Jones *et al. Genome Biology* (2025) 26:92
https://doi.org/10.1186/s13059-025-03545-2

Genome Biology

METHOD

Open Access

TF Profiler: a transcription factor inference method that broadly measures transcription factor activity and identifies mechanistically distinct networks

Taylor Jones[1,2], Rutendo F. Sigauke[1], Lynn Sanford[1], Dylan J. Taatjes[2], Mary A. Allen[1*] and Robin D. Dowell[1,3,4*]

*Correspondence:
mary.a.allen@colorado.edu;
robin.dowell@colorado.edu

[1] BioFrontiers Institute, University of Colorado Boulder, 3415 Colorado Ave., UCB 596, Boulder, CO 80309, USA
[2] Biochemistry, University of Colorado Boulder, 3415 Colorado Ave., UCB 596, Boulder, CO 80309, USA
[3] Computer Science, University of Colorado Boulder, 1111 Engineering Drive, UCB 430, Boulder, CO 80309, USA
[4] Molecular, Cellular and Developmental Biology, University of Colorado Boulder, 1945 Colorado Ave, UCB 347, Boulder, CO 80309, USA

Abstract

TF Profiler is a method of inferring transcription factor (TF) regulatory activity, i.e., when a TF is present and actively participating in the regulation of transcription, directly from nascent sequencing assays such as PRO-seq and GRO-seq. While ChIP assays have measured DNA localization, they fall short of identifying when and where the effector domain of a transcription factor is active. Our method uses RNA polymerase activity to infer TF effector domain activity across hundreds of data sets and transcription factors. TF Profiler is broadly applicable, providing regulatory insights on any PRO-seq sample for any transcription factor with a known binding motif.

Keywords: Transcription factor, Cellular regulation, Tissue specificity

Background

Transcription is a fundamental process that defines cellular function, stress response, and cell identity [1]. The regulation of gene expression patterns is driven by a myriad of sequence-specific transcription factors (TFs) that vary in activity based on both cell type and environmental factors. While there are over 1600 TFs [2] in the human genome, our understanding of how their activity is regulated remains incomplete. For example, there is no consensus on when or where individual TFs are actively altering gene expression patterns.

Transcription factors orchestrate gene regulation programs by altering the activity of cellular RNA polymerases, primarily RNA polymerase II (RNAPII). Some TFs increase transcriptional output (an activator) whereas others decrease transcriptional output (a repressor). Therefore, characterizing when and where TFs are active—not only where they bind in the genome but also when they are actively regulating RNAPII—is necessary to understand their biological function. In fact, one of the goals of the Encyclopedia

of DNA Elements (ENCODE) Consortium was to identify all functional regulatory elements in the human genome [3]. In the ENCODE project, the primary method utilized to assess TF activity was chromatin immunopreciptation (ChIP-seq). ChIP informs on the genomic localization of a TF, which reflects the function of its DNA binding domain typically interacting with DNA in a sequence-specific manner [4].

From ChIP-seq studies, it is possible to infer a position specific scoring matrix (PSSM) for a given DNA binding protein. ChIP-seq studies, however, are low throughput as one sequence-specific protein is evaluated at a time, in one cell type at a time. Furthermore, there is ample evidence that TF binding can occur without altering gene expression [5, 6], as the DNA binding domain is usually independent of the effector domain (also known as the activation domain or repressor domain). The effector domain interacts with co-regulatory factors that directly or indirectly control RNAPII function to alter gene transcription nearby; thus, TFs play crucial role in transcriptional regulation [7, 8].

However, measuring the activity of the effector domain (i.e., measuring TF regulatory activity) has historically been difficult, at least in part because TF regulatory activity can be controlled at multiple stages. For example, TF regulation may occur via changes in protein levels (e.g., TF transcription, translation, or degradation) or through post-translational modifications. Many TFs have well-established mechanisms of activation, such as the MAPK pathway phosphorylation events that result in stabilization and activation of MYC [9], or the inhibition of the ubiquitin ligase HDM2 resulting in the stabilization and activation of p53 [10–12]. In these cases, the MYC and TP53 genes are present at the mRNA and protein level in most cellular conditions, despite being repressed until activated by specific stimuli. Thus, neither transcription of the gene encoding the TF, nor TF-DNA binding guarantees that it will alter RNAPII transcription. The ultimate outcome of TF effector domain activity is a change in transcription, hence nascent transcription assays are well-suited to inform on effector domain activity.

Run-on RNA sequencing (such as precision run-on sequencing, PRO-seq [13, 14] and global run-on sequencing, GRO-seq [15]) provides a direct read out of RNA polymerase activity as RNA is captured from the actively catalyzing cellular polymerases. These nascent run-on RNA assays have revealed extensive genome-wide transcription, at genes and enhancers [16–19], and demonstrated that most sites of RNAPII initiation give rise to bidirectional transcription. While the function of the resulting RNA transcripts within enhancers is incompletely understood, a technical benefit of these transcripts is that their distinctive profile can be used to annotate active enhancers genome wide [20, 21].

Prior studies on individual TFs found that TF activation resulted in concomitant changes in transcript levels associated with a subset of ChIP measured TF binding sites [22–26]. Subsequent work generalized these findings, showing a strong co-association of TF binding sites with sites of RNAPII initiation, the majority of which occurring at enhancers [20, 21]. The model that emerged was that the regulatory activity of the TF (e.g., activity of the effector domain) results in changes to RNAPII initiation proximal to the TF binding motif [21]. Armed with this result, methods were developed to infer changes in TF activity in response to a perturbation, using nascent transcription data and known TF binding motifs [18, 21, 27–30]. The effectiveness of these methods strongly indicates that nascent transcription serves as a functional readout on the

activity of a TF's effector domain. While changes in nascent transcription levels reliably capture changes in TF effector domain activity, only a subset of human TFs are stimulus responsive. Many other TFs are required for cell identity and homeostasis, making perturbation studies more challenging. What is needed is a wide scale analysis of when a TF's effector domain is active, even in the absence of perturbation data.

To this end, we sought to develop an appropriate null hypothesis and statistical framework for predicting TF activity from a single nascent transcription sample—absent any other sample for comparison. To that end, we develop a statistical framework that compares data from an individual nascent transcription sample to a principled, biologically informed statistical expectation. When a TF recognition motif co-localizes with sites of RNAPII initiation more (or less) than expected by chance, we infer that the TF is functional as an activator (or repressor). Importantly, our algorithm can be used to identify all actively regulating TFs in a single sample, a technique we call "TF Profiler." We applied our algorithm to 287 high quality nascent RNA sequencing data sets, representing over 20 different tissues. From this compendium, we identify three classes of TFs: ubiquitous (active in all tissue types), tissue-specific, and stimulus responsive. For example, our method accurately classifies the well known TFs Oct4 and Nanog as active only in embryonic cells. Furthermore, our model uncovered unique sequence features inherent to tissue specific TFs, suggesting a role in the establishment of cell identity.

Results

An expectation model for TF motif co-occurrences

The activity of the TF effector domain alters nascent transcription proximal to sites of TF binding [7]. Based upon this characteristic, methods to infer TF activity changes from nascent transcription data and TF sequence motifs have been developed [18, 21, 27–30]. Our prior work derived a simple metric known as the motif displacement (MD) score, which quantified co-localization of TF recognition motifs in DNA sequence with sites of RNAPII initiation [21]. A subsequent extension to the MD-score approach incorporated changes in transcription levels, effectively capturing when the perturbation leads to changes in the potency of the TF effector domain [27]. In fact, increases in the MD-score were shown to generally be typical of the activation of an activator transcription factor whereas depletion in the score reflected the activation of a repressor.

However, our prior work left it unclear whether the MD-score could be utilized to infer which TFs were actively participating in regulation in homeostatic cellular conditions—i.e., in the absence of a perturbation. To answer this question, we reasoned that a comparison of the original MD-score [17] to a principled, biologically informed statistical model of motif co-localization would allow for the assessment of TF effector domain activity in a single sample. Thus, when a TF recognition motif co-localizes with sites of RNAPII initiation more (or less) than expected by chance, we infer that the TF is actively participating in RNAPII regulation (as an activator or repressor). We refer to this approach as "TF Profiler."

First, let us consider the MD-score metric in a rigorous mathematical framework. Let $X_k = \mu_1, \mu_2, \ldots, \mu_n$ be the RNAPII initiation sites (μ) for a set of bidirectional locations genome-wide for some experiment k. Importantly, sites of bidirectional transcription can be identified directly from nascent transcription data [17, 21, 28, 31, 32]. Let $Y_j =$

$y_1, y_2, \ldots, y_m$ be the set of all significant motif instances for some TF-DNA binding motif model j genome-wide, which is invariant given the genome of interest (Fig. 1A). We can then plot the motif displacement distribution (Fig. 1B) as a heatmap, where heat indicates the number of motif hits (Y_j) relative to the sites of RNAPII initiation (X_k). In this framework, we can calculate the MD-score as:

$$g(X_k, Y_j; a) = \sum_{\mu \in X_k} \sum_{y \in Y_j} \delta(|\mu - y| < a)$$
$$md_{k,j} = g(X_k, Y_j; h)/g(X_k, Y_j; H) \quad (1)$$

where $g()$ quantifies the count of motif hits for a given motif (j) across the complete set of RNAPII initiation sites (X_k). The $\delta(.)$ term is a simple indicator function that returns one if the distance between one RNAPII initiation position (μ) has an instance of the TF-DNA binding motif (y) within a specified distance (a). Hence, the MD-score ($md_{k,j}$) for a given experiment k and TF recognition motif j quantifies co-localization of motif instances near sites of RNAPII initiation ($h = 150$ bps) relative to a larger local window ($H = 1500$ bp).

