PNAS
www.pnas.org

Supplementary Information for

A Machine Learning-based Framework For Modeling Transcription Elongation

Peiyuan Feng, An Xiao, Meng Fang, Fangping Wan, Shuya Li, Peng Lang, Dan Zhao, and Jianyang Zeng

Jianyang Zeng
E-mail: zengjy321@tsinghua.edu.cn
Dan Zhao
E-mail: zhaodan2018@tsinghua.edu.cn

This PDF file includes:

Supporting Information Text

Sensitivity analyses of the criteria for defining Pol II pausing sites

The main goal of our criteria is to ensure the chosen sites to be true positive samples as much as possible. Specifically, to obtain a set of positive samples with high precision, we wanted to exclude those candidate samples that were highly sensitive to the coverage of NET-seq data as such sensitivity may indicate sequencing noise in NET-seq experiments. We first plotted the numbers of Pol II pausing sites (i.e., positive samples) under different minimum read counts in HeLa S3 and HEK293T cell lines (Fig. S1A). We observed that the number of chosen Pol II pausing sites changed drastically when setting the minimum read counts to 1, 2, 3 and 4, indicating that a large number of sensitive (potentially false positive) samples were selected for these choices. On the contrary, the change was relatively smooth when changing the minimum read counts from 4 to 5. Thus, setting the minimum read count to 4 would result in a reliable set of positive samples. We further exmained the genomic coverage of positive samples. By plotting the ratios of genes that had at least one Pol II pausing site under different minimum read counts in the datasets used in our study (Fig. S1B), we found that the ratios of the recalled expressed genes (i.e., expressed genes that contained at least one Pol II pausing site selected according to our criteria) decreased from over 90% to about 72% and 68% in HeLa S3 and HEK29T cell lines, respectively, when changing the minimum read counts from 1 to 4. Although there was a trade-off between a high precision positive set and a good genomic coverage, the above results demonstrated that choosing 4 as the minimum read count can still lead to a decent genomic coverage. In summary, our criteria is highly sensitive and can obtain high-quality dataset of Pol II pausing sites.

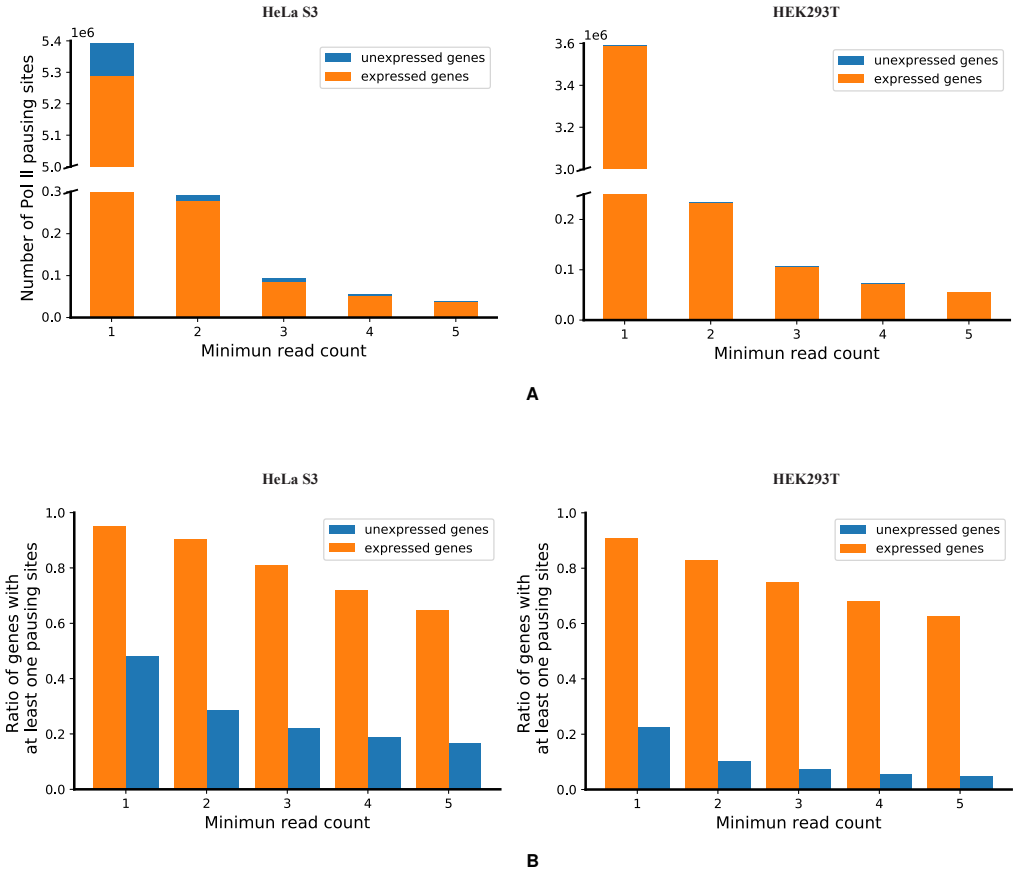


Fig. S1. Sensitivity analyses on the criteria used to define Pol II pausing sites. (A) Numbers of Pol II pausing sites in both expressed and unexpressed genes under different minimum read counts in HeLa S3 and HEK293T cell lines. (B) Ratios of both expressed and unexpressed genes with at least one Pol II pausing site according to minimum read counts.

