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

Experimental identification of DNA/RNA-binding sites of proteins is difficult for some proteins. Deep learning methods have been developed to accurately predict gene expression levels from DNA-binding sites of transcription factors identified by ChIP-seq assay. Instead of the ChIP-seq data, if the interactions between transcription factors and other genes are obtained from the correlation of mRNA expression levels in different cell and tissue types, the interactions can be used as the input data for the deep learning. The deep learning method would select important interactions to accurately predict gene expression levels by learning from the data. The selected interactions potentially include the same or similar interactions obtained from experimental assays such as ChIP-seq. I replaced some of the input data for deep learning with the sites of the same number of different DNA/RNA-binding proteins based on co-expression data. Interestingly, the correlation coefficient between actual and predicted gene expression levels increased from 0.70 to 0.80 using co-expression data. The contribution scores of the input data selected from the gene expression correlation were higher than the original, and showed high positive or low negative contribution scores. Some functional annotations that were statistically overrepresented in the predicted target genes correspond to the known and associated functions of the 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 transcription factors, DNA/RNA-binding proteins, 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. When 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 remove the binding sites of RNA-binding proteins and miRNAs, the correlation coefficient was 0.70. Although RNA-binding proteins and miRNAs bind to mRNAs and regulate their expression, enhancer-promoter interactions are also involved in the regulation of gene expression. Insulator proteins act as a barrier of enhancer–promoter interactions. CTCF and cohesin are considered to be major insulators in vertebrates. With the modified deep learning method, I found that the directional bias of CTCF binding sites identified by the statistical test of contribution scores was consistently proportional to the bias observed at chromatin interaction sites by experimental analysis (76.4% forward-reverse, 11.0% forward-forward, 10.3% reverse-reverse, 2.3% reverse-forward orientations), and discovered other insulator-associated DNA-binding proteins, most of which show the directional bias of DNA-binding sites in human genomes (Osato and Hamada, 2025).

Based on these prediction accuracies, I plan to predict DNA-binding proteins that regulate gene expression levels without using the DNA-binding sites of proteins derived

from their ChIP-seq experimental data. Instead, I intend to use the co-expression data of genes between different cell and tissue types in a public database. If a gene encoding a DNA/RNA-binding protein shows the correlations of gene expression levels with other genes between cell and tissue types, the DNA/RNA-binding protein is expected to regulate the correlation of genes directly or indirectly. Then, deep learning can learn and predict the regulations that are important for predicting gene expression levels.

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. 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. The correlation of gene expression would be regulated by direct and/or non-direct interactions of molecules. Some of the molecules may be elusive to detect by experimental analysis and hard to predict by other computational analyses. Gene expression correlation is potentially associated with regulations that have not been identified.

Here, I replace a relatively small fraction of the input data of DNA-binding sites with a low contribution score for predicting gene expression levels with DNA-binding sites based on gene expression correlation. Deep learning learns to predict gene expression levels and selects interactions between a DNA-binding site and a gene that contribute to the prediction. The regulatory target genes of a transcription factor include functionally related genes. Thus, the functional enrichment of genes is potentially discovered, even though the prediction of regulatory target genes may not be precise

enough to achieve a research purpose. I compared the overlap between the predicted regulatory target genes of a DNA/RNA-binding proteins and the DNA/RNA-binding sites by experimental assays such as ChIP-seq and eCLIP. Then, I performed functional enrichment analyses of the predicted regulatory target genes of ten DNA/RNA-binding proteins and found that the enriched functional annotations correspond to the known functional annotations in the UniProt database.

