Title

Functional prediction of DNA/RNA-binding proteins using deep learning based on gene expression correlations

Naoki Osato [1*]

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

E-mail: naokiosato11@gmail.com

*Corresponding author.

Abstract

Deep learning methods have been developed to predict gene expression levels across various cell and tissue types based on DNA-binding sites of protein identified by ChIP-seq. However, the experimental identification of DNA/RNA-binding sites and the functional characterization of certain proteins remain challenging. Here I predicted the regulatory target genes and functional roles of DNA/RNA-binding proteins using gene co-expression data. Gene co-expression correlations are derived from gene expression profiles across cells and tissues and represent interactions between a DNA/RNA-binding protein and other genes. I used these interactions to predict gene expression levels by partially replacing the input data in the deep learning analysis. Functional annotations statistically overrepresented in the predicted target genes of ten DNA/RNA-binding proteins examined in this study were consistent with their known biological roles. Furthermore, several annotations not present in current gene function databases were supported by findings in previously published studies.

Keywords

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

Introduction

Deep learning approaches have improved the accuracy of predicting gene expression levels from the binding sites of DNA/RNA-binding proteins, such as transcription factors, as well as other regulatory factors including miRNAs [1]. Transcription factors bind to promoters near transcription start sites (TSSs) and to

enhancers, which are located distally upstream or downstream of genes. The method predicts gene expression levels in 53 human tissues based on experimentally determined binding sites of DNA/RNA-binding proteins in promoter regions and on predicted binding sites for miRNAs in transcribed regions. The correlation between actual and predicted gene expression levels was 0.82. In a previous study, I modified the method by including data on DNA-binding proteins in open chromatin regions within 1 Mbp of TSSs—some of which likely correspond to enhancers—and by excluding RNA-binding proteins and miRNAs. This resulted in a correlation coefficient of 0.70 [2].

The binding sites of DNA/RNA-binding proteins are identified using experimental assays such as ChIP-seq and ChIA-PET. These assays require high-quality antibodies against the target proteins, which are difficult to obtain for some proteins. For some DNA/RNA-binding proteins, determining their binding sequences through experimental analysis may be impractical due to limitations in time, cost, and labor. Large language model (LLM)-based deep learning architectures, such as Enformer, can predict DNA sequences that are critical for regulating gene expression [3]. Massively parallel reporter assays (MPRAs) have been used to measure the regulatory activity of hundreds of thousands of sequences [4]. However, it remains challenging to determine which DNA- or RNA-binding proteins interact with these sequences.

Analysis of gene co-expression across different cell and tissue types has suggested regulatory interactions among genes, including those encoding DNA- or RNA-binding proteins. Gene expression data from RNA-seq experiments and gene co-expression analysis results are publicly available in various databases [5-9]. Numerous studies have reported gene co-expression network analyses to infer functional relationships among genes [5,9-13]. Gene expression correlations may be influenced by both direct and indirect

73 molecular interactions. Some of these molecules may be elusive and difficult to detect
74 experimentally, as well as challenging to predict using alternative computational
75 approaches. These correlations may reflect regulatory mechanisms that have yet to be
76 fully characterized.

77 In this study, to investigate the functional annotations enriched in the putative target
78 genes of DNA- or RNA-binding proteins, I utilized gene co-expression data to identify
79 interactions between these proteins and other genes. Based on gene co-expression data,
80 the DNA-binding sites of approximately 300 proteins that showed low contribution scores
81 in my previous deep learning model were replaced with those of other DNA- or RNA-
82 binding proteins. The deep learning model learned to predict gene expression levels and
83 assigned importance to specific interactions between DNA-binding sites and genes that
84 contribute to the prediction. Regulatory target genes of a transcription factor often include
85 functionally related genes [14]. Therefore, functional enrichment among the predicted target
86 genes can be identified, even if the predictions are not always accurate enough for precise
87 identification of regulatory relationships. I examined the overlap between the predicted
88 regulatory target genes of DNA- or RNA-binding proteins and the binding sites identified
89 by experimental assays such as ChIP-seq and eCLIP. I found that the predicted regulatory
90 target genes whose binding sites were identified by experimental assays for a given DNA-
91 or RNA-binding protein exhibited higher absolute contribution scores than those for other
92 proteins. I then performed functional enrichment analyses on the predicted regulatory
93 target genes of ten DNA- or RNA-binding proteins and found that the enriched functional
94 annotations corresponded to known functions documented in the UniProt database. I also
95 identified functional annotations that were enriched in the predicted regulatory targets but
96 were not annotated in the UniProt database, although they have been described in prior

97 publications.

98 These analyses can facilitate the identification of functional roles and regulatory
99 targets of DNA/RNA-binding proteins without the need for experimental protein-binding
100 data. They can also reveal both direct and indirect interactions, including those difficult
101 to detect experimentally, thereby reducing analysis costs and enabling the discovery of
102 novel functions.

103

104 **Results**

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

106 To examine the effect of gene co-expression interactions on gene expression
107 prediction, the input data for 302 out of 1,310 DNA-binding proteins with low
108 contribution scores (DeepLIFT scores) were substituted with co-expression interaction
109 data involving mRNAs encoding DNA-binding proteins and other mRNAs not encoding
110 DNA- or RNA-binding proteins. DNA-binding proteins encoded by the co-expressed
111 genes were randomly selected from gene symbols listed in the ENPD database and other
112 databases used for DNA-binding motif analysis [15]. Notably, the correlation coefficient
113 between actual and predicted gene expression levels increased from 0.70 to 0.80 in human
114 foreskin fibroblast (HFF) cells when co-expression data were used (**Fig. 2a, b**). Then,
115 among the 1,310 DNA-binding proteins, 325 proteins with low contribution scores were
116 replaced with co-expression interaction data involving genes encoding RNA-binding
117 proteins and other mRNAs. The RNA-binding proteins were randomly selected from gene
118 symbols listed in the ENPD database. The correlation coefficient increased from 0.70 to
119 0.81 when co-expression data were used as well (**Fig. 2c**). These results suggest that gene
120 co-expression interactions contribute significantly to the prediction of gene expression

levels.