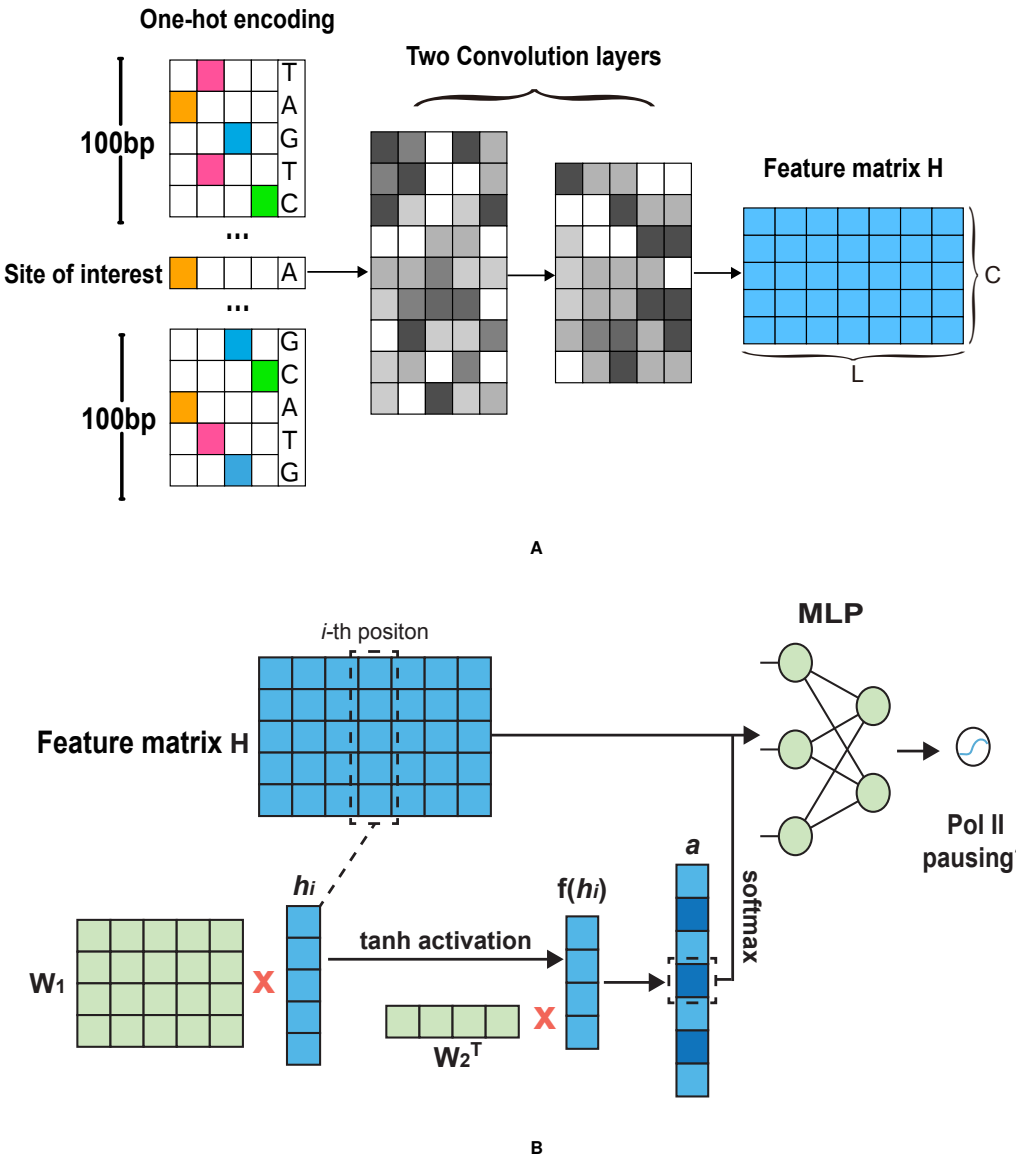
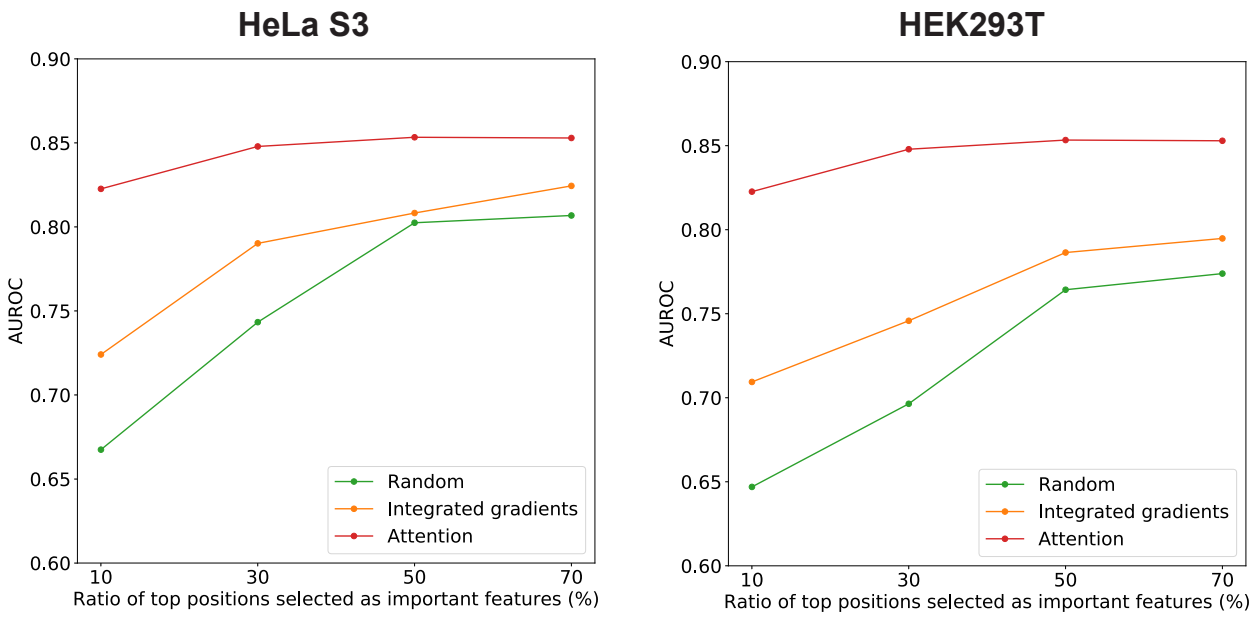