Importantly, our prior work showed that the value of the MD-score metric depends on precisely defining sites of RNAPII initiation, which is readily accomplished in nascent transcription assays [27]. Furthermore, our Tfit approach [33] to identifying sites of bidirectional transcription was previously shown to be highly precise on the position of

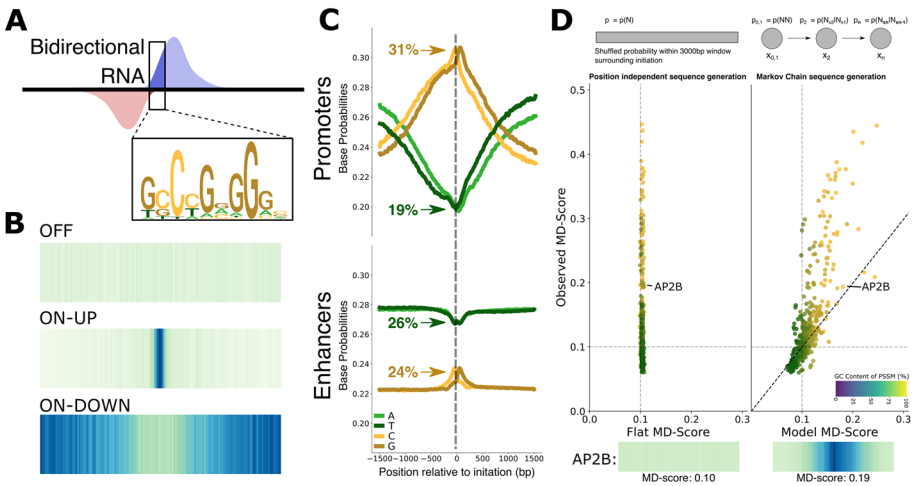

Fig. 1 Overview of the TF profiling model. **A** Cartoon representing the co-localization between bidirectional transcription observed in nascent RNA sequencing (blue and red are data on each strand) and TF motifs. The PSSM for AP2B is shown. This co-localization can be used to assess global motif displacement scores. **B** Heatmaps representing the motif displacement distribution [21] for three distinct TFs with different activation states, OFF (ZN586), ON-UP (SP3), and ON-DOWN (PAX5). The center of the heatmap is the position of the middle of the bidirectional (PolII initiation site) and the heat (darker is more) represents the number of motif instances at that position (relative to the center) genome-wide. **C** Observed promoter (top) and non-promoter (e.g., enhancers, bottom) per position base probabilities surrounding PolII initiation sites show a profound GC bias. In this data set [38], bidirectionals are 30% at promoters (top) and 70% at non-promoters (enhancers, bottom). **D** The observed motif displacement score distribution assuming a flat background and no positional information (left) compared to a position dependent di-nucleotide Markov background (right). Each dot is a single TF position specific scoring matrix, colored by its inherent GC content. The probability (p_i) is defined by the observed probabilities (N) at position i. The position and motif displacement distribution for AP2B is shown with both background models

RNAPII initiation [32]. Using Tfit, we find that nascent run-on RNA transcription assays (e.g., PRO-seq) strike a balance for defining TF activity with precision (comparable to TF ChIP) and scale (comparable to H3K27ac ChIP; Additional file 1: Fig. S1A, B).

To utilize the MD-score as a metric for TF activity, we seek to calculate an odds ratio: the MD-score observed (in a single sample) compared to the expected MD-score (from a statistical model). Conceptually, the expected MD-score must reflect the nucleotide biases of not only the TF-DNA binding motif but also the distinct non-stationary patterns of sequence inherent in genomes. In particular, mammalian genomes have GC-content enrichment at promoters [3, 34] and enhancers [21], consistent with sites of RNAPII initiation. Gene promoters are associated with open chromatin and are highly enriched for CpG islands [35–37]. Whereas the human genome is approximately 60% AT, promoters are approximately 60% GC and enhancers are more modestly GC rich, reaching a nearly equal composition of all four bases (Fig. 1C; see Methods section for promoter and enhancer classification). The difference between enhancer and promoter GC content is statistically significant (Additional file 1: Fig. S2). Importantly, in both cases (enhancers or promoters) the bias is position dependent, reaching a maximum bias coincident with the inferred position of RNAPII loading (μ in our model, inferred by Tfit [33]). We observe that this bias correlates with the overall transcription level, where regions with higher transcription levels tend to display a higher GC content over a broader initiation region (Additional file 1: Fig. S3). Because of the positional base composition bias at RNAPII initiation regions, certain motif instances will be favored (high GC) or disfavored (high AT) by chance alone. Our background expectation model must account for this inherent bias.

Therefore, we took a simulation based approach to the development of the expected MD-score. Specifically, we leverage a dinucleotide model of positional nucleotide preference (Fig. 1D), which accounts for known genomic dinucleotide biases, such as the general preference for CG in CpG islands compared to GC (Additional file 1: Fig. S4). To this end, sequences of the length of $2H$ nucleotides were generated, accounting for dinucleotide preferences in regions of RNAPII initiation. Importantly, the positions i are defined relative to the RNAPII initiation position μ (e.g., the generated sequence is $\mu \pm$ H). Let $x_n = x_1, x_2, \ldots x_{2H}$ where the probability of a specific nucleotide at each x_i is determined based on the nucleotide x_{i-1}. Thus, each position is described by the conditional probability $p(N_{x_i}|N_{x_{i-1}})$, where N represents one of the four nucleotides (A, T, C, or G). The initial dinucleotide x_1x_2 is calculated as $p(N_1, N_2)$ and all subsequent positions are based on the conditional probability of the previous position. Therefore, we generate the sequences as:

$$\begin{aligned} x_1x_2 &= p(N_1, N_2) \\ x_i &= p(N_i|N_{i-1}) \text{ for } i > 2 \end{aligned} \tag{2}$$

Importantly, we further capture the natural diversity in GC bias (both magnitude and width) by simulating from distinct promoter and enhancer dinucleotide probabilities (Additional file 1: Fig. S4). The proportion of bidirectional calls at promoters (versus enhancers) varies across data sets (Additional file 1: Fig. S5), which may be biological or could reflect ascertainment biases since promoters tend to be more highly transcribed. Since promoters are considerably more GC rich than enhancers (Additional file 1: Fig.

S2), TFs with GC rich motifs will be disproportionately enriched (false positives) in data sets with high promoter content. To control for this, we simulated sequences from the two classes (enhancers and promoters) in proportion to the observed ratio for a total of 10^6 instances (see Methods section). Using these simulated sequences, we calculated expected MD-scores (Eq. 1). This enables us to compare the expected (i.e., model derived, x-axis) to observed (i.e., experimentally observed, y-axis) MD-score for a single data set [38] as shown in Fig. 2A. Thus, the expectation model is calculated on a per data set basis to accurately reflect the composition of initiation regions inherent to that cell type and condition.

Building TF activity profiles across tissues

The next step was to assess the statistical significance of TF activity for each TF-PSSM occurrence; that is, to statistically ask which motif hits in Fig. 2A were significantly more (or less) co-localized with RNAPII initiation sites than our background model suggests (versus expectation). Logically, TF-motifs with greater (or less) than expected co-localization are the TFs we infer as ON (ON-UP and ON-DOWN, respectively) and participating actively in RNAPII regulation.

To this end, the MD-scores for 388 TF motifs (HOCOMOCO core version 11 [39]) were calculated for all control data sets of sufficient quality ($n = 126$; see Methods section). In each data set, the observed MD-score was compared to the expected MD-score. To assess statistical significance, we further assumed that the majority of TFs will be OFF across all control data sets (75% not significantly different from expectation; Additional file 1: Fig. S6A). The distribution of residuals (Additional file 1: Fig. S6B) was then used to assess significance for all TF motifs within all control data sets. This resulted in a range of 80–164 TFs that were called ON in any given data set (mean = 123.5, p value

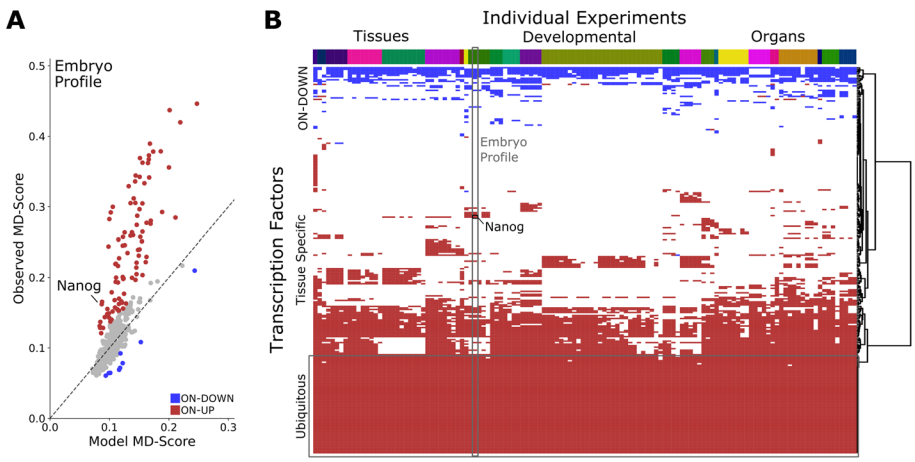

Fig. 2 Generating and clustering of TF profiles. **A** Scatter plot showing the expected (x-axis) and observed (y-axis) MD-scores for all PSSMs in HOCOMOCO for embryonic stem cells [38]. Significant differences are colored with ON-UP red and ON-DOWN blue (see Methods section). The collection of ON and OFF (gray) labels is the TF profile for this experiment. **B** Ward clustering of TF profiles (columns) in 126 samples representing over twenty tissue types. Each sample is labeled by its tissue of origin (top, colored bar) with tissue labels further classified into tissues (e.g., bone, blood), developmental (e.g., fetal, embryo), and organ (e.g., breast, kidney). The ON-DOWN TFs (at the top, blue) tend to be shared across samples. The ON-UP (red) shows a variety of patterns including tissue specific pockets (middle) and ubiquitously on and active (bottom)

< 0.05). The ON TFs can be split into two categories, on and enriched (ON-UP, activators; range of 74–148, mean = 109.9) or on and depleted (ON-DOWN, repressors; range 5–28, mean = 13.6) (Additional file 1: Fig. S6C). We refer to the collection of ON TFs (either UP or DOWN) for a given cell line as its TF activity profile.

An example TF activity profile is shown in Fig. 2A, where red and blue represent TFs that are classified as ON-UP (red) and ON-DOWN (blue) and gray represents TFs that are OFF. When applied to an embryonic stem cell data set (Fig. 2A; $n = 3$ biological replicates) [38], we called 95 enriched and 9 depleted TFs (Fig. 2A). Enriched TFs included the pluripotent factors responsible for embryonic stem cell self-renewal, Oct4 (pval $= 1.3e^{-5}$), Nanog (pval $= 2.7e^{-5}$), and SRY-Box Transcription Factors 3 and 4 (pval = 0.008, pval = 0.03, respectively). Importantly, across 3 additional independent embryonic stem cell data sets [40–42] the same pluripotency factors were consistently called as active. TFs identified as depleted include a variety of known repressors including SNAI2, CEBPA, and E2F1 [43–45].

We next sought to expand our examination of TF activity profiles by clustering the profiles across tissues and cell types that had high quality nascent run-on RNA sequencing data (see Methods section). In total, we examined 126 distinct data sets representing a total of 299 nascent run-on RNA-sequencing samples in basal conditions (i.e., normal growth or control samples). We used Ward's method to cluster the TF activity profiles (using Euclidean distance) across the DBNascent high-quality control samples. We found that the major determinant in clustering was tissue identity (Fig. 2B; Additional file 1: Fig. S7).