The contribution scores of the input features selected based on gene expression correlations were distributed across a wide range, but both their mean and median values were higher than those of the original input (**Fig. 2d and Supplementary Table S1**). I examined the distribution of Spearman's correlation coefficients between the contribution scores and gene expression levels ($\log_2$-TPMs; transcripts per million) (**Supplementary Fig. S2**), and plotted example relationships between the contribution scores of a DNA- or RNA-binding protein and the expression levels of its predicted regulatory target genes across cell types (**Fig. 2e**). Based on the interactions between co-expressed genes and DNA/RNA-binding proteins, the deep learning model adjusted its predictions of gene expression levels in a cell type–specific manner.

Overlap between high-scoring genes and experimental targets

To investigate whether the regulatory target genes of a DNA/RNA-binding protein, selected based on their contribution scores, are associated with the binding sites of that protein, I used ChIP-seq and eCLIP data. Specifically, a gene was considered associated with the protein if a ChIP-seq peak for a DNA-binding protein was located within ±5 kb of its transcription start site (TSS), or if an eCLIP peak for an RNA-binding protein was present in its mRNA-transcribed region. ChIP-seq data for 32 DNA-binding proteins in HepG2 cells and eCLIP data for 29 RNA-binding proteins in K562 cells were used. Genes with a DNA- or RNA-binding site for the protein showed high positive or low negative contribution scores among the predicted regulatory target genes of the DNA/RNA-binding protein derived from deep learning on gene co-expression data (Fig. 3 and Supplementary Fig. S1). These high and low contribution scores are likely associated

with the protein's regulatory roles in activating or repressing gene expression, as described in the documentation of DeepLIFT on its GitHub page and in previous studies using DeepLIFT [1 16].

Functional enrichment analysis of the target genes of a DNA- or RNA-binding proteins

The transcriptional target genes of a transcription factor often share functionally related annotations. Functional enrichment analysis of gene annotations can help predict the function of a transcription factor and verify that the predicted target genes are not randomly selected. To assess whether the regulatory target genes of a DNA/RNA-binding protein, 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 DNA/RNA-binding proteins [17,18]. The regulatory target genes of DNA/RNA-binding proteins exhibited functionally related annotations and significant enrichment in Gene Ontology (GO) terms. However, when the same number of genes were 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 in the regulatory target genes of DNA/RNA-binding proteins, I performed a statistical overrepresentation test using the gene list analysis tool provided by PANTHER (https://pantherdb.org/) [19]. This test utilizes functional annotation databases, including PANTHER Pathways, Protein Class, GO (Gene Ontology) Slim, GO, and Reactome Pathways. The functional annotations enriched in the putative regulatory target genes of a DNA/RNA-binding protein were compared with those of the protein

169 retrieved from the UniProt database (e.g.,
170 https://www.uniprot.org/uniprotkb/Q9Y2L8/entry) [20]. Ten examples of DNA/RNA-
171 binding proteins are presented in **Table 1** and **Supplementary Table S4**. The predicted
172 functional annotations of DNA/RNA-binding proteins correspond to those in the UniProt
173 database and published studies. Notably, among the ten examples, AKAP8 is annotated
174 with cell cycle-related functions, such as "GO:0000278 mitotic cell cycle" and
175 "GO:0044839 cell cycle G2/M phase transition," in the UniProt database
176 (https://www.uniprot.org/uniprotkb/O43823/entry). In the present study, AKAP8 was
177 also implicated in circadian clock regulation. A search of the PubMed database identified
178 two studies supporting the involvement of AKAP8 in circadian rhythms: one states that
179 "genes such as NFIL3, AKAP8, and SALL3, which are highly correlated with known
180 molecular clock genes (CRY1), can be identified as candidate circadian genes," while the
181 other reports that "AKAP5 and AKAP8 are both involved in the biosynthesis of the
182 circadian hormone melatonin, which is markedly disrupted in individuals with ASDs."
183 [21,22]. These findings demonstrate that the methods and analyses used in this study
184 successfully identified both known and novel functional annotations associated with
185 DNA/RNA-binding proteins.

186 Deep learning adjusted the contribution (i.e., neural network weights) of
187 DNA/RNA-binding proteins to the prediction of their regulatory target gene expression
188 during training. To evaluate the effect of learning on functional enrichment, I analyzed
189 all co-expressed genes (z score > 2) of ten DNA/RNA-binding proteins using Panther DB.
190 I then compared the enriched annotations in these co-expressed genes with those of the
191 regulatory target genes identified by deep learning (**Table 2** and **Supplementary Table**
192 **S5**). Although the number of enriched functional annotations increased for all co-

193 expressed genes due to the larger gene set, some annotations were uniquely or more
194 significantly enriched in the regulatory target genes of the DNA/RNA-binding protein
195 after learning. Notably, circadian rhythm-related functional annotations, as seen for
196 AKAP8, were identified only after learning (**Table 2**; yellow highlights in the
197 'S2_AKAP8' tab of **Supplementary Table S4**). Similarly, for PKM, kinase-related
198 functional annotations were detected only after learning (**Table 2** and 'S2_PKM' in
199 **Supplementary Table S4**). PKM interacts with ALDOA, as shown in the STRING
200 database for both mouse (https://string-db.org/network/10090.ENSMUSP00000034834)
201 and human [23,24]. Previous studies support the association between the functional
202 annotation "aldolase (PC00044)" and PKM in glycolysis. For example, one study
203 reported that "Aldolase A (ALDOA) is an enzyme that plays an important role in
204 glycolysis and gluconeogenesis, which is closely related to tumor metabolism." [25,26].
205 These functional annotations were enriched after the deep learning process. Thus, they
206 appear to be involved in biological processes in K562 cells, where the regulatory target
207 genes of the RNA-binding protein were identified. Circadian clocks and PKM have been
208 associated with human acute myeloid leukemia (AML) cells through glycolysis [27-30].
209 These results suggest that combining gene co-expression data with deep learning to
210 predict gene expression levels can reveal the functional associations of DNA/RNA-
211 binding proteins in a cell-type- or tissue-specific context.
212