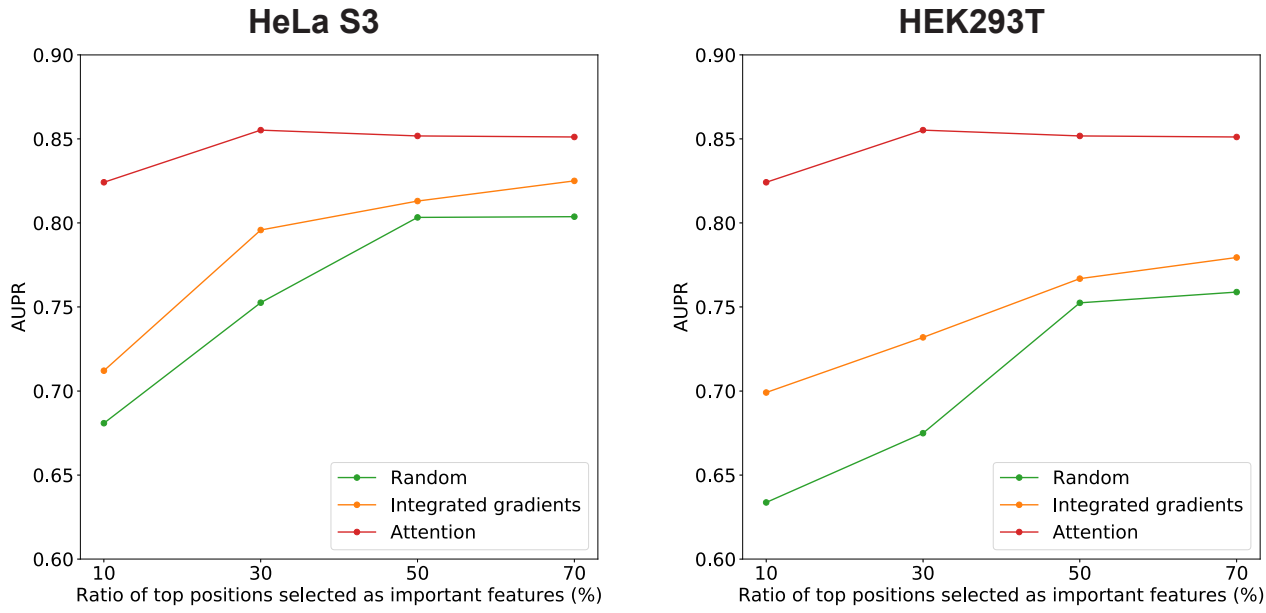