Moreover, the clustering of TF activity profiles suggested that at the extremes, some TFs are active across nearly all cell types and other TFs are tissue specific. Notably, only 4–6 TFs per tissue type were truly tissue specific; however, they were the major determinant for clustering. For example, the TF MyoD is a strong determinate in muscle differentiation [46] and was uniquely ON-UP (pval $= 7.0e^{-9}$) in the myoblast data set [47] and OFF in the other 125 data sets tested. The blood associated factor, GATA-2, was uniquely called as ON-UP across several blood samples [28, 48–52] and was notably OFF within the other data sets. In addition to these well known cell type specific TFs, we also recovered less well annotated TFs that infer uncharacterized biological functions. For example, ZNF121 in blood, or ZNF146 in organ function.

We also noticed that some TFs implicated in general cellular processes were commonly called ON-UP in the TF activity profiles. These "ubiquitously active" TFs included members of the ETS family, the E2F family, and KLF/SP family. These TFs have highly redundant binding motifs and their biological functions relate to cellular homeostasis and proliferation [53–55]. Additionally, these ubiquitous TFs may help maintain promoter accessibility and/or enable promoter-promoter looping [56, 57].

TF region selection across tissues

We next sought to further characterize the two extreme classes of TF regulatory activity: the ubiquitous and tissue-specific classes of TFs. First, we noted that the GC content of the TF recognition motifs differed between ubiquitous and tissue-specific TFs (Fig. 3A). The ubiquitous TFs bound GC-rich regions that were close to the average GC composition at promoters, consistent with prior reports [56]. By contrast, the tissue-specific TFs

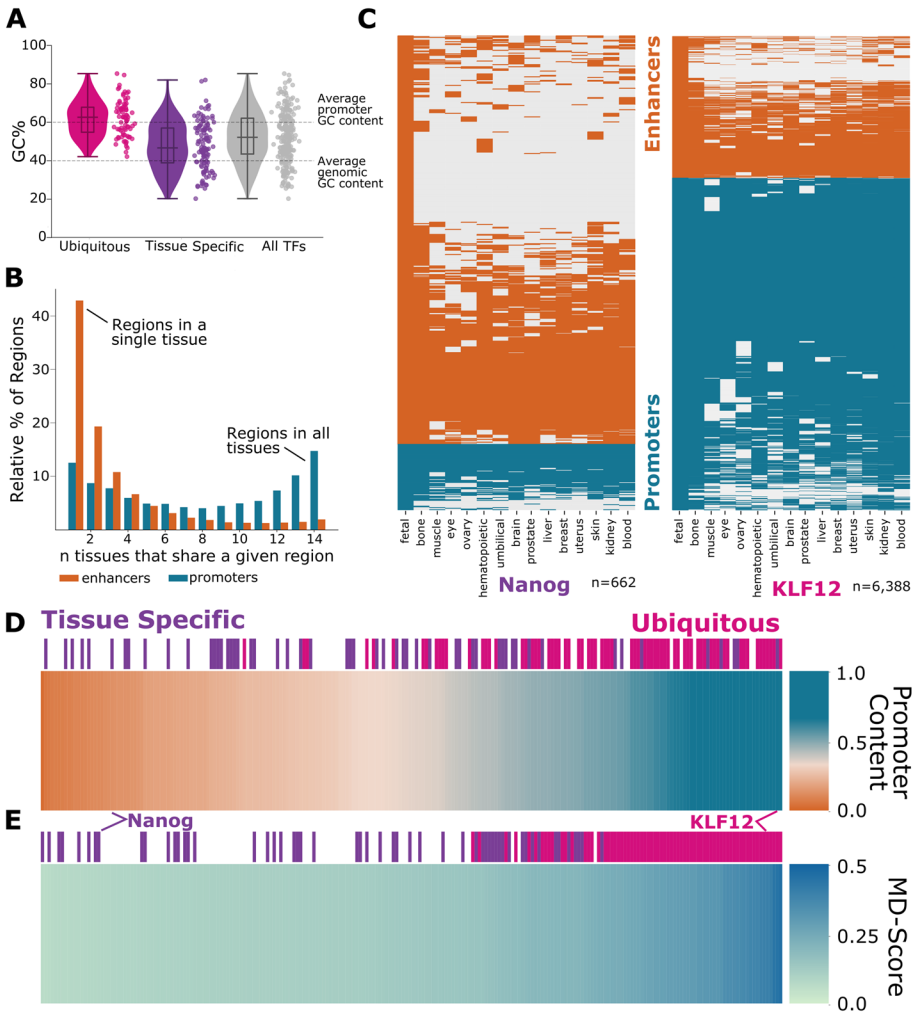

Fig. 3 Tissue specific and ubiquitous factors have distinct localization and regulation preferences. **A** Violin plots showing the GC content of the TF motif (PSSM) of ubiquitously shared TFs (pink), tissue specific TFs (purple), and all TFs (gray). **B** Bar plot showing that enhancers (orange) are far more tissue specific than promoters (blue) which tend to be on in all tissues. **C** Set of consensus bidirectional regions in embryonic stem cells containing a centered motif for the tissue specific Nanog (left) or the ubiqitous KLF12 (right), colored (promoter: blue, enhancer: orange) by presence or absence across multiple tissues (x-axis). Across 662 regions containing Nanog in ESC cells, 86.1% of regions are enhancers, with some being shared (bottom, solid orange) and others being more tissue specific (middle, mostly white). Across the 6388 KLF12 containing regions in ESCs, 70.3% are promoter regions. **D** Fraction of ChIP-seq binding sites at the 5 end of genes (promoter, blue) or at distal regulatory regions (enhancers, orange) for ubiquitous TFs (pink) and tissue specific TFs (purple) [61, 62]. **E** Motif displacement score as a heatmap (darker is higher) for ubiquitous TFs (pink) and tissue specific TFs (purple)

tended to have motif preferences closer to genomic background (Fig. 3A). Notably, this result was recapitulated in SELEX and protein binding microarray data independent of genomic context (Additional file 1: Fig. S8).

Given the sequence preferences inherent to the ubiquitous and tissue-specific TF classes, we next wondered whether these TFs would act at distinct genomic regions (e.g., promoters vs. enhancers). As previously noted [58–60], enhancer regions are more tissue specific whereas promoters are often transcribed more broadly across cell types (Fig. 3B). Thus, we examined the number of enhancer and promoter regions contributing

to each TF's activity profile. We observed that tissue specific TFs predominantly regulate at enhancers whereas ubiquitous TFs generally regulate at promoters (Fig. 3C). For example, Nanog is an embryonic specific TF. The vast majority of transcribed regions containing the Nanog motif in embryonic stem cells (86.1%) are enhancers, the majority of which are unique to the embryonic tissue samples. In contrast, the regions with the KLF12 motif, a ubiquitous TF, tend to be promoter associated (70.3%) and transcribed across most tissues.

We next wondered whether the region bias identified by TF Profiler was also present in transcription factor chromatin immunoprecipitation (ChIP). Thus, we next examined TF ChIP-seq data curated from cistromeDB [61, 62]. To select for high quality ChIP signal, we only considered TF ChIP-seq peaks within regulatory regions (n = 53,244 promoters, n = 559,150 enhancers). We then asked how often ChIP peaks fell within promoter regions versus enhancer regions. The ChIP data further supported the observation that ubiquitous TFs bind and regulate predominantly at promoters, whereas tissue-specific TFs bind and regulated predominantly at enhancers (Fig. 3D). This result is reliably captured by the MD-score approach, where ubiquitous TFs have higher MD-scores on average than tissue specific factors (Fig. 3E).

Regulation of gene encoding TF

Given the distinct binding sites and biological functions of the ubiquitous and tissue-specific TFs classes, we next asked whether the regulation of these TF classes was distinct. To this end, we first examined the transcription level of the gene encoding each TF. For example, we assessed the transcription level of GRHL2 (tissue specific) and CREB1 (ubiquitous) across the control data sets. The tissue specific TF had many samples with low gene transcription and a few samples with high gene transcription. Thus, the distribution of the TF gene transcription level followed an exponential distribution, consistent with an transcription of the gene in a limited subset of the data. In contrast, the transcription of the ubiquitous TF was normally distributed (Fig. 4A), consistent with the TF gene being transcribed in all samples. Consequently, we classified each gene encoding a TF as either fitting an exponential or normal distribution. Notably, the tissue specific TFs tended to fit an exponential distribution, like GRHL2, and ubiquitous TFs had a bias towards a normal distribution of transcription like CREB1 (Additional file 1: Fig. S9). In sum, the two classes show distinct corresponding cumulative density functions for transcription of the gene encoding the TF across a subset of high confidence TFs (Fig. 4B; see Methods section).

We next examined these distributions the context of TF regulatory activity (the MD-score). As a representative example, Fig. 4C shows a plot of the MD-score vs. the transcription level (RPKM) for the ubiquitous TF KLF12. There was no correlation between the TF gene transcription level and predicted TF activity. This pattern was observed across many ubiquitous TFs such as SP1 and ETV1 (Additional file 1: Fig. S10A) and suggests that there is no obvious relationship between a ubiquitous TF's transcription level and its activity (i.e., MD-score). In contrast, the tissue specific TF Nanog shows a positive correlation between its activity (MD-score) and gene transcription level (RPKM; Fig. 4D). Moreover, this positive correlation was observed for many tissue specific factors, including MyoD and GATA-2 (Additional file 1: Fig. S10B). This result indicates