213 ## Discussion

214 Deep learning models have been used to predict human gene expression levels
215 across cell types based on the DNA-binding sites of proteins. Here, I applied this approach
216 to predict novel regulatory interactions between DNA/RNA-binding proteins and their

target genes based on gene expression correlations across cell types. Gene expression correlations were statistically analyzed using RNA-seq data available from public databases across various cell and tissue types. Experimental identification of DNA/RNA-binding sites using assays such as ChIP-seq and eCLIP is more time-consuming, costly, and labor-intensive than RNA-seq experiments. Experimental determination of DNA/RNA-binding sites remains challenging for certain DNA/RNA-binding proteins. Deep learning analysis based on gene expression correlations predicted the regulatory target genes of DNA/RNA-binding proteins. Functional enrichment analysis of these genes highlighted potential functional associations and roles of the DNA/RNA-binding proteins. Many of the predicted annotations were consistent with known functional annotations reported in public databases. In addition, some predicted annotations not present in these databases were supported by previously published studies. Functionally related annotations were commonly observed among the regulatory target genes. Functional enrichment analysis revealed the potential roles and associations of DNA/RNA-binding proteins. Many of the predicted functional annotations were consistent with those reported in public databases. Additionally, annotations not previously described in these databases were also identified and supported by published studies.

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 may not completely overlap. Although ChIP-seq can identify tens of thousands of DNA-binding sites for a given protein, not all of them may be directly involved in gene expression regulation. In addition, some transcription factors are known to participate in multiple cellular functions [31,32]. The number of co-expressed genes (*z*

241 score > 2) for each DNA/RNA-binding protein was generally less than 1,500, which is
242 smaller than the number of genes with ChIP-seq peaks for the same proteins. However,
243 most of the statistically significant co-expressed genes are likely to be involved in gene
244 expression regulation. In our previous analysis of insulator-associated DNA-binding
245 proteins, 97 transcription factors and DNA-binding proteins were identified as being
246 potentially associated with insulator functions that influence enhancer–promoter
247 interactions [2]. Some of these proteins may play important roles in the formation of
248 chromatin loops, transcriptional condensates, and/or phase separation—mechanisms that
249 influence gene expression differently from the direct regulation by transcription factors.
250 In our deep learning analysis of insulator-associated DNA-binding proteins, we
251 incorporated eQTL data to identify DNA-binding sites associated with changes in gene
252 expression levels. However, in this study, I did not use DNA-binding motif sequences or
253 open chromatin regions identified by DNase-seq or ATAC-seq assays to determine DNA-
254 binding sites, as the objective was to investigate DNA-binding proteins with unknown
255 motifs. Moreover, the associations between eSNPs and eGenes in eQTL data may reflect
256 indirect interactions. The contribution scores of DNA-binding sites selected based on
257 gene expression correlations exhibited both high and low values, suggesting that indirect
258 interactions between DNA-binding proteins and other genes may influence these scores.
259 In the functional enrichment analysis of regulatory target genes, ten DNA/RNA-
260 binding proteins showed functional annotations consistent with those reported for the
261 respective proteins, suggesting that the number of predicted target genes was sufficient to
262 detect statistically significant functional associations. Statistical enrichment was assessed
263 using Fisher's exact test with false discovery rate (FDR) correction in PANTHER DB,
264 which generally required more than 100 regulatory target genes to achieve statistical

265 significance. For some proteins, fewer than 100 predicted target genes were identified,
266 limiting the detection of enriched functions. A potential next step in this study is to
267 enhance the detection of functional enrichment in small gene sets by incorporating
268 additional features, such as gene expression levels and gene co-expression correlations.
269 Furthermore, functional enrichment analysis using high or low-ranking regulatory target
270 genes revealed novel functions potentially associated with specific DNA/RNA-binding
271 proteins. These results suggest that regulatory target genes with high absolute
272 contribution scores can facilitate the identification of functional roles of DNA/RNA-
273 binding proteins in a cell type–specific manner.

274 When all input data in the deep learning analysis were replaced with interactions
275 between DNA/RNA-binding proteins and other genes based on gene expression
276 correlations, the number of functionally enriched annotations in the predicted regulatory
277 target genes decreased, and known functional annotations from the UniProt database were
278 no longer included. This suggests that a high replacement ratio of input data may result
279 in learning error-prone interactions between DNA/RNA-binding proteins and target
280 genes. In a previous study, when the regulatory target genes of a DNA-binding protein
281 were replaced with randomly selected genes, the number of enriched functional
282 annotations decreased as the number of random genes increased. The number of enriched
283 functions, normalized to the total number of predicted genes, reflects the accuracy of the
284 predicted regulatory target genes [33].

285 DNA-binding proteins were ranked for each gene based on the contribution scores
286 of their DNA-binding sites. When a protein's rank was within the top 50 or bottom 50
287 (i.e., rank ≤ 50 or ≥ 1,260 out of 1,310), functional enrichment of its putative target genes
288 was typically detectable using statistical tests in PANTHER DB. Some DNA/RNA-

289 binding proteins did not exhibit high- or low-ranking target genes in HepG2 cells.
290 However, these proteins may have such rankings in other cell types, allowing functional
291 enrichment analysis to reveal their roles in different cellular contexts. Some transcription
292 factors and DNA-binding proteins form protein complexes that regulate gene expression
293 by binding cooperatively to the same promoters [34-36]. In such cases, not all proteins within
294 the complex necessarily have consistently high or low ranks. Co-localization analysis of
295 DNA/RNA-binding proteins may help uncover the functional roles of these complexes.

