1608–1624 *Nucleic Acids Research, 2023, Vol. 51, No. 4* *Published online 2 February 2023*
https://doi.org/10.1093/nar/gkac1272

Predictive model of transcriptional elongation control identifies trans regulatory factors from chromatin signatures

Toray S. Akcan[1,2], Sergey Vilov[1] and Matthias Heinig[1,2,3,*]

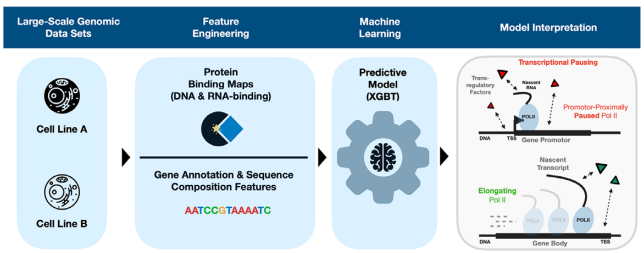
[1]Institute of Computational Biology, Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Ingolstädter Landstraße 1, 85764 Neuherberg, Germany, [2]Department of Computer Science, TUM School of Computation, Information and Technology, Technical University Munich, Munich, Germany and [3]DZHK (German Centre for Cardiovascular Research), Munich Heart Association, Partner Site Munich, 10785 Berlin, Germany

Received October 13, 2021; Revised December 09, 2022; Editorial Decision December 15, 2022; Accepted January 12, 2023

ABSTRACT

Promoter-proximal Polymerase II (Pol II) pausing is a key rate-limiting step for gene expression. DNA and RNA-binding trans-acting factors regulating the extent of pausing have been identified. However, we lack a quantitative model of how interactions of these factors determine pausing, therefore the relative importance of implicated factors is unknown. Moreover, previously unknown regulators might exist. Here we address this gap with a machine learning model that accurately predicts the extent of promoter-proximal Pol II pausing from large-scale genome and transcriptome binding maps and gene annotation and sequence composition features. We demonstrate high accuracy and generalizability of the model by validation on an independent cell line which reveals the model's cell line agnostic character. Model interpretation in light of prior knowledge about molecular functions of regulatory factors confirms the interconnection of pausing with other RNA processing steps. Harnessing underlying feature contributions, we assess the relative importance of each factor, quantify their predictive effects and systematically identify previously unknown regulators of pausing. We additionally identify 16 previously unknown 7SK ncRNA interacting RNA-binding proteins predictive of pausing. Our work provides a framework to further our understanding of the regulation of the critical early steps in transcriptional elongation.

GRAPHICAL ABSTRACT



INTRODUCTION

Transcription of genes is an essential mechanism to maintain cell homeostasis and enable adaptation to changing internal and external stimuli (1,2). It is tightly regulated by chromatin state and transcription factors (TFs) functioning in a highly coordinated fashion (3). The transcriptional cycle starts with the recruitment of the RNA polymerase into the pre-initiation complex (PIC) (4,5). During transcription initiation, a short fragment of nascent RNA is synthesized. The polymerase is then paused at the promoter before entering into productive elongation upon further regulatory signals or terminating prematurely (6). This promoter-proximal pausing is a key rate-limiting step for gene expression as it decides whether a full-length transcript will be made or not (7,8). At equilibrium, paused RNA polymerase accumulates at the promoter since the transcriptional initiation rate is faster than the rate of productive elongation or premature termination (9,10). In vivo, this accumulation can be observed in assays that monitor nascent transcription, such as global run-on sequencing (GRO-seq) (11). Based on this data, the equilibrium between transcription initiation and productive elongation, which is decisive for the regulation of gene expression, can be quantified by the pausing index (PI), also known as the traveling ratio (TR) (12–14). It is defined as the ratio of GRO-seq reads

*To whom correspondence should be addressed. Tel: +49 89 3187 2434; Email: matthias.heinig@helmholtz-muenchen.de

in a window around the promoter compared to the rest of the gene body.

Promoter proximal pausing is the default state after transcription initiation (10,15–17). In addition, the duration of pausing is regulated by the interplay of specific factors that either promote pausing or elongation (16). Pause promoting factors include the DSIF complex consisting of SUPT5H and SUPT4H1, the negative elongation factor NELF, the 7SK complex, consisting of the most highly expressed non-coding RNA *7SK* and proteins such as LARP7, and also specific features of the DNA/RNA sequence (7,18–23). The most important elongation-promoting factor is the positive transcription elongation factor B (P-TEFb), which consists of CDK9, CCNK, CCNT1 and CCNT2 (24,25). Biochemical blocking of P-TEFb showed that its activity is critically important for pause release (26–30). Positive and negative regulators are tightly interlinked. P-TEFb is bound by the inactivating 7SK complex and can be released into its active form by BRD4 (31). Once active it phosphorylates regulators of elongation, such as DSIF, as well as other regulators of chromatin state and RNA processing (32). In addition to these direct regulators, pausing is also indirectly regulated by factors that determine transcriptional initiation and transcript processing (33,34). For example, SRSF2 regulates splicing and has been demonstrated to also determine the duration of pausing (35,36).

Recruitment of P-TEFb to specific promoters through interactions with individual TFs (e.g. NFKB), Mediator, coactivators, and RNA-binding proteins (e.g. DDX2, SRSF2) has been described (35,37–39). Large-scale binding maps of hundreds of RNA binding proteins (RBPs) have recently become available from the ENCODE project (40). Together with the DNA binding maps and GRO-seq, these data allow us to systematically address several key questions about the regulation of pausing at specific promoters. First, which sequence or protein factors determine the recruitment of regulators to a specific promoter? Second, how do signals from positive and negative regulators translate into the extent of pausing quantitatively?

Here, we address these questions by training machine learning models that predict the extent of promoter-proximal pausing quantified by the pausing index from large-scale genome and transcriptome binding maps as well as gene annotation and sequence composition features. We demonstrate high accuracy and generalizability of the model by validation on an independent cell line and we show that the model can accurately predict differential pausing between cell lines indicating that the model captured general cell line independent rules of pausing regulation. Model interpretation allows for assessing the relative importance of each factor, quantifying their effects and predictive values, and systematically identifying previously unknown regulators of pausing. Grouping of factor contributions by molecular functions confirmed the strong interconnection of pausing and co-transcriptional splicing and other steps of gene expression. We additionally identified 16 previously unknown *7SK* interacting RBPs predictive of pausing. These novel pause regulators allow for a systematic and targeted investigation of the regulation of pausing at specific promoters in more detail. Moreover, they provide entry points for experimental manipulation (e.g. with knockdown experiments) to assess their downstream effects on pausing and gene expression in general.

MATERIALS AND METHODS

Transcript annotations (GENCODE)

To engineer gene-centric features of protein binding events and gene annotation and sequence composition features as predictors in our machine learning models we obtained transcript annotations for protein-coding genes and non-coding RNAs from the GENCODE (41) database for the hg19 (GrCH37) genome build. We obtained 81 745 annotated protein-coding transcripts for 20 167 genes. Of these transcripts, 30 186 (18 889 genes) were supported by RefSeq (42) annotations and selected as high-confidence transcripts for the analysis. From the annotations, we obtained 5-prime, intronic, coding exonic and 3-prime genomic regions for each transcript which served to capture interpretable binding sites when integrating CHIP-seq and eCLIP-seq data sets (see CHIP-seq data integration & eCLIPseq data integration). HUGO gene nomenclatures (HGNC) (43) from GENCODE were used to further annotate the transcripts with their respective gene symbols.

A set of non-coding transcripts was obtained through appropriate filtering of the GENCODE transcript annotation set for transcripts that were annotated as one of *miscRNA*, *miRNA*, *snoRNA*, *snRNA* and *lincRNA* which represent miscellaneous, micro, small nucleolar, small nuclear and long intervening RNA biotypes, respectively. These non-coding transcripts were used to engineer features for the machine learning task as well as other downstream analyses, especially in the context of the *7SK* non-coding RNA (see Identification of *7SK* Interacting Proteins). Analogous to the protein-coding transcripts, the genomic regions (5-prime, intronic, exonic, and 3-prime) of non-coding transcripts were used to create binding site features based on CHIP-seq and eCLIP-seq data sets.

Transcript quantifications (RNA-seq)

To ensure that only expressed transcripts are considered we obtained pre-processed transcript quantifications from total RNA-seq experiments from the ENCODE (44,45) project for the K562 and HepG2 cell lines for the hg19 (GrCH37) genome build. Each experiment had two biological replicates. The obtained transcript expressions were required to have a valid ENSEMBLE (46) ID, to be annotated in the aforementioned GENCODE and RefSeq transcript annotation set, to be expressed (fragments per kilobase million (FPKM) > 0) in both of the replicates. The FPKMs were $\log_{10}$-transformed for downstream analyses. After these filtering steps, we considered 16 403 (K562) and 16 670 (HepG2) of the 30 186 protein-coding transcripts and 2655 and 1950 non-coding transcripts for the K562 and HepG2 cell lines, respectively. The transcript quantifications data sets (tsv-files) were taken from ENCODE experiments ENCSR885DVH (K562) and ENCSR181ZGR (HepG2), with accession numbers of replicated experiments ENCFF424CXV and ENCFF073NHK for the K562

cell line and accession numbers ENCFF205WUQ and ENCFF915JUZ for the HepG2 cell line, respectively.

Transcript quantification for the Hela cell line were taken from GSM2400170 and were processed in analogy to the RNA-seq data sets of the K562 and HepG2 cell lines. We thereby obtained the expression profiles of $n = 17\ 934$ protein-coding and $n = 3331$ non-coding transcripts in the Hela cell line.

Transcription start site annotations (CAGE)

To increase the confidence in the expressed transcripts, we further integrated Cap-analysis Gene Expression Data (CAGE) (47) transcription start sites (TSS) for the K562 and HepG2 cell lines. CAGE read counts of the most correlated replicates were aggregated per cell fraction per cell line. Reads were normalized to transcripts per million reads (TPMs). Resulting TSS were then parametrically clustered (48) into CAGE transcription start site clusters (CTSS cluster) with a TPM threshold of 0.1. Singletons with TPM <0.1 were excluded. Only transcripts whose transcription start site (TSS) was also the dominant CAGE transcription start site (CTSS) in a cell-type specific CTSS cluster were retained. We thereby were left with 16 194 and 16 412 protein-coding transcripts in the K562 and HepG2 cell lines, respectively.

Quantifying promoter-proximal pol II pausing (GRO-seq)

We integrated Global-Run-On-sequencing (GRO-seq) (49) data to quantify transcriptional pausing at protein-coding genes with the commonly used pausing index (PI) also known as the traveling ratio (12,27). The PIs served as targets to be predicted in a machine learning task. GRO-seq captures the nascent fragments that build up during the transcriptional cycle and thereby allows us to assess Pol II productivity based on the nascent RNA fragment output. As it is commonly done in the field, we have defined the PI as the $\log_2$ ratio of GRO-seq read counts (number of 30 bp reads overlapping at each position) at the transcription start site (TSS) to the GRO-seq read signals in the gene body. To optimize the PI definition we have built pausing indices with varying TSS window sizes and chose the window size maximizing the negative correlation of the PI with the corresponding transcript expressions (Pearson's $\rho = -0.68$ (K562) and $\rho = -0.66$ (HepG2); see Supplementary Figure S1 pausing index optimization). This was motivated by the fact that high PIs, representative of transcriptional pausing, should result in low gene expression profiles and vice versa. This led to a sharp TSS window size of 3 bp ranging 1 bp up- and downstream of the TSS while rendering the remaining part of the transcripts as the gene body window. Read lengths of 30 bp (K562, GSM1480325) and at least 25 bp (HepG2, GSM2428726) ensure that the most frequent Pol II pause site and associated components (50) are covered. Each signal (counts of GRO-seq reads within windows) was then normalized by the respective window size. A pseudo count of 1 read was added to each resulting window for the $\log_2$ transformation when building the ratio. The PI was calculated for each of the 16194 and 16412 expressed protein-coding transcript in a strand-specific manner for the K562 and HepG2 cell line, respectively. Only transcripts that solely contained the DNA base letters (A, T, C, G) along the whole transcript were considered. This further led to the exclusion of 16 and 9 protein-coding transcripts in the K562 and HepG2 cell lines, respectively. This filtering ensures that we exclude reads that might be erroneously mapped such that we capture the full GRO-seq read signals along the remaining transcripts and thereby obtain comparable signal counts. Overlapping protein-coding transcripts were excluded given the fact that corresponding GRO-seq signals can not be uniquely ascribed to a particular transcript and consequently would result in convoluted PI signals. Transcripts that had no GRO-seq signal neither at the TSS nor in the gene body were excluded as well ($n = 129$ in K562; $n = 196$ in HepG2). This has led to the consideration of 8426 and 8260 protein coding transcripts in the K562 and HepG2 cell lines, respectively (see Supplementary Figure S2 for distribution of pausing indices). The corresponding GRO-seq wig-files can be found under GEO accessions GSM1480325 and GSM2428726 for the K562 and HepG2 cell lines, respectively.

