Received: 17 May 2023 | Revised: 11 August 2023 | Accepted: 22 August 2023

DOI: 10.1002/pmic.202200462

Proteomics
Proteomics and Systems Biology

REVIEW

Computational tools for inferring transcription factor activity

Dennis Hecker[1,2,3] | Michael Lauber[4] | Fatemeh Behjati Ardakani[1,2,3] | Shamim Ashrafiyan[1,2,3] | Quirin Manz[4] | Johannes Kersting[4,5] | Markus Hoffmann[4,6,7] | Marcel H. Schulz[1,2,3] | Markus List[4]

[1] Goethe University Frankfurt, Frankfurt am Main, Germany

[2] German Center for Cardiovascular Research, Partner site Rhein-Main, Frankfurt am Main, Germany

[3] Cardio-Pulmonary Institute, Goethe University Hospital, Frankfurt am Main, Germany

[4] Big Data in BioMedicine Group, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany

[5] GeneSurge GmbH, München, Germany

[6] Institute for Advanced Study, Technical University of Munich, Garching, Germany

[7] National Institute of Diabetes, Digestive, and Kidney Diseases, National Institutes of Health, Bethesda, Maryland, USA

Correspondence
Markus List, Big Data in BioMedicine Group, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, 85354 Freising, Germany.
Email: markus.list@tum.de

The authors wish to be known that, in their opinion, the first two and last two authors should be considered as shared first and last authors, respectively.

Funding information
German Excellence Initiative; German Centre for Cardiovascular Research, Grant/Award Number: 81Z0200101; [EXC2026]ID, Grant/Award Number: 390649896; Cardio-Pulmonary Institute, Grant/Award Number: 390649896; National Institute of Diabetes and Digestive and Kidney Diseases; Deutsche Forschungsgemeinschaft, Grant/Award Numbers: SFB1531, 456687919, TRR 267, 403584255; Swiss State Secretariat for Education, Research and Innovation, Grant/Award Number: 22.00115; Institute for Advanced Study, Technische Universität München; H2020 European Research Council, Grant/Award Number: 101057619

Abstract

Transcription factors (TFs) are essential players in orchestrating the regulatory landscape in cells. Still, their exact modes of action and dependencies on other regulatory aspects remain elusive. Since TFs act cell type-specific and each TF has its own characteristics, untangling their regulatory interactions from an experimental point of view is laborious and convoluted. Thus, there is an ongoing development of computational tools that estimate transcription factor activity (TFA) from a variety of data modalities, either based on a mapping of TFs to their putative target genes or in a genome-wide, gene-unspecific fashion. These tools can help to gain insights into TF regulation and to prioritize candidates for experimental validation. We want to give an overview of available computational tools that estimate TFA, illustrate examples of their application, debate common result validation strategies, and discuss assumptions and concomitant limitations.

KEYWORDS
bioinformatic tools, gene regulation, gene regulatory networks, transcription factor activity

Abbreviations: GRN, gene regulatory network; MPR, massively parallel reporter assay; NCA, network component analysis; PC, principal component; PCA, principal component analysis; PWM, position weight matrix; TF, transcription factor; TFA, transcription factor activity; TFBS, transcription factor binding site.

1 | INTRODUCTION

Transcription factors (TFs) are essential proteins that regulate gene expression by binding to specific DNA sequences in the promoter or enhancer regions of genes [1, 2]. They exert their regulatory activity via a diverse range of mechanisms, such as recruiting cofactors, remodeling of the chromatin state, altering epigenetic modifications, or interacting directly with the transcription machinery [3–5]. TFs are crucial actors in many cellular processes, including development, differentiation, and response to environmental stimuli [6–8]. It is believed that nearly half of all known TFs are expressed in any cell type, although only a small number of them are thought to be sufficient for establishing the cell type-defining gene expression programs [9, 10]. Due to their role in central biological processes, their dysregulation is observed in various diseases [11–13]. Consequently, it is of great interest to gain insights into transcription factor activity (TFA). We define TFA as the regulatory impact that a TF exerts on the expression of each of its target genes, which includes any form of regulation, may it be activation or repression or other effects on transcription like alternative splicing. Instead of retrieving TFA for each individual gene, it is often summarized as a TF's influence on a set of genes, or more generally as a TF's importance for a cell state or a certain condition. TFA can be influenced by various mechanisms, including epigenetic modifications, post-transcriptional regulation, post-translational modifications, protein–protein interactions, presence of cofactors, localization, or DNA structural changes [14–16] (Figure 1). Investigation of TFA on all regulatory levels represents a significant research challenge, as TFs can have numerous target genes, each controlled by potentially various enhancer regions in a cell type-specific manner [15, 17]. For many TFs, there is limited information about which genes or processes they influence and whether they act as repressors or activators. Given the prohibitive cost and time required to investigate all possible combinations of regulatory players experimentally, many computational tools have been proposed to analyze TFA. We aim to provide a comprehensive overview of these existing computational approaches.

This review first gives a brief outline of experimental protocols and data modalities for TFA inference. We then present the various computational tools partitioned into two main categories (Figure 2, Table 1). Methods in the first category, referred to as gene regulatory network (GRN)-based methods, rely on a TF to gene mapping in order to estimate TFA. They either use a pre-built network or create a network *de novo* based on tool-specific data modalities. TFA is then typically inferred for a TF and its target genes, which together form a so-called regulon [18]. Genome occupancy-based tools, on the other hand, assess TFA by the genome-wide binding behavior of a TF, independent of individual target genes. The TFA inferred by these tools can be seen as a higher level estimate of TFA that summarizes the effects on individual genes. We highlight the strengths and limitations of each approach and give a comprehensive overview of the available methods, including a decision tree separating the tools by the data they use (Figure 3). We further describe experimental setups for a more direct TFA readout, discuss prevalent validation endeavors with their shortcomings, present example applications of TFA inference, and point to key research gaps and opportunities for future developments.

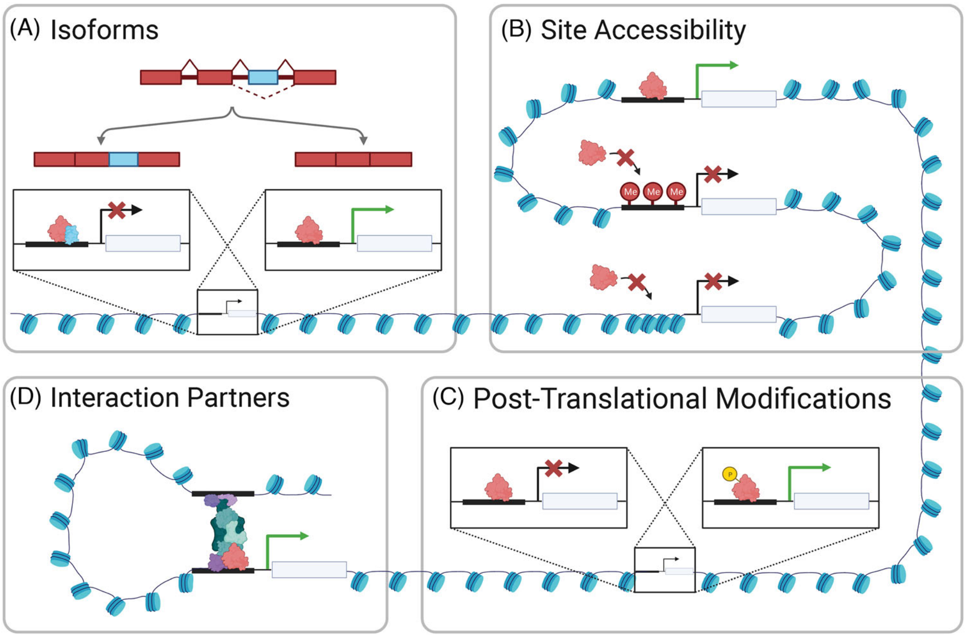

FIGURE 1 The complexity of transcription factor activity (TFA). The activity of a transcription factor (TF) can be influenced by numerous, potentially interacting factors: (A) different isoforms of the same TF can have different functions. (B) Chromatin accessibility, and thus the ability of a TF to bind its target, can be limited by DNA methylation or chromatin structure. (C) TFA can be influenced by post-translational modifications. (D) Many different interaction partners are involved in transcription, for example, other TFs, mediator proteins, and cofactors. All of which can potentially affect TFA. Created with BioRender.com.

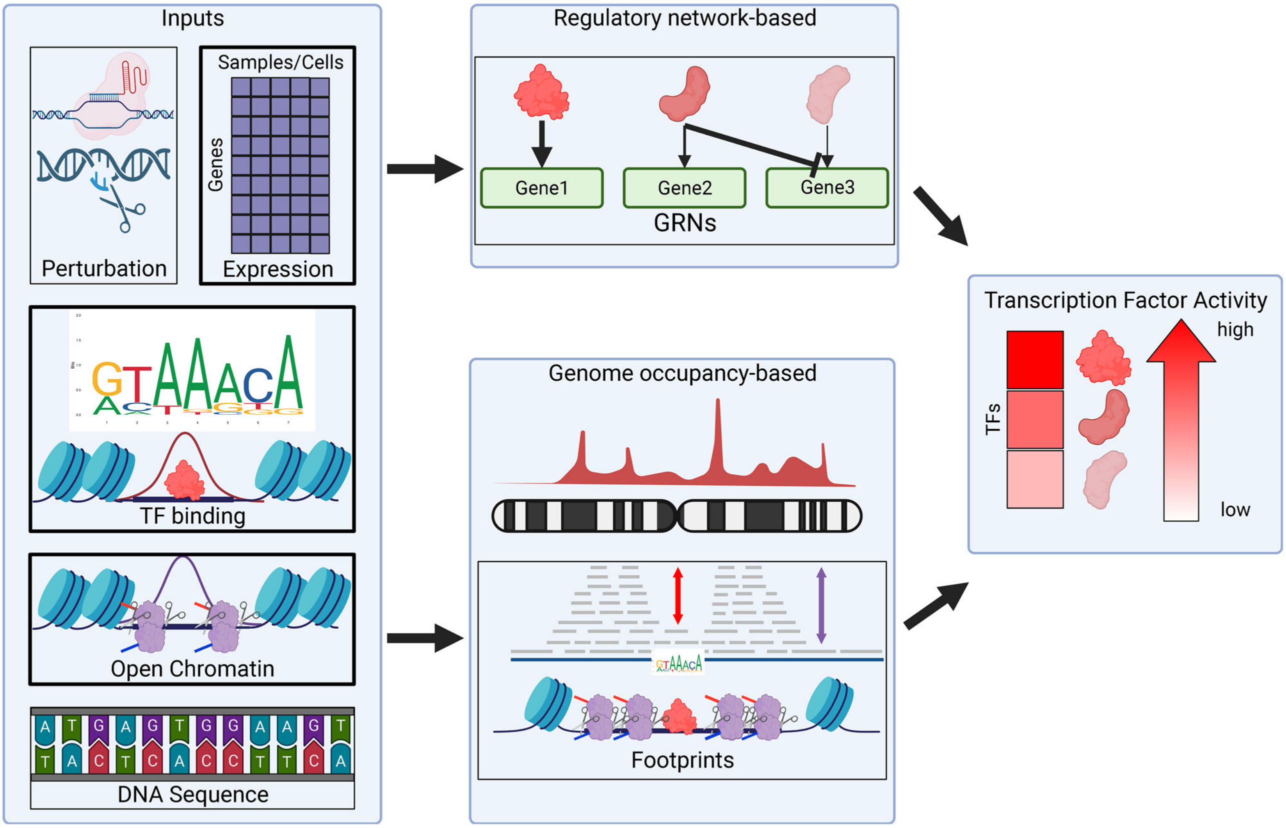

FIGURE 2 Overview of the general setup of computational tools to infer transcription factor activity (TFA). Among the most frequently used data types are gene expression, information on transcription factor (TF) binding, and open chromatin using restriction enzyme accessibility. Some approaches also incorporate perturbation data or use the DNA sequence for deep learning. Regulatory network-based methods depend on linking TFs to potential target genes, often by creating, including, or refining a gene-regulatory network. Conversely, genome occupancy-based approaches follow a target gene-agnostic paradigm that utilizes genome-wide signals, for example, TF footprints. Created with BioRender.com.

1.1 | Data modalities in TFA inference

There is a multitude of different modalities coming from various biological assays that are used as input for computational tools to infer TFA, either individually or in an integrated fashion. The most prominent modality is the transcriptome, with RNA-seq as standard assay to quantify the expression of genes. While measurements of the transcript levels can help in finding TFs which are present in a cell type, it is also frequently used to correlate TF expression with the overall gene expression across samples or to find differential expression between conditions.

Another major type of data is information on TF binding sites (TFBSs) occupancy. TFBS occupancy can either be measured experimentally, for example, with ChIP-seq [19, 20], or predicted by quantifying the agreement of a TF binding motif to the DNA sequence. The binding motif of a TF is a representation of the preferred binding sequence, commonly encoded in a position weight matrix (PWM). PWMs are the probabilistic quantification of the nucleotide frequencies observed at a TF's binding sites [15]. Since the genome contains magnitudes more potential binding motifs than are actually bound by the cognate TFs, alternative modalities are often included to yield a more accurate TFBS prediction. It was found that TF binding correlates with multiple epigenetic data types, such as measurements of chromatin accessibility, specific histone modifications, or the presence of other cofactors [21, 22]. In practice, a common approach in addressing the rate of false positive motif-based TFBS predictions is to limit the search space to regions harboring such epigenetic marks, usually via peak calling. Prominent assays in this context are DNase-seq and ATAC-seq for finding regions with open chromatin, both working with enzymes that cleave accessible DNA [23, 24]. Those two assays can additionally be used for footprint identification. Footprints are regions within open chromatin where the binding of a TF prevents the cleavage enzymes to cut, which leads to a characteristic drop in read coverage, and thus can increase accuracy of TFBS predictions [19, 25–27].

