Title

Functional prediction of DNA/RNA-binding proteins by deep learning from 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 in different cell and tissue types from DNA-binding sites of proteins including transcription factors identified by ChIP-seq assay. However, experimental identification of DNA/RNA-binding sites of proteins and functional characterization of proteins are difficult for some proteins. Gene co-expression correlation has been identified from gene expression data among cells and tissues, and includes the interactions between a DNA/RNA-binding protein and other genes. To predict the functional roles of DNA/RNA-binding proteins based on the interactions between a DNA/RNA-binding protein and other genes, I used the interactions to predict gene expression levels by partially replacing the input data in the deep leaning analysis. The correlation coefficient between actual and predicted gene expression levels increased from 0.70 to 0.80, indicating that gene co-expression data contribute to predict gene expression levels. Some of the functional annotations statistically overrepresented in the predicted target genes correspond to the known and associated functions of the DNA/RNA-binding proteins. Furthermore, functional annotations that are not registered in a database of functional annotations of genes have been confirmed in published studies. It contributes to predict the functional roles of DNA/RNA-binding proteins.

Keywords

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

Introduction

Deep learning analysis has improved the accuracy of predicting gene expression levels from the binding sites of DNA/RNA-binding proteins including transcription factors and other regulatory factors (Tasaki, et al., 2020). Transcription factors bind to promoters around transcription start sites (TSSs) and enhancers, which are located far upstream and downstream of genes and TSSs. The method predicts gene expression levels in 53 human tissues based on the experimental and predicted data of binding sites of DNA/RNA-binding proteins and miRNAs in promoter and mRNA-transcribed regions. The correlation coefficient between actual and predicted gene expression levels was 0.82. In my previous study, I modified the method to include the data of DNA-binding proteins bound at open chromatin regions in up to 1 Mbp upstream and downstream regions from TSSs, some of which regions would be associated with enhancers, and removed the binding sites of RNA-binding proteins and miRNAs, the correlation coefficient was 0.70 (Osato and Hamada, 2025).

The binding sites of DNA/RNA-binding proteins are identified by experimental assays such as ChIP-seq and ChIA-pet. These assays require high-quality antibodies of the proteins, but it is difficult to obtain high-quality antibodies for some proteins. For some DNA/RNA-binding proteins, it may be impractical to determine the binding DNA/RNA sequences by experimental analysis with the time, cost, and labor available. Large Language Model (LLM)-based deep learning models, such as Enformer, will predict DNA sequences that are important for predicting gene expression levels (Avsec, et al., 2021). Massively parallel reporter assays detected the activity of hundreds of thousands of sequences (Agarwal, et al., 2025). However, it would be difficult to identify which DNA/RNA-binding proteins bind to the DNA sequences.

Analysis of gene co-expression correlation between cell and tissue types suggested

73 regulatory interactions among genes, including mRNAs encoding DNA/RNA-binding
74 proteins. Gene expression data by RNA-seq assay and the results of gene co-expression
75 correlation analysis are available in databases (Miller and Bishop, 2021; Obayashi, et al.,
76 2023; Raina, et al., 2023; Yang, et al., 2017; Zhu, et al., 2015). Gene co-expression
77 network analysis has been reported (Lemoine, et al., 2021; Obayashi, et al., 2023; Raina,
78 et al., 2023; Stuart, et al., 2003; Warde-Farley, et al., 2010; Wong, et al., 2018). The
79 correlation of gene expression would be regulated by direct and/or non-direct interactions
80 of molecules. Some of the molecules may be elusive to detect by experimental analysis
81 and difficult to predict by other computational analyses. The correlation of gene
82 expression is potentially associated with regulations that have not been identified.

83 In this study, to investigate the functional annotations enriched in the putative
84 regulatory target genes of a DNA/RNA-binding protein, I used gene co-expression data
85 to extract the interactions between a DNA/RNA-binding protein and other genes. The
86 DNA-binding sites of about 300 proteins with a low contribution score in my previous
87 deep learning model were replaced with those of other DNA/RNA-binding proteins based
88 on gene co-expression data. Deep learning learned to predict gene expression levels and
89 selected interactions between a DNA-binding site and a gene that contribute to the
90 prediction. The regulatory target genes of a transcription factor include functionally
91 related genes (McLean, et al., 2010). Thus, the functional enrichment of genes is
92 potentially discovered, although the prediction of regulatory target genes may not be
93 precise enough to achieve a research purpose. I compared the overlap between the
94 predicted regulatory target genes of a DNA/RNA-binding proteins and the DNA/RNA-
95 binding sites by experimental assays such as ChIP-seq and eCLIP. Then, I performed
96 functional enrichment analyses of the predicted regulatory target genes of ten DNA/RNA-

97 binding proteins and found that the enriched functional annotations corresponded to the
98 known functional annotations in the UniProt database. Furthermore, I discovered that
99 some functional annotations enriched in the putative regulatory target genes of a
100 DNA/RNA-binding protein were not described in the UniProt database, but were
101 investigated in published papers.

102 These analyses would be useful to identify the functional roles and regulatory target
103 genes of DNA/RNA-binding proteins without using experimental data from protein-
104 binding assays, and are expected to predict the direct and non-direct interactions between
105 DNA/RNA-binding proteins and their regulatory target genes, where some interactions
106 may be difficult to study by experimental analysis. This will contribute to the discovery
107 of novel interactions and functions, and reduce the cost of experimental analysis.