The pausing index based on GRO-cap data for the cross-technology evaluation in the K562 cell line was calculated on data obtained from GSM1480322 and processed in analogy to the K562 and HepG2 GRO-seq data sets. The GRO-seq data for the Hela cell with read lengths of 36bp was taken from GSE62046 and also processed in analogy to the K562 and HepG2 data set, providing the pausing index for $n = 8428$ protein-coding transcripts in the Hela cell line.

DNA binding sites (CHIP-seq)

Chromatin immunoprecipitation sequencing (CHIP-seq) (51) data served to engineer features of gene-centric genomic protein binding events, which were used as input for the machine learning models. These binding sites for DNA binding proteins (DBPs) were obtained from all available CHIP-seq experiments from the ENCODE project for the K562, HepG2 and Hela cell lines for the hg19 (GrCH37) genome build through corresponding peak-called data sets (bed-files). Perturbation experiments were excluded and only optimal (according to irreproducible discovery rate (IDR)) (52) thresholded replicated peaks were considered for downstream analyses to increase the confidence in the obtained binding sites. Experiments with antibodies directly against the factor of interest and newer versioned experiments were prioritized over epitope-tagged and older versioned experiments. We thereby obtained 5041190 (K562), 4138805 (HepG2) and 1010402 (Hela) genomic binding sites for 309 (K562), 211 (HepG2) and 62 (Hela) factors (see Supplementary Tables S1–S3 for CHIP-seq factors per cell line) that served feature engineering purposes (see Feature Engineering). ENCODE CHIP-seq accession numbers for each cell line can be found in Supplementary Tables S4–S6.

RNA binding sites (eCLIP-seq)

Enhanced crosslinking and immunoprecipitation (eCLIP-seq) (53) data served to build gene-centric transcriptomic protein binding features. Binding sites of all RNA-binding proteins (RBPs) from the ENCODE project for the K562

and HepG2 cell lines were obtained for the hg19 (GrCH37) genome build through corresponding peak-called data sets (bed-files). Perturbation experiments were excluded and only optimal IDR thresholded replicated peaks were considered. Newer versioned experiments were prioritized over older versioned experiments. We thereby obtained 409839 (K562) and 435015 (HepG2) transcriptomic binding sites for 120 (K562) and 103 (HepG2) factors (see Supplementary Tables S7 and S8 of eCLIP-seq factors per cell line) for feature engineering (see Feature Engineering). ENCODE eCLIP-seq accession numbers for each cell line can be found in Supplementary Tables S10 & S11.

Transcriptomic binding sites ($n = 3\,035\,169$) in the Hela cell line were taken from the POSTAR (54) data base for all available factors ($n = 30$; see Supplementary Table S9) and lifted to the hg19 genome build.

Identification of *7SK* interacting proteins

We filtered the GENCODE transcript annotation data set for all *7SK* annotated transcripts to enable the identification of known and novel *7SK* binding proteins via observed CLIP-seq signals (eCLIP-seq or POSTAR-derived binding sites) on corresponding transcripts and assess their predictive value in the context of transcriptional pausing. In particular, *7SK* transcripts which were labeled as pseudo versions were included if they were expressed at least at the median expression level of all expressed non-coding transcripts. Their inclusion was motivated by the idea that factors that also bind these pseudo *7SK* transcripts may compete (55) for respective binding sites with factors that bind the non-pseudo version. The set of *7SK* binding factors was defined for each cell line as all factors with at least one CLIP binding site on any of the *7SK* transcripts (see Supplementary Tables S12 - S14).

Feature engineering

For the machine learning task of predicting the gene-wise pausing index of protein-coding genes we engineered features of DNA- and RNA binding events at protein-coding and the closest proximal non-coding transcripts upstream and downstream of the TSS of each protein-coding transcript. In addition DNA sequence and annotation features of protein-coding transcripts served as predictors for the models. The following features were created:

- transcript length (tx.len)
- strand specification (tx.strand)
- chromosome specification (tx.chr.loc)
- location on the linear genome (tx.loc)
- number of annotated exons (tx.ex.num)
- average exon width (tx.ex.width)
- exon density (tx.ex.ratio; ratio of the length of the transcript including introns to the number of exons)
- fraction of exonic sequence (tx.ex.seq; ratio of the length of all exonic sequences to the transcript length)
- GC content of the whole transcript including introns (tx.gc.seq)
- Width of CAGE transcription start site cluster (CTSS) (tx.tss.width)
- AT content of CTSS (tx.tss.at.cont)
- distance to most proximal CpG island (cpg.island.dist) along with information about the CpG island length (cpg.island.length), and features of the sequence: number of CpGs (cpg.island.count), percentage C or G (cpg.island.percent.cg), percentage of CpG (cpg.island.percent.cpg), and ratio of observed to expected CpG (cpg.island.percent.exp.v.obs))
- binary encoding whether the transcript is a housekeeping gene (housekeeping)
- binary encoding of RBP binding events separately for 5′/3′-UTR, introns and coding exons
- binary encoding of DBP binding events separately for 5′/3′-UTR, introns and coding exons excluding Pol II bindings as these are expected to be naturally correlated with the prediction target
- binary encoding of RBP/DBP binding events separately for 5′/3′-UTR, introns and coding exons of the two most TSS proximal non-coding RNAs excluding polymerase II bindings as these are expected to be naturally correlated with the prediction target

Binary encodings start with either ‘chip’ or ‘clip’, followed by the protein and the genomic or transcriptomic region of the proteins binding events on DNA (‘chip’) or RNA (‘clip’). For instance, ‘chip.RBFOX2.5prime’ denotes a binding event of RBFOX2 on the 5′ end of genomic regions of transcripts. Analogously, ‘clip.RBFOX2.5prime.Proxmial.ncRNA.2’ denotes a binding event of RBFOX2 on the 5′ end of transcriptomic regions of the second most TSS-proximal ncRNA of transcripts. CpG islands have previously been implicated in pausing (56), therefore we included CpG island annotations from the UCSC golden path for the hg19 genome build (cpgIslandExt.txt.gz), to engineer CpG island-centric model features. Annotations of housekeeping genes were taken from (57). The number of proximal ncRNAs was fixed to two since in combination with CHIP-seq and eCLIP-seq signals on these proximal ncRNAs the feature space would otherwise overgrow the number of genes (and therefore data points in the regression task) which would result in overfitting of the models. Numeric features not in the range [0:1] were rescaled to that range to achieve faster and more accurate model convergences. DNA- and RNA-binding signals went into the model as binary features (binding (1) or non-binding (0)) (see Supplementary Tables S15 - S17 for the number of binding events per factor on individual genomic or transcriptomic regions for each cell line). The distribution of annotation-based features for the K562 and HepG2 cell lines can be found in Supplementary Figures S3 and S4, respectively. These feature vectors served as a scaffold to build various data matrices for a machine learning regression task based on different feature sub-spaces defined by prior domain knowledge as discussed in the next section.

Feature subsets based on prior knowledge

We stratified the feature space into functionally related sets of proteins in order to characterize the relevance and quantify the importance of pre-, co- or post-transcriptional

events in the context of transcriptional pausing. These subsets of binding features of DNA- and RNA-binding factors implicated in specific biological processes were constructed by integrating Gene Ontology (GO) (58,59) annotations. Functional sets of factors (Chromatin, Initiation, Elongation, Termination, Splicing) were generated based on whether a specific factor was annotated to a biological process (BP) ontology term of any of the following sets: **Chromatin** (chromosome organization, GO:0051276; chromatin organization, GO:0006325; chromatin remodeling, GO:0006338), **Initiation** (RNA polymerase II preinitiation complex assembly, GO:0051123; transcription initiation from RNA polymerase II promoter, GO:0006367), **Elongation** (transcription elongation from RNA polymerase II promoter, GO:0006368), **Termination** (termination of RNA polymerase II transcription, GO:0006369), **Splicing** (mRNA splicing via spliceosome GO:0045292; regulation of alternative mRNA splicing via spliceosome, GO:0000381) and **Processing** (mRNA export from the nucleus, GO:0006406; mRNA 3′-end processing, GO:0031124). The set of Elongation factors was further extended by pause regulatory factors from the literature (16,60,61) if not already included in the GO-derived factor set **Elongation**. These were super elongation complex (SEC) factors CCNT1, CCNT2, ELL, ELL2, ELL3, AFF1, AFF4, MLLT1, MLLT3, established pausing factors NELFA, NELFB, NELFCD, NELFE, SUPT4H1, SUPT5H, SUPT6H, SUPT16H, BRD4, MYC, TAF1, TBP, PAF1, and CDK9 (P-TEFB), as well as 7SK ncRNA pause mediator complex binding factors LARP7, HEXIM1, HEXIM2 and MEPCE (see also Supplementary Table S13). However, we could only consider a subset (n = 19) of all established pausing factors, which were assayed in the CHIP-seq and eCLIP-seq experiments. The **Elongation** factor set thus contained POLR2A, POLR2B, POLR2G, POLR2H, MLLT1, SUPT5H, GTF2F1, BRD4, WDR43, NCBP2, HNRNPU, LARP7, MYC, TAF1, TBP, AFF1, EZH2, PAF1 and SSRP1. However, polymerase associated factors (POLR2A, POLR2B, POLR2G, POLR2H) were excluded since these are expected to correlate with the pausing signal. A set of *7SK* binding proteins derived from binding sites observed in the eCLIP-seq data was generated to quantitatively assess the importance of unknown or less well-established *7SK*-associated factors (see *7SK* non-coding RNA or Supplementary Tables S12-S14 of *7SK* binding factors per cell line). A set representative of general pausing associated factors was generated by forming the union of the Elongation and *7SK* associated factor set (**Elongation + 7SK**). For a list of factors in each resulting functional factor set per cell line see Supplementary Tables S19 & S20.

Each resulting factor set was further stratified into sequence-specific and non-sequence-specific binders. The Molecular Signatures Database (MSigDB) (62,63), a collection of annotated gene sets, the Catalog of Inferred Sequence Binding Preferences (CIS-BP) (64), a library of transcription factors and their binding motifs, and the Homo sapiens comprehensive model collection (HOCOMOCO) (65), a collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis based on binding motifs, were queried to identify sequence specific factors (see Supplementary Tables S21 & S22).

The feature vector space of binding events was then accordingly grouped by these factor sets (see Supplementary Table S23 of factor presence in feature subspaces) to form different feature matrices, always accompanied by DNA sequence and annotation features of protein-coding genes. These feature matrices based on prior domain knowledge, *7SK* ncRNA associations, and sequence-specificity served to build an array of predictive models based on features with a defined biological function. For a baseline comparison of model performances, we have further built 100 random models which randomize over the number of factors, the factors themselves, and their binding patterns. The binding patterns were randomized according to the observed binding proportions.

Model training

Models of transcriptional pausing were obtained by training Extreme Gradient Boosting Tree (XGB) regressors to predict the pausing index with each of the feature subsets (see previous section). Models were trained in each cell line and validated with (i) a 5-fold cross-validation and the application of the model on a 50% holdout test data set from the same cell line taken at random prior to training (individual models) and with (ii) data from an independent cell line with features that are common to both cell lines (synchronized models). This provided us with an unbiased estimate of the model performances as trained models have neither seen the gene's target distribution nor the specific feature distributions of the other cell line. Although the first validation approach is not based on data from an independent cell line as is the case with the synchronized models, it still provides an unbiased model performance estimate as trained models have also not seen any of the data points from the 50% hold out test data set taken prior to training (cross-validation).