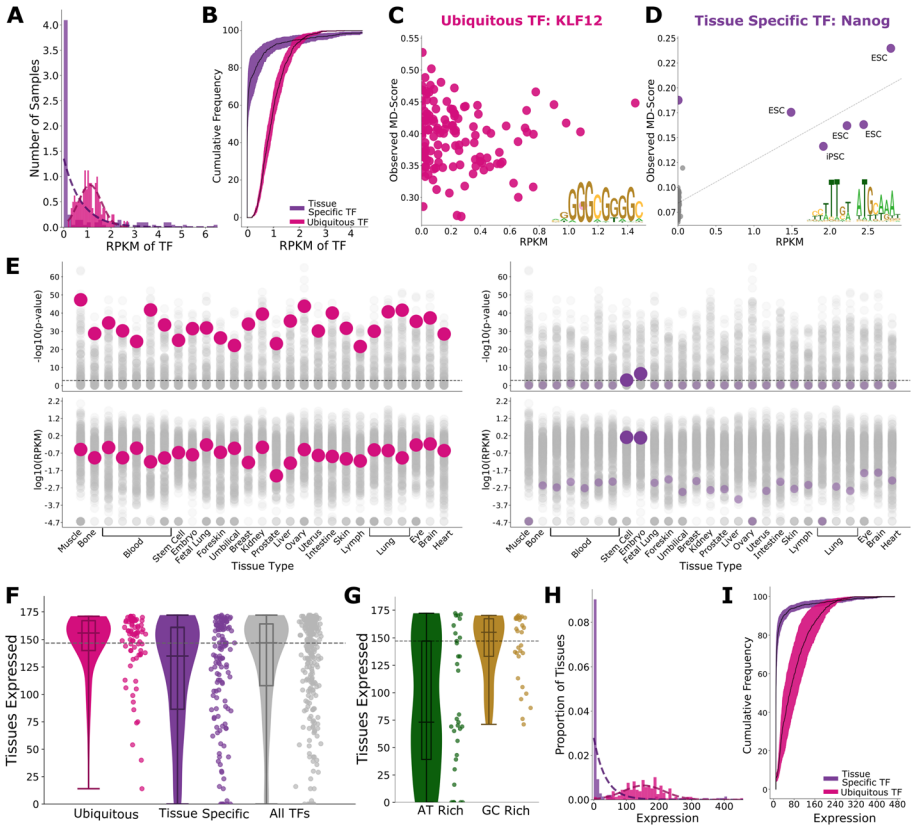

Fig. 4 Tissue specific TFs are regulated at transcription. Ubiquitous factors are post-transcripionally regulated. **A** Histogram of the transcription level (x-axis) of a tissue specific TF (GRHL2; purple) and ubiquitous TF (CREB1; pink) across the nascent RNA-seq datasets. **B** Cumulative distribution function of the transcription of the gene encoding the TF (RPKM) for a set of high confidence tissue specific factors (purple) and ubiquitous factors (pink) across 126 control experiments. The relationship between the transcription level of the gene encoding the TF (x-axis) and observed MD-score (y-axis) for (**C**) ubiquitous TF (KLF12) and (**D**) a tissue specific TF (Nanog). HOCOMOCO PSSMs shown in lower right corner. **E** Plot of the significance of the MD-score (top) and the transcription of the gene encoding the TF (bottom) for all TFs (gray), highlighting KLF12 (left, pink) and Nanog (right, purple). **F** Violin plots of frequency of expression in single cell RNA-seq [63] across 172 tissues for ubiquitous (pink), tissue specific (purple), and all TFs (gray). **G** Violin plots of frequency of expression in single cell RNA-seq [63] across 172 tissues for TFs with the bottom 10% (green) and top 10% (gold) GC content within their PSSMs. **H** Histogram of the number of tissues that a tissue specific TF (GRHL2; purple) and ubiquitous TF (CREB1; pink) are expressed in by single cell RNA-seq [63]. **I** Cumulative distribution function of the steady-state RNA level (scRNA-seq) for the same high confidence tissue specific factors (purple) and ubiquitous factors (pink) across 172 tissues from atlas of fetal gene expression [63]

that tissue specific TFs are not transcribed unless they are actively regulating within a cellular context, suggesting that repression of tissue specific TFs transcription plays a role in blocking their function. In summary, the two classes of TFs, ubiquitous and tissue specific, have categorically distinct transcription patterns and suggests biologically distinct mechanisms of TF gene activation (Fig. 4E).

To further probe these results, we sought to determine whether these trends would be recapitulated in steady-state RNA levels. We utilized single cell RNA-seq (scRNA-seq) data from the atlas of fetal gene expression [63] as it allowed us to capture expression values for these TFs across 172 distinct human tissues. We observed that the ubiquitous TFs were generally expressed in all tissue types whereas the tissue specific TFs were expressed in fewer tissues (Fig. 4F). When we assessed the expression of TFs with the

lowest GC content motifs (bottom 10%), we found that the median number of tissues with TF expression falls far below the total median. 72.7% of TFs with the low GC motif set are classified as tissue specific. The TFs with the highest GC content motifs (top 10%) are expressed in more tissues than expectation and many are ubiquitously active TFs (65.5% ubiquitous; Fig. 4G). Similar to the observed trends in nascent transcription, tissue specific TFs are expressed in fewer tissues lending to an exponential fit of their expression profiles, whereas the ubiquitous TFs have gene expression that tends to be normally distributed (Fig. 4H, I). Overall, the trends we observed at the transcriptional level (PRO-seq) are recapitulated at the steady-state RNA level (scRNA-seq) suggesting this is a fundamental regulatory strategy for ubiquitous TFs versus tissue specific TFs.

Stimulus responsive TFs

The ubiquitous and tissue specific TFs represent the extremes of ON and OFF patterns within our clustering (Fig. 2B). Yet many transcription factors were ON in groups of samples, either several tissues or more sporadically across samples. We reasoned that stimulus responsive TFs could give rise to a more sporadic pattern of activity, as the activity of the TF would depend on the fine details of the growth environment. Thus, we next sought to identify high confidence stimulus responsive TFs. To accomplish this, we identified 161 data sets in treatment conditions from corresponding publications with our control data sets [60]. We applied TF Profiler to this "perturbation" collection (Additional file 1: Fig. S11A–C), identifying 53 high confidence stimulus responsive TFs.

We next sought to characterize the 53 high confidence stimulus responsive TFs. To this end, we first probed whether the stimulus responsive TFs have a have recognition motif preferences comparable to either the ubiquitous or tissue specific TFs. We determined that stimulus responsive TFs have recognition motifs that are similar to genomic background, as seen with tissue specific TFs (Additional file 1: Fig. 12A). We next examined ChIP-seq data for the stimulus responsive TFs, finding that they bind and act primarily in enhancer regions, similar to tissue specific factors (Additional file 1: Fig. 12B, C). Among each of our classified TF groups, we found a positive correlation between the GC content of the recognition motif and the preference for binding within promoter regions, where ubiquitous TFs dominate the high GC percentage regime and the other two classes (tissue specific and stimulus responsive) behave similarly in the low GC percentage regime (Fig. 5A).

We next examined the regulation of the gene encoding the stimulus responsive TFs. Intriguingly, we found that many stimulus responsive TFs are broadly transcribed, similar to ubiquitous factors, but active in only a subset of samples, similar to tissue specific TFs (Fig. 5B, Additional file 1: Fig. 13A–C). This is consistent with the fact that many stimulus responsive factors are post-transcriptionally regulated. For example, under normal conditions, p53 is constantly transcribed and translated, but subsequently degraded via the ubiquitin ligase HDM2 [64, 65]. Consistent with post-translational regulation, we observed elevated activity scores only in samples where p53 was directly stimulated (Additional file 1: Fig. 13B).

To fully understand the distinct behavior of stimulus responsive TFs, we selected NFκB as a case study, as it has the most high-quality data (six data sets, four tissue types [26, 66–70]). NFκB is a key regulator of the inflammation response across tissue types [71].

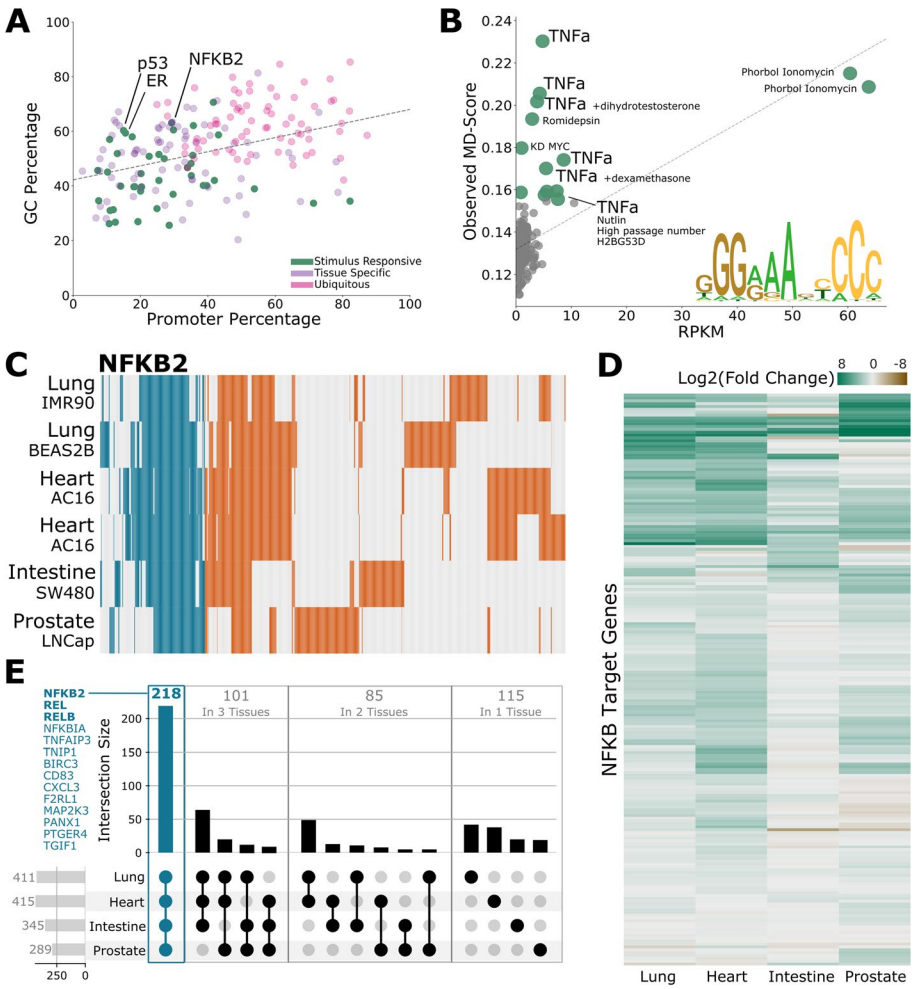

Fig. 5 Stimulus responsive TFs utilize distinct regions to achieve equitable stimulus response across tissues. **A** Scatter plot of the percentage of ChIP-seq sites within promoters (x-axis) vs. the GC content of the TF motif (PSSM, y-axis) for ubiquitous (pink), tissue specific (purple), and stimulus reponsive (green) TFs. **B** Comparing the transcription level (RPKM) of the gene encoding the TF NFκB (x-axis) and MD-score (y-axis). Significant ON-UP instances of NFκB are colored green and labeled with the stimulus. Inset is the PSSM for NFκB2. **C** Bidirectional regions with a centered NFκB2 motif from TNFα treated cells (larger font in **B**, four tissues). Enhancer regions in orange and promoters in blue. Across 2302 regions with NFκB2 motif instances, 77.5% are classified as enhancers. Of the enhancers, 1406 (78.9%) are unique to a given tissue type. Other subunits of NFκB shown in Additional file 1: Fig. S14. **D** Heatmap of NFκB target genes (y-axis) across the four tissues in (**C**) (heart, intestine, lung, prostate). Upregulated (green); gold (downregulated). **E** Upset plot of promoter regions ($n = 519$) shown in (**C**) where 42.0% are shared across all tissues (teal) and 81.0% are shared between at least two tissues. Numerous NFκB target genes (teal text) are labeled. Three target genes (NFκB2, REL, and RELB, bold) are three subunits of the NFκB TF complex

Across four of the NFκB subunits (REL, RELB, NFκB2, and T65), we noted that the TF region selection differed across tissue types, specifically within enhancer regions which represent the majority of putative binding sites (Fig. 5C, Additional file 1: Fig. S14A–C). In fact, within a tissue with multiple cell lines (lung; IMR90 and BEAS2B) the region selection varied, but within a tissue with the same cell line replicated from different publications (heart; AC16) the enhancer region selection was consistent. This suggests that the NFκB response regions are defined by the cell type. Despite this, there was a robust NFκB response in all tissues (Fig. 5D, Additional file 1: Fig. S15). While enhancer region

selection was highly variable, active promoter regions with NFκB2 motifs are more consistent across tissues. Out of 519 promoter regions across the TNFα treated samples, 218 (42.0%) are shared across all tissues (Fig. 5E). We next examined the genes associated with these promoter regions. Multiple genes were direct NFκB targets, including subunits of the NFκB TF (NFκB2, REL, and RELB).