108

109 Results

110 To examine the effect of the interaction data of co-expressed genes on the
111 prediction of gene expression levels, among 1,310 DNA-binding proteins in the input
112 data of deep learning, 302 DNA-binding proteins with low contribution scores
113 (DeepLIFT scores) were replaced with the interaction data of co-expressed genes
114 encoding DNA-binding proteins and the mRNAs other than DNA/RNA-binding proteins.
115 The DNA-binding proteins encoded by the co-expressed genes were randomly selected
116 from the gene symbols of DNA-binding proteins in the ENPD database (Tak Leung, et
117 al., 2019), and in the other databases of DNA-binding proteins used for the analysis of
118 DNA-binding motifs. Interestingly, the correlation coefficient between actual and
119 predicted gene expression levels increased from 0.70 to 0.80 in human foreskin fibroblast
120 (HFF) cells using co-expression data (**Fig. 2a, b**). Then, among 1,310 DNA-binding

proteins, 325 DNA-binding proteins with low contribution scores were replaced with the interaction data of co-expressed genes encoding RNA-binding proteins and other mRNAs. The RNA-binding proteins encoded by the co-expressed genes were randomly selected from the gene symbols of RNA-binding proteins in the ENPD database. The correlation coefficient increased from 0.70 to 0.81 using co-expression data as well (**Fig. 2c**). These results suggest that the interactions of co-expression of genes contribute to the prediction of gene expression levels.

The contribution scores of the input data selected from the gene expression correlations were distributed over a wide range between high and low, but the mean and median of the scores were larger than the original (**Fig. 2d and Supplementary Table S1**). I examined the distribution of Spearman's correlations between contribution scores and gene expression levels (log2-TPMs, transcripts per million) (**Supplementary Fig. S2**), and plotted example relationships between the contribution scores of a DNA/RNA-binding protein and the expression levels of the regulatory target genes across cell types (**Fig. 2e**). Based on the interactions between co-expressed genes and a DNA/RNA-binding protein, deep learning adjusted the prediction model of gene expression levels in each cell type.

To investigate whether the regulatory target genes of a DNA/RNA-binding protein selected based on their contribution scores are associated with the DNA/RNA-binding sites of the protein, I used the DNA-binding sites by ChIP-seq assay and the RNA-binding sites by eCLIP assay around the genes, where a DNA-binding site of the protein by ChIP-seq assay within ±5 kb of TSSs or an RNA-binding site of the protein by eCLIP assay in mRNA-transcribed regions of a gene (ChIP-seq data of 32 DNA-binding proteins in HepG2 cell and eCLIP data of 29 RNA-binding proteins in K562 cell). The genes with a

145 DNA-binding site of the protein or an RNA-binding site of the protein corresponded to
146 high positive or low negative contribution scores of the regulatory target genes of the
147 DNA/RNA-binding protein derived from the deep leaning of gene co-expression data
148 (**Fig. 3 and Supplementary Fig. S1**). The high and low contribution scores would be
149 associated with the regulatory functions of activation and repression of gene expression
150 from the description of the contribution score (DeepLIFT score) on the GitHub webpage
151 of DeepLIFT (Shrikumar, et al., 2017) and a previous study using DeepLIFT (Tasaki, et
152 al., 2020).

153 The transcriptional target genes of a transcription factor include associated
154 functions of genes. The functional enrichment of genes would be useful to predict the
155 function of a transcription factor and to confirm that the transcriptional target genes are
156 not randomly selected. For the functional enrichment analysis of regulatory target genes
157 of DNA/RNA-binding proteins, I used the statistical overrepresentation test of the gene
158 list analysis of Panther DB (https://pantherdb.org/) (Thomas, et al., 2022). The test uses
159 functional annotation databases such as PANTHER Pathways, Protein Class, GO (Gene
160 Ontology)-Slim, GO, and Reactome Pathways. The functional annotations enriched in
161 putative regulatory target genes of a DNA/RNA-binding protein were compared with
162 those of the DNA/RNA-binding protein obtained from the UniProt database (e.g.
163 https://www.uniprot.org/uniprotkb/Q9Y2L8/entry) (UniProt, 2025). The ten examples of
164 DNA/RNA-binding proteins are shown in **Table 1** and **Supplementary Table S4**. The
165 putative functional annotations of the DNA/RNA-binding proteins correspond to and are
166 associated with those in the UniProt database and publications. Notably, among the ten
167 examples, AKAP8 has cell cycle-related functional annotations such as "GO:0000278
168 mitotic cell cycle" and "GO:0044839 cell cycle G2/M phase transition" in the UniProt

169 database (https://www.uniprot.org/uniprotkb/O43823/entry). In this study, AKAP8 was
170 suggested to be involved in circadian clock systems. By searching Google and the
171 PubMed database, I found the following papers showing that AKAP8 is involved in
172 circadian rhythms: "But the genes, such as NFIL3, AKAP8, and SALL3, highly correlated
173 with known molecular clock genes (CRY1) can be identified as the candidate circadian
174 genes." and "AKAP5 and AKAP8 are both involved in the biosynthesis of the circadian
175 hormone melatonin, which is markedly disturbed in people with ASDs." (Poelmans, et
176 al., 2013; Tan, et al., 2024). These results demonstrate that the method and analysis in this
177 study detected known and novel functional annotations associated with DNA/RNA-
178 binding proteins.

179 The deep learning analysis changed the contribution (i.e., weights in the neural
180 network) of a DNA/RNA-binding protein to the prediction of the expression levels of the
181 regulatory target genes of the DNA/RNA-binding protein during the learning phase. To
182 investigate the effect of learning on the functional enrichment analysis of the target genes,
183 I performed the functional enrichment analysis of all co-expressed genes (z score > 2) of
184 the ten DNA/RNA-binding proteins using Panther DB, and compared the functional
185 annotations enriched in all co-expressed genes of a DNA/RNA-binding protein and those
186 of the regulatory target genes of the DNA/RNA-binding protein in the deep learning
187 analysis shown in **Table 2** and **Supplementary Table S5**. Although the number of
188 functional annotations enriched in all co-expressed genes of a DNA/RNA-binding protein
189 increased due to the increase in genes, some functional annotations were only or highly
190 enriched in the regulatory target genes of the DNA/RNA-binding protein after learning.
191 Notably, as in the case of AKAP8 functional annotations, circadian rhythm-related
192 functional annotations were detected only after learning (**Table 2** and annotations