Regression with squared loss was chosen for the learning objective. The coefficient of determination (*R*-squared, R^2) was used as the evaluation metric to compare and evaluate trained models. See Supplementary Table S24 for hyperparameter specification and the Zenodo repository for R-Data structures with all model matrices (model.matrices.RDS).

Feature scoring

Shapley additive explanations (SHAP) (66,67) were used as a scoring metric for feature contributions. SHAP is a game theoretic approach to explaining the output of any machine learning model. In contrast to the well-known variable importance metric, it is able to show the positive or negative relationship for each feature with the target. As opposed to most feature importance metrics that average over all genes, each gene receives its own set of SHAP values, greatly enhancing the prediction transparency. SHAP values are additive and allow us to aggregate over contributions of subsets of features which enabled us to capture contributions of binding features per protein and subsequently group these proteins into sets of positive and negative regulatory factors. For instance, we obtain contribution scores for a transcription factor binding on the 5′UTR, exons, introns, and 3′UTR on the genome and transcriptome as identified by

CHIP-seq and eCLIP-seq, respectively. We derived total factor contributions by aggregating the SHAP scores per factor over each gene region which enabled us to identify specific pause regulatory factors by selecting factors with high effect sizes.

RESULTS

Predictive models of transcriptional pausing

The transitioning of promoter-proximally paused Pol II (Figure 1A, promoter-proximally paused Pol II) into its elongating phase of nascent RNA synthesis (Figure 1A, elongating Pol II) is regulated by trans-acting protein co-factors as well as cis-regulatory DNA and RNA sequence features (16,18) which we refer to as chromatin signatures.

For the identification of such specific regulatory chromatin signatures, we used large-scale genomic and transcriptomic protein binding maps from ENCODE and compiled gene annotation and sequence composition features. We then followed a systematic machine learning approach to predict the degree of transcriptional pausing at protein-coding genes (Figure 1B) through the integration of these chromatin signatures in a regression model with Extreme Gradient Boosting trees (XGB) with the potential to reveal explanatory factors (Figure 1C).

To facilitate the validation in independent cell lines we obtained relevant data sets for two different cell lines (K562 and HepG2). The prediction target was defined as the gene-wise *pausing index* (see Materials and Methods; see Supplementary Figure S2 for pausing index distributions). It quantifies the degree to which a gene is paused (high pausing index) or elongated (low pausing index). As compared to traditional definitions of the PI, our flexible definition seeks to identify the threshold that is best aligned to the expected relation to transcript levels and covers the more clearly distinguishable peak that lies more proximal to the TSS (see Supplementary Figure S2C–E). To construct the feature matrix of predictors as input for our models we systematically integrated genome-wide CHIP-seq (see Materials & Methods) and eCLIP-seq (see Materials and Methods) data from the ENCODE project, providing DNA and RNA binding sites on the genome and transcriptome respectively (see Supplementary Tables S15 and S16). Gene-centric annotation and composition features were mainly engineered based on GENCODE transcript annotations (see Materials and Methods, Supplementary Figures S3 and S4). CAGE transcription start sites were integrated (see Materials & Methods) to define high confidence TSS and further validate the expression of transcripts. We thereby obtained a total of 2503 features of 2485 DNA & RNA binding and 18 gene annotation features in the K562 cell line and 1832 features of 1814 DNA and RNA binding and 18 gene annotation features in the HepG2 cell line. We then trained an Extreme Gradient Boosting Tree regressor (see Materials and Methods and Supplementary Table S25) to predict the pausing index of protein-coding genes ($n = 8426$ in K562) with high accuracy and explain up to 68% of the observed variance ($R^2 \sim 0.68$ on 50% hold-out test data set, K562) of the pausing index (Figure 2A).

The model performances can be further evaluated through (i) the application of a model trained on one cell line and applied to the full data of the other cell line (Figure 2B), (ii) the application of a model trained on one cell line and applied to genes that are only expressed in the other cell line (Figure 2D) and (iii) the application of a model trained on one cell line and applied to genes present in both cell lines with significantly different pausing indices representing extreme observation specific to the other cell line (Figure 2F). See Supplementary Figure S5 for model performances of a model trained on the HepG2 cell line and validated on the K562 cell line.

The predictive power and generalizability of the model were supported by the high prediction performance on the independent cross-cell type test data set (Figure 2B, performance on HepG2 data of K562 model) in which it was still able to explain up to 53% of the variance. The decreased model performance with an R^2 of 0.53 as compared to 0.68 (Figure 2A) is likely due to the reduced amount of features that are available in the HepG2 cell line (39% of all features ($n = 987$) of $n = 2503$ features available in the K562 cell line).

A good performance in the cross-cell type prediction task (Figure 2B) can have two reasons: (i) the model captures the signal of ubiquitously expressed genes that are similar between cell types, as might be the case with housekeeping genes, or (ii) it learned general rules that would also allow for predicting cell type-specific pausing indices from cell type-specific chromatin signatures. To distinguish these scenarios we identified the sets of exclusively expressed genes (Figure 2C) and assessed the performances of models trained on one of the cell lines on the genes exclusively expressed in the other cell line (Figure 2D). The K562 model was able to explain up to 57% and the HepG2 model up to 58% of the observed variance in the pausing indices in the HepG2 and K562 cell line respectively.

We further validated that our model can also identify quantitative changes on transcripts which showed differential (fold change ≥ 2) cell type-specific distributions of the pausing indices. For these sets of transcripts (Figure 2E, blue, green) we evaluated the concordance of observed pausing index differences between the cell lines against the differences in predictions of the pausing indices using models trained in one of the cell lines and applied them to data in the other cell line (Figure 2F). Although we can recognize a substantial decrease in model performances with a correlation of 0.24 (Figure 2F, HepG2 specific pausing indices; green) as compared to 0.73 for the prediction on the entire HepG2 cell type data (Figure 2B) or 0.76 on HepG2 cell type-specific genes (Figure 2D), the model not only predicts extreme cases but also captures quantitative differences of pausing indices specific to the cross cell type to a certain extend. This further underlines the ability of the model to generalize to other cell lines and shows that cross cell type predictions are not only driven by ubiquitously expressed genes.

To further increase the confidence in the obtained modeling results we have additionally investigated (i) data on a third cancer cell line (HeLa), (ii) three additional machine learning methods (Ridge Regression (RR), Random Forests (RF), Gradient Boosting Trees (GBDT)) and (iii) a model based on the pausing index calculated on a different run-on-assay (GRO-cap). These served to additionally

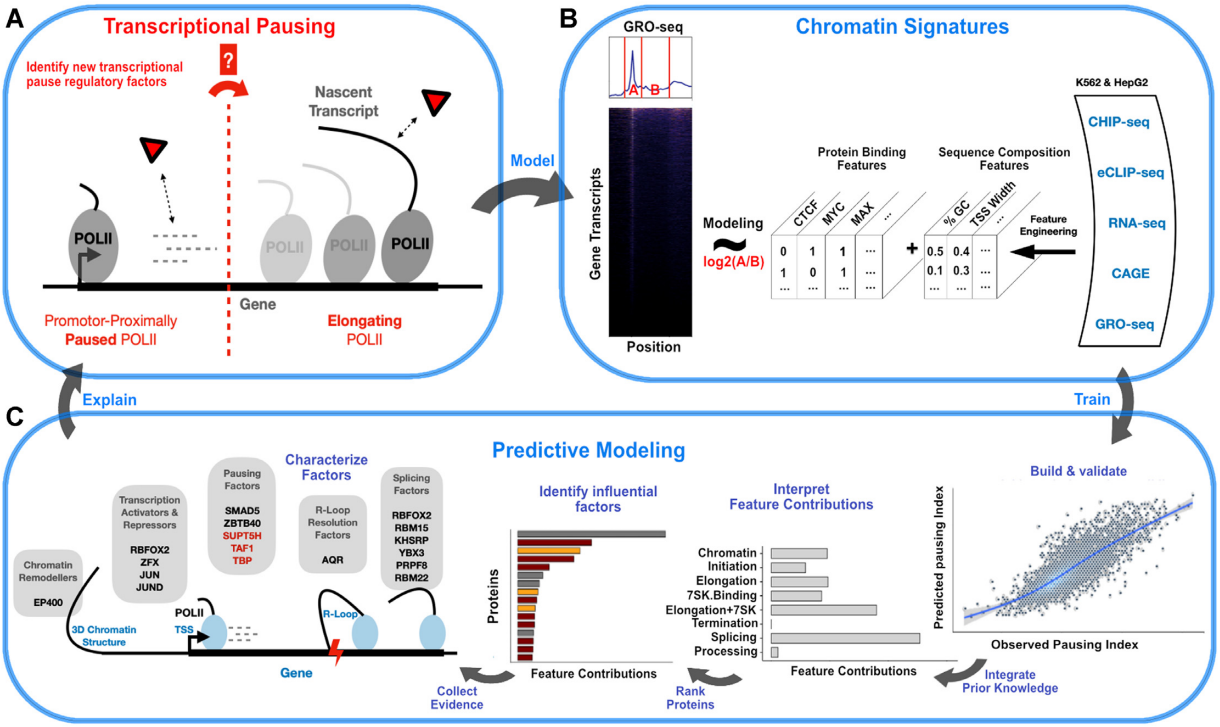


Figure 1. (**A**) Central question as to which specific factors are implicated in the transitioning of promoter-proximally paused polymerase II into its elongating phase of nascent RNA synthesis. (**B**) Integration of large-scale genomic data sets to build the chromatin context of transcriptional pausing (A) with protein binding events and gene annotation and sequence composition features for the prediction task of promoter-proximal pausing of the polymerase II. Pausing is quantified by relating GRO-seq read densities at the TSS to GRO-seq read densities in the gene body. (**C**) Machine learning approach to predict promoter-proximal Pol II pausing with chromatin signatures (B), followed by the integration of prior knowledge and selection of factors as regulators of promoter-proximal Pol II pausing.

validate the cross-cell line prediction performances, rule out prediction performance differences potentially resulting from the selection of the architecture of the machine learning model, and rule out technological bias.

A 5-fold cross-validated and regularized XGB regression model in the Hela cell line ($n = 92$ DNA- and RNA-binding factors) achieves an R-squared of 0.56 (Pearson's rho = 0.75) when applied to an independent 50% hold-out test dataset from the same cell line taken prior to training (see Supplementary Figure S6A). This performance is lower than the other full model's performances (R-squared HepG2: 0.62, K562: 0.68). The difference can be attributed to the fact that the Hela model includes four times fewer factors than the full K562 model ($n = 92$ versus $n = 404$). Models trained on each cell line using only features present in the HeLa data achieve comparable and expectedly lower model performances of R-squared between 0.53–0.56 (see Supplementary Figure S6B). Nevertheless, this reduced set of factors (37/295, only 12% of available factors in HepG2 and 47/404, only 11% of available factors in K562) is still predictive of pausing.

To evaluate the impact of the type of model on prediction performance, we have conducted a systematic comparison with three alternative methods based on the full K562 data set (see Supplementary Table S26), namely Ridge regression (RR), Random Forests (RF), and Gradient Boosting Decision Trees (GBDT). As expected, RR analysis performs worst (R-squared 0.6 on K562 50% hold-out test data). The tree-based RF and GBDT models perform similarly well, also compared to the XGB model (R-squared RF: 0.69, GBDT: 0.71, XGB: 0.68) and greatly outperform linear regression analysis, as these algorithms can take non-linear relationships into account. The fact that all models perform reasonably well underlines the predictive power of underlying features. These results suggest that the tree based models can be used interchangeably.