Fig. S2. A detailed illustration of the deep learning framework employed in PEPMAN. (A) The input contextual sequence is one-hot encoded and passed through a two-layer convolutional neural network, resulting in an L-by-C feature matrix $\boldsymbol{H}$. (B) The feature matrix $\boldsymbol{H}$ is used to calculate an attention vector $\boldsymbol{a}$, storing the attributions of individual positions to the final prediction. After that, $\boldsymbol{H}$ is multiplied with $\boldsymbol{a}$ and then fed into a multi-layer perceptron (MLP) network to obtain the final output. See the main text for more details.



A



B

Fig. S3. Performance evaluation of different feature attribution methods with respect to the fractions of the top positions selected as important features in HeLa S3 and HEK293T cell lines, respectively. (A) Area under receiver-operating characteristic (AUROC) scores. (B) Area under precision recall (AUPR) scores.

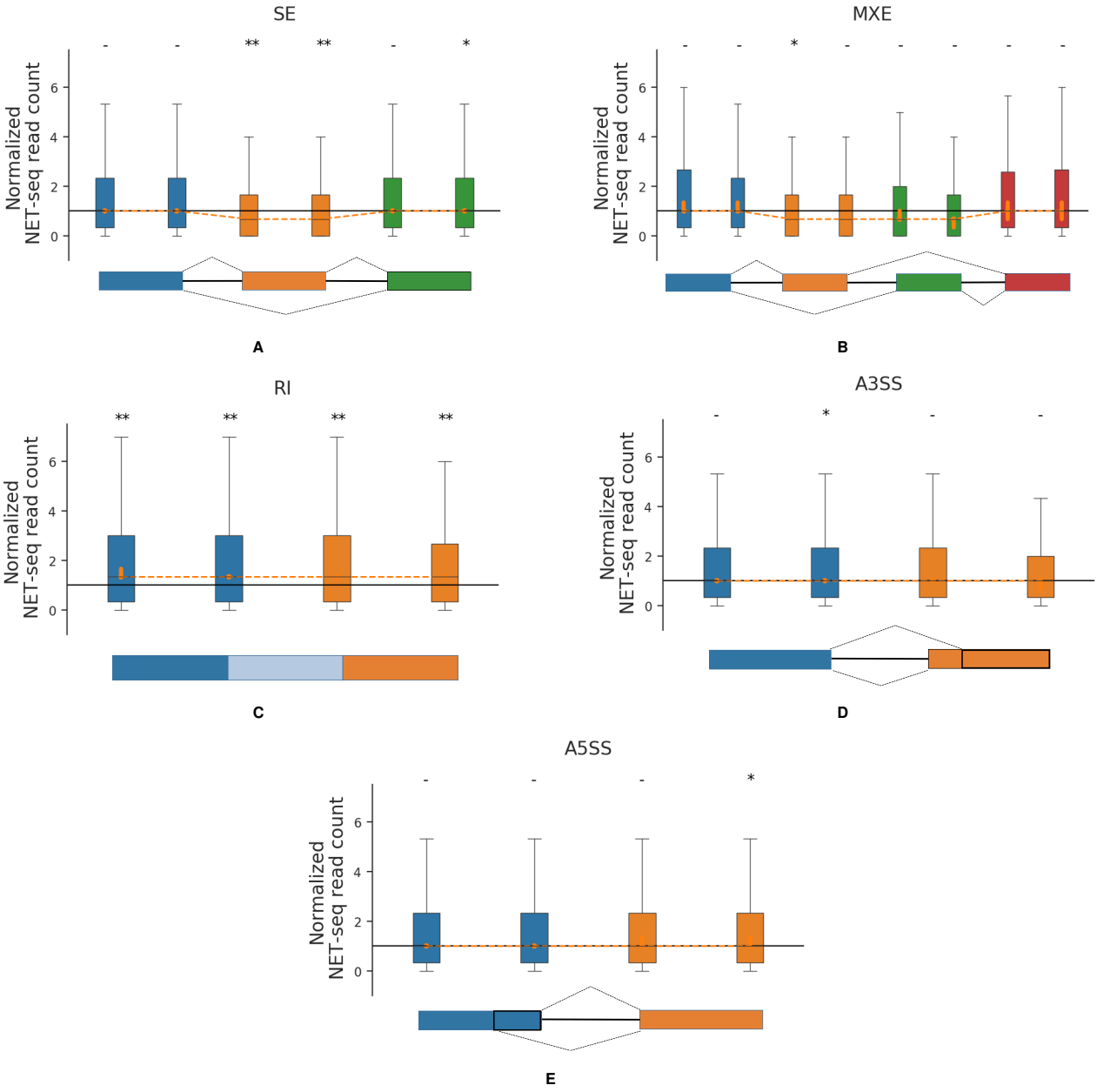