highlighted in yellow in the 'S2_AKAP8' tab of **Supplementary Table S4**). As for PKM functional annotations, kinase etc. were identified only after learning (**Table 2** and 'S2_PKM' in **Supplementary Table S4**). PKM interacts with ALDOA (Snel, et al., 2000; Szklarczyk, et al., 2023) (STRING database interaction, mouse PKM https://string-db.org/network/10090.ENSMUSP00000034834 ; human PKM also has an interaction with ALDOA in the database), and I found research papers supporting that the functional annotation "aldolase (PC00044)" is associated with PKM involved in glycolysis: "Aldolase A (ALDOA) is an enzyme that plays an important role in glycolysis and gluconeogenesis, which is closely related to tumor metabolism" (Tian, et al., 2022; Zhang, et al., 2017). These functional annotations were enriched after learning. Thus, they would be involved in the functions in K562 cell where the regulatory target genes of an RNA-binding protein were selected. Circadian clocks and PKM would be associated with human acute myeloid leukemia (AML) cells via glycolysis (Lee, 2021; Puram, et al., 2016; Yang, et al., 2024; Zlacka and Zeman, 2021). These results suggest that combining gene co-expression data with deep learning to predict gene expression levels would reveal the functional associations of DNA/RNA-binding proteins in a cell or tissue type.

Discussion

The expression levels of human genes in cell types have been predicted using deep learning based on the DNA-binding sites of proteins. Here, I applied this method to predict unknown interactions between a DNA/RNA-binding protein and genes based on gene expression correlations between cell types. Gene expression correlations were statistically analyzed using gene expression data from RNA-seq experiments in cell and tissue types. Experimental analysis of DNA/RNA-binding sites by ChIP-seq, eCLIP

assays, etc. is more time-consuming, expensive, and laborious than RNA-seq experiments. Experimental analysis of DNA/RNA-binding sites is still difficult for some DNA/RNA-binding proteins. Deep learning analysis of DNA/RNA-binding proteins based on correlations of their expression levels with other genes predicted regulatory target genes of the proteins. Regulatory target genes include associated functions of genes. Enrichment analysis of the functional annotations of the regulatory target genes revealed the functional associations and roles of the DNA/RNA-binding proteins. The predicted functional annotations correspond to the known functional annotations of the proteins. Furthermore, functional annotations not described in a database were also discovered and confirmed in published studies.

When comparing DNA-binding sites predicted based on the contribution scores for predicting gene expression levels with DNA-binding sites by ChIP-seq assay, the two types of DNA-binding sites may not overlap completely. Although tens of thousands of DNA-binding sites of proteins are identified by ChIP-seq assay, some of them may not be directly involved in the regulation of gene expression. Some transcription factors are involved in multiple functions (Ladomery, 1997) (Kim, et al., 2015). The number of co-expressed genes of a DNA/RNA-binding protein with z score ≥ 2 was about less than 1,500, and was less than the number of genes with ChIP-seq peaks of the DNA/RNA-binding proteins, but the most of the co-expressed genes with a statistical significance would be involved in the regulation of gene expression. Our previous analysis of insulator-associated DNA-binding proteins revealed that 97 transcription factors and DNA-binding proteins would be associated with insulator functions that affect enhancer-promoter interactions (Osato and Hamada, 2025). Some of them would be important for the formation of chromatin loops, transcriptional condensates and/or phase separation,

241 which are different mechanisms to affect gene expression from the direct regulation of
242 genes by transcription factors. In the deep learning analysis of insulator-associated DNA-
243 binding proteins, we used eQTL data to find DNA-binding sites of proteins that affect
244 gene expression levels. However, in this study, I did not use the DNA-binding motif
245 sequences of proteins and open chromatin regions by DNase-seq or ATAC-seq assay to
246 identify their DNA-binding sites because the purpose of this study is to investigate DNA-
247 binding proteins whose DNA-binding motifs are unknown. Moreover, the association
248 between eSNP and eGene of eQTL data may include indirect interactions. The
249 contribution scores of DNA-binding sites of proteins selected based on gene expression
250 correlation showed higher or lower scores, suggesting that indirect interactions between
251 DNA-binding proteins and other genes would affect the scores.

252 As for the analysis of functional enrichment of regulatory target genes of a DNA-
253 binding protein, ten DNA/RNA-binding proteins correspond to the known functional
254 annotations of the DNA/RNA-binding proteins, suggesting that the number of predicted
255 regulatory target genes is sufficient to identify the functional annotations of the proteins
256 by a statistical test. The identification of functional enrichments of regulatory target genes
257 of a protein seemed to require more than one hundred of genes by a Fisher's exact test
258 adjusted with a false discovery rate (FDR) in Panther DB. The number of regulatory target
259 genes of some proteins was less than a hundred, and their functional enrichments seemed
260 to be difficult to identify. One of the next steps of this study would be to improve the
261 method for identifying the functional enrichments of a small number of genes by
262 considering other information such as gene expression levels and correlation of gene
263 expression.

264 When I replaced all the input data of the deep learning analysis with the interactions

265 between DNA/RNA-binding proteins and other genes based on gene expression
266 correlation, the number of functional enrichments of predicted regulatory target genes of
267 a DNA/RNA-binding protein was reduced, and the result of the functional enrichment
268 analysis did not include the same functional annotations as the known functional
269 annotations of the DNA/RNA-binding protein in the UniProt database. Thus, the high
270 replacement ratio of the input data would lead to learning error-prone interactions
271 between a DNA/RNA-binding protein and regulatory target genes. In my previous study,
272 when the regulatory target genes of a DNA-binding protein were replaced by randomly
273 selected genes, the number of functional enrichments decreased as the number of
274 randomly selected genes increased. The number of functional enrichments reflects the
275 accuracy of the regulatory target genes (Osato, 2018).