To assess whether the model learned a technology bias inherent to GRO-seq, we trained analogous models based on GRO-cap data from K562. The GRO-cap model showed even slightly higher performance (R-squared = 0.72, Pearson's rho = 0.85; see Supplementary Figure S7A) than the GRO-seq data (R-squared = 0.68, rho = 0.83; see Supplementary Figure S7A) on a hold-out data set of GRO-cap pausing indices. To distinguish if both models learned patterns related to pausing or a technology bias, we applied models trained on one technology to predict the pausing index of the genes in the hold-out test set and compared these predictions to the observed pausing index measured with the second technology (cross-technology evaluation). In this comparison, the GRO-seq model can explain 53% of the variance in the GRO-cap measurements and the GRO-cap model can explain 44% of the variance of the GRO-seq measurements respectively. Given the noise introduced by the different technologies (R-squared between GRO-seq and GRO-cap: 0.74) and the uncertainty of the model predictions (R-squared GRO-seq model: 0.68), we can cal-

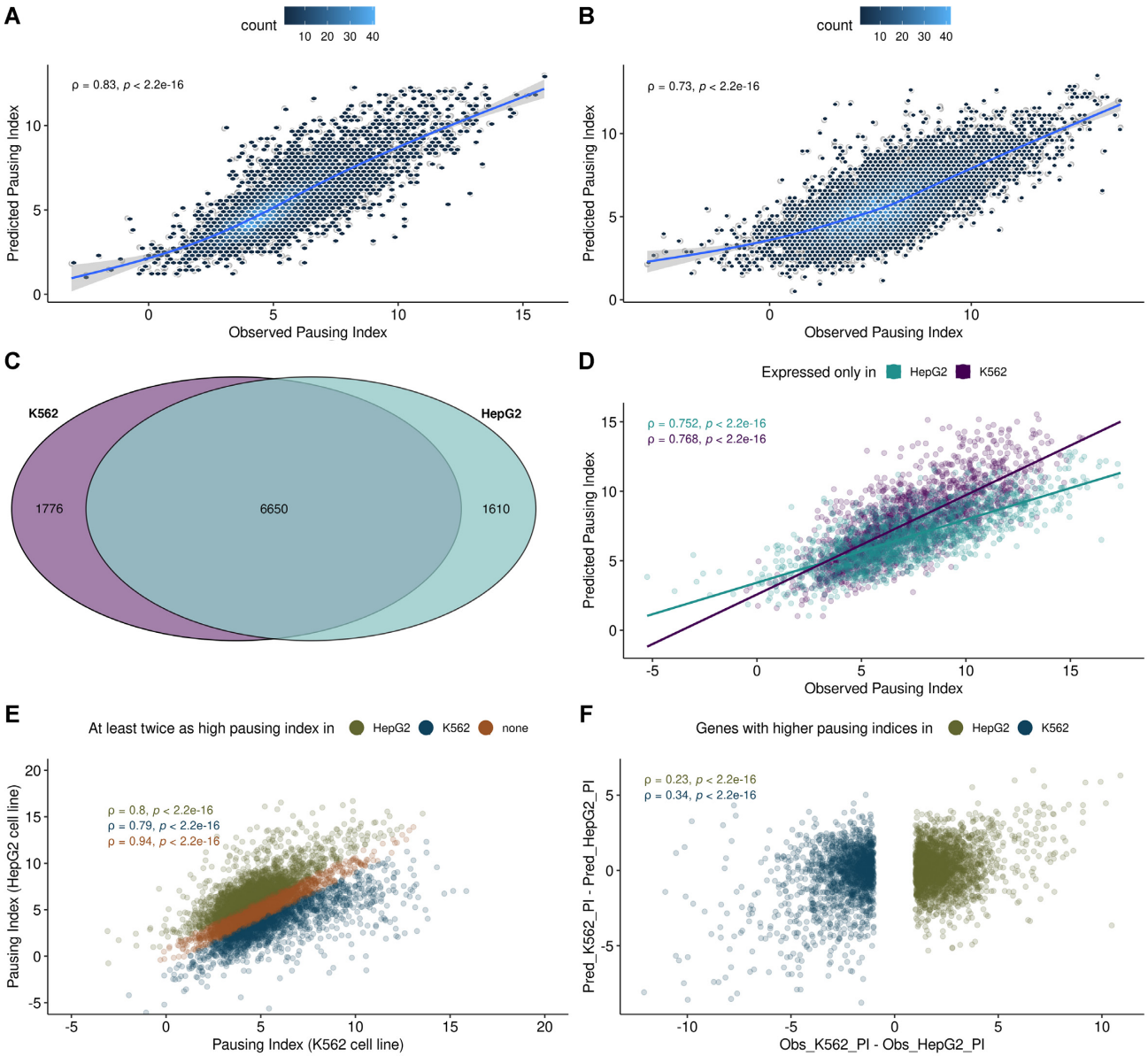


Figure 2. (**A**) Observed versus predicted pausing indices (log2 scale) of a 5-fold cross-validated and regularized XGB regression model in the K562 cell line applied to an independent 50% hold-out test dataset from the same cell line taken prior to training. Pearson's correlation coefficient rho (ρ) with the associated p-value is depicted in the upper left. (**B**) Observed vs. predicted pausing indices of a 5-fold cross-validated and regularized XGB regression model in the K562 cell line applied to the independent test dataset from the cross cell line (HepG2). The model was trained with features common to both cell lines. Pearson's correlation coefficient rho (ρ) with the associated p-value is depicted in the upper left. (**C**) Venn diagram of transcripts between cell lines. (**D**) Observed vs. predicted pausing indices of a 5-fold cross-validated and regularized XGB regression model from each cell line applied to data of genes exclusively expressed in the cross cell line. Pearson's correlation coefficient rho (ρ) with the associated *P*-values are depicted in the upper left. (**E**) Observed pausing indices from the K562 versus HepG2 cell line. Transcripts with at least a 2-fold higher pausing index in one but not the other cell line are colored either green (HepG2 specific transcripts) or blue (K562 specific transcripts). Transcripts with similar pausing indices (less than a 2-fold change) in both cell lines, thus not specific to any of the cell lines, are colored in orange. Pearson's correlation coefficients (ρ) for each of the groups with associated p-values are depicted in the upper left. (**F**) Observed pausing index differences between cell lines against differences of predicted pausing indices obtained from models trained in each cell line and applied to data from the cross cell line. Models were trained on features common to both cell lines. Differences are shown for genes which showed a 2-fold change between cell lines as identified in E).

culate the expected proportion of variances explained by the product $0.74 \cdot 0.68 = 0.50$. Therefore, the observed *R*-squared of 0.53 (GRO-seq) is well in line with our expectation. These results suggest that the models can generalize between technologies and prediction are not dominated by technology biases.

Given the high predictive power of the obtained model not only on intra-cell type holdout test data sets (Figure 2A), the inter-cell type test data set (Figure 2B) as well as its ability to predict pausing indices of cell type-specific genes (Figure 2D) and cell type-specific differential pausing indices (Figure 2F), we concluded that our model captured general rules of pausing regulation independent of the cell type and that the underlying chromatin signatures of the models would have sufficient discriminatory power to explain the observed variance in the pausing index. The successful validation of model performances on data of a third cell line (see Supplementary Figure S6), with alternative model architectures (see Supplementary Table S26) and an alternative Pol II run-on-assay (see Supplementary Figure S7) further increased the confidence in the obtained modeling results. We thus continued with downstream feature interpretation and selection approaches to suggest potential novel regulators of transcriptional pausing. Downstream analyses were performed on data from the K562 cell line due to the increased amount of features available.

Contribution of individual transcript processing steps to the prediction of pausing

We next aimed to gain a mechanistic understanding of the underlying predictive contributions. To measure the contributions of model features we have used Shapley Additive Explanations (SHAP) (67,68) as a feature scoring metric (see Materials and Methods) which captures the directional contribution of each model feature specifically for each gene on the target variable. A model feature may increase or decrease the pausing index or exert no effect at all depending on the factors relevance for pausing and their interaction with other features of each gene (Figure 3A). Their combined effects converge in predicted pausing indices which in turn represent the average output whether a gene is paused or not.

Because transcriptional pausing is connected with other steps of gene expression from chromatin organization (69–71), transcription initiation (8,50,72), to splicing (33,73,74) and post-transcriptional transcript processing (34,75,76), we assessed the regulators of these pre-, co- or post-transcriptional events according to their importance in predicting pausing. To that end, we have generated sets of regulators (see Methods and Supplementary Tables S18 - S20) representative of specific RNA processing events (*Chromatin*, *Initiation*, *Elongation*, *Splicing*, *Termination*, *Processing*) based on Gene Ontology (GO) annotations. The *Elongation* factor set was further extended by established pausing factors from the literature. The *7SK* non-coding RNA complex is a key regulator of pausing (35,77–81). To assess the role of RNA binding proteins participating in the 7SK complex for pausing, we additional built a set of factors that bind the *7SK* ncRNA in the eCLIP-seq datasets (see Methods and Supplementary Tables S12 and S13 for *7SK* binding factors per cell line). This set included the well-known *7SK* binder LARP7, the pausing-related regulator AQR previously not associated with the *7SK* as well as the following factors not previously associated with pausing: SSB (LARP3), HNRNPK, DGCR8, PCBP1, ATF, ZNF800, XRCC6, NCBP2, SBDS, YWHAG, GRWD1, ZNF622, SRSF7, TARDBP and BUD13. A set consisting of the union of Elongation and *7SK*-associated factors were generated as well (*Elongation* + *7SK*). All sets of regulators were further stratified into known sequence-specific and non-sequence-specific binders (see Supplementary Tables S21 and S22) in order to assess the relevance of sequence-specific binding events. For each factor in the resulting functional set of regulators we aggregated their feature contributions (Figure 3A) per functional process (Figure 3B).

Splicing factors had the highest contributions followed by elongation and *7SK* binding proteins. This strongly supported the intricate connection to co-transcriptional splicing events (36,73,82) and strengthened the role of the newly identified *7SK* binding proteins as transcriptional pause regulatory factors. The *Elongation* factor set of established pausing factors served as a validation of our approach.

We next asked how models would perform if they are trained exclusively on the features defined by each of the previously defined sets of regulators. For a baseline comparison models were also trained on randomized input data (see Materials and Methods). Figure 3C shows the model performances (R^2 values) for each of the feature subspaces of cross-validated models in the K562 cell line on the independent 50% holdout test data sets (see also Supplementary Table S25 for all model results). In general, all models perform reasonably well relative to the number of features they incorporate. As an example, the splicing factor based model (*Splicing*) incorporates only 14% ($n = 57$) of all available factors yet performs almost equally well as the full model (*All*) incorporating all available factors ($n = 398$). Likewise, the *Initiation* model considers only about half the number of factors than the chromatin-associated model (*Chromatin*) yet performs slightly better (R^2 of 0.54 versus 0.53).

As expected, the *7SK* ncRNA-associated factor model (*7SK.Binding*) and the model with previously established pausing factors (*Elongation*) perform very well despite the low number of factors considered in those models. The predictive power of pausing/elongation factors becomes further evident when we consider the model of the union of 7SK and established elongation factors (*Elongation* + *7SK*) which outperforms (R^2 0.62) each individual factor set alone (*7SK.Binding*: $R^2 = 0.55$, *Elongation*: $R^2 = 0.56$) and performs almost equally well as the full model ($R^2 = 0.62$ versus 0.68). This result highlights the relevance of the novel set of *7SK* binders identified by protein-RNA interactions as putative pause regulators. Taken together, the majority of factor sets show high predictive power relative to the number of factors they incorporate but their performances should not be compared directly to each other due to the variable amount of factors considered in the models. Their predictive value demonstrates the interconnectedness of underlying processes with the transcriptional pausing outcome. It further supported and strengthened the role of the *7SK* ncRNA as a transcriptional pause mediator complex