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

Results

To examine the effect of the interaction data of co-expressed genes on the prediction of gene expression levels, among 1,310 DNA-binding proteins in the input data of deep learning, 302 DNA-binding proteins with low contribution scores were replaced with the interaction data of co-expressed genes encoding DNA-binding proteins and the mRNAs other than DNA/RNA-binding proteins. The DNA-binding proteins encoded by the co-expressed genes were randomly selected from the gene symbols of DNA-binding proteins in the ENPD database (Tak Leung, et al., 2019), and in the other databases of DNA-binding proteins used for the analysis of DNA-binding motifs. Interestingly, the correlation coefficient between actual and predicted gene expression

121 levels increased from 0.70 to 0.80 in human foreskin fibroblast (HFF) cells using co-
122 expression data (**Fig. 2**). The contribution scores of the input data selected from the gene
123 expression correlation were larger than the original (**Table 1**). Then, among 1,310 DNA-
124 binding proteins, 325 DNA-binding proteins with low contribution scores were replaced
125 with the interaction data of co-expressed genes encoding RNA-binding proteins and other
126 mRNAs. The RNA-binding proteins encoded by the co-expressed genes were randomly
127 selected from the gene symbols of RNA-binding proteins in the ENPD database. The
128 correlation coefficient increased from 0.70 to 0.81 using co-expression data as well. These
129 results suggest that the interactions of co-expression of genes contribute to the prediction
130 of gene expression levels.

131 Then, I compared the DNA/RNA-binding sites selected based on their contribution
132 scores with the DNA-binding sites by ChIP-seq assay and the RNA-binding sites by
133 eCLIP assay around genes. The contribution scores of the DNA/RNA-binding sites were
134 consistent with the DNA/RNA-binding sites of the experimental data (ChIP-seq data of
135 32 DNA-binding proteins and eCLIP data of 29 RNA-binding proteins) (**Fig. 3 and**
136 **Supplementary Fig. S1**). High positive and low negative contribution scores were
137 observed for the binding regulatory target genes of DNA/RNA-binding proteins derived
138 from the gene co-expression data. High and low contribution scores seem to correspond
139 to regulatory functions of activation and repression of gene expression from the
140 explanation of contribution score on the GitHub webpage of DeepLIFT (Shrikumar, et al.,
141 2017).

142 The ratios of the number of high-ranking promoters and promoters+enhancers with
143 a ChIP-seq peak to the number of all high-ranking promoters and promoters+enhancers
144 were respectively calculated (**Table 2 and Supplementary Table S1**). Promoters were

145 defined as genomic regions around ±5 kbp from TSS, and promoters+enhancers were
146 defined as genomic regions around ±100 kbp from TSS, because 84% of enhancers are
147 estimated to be located within the distance of TSS based on the proportions of high-
148 confidence enhancer-gene pairs. (Gasperini, et al., 2019) (Avsec, et al., 2021). The mean
149 and median ratios of the number of high-ranking promoters+enhancers (±100 kbp of TSS)
150 with a ChIP-seq peak to the number of all high-ranking promoters+enhancers were 0.69
151 and 0.76 (Ration A2 / B2). The mean and median ratios of the number of high-ranking
152 promoters (±5 kbp of TSS) with a ChIP-seq peak to the number of all high-ranking
153 promoters were 0.33 and 0.34 (Ration A2 / B2). Some DNA-binding proteins such as
154 FBXL19, POGK, SPEN, THAP7, ZNF142, and ZNF800 showed high ratios (> 0.5) in
155 promoters. Some ChIP-seq peaks of DNA-binding proteins may not be associated with
156 the regulation of gene expression. Thus, the DNA-binding sites of proteins with high
157 contribution scores will not completely overlap with the ChIP-seq peaks of the same
158 DNA-binding proteins.

159 The transcriptional target genes of a transcription factor include associated
160 functions of genes. The functional enrichment of genes would be useful to predict the
161 function of a transcription factor and to confirm that the transcriptional target genes are
162 not randomly selected. For the functional enrichment analysis of regulatory target genes
163 of DNA/RNA-binding proteins, I used the statistical overrepresentation test of the gene
164 list analysis of the Panther DB (https://pantherdb.org/) (Thomas, et al., 2022). The test
165 uses functional annotation databases such as PANTHER Pathways, Protein Class, GO-
166 Slim, GO, and Reactome Pathways. The functional annotations enriched in putative
167 regulatory target genes of a DNA/RNA-binding protein were compared with those of the
168 DNA/RNA-binding protein obtained from the UniProt database (e.g.