Fig. S4. Associations between different types of alternative splicing sites and Pol II pausing tendencies derived from NET-seq data. For individual types of alternative splicing events, the NET-seq read counts of 3′ and 5′ splice sites were normalized by the median read counts of the same 5000 randomly selected constitutive exons. *: $1 \times 10^{-10} < P < 0.001$, **: $P < 1 \times 10^{-10}$, -: $P > 0.001$, two-sided Wilcoxon rank-sum test.

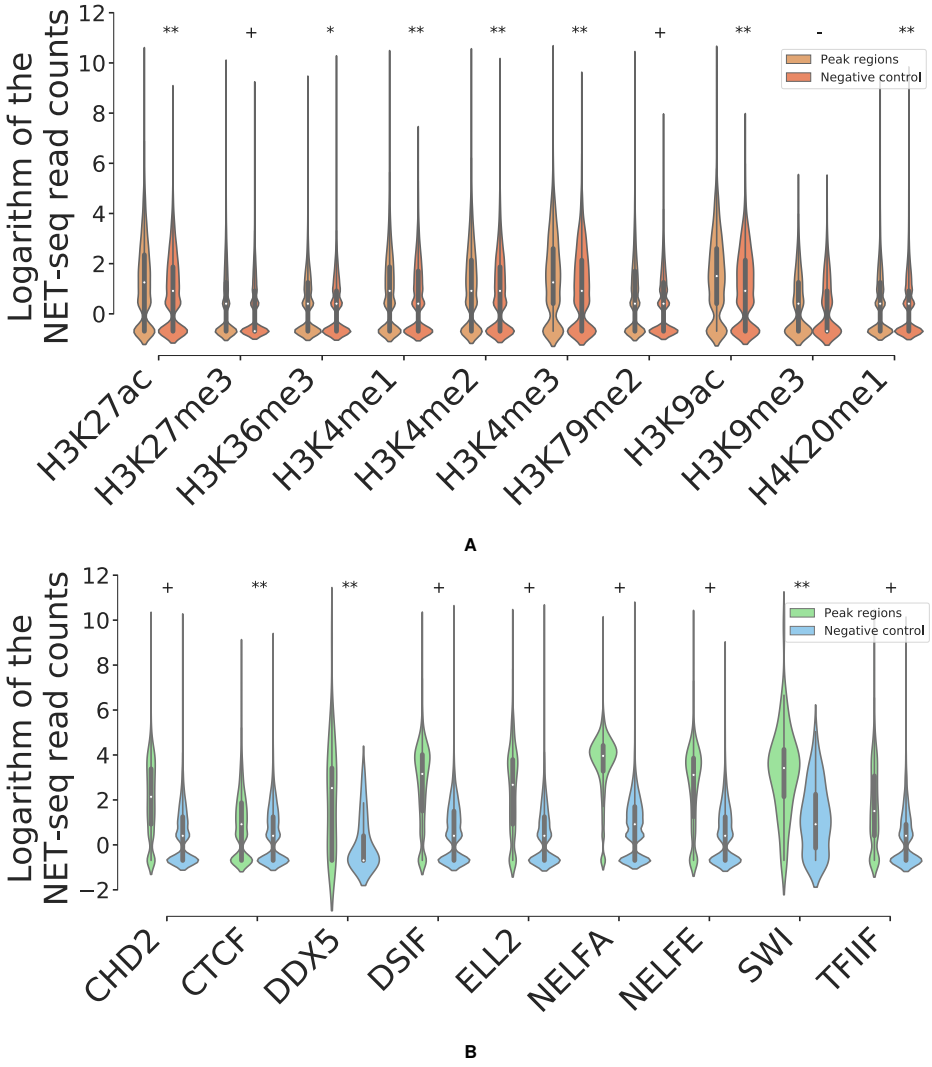


Fig. S5. Associations between histone modifications, transcription factors and Pol II pausing tendencies derived from NET-seq data. (A) The association analysis of histone modifications. (B) The association analysis of transcription factors. The NET-seq read counts were added with 0.5 pseudocounts and taken logarithm. Negative control: randomly sampled genomic regions from the same genes, each of which did not overlap with any binding area of the corresponding epigenetic factor and had the same length of an epigenetic binding peak. *: $1 \times 10^{-5} < P < 0.05$, **: $1 \times 10^{-50} < P < 1 \times 10^{-5}$, +: $P < 1 \times 10^{-50}$; -: $P > 0.05$, two-sided Wilcoxon rank-sum test.