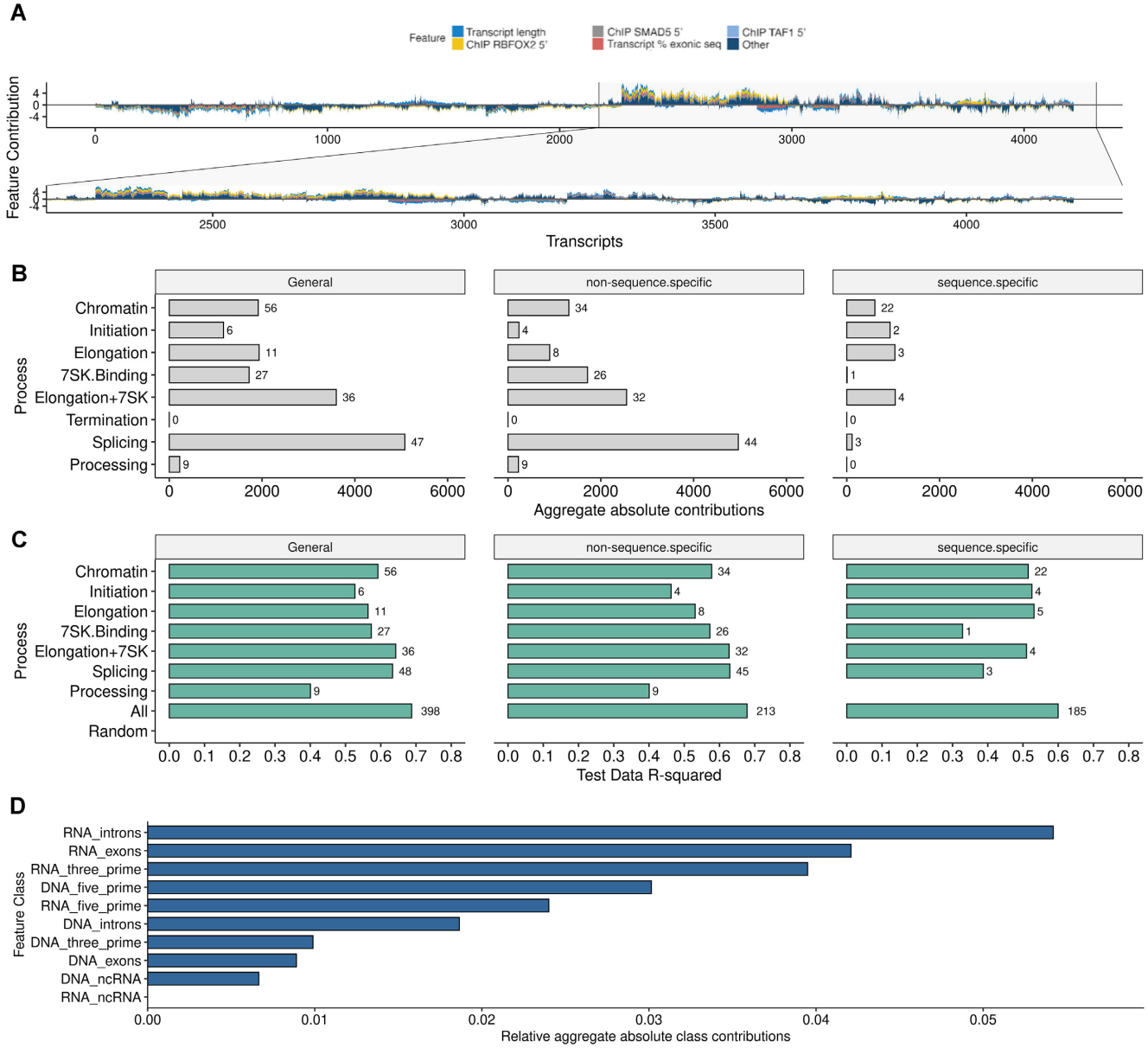


Figure 3. (**A**) Individual feature contributions (SHAP feature contributions, y-axis) on each transcript (x-axis) with a sample zoom-in on a subset of transcripts for better visual investigation. Only the top 5 most influential features are colored and remaining features aggregated in 'Other', see legend. Feature 'ChIP RBFOX2 5′' refers to the binary indicator variable for a RBFOX2 binding site determined by ChIP-seq being present in the 5′ region of the transcript (see Methods section on feature engineering). The other ChIP-seq data sets are labeled analogously. (**B**) Aggregate absolute contributions of factor classes based on prior knowledge, further divided by sequence and non-sequence specific binding factors. The process '*Processing*' refers to mRNA polyadenylation and export from the nucleus. Number of factors are given behind the bars, only factors with non-zero contributions were counted. (**C**) R^2 performances of individual models of factor classes based on prior knowledge on 50% holdout test data set. Number of factors associated with each functional process are given behind the bars, irrespective of their contributions scores, i.e. same factor sets as in (**B**) which in turn shows only factors with contributions >0. (**D**) Aggregate absolute contributions of factors based on their binding modes.

and allowed us to suggest the factors from the set of *7SK* associated factors (*7SK.Binding*) (see Supplementary Tables S12 and S13) as additional *7SK* ncRNA binding proteins to be implicated in the regulation of pausing based on their predictive value.

We next asked whether protein-DNA or protein-RNA binding events contributed to the explanatory power of the models. We found that the individual contributions of RNA binding events are generally higher than those of DNA binding events (Figure 3D). Investigating the contributions of factors by their functional classes within the highest ranked class (RNA introns) (see Supplementary Figure S9 & S10) reveals that splicing factors are enriched for RNA intron binding sites (Fisher's exact test, one-sided (greater), $P = 0.034$, odds ratio 4.45, confidence interval [1.11;Inf] in K562 and $P = 0.032$, odds ratio 7.1 [1.15;Inf] in HepG2). The high contributions of genomic binding events on the 5′ region of transcripts (Figure 3D, DNA_five_prime) are in

line with observed 5′ modulated transcriptional pause states (83).

Overall the results for the HepG2 cell line are very similar and support the conclusions (Supplementary Figure S8). Although gene annotation and composition features account for 26% of all feature contributions (see Supplementary Figure S11–S14) they are static in their nature and cannot explain the variation of pausing between cell lines. Therefore, we focus the discussion on individual proteins and their binding events as they are dynamic between cell lines.

Modulators of transcriptional pausing

Based on our model, we aimed to identify specific pause regulatory factors. To obtain a ranking of the importance of individual DNA- and RNA-binding factors for predicting Pol II pausing, we aggregated the SHAP contributions (see Supplementary Figures S15 and S16 for individual feature contributions per cell line) into a single contribution score per factor and selected the minimal set of most influential factors (16 out of 398) that makes up 50% of all feature contributions (Figure 4A). Established pausing factors from the literature (Figure 4A, highlighted in red) are ranked among these top influential factors, validating our factor ranking approach. Three factors not primarily related to pausing were ranked higher than the established pausing factors and are potential novel modulators of pausing with at least the effect size of the established factors. However, all other factors have similarly high contributions and can be considered almost equally important.

A minimal model that only operates on the features of these 16 most influential factors (including gene annotation and composition features) which includes only five known pausing or *7SK*-related factors (AQR, BRD4, SUPT5H, TAF1, TBP) achieves an R^2 of 0.65 (on 50% holdout test data set; see Supplementary Figures S17 and S18 performances of minimal models per cell line) and thus performs almost equally well as the full model with all 398 factors and an R^2 of 0.68. Additionally, it outperforms the *Elongation + 7SK* model (Figure 3) which incorporates almost twice as many factors (n = 27) of *7SK*-associated and established elongation factors which, although highly predictive, only achieved an R^2 of 0.61 as compared to an R^2 of 0.65 of the minimal model which indicates that not all pausing related factors were captured in the *Elongation + 7SK* set. The minimal model ($n = 9$) of the HepG2 data consisted of RBFOX2, AQR, TAF1, TBP, RBM15, RBM22 KHSRP, PRPF8 and YBX3, which are all included in the minimal model identified in K562.

To obtain a reference of the predictive power of obtained factors we trained another model solely based on Pol II CHIP-seq binding data (binding patterns of POLR2A, POLR2AphosphoS2, POLR2AphosphoS5, POLR2B, POLR2G and POLR2H) since promoter proximal pausing is tightly related to the phosphorylation state of RNA Pol II which should contain the information necessary to explain the extent of promoter proximal pausing of each transcript and hence be able to predict the extent of pausing defined by the pausing index computed from GRO-seq data. This model (see Supplementary Figure S19) can explain up to 60% of variance in the pausing index as compared to the full K562 model ($n = 398$ factors) which can explain up to 68% or the minimal model ($n = 16$ factors) which can explain up to 65% of variance in the pausing index. The Pol II-only model is missing further subunits (like POLR2C, POLR2E etc.) which is likely the reason why not more of the variance can be explained. However, an additional 8% can be explained by non-polymerase II-associated factors. This is in addition to the 60% of variance explained by polymerase II-associated factors, which underlines non-polymerase II-associated factor's predictive power for transcriptional pausing. This comparison validates our feature engineering, model selection and model training approach, as the model behaves as expected, providing high predictive power. However, for a mechanistic understanding of the regulatory networks in which a variety of factors and co-factors beyond polymerase subunits are interacting to co-modulate different transcriptional processes (e.g. pausing, elongation, splicing) the integration of only polymerase-associated subunits would lead to a circular reasoning and limit the discovery of additional explanatory factors, which necessitates the integration of protein binding data for a broad range of proteins, ideally on the DNA and RNA level.

Lastly, investigating the factor rankings of the previously trained models based on (i) data on of a third cancer cell line (HeLa), (ii) three additional machine learning methods (Ridge Regression (RR), Random Forests (RF), Gradient Boosting Trees (GBDT)) and (iii) the pausing index calculated on a different run-on-assay (GRO-cap), shows that certain factors are consistently ranked high across all validation settings which greatly increases the confidence in the factor's regulatory role in transcriptional pausing. To begin with, a comparison of the top 15 ranking factors (see Supplementary Figure S6 C) from model's trained in each cell line and validated on cross-cell type data with pairwise shared features between the cell lines (see Supplementary Figure S6 B) shows that six factors (TAF1, TBP, UPF1, TIA1, PTBP1 and U2AF2), including the well-established pausing factors (TAF1, TBP) are consistently ranked high across all models. Both included in our minimal factor set ($n = 16$). To continue, a comparison of the tree-based models trained during the evaluation of model architectures (see Supplementary Table S26) shows that 56% (9/16) factors are common among the top 16 ranking factors of trained models. This set consists of RBFOX2, AQR, SMAD5, TAF1, SUPT5H, YBX3, RBM15, KHSRP, and PRPF8. The presence of two well-established pausing factors (TAF1 and SUPT5H) again validates our factor ranking, model building, and selection approach. Moreover, all of these factors are included in our minimal factor model which further increases the robustness of our results as different model architectures converge on a similar ranking of explanatory predictors. Finally, a comparison of the 16 top contributing factors (see Supplementary Figure S7 B) of models trained to predict pausing indices based on different sequencing protocols (GRO-seq versus GRO-cap) (see Supplementary Figure S7 A), with the top ranking factors from the minimal K562 model, shows that 68% (11/16) of these factors are common across both models providing predictive power across both sequencing protocols. This set of

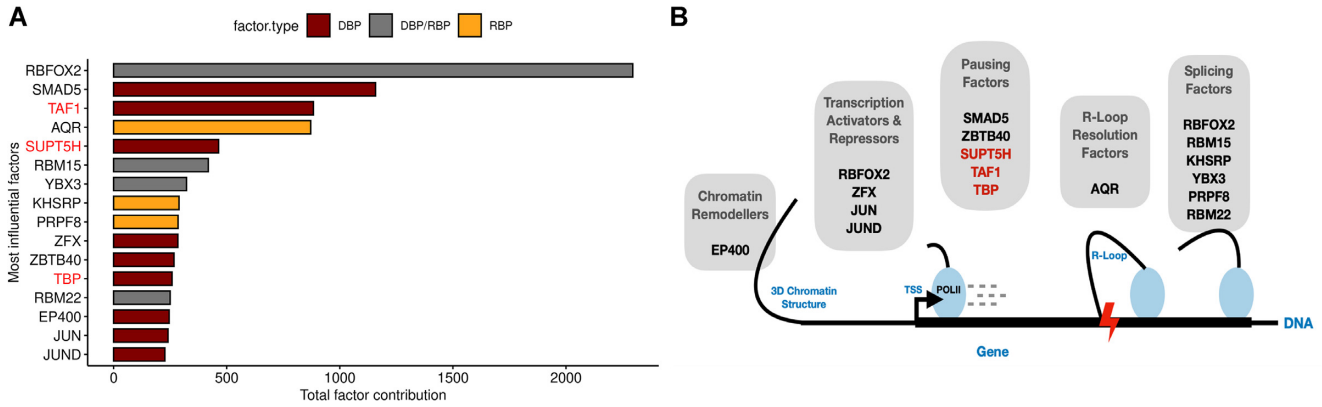


Figure 4. (**A**) Increasingly ordered aggregate factor contributions of factors that make up at least 50% of model contributions. Established pausing/elongation factors are colored red. The bar fill colors identify DNA-binding (DBP; dark red), RNA-binding (RBP; orange), or DNA- and RNA-binding (DBP/RBP; grey) factors. (**B**) A conceptual view on the interconnection and interplay of identified transcriptional pause regulatory proteins with associated transcriptional regulatory processes (Chromatin Remodelling, Transcription Activation/Repression, Transcriptional Pausing, R-Loop resolution and Splicing).

factors consists of RBFOX2, SMAD5, TAF1, AQR, SUPT5H RBM15, YBX3, KHSRP, ZFX, TBP and EP400. These factors represent confident regulators of transcriptional pausing as they are selected across different sequencing protocols.