169 https://www.uniprot.org/uniprotkb/Q9Y2L8/entry) (UniProt, 2025). The ten examples of
170 DNA/RNA-binding proteins are shown in **Table 3** and **Supplementary Table S2**. The
171 putative functional annotations of the DNA/RNA-binding proteins correspond to and are
172 associated with those in the UniProt database and publications. Notably, among the ten
173 examples, AKAP8 has cell cycle-related functional annotations such as "GO:0000278
174 mitotic cell cycle" and "GO:0044839 cell cycle G2/M phase transition" in the UniProt
175 database (https://www.uniprot.org/uniprotkb/O43823/entry). In this study, AKAP8 was
176 suggested to be involved in circadian clock systems. By searching Google and the
177 PubMed database, I found the following papers showing that AKAP8 is involved in
178 circadian rhythms: "But the genes, such as NFIL3, AKAP8, and SALL3, highly correlated
179 with known molecular clock genes (CRY1) can be identified as the candidate circadian
180 genes." and " AKAP5 and AKAP8 are both involved in the biosynthesis of the circadian
181 hormone melatonin, which is markedly disturbed in people with ASDs." (Tan, et al.,
182 2024) (Poelmans, et al., 2013). These results demonstrate that the method and analysis in
183 this study detected known and novel functional annotations associated with DNA/RNA-
184 binding proteins.
185

186 ## Discussion

187 The expression levels of human genes in cell types have been predicted using deep
188 learning based on the DNA-binding sites of proteins. Here, I applied this method to
189 predict unknown interactions between a DNA/RNA-binding protein and genes based on
190 gene expression correlations between cell types. Gene expression correlations were
191 statistically analyzed using gene expression data from RNA-seq experiments in cell and
192 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.

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, which are different mechanisms to affect gene expression from the direct regulation of genes by transcription factors. In the deep learning analysis of insulator-associated DNA-

binding proteins, we used eQTL data to find DNA-binding sites of proteins that affect gene expression levels. However, in this study, I did not use the DNA-binding motif sequences of proteins and open chromatin regions by DNase-seq or ATAC-seq assay to identify their DNA-binding sites because the purpose of this study is to investigate DNA-binding proteins whose DNA-binding motifs are unknown. Moreover, the association between eSNP and eGene of eQTL data may include indirect interactions. The contribution scores of DNA-binding sites of proteins selected based on gene expression correlation showed higher or lower scores, suggesting that indirect interactions between DNA-binding proteins and other genes would affect the scores.

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

When I replaced all the input data of the deep learning analysis with the interactions between DNA/RNA-binding proteins and other genes based on gene expression correlation, the number of functional enrichments of predicted regulatory target genes of

241 a DNA/RNA-binding protein was reduced, and the result of the functional enrichment
242 analysis did not include the same functional annotations as the known functional
243 annotations of the DNA/RNA-binding protein in the UniProt database. Thus, the high
244 replacement ratio of the input data would lead to learning error-prone interactions
245 between a DNA/RNA-binding protein and regulatory target genes. In my previous study,
246 when the regulatory target genes of a DNA-binding protein were replaced by randomly
247 selected genes, the number of functional enrichments decreased as the number of
248 randomly selected genes increased. The number of functional enrichments reflects the
249 accuracy of the regulatory target genes (Osato, 2018).