276 The DNA-binding proteins were ranked according to the contribution scores of
277 their DNA-binding sites for each gene. When the rank of a DNA-binding protein is higher
278 or equal to 50 or lower or equal to 1,260 out of a total of 1,310, the functional enrichments
279 of putative target genes appeared to be identified by the statistical test in Panther DB.
280 Some DNA/RNA-binding proteins did not have high or low-ranking putative regulatory
281 target genes in HepG2 cells. These proteins may have high or low-ranking target genes
282 in other cell types, and thus functional enrichment analysis of target genes could be
283 performed to reveal their functional roles. Some transcription factors and DNA-binding
284 proteins would bind to each other to form the protein complex and regulate gene
285 expression levels (Burley and Kamada, 2002) (Goos, et al., 2022) (Berenson, et al., 2023).
286 These proteins would bind to the same promoters of genes and the ranking of each protein
287 in the complex may not always be high or low. Analysis of the co-localization of
288 DNA/RNA-binding proteins would reveal the functional roles of the protein complex.

The deep learning model, model parameters, and analysis methods used in this study could be adapted and improved for the purpose of this study. I examined a small number of DNA/RNA-binding proteins in a few cell types in this study, and would be able to predict the functional roles of a larger number of DNA/RNA-binding proteins in cell and tissue types. The results of the analysis would be useful to understand and study the biological and molecular functions of unknown functions of DNA/RNA-binding proteins and their regulatory target genes.

Methods

Deep learning analysis

I modified the deep learning method for predicting gene expression levels among cell and tissue types based on the DNA-binding sites of transcription factors and other regulators, and included the DNA-binding sites of transcription factors in regions far upstream and downstream of TSSs (Osato and Hamada, 2025; Tasaki, et al., 2020) (**Fig. 1**). To reduce the heterogeneity of cells in a tissue, I used gene expression data of cell types. The gene expression data of different cell types were obtained from public databases: human foreskin fibroblasts (HFFs) (GSM941744), human mammary epithelial cells (HMECs) (Encyclopedia of DNA Elements at UCSC http://genome.ucsc.edu/encode/, 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam') (Barrett, et al., 2013; Edgar, et al., 2002), neural progenitor cells (NPCs) (GSM915326, ENCFF529SIO), human hepatoma cell line (HepG2) (ENCFF670LIE), and human leukemia cell line (K562) (ENCFF584WML). FPKM values of gene expression levels were calculated based on GENCODE V19 gene annotations in the ENCODE portal database

313 (https://www.encodeporoject.org/) (Luo, et al., 2020; Mudge, et al., 2025) using the
314 RSeQC tool (http://rseqc.sourceforge.net/) (Wang, et al., 2012). The FPKM values were
315 transformed to log2-fold with the addition of 0.25 to avoid negative infinity results. Log2-
316 fold changes relative to the median log2-FPKM across the four cell types were calculated
317 according to the original deep learning method (transcriptome data processing as
318 described by Tasaki et al.) (Osato and Hamada, 2025; Tasaki, et al., 2020).

319 The DNA-binding sites overlapping with open chromatin regions in putative
320 enhancers located outside of promoter regions were added as input data in the deep
321 learning method. To identify the DNA-binding sites of DNA-binding proteins (DBPs) in
322 promoter and enhancer regions, ChIP-seq data of the DBPs were obtained from GTRD
323 database release v20.06 (https://gtrd.biouml.org) (Kolmykov, et al., 2021). The binding
324 site of a DBP that regulates the expression level of each transcript was selected based on
325 eQTL data (GTEx_Analysis_v8_eQTL.tar) from GTEx portal release v8
326 (https://gtexportal.org). The DNA-binding site located within a 100-bp region around a
327 single-nucleotide polymorphism (SNP) of an eQTL was supposed to regulate the
328 expression of the transcript corresponding to the eQTL. I did not use *trans*-eQTL data,
329 where the SNP of an eQTL was not located on the same chromosome as the transcribed
330 region of the eQTL. The genomic locations of the eQTLs were converted to the
331 corresponding locations in the hg19 version of the human genome using the LiftOver tool
332 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover) (Hinrichs, et al., 2006).

333 The original model of the input data of DNA-binding sites in promoter regions was
334 used for the deep learning analysis: the lengths of a DBP ChIP-seq peak in each 100-base
335 bin in a region between 2 kbp upstream and 1 kbp downstream of the TSS of each gene
336 or transcript (**Fig. 1**). To add the input data of DNA-binding sites of DBPs in more

upstream and downstream regions of a TSS and reduce the number of bins, I first computed the length of a DBP ChIP-seq peak in a 100-base bin and then selected the maximum length in each interval of 100 bins in a region between 1 megabase upstream and downstream of a TSS. This step outputted the maximum 200 longest DBP ChIP-seq peaks for each transcript. I modified the original model and extended it to include the data.

In the training of the modified deep learning model, I used the same test, training, and validation data as in the original method: all 2,705 genes encoded on chromosome 1 as test data, training data (22,251 genes), and validation data (2,472 genes) randomly selected from the other genes (Osato and Hamada, 2025; Tasaki, et al., 2020). The DeepLIFT tool was used to calculate the contribution scores of DNA-binding sites of proteins, which are important for predicting gene expression levels (Shrikumar, et al., 2017). The input data for promoters and putative enhancer regions were concatenated into a total of 230 bins of the lengths of 1,310 DBP ChIP-seq peaks, consisting of 30 bins for promoters and 200 bins for putative enhancers, because the DeepLIFT tool 'RevealCancel rule on dense layers and the Rescale rule on conv layers (genomics default)', with high scoring accuracy, cannot be applied to a neural network with multiple input layers. Ubuntu (version 20.04), Python (version 3.8.10), TensorFlow (version 2.6.0), NVIDIA CUDA toolkit (version 11.4.2), GeForce RTX 3080 GPU with 10 GB memory, and Intel Core i7-10700 CPU with 128 GB memory were used for training and prediction, and the original scripts were modified to run on newer versions of the software. The training method was the same as the original, and the training time was two to four hours.