Discussion

Here we present TF Profiler, a method of TF activity inference that identifies which TFs, among hundreds with well-characterized motifs, are actively regulating RNAPII from a single nascent RNA-sequencing experiment. The method relies on a robust sequence based expectation model derived from the base probabilities at RNAPII initiation regions. Using this method, we can identify which TFs are ON and active, regardless of whether the TF is an activator or repressor. We anticipate that this method will be broadly useful for assessing the set of TFs active in any cell type, provided that high-quality nascent sequencing data is available. Interestingly, the TF Profiler method identified three classes of TFs: ubiquitous TFs which are always on regardless of cell type or condition, tissue specific TFs that drive cell identity and stimulus responsive TFs that are poised to alter transcription in response to a perturbation. We also showed that these TF classes have distinct DNA binding preferences and are regulated via distinct mechanisms. Because TFs drive all biological processes and are among the most important class of proteins in biology, it is critical to develop tools to reliably assess TF activity.

The ubiquitous TFs have GC-rich recognition motifs and bind preferentially at promoters. The ubiquitous TFs are represented in part by the ETS, KLF, E2F, ATF, and SP1 families. We note that among the ubiquitous TFs ($n = 78$), many motif preferences are similar and therefore difficult to distinguish from each other. While there may be subtle differences in which TFs are active in a given cell line, a subset of these ubiquitous TFs are always active regardless of cell line or condition. In agreement, most of the genes encoding ubiquitous TFs were transcribed in nearly all data sets tested, suggesting they function cooperatively or redundantly. Moreover, individual ubiquitous TFs are typically not essential, suggesting they behave collectively to regulate RNAPII function, perhaps to help maintain nucleosome-free promoters [56], though genomic regions with high GC content naturally exclude nucleosomes [34]. Finally, ubiquitous TFs regulate genes important for cellular proliferation, metabolism, and homeostasis [72–74], consistent with their general requirement across cell types.

Distinct from the ubiquitous TFs, the tissue specific TFs preferentially bind enhancers, which are lower transcribed, with binding motifs that have a nucleotide composition similar to genomic background, i.e., more AT-rich. It is difficult to disentangle which of these features—enhancer/promoter preference, sequence content, or transcription levels—is critical to the tissue specific nature of these TFs. Importantly, there are a subset of TFs that bind AT-rich regions but were not called ON in any of our data sets. Many of these never ON TFs are implicated in cell identity for cell types with no nascent RNA sequencing data. For example, UNCX is a TF implicated in regulation of the cerebellum with an AT-rich binding preference. Yet no cerebellum data is present within DBNascent [60], which could explain why we do not see UNCX as ON in any of these data sets.

Many of the tissue specific group of TFs are not transcribed unless they are ON within a given cellular context, e.g., their activity may be regulated by their transcription.

Many TFs are neither ubiquitous or tissue specific. This includes TFs that are on in subsets of, often related, tissues. It also includes the stimulus responsive TFs, which share many of the same recognition properties as the tissue specific TFs. Namely, they bind predominantly at enhancer regions and have recognition motifs similar to genomic background composition. Yet unlike tissue specific TFs, the gene encoding stimulus responsive TFs are typically transcribed across a broad range of tissue types and conditions (similar to the ubiquitous TFs). This pattern is consistent with post-transcriptional regulation of these TFs, allowing them to be poised for activation but are not always ON; instead, post-transcriptional mechanisms regulate their activity.

We speculate that the binding preferences and mechanism of regulation for a given TF may be predicted based on the TFs function. While it is known that tissue specific TFs play a crucial role in defining cell identity, we postulate that these TFs are not transcribed unless actively regulating transcription as they are key players in establishing tissue specific enhancer regions. Furthermore, it is tempting to speculate that these tissue specific enhancer patterns would then directly explain the subset of tissue specific stimulus responsive regulatory sites. However, the set of tissue specific TFs are not enriched directly in or adjacent to the cell type specific stimulus responsive sites. This contradiction suggests that tissue specific stimulus responses may arise from some complex interaction between tissue specific TFs at some sites and other factors such as chromatin state or transcriptionally active domains.

Importantly, our approach detects TF effector domain activity because of the co-localization of binding motif instances with sites of RNA polymerase II initiation. However, some transcription factors may alternatively function as chromatin modifiers. If a TF functions primarily to modify chromatin without direct effects on transcription, our approach may not identify this activity. In fact, our prior work identified a small number of TFs whose motif co-localization was consistently offset from sites of RNA polymerase II initiation, with many of these TFs annotated as chromatin modifiers [21]. Detecting the regulatory activity of chromatin modifying TFs will likely require accessibility and conceptually similar methods developed for that data, such as ChromVAR [75]. It is intriguing to speculate whether the combination of these approaches, on matched nascent and accessibility data, would uncover novel classes of transcription factor function.

Conclusion

In summary, TF Profiler is a broadly applicable method of inferring transcription factor regulatory activity directly from nascent run-on sequencing assays. TF Profiler provides a method of assessing the activity of a TF's effector domain directly from the co-occurance of TF recognition motif instances and sites of RNAPII initiation.

Methods

Code availability

The stand-alone TF Profiler application can be found on github (https://github.com/Dowell-Lab/TF_profiler) and Zenodo [76]. TF Profiler takes an annotation file for bidirectional regions from a nascent sequencing experiment and derives a TF activity

profile. This includes generating simulated sequences based on the base composition of the regions provided, scanning for PSSM hits within the genome and statistically assessing TF enrichment and depletion.

Additional stand-alone scripts and useful data files associated with this work can be found on github (https://github.com/Dowell-Lab/TF_profiler_additonal_scripts) and Zenodo [77].

Curating data from DBNascent

All nascent RNA sequencing data, which includes both precision run-on sequencing (PRO-seq) and global run-on sequencing (GRO-seq), were obtained from DBNascent [60]. Briefly, the database contains 502 human PRO-seq samples from 60 publications and 780 human GRO-seq samples from 106 publications. All data were aligned to the human reference genome (hg38; https://github.com/Dowell-Lab/Nascent-Flow; archived at Zenodo [78]) and were subjected to extensive quality control. Data was additionally processed for identifying sites of bidirectional transcription (also known as transcribed regulatory elements) via a NextFlow pipeline built upon Tfit [17] (https://github.com/Dowell-Lab/Bidirectional-Flow) and archived at Zenodo [79]. Standardized Nextflow pipelines are described in detail in Sigauke et al. [60].

To prepare the data for TF Profiler, high quality nascent RNA samples were selected from the database, with minimum quality score of 4 (minimum of 5 million reads, over 50% of reads map to the reference genome and less than 95% duplication). Within these samples, Tfit [17] was utilized to annotate bidirectional regions. All Tfit calls between biological replicates of a given cell type within a given paper were merged using muMerge [27]. If only one biological replicate passed the quality score cut-off, it was still used for subsequent analysis as a single replicate data set. The grouped biological replicates within a cell type and paper are referred to as "data sets" in this study. All data sets are described with cell type, perturbations, tissue identity, and the utilized SRR identifier numbers in Additional file 2. Bidirectional annotations were merged on a per-cell and per-paper basis to maximize number of reliable calls per data set.

To account for low complexity in some data sets, which can arise from either poor pull down efficiency or high sequencing noise, we also filtered data sets based on the quality of the bidirectional calls. To this end, we required that the region within $2h$ of μ had a base composition of at least 50% GC content. The final requirement is that at least 50% of called regions must not be at an annotated promoters. If promoter regions are overrepresented, then we lose sensitivity when calling many TFs, as most TFs bind predominantly at distal regulatory elements, such as enhancers.

Curating a master bidirectional region list

After this two-step quality control process, we ended up with 126 distinct data sets from 88 publications that represent 79 unique cell lines under basal conditions (e.g., basal, normal growth conditions; $n = 299$ unique biological samples). Samples were merged step-wise, with all samples of a given tissue type were merged into a tissue specific regions file. Any region less than 20 nt were windowed to be at least 20 nt in length. The tissue specific region files were then merged into the master file. From the same publications as the control samples an additional 161 data sets with identifiable perturbation

or genome modified conditions also passed the quality control process. These samples come from 65 of the 88 control publications and represent 46 unique cell lines ($n = 411$ unique biological samples).

Regions within the master file (control samples only) were divided into two sets: promoters and enhancers. Promoters were defined as all regions (windowed by $h = 150$) within 1000 bp (300 bp upstream, 700 bp downstream) of RefSeq (hg38 release GCF_(109.20190607_2019_06) annotated transcription start site (TSS). All other regions were labeled as enhancer. This resulted in a total of 53,244 promoter regions and 611,963 enhancer regions within the master file.

Calculating positional probabilities

Both regions types (promoters and enhancers) used to independently extract two sets of positional probabilities surrounding μ using a window size of $H = 1500$ (e.g., $\mu \pm 1500$). Sequences were extracted using bedtools getfasta (bedtools/2.25.0). All ambiguous bases were replaced with randomly sampled nucleotides (A, C, G, T) using a flat distribution (all bases equal probable). Two distinct probability distributions are then tallied from the sequences. First, the dinucleotide ($n = 16$; AA, AT, CA, CG, etc.) frequencies at the start of each sequence (e.g., at -1500 from μ). These probabilities are used to initiate the sequence generator. The second distribution obtained from the sequence data is the per position conditional probabilities (e.g., $P(n_i|n(i-1))$) (see Additional file 1: Fig. S4). The dinucleotide and conditional probabilities are described by Eq. 2. Position specific mononucleotide frequencies, which simply reflect the probabilities of a given nucleotide at a given position across the window, were calculated for Fig. 1C. Position independent (also referred to as "flat") probabilities (shown in Fig. 1D, left) were generated by taking the mononucleotide probabilities and averaging them across the window ($H = 1500*2$).

Note that base composition plots (Additional file 1: Fig. S4A and Fig. 1C) are smoothed for clear visualization (scipy savgol filter version 1.5.4). Code used to calculate the position specific probabilities can be found within the sequence_generator module of TF Profiler.