296 In conclusion, the deep learning model, its parameters, and analytical framework
297 developed in this study provide a foundation for predicting the functional roles of
298 DNA/RNA-binding proteins based on gene expression correlations. Although the present
299 analysis focused on a limited number of proteins and cell types, the approach
300 demonstrated its potential for identifying both known and novel functional annotations.
301 These findings may contribute to a deeper understanding of the biological and molecular
302 functions of DNA/RNA-binding proteins and their regulatory target genes, especially
303 those with previously uncharacterized roles.

304

305 # Methods

306 **Deep learning analysis**

307 I modified a previously developed deep learning approach to predict gene
308 expression levels across different cell and tissue types using DNA-binding site
309 information of transcription factors and other regulatory proteins. In this study, DNA-
310 binding sites located in regions far upstream and downstream of transcription start sites
311 (TSSs) were included in the input data [1,2] (**Fig. 1**). To reduce the effects of cellular
312 heterogeneity within tissues, gene expression data were collected at the cell-type level.

313 RNA-seq data from various human cell types were obtained from public databases:
314 human foreskin fibroblasts (HFFs; GSM941744), human mammary epithelial cells
315 (HMECs; ENCODE at UCSC Genome Browser,
316 wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam) [37,38], neural progenitor cells
317 (NPCs; GSM915326 and ENCFF529SIO), HepG2 hepatoma cells (ENCFF670LIE), and
318 K562 leukemia cells (ENCFF584WML). FPKM values of gene expression were
319 calculated using the RSeQC tool (http://rseqc.sourceforge.net/) based on GENCODE v19
320 gene annotations in the ENCODE portal (https://www.encodeproject.org/). FPKM values
321 were transformed to $\log_2$ scale with the addition of 0.25 to avoid negative infinity. Log_2-
322 fold changes were calculated relative to the median $\log_2$-FPKM across the four cell types,
323 following the transcriptome data processing method described by Tasaki et al 2020 [1,2].

324 DNA-binding sites overlapping open chromatin regions in putative enhancers
325 located outside of promoter regions were included as input features in the deep learning
326 model. To identify the DNA-binding sites of DNA-binding proteins (DBPs) in promoter
327 and enhancer regions, ChIP-seq data for DBPs were obtained from the GTRD database
328 (release v20.06; https://gtrd.biouml.org) [42]. The binding sites of DBPs regulating the
329 expression of each transcript were selected based on expression quantitative trait loci
330 (eQTL) data (GTEx_Analysis_v8_eQTL.tar) from the GTEx portal (release v8;
331 https://gtexportal.org). A DNA-binding site located within a 100-bp window centered on
332 a single nucleotide polymorphism (SNP) of an eQTL was considered to regulate the
333 corresponding transcript. Only cis-eQTLs were used; trans-eQTLs, where the SNP and
334 the transcript are located on different chromosomes, were excluded. The genomic
335 coordinates of the eQTLs were converted to the hg19 human genome assembly using the
336 UCSC LiftOver tool (http://hgdownload.soe.ucsc.edu/downloads.html#liftover) [43].

The original deep learning model used DNA-binding site data in promoter regions as input features. Specifically, the input was defined as the length of ChIP-seq peaks of a DNA-binding protein (DBP) within each 100-base-pair bin in the region spanning from 2 kilobases (kb) upstream to 1 kb downstream of the transcription start site (TSS) of each gene or transcript (**Fig. 1**). To incorporate DNA-binding sites located in more distant upstream and downstream regions and to reduce the total number of input bins, I first computed the peak length of a DBP within each 100-bp bin across a ±1 megabase (Mb) region surrounding the TSS. I then selected the maximum peak length within each window of 100 consecutive bins. This procedure resulted in up to 200 input values per transcript, representing the longest ChIP-seq peaks in the extended region. The original model was modified to accommodate this additional data.

The modified deep learning model was trained using the same test, training, and validation datasets as in the original method [1,2]: all 2,705 genes located on chromosome 1 were used as the test set; 22,251 genes were randomly selected from other chromosomes for training, and 2,472 genes were selected for validation. To interpret the model, DeepLIFT [16] was used to calculate the contribution scores of DNA-binding sites, indicating their importance in gene expression prediction. Input data representing promoter and putative enhancer regions were concatenated into a single input matrix comprising 230 bins: 30 bins for promoter regions and 200 for putative enhancers, each bin representing the length of ChIP-seq peaks for 1,310 DNA-binding proteins (DBPs). This concatenation was necessary because the DeepLIFT configuration used ('RevealCancel rule on dense layers and Rescale rule on conv layers', recommended for genomics) does not support models with multiple input layers. Model training and inference were performed on a system running Ubuntu 20.04 with Python 3.8.10,

361 TensorFlow 2.6.0, the NVIDIA CUDA toolkit 11.4.2, a GeForce RTX 3080 GPU (10 GB
362 memory), and an Intel Core i7-10700 CPU with 128 GB RAM. Original scripts were
363 updated to support the newer software versions. The training procedure followed the
364 original method, and training time ranged from two to four hours.
365
366 **Functional prediction of DNA/RNA-binding proteins**
367 To predict the regulatory target genes of DNA/RNA-binding proteins without
368 relying on protein binding assay data such as ChIP-seq or eCLIP, I modified the deep
369 learning input by substituting low-contribution DNA-binding proteins with other proteins.
370 Among the 1,310 DNA-binding proteins originally used in the model, 302 proteins with
371 lower contribution scores for predicting gene expression levels were replaced with other
372 DNA-binding proteins. For RNA-binding protein analysis, 325 DNA-binding proteins
373 were replaced with RNA-binding proteins. Gene symbols of RNA-binding proteins were
374 obtained from the ENPD database [15]. Interactions between the substituted DNA/RNA-
375 binding proteins and their putative target genes were estimated based on gene co-
376 expression pairs across different cells and tissues, as obtained from COXPRESdb [5].
377 Gene co-expression refers to the similarity in expression profiles between pairs of
378 genes. Functionally related genes, as well as genes under common regulatory control, are
379 generally co-expressed across different cell and tissue types. When a gene encoding a
380 DNA/RNA-binding protein exhibits expression patterns similar to those of other genes,
381 those genes are likely to be regulated by the protein and functionally associated with it.
382 Co-expression interactions with a z-score > 2 between genes encoding DNA/RNA-
383 binding proteins and other genes (excluding DNA/RNA-binding proteins) were used to
384 replace those with lower contribution scores in the input data. The binding sites of the