Functional prediction of DNA/RNA-binding proteins

361 To predict the regulatory target genes of DNA/RNA-binding proteins without using
362 data from protein binding assays such as ChIP-seq or eCLIP, among 1,310 DNA-binding
363 proteins in the input data of deep learning, 302 DNA-binding proteins with lower
364 contribution scores for predicting gene expression levels were replaced by other DNA-
365 binding proteins. For the analysis of RNA-binding proteins, 325 DNA-binding proteins
366 were replaced by RNA-binding proteins. The gene symbols of RNA-binding proteins
367 were obtained from the ENPD database (Tak Leung, et al., 2019). The interactions
368 between replaced DNA/RNA-binding proteins and genes regulated by the proteins were
369 estimated based on the co-expression pairs of genes between cells and tissues obtained
370 from the COEXPRESdb (Obayashi, et al., 2023). Gene co-expression is the similarity of
371 expression profiles between a pair of genes. Functionally related genes and genes with
372 the same regulation are generally expressed together between cell and tissue types. If a
373 gene encoding a DNA/RNA-binding protein shows a similar expression pattern with
374 other genes, they are potentially regulated by the protein and functionally related. The co-
375 expression interactions with z score > 2 between the genes encoding other DNA/RNA-
376 binding proteins and the genes other than DNA/RNA-binding proteins were used as input
377 data, replacing those with lower contribution scores. The binding sites of the DNA- and
378 RNA-binding protein were hypothetically located at 2 kb upstream of the TSSs and TSSs
379 of the co-expressed genes, respectively. Deep learning and the DeepLIFT tool with the
380 new input data calculated the contribution scores of the binding sites of DNA/RNA-
381 binding proteins for predicting gene expression levels.

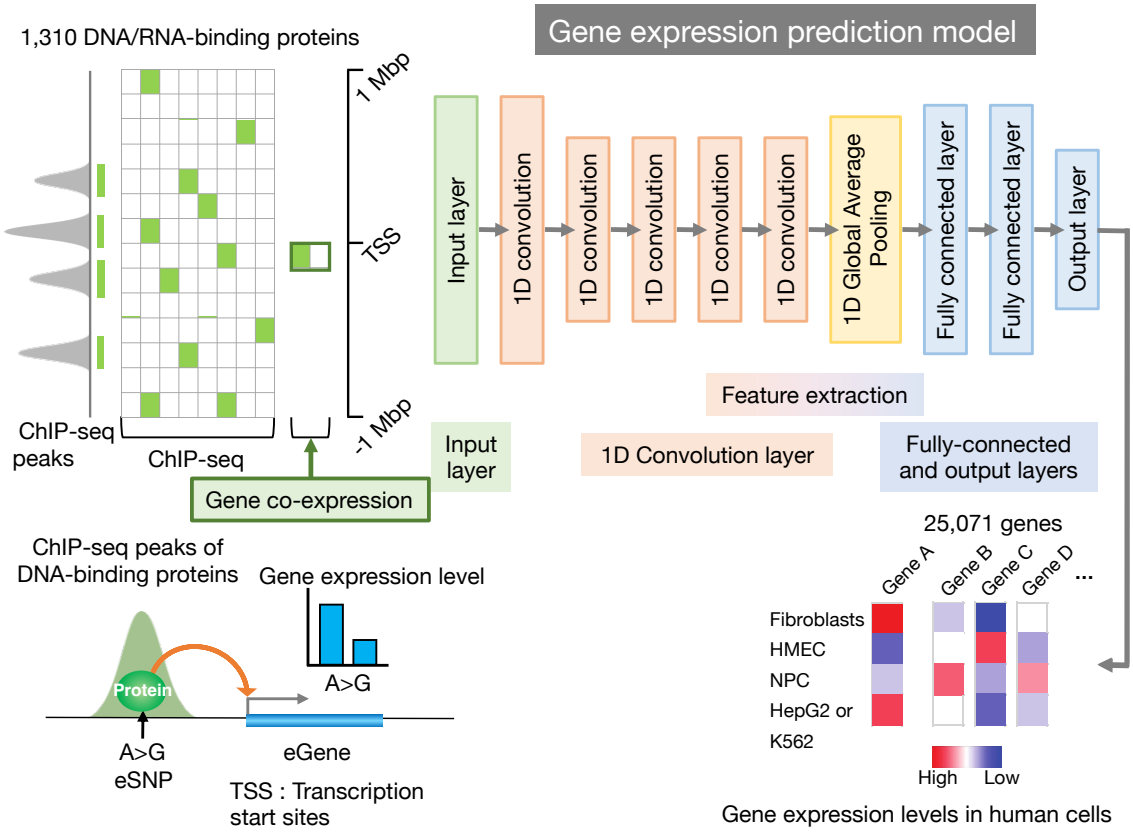
382 For each DNA/RNA-binding protein, I examined the rank of the contribution score
383 of the DNA/RNA-binding site by comparing it with the contribution scores of the other
384 DNA/RNA-binding proteins for each gene. According to the tendency of high or low

385 ranks (i.e., high positive or low negative contribution scores) of DNA/RNA-binding sites
386 of a protein among its regulatory target genes, I selected regulatory target genes with high
387 or low ranks, where the rank within the top five or bottom five for each gene, of
388 DNA/RNA-binding sites of the protein as the primary target genes regulated by the
389 protein. The rank threshold for selecting primary target genes was adjusted for each
390 DNA/RNA-binding protein to detect functional enrichments. To obtain functional
391 enrichments associated with the DNA/RNA-binding protein, the rank threshold was
392 restricted within the top five or bottom five. The functional enrichment analysis of the
393 primary target genes regulated by the DNA/RNA-binding protein was performed using
394 'Statistical overrepresentation test' of Panther DB (https://pantherdb.org/) (Thomas, et al.,
395 2022). The Ensemble Gene IDs of the primary target genes were used as the analyzed
396 gene set, and the Ensemble Gene IDs in the gene expression data for deep leaning were
397 used as the reference gene set. Fisher's exact test was performed with adjustment for the
398 Benjamini-Hochberg false discovery rate.