Generating simulated sequences around RNAPII initiation

Using the dinucleotide training data described in Calculating positional probabilities section, we employ a Markov chain to generate 10^6 sequences each from the promoter and enhancer probability sets. This was achieved by using numpy (version 1.19.5) random number generator based on (1) the initial dinucleotide probability and (2) the subsequent conditional probabilities that account for position X-1 to select the nucleotide in position X. The sequences were checked to ensure there was no identical sequences within the $2*10^6$ sequences generated. The validity of sequence generation was confirmed by ensuring that generated sequence recapitulate the probability distributions used in their generation (within $\pm$ 0.0001). Sequences were generated in batches with distinct numpy seeds (seeds used: 38–50, 107–119, 275–287, 395–407, 462–474, 523–535, 687–699, 721–733, 831–843, 986–998) and the probabilities used are available on the additional data github page. The generation of mononucleotide and flat simulated sequences were generated in a similar manner (including the same seeds), only using the mononucleotide and position independent probabilities, respectively. Code used

to generate all sequences can be found within the sequence_generator module of TF Profiler.

Counting over genes and bidirectionals

RefSeq gene counts for human sample within DBNascent were counted over hg38 RefSeq genes (hg38 release GCF_000001405.40-RS_2023_03) using Rsubread, featurecounts (version 2.12.3) [60, 80]. For all samples within a biological replicate for a given data set (both control, $n = 299$ and perturbation, $n = 411$ biological replicates), the mean RPKM was calculated for every gene isoform. Only the highest mean RPKM isoform for every gene was retained. Gene counts were used for additional analyses, including the transcription level of the gene encoding the TFs across tissue types, the transcription level of TF genes vs. TF activity and DESeq2 analyses between control and perturbation conditions.

Bidirectional counts were also measured to assess (1) whether GC content of bidirectionals relates to the transcription level and (2) how this relates to enhancer and promoter content. This data is shown in Additional file 1: Fig. S3. To count over all bidirectionals, the master bidirectional file was utilized (generation of this file described in Calculating positional probabilities section). This file contains all bidirectionals called within the 126 control data sets. To ensure that the regions were wide enough, the regions within the master bed file were windowed ± 150 bp surrounding μ. This could cause some regions to overlap, therefore the bedfile was sorted (sort -k1,1 -k2,2n) and merged (bedtools merge, version 2.28.0). Feature counts was used to count over the windowed master file using Rsubread, featurecounts (version 2.0.1). Like with genes, all individual control biological replicates ($n = 299$ independent samples from $n = 126$ control data sets) were used to count over the windowed-master bed file.

Motif scanning

Motif scanning was performed using the MEME suite (version 5.0.3) function FIMO scan [81]. This scan was performed using a flat background model (equal distribution assumed of the four canonical nucleotides). The threshold was set to 1e−5. The motif files used were from HOCOMOCO version 11 [39]. The scan was performed across the human genome (hg38) and these motif hits were used for subsequent analysis. Internal to the TF Profiler program, the motif scan can also be performed de novo across only the bidirectional regions provided, or take in pre-scanned regions genome wide. Motif scanning was performed on simulated sequences using the same parameters. Code used to perform motif scanning can be found within the fimo_scanner module of TF Profiler.

Calculating distances between RNAPII initiation and motif hits

To measure TF co-localization, the relative distance between a motif hit and the center of the bidirectional transcript must be assessed. The distance for all motif hits within the large window ($H = 1500$) of a given region was calculated, using the center of the motif and the center of the bidirectional. For motifs and regions of odd length the center is rounded to the nearest even integer per the native python rounding function. Each motif hit is associated with the distance to the center of the bidirectional as well as two ranking metrics. The two ranking metrics are a distance rank metric (e.g., which motif is closest

to the center of the bidirectional, where 1 is closest) and a quality rank (e.g., defined by FIMO score where 1 is the highest quality hit in the region). All motif hits within the large window are stored within the distance tables.

For this study only a single motif hit for a given PSSM per bidirectional was retained for further analysis. Hence, for each bidirectional region and *distinct* PSSM, only a single hit per *unique* PSSM is considered for further analysis. In the case of multiple motif hits for a single PSSM within one bidirectional, the motif hit used for further analysis was the motif hit closest to the center of the bidirectional (i.e., distance rank = 1). Code used to generate these distance tables can be found within the distance_module of TF Profiler.

PSSM GC content analyses

To calculate the GC content of the PSSMs, we extracted all probabilities for both G and C across the length of the PSSM and summed them together. This was then divided this by the length of the PSSM to give the overall probability of a GC within the PSSM itself. This was done for all HOCOMOCO core TFs. To validate that the GC percentage is associated with a given TF rather than genomic context (nucleosome arrangement, for example), we looked at both SELEX and protein binding microarray data (CIS-BP version 2.00). The GC content was calculated for PBM and SELEX in the same manner as HOCOMOCO PSSMs (Additional file 1: Fig. S8A, B).

Calculating MD-scores

The calculation of MD-scores was originally defined in Azofeifa et al. [21] and is described mathematically by Eq. 1. Briefly, the MD-score quantifies co-localization of motif instances (hits) near sites of RNAPII initiation ($h = 150$ bps) relative to a larger local window ($H = 1500$ bp) genome wide. Precision in position of RNAPII initiation is required for robust MD-Score calculation [27, 82].The MD-score was calculated for all motifs within HOCOMOCO core version 11 ($n = 388$ motifs) [39], in every data set in this study ($n = 287$ data sets).

To calculate expected MD-score from simulated data, we leverage each data set's distinct proportion of enhancer to promoter bidirectionals (Additional file 1: Fig. S5)—thus accounting for each data sets' distinct composition profile. To this end, we calculate the proportion of promoter associated bidirectionals (see Curating a master bidirectional region list section for labeling promoter bidirectionals). The proportion of promoter associated bidirectionals ranges from 0.14 to 0.49 across the 287 data sets. Based on this for every 0.02 step from 0.14 to 0.5 we calculated simulated MD-scores using 10^6 simulated sequences (and associated motif hits) total. To do so we used numpy random number generator to randomly select a given proportion of promoters from the 10^6 promoter sequences and the remainder from the 10^6 enhancer sequences. For example, if a given data set had a proportion of 0.26 promoters, then 260,000 promoter sequences and 740,000 enhancer sequences were be selected from the dinucleotide simulated sequence data. From these 10^6 sequences the expectation MD-scores were calculated and used for the basis of comparison for subsequent analyses. For each data set the MD-score proportion was rounded to the nearest 0.02 (a proportion of 0.255 rounds to 0.26; a proportion of 0.245 rounds to 0.24), and the expected MD-scores are selected from that set as the background model. Five seeds (96, 118, 559, 603, 961) were used for the numpy

random number generator to subset the sequences. All resulted in similar expectation MD-scores for a given TF within a promoter proportion set. The seed used for selecting promoter and enhancer sequences for all subsequent analysis was 118.

Statistically assessing TF activity profiles

We sought to statistically assess whether the MD-score for a given TF was higher (ON-UP) or lower (ON-DOWN) than expectation for each data set. For the meta-analysis, TFs across all data sets were combined for subsequent linear fits. Two separate fits were conducted, one on the control condition data (Additional file 1: Fig. S6A, $n = 48{,}888$ points, 126 data sets with 388 TFs) and one on perturbation conditions (Additional file 1: Fig. S11A, $n = 62{,}468$ points, 161 data sets with 388 TFs).

In each case when all data is fit the slope is greater than 1, indicating higher activity in the experimental data than the expectation model. This is an expected result as some TFs should be ON and active in a given cellular context. Therefore, we opted to use an inlier method, where we fit a set proportion of inlier TF MD-scores to a linear regression. The proportion of inliers was optimized for each set independently by testing every 5% inlier proportion from 5% (almost no TFs being fit) to 100% inliers (all of the data). The proportion closest to slope of 1.0 and intercept of 0.0 is assumed to identify the set of TFs unchanging within the set of data. The normal distribution of the residuals of the inliers was then used to attribute a p value for each TF across all data sets (Additional file 1: Fig. S6B, Additional file 1: Fig. S11B). While the data was fit all together to get a better estimate the distribution of the residuals, the TF Profiler program fits the residuals of the inliers for a single data set at a time by default for single case uses. Code used to generate these distance tables can be found within the statistics_module of TF Profiler.

Clustering TF activity profiles

To generate the highest confidence TF activity profiles for a given TF, we required a degree of replication across tissue types in control samples. For tissues with at least four data sets, a TF within the TF activity profile needed to be called as ON in at least 50% of the samples. For tissues with sparse data (less than four data sets), this replication was not required. There were a total of 26 tissue classifications for defining the high confidence TF activity profiles: blood (hematopoieticprogenitor), blood (K562), blood (lymphoid), blood (marrow), blood (myeloid), bone, brain, breast, embryo, eye, heart, intestine, kidney, liver, lung, lung (fetal), lung (muscle), muscle, ovary, prostate, skin, skin (foreskin), skin (lymph), stem cell, umbilical, uterus. Which data set belongs in which tissue set is defined in the sample_metadata_table found on in the additional data github. The tissues were defined in narrow categories as TFs vary between cell types as well as between tissues. The narrow classification permits for greater sensitivity when defining high confidence TF activity profiles.

Once tissue specific profiles were rigorously defined, they were clustered using the R (version 3.6.0) package ComplexHeatmaps (version 2.2.0) which utilizes the native R function hclust. For clustering purposes, the TF activity profiles were numerically represented as 1 (ON-UP), −1 (ON-DOWN), and 0 (OFF). The Euclidean distances were used followed by Wards method to cluster the profiles. A full cluster map was generated and shown in Additional file 1: Fig. S7. For the main text figure, the tissues were manually

divided into three categories, tissue, organ, and developmental. The tissues within those categories were clustered to assess which are most closely related. This ordering was used for Fig. 2B.

Classifying TFs

Here we define three categories of TFs: ubiquitous, tissue specific, and stimulus responsive. All classifications were defined using the high confidence TF activity profiles. High confidence TF activity profile generation is described in Clustering TF activity profiles section.

To classify a TF as ubiquitous, it was required to be ON-UP in at least 95% of the control data sets. To classify a TF as tissue specific, it needed to be uniquely ON in a given tissue. The exception for this is blood TFs (such as GATA and STAT TFs). These TFs had strong signatures in blood but also tended to be called ON in many organ samples. For this reason, blood specific TFs were excluded from being called ON in developmental or organ sets. Additionally, many organ TFs were shared due to similar function across tissues. If a TF in organ samples was only shared across two organs, it was still defined as tissue specific. One category not discussed in depth is shared, but not ubiquitously shared TFs. This general group is classified as TFs that are on in more than two organs, more than one blood cell type, or more than one developmental cell type. Finally, the stimulus responsive TFs defined by (1) TFs that were called ON in the perturbation sample but not called ON in the control sample and (2) not a ubiquitous or tissue specific TF within the tissue of the experiment tested. TF classifications are outlined in the TF_classes_table found on in the additional data github.