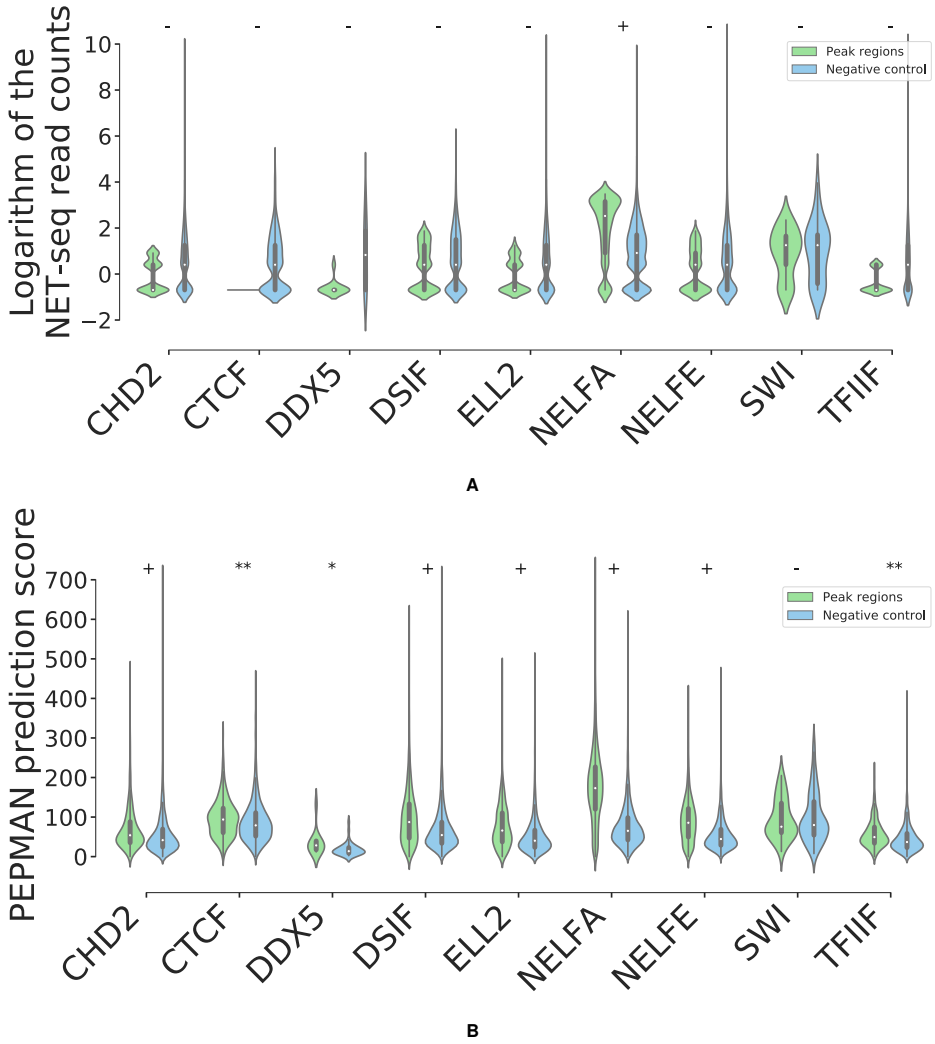


Fig. S6. Associations between transcription factors and Pol II pausing tendencies derived from NET-seq data (A) and PEPMAN prediction scores (B), respectively, for those low-coverage NET-seq regions (i.e., with bottom 30% coverage). The NET-seq read counts were added with 0.5 pseudocounts and taken logarithm. Negative control: randomly sampled genomic regions from the same genes, each of which did not overlap with any binding area of the corresponding transcription factor and had the same length as a binding peak. *: $1 \times 10^{-10} < P < 0.05$, **: $1 \times 10^{-30} < P < 1 \times 10^{-10}$, +: $P < 1 \times 10^{-30}$; -: $P > 0.001$, two-sided Wilcoxon rank-sum test.