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

263 The deep learning model, model parameters, and analysis methods used in this
264 study could be adapted and improved for the purpose of this study. I examined a small

265 number of DNA/RNA-binding proteins in a few cell types in this study, and would be
266 able to predict the functional roles of a larger number of DNA/RNA-binding proteins in
267 cell and tissue types. The results of the analysis would be useful to understand and study
268 the biological and molecular functions of unknown functions of DNA/RNA-binding
269 proteins and their regulatory target genes.
270

271 # Methods

272 **Deep learning analysis**

273 I modified the deep learning method for predicting gene expression levels among
274 cell and tissue types based on the DNA-binding sites of transcription factors and other
275 regulators, and included the DNA-binding sites of transcription factors in regions far
276 upstream and downstream of TSSs (Tasaki, et al., 2020) (Osato and Hamada, 2025) (**Fig.**
277 **1**). To reduce the heterogeneity of cells in a tissue, I used gene expression data of cell
278 types. The gene expression data of different cell types were obtained from public
279 databases: HFFs (GSM941744), human mammary epithelial cells (HMECs)
280 (Encyclopedia of DNA Elements at UCSC http://genome.ucsc.edu/encode/,
281 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam') (Barrett, et al., 2013) (Edgar,
282 et al., 2002), neural progenitor cells (NPCs) (GSM915326, ENCFF529SIO), human
283 hepatoma cell line (HepG2) (ENCFF670LIE), and human leukemia cell line (K562)
284 (ENCFF584WML). FPKM values of gene expression levels were calculated based on
285 GENCODE V19 gene annotations in the ENCODE portal database
286 (https://www.encodeporoject.org/) (Luo, et al., 2020) (Mudge, et al., 2025) using the
287 RSeQC tool (http://rseqc.sourceforge.net/) (Wang, et al., 2012). The FPKM values were
288 transformed to log2-fold with the addition of 0.25 to avoid negative infinity results. Log2-

289 fold changes relative to the median log2-FPKM across the four cell types were calculated
290 according to the original deep learning method (transcriptome data processing as
291 described by Tasaki et al.) (Tasaki, et al., 2020) (Osato and Hamada, 2025).

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

306 The original model of the input data of DNA-binding sites in promoter regions was
307 used for the deep learning analysis: the lengths of a DBP ChIP-seq peak in each 100-base
308 bin in a region between 2 kbp upstream and 1 kbp downstream of the TSS of each gene
309 or transcript (**Fig. 1**). To add the input data of DNA-binding sites of DBPs in more
310 upstream and downstream regions of a TSS and reduce the number of bins, I first
311 computed the length of a DBP ChIP-seq peak in a 100-base bin and then selected the
312 maximum length in each interval of 100 bins in a region between 1 megabase upstream

313 and downstream of a TSS. This step outputted the maximum 200 longest DBP ChIP-seq
314 peaks for each transcript. I modified the original model and extended it to include the
315 data.

316 In the training of the modified deep learning model, I used the same test, training,
317 and validation data as in the original method: all 2,705 genes encoded on chromosome 1
318 as test data, training data (22,251 genes), and validation data (2,472 genes) randomly
319 selected from the other genes (Tasaki, et al., 2020) (Osato and Hamada, 2025). The
320 DeepLIFT tool was used to calculate the contribution scores of DNA-binding sites of
321 proteins, which are important for predicting gene expression levels (Shrikumar, et al.,
322 2017). The input data for promoters and putative enhancer regions were concatenated into
323 a total of 230 bins of the lengths of 1,310 DBP ChIP-seq peaks, consisting of 30 bins for
324 promoters and 200 bins for putative enhancers, because the DeepLIFT tool 'RevealCancel
325 rule on dense layers and the Rescale rule on conv layers (genomics default)', with high
326 scoring accuracy, cannot be applied to a neural network with multiple input layers.
327 Ubuntu (version 20.04), Python (version 3.8.10), TensorFlow (version 2.6.0), NVIDIA
328 CUDA toolkit (version 11.4.2), GeForce RTX 3080 GPU with 10 GB memory, and Intel
329 Core i7-10700 CPU with 128 GB memory were used for training and prediction, and the
330 original scripts were modified to run on newer versions of the software. The training
331 method was the same as the original, and the training time was two to four hours.
332