Upon investigation of the identified most influential pausing factors (n = 16, K562) defined by our model the interconnection of pausing with other RNA-processing events becomes further apparent. An interesting picture emerges considering the functional background of these factors (Figure 4B).

Pausing factors

Several pausing factors are well established (TAF1, TBP, SUPT5H) and occupy high ranks in our models. TAF1 and TBP are components of the pre-initiation complex (PIC). Its formation inherently leads to pausing (61). This behavior can be modulated by other pausing factors, especially the protein complexes NELF and DSIF (SUPT5H) increase pausing whereas the P-TEFb complex associates with pause release.

Chromatin remodelers

The chromatin remodeler EP400 had a large impact on our model. Chromatin state is defined by nucleosome positioning and posttranslational modification of its histones. It is tightly linked to transcription initiation, elongation, and co-transcriptional splicing and can be actively modulated by chromatin remodelers (84–87). EP400 is a histone acetyltransferase and promotes gene activation after PIC assembly through the depositioning of H3.3/H2.AZ into promoters and enhancers (88). It interacts with the well-known pausing factor MYC (27,88,89) and might be linked to transcriptional pausing through this association. In fact, regulation of Pol II pausing at promoter-proximal nucleosomes by chromatin remodelers like for instance CHD1 (90) has already been established.

Transcriptional repressors and activators

Among the top influential factors we can find activating transcription factors ZFX, JUN, and JUND as well RBFOX2 as a repressive transcription factor. ZFX family members exert a transcription-activating function in multiple types of human tumors and bind downstream from the TSS at the majority of CpG island promoters regulating genes for essential housekeeping functions. ZFX family members have been suggested to act in a similar manner as the MYC family of transcription factors due to their shared pervasive binding at promoter sites as well as similar profound proliferation defects upon knockdown (91,92). Given that MYC plays an important role in transcriptional pause release through the recruitment of P-TEFb (27,93), a similar connection could exist for ZFX. Moreover, a comparison of the binding patterns of ZFX with Pol II and H3K4me3 has shown that ZFX is slightly downstream from the most frequent Pol II pause site and slightly upstream of the downstream peak of the H3K4me3 signal (91,92), further suggesting a role of ZFX in regulating Pol II pausing.

JUN and JUND are subcomponents of the activating protein 1 (AP-1) (94,95) which in turn controls cell proliferation, neoplastic transformation, apoptosis, and the expression of immune mediators. AP-1 is suppressed by the negative elongation factor NELF (96), but so far no regulation of transcriptional pausing by AP-1 has been reported.

RBFOX2 acts both, as a regulator of alternative splicing as discussed later, and transcriptional repressor through the binding to chromatin-associated RNA, especially promoter-proximal nascent RNA, through the recruitment of the polycomb-complex 2 (PRC2) to its site of action (91,97,98). In fact, knockout of *RBFOX2* in cardiomyocytes leads to decreased pausing indices and suggests that RBFOX2 and PRC2 enhance coordinated transcriptional pausing at gene promoters (98).

Co-transcriptional splicing and mRNA regulatory factors

The presence of several splicing-associated factors (RBFOX2, PRPF8, RBM15, RBM22, KHSRP, YBX3,

AQR) further strengthens the intricate connection to co-transcriptional splicing events (33,74,99,100). Co-transcriptional splicing of pre-mRNAs is dependent on the availability of the nascent RNA that forms during the transcriptional cycle which in turn is a function of Pol II pausing. In fact, it has been shown that active spliceosomes are complexed to the Pol II S5P C-terminal domain during elongation and co-transcriptional splicing (101). In particular, it has also been shown that transcription kinetics strongly impact splicing decisions, such that slow Pol II elongation rates allow more time for spliceosome assembly and thereby favor splicing. Moreover, the inhibition of the spliceosomal U2 snRNP function has been shown to enhance Pol II pausing in promoter-proximal regions, impair the recruitment of P-TEFb and thereby reduce Pol II elongation velocity at the beginning of genes (82). These indicated that the release of paused Pol II requires the formation of functional spliceosomes and that positive feedback from the splicing machinery to the transcription machinery exists. In this context, RBFOX2 acts as a well-established regulator of alternative splicing (102–104) with an integral role in transcriptional pausing (98). Likewise, RBM15 (105), RBM22 (106,107), PRPF8 (108), KHSRP (109), and YBX3 (110) as pre-mRNA splicing factors or spliceosome components are likely to have a similar connection to pausing as is the case for RBFOX2 and splicing in general.

AQR is a high-ranking R-loop resolution factor (111). R-loops are RNA/DNA structures in which nascent RNA anneals back to the template DNA (112–115). It has also been suggested that R-loop formation is likely part of the mechanism for Pol II pausing (114) to hold back the elongation of Pol II (116) and the DNA replisome (117). The importance of splicing events for pausing is further strengthened by splice defect-induced R-loop formations as a result of increased RNA-DNA hybrid annealing due to the lack of splicing-dependent nascent RNA processing which would otherwise prevent the formation of such structures through timely splicing events.

Novel pausing factors

For the factors ZBTB40 and SMAD5 not previously associated with the regulation of pausing we suggest a novel link. ZBTB40 is not well characterized but has been established to be a regulator of osteoblast activity and bone mass (118). SMAD5, together with other SMAD proteins, is a signal transducer and is activated in the cytoplasm and accumulated in the nucleus where it regulates transcription via remodeling of the chromatin architecture through the recruitment of a variety of coactivators and corepressors to the chromatin (91,97), suggesting a role regulating transcriptional pausing outcomes through a series of chromatin remodeling events and recruitment of transcription factors.

DISCUSSION

The understanding of promoter-proximal Pol II pause regulatory elements is an important step towards disentangling the gene regulatory mechanisms underlying cell homeostasis and plasticity. We improved our understanding by training machine learning models that predict the extent of promoter proximal pausing from large-scale genome and transcriptome binding maps, as well as gene annotation and sequence composition features providing insights into cis- and trans-acting regulatory elements underlying transcriptional pausing. Recent models of transcriptional pausing based on random forests in the Hela cell line (18) focused on NET-seq derived pause sites that are not necessarily promoter proximal. This model solely incorporated DNA sequence features like DNA structures (Z-DNA, repeats etc.), methylation states or transcription factor binding motifs. This is similar to another recent machine-learning approach with a deep-learning architecture called PEPMAN (Feng et al. 2021) to systematically model Pol II pausing events from high-throughput sequencing data based on raw DNA sequence input features. The author's also suggest a strong connection of transcriptional pausing to co-transcriptional splicing events which is very much in line with our results. In contrast to both approaches, our model relies on experimentally determined binding sites of both DNA and RNA binding proteins, which integrate information on the presence of binding sites but also on the cellular context. For example, not all binding motifs are necessarily bound by trans-acting factors in all cell lines.

Our model achieves high predictive accuracy ($R^2 \sim 0.68$ with $n = 389$, factors; $R^2 \sim 0.65$ with only $n = 16$ factors), indicating that the binding of identified trans-acting protein factors to DNA and RNA explains a large part of the variability of the extent of pausing. The accurate prediction of differential pausing based on cross-cell type specific binding data ($R^2 \sim 0.52$) demonstrated that the model learned general rules, which are not cell type specific. This is in line with the observation that the pausing of genes is consistent across a large proportion of cell types (12). Models built from subsets of proteins implicated in all steps of gene expression, including chromatin remodeling, transcription initiation, elongation, splicing, and further downstream transcript processing demonstrated high predictive power. This confirms the intimate cross-talk between these processes (8,16,50,69,70,73,119,120). Of note, factors implicated in splicing have the highest predictive power for pausing. This is in line with many studies that show dual roles for individual proteins such as RBFOX2 (102–104), SRSF2 (33), U2AF65 (82) or MAGOH (82) providing a direct causal link between the two processes. One important goal of our analysis was to identify novel potential pausing regulators. We achieved this using two approaches. First, we identified novel *7SK* binding RBPs and showed that their binding patterns are highly predictive of pausing. Second, we analyzed the feature importance in our model and pinpointed protein factors with higher feature importance than established pausing factors. Many of these factors such as RBOFX2 (102–104), AQR (111), JUN, and JUND (94) have been demonstrated to affect pausing or are implicated in processes that have already been associated with pausing. These factors constitute interesting targets for further experimental validation, as our results already provide some initial mechanistic hypotheses.

We chose to analyze data from the HepG2 and K562 cell lines since they have been extensively characterized in the ENCODE project. The number of DNA and RNA bind-

ing maps available is unparalleled and enables the identification of previously unknown regulators of promoter proximal pausing. These data sets come with the limitation that not all previously characterized regulators of pausing are available. The second limitation is that only GRO-seq data and similar variations are available to quantify promoter proximal pausing. Recent multi-omics approaches based on TT-seq (121) and mNET-seq (7,8,122) have been applied to K562 and Raji B cell lines to estimate the kinetic rates of initiation and pause duration more precisely. These approaches provide ground for future studies of transcriptional pausing with greater precision and detail once broadly available across cell lines which would enable elaborate validation procedures. Unfortunately, such data are not available for a second ENCODE cell line such that a cross-validation of the model would not be possible. Taken together, our work provides a framework to further our understanding of the regulation of the critical early steps in transcriptional elongation. We expect further improvements with better kinetic profiling of the polymerase and increasing availability of binding maps or improved prediction of binding sites from sequence.

DATA AVAILABILITY

The code is available at https://github.com/heiniglab/POLII_pausing. All data and results are also available at 10.5281/zenodo.5236311.

ACCESSION NUMBERS

GRO-seq data for the K562, HepG2, and Hela cell lines were obtained from studies with GEO accessions ***GSM1480325***, ***GSM2428726,*** and **GSE62046** respectively. GRO-cap data for K562 cell line was taken from ***GSM1480322*.** RNA-seq data transcript quantifications data sets (tsv-files) were taken from ENCODE from the experiment **ENCSR885DVH** with accession numbers of replicated experiments **ENCFF424CXV** and **ENCFF073NHK** for the K562 cell line, as well as the experiment **ENCSR181ZGR** with accession numbers of replicated experiments **ENCFF205WUQ**, **ENCFF915JUZ** for the HepG2 cell line. RNA-seq data for the Hela cell line was taken from **GSM2400170**. ENCODE Accession number of CHIP-seq and eCLIP-seq data sets can be found in supplementary tables S4 - S6 and S10 & S11, respectively. Annotations of housekeeping genes were taken from (57) (see Supplementary Table S27; housekeeping.RDS in zenodo repository). CpG island annotations were taken from the UCSC golden path for the hg19 genome build (cpgIslandExt.txt.gz) (see Supplementary Table S27; cpg.islands.RDS in zenodo repository). Gene annotations along with HGNC and RefSeq metadata files were taken from GENCODE (see Supplementary Table S27). CAGE transcription start sites for all cell lines are provided in the zenodo repository as an R-data structure (CTSS.RDS).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge the ENCODE Consortium and the ENCODE production laboratories generating the particular datasets. We would like to acknowledge insightful discussions with Dr Jernej Murn.

FUNDING

MH is supported by funding from the Federal Ministry of Education and Research (BMBF) within the e:med program [01ZX1408D, 01ZX1708G to M.H.]; German Center for Cardiovascular Research (DZHK) [81Z0600106, 81Z0600105]. Funding for open access charge: own institutional funds.

Conflict of interest statement. None declared.