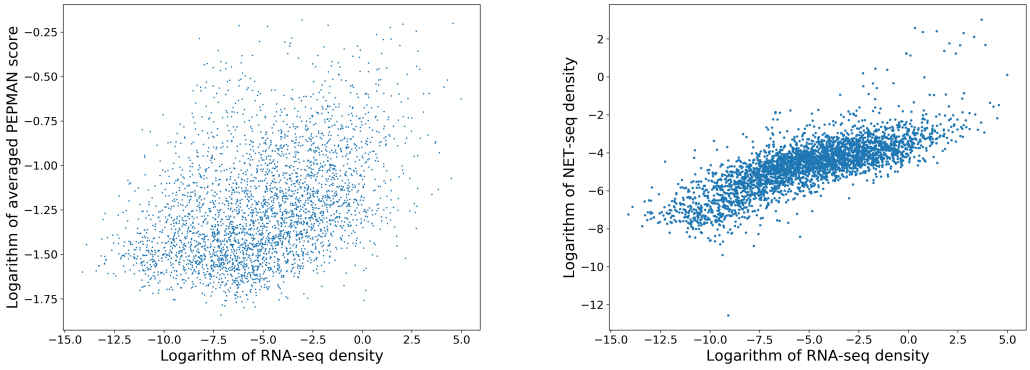


Fig. S7. Correlations between gene expression (represented as RNA-seq read count densities) and PEPMAN prediction scores (A) and read counts in NET-seq data (B) for individual genes. The spearman correlation between PEPMAN prediction scores and expression levels (A) is 0.41 and the spearman correlation between NET-seq read counts and expression levels (B) is 0.78.

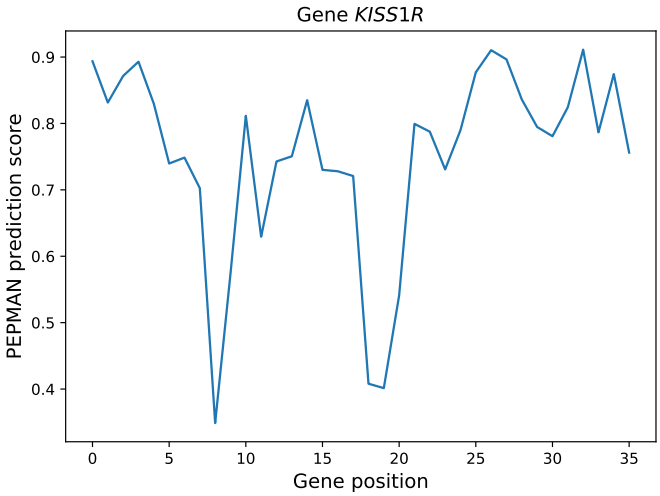


Fig. S8. The PEPMAN prediction scores on the gene *KISS1R*. Each position represents a 201-bp non-overlapping window. There were only five NET-seq read counts aligned on this gene.

Fig. S9. Motifs enriched in those regions with high PEPMAN prediction scores on lowly expressed genes, called by program HOMER (1) with Benjamini q-values < 0.05.