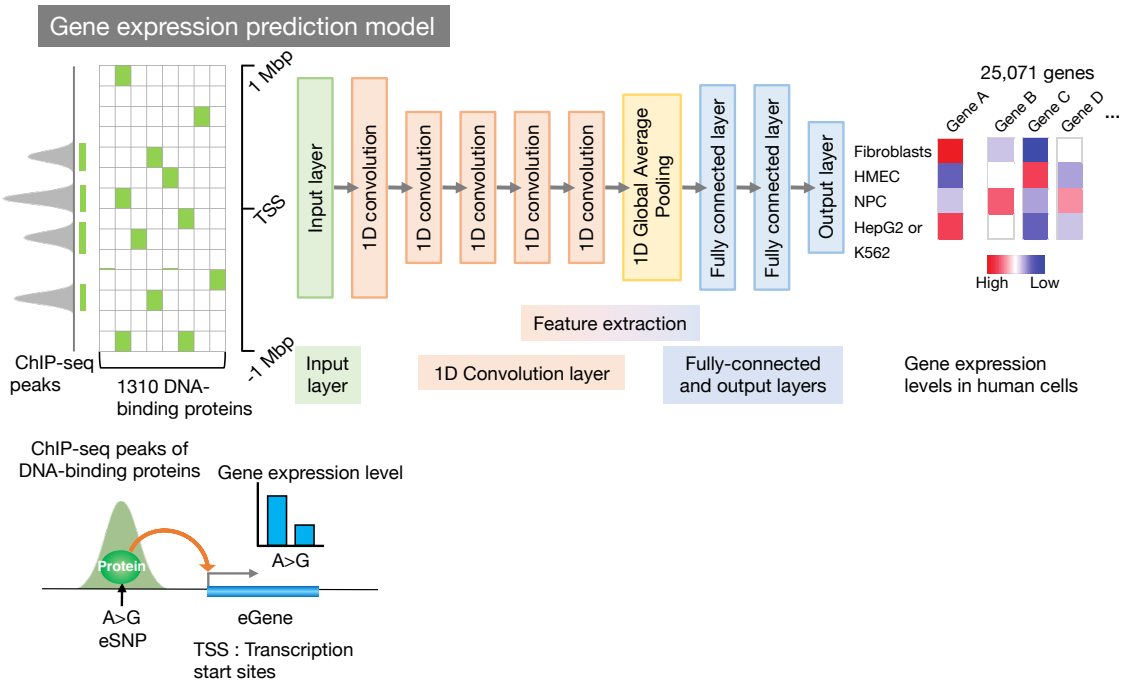
333 **Functional prediction of DNA/RNA-binding proteins**

334 To predict the regulatory target genes of DNA/RNA-binding proteins without using
335 data from protein binding assays such as ChIP-seq or eCLIP, among 1,310 DNA-binding
336 proteins in the input data of deep learning, 302 DNA-binding proteins with lower

contribution scores for predicting gene expression levels were replaced by other DNA-binding proteins. For the analysis of RNA-binding proteins, 325 DNA-binding proteins were replaced by RNA-binding proteins. The gene symbols of RNA-binding proteins were obtained from the ENPD database (Tak Leung, et al., 2019). The interactions between other DNA/RNA-binding proteins and genes regulated by the proteins were estimated based on the co-expression pairs of genes between cells and tissues obtained from the COEXPRESdb (Obayashi, et al., 2023). Gene co-expression is the similarity of expression profiles between a pair of genes. Functionally related genes and genes with the same regulation are generally expressed together between cell and tissue types. If a gene encoding a DNA/RNA-binding protein shows a similar expression pattern with other genes, they are potentially regulated by the protein and functionally related. The co-expression interactions with z score ≥ 2 between the genes encoding other DNA/RNA-binding proteins and the genes other than DNA/RNA-binding proteins were used as input data, replacing those with lower contribution scores. The binding sites of the DNA- and RNA-binding protein were hypothetically located at 2 kb upstream of the TSSs and TSSs of the co-expressed genes, respectively. Deep learning and the DeepLIFT tool with the new input data calculated the contribution scores of the binding sites of DNA/RNA-binding proteins for predicting gene expression levels. For each DNA/RNA-binding protein, I examined the rank of the contribution score of the DNA/RNA-binding site by comparing it with the contribution scores of the other DNA/RNA-binding proteins for each gene. Functional enrichment analysis of putative regulatory target genes of the DNA/RNA-binding protein was performed by the Panther DB (https://pantherdb.org/) (Thomas, et al., 2022). To examine the DNA/RNA-binding sites associated with the binding sites of ChIP-seq data and eCLIP data, I selected the DNA-binding sites within