Comparing bidirectional regions across tissue types

To compare region usage in control conditions, each tissue was assigned a consensus region set. To generate consensus regions across tissues, the master bed file (described in Curating a master bidirectional region list section) was used. For each region within the master bed file, the data set that contributed to that region was noted. The tissues were broken into broad categories ($n = 15$; brain, blood, muscle, fetal, liver, ovary, hematopoietic, breast, skin, kidney, eye, umbilical, bone, prostate, and uterus) for region selection to increase the total number of regions accounted for in subsequent comparisons (see Additional file 2). For a region to be called within the consensus profile, the region needed to be attributed to at least 50% of the data sets within the broad tissue set. This ensures that the regions called are truly active bidirectional regions within the broad tissue category. Distances between the consensus regions and all HOCOMOCO motif hits were calculated using the distance calculations described in Calculating distances between RNAPII initiation and motif hits section. Motif hits within ±150 bp of initiation for a given bidirectional were considered a positive hit. Positive motif hits and consensus regions were used to systematically assess TF region selection across tissues, as shown in Fig. 3C.

The perturbation condition TNFα is one of the most highly represented perturbations, with 6 data sets in 6 independent publications applied across a myriad of tissue types. To study region selection across tissue type in TNFα conditions, we used *muMerge* across the 6 independent TNFα data sets to create a master TNFα bed file. Contributions to

each region per data set were retained. These regions were used for distance calculations with all HOCOMOCO motif hits. As previously, motif hits within ±150 bp of initiation for a given bidirectional were considered a positive hit. This data was used to assess NFκB subunit region selection as shown in Fig. 5C and Additional file 1: Fig. S14.

Using ChIP-seq data from CistromeDB

Both TF and histone ChIP-seq region data was obtained from CistromeDB [61, 62]. Within this database there are six total quality parameters assessed for every ChIP-seq experiment. These can be broken into two main categories, mapping and peak quality. The mapping scores account for sequence quality, number of unique sequences, and unique molecule representation after sub-sampling the data. The peak scores account for the number of peaks, the signal to noise ratio, and the overlap of peaks with accessible regions. In order for the ChIP-seq sample to be used here, we required the sample to pass at least one parameter within both mapping and peak scores. CistromeDB contained TF ChIP-seq data for 316 unique TFs within HOCOMOCO v11 core set ($n = 388$ total) that passed the defined QC standards.

Promoter associated ChIP-sites were defined as a ChIP site falling within 1000 bp (300 bp upstream, 700 bp downstream) of RefSeq (hg38 release GCF_(109.20190607_2019_06) annotated transcription start site (TSS), as with bidirectional calls. The percentage of promoter-associated regions was calculated by the total number of promoter associated ChIP-sites over the total number of ChIP-sites that fall within a bidirectional region from the master bed file. This reduces noise and regions where a TF is bound but not actively regulating. Thus, the ChIP promoter percentage reflects the percentage of functional TF binding events that occur within promoter regions versus all functional binding events.

In many cases, there were multiple ChIP-seq samples for a single TF. In this case, the median calculated promoter percentage was used. Heatmaps in Fig. S1 use cistromeDB regions from five independent samples per condition detailed in Additional file 3 (extended information on these samples resides in a file called ChIP_metadata_table on the additional data github page). Distance tables were generated using the TF Profiler program as previously described in Calculating distances between RNAPII initiation and motif hits section. The R (3.6.0) package ComplexHeatmaps (version 2.2.0) was used to plot the motif localization using the generated distance tables.

Fit classification for transcription level of TFs

We construct a simple classifier to assess the distribution of the transcription level for all genes encoding TFs across all data sets. To do this, we used Fitter (version 1.5.2) built on scipy (version 1.5.4). This program takes an array RPKM normalized counts across control data sets (as described in Counting over genes and bidirectionals section) and assesses how well that array fits a given data distribution. We used Fitter to classify the transcription level distributions as either exponential or normal, with the best fit defined as the minimum sum of the error metric squared (parameter fitterf.get_best(method = 'sumsquare_error')). If the KS p value was greater than 0.1, this indicated a poor fit for

both exponential and normal, thus the TF was classified as "other" (Additional file 1: Fig. S9).

We generated the cumulative frequency plots for the TF gene transcription across control samples using a bin size of 0.001 RPKM. The mean cumulative frequency across high confidence TFs was plotted as a black line and the standard error is shown as a colored region across the high confidence TFs (Fig. 4B). The high confidence TFs were selected by significance values (p value ≤ 0.001). These TFs were parsed down to a group of TFs with a similar range of transcription level such that they could be plotted on the same x-axis for the cumulative frequency plots. The most confident tissue specific factors within their respective tissue types were identified as MyoD, GRHL2, TEAD4, p63, GATA1, and Oct4. The most confident ubiquitous factors across all tissue types were identified as NFYB, ELK1, CREB1, SP1, ATF1, and SP3.

Single cell RNA-seq data

For single cell RNA-seq data (scRNA-seq), we used data published from the "human cell atlas of fetal gene expression" [63]. This data was accessed from NCBI GEO accession number GSE156793. We used the publicly available file titled: "GSE156793_S6_gene_expression_celltype." To be considered "expressed" in a given tissue we used the expression cutoff of 0.01. scRNA-seq expression values were fit to either an exponential or normal distribution by Fitter as described in Fit classification for transcription level of TFs.

Differential expression

Gene counts previously quantified from DBNascent were used for differential expression analysis (see Counting over genes and bidirectionals section). We focused on 6 data sets in which there was TNFα treatment and their corresponding controls. Data sets from the same tissue type were grouped within a single DESeq2 (version 1.26.0) object. Differential gene expression was assessed between TNFα vs. control separately for the 4 tissues represented (heart [26, 66] $n = 8$ control samples, $n = 7$ treatment samples; lung [68, 69] $n = 5$ control samples, $n = 5$ treatment samples; intestine [67], $n = 2$ control samples, $n = 2$ treatment samples; prostate [70], $n = 2$ control samples, $n = 2$ treatment samples). Additional sample information is defined within Additional file 2 and extended information on these samples is on github in a file called sample_metadata_table and differential expression results can found on the additional data github page. NFκB targets were defined from the GSEA Hallmarks (version 5.0) pathway: TNFα signaling via NFκB ($n = 200$ genes).

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03545-2.

Additional file 1. Contains Supplementary Figures.

Additional file 2. Contains information on all nascent run-on sequencing samples utilized. For each sample we provide cell type, original paper, protocol, organism, sample type, tissue, treatment and GEO SRR accession numbers.

Additional file 3. Contains information for all ChIP data utilized. For each sample we include the Cistrome ID, species, factor ID, cell line, cell type, tissue and corresponding HOCOMOCO TF identifier.

Additional file 4. Peer review history.

Acknowledgements
We are grateful to the BioFrontiers IT department for their support in building the database and to the members of the Dowell and Allen labs for curating the nascent database.

Review history
The review history is available as Additional file 4.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
TJ, MAA, and RDD conceptualized the body of work and interpreted results. MAA and RDD supervised the study. DJT aided with conceptualization and interpretation. RFS and LS collected and derived quality control metrics for the nascent sequencing data within the DBNascent construction. TJ conducted all data analysis. MAA aided in code development. TJ and RDD wrote the paper. All authors revised the final manuscript.

Funding
This work was funded by the National Science Foundation under grants ABI1759949 and the National Institutes of Health grant GM125871 and HL156475.

Data availability
The stand-alone TF Profiler application is available at on github (https://github.com/Dowell-Lab/TF_profiler) and Zenodo [76]. Additional stand-alone scripts and useful data files associated with this publication can be found on github (https://github.com/Dowell-Lab/TF_profiler_additonal_scripts) and Zenodo [77]. TF Profiler and the TF Profiler additional scripts are licensed under the GPL v3.0.

Declarations

Ethics approval and consent to participate
Ethics approval is not applicable.

Competing interests
Drs. Dowell and Allen are on a patent that uses enhancer RNAs to infer transcription factor activity. The patent does not place any restriction on reuse and reproducibility of this manuscript. The other authors declare that they have no competing interests.

Received: 8 January 2024 Accepted: 17 March 2025
Published online: 09 April 2025

References

1. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013;152(6):1237–51.
2. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. Cell. 2018;172(4):650–65.
3. ENCODE Project Consortium, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.
4. Inukai S, Kock KH, Bulyk ML. Transcription factor-DNA binding: beyond binding site motifs. Curr Opin Genet Dev. 2017;43:110–9.
5. Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, et al. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. 2008;6(2):e27.
6. Brackley CA, Johnson J, Kelly S, Cook PR, Marenduzzo D. Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains. Nucleic Acids Res. 2016;44(8):3503–12.
7. Frietze S, Farnham PJ. Transcription factor effector domains. Handbook of Transcription Factors. 2011;261–77. https://doi.org/10.1007/978-90-481-9069-0.
8. Soto LF, Li Z, Santoso CS, Berenson A, Ho I, Shen VX, et al. Compendium of human transcription factor effector domains. Mol Cell. 2022;82(3):514–26.
9. Salghetti SE, Kim SY, Tansey WP. Destruction of Myc by ubiquitin-mediated proteolysis: cancer-associated and transforming mutations stabilize Myc. EMBO J. 1999;18(3):717–26.
10. Haupt Y, Maya R, Kazaz A, Oren M. Mdm2 promotes the rapid degradation of p53 [journal article]. Nature. 1997;387(6630):296–9. https://doi.org/10.1038/387296a0.
11. Honda R, Tanaka H, Yasuda H. Oncoprotein MDM2 is a ubiquitin ligase E3 for tumor suppressor p53. FEBS Lett. 1997;420(1):25–7. https://doi.org/10.1016/s0014-5793(97)01480-4.
12. Kubbutat MHG, Jones SN, Vousden KH. Regulation of p53 stability by Mdm2. Nature. 1997;387(66306630):299–303. https://doi.org/10.1038/387299a0.
13. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013;339(6122):950–3.
14. Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc. 2016;11(8):1455–76.