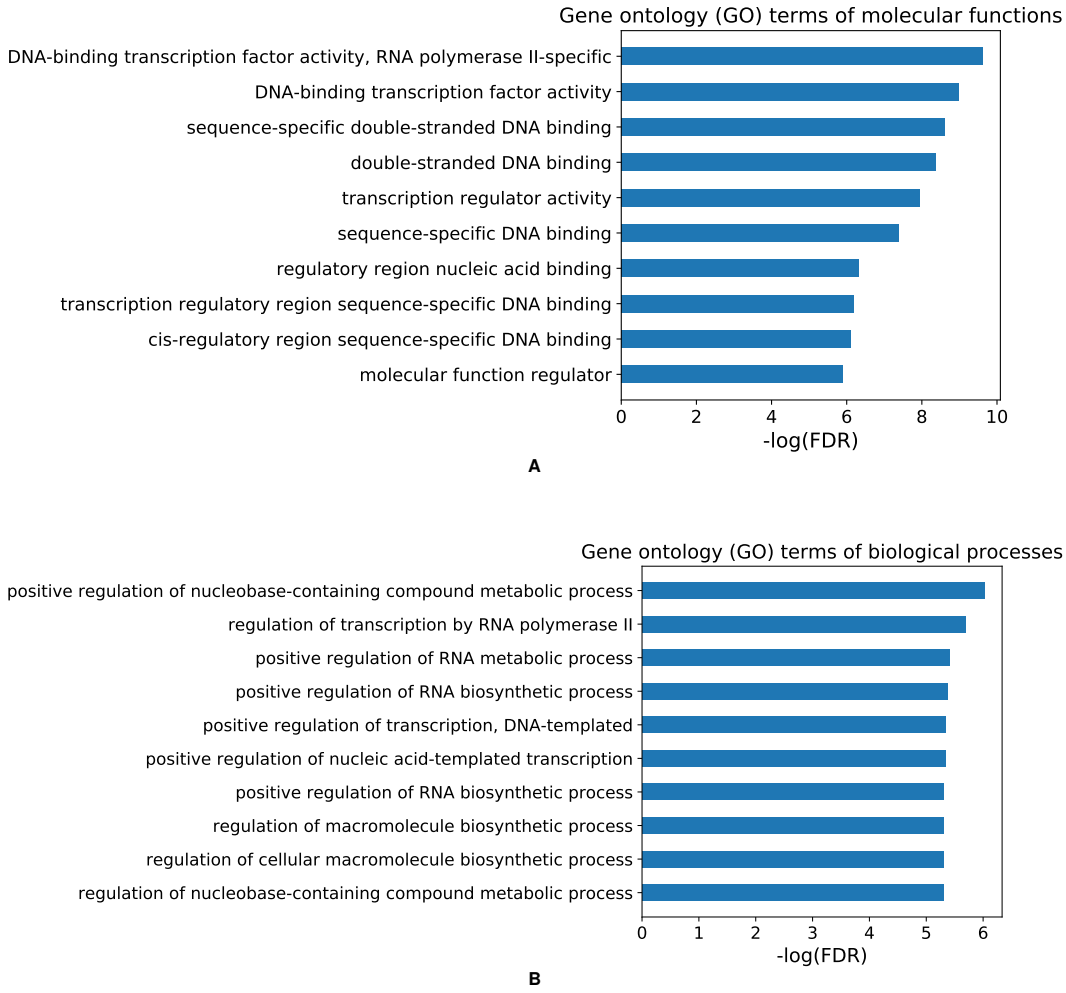


Fig. S10. Top 10 gene ontology (GO) terms on the Pol II pausing associated genes (defined as genes whose average PEPMAN prediction scores were larger than 0.5, $N = 175$). (A) GO analysis of molecular functions. (B) GO analysis of biological processes.

Table S1. Detailed AUROC and AUPR scores of PEPMAN and baselines in HeLa S3 and HEK293T cell lines. The scores are shown in average±standard deviation over ten repeats and the highest ones are highlighted in bold.

Method	HeLa S3		HEK293T	
	AUROC	AUPR	AUROC	AUPR
PEPMAN	**0.8696±0.0014**	**0.4783±0.0042**	**0.8454±0.0016**	**0.4075±0.0044**
prePEPMAN	0.8486±0.0034	0.4209±0.0099	0.8075±0.0052	0.3150±0.0095
CNN+LSTM	0.8184±0.0042	0.3784±0.0107	0.7934±0.0031	0.2995±0.0059
CNN	0.7973±0.0096	0.3365±0.0127	0.7522±0.0123	0.2450±0.0150
LS-GKM	0.7259±0.0018	0.2841±0.0016	0.6749±0.0006	0.2013±0.0010

Table S2. AUROC and AUPR scores of PEPMAN, PEPMAN− and PEPMAN+, which represent the original PEPMAN model, a PEPMAN model trained on the samples in which the regions with the low attention scores were masked out, and a PEPMAN model trained on the samples in which the high attention regions were masked out, respectively.
The scores are shown in average±standard deviation over ten repeats and the highest ones are highlighted in bold.

Method	HeLa S3		HEK293T	
	AUROC	AUPR	AUROC	AUPR
PEPMAN	**0.8696±0.0014**	**0.4783±0.0042**	**0.8454±0.0016**	**0.4075±0.0044**
PEPMAN−	0.8535±0.0015	0.4294±0.0059	0.8314±0.0020	0.3801±0.0047
PEPMAN+	0.6980±0.0021	0.2351±0.0017	0.6303±0.0017	0.1589±0.0017

Table S3. AUROC and AUPR scores of PEPMAN using different window sizes for defining Pol II pausing sites in HeLa S3 and HEK293T cell lines. The scores are shown in average±standard deviation over ten repeats. PEPMAN150, PEPMAN100, PEPMAN50 represent the PEPMAN model using 151, 101 and 51 bp window sizes, respectively.

Method	HeLa S3		HEK293T	
	AUROC	AUPR	AUROC	AUPR
PEPMAN	0.8696±0.0014	0.4783±0.0042	0.8454±0.0016	0.4075±0.0044
PEPMAN150	0.8714±0.0020	0.4805±0.0056	0.8459±0.0017	0.4110±0.0051
PEPMAN100	0.8729±0.0017	0.4849±0.0045	0.8473±0.0012	0.4116±0.0035
PEPMAN50	0.8714±0.0011	0.4813±0.0034	0.8467±0.0010	0.4098±0.0034

Table S4. The best hyperparameter settings of our model determinded using a grid search strategy

Hyperparamter	Value
batch size	32
learning rate	0.002
kernel number	128
kernel size	7
attention layer T	128
dropout (convolution)	0.25
dropout (fc layer)	0.4
attention units	128
epoch	35

References

1. S Heinz, et al., Simple combinations of lineage-determining transcription factors prime-regulatory elements required for macrophage and b cell identities. *Mol. Cell* **38**, 576–589 (2010).