

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

34 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).

35 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).

36 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).

37 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).

38 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).

39 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).

40 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).

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

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

43 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).

44 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).

45 Ladomery, M. Multifunctional proteins suggest connections between transcriptional and post-transcriptional processes. *Bioessays* **19**, 903-909, doi:10.1002/bies.950191010 (1997).

46 Kim, S., Yu, N. K. & Kaang, B. K. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp Mol Med* **47**, e166, doi:10.1038/emm.2015.33 (2015).

47 Osato, N. Characteristics of functional enrichment and gene expression level of human putative transcriptional target genes. *BMC Genomics* **19**, 957, doi:10.1186/s12864-017-4339-5 (2018).

48 Burley, S. K. & Kamada, K. Transcription factor complexes. *Curr Opin Struct Biol* **12**, 225-230, doi:10.1016/s0959-440x(02)00314-7 (2002).

49 Goos, H. *et al.* Human transcription factor protein interaction networks. *Nat Commun* **13**, 766, doi:10.1038/s41467-022-28341-5 (2022).

50 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).

51 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).

52 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).

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

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

55 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).

56 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).