15. Core L, Lis J. Transcription regulation through promoter-proximal pausing of RNA polymerase II. Science. 2008;319:1791.
16. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322(5909):1845–8.
17. Azofeifa JG, Dowell RD. A generative model for the behavior of RNA polymerase. Bioinformatics. 2016;33(2):227–34.
18. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465(7295):182–7.
19. Schwalb B, Michel M, Zacher B, Frühauf K, Demel C, Tresch A, et al. TT-seq maps the human transient transcriptome. Science. 2016;352(6290):1225–8.
20. Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, et al. Identification of active transcriptional regulatory elements from GRO-seq data. Nat Meth. 2015;12(5):433–8.
21. Azofeifa JG, Allen MA, Hendrix JR, Read T, Rubin JD, Dowell RD. Enhancer RNA profiling predicts transcription factor activity. Genome Res. 2018;28(3):334–44. https://doi.org/10.1101/gr.225755.117.
22. Hah N, Murakami S, Nagari A, Danko CG, Kraus WL. Enhancer transcripts mark active estrogen receptor binding sites. Genome Res. 2013;23(8):1210–23.
23. Hardin PE, Panda S. Circadian timekeeping and output mechanisms in animals. Curr Opin Neurobiol. 2013;23(5):724–31.
24. Allen MA, Mellert H, Dengler V, Andryzik Z, Guarnieri A, Freeman JA, et al. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. eLife. 2014;3:e02200.
25. Puc J, Kozbial P, Li W, Tan Y, Liu Z, Suter T, et al. Ligand-dependent enhancer activation regulated by topoisomerase-I activity. Cell. 2015;160(3):367–80.
26. Luo X, Chae M, Krishnakumar R, Danko CG, Kraus WL. Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFα signaling revealed by integrated genomic analyses. BMC Genomics. 2014;15:155.
27. Rubin JD, Stanley JT, Sigauke RF, Levandowski CB, Maas ZL, Westfall J, et al. Transcription factor enrichment analysis (TFEA): quantifying the activity of hundreds of transcription factors from a single experiment. Nat Commun Biol. 2021. https://doi.org/10.1101/2020.01.25.919738.
28. Wang Z, Chu T, Choate LA, Danko CG. Identification of regulatory elements from nascent transcription using dREG. Genome Res. 2019;29(2):293–303.
29. Kristjánsdóttir K, Dziubek A, Kang HM, Kwak H. Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. Nat Commun. 2020;11(1):5963.
30. Bae S, Kim K, Kang K, Kim H, Lee M, Oh B, et al. RANKL-responsive epigenetic mechanism reprograms macrophages into bone-resorbing osteoclasts. Cell Mol Immunol. 2023;20(1):94–109.
31. Allison KA, Kaikkonen MU, Gaasterland T, Glass CK. Vespucci: a system for building annotated databases of nascent transcripts. Nucleic Acids Res. 2014;42(4):2433–47.
32. Yao L, Liang J, Ozer A, Leung AKY, Lis JT, Yu H. A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. Nat Biotechnol. 2022;40:1–10. https://doi.org/10.1038/s41587-022-01211-7.
33. Azofeifa JG, Allen MA, Lladser ME, Dowell RD. An annotation agnostic algorithm for detecting nascent RNA transcripts in GRO-Seq. IEEE/ACM Trans Comput Biol Bioinforma. 2017;14(5):1070–81.
34. Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, et al. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. 2012;22(12):2399–408.
35. Antequera F, Bird A. Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci. 1993;90(24):11995–9.
36. Bird AP. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 1980;8(7):1499–504.
37. Ioshikhes IP, Zhang MQ. Large-scale human promoter mapping using CpG islands. Nat Genet. 2000;26(1):61–3.
38. Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, et al. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells [journal article]. Proc Natl Acad Sci U S A. 2013;110(8):2876–81. https://doi.org/10.1073/pnas.1221904110.
39. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018;46(D1):D252–9.
40. Wang A, Yue F, Li Y, Xie R, Harper T, Patel NA, et al. Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. Cell Stem Cell. 2015;16(4):386–99.
41. Bouvy-Liivrand M, Hernández de Sande A, Pölönen P, Mehtonen J, Vuorenmaa T, Niskanen H, et al. Analysis of primary microRNA loci from nascent transcriptomes reveals regulatory domains governed by chromatin architecture. Nucleic Acids Res. 2017;45(17):9837–9749.
42. Estarás C, Benner C, Jones KA. SMADs and YAP compete to control elongation of β-catenin:LEF-1-recruited RNAPII during hESC differentiation. Mol Cell. 2015;58(5):780–93.
43. Nieto MA. The snail superfamily of zinc-finger transcription factors. Nat Rev Mol Cell Biol. 2002;3(3):155–66.
44. Nerlov C. The C/EBP family of transcription factors: a paradigm for interaction between gene expression and proliferation control. Trends Cell Biol. 2007;17(7):318–24.
45. Chong JL, Wenzel PL, Sáenz-Robles MT, Nair V, Ferrey A, Hagan JP, et al. E2f1-3 switch from activators in progenitor cells to repressors in differentiating cells. Nature. 2009;462(7275):930–4.
46. Davis RL, Weintraub H, Lassar AB. Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell. 1987;51(6):987–1000. https://doi.org/10.1016/0092-8674(87)90585-X.
47. LeRoy G, Oksuz O, Descostes N, Aoi Y, Ganai RA, Kara HO, et al. LEDGF and HDGF2 relieve the nucleosome-induced barrier to transcription in differentiated cells. Sci Adv. 2019;5(10):eaay3068.
48. Blumberg A, Zhao Y, Huang YF, Dukler N, Rice EJ, Chivu AG, et al. Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. BMC Biol. 2021;19(1):30.

49. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet. 2014;46(12):1311–20.
50. Dukler N, Booth GT, Huang YF, Tippens N, Waters CT, Danko CG, et al. Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. Genome Res. 2017;27(11):1816–29.
51. Niskanen EA, Malinen M, Sutinen P, Toropainen S, Paakinaho V, Vihervaara A, et al. Global SUMOylation on active chromatin is an acute heat stress response restricting transcription. Genome Biol. 2015;16:153.
52. Vihervaara A, Mahat DB, Guertin MJ, Chu T, Danko CG, Lis JT, et al. Transcriptional response to stress is pre-wired by promoter and enhancer architecture. Nat Commun. 2017;8(1):255.
53. Hsieh PN, Fan L, Sweet DR, Jain MK. The Krüppel-like factors and control of energy homeostasis. Endocr Rev. 2019;40(1):137–52.
54. Wu L, Timmers C, Maiti B, Saavedra HI, Sang L, Chong GT, et al. The E2F1-3 transcription factors are essential for cellular proliferation. Nature. 2001;414(6862):457–62.
55. Sharrocks AD. The ETS-domain transcription factor family. Nat Rev Mol Cell Biol. 2001;2(11):827–37.
56. Zhao Y, Vartak SV, Conte A, Wang X, Garcia DA, Stevens E, et al. "Stripe" transcription factors provide accessibility to co-binding partners in mammalian genomes. Mol Cell. 2022;82(18):3398-3411.e11. https://doi.org/10.1016/j.molcel.2022.06.029.
57. Dejosez M, Dall'Agnese A, Ramamoorthy M, Platt J, Yin X, Hogan M, et al. Regulatory architecture of housekeeping genes is driven by promoter assemblies. Cell Rep. 2023;42(5):112505. https://doi.org/10.1016/j.celrep.2023.112505.
58. Lidschreiber K, Jung LA, von der Emde H, Dave K, Taipale J, Cramer P, et al. Transcriptionally active enhancers in human cancer cells. Mol Syst Biol. 2021;17(1):e9873.
59. Lee SA, Kristjánsdóttir K, Kwak H. eRNA co-expression network uncovers TF dependency and convergent cooperativity. Sci Rep. 2023;13(1):19085.
60. Sigauke RF, Sanford L, Maas ZL, Jones T, Stanley JT, Townsend HA, et al. Atlas of nascent RNA transcripts reveals enhancer to gene linkages. bioRxiv. 2023;2023–12. https://doi.org/10.1101/2023.12.07.570626.
61. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res. 2017;45:D658–62. https://doi.org/10.1093/nar/gkw983.
62. Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res. 2019;47(D1):D729–35.
63. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, et al. A human cell atlas of fetal gene expression. Science. 2020;370(6518):eaba7721.
64. Wade M, Wong ET, Tang M, Stommel JM, Wahl GM. Hdmx modulates the outcome of p53 activation in human tumor cells. J Biol Chem. 2006;281(44):33036–44.
65. Huang L, Yan Z, Liao X, Li Y, Yang J, Wang ZG, et al. The p53 inhibitors MDM2/MDMX complex is required for control of p53 activity in vivo. Proc Natl Acad Sci. 2011;108(29):12001–6.
66. Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Mol Cell. 2013;50(2):212–22.
67. Rahnamoun H, Lu H, Duttke SH, Benner C, Glass CK, Lauberth SM. Mutant p53 shapes the enhancer landscape of cancer cells in response to chronic immune signaling. Nat Commun. 2017;8(1):754.
68. Sasse SK, Gruca M, Allen MA, Kadiyala V, Song T, Gally F, et al. Nascent transcript analysis of glucocorticoid crosstalk with TNF defines primary and cooperative inflammatory repression. Genome Res. 2019;29(11):1753–65.
69. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503(7475):290–4.
70. Malinen M, Niskanen EA, Kaikkonen MU, Palvimo JJ. Crosstalk between androgen and pro-inflammatory signaling remodels androgen receptor and NF-κB cistrome to reprogram the prostate cancer cell transcriptome. Nucleic Acids Res. 2017;45(2):619–30.
71. Capece D, Verzella D, Flati I, Arboretto P, Cornice J, Franzoso G. NF-κB: blending metabolism, immunity, and inflammation. Trends Immunol. 2022;43(9):757–75.
72. Zhang JP, Zhang H, Wang HB, Li YX, Liu GH, Xing S, et al. Down-regulation of Sp1 suppresses cell proliferation, clonogenicity and the expressions of stem cell markers in nasopharyngeal carcinoma. J Transl Med. 2014;12(1):1–12.
73. Rabacal W, Pabbisetty SK, Hoek KL, Cendron D, Guo Y, Maseda D, et al. Transcription factor KLF2 regulates homeostatic NK cell proliferation and survival. Proc Natl Acad Sci. 2016;113(19):5370–5.
74. Hsu T, Trojanowska M, Watson DK. Ets proteins in biological control and cancer. J Cell Biochem. 2004;91(5):896–903.
75. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods. 2017;14(10):975–8.
76. Jones T, Allen M, Dowell R. TF Profiler. 2025. https://doi.org/10.5281/zenodo.14953546.
77. Jones T, Sigauke RF, Sanford L, Allen MA, Dowell R. TF Profiler supporting files. 2025. https://doi.org/10.5281/zenodo.14953532.
78. Sanford L, Tripodi I, Gruca M, Allen M, Dowell R. Nascent-Flow. 2025. https://doi.org/10.5281/zenodo.14953945.
79. Sigauke R, Sanford L, Allen M, Dowell R. Nascent-Flow. 2025. https://doi.org/10.5281/zenodo.14953943.
80. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
81. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43(W1):W39–49.
82. Hunter S, Sigauke RF, Stanley JT, Allen MA, Dowell RD. Protocol variations in run-on transcription dataset preparation produce detectable signatures in sequencing libraries. BMC Genomics. 2022;23(1):1–18.

Publisher's Note