385 DNA/RNA-binding proteins were hypothetically assigned to 2 kb upstream of the
386 transcription start sites (TSSs) and at the TSSs of their co-expressed genes, respectively.
387 Using this updated input, deep learning analysis and the DeepLIFT tool were applied to
388 calculate the contribution scores of the binding sites of DNA/RNA-binding proteins for
389 predicting gene expression levels.

390 For each DNA/RNA-binding protein, I evaluated the rank of its contribution score
391 by comparing the contribution scores of its binding sites with those of other DNA/RNA-
392 binding proteins for each gene. Based on the tendency of high or low ranks (i.e., high
393 positive or low negative contribution scores) across the predicted regulatory target genes,
394 I selected genes in which the protein's binding site ranked within the top five or bottom
395 five as its primary regulatory target genes. To detect functional enrichment, the rank
396 threshold was adjusted individually for each DNA/RNA-binding protein. In most cases,
397 the threshold was limited to the top five or bottom five ranks. For RPLP0, the threshold
398 was extended to the top eight to increase the number of selected genes.

399 Functional enrichment analysis of the primary regulatory target genes of each
400 DNA/RNA-binding protein was performed using in-house scripts and the Statistical
401 Overrepresentation Test provided by PANTHER (https://pantherdb.org/) [19]. To evaluate
402 the specificity of enriched Gene Ontology (GO) annotations, I compared the number of
403 significantly enriched GO terms among three gene sets: (1) the primary target genes of
404 each DNA/RNA-binding protein, (2) all co-expressed genes with a z-score > 2, and (3)
405 randomly selected genes matched in number to the co-expressed genes. GO annotations
406 from the three major categories—biological process, molecular function, and cellular
407 component—were combined for the analysis. The GO ontology file
408 (https://purl.obolibrary.org/obo/go.obo) and the human gene annotation file

409 (https://current.geneontology.org/annotations/goa_human.gaf.gz) were downloaded from
410 the Gene Ontology resource (https://geneontology.org/docs/download-ontology/) [17,18].
411 Human gene symbols were converted to Ensembl Gene IDs using the HUGO Gene
412 Nomenclature Committee (HGNC) database (https://www.genenames.org/). For
413 PANTHER analysis, Ensembl Gene IDs of the primary target genes were used as the test
414 gene set, and Ensembl Gene IDs from the full gene expression dataset used for deep
415 learning were used as the reference set. Functional enrichment was assessed using
416 Fisher's exact test, with multiple testing correction applied using the Benjamini–
417 Hochberg false discovery rate (FDR), with significance defined as FDR < 0.05.

418 To investigate the overlap between experimentally identified binding sites and the
419 predicted regulatory target genes of DNA/RNA-binding proteins derived from deep
420 learning on gene co-expression data, I selected DNA-binding sites located within ±5 kb
421 of a gene's transcription start site (TSS) based on ChIP-seq data, and RNA-binding sites
422 located between the TSS and transcription end site (TES) based on eCLIP data. Gene
423 annotations were based on the human GENCODE v19 release [40]. For this analysis, ChIP-
424 seq data for 32 DNA-binding proteins and eCLIP data for 29 RNA-binding proteins were
425 obtained from the ENCODE portal and a published study (**Supplementary Table S6**) [44].
426 I also evaluated the proportions of genes with high or low contribution scores, as defined
427 above, among those with ChIP-seq or eCLIP binding sites near the TSS or transcribed
428 regions, respectively.

429 To examine the binding sites of ChIP-seq data and eCLIP data around the
430 regulatory target genes of DNA/RNA-binding proteins predicted by deep learning of gene
431 co-expression data, I selected the DNA-binding sites within ±5 kbp of a TSS of a gene by
432 ChIP-seq assay or the RNA-binding sites between the transcription start and end of a gene

433 by eCLIP assay based on the Human GENCODE v19 gene annotation [40]. For this analysis,
434 ChIP-seq data of 32 DNA-binding proteins and eCLIP data of 29 RNA-binding proteins
435 were obtained from the ENCODE portal and the paper (**Supplementary Table S6**) [44]. I
436 also examined the ratios of the number of genes with high and low contribution scores of
437 the DNA/RNA-binding site around genes with the DNA/RNA-binding site by ChIP-seq
438 or eCLIP assays.
439
440 Statistical analysis
441 Functional enrichment was assessed using Fisher's exact test, with multiple testing
442 correction applied using the Benjamini–Hochberg false discovery rate (FDR) procedure.
443 Statistical significance was defined as FDR < 0.05. Distributions of DeepLIFT
444 contribution scores for DNA/RNA-binding sites derived from gene co-expression data
445 and ChIP-seq data were compared using the Mann-Whitney U test (also known as the
446 Wilcoxon rank sum test). A P value < 0.05 was considered statistically significant.
447
448 **Data availability**
449 RNA-seq data of human foreskin fibroblast (HFF) cells were obtained from the Gene
450 Expression Omnibus (GEO) under accession number GSM941744. RNA-seq data of
451 human mammary epithelial cells (HMEC) were retrieved from ENCODE at the UCSC
452 Genome Browser (http://genome.ucsc.edu/encode/) under the file name
453 *wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam*. RNA-seq data of human neural
454 progenitor cells (NPC) were obtained from GSM915326. RNA-seq data for HepG2 and
455 K562 cells were obtained from the ENCODE portal under accession numbers
456 ENCFF670LIE and ENCFF584WML, respectively. Co-expression data of mRNAs were