REFERENCES

1. Lin,J. and Amir,A. (2018) Homeostasis of protein and mRNA concentrations in growing cells. *Nat. Commun.*, **9**, 4496.
2. Sallie,R. (2004) Transcriptional homeostasis: a mechanism of protein quality control. *Med. Hypotheses*, **63**, 232–234.
3. Mitsis,T., Efthimiadou,A., Bacopoulou,F., Vlachakis,D., Chrousos,G.P. and Eliopoulos,E. (2020) Transcription factors and evolution: an integral part of gene expression (Review). *World Acad. Sci. J.*, **2**, 3–8.
4. Schier,A.C. and Taatjes,D.J. (2020) Structure and mechanism of the RNA polymerase II transcription machinery. *Genes Dev.*, **34**, 465–488.
5. Malik,S., Molina,H. and Xue,Z. (2017) PIC activation through functional interplay between mediator and TFIIH. *J. Mol. Biol.*, **429**, 48–63.
6. Wissink,E.M., Vihervaara,A., Tippens,N.D. and Lis,J.T. (2019) Nascent RNA analyses: tracking transcription and its regulation. *Nat. Rev. Genet.*, **20**, 705–723.
7. Gressel,S., Schwalb,B., Decker,T.M., Qin,W., Leonhardt,H., Eick,D. and Cramer,P. (2017) CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife*, **6**, e29736.
8. Gressel,S., Schwalb,B. and Cramer,P. (2019) The pause-initiation limit restricts transcription activation in human cells. *Nat. Commun.*, **10**, 3603.
9. Cheng,C. and Sharp,P.A. (2003) RNA polymerase II accumulation in the promoter-proximal region of the dihydrofolate reductase and gamma-actin genes. *Mol. Cell. Biol.*, **23**, 1961–1967.
10. Adelman,K. and Lis,J.T. (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.*, **13**, 720–731.
11. Gardini,A. (2017) Global run-On sequencing (GRO-Seq). *Methods Mol. Biol.*, **1468**, 111–120.
12. Day,D.S., Zhang,B., Stevens,S.M., Ferrari,F., Larschan,E.N., Park,P.J. and Pu,W.T. (2016) Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. *Genome Biol.*, **17**, 120.
13. Bartman,C.R., Hamagami,N., Keller,C.A., Giardine,B., Hardison,R.C., Blobel,G.A. and Raj,A. (2019) Transcriptional burst initiation and polymerase pause release are key control points of Transcriptional regulation. *Mol. Cell*, **73**, 519–532.
14. Reppas,N.B., Wade,J.T., Church,G.M. and Struhl,K. (2006) The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting. *Mol. Cell*, **24**, 747–757.
15. Dollinger,R. and Gilmour,D.S. (2021) Regulation of promoter proximal pausing of RNA polymerase II in metazoans. *J. Mol. Biol.*, **433**, 166897.
16. Core,L. and Adelman,K. (2019) Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes Dev.*, **33**, 960–982.
17. Krumm,A., Hickey,L.B. and Groudine,M. (1995) Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev.*, **9**, 559–572.

18. Gajos,M., Jasnovidova,O., van Bömmel,A., Freier,S., Vingron,M. and Mayer,A. (2021) Conserved DNA sequence features underlie pervasive RNA polymerase pausing. *Nucleic Acids Res.*, **49**, 4402–4420.
19. Watts,J.A., Burdick,J., Daigneault,J., Zhu,Z., Grunseich,C., Bruzel,A. and Cheung,V.G. (2019) cis elements that mediate RNA polymerase II pausing regulate Human gene expression. *Am. J. Hum. Genet.*, **105**, 677–688.
20. Castelo-Branco,G., Amaral,P.P., Engström,P.G., Robson,S.C., Marques,S.C., Bertone,P. and Kouzarides,T. (2013) The non-coding snRNA 7SK controls transcriptional termination, poising, and bidirectionality in embryonic stem cells. *Genome Biol.*, **14**, R98.
21. Diribarne,G. and Bensaude,O. (2009) 7SK RNA, a non-coding RNA regulating P-TEFb, a general transcription factor. *RNA Biol.*, **6**, 122–128.
22. Peterlin,B.M., Brogie,J.E. and Price,D.H. (2012) 7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription. *Wiley Interdiscip. Rev. RNA*, **3**, 92–103.
23. Feng,P., Xiao,A., Fang,M., Wan,F., Li,S., Lang,P., Zhao,D. and Zeng,J. (2021) A machine learning-based framework for modeling transcription elongation. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2007450118.
24. Paparidis,N.F.D.S., Durvale,M.C. and Canduri,F. (2017) The emerging picture of CDK9/P-TEFb: more than 20 years of advances since PITALRE. *Mol. Biosyst.*, **13**, 246–276.
25. Brès,V., Yoh,S.M. and Jones,K.A. (2008) The multi-tasking P-TEFb complex. *Curr. Opin. Cell Biol.*, **20**, 334–340.
26. Chao,S.H. and Price,D.H. (2001) Flavopiridol inactivates P-TEFb and blocks most RNA polymerase II transcription in vivo. *J. Biol. Chem.*, **276**, 31793–31799.
27. Rahl,P.B., Lin,C.Y., Seila,A.C., Flynn,R.A., McCuine,S., Burge,C.B., Sharp,P.A. and Young,R.A. (2010) c-Myc regulates transcriptional pause release. *Cell*, **141**, 432–445.
28. Henriques,T., Gilchrist,D.A., Nechaev,S., Bern,M., Muse,G.W., Burkholder,A., Fargo,D.C. and Adelman,K. (2013) Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol. Cell*, **52**, 517–528.
29. Jonkers,I., Kwak,H. and Lis,J.T. (2014) Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife*, **3**, e02407.
30. Ni,Z., Saunders,A., Fuda,N.J., Yao,J., Suarez,J.-R., Webb,W.W. and Lis,J.T. (2008) P-TEFb is critical for the maturation of RNA polymerase II into productive elongation in vivo. *Mol. Cell. Biol.*, **28**, 1161–1170.
31. Schröder,S., Cho,S., Zeng,L., Zhang,Q., Kaehlcke,K., Mak,L., Lau,J., Bisgrove,D., Schnölzer,M., Verdin,E. *et al.* (2012) Two-pronged binding with bromodomain-containing protein 4 liberates positive transcription elongation factor b from inactive ribonucleoprotein complexes. *J. Biol. Chem.*, **287**, 1090–1099.
32. Sansó,M., Levin,R.S., Lipp,J.J., Wang,V.Y.-F., Greifenberg,A.K., Quezada,E.M., Ali,A., Ghosh,A., Larochelle,S., Rana,T.M. *et al.* (2016) P-TEFb regulation of transcription termination factor Xrn2 revealed by a chemical genetic screen for Cdk9 substrates. *Genes Dev.*, **30**, 117–131.
33. Akhtar,J., Kreim,N., Marini,F., Mohana,G., Brüne,D., Binder,H. and Roignant,J.-Y. (2019) Promoter-proximal pausing mediated by the exon junction complex regulates splicing. *Nat. Commun.*, **10**, 521.
34. Yonaha,M. and Proudfoot,N.J. (1999) Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. *Mol. Cell*, **3**, 593–600.
35. Ji,X., Zhou,Y., Pandit,S., Huang,J., Li,H., Lin,C.Y., Xiao,R., Burge,C.B. and Fu,X.-D. (2013) SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. *Cell*, **153**, 855–868.
36. Guo,Y.E., Manteiga,J.C., Henninger,J.E., Sabari,B.R., Dall'Agnese,A., Hannett,N.M., Spille,J.-H., Afeyan,L.K., Zamudio,A.V., Shrinivas,K. *et al.* (2019) Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature*, **572**, 543–548.
37. Peterlin,B.M. and Price,D.H. (2006) Controlling the elongation phase of transcription with P-TEFb. *Mol. Cell*, **23**, 297–305.
38. Takahashi,H., Parmely,T.J., Sato,S., Tomomori-Sato,C., Banks,C.A.S., Kong,S.E., Szutorisz,H., Swanson,S.K., Martin-Brown,S., Washburn,M.P. *et al.* (2011) Human mediator subunit MED26 functions as a docking site for transcription elongation factors. *Cell*, **146**, 92–104.
39. Calo,E., Flynn,R.A., Martin,L., Spitale,R.C., Chang,H.Y. and Wysocka,J. (2015) RNA helicase DDX21 coordinates transcription and ribosomal RNA processing. *Nature*, **518**, 249–253.
40. Van Nostrand,E.L., Freese,P., Pratt,G.A., Wang,X., Wei,X., Xiao,R., Blue,S.M., Chen,J.-Y., Cody,N.A.L., Dominguez,D. *et al.* (2020) A large-scale binding and functional map of human RNA-binding proteins. *Nature*, **583**, 711–719.
41. Frankish,A., Diekhans,M., Ferreira,A.-M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J., Armstrong,J. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.
42. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
43. Tweedie,S., Braschi,B., Gray,K., Jones,T.E.M., Seal,R.L., Yates,B. and Bruford,E.A. (2021) Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.*, **49**, D939–D946.
44. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
45. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
46. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J. *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.
47. Noguchi,S., Arakawa,T., Fukuda,S., Furuno,M., Hasegawa,A., Hori,F., Ishikawa-Kato,S., Kaida,K., Kaiho,A., Kanamori-Katayama,M. *et al.* (2017) FANTOM5 CAGE profiles of human and mouse samples. *Sci Data*, **4**, 170112.
48. Frith,M.C., Valen,E., Krogh,A., Hayashizaki,Y., Carninci,P. and Sandelin,A. (2008) A code for transcription initiation in mammalian genomes. *Genome Res.*, **18**, 1–12.
49. Lopes,R., Agami,R. and Korkmaz,G. (2017) GRO-seq, A tool for identification of transcripts regulating gene expression. *Methods Mol. Biol.*, **1543**, 45–55.
50. Shao,W. and Zeitlinger,J. (2017) Paused RNA polymerase II inhibits new transcriptional initiation. *Nat. Genet.*, **49**, 1045–1051.
51. Park,P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
52. Li,Q., Brown,J.B., Huang,H. and Bickel,P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann Appl Stat*, **5**, 1752–1779.
53. Van Nostrand,E.L., Pratt,G.A., Shishkin,A.A., Gelboin-Burkhart,C., Fang,M.Y., Sundararaman,B., Blue,S.M., Nguyen,T.B., Surka,C., Elkins,K. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.
54. Hu,B., Yang,Y.-C.T., Huang,Y., Zhu,Y. and Lu,Z.J. (2017) POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res.*, **45**, D104–D114.
55. Salmena,L., Poliseno,L., Tay,Y., Kats,L. and Pandolfi,P.P. (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?*Cell*, **146**, 353–358.
56. Kellner,W.A., Bell,J.S.K. and Vertino,P.M. (2015) GC skew defines distinct RNA polymerase pause sites in CpG island promoters. *Genome Res.*, **25**, 1600–1609.
57. Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet.*, **29**, 569–574.
58. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
59. Gene Ontology Consortium (2021) The Gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
60. Luo,Z., Lin,C. and Shilatifard,A. (2012) The super elongation complex (SEC) family in transcriptional control. *Nat. Rev. Mol. Cell Biol.*, **13**, 543–547.