Notably, other modalities that contribute to TFA, such as the quantity of proteins and their post-translational modifications, [15, 16, 28], are not yet widely utilized in TFA inference. Handling such modalities is challenging, due to the scarcity of data, missing knowledge on their precise biological role, and how to properly incorporate them in computational tools. Therefore, we do not further detail them here.

TABLE 1 Overview of computational tools for TFA inference.

Tool	Reference	bulk RNA	scRNA	Bulk DNase/ ATAC	scATAC	Other resource	Method/ algorithm	Quantitative score	Significance analysis	Output	Package	Language
Regulatory network-based												
Ma and Brent	[42]	✓	X	X	X	Prior network, (motif/ChIP-seq, perturbation data)	Bilinear model	✓	X	TFA per sample, general GRN with mode of action	X	Python
RACER	[43]	✓	X	X	X	ChIP-seq, (CNV, 5mC, miRNA)	Two-stage framework	✓	✓	TFA per sample, TF-gene inter- actions	X	R
RABIT	[44]	✓	X	X	X	ChIP-seq, (CNV, 5mC, somatic mutations)	Linear regression	✓	✓	General TFA	(✓)	C++
biRte	[45]	✓	X	X	X	Prior network, (CNV, miRNA and TF-TF interactions)	Bayesian framework using Markov-Chain-Monte-Carlo	✓	X	Set of TFs explaining differential expression	✓	R
SEPIRA	[47]	✓	X	X	X	Collection of bulk RNA-seq, (5mC)	Linear regression	✓	✓	TFA per sample, mode of action	✓	R
SCIRA	[63]	X	✓	X	X	Collection of bulk RNA-seq	Linear regression	✓	✓	TFA per cell, mode of action	(✓)	R
VIPER	[18]	✓	X	X	X		Regulon (aREA) enrichment	✓	✓	Differential TFA, mode of action	✓	R
metaVIPER	[64]	✓	✓	X	X		Regulon (aREA) regulon (aREA)	✓	✓	Differential TFA, mode of action	✓	R
NetProphet 3	[49]	✓	X	X	X	Any TF-gene evidence	XGBoost [154] on multiple approaches	✓	X	Probability of direct functional binding of a TF to a gene	✓	R, Shell, Python
ISMARA	[52]		X	X	X	Motifs, (ChIP-seq in- stead of RNA-seq)	Linear model with Bayesian procedure	✓	X	General TFA, TF motif importance across samples, TF-gene interactions	X	
FindIT2	[53]	✓	X	✓	X	Motifs/ChIP-seq	Regulatory potential [54], enrichment in subsets	✓	X	General TFA, TF-gene scores	✓	R
Taiji	[56]	✓	X	✓	X	Motifs	Personalized PageRank	X	X	Differential TFA, TFs for transdifferentiation	✓	Haskell

(Continues)

TABLE 1 (Continued)

Tool	Reference	bulk RNA	scRNA	Bulk DNase/ ATAC	scATAC	Other resource	Method/ algorithm	Quantitative score	Significance analysis	Output	Package	Language
TEPIC	[156]	✓]	X	✓	X	Motifs/ChIP-seq, (chromatin interactions)	Elastic net regression	✓	X	TF-gene scores, TFA in sample, differential TFA, TFA in time series, mode of action	X	Python, R, Shell
TF-Prioritizer	[59]	✓	X	✓	X	Motifs/ChIP-seq	Logistic regression	✓	✓	TF-gene scores, footprints, differential TFA, mode of action	✓	Java, Python, Shell
TRIANGU-LATE	[62]	X	✓	✓	X	Motifs/ChIP-seq	Multi-task learning regression	✓	X	TFA per cell	X	R
BITFAM	[65]	X	✓	X	X	TF-binding database	Bayesian factor analysis	✓	X	TFA per cell, cell clustering, trajectory generation	(✓)	R
SCENIC	[66]	X	✓	X	X	Motifs	Regulon enrichment (AU-Cell)	X	X	Active regulons per cell, cell clustering	✓	Python, R
SCENIC+	[68]	X	✓	X	✓	Motifs	Gradient Boosting Machine regression, correlation, AU-Cell	✓	X	Regulons of TF-regions-genes, in silico TF perturbation	✓	Python
Inferelator 3.0	[69]	✓	✓	X	X	Prior network	Regularized regression with different model options	✓	X	Cell type-specific GRNs	✓	Python
TIGER	[70]	✓	✓	X	X	Prior network	Bayesian matrix factorization, Variational Bayes	✓	X	TFA per sample/cell, general GRN with mode of action	(✓)	R
Genome occupancy-based												
HINT	[27]	X	X	✓	X	Motifs	HMMs	✓	✓	TFA per sample, TF footprints, differential TFA	✓	Python
TOBIAS	[19]	X	X	✓	X	Motifs	Correlation of footprint score with TFBS and cut- off selection	✓	✓	TFA per sample, TF footprints, differential TFA, TF-gene interactions, TF-TF networks	✓	Python, Cython, Shell
BaGFoot	[72]	X	X	✓	X	Motifs	Quantification of footprint features between conditions	✓	✓	TFA per sample, TF footprints,differential TFA	X	R
diffTF	[73]	(✓)	X	✓	X	Motifs/ChIP-seq	Statistics on accessibility changes	✓	✓	Differential TFA, mode of action	✓	R, Python, Shell

(Continues)

TABLE 1 (Continued)

Tool	Reference	bulk RNA	scRNA	Bulk DNase/ ATAC	scATAC	Other resource	Method/ algorithm	Quantitative score	Significance analysis	Output	Package	Language
chromVAR	[74]	*X*	*X*	✓	✓	Motifs/ChIP-seq	Accessibility deviation	✓	✓	TFA per sample/cell, differential TFA	✓	R, C++
chromVAR-Multiome	[75]	*X*	✓	✓	✓	Motifs	Modified chromVAR	✓	✓	TFA per sample/cell, differential TFA, in silico TF perturbations	*X*	R, Python
BROCKMAN	[76]	*X*	*X*	✓	✓	Motifs	Matrix factorization and PCA on gapped k-mer frequencies	✓	✓	Differential TFA, TF-TF interactions	✓	R
scFAN	[77]	*X*	*X*	✓	✓	Motifs	CNN	✓	*X*	TFA per cell	✓	Python
DeepAccess	[79, 80]	*X*	*X*	✓	*X*	Motifs	CNN	✓	✓	TFA per sample, differential TFA, in silico motif perturbation	✓	Python, Shell
scBasset	[81]	*X*	*X*	*X*	✓	Motifs	CNN	✓	*X*	TFA per cell, in silico motif perturbation	✓	Python

Note: Optional inputs in *Other resource* are enclosed by parentheses. If a tool requires a prior network, that network can be constructed from any source of data, such as literature-based databases or ChIP-seq catalogs. Mode of action means that tools predict whether a TF has an activating or repressive influence. *Quantitative score* indicates whether a tool returns a numeric TFA. *Significance analysis* shows whether a *p*-value is provided. We consider a software to provide a *Package* if it is either installable via common package management systems such as bioconda, or via container platforms like Docker. If such an environment is not provided, but the code comes with an integrated documentation (e.g., .Rd-files for R), it is labeled with a checkmark in parentheses. For programming languages, we list the three most used ones. CNN, convolutional neural network; CNV, copy number variation; 5mC, cytosine methylation; HMMs, Hidden Markov model; TF, transcription factor; TFA, transcription factor activity.

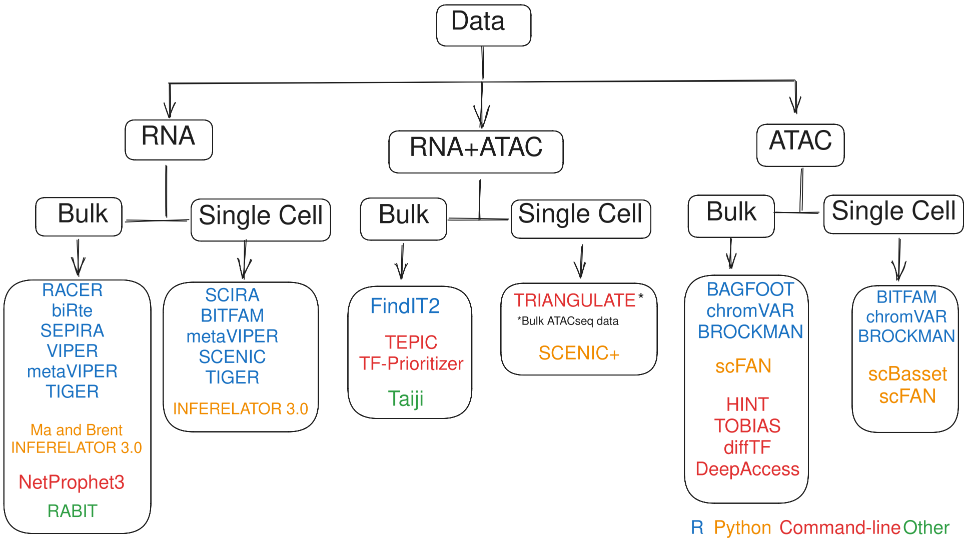

FIGURE 3 Decision tree representation of input data and methods. The nodes represent the input data, while the leaves correspond to methods with available packages. Each method is color-coded based on its corresponding programming language.

2 | COMPUTATIONAL TOOLS FOR TFA INFERENCE

2.1 | Regulatory network-based approaches

A plethora of tools define TFA via the interaction of TFs with putative target genes and construct TF- or gene-specific regulons or try to assemble whole regulatory networks. We only consider GRNs as estimate of TFA, if the strengths of the TF-gene interactions are quantified, that is, as edge weights in a graph.

The first methods approximating TFA emerged in the early 2000s and were based on linear regression. With the limited amount of data available at that time, studies were mainly focused on yeast strains, which contain a much lower number of genes and TFs than mammalian cells [29]. In general, the methods searched the upstream region of differentially expressed genes for shared sequence motifs and tried to explain expression log-ratios based on motif score and occurrence. The coefficients of a motif that implicitly represented a TF acted as a proxy for a TF's activity [30–33]. With the rising availability of ChIP data subsequent studies integrated the ChIP signal into the model [34]. A popular method that was developed in 2003 was the network component analysis (NCA) technique [35]. NCA generates a matrix of log-ratios of expression values for multiple samples as the product of a matrix of control strength and a matrix of TFAs. Each row in the control strength matrix represents the potential influence of TFs on a gene, while a column in the TFA matrix represents the activity of all TFs per sample. Prior information derived from ChIP experiments was incorporated into the control strength matrix where TF-gene interactions without evidence were set to zero. To be able to uniquely decompose the expression matrix, a set of constraints was imposed onto the TFA and control strength matrices. In its original implementation, NCA had several limitations which were tackled by subsequent methods. *FastNCA* provided an implementation with an improved computational complexity and run time [36]. *gNCA* allowed to incorporate prior information from knockout experiments [37]. *ROBNCA* improved the robustness of the algorithms by explicitly integrating noise and outliers of the expression data into the model [38]. As NCA only checks upon initialization if all constraints are satisfied, *gfNCA* ensures that no violations occur during the iteration steps [39]. *sparseNCA* extended NCA to be able to handle the incompleteness of the prior information [40]. *LNCA* was adapted to cope with expression data sets that show high heterogeneity such as cancer samples or samples covering different cell states. In order to accomplish this, *LNCA* creates local expression profiles with their corresponding control strength matrices by partitioning the expression data using the k-nearest neighbor algorithm and then finds an optimal global solution [41].

Similarly, Ma and Brent also use a control strength and TFA matrix and describe TF regulation as the product of these matrices with a bilinear model [42]. They further specify positive control strengths to indicate an activating effect and negative values as a repressive effect, while the TFA matrix is constrained to positive values. Their model performed best when the signs of the control strength were predetermined using TF perturbation data. Initializing the control strength matrix with ChIP-seq data led to a drop in performance, which dropped even further when using motif-based TFBS prediction.

TFA analysis in heterogeneous samples, such as those derived from cancer patients, is complicated by confounding factors. For instance, copy number variations can lead to differential gene expression even when the activity of the regulatory factor remains constant between two samples. To overcome this issue, *RACER* was developed to estimate TFAs in such conditions [43]. The authors use a regularized linear regression model in which gene expression depends on DNA methylation levels, copy number variations, miRNA levels, and the product of TFA and TF binding strength. A feature selection procedure provides a further metric to assess the importance of the TFs by comparing the model's performance when a TF is left out.

Similarly, Jiang et al. control for the influence of copy number, DNA methylation, and TF somatic mutation in their regression framework called *RABIT* [44]. However, their method focuses on differences in TFAs between samples by using differential gene expression as the dependent variable, and the regulatory activity of a TF is based on a significance test of the regression coefficients.

biRte is another method that is able to utilize data beyond transcriptome information [45]. By integrating differential gene expression data and a regulatory network into a joint probabilistic framework, the model uncovers TFs that drive differences between two groups. mRNA expression levels are modeled using a sparse Bayesian linear regression and Markov-Chain Monte-Carlo sampling infers the activity states of the regulatory factors. Additionally, *biRte* can incorporate further information like prior knowledge of TFAs, TF-TF interactions, and miRNA expression data into the model.