399 To examine the binding sites of ChIP-seq data and eCLIP data around the
400 regulatory target genes of DNA/RNA-binding proteins predicted by deep learning of gene
401 co-expression data, I selected the DNA-binding sites within ±5 kbp of a TSS of a gene by
402 ChIP-seq assay or the RNA-binding sites between the transcription start and end of a gene
403 by eCLIP assay based on the Human GENCODE v19 gene annotation (Mudge, et al.,
404 2025). For this analysis, ChIP-seq data of 32 DNA-binding proteins and eCLIP data of
405 29 RNA-binding proteins were obtained from the ENCODE portal and the paper
406 (**Supplementary Table S6**) (Van Nostrand, et al., 2020). I also examined the ratios of the
407 number of genes with high and low contribution scores of the DNA/RNA-binding site
408 around genes with the DNA/RNA-binding site by ChIP-seq or eCLIP assays.

409

410 # Figures



411

412 **Figure 1. Overview of prediction of gene expression levels.** The output data of gene
413 expression levels in human cells were predicted from the input data of genomic locations
414 of DNA-binding sites of proteins at promoters and more distant regions. In the machine
415 learning phase, a DNA-binding protein bound to a distant region overlapping an eSNP
416 was associated with a gene (eGene) that was potentially regulated by the DNA-binding
417 protein based on the eSNP and eGene of the eQTL data. Based on the association of each
418 gene, the locations of ChIP-seq peaks of DNA-binding proteins were binned (green box)
419 and summarized in a matrix as input data.

420

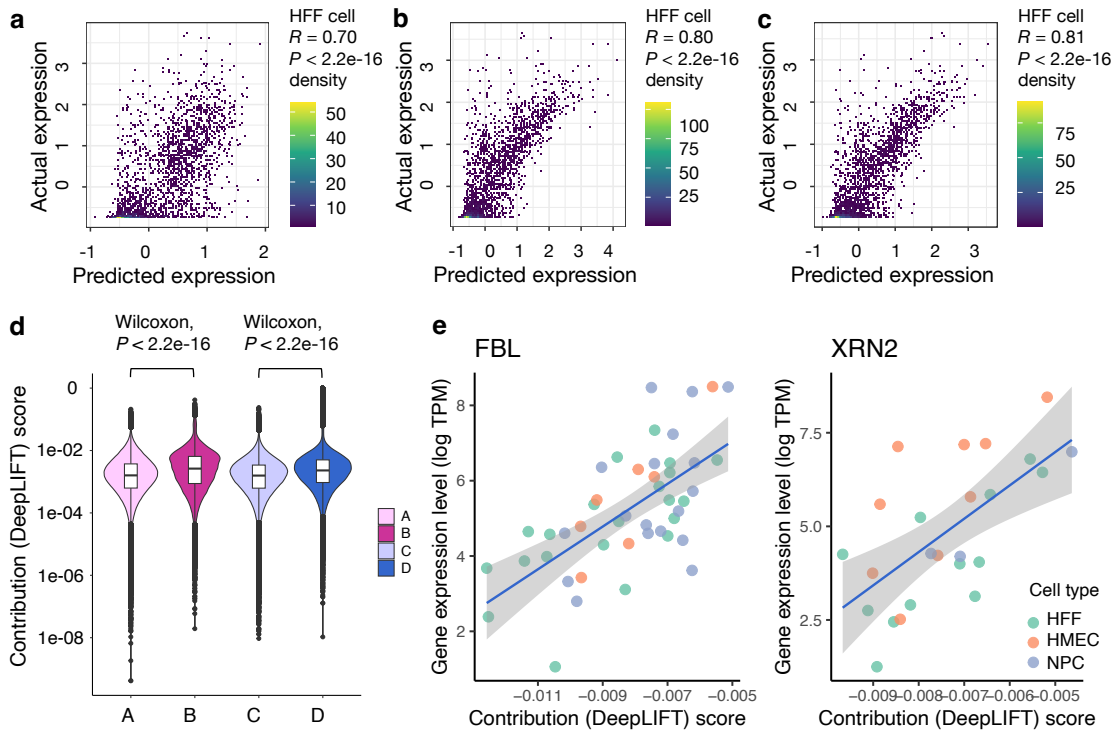


Figure 2. Comparison of actual and predicted gene expression levels.

The predicted log2-TPM (transcripts per million) values of genes were compared with the actual expression levels in HFF cells using Spearman's rank correlation. **a** The input data for deep learning consists of the DNA-binding sites of 1,310 DNA-binding proteins from ChIP-seq data. **b** Among 1,310 proteins, 302 DNA-binding proteins with low contribution scores were replaced by the same number of DNA-binding proteins selected based on gene expression correlation data. **c** Among 1,310 proteins, 325 DNA-binding proteins with low contribution scores were replaced by the same number of RNA-binding proteins selected based on gene expression correlation data. **d** DNA-binding sites based on co-expression data showed higher contribution (DeepLIFT) scores. A and C: Binding sites of DNA-binding proteins were obtained from ChIP-seq data. B and D: Binding sites based on co-expression data between genes. A and B: Positive contribution scores. C and D: Absolute values of negative contribution scores. (Wilcoxon signed-rank test, $P < 2.2e-$

16) **e** Example relationships between the predictive contributions of a regulator and its expression levels across tissues. A line of best fit based on linear regression was depicted with 95 percent confidence intervals.