61. Fant,C.B., Levandowski,C.B., Gupta,K., Maas,Z.L., Moir,J., Rubin,J.D., Sawyer,A., Esbin,M.N., Rimel,J.K., Luyties,O. *et al.* (2020) TFIID enables RNA polymerase II promoter-Proximal pausing. *Mol. Cell*, **78**, 785–793.
62. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
63. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
64. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
65. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. *et al.* (2017) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
66. Lundberg,S.M. and Lee,S.-I. (2017) A Unified Approach to Interpreting Model Predictions. In: Guyon,I., Von Luxburg,U., Bengio,S., Wallach,H., Fergus,R., Vishwanathan,S. and Garnett,R. (eds). *Advances in Neural Information Processing Systems*. Curran Associates, Inc, Vol. **30**.
67. Lundberg,S.M., Erion,G., Chen,H., DeGrave,A., Prutkin,J.M., Nair,B., Katz,R., Himmelfarb,J., Bansal,N. and Lee,S.-I. (2020) From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, **2**, 56–67.
68. Lundberg,S.M., Nair,B., Vavilala,M.S., Horibe,M., Eisses,M.J., Adams,T., Liston,D.E., Low,D.K.-W., Newman,S.-F., Kim,J. *et al.* (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.*, **2**, 749–760.
69. Gilchrist,D.A. and Adelman,K. (2012) Coupling polymerase pausing and chromatin landscapes for precise regulation of transcription. *Biochim. Biophys. Acta*, **1819**, 700–706.
70. Gilchrist,D.A., Dos Santos,G., Fargo,D.C., Xie,B., Gao,Y., Li,L. and Adelman,K. (2010) Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation*Cell*, **143**, 540–551.
71. Vaid,R., Wen,J. and Mannervik,M. (2020) Release of promoter-proximal paused pol II in response to histone deacetylase inhibition. *Nucleic Acids Res.*, **48**, 4877–4890.
72. Lerner,E., Ingargiola,A., Lee,J.J., Borukhov,S., Michalet,X. and Weiss,S. (2017) Different types of pausing modes during transcription initiation. *Transcription*, **8**, 242–253.
73. Saldi,T., Cortazar,M.A., Sheridan,R.M. and Bentley,D.L. (2016) Coupling of RNA polymerase II transcription elongation with pre-mRNA splicing. *J. Mol. Biol.*, **428**, 2623–2635.
74. Carrillo Oesterreich,F., Bieberstein,N. and Neugebauer,K.M. (2011) Pause locally, splice globally. *Trends Cell Biol.*, **21**, 328–335.
75. Fusby,B., Kim,S., Erickson,B., Kim,H., Peterson,M.L. and Bentley,D.L. (2016) Coordination of RNA polymerase II pausing and 3′ End processing factor recruitment with alternative polyadenylation. *Mol. Cell. Biol.*, **36**, 295–303.
76. Ishov,A.M., Gurumurthy,A. and Bungert,J. (2020) Coordination of transcription, processing, and export of highly expressed rnas by distinct biomolecular condensates. *Emerg Top Life Sci*, **4**, 281–291.
77. McNamara,R.P., Bacon,C.W. and D'Orso,I. (2016) Transcription elongation control by the 7SK snRNP complex: releasing the pause. *Cell Cycle*, **15**, 2115–2123.
78. Studniarek,C., Tellier,M., Martin,P.G.P., Murphy,S., Kiss,T. and Egloff,S. (2021) The 7SK/P-TEFb snRNP controls ultraviolet radiation-induced transcriptional reprogramming. *Cell Rep.*, **35**, 108965.
79. Quaresma,A.J., Bugai,A. and Barboric,M. (2016) Cracking the control of RNA polymerase II elongation by 7SK snRNP and P-TEFb. *Nucleic Acids Res.*, **44**, 7527–7539.
80. Barboric,M., Lenasi,T., Chen,H., Johansen,E.B., Guo,S. and Peterlin,B.M. (2009) 7SK snRNP/P-TEFb couples transcription elongation with alternative splicing and is essential for vertebrate development. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 7798–7803.
81. Egloff,S., Vitali,P., Tellier,M., Raffel,R., Murphy,S. and Kiss,T. (2017) The 7SK snRNP associates with the little elongation complex to promote snRNA gene expression. *EMBO J.*, **36**, 934–948.
82. Caizzi,L., Monteiro-Martins,S., Schwalb,B., Lysakovskaia,K., Schmitzova,J., Sawicka,A., Chen,Y., Lidschreiber,M. and Cramer,P. (2021) Efficient RNA polymerase II pause release requires U2 snRNP function. *Mol. Cell*, **81**, 1920–1934.
83. Sheridan,R.M., Fong,N., D'Alessandro,A. and Bentley,D.L. (2019) Widespread Backtracking by RNA Pol II Is a Major Effector of Gene Activation, 5′ Pause Release, Termination, and Transcription Elongation Rate. *Mol. Cell*, **73**, 107–118.
84. Lorch,Y. and Kornberg,R.D. (2017) Chromatin-remodeling for transcription. *Q. Rev. Biophys.*, **50**, e5.
85. Smolle,M., Workman,J.L. and Venkatesh,S. (2013) reSETting chromatin during transcription elongation. *Epigenetics*, **8**, 10–15.
86. Zraly,C.B. and Dingwall,A.K. (2012) The chromatin remodeling and mRNA splicing functions of the Brahma (SWI/SNF) complex are mediated by the SNR1/SNF5 regulatory subunit. *Nucleic Acids Res.*, **40**, 5975–5987.
87. Schwartz,S. and Ast,G. (2010) Chromatin density and splicing destiny: on the cross-talk between chromatin structure and splicing. *EMBO J.*, **29**, 1629–1636.
88. Pradhan,S.K., Su,T., Yen,L., Jacquet,K., Huang,C., Côté,J., Kurdistani,S.K. and Carey,M.F. (2016) EP400 Deposits H3.3 into Promoters and Enhancers during Gene Activation. *Mol. Cell*, **61**, 27–38.
89. Fuchs,M., Gerber,J., Drapkin,R., Sif,S., Ikura,T., Ogryzko,V., Lane,W.S., Nakatani,Y. and Livingston,D.M. (2001) The p400 complex is an essential E1A transformation target. *Cell*, **106**, 297–307.
90. Chiu,A.C., Suzuki,H.I., Wu,X., Mahat,D.B., Kriz,A.J. and Sharp,P.A. (2018) Transcriptional Pause Sites Delineate Stable Nucleosome-Associated Premature Polyadenylation Suppressed by U1 snRNP*Mol. Cell*, **69**, 648–663.
91. Rhie,S.K., Yao,L., Luo,Z., Witt,H., Schreiner,S., Guo,Y., Perez,A.A. and Farnham,P.J. (2018) ZFX acts as a transcriptional activator in multiple types of human tumors by binding downstream of transcription start sites at the majority of CpG island promoters. *Genome Res.*, **28**, 310–320.
92. Ni,W., Perez,A.A., Schreiner,S., Nicolet,C.M. and Farnham,P.J. (2020) Characterization of the ZFX family of transcription factors that bind downstream of the start site of CpG island promoters. *Nucleic Acids Res.*, **48**, 5986–6000.
93. Rahl,P.B. and Young,R.A. (2014) MYC and transcription elongation. *Cold Spring Harb. Perspect. Med.*, **4**, a020990.
94. Shaulian,E. and Karin,M. (2001) AP-1 in cell proliferation and survival. *Oncogene*, **20**, 2390–2400.
95. Gazon,H., Barbeau,B., Mesnard,J.-M. and Peloponese,J.-M. Jr (2017) Hijacking of the AP-1 signaling pathway during development of ATL. *Front Microbiol*, **8**, 2686.
96. Yu,L., Zhang,B., Deochand,D., Sacta,M.A., Coppo,M., Shang,Y., Guo,Z., Zeng,X., Rollins,D.A., Tharmalingam,B. *et al.* (2020) Negative elongation factor complex enables macrophage inflammatory responses by controlling anti-inflammatory gene expression. *Nat. Commun.*, **11**, 2286.
97. Hill,C.S. (2016) Transcriptional control by the SMADs. *Cold Spring Harb. Perspect. Biol.*, **8**, a022079.
98. Wei,C., Xiao,R., Chen,L., Cui,H., Zhou,Y., Xue,Y., Hu,J., Zhou,B., Tsutsui,T., Qiu,J. *et al.* (2016) RBFox2 Binds nascent RNA to globally regulate polycomb complex 2 targeting in mammalian genomes. *Mol. Cell*, **62**, 875–889.
99. Alexander,R.D., Innocente,S.A., Barrass,J.D. and Beggs,J.D. (2010) Splicing-dependent RNA polymerase pausing in yeast. *Mol. Cell*, **40**, 582–593.
100. Andersen,P.K. and Jensen,T.H. (2010) A pause to splice. *Mol. Cell*, **40**, 503–505.
101. Nojima,T., Rebelo,K., Gomes,T., Grosso,A.R., Proudfoot,N.J. and Carmo-Fonseca,M. (2018) RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing*Mol. Cell*, **72**, 369–379.
102. Braeutigam,C., Rago,L., Rolke,A., Waldmeier,L., Christofori,G. and Winter,J. (2013) The RNA-binding protein Rbfox2: an essential

regulator of EMT-driven alternative splicing and a mediator of cellular invasion. *Oncogene*, **33**, 1082–1092.
103. Ying,Y., Wang,X.-J., Vuong,C.K., Lin,C.-H., Damianov,A. and Black,D.L. (2017) Splicing Activation by Rbfox Requires Self-Aggregation through Its Tyrosine-Rich Domain*Cell*, **170**, 312–323.
104. Quentmeier,H., Pommerenke,C., Bernhart,S.H., Dirks,W.G., Hauer,V., Hoffmann,S., Nagel,S., Siebert,R., Uphoff,C.C., Zaborski,M. *et al.* (2018) RBFOX2 and alternative splicing in B-cell lymphoma. *Blood Cancer J.*, **8**, 77.
105. Zhang,L., Tran,N.-T., Su,H., Wang,R., Lu,Y., Tang,H., Aoyagi,S., Guo,A., Khodadadi-Jamayran,A., Zhou,D. *et al.* (2015) Cross-talk between PRMT1-mediated methylation and ubiquitylation on RBM15 controls RNA splicing. *Elife*, **4**, e07938.
106. Xiao,R., Chen,J.-Y., Liang,Z., Luo,D., Chen,G., Lu,Z.J., Chen,Y., Zhou,B., Li,H., Du,X. *et al.* (2019) Pervasive chromatin-RNA binding protein interactions enable RNA-Based regulation of transcription. *Cell*, **178**, 107–121.
107. Rasche,N., Dybkov,O., Schmitzová,J., Akyildiz,B., Fabrizio,P. and Lührmann,R. (2012) Cwc2 and its human homologue RBM22 promote an active conformation of the spliceosome catalytic centre. *EMBO J.*, **31**, 1591–1604.
108. Wickramasinghe,V.O., Gonzàlez-Porta,M., Perera,D., Bartolozzi,A.R., Sibley,C.R., Hallegger,M., Ule,J., Marioni,J.C. and Venkitaraman,A.R. (2015) Regulation of constitutive and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5′ splice site strength. *Genome Biol.*, **16**, 201.
109. Briata,P., Bordo,D., Puppo,M., Gorlero,F., Rossi,M., Perrone-Bizzozero,N. and Gherzi,R. (2016) Diverse roles of the nucleic acid-binding protein KHSRP in cell differentiation and disease. *Wiley Interdiscip. Rev. RNA*, **7**, 227–240.
110. Rambout,X., Dequiedt,F. and Maquat,L.E. (2018) Beyond transcription: roles of transcription factors in pre-mRNA splicing. *Chem. Rev.*, **118**, 4339–4364.
111. Sollier,J., Stork,C.T., García-Rubio,M.L., Paulsen,R.D., Aguilera,A. and Cimprich,K.A. (2014) Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. *Mol. Cell*, **56**, 777–785.
112. Vidal,M. and Starowicz,K. (2017) Polycomb complexes PRC1 and their function in hematopoiesis. *Exp. Hematol.*, **48**, 12–31.
113. Pherson,M., Misulovin,Z., Gause,M., Mihindukulasuriya,K., Swain,A. and Dorsett,D. (2017) Polycomb repressive complex 1 modifies transcription of active genes. *Sci. Adv.*, **3**, e1700944.
114. Chen,L., Chen,J.-Y., Zhang,X., Gu,Y., Xiao,R., Shao,C., Tang,P., Qian,H., Luo,D., Li,H. *et al.* (2017) R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters*Mol. Cell*, **68**, 745–757.
115. Chédin,F. (2016) Nascent connections: r-Loops and chromatin patterning. *Trends Genet.*, **32**, 828–838.
116. Huertas,P. and Aguilera,A. (2003) Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. *Mol. Cell*, **12**, 711–721.
117. Tuduri,S., Crabbé,L., Conti,C., Tourrière,H., Holtgreve-Grez,H., Jauch,A., Pantesco,V., De Vos,J., Thomas,A., Theillet,C. *et al.* (2009) Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. *Nat. Cell Biol.*, **11**, 1315–1324.
118. Doolittle,M.L., Calabrese,G.M., Mesner,L.D., Godfrey,D.A., Maynard,R.D., Ackert-Bicknell,C.L. and Farber,C.R. (2020) Genetic analysis of osteoblast activity identifies Zbtb40 as a regulator of osteoblast activity and bone mass. *PLoS Genet.*, **16**, e1008805.
119. Gromak,N., West,S. and Proudfoot,N.J. (2006) Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol. Cell. Biol.*, **26**, 3986–3996.
120. Nag,A., Narsinh,K. and Martinson,H.G. (2007) The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. *Nat. Struct. Mol. Biol.*, **14**, 662–669.
121. Schwalb,B., Michel,M., Zacher,B., Frühauf,K., Demel,C., Tresch,A., Gagneur,J. and Cramer,P. (2016) TT-seq maps the human transient transcriptome. *Science*, **352**, 1225–1228.
122. Prudêncio,P., Rebelo,K., Grosso,A.R., Martinho,R.G. and Carmo-Fonseca,M. (2020) Analysis of mammalian native elongating transcript sequencing (mNET-seq) high-throughput data. *Methods*, **178**, 89–95.