361 ±5 kbp of a TSS of a gene by ChIP-seq assay or the RNA-binding sites between the
362 transcription start and end of a gene by eCLIP assay based on the Human GENCODE
363 v19 gene annotation (Mudge, et al., 2025). For this analysis, ChIP-seq data of 32 DNA-
364 binding proteins and eCLIP data of 29 RNA-binding proteins were obtained from the
365 ENCODE portal and the paper (**Supplementary Table S3**) (Van Nostrand, et al., 2020).
366 I also examined the ratios of the number of genes with high and low contribution scores
367 of the DNA/RNA-binding site around genes with the DNA/RNA-binding site by ChIP-
368 seq or eCLIP assays.
369

370 ## Figures



371

372 **Figure 1. Overview of prediction of gene expression levels.** The output data of gene
373 expression levels in human cells were predicted from the input data of genomic locations
374 of DNA-binding sites of proteins at promoters and more distant regions. In the machine
375 learning phase, a DNA-binding protein bound to a distant region overlapping an eSNP

was associated with a gene (eGene) that was potentially regulated by the DNA-binding protein based on the eSNP and eGene of the eQTL data. Based on the association of each gene, the locations of ChIP-seq peaks of DNA-binding proteins were binned (green box) and summarized in a matrix as input data.

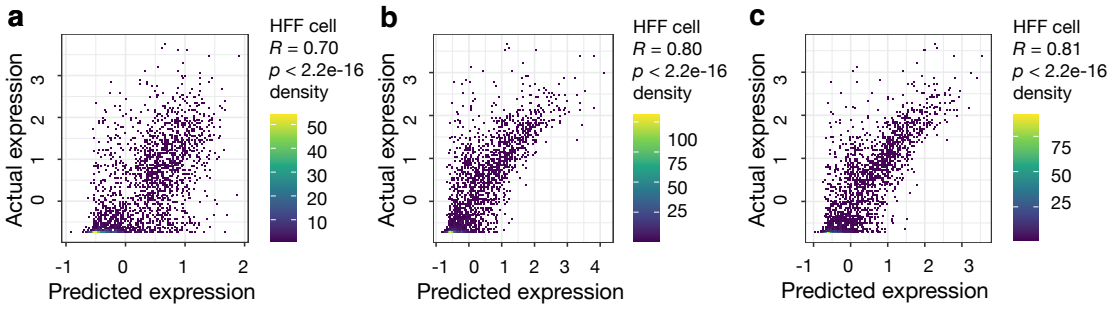


Figure 2. Comparison of actual and predicted gene expression levels.

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.