457 downloaded from COXPRESdb (https://zenodo.org/record/6861444). The list of
458 DNA/RNA-binding proteins was obtained from the ENPD database
459 (http://qinlab.sls.cuhk.edu.hk/ENPD/). ChIP-seq data of transcription factors were
460 downloaded from the v20.06 release of the GTRD database (https://gtrd.biouml.org).
461 eQTL data (GTEx_Analysis_v8_eQTL.tar) were obtained from the GTEx portal (v8
462 release) (https://gtexportal.org). eCLIP data of RNA-binding proteins were retrieved from
463 ENCSR456FVU in the ENCODE portal and from Supplementary Data 2 of Van Nostrand
464 *et al.*, *Nature* (2020), https://doi.org/10.1038/s41586-020-2077-3.
465
466 **Code availability**
467 The deep learning model used in this study is available at:
468 https://github.com/reposit2/coexgene .
469
470
471 **Figures**

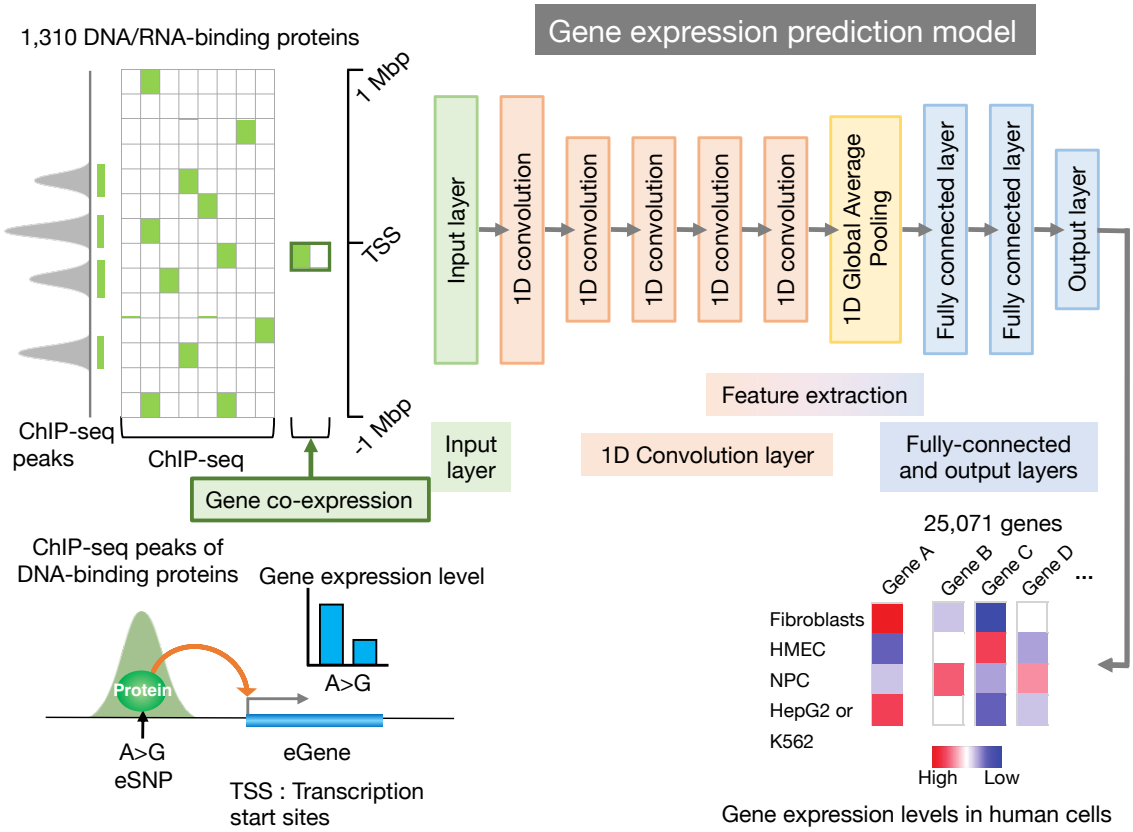


Figure 1. Overview of gene expression prediction. Gene expression levels in human cells were predicted from the genomic locations of DNA-binding sites at promoters and distal regions. During model training, a DNA-binding protein bound to a distal region overlapping an expression-associated SNP (eSNP) was linked to a potentially regulated gene (eGene) based on eQTL data. For each gene–protein association, ChIP-seq peak locations were binned (green box) and summarized into a matrix as input for deep learning.