SEPIRA relies on a large collection of expression data from public compendia, such as GTEx [46], to construct sample-specific TF-gene networks [47]. These networks are formed by highly expressed TFs and their target genes, in which interactions are inferred via co-expression and encoded as activating, repressive or non-interacting. Then, a linear regression is used to predict the expression profile from the ternary interaction matrix (genes *x* TFs), and the t-statistic of that regression represents the TFA. Instead of predicting gene expression, Chen et al. also tested to estimate the average methylation level of promoters.

The authors of *VIPER* extend the idea of TFA to any type of protein, allowing for the identification of indirect regulators of expression, such as signaling proteins [18]. *VIPER* works in a two-step fashion by first using an extension of the *ARACNe* [48] algorithm for network construction and then applying analytic rank-based enrichment analysis (aREA) to infer the activity of a protein. aREA checks for the enrichment of a regulon within genes that are differentially expressed. Each gene in a regulon is weighted based on the confidence of the regulator-gene interaction and its mode of action. To infer the mode and strength of action, Alvarez et al. modeled the Spearman's correlation coefficient density between each regulator and its target as a three-Gaussian mixture, which also allows each regulator-target pair to be represented by a continuous value.

NetProphet 3 predicts the probability of functional TF binding events in a gene's promoter by combining multiple weighted networks derived from expression data and sequence information with regularized gradient boosting [49]. Compared to its predecessor *NetProphet 2.0* [50], *NetProphet 3* was developed to be more flexible and allows users to incorporate any number and type of evidence scores for TF-gene interactions. However, its performance relies heavily on the usage of TF perturbation data. The authors showed that without the perturbation data, the model was not able to outperform even simpler regression-based approaches.

ISMARA is a webserver version of Balwierz et al.'s tool *MARA* [51], which aims to identify key regulators by predicting either gene expression or chromatin states across samples with a linear model using a Bayesian procedure [52]. TF information is provided via a collection of precalculated TFBS in promoters, and is used to find informative TF motifs, as well as to identify TF-gene interactions that explain the changes across samples. TF–TF interactions are also considered by looking at which motifs are found in TF promoters.

FindIT2 combines multiple tools into an R package, among which is the aforementioned *MARA* [53]. The TF-related functions include linking ChIP-seq peaks to genes, which is either based on closest distance or on a defined window, and allows one to quantify the correlation of peak accessibility with gene expression or promoter accessibility. *FindIT2* also calculates the regulatory potential [54] per gene, which summarizes surrounding peaks and, if applied on ChIP-seq data or TFBS predictions, can score TF target genes. Another functionality is looking for enrichment of TFs in peak subsets or among regulators of specific genes.

Another example of a tool compiling existent approaches is *decoupleR*, which contains eleven methods to estimate the activity of TFs or of any other biological factor, based on prior knowledge and omics measurements [55].

Zhang et al. developed *Taiji*, a method based on the personalized PageRank algorithm to rank TFs by their importance in a network [56]. The tool utilizes epigenome data to link active regulatory regions to their putative target genes by applying the method *EpiTensor* [57]. A cell type-specific GRN is then generated based on the TFBS predictions in the regulatory regions of each gene. The node weight of a TF in its GRN corresponds to the number of differentially expressed genes regulated by the TF, while the edge weight is proportional to a TF's expression level. Applying the PageRank algorithm on the network gives the overall importance of a TF. In an extension called *Taiji-reprogram* the method predicts the top TFs whose differential activity explains the transcriptional differences between two conditions [58].

TEPIC also uses TF motifs in regulatory regions of genes, but in a non-hit-based fashion. It calculates continuous binding site affinities of TFs per gene and uses them in an elastic net regression model to predict gene expression. *TEPIC* has various extensions and can integrate chromatin accessibility, chromatin footprints, chromatin interactions, or ChIP-seq data to find TFs predictive for gene expression within a sample, for differential expression, or time series data [155–158]. Recently, Hoffmann et al. published *TF-Prioritizer* [59], an automated pipeline which is based on *TEPIC*'s functionalities. It was designed to prioritize TFs explaining differential expression by using ChIP-seq, ATAC- seq, or DNase-seq data combined with RNA-seq data. If ATAC-seq or DNase-seq peaks are provided, footprints are called with *HINT-ATAC* [27] (see also Section 2.2). *TEPIC* then calculates TF affinity scores, followed by *DYNAMITE* [60] to employ a logistic regression model for each condition or time point, including empirical *p*-values based on a background distribution of scores.

2.1.1 | Network-based methods in single cell

The previously listed methods were developed for the use of bulk data and were hence limited to investigating tissues, cell lines, or FACS-sorted cells. However, the emergence of single-cell technologies has enabled the study of individual cells, resulting in the creation of many new computational tools, in particular for GRN inference (recently reviewed in Ref. [61]).

Behjati Ardakani et al. proposed *TRIANGULATE*, an extension of *TEPIC* to scRNA-seq where the TFAs of individual cells can be inferred using a tree-guided multitask regression model [62]. *TRIANGULATE* takes as input the gene expression measurements in single cells along with the per gene TFAs generated by *TEPIC*. The *TRIANGULATE* setup is designed in such a way that the single cells appearing in the same lineage tree—inferred from the single cell expression data—are penalized similarly. The coefficients of this multitask learning model are then used to deduce the relevance of TFs in regulating each individual cell.

Similarly, Teschendorff and Wang built a new version of *SEPIRA* [47], called *SCIRA* [63]. Regulons are still inferred from a collection of bulk RNA samples, but *SCIRA* allows one to successively estimate TFA in single cells.

VIPER also received a successor version, called *metaVIPER*, which aims to overcome the requirement for a large number of gene expression data sets from the same tissue. It tries to solve this problem by constructing tissue-specific networks from a set of heterogeneous samples and thus, allows its usage on single cell data [64]. The approach is based on the assumption that regulator–target interactions may be partially conserved even across distinct lineages. Hence, given a sufficient number of different tissue-specific networks, there is a high probability that a protein shares the same targets in a subset of the networks. Further, the algorithm assumes that only the context-specific regulons will show a significant enrichment score when comparing genes that are differentially expressed in the tissue of interest. The method *BITFAM* applies Bayesian factor analysis to infer TFA from scRNA-seq data by decomposing an expression matrix into a weight and TFA matrix [65]. While the TFA matrix represents the TFA in each individual cell, each column in the weight matrix represents the potential targets of a TF. To derive the posterior distributions of the matrices, the method leverages a collection of non-tissue-specific ChIP-seq data which is used to incorporate prior probabilities into the weight matrix. The learned TFA values can be further used for downstream analysis tasks such as cell clustering or trajectory inference.

SCENIC is an approach that infers GRNs from scRNA-seq data and characterizes regulons in a single cell as active or not [66]. To achieve its goal, the method first uses *GENIE3* [67] to find genes that are co-expressed with TFs across a large number of single cells. To minimize false positives, *SCENIC* then checks the putative genes of a regulon for motif enrichment of their regulator. Finally, *SCENIC*'s enrichment algorithm, called *AUCell*, classifies the regulons in each cell as active or not.

In its latest version, named *SCENIC+*, the method tries to leverage epigenome data by taking both scRNA-seq and scATAC-seq data as input [68]. Using topic modeling on co-accessible regions, they identify candidate enhancers in specific cell types and states. For each TF, *SCENIC+* then infers all its target genes and the *cis*-regulatory regions through which it exerts its effects by using Pearson correlation and gradient boosting machines. Based on the calculated GRN, *SCENIC+* can also perform in silico TF perturbations and identify the most influential TF for each cell state.

Another tool for constructing cell type-specific GRNs from scRNA-seq data is *Inferelator 3.0* [69]. It requires a prior network of TF–gene interactions to calculate TFA from the expression of a TF's target genes, and infers new GRNs via a selection of regularized regression models that estimate expression from the TFA. The prior network can be built from existing databases or with their tool *inferelator-prior* that predicts TFBS with PWMs in a gene's regulatory regions.

Similar to Ma and Brent's approach, *TIGER* is based on matrix factorization and incorporates a sign constraint on the regulatory network, as well as restricts the TFA matrix to non-negative values [70]. In contrast to Ma and Brent's method, *TIGER* can use scRNA-seq data and uses a Bayesian approach which relies on a literature-curated network to impose prior distributions on the variables.

2.2 | Genome occupancy-based approaches

As a second category of tools for TFA inference, we want to summarize those that do not require any mapping of TFs to genes. They quantify the genome-wide binding behavior of TFs and estimate their importance for the sample at hand or try to identify those responsible for changes between conditions. Usually, a TF's binding behavior is assessed by the chromatin accessibility at the binding sites. One class of such methods focuses on a specific shape of the accessibility profile, the so-called footprints. Footprint calling is often used to narrow down TFBS and can be performed on DNase-seq as well as ATAC-seq data. Both of these sequencing methods introduce their own biases which need to be corrected for, but the detectable footprints are largely shared [71]. To map footprints to TFs one can either first find TF motifs and look for footprint signatures around those, or perform a posterior TF motif search in already identified footprints' sequences.

One example of such a footprinting tool is *HINT*, which uses hidden Markov models to identify TF footprints by using strand-specific open chromatin signals with correction for protocol-specific biases [26, 27]. It was adapted to work both with DNase-seq, as well as ATAC-seq data. *HINT* employs position dependency models for the correction of cleavage bias, which was shown to be crucial for its performance. TFA is quantified by averaging the depth of a TF's footprints and the number of reads in its flanking regions, and can be differentially compared between samples.

Similarly, *TOBIAS* enables genome-wide investigation of TF binding dynamics via footprint calling from ATAC-seq and offers additional analyses and visualization tools [25]. Bias correction is done by calculating a dinucleotide weight matrix of the cleavage enzyme to then estimate the enzyme's expected influence and subtract it from the measured signal. Footprint scores are generated using a scoring function that considers the accessibility and depth of the local footprint. This score is then correlated with the presence of TFBS, and a threshold is set to distinguish between bound and unbound sites. Moreover, *TOBIAS* allows contrasting footprints across conditions, comparison of binding specificity between individual TFs, and TF network prediction.

BaGFoot also utilizes chromatin accessibility data (DNase-seq or ATAC-seq) for TF footprinting and aims to detect changes between conditions [72]. The approach focuses on the footprint depth, as well as the accessibility of the flanking region at all motif occurrences of a TF. Using these two metrics at all motif sites allows for measuring changes or TFs that do not show a measurable footprint pattern, meaning TFs where the footprint signature is too variable to confidently determine footprints.

Similarly, *diffTF* also aims to estimate changes in TFA between two conditions via accessibility changes at potential TFBS [73]. The algorithm scans the genome for TF binding motifs that overlap with a consensus peak set of all samples called on ATAC-seq data. For peaks containing multiple motifs of the same TF, the binding site with the highest read count across all samples is chosen as a representative. While controlling for GC content, log2 fold-changes for all peaks of a TF are calculated and compared to a background distribution. TFA is then represented as the mean difference to the background. If additional expression data is provided, *diffTF* classifies TFs into activators and repressors, based on the Pearson correlation between the RNA-seq counts of a TF and the ATAC-seq signal of all its putative binding sites.

2.2.1 | Occupancy-based methods in single cell

Tools examining genome-wide occupancy of TFs are also increasingly developed specifically for single cell data, or designed to work with bulk as well as single cell. For example, Schep et al. created a method to calculate accessibility deviations for peaks that share the same motif in single cell (or sparse) chromatin accessibility data [74]. Their method, *chromVar*, analyzes the gain or loss of accessibility of motifs within peaks by calculating a z-score of the number of fragments that map to a motif in a cell, minus the expected number of fragments based on all cells. The mean and standard deviation for the scaling procedure are based on a background peak set that matches the GC content and accessibility, thus controlling for technical biases introduced by PCR amplification or variable transposase tagmentation conditions.

Like previous methods, *chromVar* suffers from two major limitations: an open TF motif does not necessarily represent a binding event, and the same motifs can be shared by many different TFs. Argelaguet et al. attempt to tackle these problems by introducing a modified version of *chromVar*, called *chromVAR-Multiome* [75]. In this method, binding sites are based on an in silico binding score that incorporates information from single cell RNA expression. For a motif to be considered, the correlation coefficient between its accessibility and the gene expression of its corresponding TF must pass a threshold.

de Boer and Regev developed an R package, *BROCKMAN*, to unravel the dependencies between TF binding and chromatin accessibility or chromatin marks using gapped k-mer frequencies across multiple samples or single cells [76]. They first construct a k-mer *x* samples matrix that contains the frequency of a particular k-mer associated with an open chromatin region or chromatin mark measured in each sample. Next, using principal component analysis (PCA), they decompose this matrix into two other matrices: (1) k-mer *x* principal components (PCs)—indicating the contribution of k-mers to each PC and (2) PCs *x* samples—projection of samples into PCs, where the number of PCs is determined through a permutation test. In this manner, the k-mers represent a group of co-varying TFs that are identified through those recognizing multiple related k-mers. Finally, to infer the differential TFA, they examine the significant PCs associated with k-mers that were classified into “bound” or “unbound” for each TF. Applying a hypergeometric test helps assess the enriched or depleted status of a bound k-mer to a particular TF, enabling the identification of differential TFs between various conditions or cell types.de Boer and Regev also predict TF–TF interactions by identifying TFs that show covariation on the same PC.

scFAN uses deep learning on scATAC data and estimates TF binding in single cells [77]. The convolutional neural network is fist trained on bulk ATAC-seq and ChIP-seq data to then predict TF binding in open chromatin regions of individual cells using their continuous scATAC-seq profile as input. The ATAC signal of similar cells is aggregated to reduce sparsity. TFA per cell is then quantified by summarizing the predicted occurrence of each TF across all scATAC-peaks.