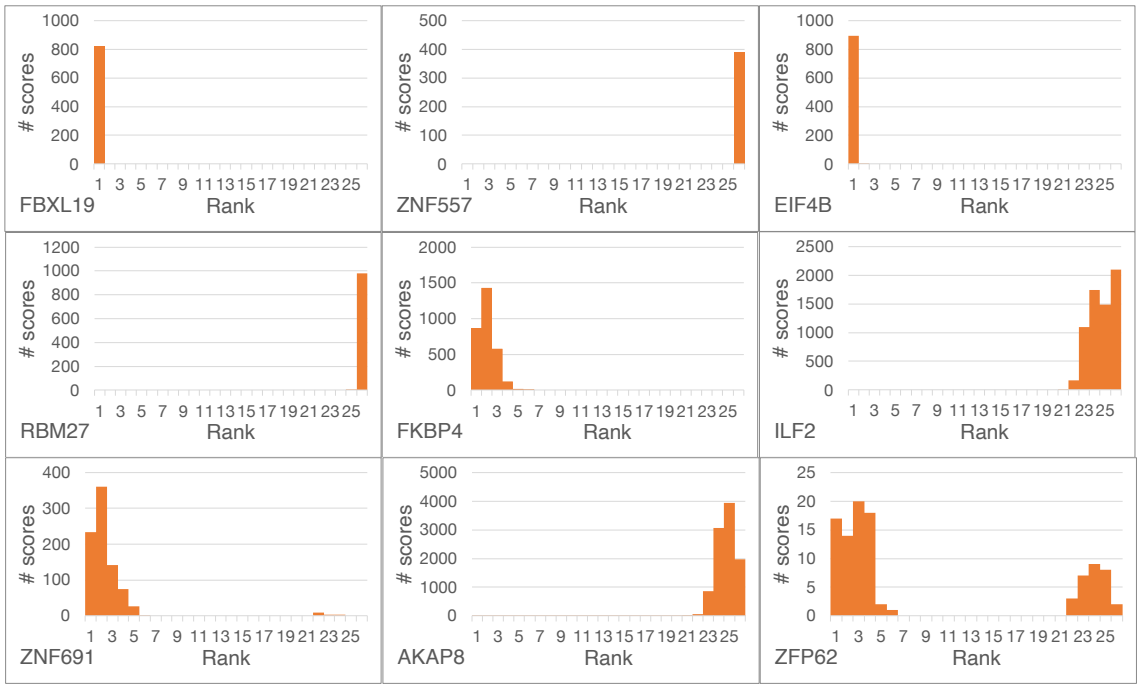


Figure 3. Comparison of predicted target genes with experimental data. The X-axis shows the rank of the contribution score of a DNA/RNA-binding site for predicting a gene expression level. Contribution scores of DNA/RNA-binding sites were ranked from 1 (high positive contribution scores, left end) to 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.

407

408 # Tables

409 **Table 1. Contribution scores of DNA/RNA-binding sites increased with co-**

410 **expression of genes.**

DNA-binding proteins (HepG2 cell)			
Correlation genes		Non-correlation genes	
	Plus score		Plus score
Mean	0.005673	Mean	0.003391
Median	0.003829	Median	0.001893
	Minus score		Minus score
Mean	-0.017000	Mean	-0.002502
Median	-0.002286	Median	-0.001583

RNA-binding proteins (K562 cell)			
Correlation genes		Non-correlation genes	
	Plus score		Plus score
Mean	0.004916	Mean	0.002962
Median	0.002422	Median	0.001689
	Minus score		Minus score
Mean	-0.011987	Mean	-0.002485
Median	-0.001521	Median	-0.001414

411

412

413 **Table 2. Ratio of DNA-binding sites of a protein with high contribution scores to**

414 **ChIP-seq DNA-binding sites.**

DBP	# High-ranking regulatoly target genes with a ChIP-seq peak (A1)	# High-ranking regulatoly target genes (B1)	Ratio (A1 / B1)	# High-ranking regulatoly target genes with a ChIP-seq peak (A2)	# High-ranking regulatoly target genes (B2)	Ratio (A2 / B2)
AKAP8	297	590	0.503	2001	3609	0.554
BAZ2A	20	61	0.328	402	749	0.537
CAMTA2	187	250	0.748	625	784	0.797
DMTF1	74	207	0.357	188	453	0.415
FBXL19	595	687	0.866	823	937	0.878
FOXJ3	27	38	0.711	84	120	0.700
FUBP1	91	146	0.623	131	206	0.636
MATR3	0	53	0.000	25	568	0.044
MTF1	34	89	0.382	147	353	0.416
POGK	246	274	0.898	746	786	0.949
SPEN	938	964	0.973	980	1010	0.970
THAP4	743	907	0.819	768	937	0.820
THAP7	564	638	0.884	959	1057	0.907
TOE1	0	3	0.000	12	43	0.279
ZBTB34	680	930	0.731	700	957	0.731
ZC3H13	62	102	0.608	380	520	0.731
ZC3H4	631	934	0.676	631	934	0.676
ZFP62	24	139	0.173	61	302	0.202
ZKSCAN5	506	1062	0.476	511	1069	0.478
ZNF12	253	315	0.803	656	798	0.822
ZNF142	107	125	0.856	604	673	0.897
ZNF276	174	212	0.821	531	603	0.881
ZNF446	766	858	0.893	785	881	0.891
ZNF526	46	72	0.639	379	515	0.736
ZNF557	643	908	0.708	643	908	0.708
ZNF576	0	2	0.000	21	33	0.636
ZNF646	687	904	0.760	765	1000	0.765
ZNF691	110	131	0.840	305	351	0.869
ZNF786	29	38	0.763	199	262	0.760
ZNF800	274	357	0.768	706	883	0.800
ZNF865	40	44	0.909	181	198	0.914
ZSCAN25	18	26	0.692	97	124	0.782