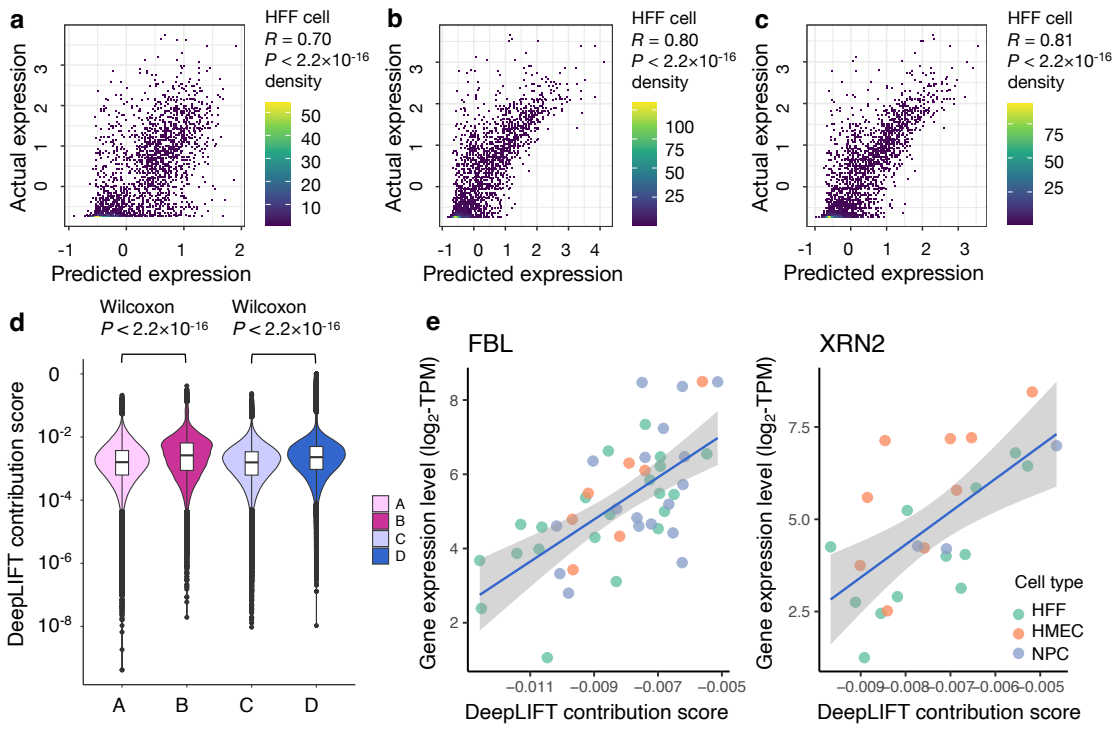
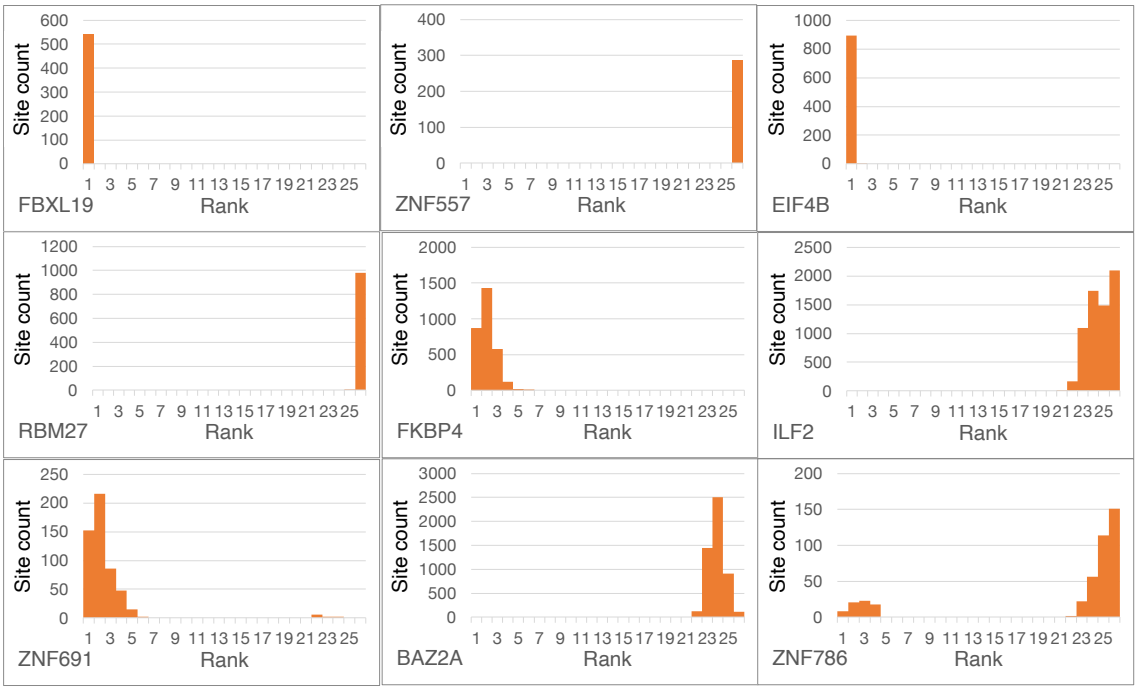


Figure 2. Relationship between gene expression levels and contribution scores. Predicted gene expression levels ($\log_2$-TPM) were compared with actual values in HFF cells using Spearman's rank correlation. **a** Input data included DNA-binding sites of 1,310 DNA-binding proteins obtained from ChIP-seq data. **b** Among the 1,310 proteins, 302 with low contribution scores were replaced with an equal number of DNA-binding proteins selected based on gene co-expression data. **c** Similarly, 325 low-contribution DNA-binding proteins were replaced with RNA-binding proteins selected based on gene co-expression data. **d** DNA-binding sites based on co-expression data showed higher DeepLIFT contribution scores. A and C: Binding sites of DNA-binding proteins derived from ChIP-seq data. B and D: Binding sites derived from co-expression data. A and B: Positive contribution scores. C and D: Absolute values of negative contribution scores (Wilcoxon signed-rank test, $P < 2.2 \times 10^{-16}$). **e** Example of the relationship between the contribution scores of a regulator and its expression levels across tissues. A linear

495 regression line with 95% confidence intervals is shown.

496



497

498 **Figure 3. Comparison of predicted target genes with experimental data.**

499 The x-axis represents the rank of contribution scores of DNA/RNA-binding sites used to
500 predict gene expression levels. Scores were ranked from 1 (highest positive contribution,
501 left) to 1,310 (lowest negative contribution, right) and binned into 50-rank intervals. Rank
502 1 corresponds to ranks 1–50, rank 2 to ranks 51–100, and rank 26 to ranks 1,251–1,310.
503 DNA/RNA-binding sites were categorized as being located in the promoters or
504 transcribed regions of genes experimentally bound by the DNA/RNA-binding protein
505 shown in each graph. These sites showed higher or lower contribution scores than those
506 across all ChIP-seq binding sites. The y-axis indicates the number of sites within each 50-
507 rank bin. Positive contribution scores are associated with gene activation, while negative
508 scores correspond to gene repression.