2.2.2 | Sequence-based deep learning models

There has been a tremendous advancement in deep learning models that predict epigenetic signals or gene expression from DNA sequence alone. While these models do not initially use TF information, they can successively be interrogated to find which motifs drive the predictions, or, be fed with artificial sequences. Their use in practice is hampered by the required computational resources, the cell type-specificity of the predictions, and that they have difficulties to predict the impact of the genomic environment [78]. Nevertheless, these methods offer complementary insights into the syntax of the regulatory code.

Hammelman and Gifford created such a deep learning approach and used it for the identification of cell state-specific TFs in chromatin accessibility data [79, 80]. Their method is part of the framework *DeepAccess* and is based on an ensemble of convolutional neural networks. Upon training with cell type-specific open chromatin regions and randomly sampled closed DNA sequences, *DeepAccess* predicts whether a sequence is accessible. For estimating the influence of TF motifs, the predicted accessibility of a sequence set with a TF motif is compared to the same sequences without the motif. Applying a signed-rank test statistic on the predicted difference gives the so-called expected pattern effect. This metric can also be calculated for the difference between conditions, which the authors term differential expected pattern effect. The flexibility of the approach allows scientists to investigate combinations of TF motifs and varying spacing between motifs.

Yuan and Kelley also model chromatin accessibility from sequence with convolutional neural networks, but in single cell data, specifically scATAC-seq [81]. Similar to Behjati Ardakani et al., their tool *scBasset* treats single cells as the tasks in their multitask setting. The final layer of their network represents a latent cell embedding which is then combined with a linear transformation matrix to predict accessibility. In order to derive the single cell TFA, perturbed DNA sequences—with or without the TF motif of interest—are given to the model. The predicted activity delivered by the output layer of *scBasset* allows the user to investigate the role of the TF in each single cell. Meaning, if the TF was positively involved in regulating a cell, the sequence with the motif is expected to return an increased accessibility.

Worth mentioning is also *BPNet* which does not predict accessibility, but the binding profile of specific TFs from CHIP-exo data [82]. Although it needs to be trained per TF, *BPNet* allows one to examine TF-specific motif syntax rules, specifically how the presence of other motifs and their spacing affects the predicted TF binding profile.

3 | APPLICATIONS

TFA inference methods have been utilized in diverse settings to advance our comprehension of fundamental biological mechanisms. Here, we showcase the application of TFA methods across various contexts, with a focus on examples from development and differentiation, cancer and aging. This is by all means not a complete list, but supposed to illustrate potential applications of TFA inference.

3.1 | Development and cell differentiation

TFs are essential regulators during development for sustaining cellular potency, as well as for establishing specific cell lineages. As an example application in this context, Kamimoto et al. used their in silico perturbation approach to predict TFs that are important for axial mesoderm development in zebrafish, followed by experimental validation [83]. Lefebvre et al. leveraged TFA inference to identify a set of TFs that promote the transition from naive B cells to centroblasts in the germinal center [84]. Similarly, Liu et al. used TFA analysis to identify TFs that promote the differentiation of T cells into effector and memory cells [85]. A specific application of TFs' differentiation potential is directed reprogramming. It is a process whereby a pluripotent or somatic cell is converted to another cell type by exogenously expressing a small set of TFs. Since reprogramming can skip intermediate differentiation steps and produce arbitrary cell types, this technique holds great promise in the field of regenerative medicine, as it allows for the repair of damaged tissues and also enables researchers to investigate primary cell types that are difficult to obtain, such as specific types of neuronal cells [86]. However, the success of reprogramming is currently limited to certain cell types, and even these transformations are often incomplete, only achieving partial characteristics of the target cell type. The process itself is highly time-consuming as the possible combinations to test are enormous. Hence, developing computational methods capable of accurately predicting effective TF sets is crucial for advancing this field. While many of the established approaches are built on the identification of TFs that show differential activity between two types of cells [80, 87–89], others are specifically tailored to infer new reprogramming strategies, like *Taiji-reprogram* [58]. Hammelman et al. provide a comparative benchmark on tools for ranking reprogramming factors [86]. In an applied example, Patel et al. measured transcriptomic and chromatin accessibility changes in the lifetime of murine spinal motor neurons and used *DeepAccess* to find TF candidates regulating different stages of neuronal maturation, to then recapitulate this maturation in cultured neurons [90].

3.2 | Cancer

Aberrant activity of TFs has been shown to contribute to cancer initiation, maintenance, progression, and drug resistance in various ways [91]. Among the most prominent examples of cancer-driving TFs is the oncoprotein *c-myc*, whose activation increases overall transcription elongation [92]. Many other classes of TFs have members that contribute to malignancies, such as forkhead box proteins [93] or the *ETS* family [94]. The change in TFA in a cancer setting can be caused by direct effects on the TF itself, but also by indirect effects like mutations in TFBSs or altered levels of cofactors and miRNAs. The complexity and multitude of different drivers make it particularly challenging to infer altered activity of a TF. Nonetheless, TFA analyses have been used to estimate the impact of somatic mutations [18], to find interactors of genes promoting tumorigenesis [95], to serve as prognostic markers in association with survival rates [96], and to predict drug response [97].

3.3 | Aging

Aging is the most significant risk factor for a wide range of diseases and is highly correlated with morbidity and mortality. Compared to their younger counterparts, aged cells exhibit changes at the transcriptional level, including a loss of cell type-specific profiles and dysregulation of developmental genes [98], as well as increased cell-to-cell variation in their expression profiles [99]. TFs play a central role in the mechanisms underlying these changes [100] and are known to contribute to the initiation and progression of age-related diseases [101]. Inference of TFA patterns can, therefore, provide valuable insights into the mechanisms of aging. Maity et al. used their *SCIRA* algorithm to analyze scRNA-seq data from a public murine aging atlas and identified TFs with differential activity levels in aging, which were linked to the dysregulation of the circadian rhythm. Further, they found TFs that could explain different macrophage subtype ratios observed in aging and potential contributors to leukemia [102]. In a similar fashion, Karakaslar et al. examined the effects of aging in peripheral blood leukocytes and splenic cells and compared the transcriptome and epigenome between young and old mice [103]. They applied footprinting analysis, including *HINT* [27], to describe TFs associated with increased inflammation upon aging.

4 | EXPERIMENTAL MEASUREMENT OF TFA VIA PERTURBATION

TF perturbation promises a more direct readout of TFA and often serves as a resource for validation data. Hence, we want to give an overview of the huge range of experimental studies and their approaches for analysis. The list of presented works here focuses more on larger scaled setups and is not exhaustive, but is supposed to illustrate the variety of methods and designs. The role of this type of data in TFA validation will be discussed later in a separate section.

A common procedure is the perturbation of a TF via knockout, knockdown or overexpression, followed by a readout of the caused changes—mostly by transcriptome measurements. Dixit et al. published Perturb-seq, which combines scRNA-seq with pooled CRISPR-based perturbation, and applied it in murine immune cells and human cell lines to knock out TFs and other regulators [104]. Their method uses lentiviral vectors to deliver the sgRNA together with an expressed guide barcode for identification. Their work comes with its own computational tool *MIMOSCA* to estimate the effect of sgRNAs on gene expression with a regularized linear model, also allowing to account for covariates like the number of transcripts in a cell or the cell state. Hackett et al. present an atlas of gene expression dynamics (*IDEA*: Induction Dynamics gene Expression Atlas), that provides data on the induction of more than 200 TFs via a synthetic promoter in yeast [105]. They measured over multiple time points with the aim to increase the detection of direct regulation with rapid changes in expression, as opposed to indirect effects supposedly taking place at later time points. It also enabled examination of the dynamics of expression changes. Another example comes from Alda-Catalinas et al., who used CRISPR activation in combination with single cell transcriptomics in mouse embryonic stem cells. Their goal was to find TFs whose inductions promote a cell state expected in zygotic genome activation [106]. Using *MOFA+* [107] allowed them to jointly analyze the expression of genes and of repeat elements to derive latent factors that explain variation across cells with different sgRNAs. Nakatake et al. induced hundreds of genes, including 481 TFs, in hESCs followed by transcriptomic readout after 48 h [108]. On top of measuring expression changes, they recorded microscopic images which enabled them to link induced genes to morphological changes. To identify TFs which drive differentiation into certain cell types, they correlated the perturbed transcriptome to public transcriptome data, assuming that a high similarity indicates a TF's capability to differentiate hESC into that cell type. Similarly, Joung et al. created a TF Atlas of expression profiles of hESCs overexpressing all annotated human TF isoforms (>3500) via a barcoded ORF library, coupled with scRNA-seq [109]. They found drastic differences in the differentiation potential between splice isoforms for many TFs. Other works do not focus on expression changes upon TF perturbation, like Rubin et al., who combined CRISPR interference with ATAC-seq into Perturb-ATAC. They observe changes in the accessibility of chromatin in single cells and in particular the accessibility at TFBS [110]. Another important resource informing on a TF's activity are cell viability screens, which frequently include loss-of-function of TFs, and can thus provide an estimate of the importance of a TF in a cell line [111, 112].

Beside methods interfering with a TF's gene itself, TFA can be assessed by using a reporter gene which holds a respective TFBS in its promoter. Massively parallel reporter assays (MPRAs) measure the regulatory activity of sequences of candidate regions like enhancers, and give information on TFA via the TFBS included in the sequence. They exists in a variety of designs, but have the downside of testing regions outside of their native chromatin context and do not allow to identify endogenous target genes of a TF [14, 113]. But since the reporter is usually not functional, MPRAs promise to reduce indirect effects of perturbation [114]. Abe and Abe aimed to measure endogenous TFA via a viral-vector-based TF reporter battery using a bipromoter containing a reference gene and a TFA reporter gene which holds the binding motif of the TF [115]. The reference gene was used to correct for transfection efficiency. They tested their constructs in human and murine cell cultures, as well as in vivo in the mouse brain. Additionally, they measured different environmental conditions and stimuli to visualize dynamic changes with their so-called TFA profile. Another example is work from Kreimer et al., who conducted lentiviral-based MPRAs during neural differentiation and tested selected TF motifs in regulatory sequences [114, 116]. The sequence with the motif is placed in front of a transcribed barcode, so that the ratio of barcode to the number of coding sequences can be interpreted as activity of the sequence.

5 | LIMITATIONS OF TFA VALIDATION

As of now, there is no way to directly measure the activity of a TF. Every quantification is only an estimate capturing a certain modality, like the protein level, mRNA level, or availability of binding sites. TF perturbation is the most direct readout, but converts cells into an artificial state, accompanied by various confounding mechanisms which are discussed later on. An aggravating factor is the limited availability of perturbation data. Thus, it is also not possible to directly evaluate the quality of a TFA estimate, which has led to a variety of indirect approaches trying to describe the plausibility of results or to compare different methods. Here, we want to discuss those validation practices, potential issues, and point to aspects of TF regulation that are underrepresented in computational tools of TFA inference. We defined TFA as a TF's regulatory impact on its target genes, which is often summarized for a cell state or condition, and, hence, we focus primarily on validating TF-gene relationships. These relationships serve as the basis for—or are the result of—tools we categorized as regulatory network-based tools (Section 2.1).

A common approach is to search for support of TF-gene interactions, which has sparked efforts in assembling databases like *HTRIdb* [117], *IntAct* [118], or *TRRUST* [119], which gather TF interactions via text mining, manual curation or integration of other databases and resources. Due to their nature, those databases are biased toward well-studied TFs and collect data with variable level of evidence [16, 120]. Others try to support the importance of their identified TFs and inferred target genes via GO enrichment [49, 121], eQTLs in binding sites [122], or via measured TF binding [49, 104, 122]. Concerning changes in TF binding behavior, condition-specific TF binding experiments can be used for evaluation and further substantiated by measurements of chromatin accessibility or other assays for regulatory activity. Frequently referred to as the gold standard, however, is TF perturbation data. Numerous publications validate their findings on public data or perform their own experiments [42, 72, 87–89, 119, 123–125]. Others assembled databases, like *KnockTF* [126]. Although those different approaches for validation are commonly used, there appears to be little agreement between them. In a benchmark study, Garcia-Alonso et al. gathered TF-gene interactions from four types of approaches (literature-based databases, ChIP-seq data, TFBS calling based on PWMs and inference from expression data) into a joint database called *DoRothEA*. They found the vast majority of TF-gene interactions to be supported by only one approach (96.3%) [120]. Even across literature-based databases there was little overlap. Benchmarked against three collected TF perturbation data sets, the different types of resources varied heavily in their accuracy.