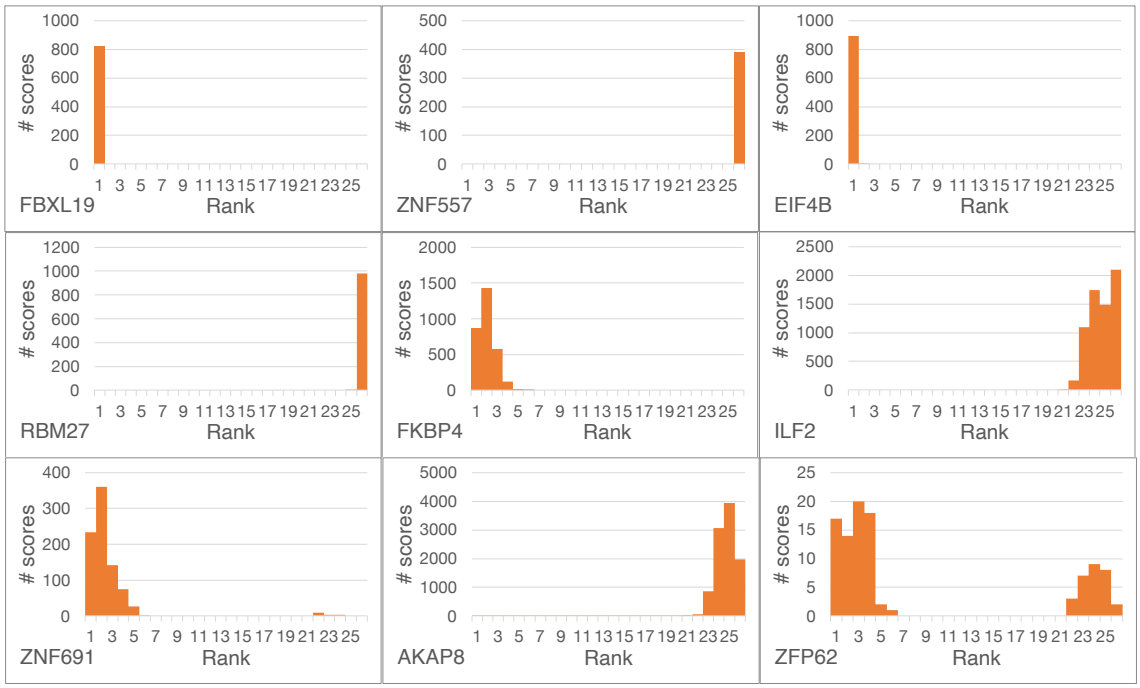


Figure 3. Comparison of predicted target genes with experimental data. The *x* axis shows the rank of the contribution score of a DNA/RNA-binding site for predicting a gene expression level. Contribution scores of DNA/RNA-binding sites were ranked from 1 (high positive contribution scores, left end) to 1,310 (low negative contribution scores, right end), binned with fifty rank intervals. Rank 1 on the *x* axis represents 1 to 50 ranks, rank 2 represents 51 to 100 ranks, and the last rank 26 represents 1,251 to 1,310 ranks. The DNA/RNA-binding sites were hypothetically located in promoters or genes bound by the DNA/RNA-binding protein of each graph. The DNA/RNA-binding sites located in promoters or genes bound by the DNA/RNA-binding protein show higher or lower contribution scores than the DNA-binding sites of the ChIP-seq data as a whole. The *y*

axis shows the number of contribution scores in each bin of fifty rank intervals. Positive contribution scores would correspond to activation and negative scores would correspond to repression of gene expression.

Tables

Table 1. Common functional annotations between a DNA/RNA-binding protein and its putative regulatory target genes.

The functional annotations of DNA/RNA-binding proteins were obtained from the UniProt database. The functional annotations statistically enriched in their putative regulatory target genes were derived from Panther DB. Gray lines separate each annotation.

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

Table 2. Functional enrichments detected in the regulatory target genes of an RNA-binding protein.

The functional enrichment analysis was performed by Panther DB. The table shows the functional annotations statistically enriched in the regulatory target genes of an RNA-binding protein after deep learning based on co-expressed genes, but not statistically enriched in all the co-expressed genes (z score > 2) including the RNA-binding protein. 'Reference' is the number of human genes, 'Analyzed' is the number of human regulatory target genes of an RNA-binding protein, 'Expected' is the number of genes with the

473 functional annotation (e.g. Circadian clock system) estimated from the ratio of genes with
474 the functional annotation in the reference genes, 'Over/Under' stands for
475 overrepresentation and underrepresentation, 'Fold Enrichment' is the ratio of the number
476 of analyzed genes with the functional annotation to the expected number of genes, 'raw
477 P-value' is the P-value of Fisher's exact test, and FDR is the Benjamini-Hochberg false
478 discovery rate of the test.

479

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

480

481

Acknowledgements

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

Author contributions

N.O. designed and performed 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], Osaka University start-up research grant, KIOXIA young scientist research grant at Waseda University; and Sumitomo basic science research projects to NO.

Data availability

Human foreskin fibroblast (HFF) cell RNA-seq data were obtained from GSM941744 in the Gene Expression Omnibus (GEO) database. The RNA-seq data of human mammary epithelial cells (HMEC) were found on ENCODE at UCSC

http://genome.ucsc.edu/encode/ at 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam'. Human neural progenitor cell (NPC) RNA-seq data were obtained from GSM915326. HepG2 and K562 RNA-seq data were obtained from ENCFF670LIE.bam and ENCFF584WML.bam in the ENCODE portal (https://www.encodeproject.org/). Co-expression data of mRNAs were downloaded from COXPRESdb (https://zenodo.org/record/6861444/files/Hsa-u.v22-05.G16651-S245698.combat_pca.subagging.z.d.zip). The list of DNA/RNA-binding proteins were obtained from ENPD database (http://qinlab.sls.cuhk.edu.hk/ENPD/). ChIP-seq data of TFs were downloaded from the v20.06 release of the GTRD database (https://gtrd.biouml.org). eQTL data (GTEx_Analysis_v8_eQTL.tar) were downloaded from the v8 release of the GTEx portal (https://gtexportal.org). eCLIP data of RNA-binding proteins were obtained from ENCSR456FVU in the ENCODE portal and from Supplementary Data 2 in the paper (Van Nostrand EL et al. *Nature* 2020 doi: 10.1038/s41586-020-2077-3).