509

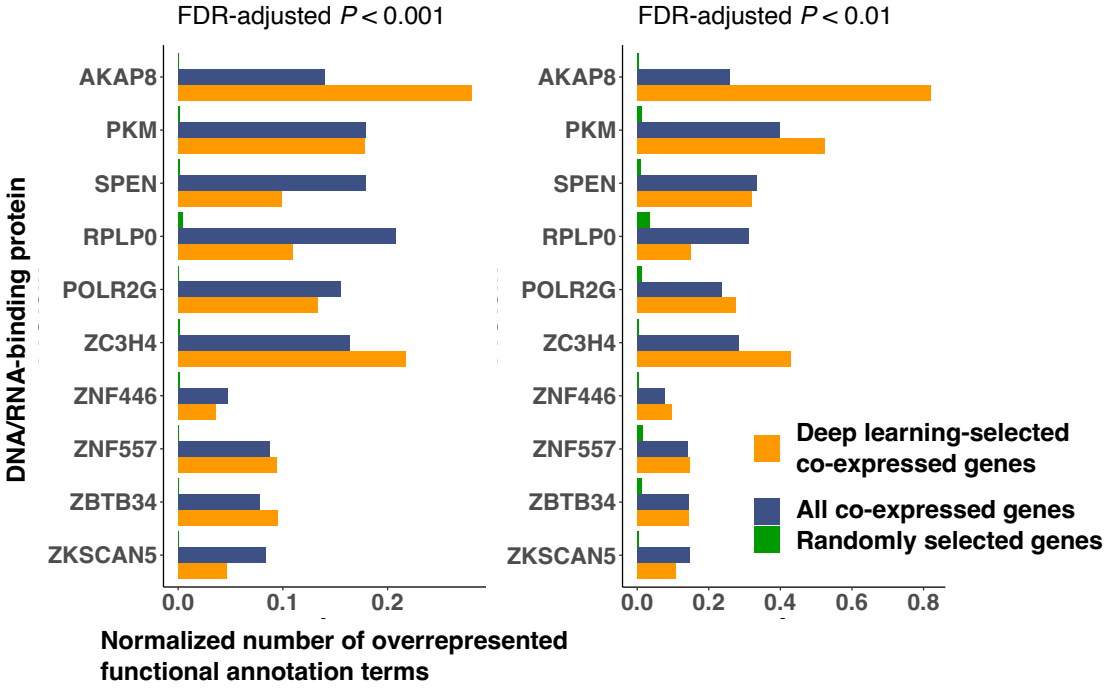


Figure 4. Number of overrepresented functional annotation terms in the target genes of DNA/RNA-binding proteins.

Functional enrichment analysis was performed on (1) co-expressed genes contributing to gene expression prediction in the deep learning model, (2) all co-expressed genes with a z-score > 2, and (3) randomly selected genes matched in number to the total co-expressed genes. Gene Ontology (GO) functional annotations from the three categories—biological process, molecular function, and cellular component—were combined for the analysis. *P*-values from Fisher's exact test were adjusted using the false discovery rate (FDR) method. The target genes of each DNA/RNA-binding protein were found to include functionally associated genes.

Tables

Table 1. Common functional annotations between DNA/RNA-binding proteins and

525 **their putative regulatory target genes.**

526 Functional annotations of DNA/RNA-binding proteins were obtained from the UniProt
527 database. Annotations significantly enriched in their putative regulatory target genes were
528 identified using the PANTHER database. Gray lines separate individual annotation
529 entries.

530

ZKSCAN5
DNA-binding transcription factor activity
C2H2 zinc finger transcription factor
RNA polymerase II cis-regulatory region sequence-specific DNA binding
metal ion binding

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

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

ZNF446
C2H2 zinc finger transcription factor
extracellular space

ZC3H4
chromatin/chromatin-binding, or -regulatory protein
transcription regulator activity
metal ion binding

POLR2G
RNA polymerase II transcription regulatory region sequence-specific DNA binding
general transcription factor
translational protein

RPLP0
large ribosomal subunit
ribonucleoprotein complex
cytosolic large ribosomal subunit

SPEN
mRNA binding
chromatin binding
transcription coregulator activity

PKM
kinase
glycolytic process
ATP metabolic process

AKAP8
negative regulation of RNA biosynthetic process
Circadian clock system
Cry-Per complex

531

532

533 **Table 2. Functional enrichments detected in the regulatory target genes of RNA-**
534 **binding proteins.**

535 Functional enrichment analysis was performed using the PANTHER database. This table
536 lists functional annotations that were statistically enriched in the regulatory target genes
537 of RNA-binding proteins after deep learning based on gene co-expression data, but not
538 enriched in the complete set of co-expressed genes (z-score > 2) including the RNA-
539 binding protein. 'Reference' indicates the total number of human genes. 'Analyzed' refers
540 to the number of regulatory target genes analyzed. 'Expected' is the number of genes with
541 the annotation expected by chance, based on its frequency in the reference set.
542 'Over/Under' indicates overrepresentation or underrepresentation. 'Fold Enrichment' is the
543 ratio of observed to expected counts. 'Raw P-value' represents the *P*-value from Fisher's
544 exact test, and 'FDR' indicates the Benjamini–Hochberg false discovery rate.
545

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 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).

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

3 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).

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

5 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).

6 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).

7 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).

8 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).

9 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).

10 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).

11 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).

12 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).

13 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).

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

15 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).

16 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).

17 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).

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

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

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

21 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).

22 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).

23 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).

24 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).

25 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).

26 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).

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

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

29 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).

30 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).

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

32 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).

33 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).

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

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

36 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).

37 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).

38 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).

39 Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic acids research* **48**, D882-D889, doi:10.1093/nar/gkz1062 (2020).

40 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).

41 Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics (Oxford, England)* **28**, 2184-2185, doi:10.1093/bioinformatics/bts356 (2012).

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

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

44 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).