A substantial discrepancy exists between TF ChIP-seq and TF perturbation data. A common assumption is that TF binding in the promoter is required for a functional TF-gene interaction, meaning that the TF is important for the gene's expression. It is the basis for many network-driven tools (Section 2.1). However, this notion that every binding event of a TF in a promoter leads to a transcriptional response was challenged by a TF perturbation screen in yeast, where only 3% of genes with a measured TFBS in their promoter were affected by the TF's knockout [127], and reproduced in more recent perturbation studies also in human and mice with varying but still small fractions of overlap [104, 105, 128, 129]. Conversely, the majority of responsive genes were not bound by the perturbed TF, posing a predicament for occupancy-based approaches (Section 2.2), since TF binding at sites of known regulatory importance appeared mostly ineffective. Multiple components have to be considered when it comes to the, seemingly, lack of regulatory function of TF binding. Aspects on the level of TFBS, the TFs themselves, as well as the target genes could influence the response to perturbation and contribute to this gap between bound and responsive genes. With regard to TFBS, low-affinity binding sites are frequently neglected, due to the difficulty in annotating them, although they can actually be informative and functional [15, 17, 130]. They could account for a portion of responsive genes where no strong binding sites were found and thus were overlooked for being regulated by a TF. Also rarely considered are more distal TFBS outside the promoter. On the other side of the scale, high-affinity binding sites and regions with more bound TFs appear to be more sensitive to perturbations [128]. Another layer of complexity is added by the redundancy of binding motifs. Compared to the number of TFs the amount of different DNA-binding domains is small, and binding specificity is additionally driven by mechanisms like combination of multiple DNA-binding domains, interaction with other proteins, DNA shape, epigenetic modifications at the target site, or compartmentalization [3, 14–17]. Despite these specificity mechanisms, TF binding and function can still be redundant, and thus, the regulatory importance of a TF might only become evident if any compensatory mechanisms are abolished [131]. For instance, Gitter et al. found a fourfold higher agreement between bound and responsive genes upon knockout when excluding TFs that had a redundant paralog [132]. On top, they could show an increase of responsive genes when a potentially compensating TF was removed in double knockout experiments. Others also found compensation for TFs which were co-expressed with other TFs, or shared functional annotation terms [133, 134]. Such buffering mechanisms of TFs point to an aspect of transcriptional regulation which is heavily neglected in computational approaches: combinatorial action of TFs. Most prominently represented by the formation of multimers, TFs interact and depend on each other, as well as on other cofactors [14, 15]. Although some tools allow for the quantification of TF co-occurrence [135, 136], most cannot estimate how a combination of TFs might affect regulation. Often mentioned in this context are the enhanceosome model, assuming synergistic cooperation, and the billboard model, describing additive cooperation [137]. Neither model appears to be universal, and experiments perturbing pairs and triplets of TFs indicate that reality is a mixture of both [104, 109, 110, 114].

On the level of target genes it is of relevance to the outcome of TF perturbation experiments, that certain features appear to make genes either more sensitive or insensitive to perturbations. Wu and Lai examined genes that did not respond to perturbation despite TF binding in their promoter in yeast, and found them to show distinct properties: low expression, low expression variation across experiments, no TATA box, having a nucleosome-free region directly upstream of the TSS, low number of bound TFs and binding sites, and short distance between binding sites and TSS [134]. Nakatake et al. observed in human that genes in regions with histone marks associated to heterochromatin responded only to very few TF perturbations, while genes with more active histone modifications responded broadly [108]. In their MPRA Kreimer et al. saw a correlation between the baseline expression from the unperturbed sequence and the effect of mutating the sequence [114]. Kang et al. gathered data from multiple TF perturbation experiments and built gradient boosted trees to predict which genes will change expression [138]. Gene expression and gene expression variation were the most informative features. Similarly, but in a broader scope and not focused on TF perturbations, Sigalova et al. found that expression variation was predictive for differential expression between conditions, independent of the experimental design [139]. Taken together, inherent properties of genes may explain non-responsiveness to perturbation of bound TFs, as well as responsiveness despite lack of TF binding.

Even across perturbation experiments it can be expected to detect different effects dependent on their design. Knockdown and overexpression studies can be variable in their efficiency of changing a TF's availability, while a full knockout consistently removes a TF. However, knockout experiments have been repeatedly found to elicit less profound changes than knockdown, as the loss-of-function was accompanied by compensatory transcription of related genes [140–143]. This compensation can also serve as explanation as to why many healthy humans exhibit several loss-of-function mutations. Mechanistically it is described to take place on a transcriptional level, independent of functional compensation on protein level. Different studies suggest the key player to be premature termination codons that are generated from the mutated transgene sequence [140–143]. For gain-of-function protocols it is also argued that they can cause more transcriptomic changes than loss-of-function [105, 108, 144]. Another aspect potentially affecting the detectable effects is the timepoint of measurement after perturbation, which differs heavily between studies. While some works aim to capture more direct effects and measure shortly after perturbation (e.g., 5 min [105]), others wait longer to focus more on differentiation effects (e.g., 7 days [109]).

All in all, TF perturbation experiments are still the most evident and insightful source of TFA validation. It should be kept in mind however, that experimental design, compensatory mechanisms, gene states, and the chromatin environment impact their outcome. Particular care should be taken, when interpreting non-responsiveness of genes despite TF binding as non-functional. In other words, if perturbing a TF does not change the expression of its bound target genes, it does not necessarily mean that it is not important for the genes' regulation in homeostasis, but that its role might be taken over by other TFs, or that other compensatory mechanisms mask its function.

6 | DISCUSSION

In this review, we presented an overview of available computational tools to estimate TFA, described examples for their usage, and discussed their validation and the concomitant shortcomings. Here, we want to further detail limitations of frequently used data and assumptions, underrepresented data, shortly summarize findings from benchmark studies, and give a perspective of how the field might develop in the future.

Plenty of methods rely on experimentally measured or predicted TF binding information, either hit-based or non-hit-based (e.g., via motif enrichment). As consequence, they are restricted to TFs where such data is obtainable. Current assays for annotating TFBSs like ChIP-seq, ChIP-exo, or CUT&RUN are limited to TFs where an antibody with high affinity is available. They require relatively large amounts of homogeneous cells, which makes it particularly challenging for tissues with a high diversity of cell types or cell states. In addition, only one TF can be measured at a time, hampering the acquisition of a complete TF binding annotation [19, 20]. Motif-based TFBS prediction, on the other hand, does not require data in the cell type of interest, but already defined motifs which were identified so far only for a fraction of all TFs. Using motifs on the DNA sequence alone comes with the downside of being prone to false-positives, which can be mitigated by the usage of chromatin accessibility data, but in turn requires an additional cell type-specific data modality [22]. Furthermore, TFs show specific binding patterns. This led to classification systems like the distinction between pioneers (can bind closed chromatin and reshape chromatin), settlers (bind majorly to their motifs in open chromatin), and migrants (bind only a fraction of their accessible motifs) [145, 146], but these classes are rarely considered for TFA inference methods. Redundancy of binding motifs of different TFs further hamper an accurate motif-based TFBS prediction.

A recurring assumption in models is to expect a high TF expression to indicate regulatory importance. This could not be confirmed by perturbation studies [108, 114], and neglects the discrepancy between RNA and protein levels and the impact of post-translational modifications [16, 147, 148]. Analogously, the majority of models assume a linear influence of TFA on gene regulation, which might be insufficient to describe biological complexity. Barely included in any model are specific characteristics of TFs, such as DNA-binding domains, other protein domains, their structure, or subcellular localization. Despite growing knowledge on such kind of information, TFs are treated as uniform features. On top, different isoforms of TFs are rarely considered, although they apparently differ in their regulatory action [109]. Another aspect is the cooperative action of TFs, which is inherently complicated to capture in a model. The majority of tools define TFs as independent or assume an additive effect. While sequence-based deep learning models start to give insights into the motif syntax of TFs, including how motif spacing and co-occurrence could affect a region's activity or a target gene's expression, their computational cost and the laborious investigation of the syntax rules were not yet transferred to more general TFA tools. The dependency on other non-TF factors is also not well understood and is lacking in models.

Chromatin compartmentalization and phase-separated condensates are another relevant factor in gene regulation, as they create microenvironments and transcriptional hubs with specific conditions. While TFs are thought to be important for the formation of such compartments, their function is also likely heavily affected by them, due to localized concentration of TFs and cofactors [15]. How exactly condensates form and how they are composed is subject of ongoing research, and thus, TFA inference still assumes a uniform availability of TFs across the genome.

Incorporation of more data modalities is a central challenge for TFA inference. Currently, tools focus on few data types—most prominently on transcriptome and chromatin accessibility—and thus, can only capture a small fraction of all mechanisms that affect gene regulation. DNA methylation, post-transcriptional and -translational modifications, protein levels and their stability and localization, hold valuable information, but are currently underrepresented, due to lacking availability and missing knowledge of usability.

Although TFA inference tools vary in the data they use and albeit the difficult validation (Section 5), there have been efforts to benchmark their performance, especially for methods quantifying or identifying TF–gene interactions. Although unweighted GRNs do not represent TFA by our definition, their comparison can still be insightful, since they form the basis for a large fraction of tools. Some works provide dedicated frameworks for benchmarking, such as *BEELINE* [149] or—not limited to TFA—*decoupleR* [55]. While the scope of tested data and acquisition of the ground truth between benchmarks differ, a common finding is that tools perform moderately at best and sometimes worse than random [55, 120, 149–153]. Variable efforts for parameter optimization could contribute to the modest performance. Some studies showed that accuracy can be increased by jointly integrating the output of multiple tools [55, 150]. Further, the intersection of highly ranked

TFs or predicted regulons across tools is often low. The low performance and similarity emphasize the necessity for standard benchmarks and a critical view on the generation of simulation data and the accompanied limitations and assumptions. It would be insightful to further investigate how general strategies and principles, for example, linear versus non-linear models, affect the performance. Nonetheless, the results of computational TFA tools are frequently backed up by findings from the literature, and top-ranked TFs are often in line with their proposed role in the condition at hand [43–45, 52, 83, 152]. Also, even if the highly ranked TFs contain false positives, it restricts the number of potential candidates and facilitates the prioritization TFs for experimental validation.

One of the key advancements in the field is the ongoing development and availability of single cell technologies, accompanied by tools that analyze such data [61]. While there are challenges regarding the sparsity and integration of multiple modalities, single cell resolution promises to overcome the inaccuracy of bulk data and to give insights into cell-specific mechanisms. It has the convenient advantage of providing a high number of samples for models to train on, given that the sparsity does not require a high level aggregation of individual cells. Another positive development is the increased feasibility and availability of large-scale TF perturbation data. While their interpretation and analysis is complicated by multiple aspects, as discussed in Section 5, they are still providing the most direct data for validation of TFA inference tools.

Sequence-based deep learning models also hold great potential for shaping the field, as they provide the possibility to identify TF motif syntax rules. However, such knowledge has yet to be transferred to more generalizable tools, that do not require high computational power or need to be trained on the sample at hand.

ACKNOWLEDGMENTS

The authors would like to thank Pau Badia-i-Mompel for providing valuable feedback for the manuscript. This work was supported by the Technical University Munich—Institute for Advanced Study, funded by the German Excellence Initiative. This work was supported in part by the Intramural Research Programs (IRPs) of the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This work has further been supported by the DZHK (German Centre for Cardiovascular Research, 81Z0200101), the Cardio-Pulmonary Institute (CPI) [EXC 2026] ID: 390649896, the DFG SFB (TRR 267) Noncoding RNAs in the cardiovascular system (Z03, project ID 403584255), and the DFG SFB1531 (S03, project number 456687919). This project was partly funded by the European Union under grant agreement no. 101057619. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. This work was also partly supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract no. 22.00115.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

This review does not make use of any public or new data.

ORCID

Quirin Manz https://orcid.org/0000-0002-5706-2718
Markus Hoffmann https://orcid.org/0000-0002-1920-288X
Marcel H. Schulz https://orcid.org/0000-0002-1252-3656
Markus List https://orcid.org/0000-0002-0941-4168

REFERENCES

1. Schmitz, R. J., Grotewold, E., & Stam, M. (2022). Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. *Plant Cell*, *34*(2), 718–741.
2. Vijayabaskar, M. S., Goode, D. K., Obier, N., Lichtinger, M., Emmett, A. M. L., Abidin, F. N. Z, Shar, N., Hannah, R., Assi, S. A., Lie-A-Ling, M., Gottgens, B., Lacaud, G., Kouskoff, V., Bonifer, C., & Westhead, D. R. (2019). Identification of gene specific cis-regulatory elements during differentiation of mouse embryonic stem cells: An integrative approach using high-throughput datasets. *PLoS Computational Biology*, *15*(11), e1007337.
3. Gonzalez, D. H. (2016). Introduction to transcription factor structure and function. In *Plant transcription factors* (pp. 3–11). Elsevier. https://linkinghub.elsevier.com/retrieve/pii/B9780128008546000014
4. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The human transcription factors. *Cell*, *172*(4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029
5. Pabo, C. O., & Sauer, R. T. (1992). Transcription factors: Structural families and principles of DNA recognition. *Annual Review of Biochemistry*, *61*, 1053–1095.
6. Huilgol, D., Venkataramani, P., Nandi, S., & Bhattacharjee, S. (2019). Transcription factors that govern development and disease: An Achilles heel in cancer. *Genes*, *10*(10), 794.
7. Staal, F. J. T., Weerkamp, F., Langerak, A. W., Hendriks, R. W., & Clevers, H. C. (2001). Transcriptional control of t lymphocyte differentiation. *Stem Cells*, *19*(3), 165–179.
8. Sukumari Nath, V., Kumar Mishra, A., Kumar, A., Matoušek, J., & Jakše, J. (2019). Revisiting the role of transcription factors in coordinating the defense response against citrus bark cracking viroid infection in commercial hop (humulus lupulus l.). *Viruses*, *11*(5), 419.
9. Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., & Young, R. A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, *122*(6), 947–956. https://doi.org/10.1016/j.cell.2005.08.020
10. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., & Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419. https://doi.org/10.1126/science.1260419
11. Andersson, E. I., Tanahashi, T., Sekiguchi, N., Gasparini, V. R., Bortoluzzi, S., Kawakami, T., Matsuda, K., Mitsui, T., Eldfors, S., Bortoluzzi, S., Coppe, A., Binatti, A., Lagström, S., Ellonen, P., Fukushima, N., Nishina, S., Senoo, N., Sakai, H., Nakazawa, H., … Ishida, F. (2016). High incidence of activating STAT5B mutations in CD4-positive t-cell large granular lymphocyte leukemia. *Blood*, *128*(20), 2465–2468.