References

Agarwal, V., *et al.* Massively parallel characterization of transcriptional regulatory elements. *Nature* 2025.

Avsec, Z., *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods* 2021;18(10):1196-1203.

Barrett, T., *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research* 2013;41(Database issue):D991-995.

Berenson, A., *et al.* Paired yeast one-hybrid assays to detect DNA-binding cooperativity and antagonism across transcription factors. *Nat Commun* 2023;14(1):6570.

Burley, S.K. and Kamada, K. Transcription factor complexes. *Curr Opin Struct Biol*

2002;12(2):225-230.
Edgar, R., Domrachev, M. and Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* 2002;30(1):207-210.
Goos, H.*, et al.* Human transcription factor protein interaction networks. *Nat Commun* 2022;13(1):766.
Hinrichs, A.S.*, et al.* The UCSC Genome Browser Database: update 2006. *Nucleic acids research* 2006;34(Database issue):D590-598.
Kim, S., Yu, N.K. and Kaang, B.K. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp Mol Med* 2015;47(6):e166.
Kolmykov, S.*, et al.* GTRD: an integrated view of transcription regulation. *Nucleic acids research* 2021;49(D1):D104-D111.
Ladomery, M. Multifunctional proteins suggest connections between transcriptional and post-transcriptional processes. *Bioessays* 1997;19(10):903-909.
Lee, Y. Roles of circadian clocks in cancer pathogenesis and treatment. *Exp Mol Med* 2021;53(10):1529-1538.
Lemoine, G.G.*, et al.* GWENA: gene co-expression networks analysis and extended modules characterization in a single Bioconductor package. *BMC Bioinformatics* 2021;22(1):267.
Luo, Y.*, et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic acids research* 2020;48(D1):D882-D889.
McLean, C.Y.*, et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* 2010;28(5):495-501.
Miller, H.E. and Bishop, A.J.R. Correlation AnalyzeR: functional predictions from gene co-expression correlations. *BMC Bioinformatics* 2021;22(1):206.
Mudge, J.M.*, et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic acids research* 2025;53(D1):D966-D975.
Obayashi, T.*, et al.* COXPRESdb v8: an animal gene coexpression database navigating from a global view to detailed investigations. *Nucleic acids research* 2023;51(D1):D80-D87.
Osato, N. Characteristics of functional enrichment and gene expression level of human putative transcriptional target genes. *BMC Genomics* 2018;19(Suppl 1):957.
Osato, N. and Hamada, M. Systematic discovery of directional regulatory motifs associated with human insulator sites. *bioRxiv* 2025.
Poelmans, G.*, et al.* AKAPs integrate genetic findings for autism spectrum disorders. *Transl Psychiatry* 2013;3(6):e270.
Puram, R.V.*, et al.* Core Circadian Clock Genes Regulate Leukemia Stem Cells in AML. *Cell* 2016;165(2):303-316.
Raina, P.*, et al.* GeneFriends: gene co-expression databases and tools for humans and model

organisms. *Nucleic acids research* 2023;51(D1):D145-D158.
Shrikumar, A., Greenside, P. and Kundaje, A. Learning Important Features Through Propagating Activation Differences. In: Doina, P. and Yee Whye, T., editors, *Proceedings of the 34th International Conference on Machine Learning*. Proceedings of Machine Learning Research: PMLR; 2017. p. 3145--3153.
Snel, B.*, et al.* STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research* 2000;28(18):3442-3444.
Stuart, J.M.*, et al.* A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)* 2003;302(5643):249-255.
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* 2023;51(D1):D638-D646.
Tak Leung, R.W.*, et al.* ENPD - A Database of Eukaryotic Nucleic Acid Binding Proteins: Linking Gene Regulations to Proteins. *Nucleic acids research* 2019;47(D1):D322-D329.
Tan, X.*, et al.* Novel Insights into the Circadian Rhythms Based on Long Noncoding and Circular RNA Profiling. *Int J Mol Sci* 2024;25(2).
Tasaki, S.*, et al.* Deep learning decodes the principles of differential gene expression. *Nat Mach Intell* 2020;2(7):376-386.
Thomas, P.D.*, et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci* 2022;31(1):8-22.
Tian, W.*, et al.* Bioinformatics analysis of the role of aldolase A in tumor prognosis and immunity. *Sci Rep* 2022;12(1):11632.
UniProt, C. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic acids research* 2025;53(D1):D609-D617.
Van Nostrand, E.L.*, et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* 2020;583(7818):711-719.
Wang, L., Wang, S. and Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics (Oxford, England)* 2012;28(16):2184-2185.
Warde-Farley, D.*, et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research* 2010;38(Web Server issue):W214-220.
Wong, A.K., Krishnan, A. and Troyanskaya, O.G. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic acids research* 2018;46(W1):W65-W70.
Yang, S.*, et al.* COEXPEDIA: exploring biomedical hypotheses via co-expressions associated with medical subject headings (MeSH). *Nucleic acids research* 2017;45(D1):D389-D396.
Yang, Y., Pu, J. and Yang, Y. Glycolysis and chemoresistance in acute myeloid leukemia. *Heliyon*

605 2024;10(15):e35721.
606 Zhang, F., *et al.* Elevated transcriptional levels of aldolase A (ALDOA) associates with cell cycle-
607 related genes in patients with NSCLC and several solid tumors. *BioData Min* 2017;10:6.
608 Zhu, Q., *et al.* Targeted exploration and analysis of large cross-platform human transcriptomic
609 compendia. *Nature methods* 2015;12(3):211-214, 213 p following 214.
610 Zlacka, J. and Zeman, M. Glycolysis under Circadian Control. *Int J Mol Sci* 2021;22(24).

611