The ratio of the number of high-ranking promoters+enhancers (±100 kbp of TSS) with a ChIP-seq peak to the number of all high-ranking promoters+enhancers was calculated. Many hypothetical DNA-binding sites with high contribution scores overlapped with the ChIP-seq peaks of the same DNA-binding protein around the same regulatory target genes (±100 kbp of TSS). Among a total of 1,310 DNA-binding proteins, higher than 20

421 or lower than 1,290 ranks (A1 and B1) and higher than 50 or lower than 1,260 ranks (A2
422 and B2).
423
424 **Table 3. Common functional annotations between a DNA/RNA-binding protein and**
425 **its putative regulatory target genes.**
426 The functional annotations of DNA/RNA-binding proteins were obtained from the
427 UniProt database. The functional annotations statistically enriched in their putative
428 regulatory target genes were derived from the Panther DB. Gray lines separate each
429 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

430

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

455 http://genome.ucsc.edu/encode/ at
456 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam'. Human neural progenitor cell
457 (NPC) RNA-seq data were obtained from GSM915326. HepG2 and K562 RNA-seq data
458 were obtained from ENCFF670LIE.bam and ENCFF584WML.bam in the ENCODE
459 portal (https://www.encodeproject.org/). Co-expression data of mRNAs were
460 downloaded from COXPRESdb (https://zenodo.org/record/6861444/files/Hsa-u.v22-
461 05.G16651-S245698.combat_pca.subagging.z.d.zip). The list of DNA/RNA-binding
462 proteins were obtained from ENPD database (http://qinlab.sls.cuhk.edu.hk/ENPD/).
463 ChIP-seq data of TFs were downloaded from the v20.06 release of the GTRD database
464 (https://gtrd.biouml.org). eQTL data (GTEx_Analysis_v8_eQTL.tar) were downloaded
465 from the v8 release of the GTEx portal (https://gtexportal.org). eCLIP data of RNA-
466 binding proteins were obtained from ENCSR456FVU in the ENCODE portal and from
467 Supplementary Data 2 in the paper (Van Nostrand EL et al. *Nature* 2020 doi:
468 10.1038/s41586-020-2077-3).
469
470
471 **References**
472
473 Agarwal, V., *et al.* Massively parallel characterization of transcriptional regulatory elements.
474 *Nature* 2025.
475 Avsec, Z., *et al.* Effective gene expression prediction from sequence by integrating long-range
476 interactions. *Nature methods* 2021;18(10):1196-1203.
477 Barrett, T., *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids*
478 *research* 2013;41(Database issue):D991-995.
479 Berenson, A., *et al.* Paired yeast one-hybrid assays to detect DNA-binding cooperativity and
480 antagonism across transcription factors. *Nat Commun* 2023;14(1):6570.
481 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.
Gasperini, M., *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* 2019;176(6):1516.
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.
Luo, Y., *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic acids research* 2020;48(D1):D882-D889.
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.
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.
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.

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.