12. Hwa, V. (2016). STAT5B deficiency: Impacts on human growth and immunity. *Growth Hormone & IGF Research*, *28*, 16–20.
13. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., & Luscombe, N. M. (2009). A census of human transcription factors: Function, expression and evolution. *Nature Reviews Genetics*, *10*(4), 252–263.
14. Inukai, S., Kock, K. H., & Bulyk, M. L. (2017). Transcription factor–DNA binding: Beyond binding site motifs. *Current Opinion in Genetics & Development*, *43*, 110–119. https://doi.org/10.1016/j.gde.2017.02.007
15. Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J., & Mann, R. S. (2019). Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Annual Review of Cell and Developmental Biology*, *35*(1), 357–379. https://doi.org/10.1146/annurev-cellbio-100617-062719 pMID: 31283382
16. Weidemüller, P., Kholmatov, M., Petsalaki, E., & Zaugg, J. B. (2021). Transcription factors: Bridge between cell signaling and gene regulation. *Proteomics*, *21*(23-24), 2000034. https://doi.org/10.1002/pmic.202000034
17. Zeitlinger, J. (2020). Seven myths of how transcription factors read the cis-regulatory code. *Current Opinion in Systems Biology*, *23*, 22–31. https://doi.org/10.1016/j.coisb.2020.08.002
18. Alvarez, M. J., Shen, Y., Giorgi, F. M., Lachmann, A., Ding, B. B., Ye, B. H., & Califano, A. (2016). Network-based inference of protein activity helps functionalize the genetic landscape of cancer. *Nature Genetics*, *48*(8), 838–847. https://doi.org/10.1038/ng.3593
19. Furey, T. S. (2012). ChIP–seq and beyond: New and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, *13*(12), 840–852. https://doi.org/10.1038/nrg3306
20. Park, P. J. (2009). ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, *10*(10), 669–680.
21. Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., & Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, *21*(3), 447–455. https://doi.org/10.1101/gr.112623.110
22. Consortium, T. E. P. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816. https://doi.org/10.1038/nature05874
23. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218.
24. Song, L., & Crawford, G. E. (2010). DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, *2010*(2), pdb.prot5384.
25. Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., Fust, A., Preussner, J., Kuenne, C., Braun, T., Kim, J., & Looso, M. (2020). ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nature Communications*, *11*(1), 4267. https://doi.org/10.1038/s41467-020-18035-1
26. Gusmao, E. G., Allhoff, M., Zenke, M., & Costa, I. G. (2016). Analysis of computational footprinting methods for DNase sequencing experiments. *Nature Methods*, *13*(4), 303–309. https://doi.org/10.1038/nmeth.3772
27. Li, Z., Schulz, M. H., Look, T., Begemann, M., Zenke, M., & Costa, I. G. (2019). Identification of transcription factor binding sites using ATAC-seq. *Genome Biology*, *20*(1), 45. https://doi.org/10.1186/s13059-019-1642-2
28. Filtz, T. M., Vogel, W. K., & Leid, M. (2014). Regulation of transcription factor activity by interconnected post-translational modifications. *Trends in Pharmacological Sciences*, *35*(2), 76–85. https://doi.org/10.1016/j.tips.2013.11.005
29. Dujon, B. (1996). The yeast genome project: What did we learn? *Trends in Genetics*, *12*(7), 263–270. https://doi.org/10.1016/0168-9525(96)10027-5
30. Bussemaker, H. J., Li, H., & Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, *27*(2), 167–171. https://doi.org/10.1038/84792
31. Conlon, E. M., Liu, X. S, Lieb, J. D., & Liu, J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(6), 3339–3344. https://doi.org/10.1073/pnas.0630591100
32. Keleş, S., Van Der Laan, M., & Eisen, M. B. (2002). Identification of regulatory elements using a feature selection method. *Bioinformatics*, *18*(9), 1167–1175. https://doi.org/10.1093/bioinformatics/18.9.1167
33. Wang, W., Cherry, J. M, Botstein, D., & Li, H. (2002). A systematic approach to reconstructing transcription networks in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(26), 16893–16898. https://doi.org/10.1073/pnas.252638199
34. Gao, F., Foat, B. C., & Bussemaker, H. J. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, *5*(1), 31. https://doi.org/10.1186/1471-2105-5-31
35. Liao, J. C., Boscolo, R., Yang, Y. L., Tran, L. M., Sabatti, C., & Roychowdhury, V. P. (2003). Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(26), 15522–15527. https://doi.org/10.1073/pnas.2136632100
36. Chang, C., Ding, Z., Hung, Y. S., & Fung, P. C. W. (2008). Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data. *Bioinformatics*, *24*(11), 1349–1358. https://doi.org/10.1093/bioinformatics/btn131
37. Tran, L. M., Brynildsen, M. P., Kao, K. C., Suen, J. K., & Liao, J C. (2005). gNCA: A framework for determining transcription factor activity based on transcriptome: Identifiability and numerical implementation. *Metabolic Engineering*, *7*(2), 128–141. https://doi.org/10.1016/j.ymben.2004.12.001
38. Noor, A., Ahmad, A., Serpedin, E., Nounou, M., & Nounou, H. (2013). ROBNCA: Robust network component analysis for recovering transcription factor activities. *Bioinformatics*, *29*(19), 2410–2418. https://doi.org/10.1093/bioinformatics/btt433
39. Boscolo, R., Sabatti, C., Liao, J. C., & Roychowdhury, V. P. (2005). A generalized framework for network component analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *2*(4), 289–301. https://doi.org/10.1109/TCBB.2005.47
40. Noor, A., Ahmad, A., & Serpedin, E. (2018). SparseNCA: Sparse network component analysis for recovering transcription factor activities with incomplete prior information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *15*(2), 387–395. https://doi.org/10.1109/TCBB.2015.2495224
41. Shi, Q., Zhang, C., Guo, W., Zeng, T., Lu, L., Jiang, Z., Wang, Z., Liu, J., & Chen, L. (2017). Local network component analysis for quantifying transcription factor activities. *Methods (San Diego, Calif.)*, *124*, 25–35. https://doi.org/10.1016/j.ymeth.2017.06.018
42. Ma, C. Z., & Brent, M. R. (2020). Inferring TF activities and activity regulators from gene expression data with constraints from TF perturbation data. *Bioinformatics*, *37*(9), 1234–1245. https://doi.org/10.1093/bioinformatics/btaa947
43. Li, Y., Liang, M., & Zhang, Z. (2014). Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Computational Biology*, *10*(10), e1003908. https://doi.org/10.1371/journal.pcbi.1003908
44. Jiang, P., Freedman, M. L., Liu, J. S., & Liu, X. S. (2015). Inference of transcriptional regulation in cancers. *Proceedings of the National*

Academy of Sciences of the United States of America, *112*(25), 7731–7736. https://doi.org/10.1073/pnas.1424272112

45. Fröhlich, H. (2015). biRte: Bayesian inference of context-specific regulator activities and transcriptional networks. *Bioinformatics*, *31*(20), 3290–3298. https://doi.org/10.1093/bioinformatics/btv379
46. Aguet, F., Anand, S., Ardlie, K. G., Gabriel, S., Getz, G. A., Graubert, A., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., Li, X., Macarthur, D. G., Meier, S. R., Nedzel, J. L., Nguyen, D. T., Segrè, A. V., Todres, E., Balliu, B., Barbeira, A. N., & Volpi, S. (2020). The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776
47. Chen, Y., Widschwendter, M., & Teschendorff, A E. (2017). Systems-epigenomics inference of transcription factor activity implicates aryl-hydrocarbon-receptor inactivation as a key event in lung cancer development. *Genome Biology*, *18*(1), 236. https://doi.org/10.1186/s13059-017-1366-0
48. Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, *7*(1), S7. https://doi.org/10.1186/1471-2105-7-S1-S7
49. Abid, D., & Brent, M. R. (2023). NetProphet 3: A machine learning framework for transcription factor network mapping and multi-omics integration. *Bioinformatics*, *39*(2), btad038. https://doi.org/10.1093/bioinformatics/btad038
50. Kang, Y., Liow, H. H., Maier, E. J., & Brent, M. R. (2018). NetProphet 2.0: Mapping transcription factor networks by exploiting scalable data resources. *Bioinformatics*, *34*(2), 249–257. https://doi.org/10.1093/bioinformatics/btx563
51. The FANTOM Consortium & Riken Omics Science Center. (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genetics*, *41*(5), 553–562. https://doi.org/10.1038/ng.375
52. Balwierz, P. J., Pachkov, M., Arnold, P., Gruber, A. J., Zavolan, M., & Van Nimwegen, E. (2014). Ismara: Automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Research*, *24*(5), 869–884. https://doi.org/10.1101/gr.169508.113
53. Shang, G. D., Xu, Z. G., Wan, M.-C., Wang, F.-X., & Wang, J. W. (2022). FindIT2: An R/Bioconductor package to identify influential transcription factor and targets based on multi-omics data. *BMC Genomics*, *23*(S1), 272. https://doi.org/10.1186/s12864-022-08506-8
54. Tang, Q., Chen, Y., Meyer, C., Geistlinger, T., Lupien, M., Wang, Q., Liu, T., Zhang, Y., Brown, M., & Liu, X. S. (2011). A comprehensive view of nuclear receptor cancer cistromes. *Cancer Research*, *71*(22), 6940–6947. https://doi.org/10.1158/0008-5472.CAN-11-2091
55. Badia-I-Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C. H., Ramirez Flores, R. O., & Saez-Rodriguez, J. (2022). decoupleR: Ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, *2*(1), vbac016. https://doi.org/10.1093/bioadv/vbac016
56. Zhang, K., Wang, M., Zhao, Y., & Wang, W. (2019). Taiji: System-level identification of key transcription factors reveals transcriptional waves in mouse embryonic development. *Science Advances*, *5*(3), eaav3262. https://doi.org/10.1126/sciadv.aav3262
57. Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J. W., Ding, B., Li, N., Zheng, L., & Wang, W. (2016). Constructing 3d interaction maps from 1d epigenomes. *Nature Communications*, *7*(1), 10812. https://doi.org/10.1038/ncomms10812
58. Wang, J., Liu, C., Chen, Y., & Wang, W. (2021). Taiji-reprogram: A framework to uncover cell-type specific regulators and predict cellular reprogramming cocktails. *NAR Genomics and Bioinformatics*, *3*(4), lqab100. https://doi.org/10.1093/nargab/lqab100
59. Hoffmann, M., Trummer, N., Schwartz, L., Jankowski, J., Lee, H. K., Willruth, L. L., Lazareva, O., Yuan, K., Baumgarten, N., Schmidt, F., Baumbach, J., Schulz, M. H., Blumenthal, D. B., Hennighausen, L., & List, M. (2023). TF-Prioritizer: A Java pipeline to prioritize condition-specific transcription factors. *GigaScience*, *12*, giad026. https://doi.org/10.1093/gigascience/giad026
60. Durek, P., Nordström, K., Gasparoni, G., Salhab, A., Kressler, C., de Almeida, M., Bassler, K., Ulas, T., Schmidt, F., Xiong, J., Glažar, P., Klironomos, F., Sinha, A., Kinkley, S., Yang, X., Arrigoni, L., Amirabad, A. D., Ardakani, F. B., Feuerbach, L., … Polansky, J. K. (2016). Epigenomic profiling of human cd4 < sup >+ < /sup >t cells supports a linear differentiation model and highlights molecular regulators of memory development. *Immunity*, *45*(5), 1148–1161. https://doi.org/10.1016/j.immuni.2016.10.022
61. Badia-I-Mompel, P., Wessels, L., Müller-Dott, S., Trimbour, R., Ramirez Flores, R. O., Argelaguet, R., & Saez-Rodriguez, J. (2023). Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-023-00618-5
62. Behjati Ardakani, F., Kattler, K., Heinen, T., Schmidt, F., Feuerborn, D., Gasparoni, G., Lepikhov, K., Nell, P., Hengstler, J., Walter, J., & Schulz, M. H. (2020). Prediction of single-cell gene expression for transcription factor analysis. *GigaScience*, *9*(11), giaa113. https://doi.org/10.1093/gigascience/giaa113
63. Teschendorff, A. E., & Wang, N. (2020). Improved detection of tumor suppressor events in single-cell RNA-Seq data. *npj Genomic Medicine*, *5*(1), 1–14. https://doi.org/10.1038/s41525-020-00151-y
64. Ding, H., Douglass, E. F., Sonabend, A. M., Mela, A., Bose, S., Gonzalez, C., Canoll, P. D., Sims, P. A., Alvarez, M. J., & Califano, A. (2018). Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nature Communications*, *9*(1), 1471. https://doi.org/10.1038/s41467-018-03843-3
65. Gao, S., Dai, Y., & Rehman, J. (2021). A Bayesian inference transcription factor activity model for the analysis of single-cell transcriptomes. *Genome Research*. https://doi.org/10.1101/gr.265595.120
66. Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J. C., Geurts, P., Aerts, J., Van Den Oord, J., Atak, Z. K., Wouters, J., & Aerts, S. (2017). SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods*, *14*(11), 1083–1086. https://doi.org/10.1038/nmeth.4463
67. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, *5*(9), e12776. https://doi.org/10.1371/journal.pone.0012776
68. Bravo González-Blas, C., De Winter, S., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V., Poovathingal, S., Wouters, J., Aibar, S., & Aerts, S. (2023). SCENIC+: Single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods*, *20*, 1355–1367. https://doi.org/10.1038/s41592-023-01938-4
69. Skok Gibbs, C., Jackson, C. A., Saldi, G. A., Tjärnberg, A., Shah, A., Watters, A., De Veaux, N., Tchourine, K., Yi, R., Hamamsy, T., Castro, D. M., Carriero, N., Gorissen, B. L., Gresham, D., Miraldi, E. R., & Bonneau, R. (2022). High-performance single-cell gene regulatory network inference at scale: The Inferelator 3.0. *Bioinformatics*, *38*(9), 2519–2528. https://doi.org/10.1093/bioinformatics/btac117
70. Chen, C., & Padi, M. (2022). Joint inference of transcription factor activity and context-specific regulatory networks. https://www.biorxiv.org/content/10.1101/2022.12.12.520141v1
71. Karabacak Calviello, A., Hirsekorn, A., Wurmus, R., Yusuf, D., & Ohler, U. (2019). Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome Biology*, *20*(1), 42. https://doi.org/10.1186/s13059-019-1654-y
72. Baek, S., Goldstein, I., & Hager, G. L. (2017). Bivariate genomic footprinting detects changes in transcription factor activity. *Cell Reports*, *19*(8), 1710–1722. https://doi.org/10.1016/j.celrep.2017.05.003

73. Berest, I., Arnold, C., Reyes-Palomares, A., Palla, G., Rasmussen, K. D., Giles, H., Bruch, P. M., Huber, W., Dietrich, S., Helin, K., & Zaugg, J B. (2019). Quantification of differential transcription factor activity and multiomics-based classification into activators and repressors: diffTF. *Cell Reports*, *29*(10), 3147–3159.e12. https://doi.org/10.1016/j.celrep.2019.10.106

74. Schep, A. N., Wu, B., Buenrostro, J. D., & Greenleaf, W. J. (2017). chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods*, *14*(10), 975–978. https://linkinghub.elsevier.com/retrieve/pii/S0092867418315149

75. Argelaguet, R., Lohoff, T., Li, J. G., Nakhuda, A., Drage, D., Krueger, F., Velten, L., Clark, S. J., & Reik, W. (2022). Decoding gene regulation in the mouse embryo using single-cell multi-omics. https://www.biorxiv.org/content/10.1101/2022.06.15.496239v2

76. De Boer, C. G., & Regev, A. (2018). BROCKMAN: Deciphering variance in epigenomic regulators by k-mer factorization. *BMC Bioinformatics*, *19*(1), 253. https://doi.org/10.1186/s12859-018-2255-6

77. Fu, L., Zhang, L., Dollinger, E., Peng, Q., Nie, Q., & Xie, X. (2020). Predicting transcription factor binding in single cells through deep learning. *Science Advances*, *6*(51), eaba9031. https://doi.org/10.1126/sciadv.aba9031

78. Karollus, A., Mauermeier, T., & Gagneur, J. (2023). Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, *24*(1), 56. https://doi.org/10.1186/s13059-023-02899-9

79. Hammelman, J., & Gifford, D. K. (2021). Discovering differential genome sequence activity with interpretable and efficient deep learning. *PLoS Computational Biology*, *17*(8), e1009282. https://doi.org/10.1371/journal.pcbi.1009282

80. Hammelman, J., Krismer, K., Banerjee, B., Gifford, D. K., & Sherwood, R. I. (2020). Identification of determinants of differential chromatin accessibility through a massively parallel genome-integrated reporter assay. *Genome Research*, *30*(10), 1468–1480. https://doi.org/10.1101/gr.263228.120

81. Yuan, H., & Kelley, D. R. (2022). scBasset: Sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nature Methods*, *19*(9), 1088–1096. https://doi.org/10.1038/s41592-022-01562-8

82. Avsec, U., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., & Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, *53*(3), 354–366. https://doi.org/10.1038/s41588-021-00782-6

83. Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., & Morris, S. A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, *614*(7949), 742–751. https://doi.org/10.1038/s41586-022-05688-9

84. Lefebvre, C., Rajbhandari, P., Alvarez, M. J., Bandaru, P., Lim, W. K., Sato, M., Wang, K., Sumazin, P., Kustagi, M., Bisikirska, B. C., Basso, K., Beltrao, P., Krogan, N., Gautier, J., Dalla-Favera, R., & Califano, A. (2010). A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology*, *6*(1), 377. https://doi.org/10.1038/msb.2010.31

85. Liu, C., Omilusik, K., Toma, C., Kurd, N. S., Chang, J. T., Goldrath, A. W., & Wang, W. (2022). Systems-level identification of key transcription factors in immune cell specification. *PLOS Computational Biology*, *18*(9), e1010116. https://doi.org/10.1371/journal.pcbi.1010116

86. Hammelman, J., Patel, T., Closser, M., Wichterle, H., & Gifford, D. (2022). Ranking reprogramming factors for cell differentiation. *Nature Methods*, *19*(7), 812–822. https://doi.org/10.1038/s41592-022-01522-2

87. D'alessio, A. C., Fan, Z. P., Wert, K. J., Baranov, P., Cohen, M. A., Saini, J. S., Cohick, E., Charniga, C., Dadon, D., Hannett, N. M., Young, M. J., Temple, S., Jaenisch, R., Lee, T. I., & Young, R. A. (2015). A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Reports*, *5*(5), 763–775. https://doi.org/10.1016/j.stemcr.2015.09.016

88. Jung, S., Appleton, E., Ali, M., Church, G. M., & Del Sol, A. (2021). A computer-guided design tool to increase the efficiency of cellular conversions. *Nature Communications*, *12*(1), 1659. https://doi.org/10.1038/s41467-021-21801-4

89. Rackham, O. J. L., Firas, J., Fang, H., Oates, M. E., Holmes, M. L., Knaupp, A. S., Suzuki, H., Nefzger, C. M., Daub, C. O., Shin, J. W., Petretto, E., Forrest, A. R. R., Hayashizaki, Y., Polo, J. M., & Gough, J. (2016). A predictive computational framework for direct reprogramming between human cell types. *Nature Genetics*, *48*(3), 331–335. https://doi.org/10.1038/ng.3487

90. Patel, T., Hammelman, J., Aziz, S., Jang, S., Closser, M., Michaels, T. L., Blum, J. A., Gifford, D. K., & Wichterle, H. (2022). Transcriptional dynamics of murine motor neuron maturation in vivo and in vitro. *Nature Communications*, *13*(1), 5427. https://doi.org/10.1038/s41467-022-33022-4

91. Bushweller, J. H. (2019). Targeting transcription factors in cancer – from undruggable to reality. *Nature Reviews Cancer*, *19*(11), 611–624. https://doi.org/10.1038/s41568-019-0196-7

92. Dang, C. V. (2012). MYC on the path to cancer. *Cell*, *149*(1), 22–35. https://doi.org/10.1016/j.cell.2012.03.003

93. Lam, E. W.-F., Brosens, J. J., Gomes, A. R., & Koo, C. Y. (2013). Forkhead box proteins: Tuning forks for transcriptional harmony. *Nature Reviews Cancer*, *13*(7), 482–495. https://doi.org/10.1038/nrc3539

94. Sizemore, G. M., Pitarresi, J. R., Balakrishnan, S., & Ostrowski, M. C. (2017). The ETS family of oncogenic transcription factors in solid tumours. *Nature Reviews Cancer*, *17*(6), 337–351. https://doi.org/10.1038/nrc.2017.20

95. Chudnovsky, Y., Kim, D., Zheng, S., Whyte, W. A., Bansal, M., Bray, M. A., Gopal, S., Theisen, M. A., Bilodeau, S., Thiru, P., Muffat, J., Yilmaz, O. H., Mitalipova, M., Woolard, K., Lee, J., Nishimura, R., Sakata, N., Fine, H. A., Carpenter, A. E., … Chheda, M. G. (2014). ZFHX4 Interacts with the NuRD core member CHD4 and regulates the glioblastoma tumor-initiating cell state. *Cell Reports*, *6*(2), 313–324. https://doi.org/10.1016/j.celrep.2013.12.032

96. Falco, M. M., Bleda, M., Carbonell-Caballero, J., & Dopazo, J. (2016). The pan-cancer pathological regulatory landscape. *Scientific Reports*, *6*, 39709. https://doi.org/10.1038/srep39709

97. Garcia-Alonso, L., Iorio, F., Matchan, A., Fonseca, N., Jaaks, P., Peat, G., Pignatelli, M., Falcone, F., Benes, C. H., Dunham, I., Bignell, G., McDade, S. S., Garnett, M. J., & Saez-Rodriguez, J. (2018). Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer Research*, *78*(3), 769–780. https://doi.org/10.1158/0008-5472.CAN-17-1679

98. Bartz, J., Jung, H., Wasiluk, K., Zhang, L., & Dong, X. (2023). Progress in discovering transcriptional noise in aging. *International Journal of Molecular Sciences*, *24*(4), 3701. https://doi.org/10.3390/ijms24043701

99. Mendenhall, A. R., Martin, G. M., Kaeberlein, M., & Anderson, R. M. (2021). Cell-to-cell variation in gene expression and the aging process. *GeroScience*, *43*(1), 181–196. https://doi.org/10.1007/s11357-021-00339-9

100. Zhou, X., Sen, I., Lin, X. X., & Riedel, C. G. (2018). Regulation of age-related decline by tran- scription factors and their crosstalk with the epigenome. *Current Genomics*, *19*(6), 464–482. https://doi.org/10.2174/1389202919666180503125850

101. Du, S., & Zheng, H. (2021). Role of FoxO transcription factors in aging and age-related metabolic and neurodegenerative diseases. *Cell & Bioscience*, *11*(1), 188. https://doi.org/10.1186/s13578-021-00700-7

102. Maity, A. K., Hu, X., Zhu, T., & Teschendorff, A. E. (2022). Inference of age-associated transcription factor regulatory activity changes in single cells. *Nature Aging*, *2*(6), 548–561. https://doi.org/10.1038/s43587-022-00233-9

103. Karakaslar, E. O., Katiyar, N., Hasham, M., Youn, A., Sharma, S., Chung, C.-H., Marches, R., Korstanje, R., Banchereau, J., & Ucar, D. (2023). Transcriptional activation of Jun and Fos members of the AP-1 complex is a conserved signature of immune aging that contributes to inflammaging. *Aging Cell*. https://doi.org/10.1111/acel.13792

104. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled penetic screens. *Cell*, *167*(7), 1853–1866.e17. https://doi.org/10.1016/j.cell.2016.11.038

105. Hackett, S. R., Baltz, E. A., Coram, M., Wranik, B. J., Kim, G., Baker, A., Fan, M., Hendrickson, D. G., Berndl, M., & Mcisaac, R. S. (2020). Learning causal networks using inducible transcription factors and transcriptome-wide time series. *Molecular Systems Biology*, *16*(3). 10.15252/msb.20199174

106. Alda-Catalinas, C., Bredikhin, D., Hernando-Herraez, I., Santos, F., Kubinyecz, O., Eckersley-Maslin, M. A., Stegle, O., & Reik, W. (2020). A single-cell transcriptomics CRISPR-activation screen identifies epigenetic regulators of the zygotic genome activation program. *Cell Systems*, *11*(1), 25–41.e9. https://doi.org/10.1016/j.cels.2020.06.004

107. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: A statis- tical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, *21*(1), 111. https://doi.org/10.1186/s13059-020-02015-1

108. Nakatake, Y., Ko, S. B. H., Sharov, A. A., Wakabayashi, S., Murakami, M., Sakota, M., Chikazawa, N., Ookura, C., Sato, S., Ito, N., Ishikawa-Hirayama, M., Mak, S. S., Jakt, L. M., Ueno, T., Hiratsuka, K., Matsushita, M., Goparaju, S. K., Akiyama, T., Ishiguro, K. I., … Ko, M. S. H. (2020). Generation and profiling of 2,135 human ESC lines for the systematic analyses of cell states perturbed by inducing single transcription factors. *Cell Reports*, *31*(7), 107655. https://doi.org/10.1016/j.celrep.2020.107655

109. Joung, J., Ma, S., Tay, T., Geiger-Schuller, K. R., Kirchgatterer, P. C., Verdine, V. K., Guo, B., Arias-Garcia, M. A., Allen, W. E., Singh, A., Kuksenko, O., Abudayyeh, O. O., Gootenberg, J. S., Fu, Z., Macrae, R. K., Buenrostro, J. D., Regev, A., & Zhang, F. (2023). A transcription factor atlas of directed differentiation. *Cell*, *186*(1), 209–229.e26. https://doi.org/10.1016/j.cell.2022.11.026

110. Rubin, A. J., Parker, K. R., Satpathy, A. T., Qi, Y., Wu, B., Ong, A. J., Mumbach, M. R., Ji, A. L., Kim, D. S., Cho, S. W., Zarnegar, B. J., Greenleaf, W. J., Chang, H. Y., & Khavari, P. A. (2019). Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell*, *176*(1-2), 361–376.e17. https://doi.org/10.1016/j.cell.2018.11.022

111. Mcdonald, E. R, De Weck, A., Schlabach, M. R., Billy, E., Mavrakis, K. J., Hoffman, G. R., Belur, D., Castelletti, D., Frias, E., Gampa, K., Golji, J., Kao, I., Li, L., Megel, P., Perkins, T. A., Ramadan, N., Ruddy, D. A., Silver, S. J., Sovath, S., … Sellers, W. R. (2017). Project DRIVE: A compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep RNAi screening. *Cell*, *170*(3), 577–592.e10. https://doi.org/10.1016/j.cell.2017.07.005

112. Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., Meyers, R. M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W. F. J., Howell, S., Merkel, E., … Hahn, W. C. (2017). Defining a cancer dependency map. *Cell*, *170*(3), 564–576.e16. https://doi.org/10.1016/j.cell.2017.06.010

113. Klein, J. C., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., & Shendure, J. (2020). A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nature Methods*, *17*(11), 1083–1091. https://doi.org/10.1038/s41592-020-0965-y

114. Kreimer, A., Ashuach, T., Inoue, F., Khodaverdian, A., Deng, C., Yosef, N., & Ahituv, N. (2022). Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. *Nature Communications*, *13*(1), 1504. https://doi.org/10.1038/s41467-022-28659-0

115. Abe, H., & Abe, K. (2022). PCR-based profiling of transcription factor activity in vivo by a virus-based reporter battery, *iScience*, *25*(3), 103927. https://doi.org/10.1016/j.isci.2022.103927

116. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N., & Yosef, N. (2019). Identification and massively parallel characterization of regulatory elements driving neural induction. *Cell Stem Cell*, *25*(5), 713–727.e10. https://doi.org/10.1016/j.stem.2019.09.010

117. Bovolenta, L. A., Acencio, M. L., & Lemke, N. (2012). Htridb: An open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, *13*(1), 405. https://doi.org/10.1186/1471-2164-13-405

118. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., … Hermjakob, H. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, *42*(D1), D358–D363. https://doi.org/10.1093/nar/gkt1115

119. Han, H., Cho, J. W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C. Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H. N., Jung, H., Nam, S., Chung, M., Kim, J. H., & Lee, I. (2018). TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, *46*(D1), D380–D386. https://doi.org/10.1093/nar/gkx1013

120. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., & Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, *29*(8), 1363–1375. https://doi.org/10.1101/gr.240663.118

121. Patel, N., & Bush, W. S. (2021). Modeling transcriptional regulation using gene regulatory networks based on multi-omics data sources. *BMC Bioinformatics*, *22*(1), 200. https://doi.org/10.1186/s12859-021-04126-3

122. Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., & Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, *13*(4), 366–370. https://doi.org/10.1038/nmeth.3799

123. Qin, Q., Fan, J., Zheng, R., Wan, C., Mei, S., Wu, Q., Sun, H., Brown, M., Zhang, J., Meyer, C. A., & Liu, X. S (2020). Lisa: Inferring transcriptional regulators through integrative modeling of public chromatin accessibility and chip-seq data. *Genome Biology*, *21*(1), 32. https://doi.org/10.1186/s13059-020-1934-6

124. Tchourine, K., Vogel, C., & Bonneau, R. (2018). Condition-specific modeling of biophysical parameters advances inference of regulatory networks. *Cell Reports*, *23*(2), 376–388. https://doi.org/10.1016/j.celrep.2018.03.048

125. Tripodi, I., Allen, M., & Dowell, R. (2018). Detecting differential transcription factor activity from ATAC-Seq data. *Molecules (Basel, Switzerland)*, *23*(5), 1136. https://doi.org/10.3390/molecules23051136

126. Feng, C., Song, C., Liu, Y., Qian, F., Gao, Y., Ning, Z., Wang, Q., Jiang, Y., Li, Y., Li, M., Chen, J., Zhang, J., & Li, C. (2020). KnockTF: A comprehensive human gene ex- pression profile database with knockdown/knockout of transcription factors. *Nucleic Acids Research*, *48*(D1), D93–D100. https://doi.org/10.1093/nar/gkz881

127. Hu, Z., Killion, P. J., & Iyer, V. R. (2007). Genetic reconstruction of a functional transcriptional regulatory network. *Nature Genetics*, *39*(5), 683–687. https://doi.org/10.1038/ng2012

128. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K., & Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. *PLoS Genetics*, *10*(3), e1004226. https://doi.org/10.1371/journal.pgen.1004226

129. Kang, Y., Patel, N. R., Shively, C., Recio, P. S., Chen, X., Wranik, B. J., Kim, G., Mcisaac, R. S, Mitra, R., & Brent, M. R. (2020). Dual threshold optimization and network inference reveal convergent evidence from TF binding locations and TF perturbation responses. *Genome Research*, *30*(3), 459–471. https://doi.org/10.1101/gr.259655.119

130. Ramos, A. I., & Barolo, S. (2013). Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1632), 20130018. https://doi.org/10.1098/rstb.2013.0018

131. Spivakov, M. (2014). Spurious transcription factor binding: Non-functional or genetically redundant? *BioEssays*, *36*(8), 798–806. https://doi.org/10.1002/bies.201400036

132. Gitter, A., Siegfried, Z., Klutstein, M., Fornes, O., Oliva, B., Simon, I., & Bar-Joseph, Z. (2009). Backup in gene regulatory networks explains differences between binding and knockout results. *Molecular Systems Biology*, *5*(1), 276. https://doi.org/10.1038/msb.2009.33

133. Dai, Z., Dai, X., Xiang, Q., & Feng, J. (2009). Robustness of transcriptional regulatory program influences gene expression variability. *BMC Genomics*, *10*(1), 573. https://doi.org/10.1186/1471-2164-10-573

134. Wu, W. S., & Lai, F. J. (2015). Functional redundancy of transcription factors explains why most binding targets of a transcription factor are not affected when the transcription factor is knocked out. *BMC Systems Biology*, *9*(Suppl 6), S2. *Nature*, 568(7751), 179–180. https://doi.org/10.1186/1752-0509-9-S6-S2

135. Bentsen, M., Heger, V., Schultheis, H., Kuenne, C., & Looso, M. (2022). TF-COMB—Discovering grammar of transcription factor binding sites. *Computational and Structural Biotechnology Journal*, *20*, 4040–4051. https://doi.org/10.1016/j.csbj.2022.07.025

136. Zhang, Q., Liu, W., Zhang, H. M., Xie, G. Y., Miao, Y. R., Xia, M., & Guo, A. Y. (2020). hTFtarget: A comprehensive database for regulations of human transcription factors and their targets. *Genomics Proteomics & Bioinformatics*, *18*(2), 120–128. https://doi.org/10.1016/j.gpb.2019.09.006

137. Arnosti, D. N., & Kulkarni, M. M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible bill- boards? *Journal of Cellular Biochemistry*, *94*(5), 890–898. https://doi.org/10.1002/jcb.20352

138. Kang, Y., Jung, W. J., & Brent, M. R. (2022). Predicting which genes will respond to transcription factor perturbations. *G3: Genes, Genomes, Genetics*, *12*(8), jkac144. https://doi.org/10.1093/g3journal/jkac144

139. Sigalova, O. M., Shaeiri, A., Forneris, M., Furlong, E. E., & Zaugg, J. B. (2020). Predictive features of gene expression variation reveal mechanistic link with differential expression. *Molecular Systems Biology*, *16*(8). 10.15252/msb.20209539

140. El-Brolosy, M. A., Kontarakis, Z., Rossi, A., Kuenne, C., Günther, S., Fukuda, N., Kikhi, K., Boezio, G. L. M., Takacs, C. M., Lai, S. L., Fukuda, R., Gerri, C., Giraldez, A J., & Stainier, D. Y. R. (2019). Genetic compensation triggered by mutant mRNA degradation. *Nature*, *568*(7751), 193–197. https://doi.org/10.1038/s41586-019-1064-z

141. El-Brolosy, M. A., & Stainier, D. Y. R. (2017). Genetic compensation: A phenomenon in search of mechanisms. *PLoS Genetics*, *13*(7), e1006780. https://doi.org/10.1371/journal.pgen.1006780

142. Ma, Z., Zhu, P., Shi, H., Guo, L., Zhang, Q., Chen, Y., Chen, S., Zhang, Z., Peng, J., & Chen, J. (2019). PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. *Nature*, *568*(7751), 259–263. https://doi.org/10.1038/s41586-019-1057-y

143. Wilkinson, M. F. (2019). Genetic paradox explained by nonsense. http://www.nature.com/articles/d41586-019-00823-5

144. Prelich, G. (2012). Gene overexpression: Uses, mechanisms, and interpretation. *Genetics*, *190*(3), 841–854. https://doi.org/10.1534/genetics.111.136911

145. Ehsani, R., Bahrami, S., & Drabløs, F. (2016). Feature-based classification of human transcription factors into hypothetical subclasses related to regulatory function. *BMC Bioinformatics*, *17*(1), 459. https://doi.org/10.1186/s12859-016-1349-2

146. Sherwood, R. I., Hashimoto, T., O'donnell, C. W., Lewis, S., Barkal, A. A., Van Hoff, J. P., Karun, V., Jaakkola, T., & Gifford, D. K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology*, *32*(2), 171–178. https://doi.org/10.1038/nbt.2798

147. Gillespie, M. A., Palii, C. G., Sanchez-Taltavull, D., Shannon, P., Longabaugh, W. J. R., Downes, D. J., Sivaraman, K., Espinoza, H. M., Hughes, J. R., Price, N. D., Perkins, T. J., Ranish, J. A., & Brand, M. (2020). Absolute quantification of transcription factors reveals principles of gene regulation in erythropoiesis. *Molecular Cell*, *78*(5), 960–974.e11. https://doi.org/10.1016/j.molcel.2020.03.031

148. Larsen, S. J., Röttger, R., Schmidt, H. H. H. W., & Baumbach, J. (2019). E. coli gene regulatory networks are inconsistent with gene expression data. *Nucleic Acids Research*, *47*(1), 85–92. https://doi.org/10.1093/nar/gky1176

149. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, *17*(2), 147–154. https://doi.org/10.1038/s41592-019-0690-6

150. Chen, S., & Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, *19*(1), 232. https://doi.org/10.1186/s12859-018-2217-z

151. Nguyen, H., Tran, D., Tran, B., Pehlivan, B., & Nguyen, T. (2021). A comprehensive survey of regulatory net- work inference methods using single cell RNA sequencing data. *Briefings in Bioinformatics*, *22*(3), bbaa190. https://doi.org/10.1093/bib/bbaa190

152. Trescher, S., & Leser, U. (2019). Estimation of transcription factor activity in knockdown studies. *Scientific Reports*, *9*(1), 9593. https://doi.org/10.1038/s41598-019-46053-7

153. Trescher, S., Münchmeyer, J., & Leser, U. (2017). Estimating genome-wide regulatory activity from multi-omics data sets using mathematical optimization. *BMC Systems Biology*, *11*(1), 41. https://doi.org/10.1186/s12918-017-0419-z

154. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining* (pp. 785–794). ACM. https://dl.acm.org/doi/10.1145/2939672.2939785

155. Hecker, D., Behjati Ardakani, F., Karollus, A., Gagneur, J., & Schulz, M. H. (2023). The adapted Activity-By-Contact model for enhancer-gene assignment and its application to single-cell data. *Bioinformatics (Oxford, England)*, *39*, btad062. https://doi.org/10.1093/bioinformatics/btad062

156. Schmidt, F., Gasparoni, N., Gasparoni, G., Gianmoena, K., Cadenas, C., Polansky, J. K., Ebert, P., Nordström, K., Barann, M., Sinha, A., Fröhler, S., Xiong, J., Dehghani Amirabad, A., Behjati Ardakani, F., Hutter, B., Zipprich, G., Felder, B., Eils, J., Brors, B., … Schulz, M. H. (2017). Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Research*, *45*(1), 54–66. https://doi.org/10.1093/nar/gkw1061

157. Schmidt, F., Kern, F., Ebert, P., Baumgarten, N., & Schulz, M. H. (2019). TEPIC 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis. *Bioinformatics*, *35*(9), 1608–1609. https://doi.org/10.1093/bioinformatics/bty856

158. Schmidt, F., Kern, F., & Schulz, M. H. (2020). Integrative prediction of gene exspression with chromatin accessibility and conformation data. *Epigenetics & Chromatin*, *13*(1), 4. https://doi.org/10.1186/s13072-020-0327-0

How to cite this article: Hecker, D., Lauber, M., Behjati Ardakani, F., Ashrafiyan, S., Manz, Q., Kersting, J., Hoffmann, M., Schulz, M. H., & List, M. (2023). Computational tools for inferring transcription factor activity. *Proteomics*, *23*, e2200462. https://doi.org/10.1002/pmic.202200462