Transcription factors: Bridge between cell signaling and gene regulation

Paula Weidemüller 2  |  Maksim Kholmatov 1  |  Evangelia Petsalaki 2  |  Judith B. Zaugg 1

1 Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstraße 1, Heidelberg 69117, Germany
2 European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton CB10 1SD, UK

Correspondence
Evangelia Petsalaki, European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton CB10 1SD, UK.
Email: petsalaki@ebi.ac.uk
Judith B. Zaugg, Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstraße 1, Heidelberg 69117, Germany.
Email: judith.zaugg@embl.de

Abstract
Transcription factors (TFs) are key regulators of intrinsic cellular processes, such as differentiation and development, and of the cellular response to external perturbation through signaling pathways. In this review we focus on the role of TFs as a link between signaling pathways and gene regulation. Cell signaling tends to result in the modulation of a set of TFs that then lead to changes in the cell’s transcriptional program. We highlight the molecular layers at which TF activity can be measured and the associated technical and conceptual challenges. These layers include post-translational modifications (PTMs) of the TF, regulation of TF binding to DNA through chromatin accessibility and epigenetics, and expression of target genes. We highlight that a large number of TFs are understudied in both signaling and gene regulation studies, and that our knowledge about known TF targets has a strong literature bias. We argue that TFs serve as a perfect bridge between the fields of gene regulation and signaling, and that separating these fields hinders our understanding of cell functions. Multi-omics approaches that measure multiple dimensions of TF activity are ideally suited to study the interplay of cell signaling and gene regulation using TFs as the anchor to link the two fields.

1 | INTRODUCTION

Transcription factors (TFs) are key regulators of cellular processes, both intrinsic, such as development and differentiation [1], as well as extrinsic, such as response to external signals [2]. Differentiation and reprogramming processes are typically driven by the transcriptional induction of a set of TFs, which then drive the required gene regulatory programs, as exemplified by the induction of Yamanaka factors (Oct4, Sox2, Klf4 and c-Myc) that are sufficient to reprogram fibroblasts into induced pluripotent stem cells (iPSCs) [3]. In contrast, the response to external cues is typically initiated by receptor activation and cell signaling, which acts on a time-scale of minutes, often through cascades of post-translational modifications (PTMs) resulting in the modulation of a set of TFs [4, 5]. The activity of a TF can be regulated either by modulating the abundance of its active form (including transcription, translation, and post-translational regulation), or by...
modulating the accessibility of its binding sites (including epigenetic processes and cell-type specific chromatin states). Once bound, TFs can open the chromatin for other factors to bind or prevent other factors from binding, and activate or repress the transcription of genes. Consequently, TFs can be studied at any of these levels, each associated with its own challenges and limited in the specific insights that can be gained.

TFs tend to be lowly abundant at the protein level [6–9], which makes them challenging to detect in proteomic assays. Furthermore, the function of TFs depends on PTMs and binding to interaction partners. Thus, their expression level does not necessarily correlate with their functional activity. On the other hand, the binding of TFs to chromatin can be readily assayed, for example, using chromatin immunoprecipitation followed by sequencing (ChIP-seq) and similar assays, which results in genome-wide maps of TF binding. Yet, it remains a challenge to delineate functional binding sites from unspecific binding events [10]. Furthermore, these binding assays are typically limited to one TF at the time, and thus grossly underestimate the complexity of gene regulation. Finally, TF activity can be inferred from assessing the chromatin accessibility around their predicted binding sites [11] or the expression of the genes they regulate [12–15]. The challenge for the former is that TF binding predictions suffer from a lack of specificity, and for the latter that our knowledge of TF-to-gene mapping is very incomplete.

In this review we focus on how TFs respond to cellular signaling, by adopting a broad definition of TFs: We used a carefully curated list by [2] and, since one aim of this review is to compare different resources that collect TF, we extended this list with proteins that are judged as TFs by commonly used databases such as TRRUST [16] and Dorothea [13]. Both databases are more lenient in their definition of TFs: TFs bind DNA within a complex and/or through a DNA-binding domain to regulate gene expression. We review the regulatory and functional aspects of TFs within signaling pathways and gene regulation with a focus on technological and data analysis challenges, and highlight the existence of a strong literature bias in the TF literature. We argue that gene regulation is part of cell signaling and propose that the most comprehensive way of studying the functional role of TFs in signaling is by combining the power of multiple assays that detect signaling activity, TF activity, their genomic localization and potential interaction partners. Therefore, TFs, which can be assayed both in terms of signaling and in terms of their impact on transcriptional regulation, offer a perfect bridge between the signaling and transcriptional regulation fields (Figure 1).

2 | POST TRANSLATIONAL MODIFICATION-BASED REGULATION OF TF ACTIVITY IN CELL SIGNALING

Signaling pathways tend to result in the activation or inactivation of TFs, often through PTMs of the TF, typically without altering their DNA binding specificity. Examples of well-known signaling cascades that lead to the activation of TFs are TGFbeta signaling leading to activation the of SMAD family TFs [17–19], Jak-STAT signaling activating the STAT TFs [20–23], Erbb2 signaling typically activating Jun and Myc [24, 25], Hippo signaling targeting the TEA-domain-containing (TEAD) family (TEAD1–TEAD4) of TFs [26, 27] and Notch signaling that induces dissociation of DNA-bound RBPJ from a corepressor complex and recruitment of a coactivator complex instead [28, 29]. Examples of TFs that are inactivated by signaling include the FOXO family, a subclass of Forkhead TFs. In the absence of insulin, FOXO TFs are bound to DNA and activate gene expression. Upon insulin presence, FOXO TFs are phosphorylated by kinases downstream of the PI3K-AKT signaling pathway, which leads to exclusion of TFs from the nucleus and hence repression of their target gene [30, 31].

A special family of TFs are nuclear receptors [32] that act as both receptors for lipid-soluble ligands (e.g., steroid hormones) and, subsequently, as TFs to regulate gene expression without the need for
an intermediary signaling transduction (reviewed in [33–35]). Nuclear receptors are characterized by the presence of a ligand binding domain, a nuclear localization domain, a transactivation domain and a specific DNA-binding domain [36, 37]. One class of nuclear receptors, such as estrogen receptor and glucocorticoid receptor, reside in the cytoplasm until binding to their ligand allows translocation to the nucleus and expression of target genes. Another class of receptors, like the thyroid hormone receptor, are already bound to their specific DNA-sequence as a heterodimer with retinoid X receptor and binding of their ligand replaces coresspressors with coactivator complexes to initiate gene expression [33, 35]. Coregulators of nuclear receptors serve as important targets, propagators and integrators of PTM’s to drive specific gene expression programs [38, 39].

The most studied PTM in signaling is phosphorylation, but other PTMs such as sumoylation, ubiquitination, acetylation, glycosylation, and methylation also play a role in signal transduction. PTMs can affect TF localization, stability, activity and interaction with other proteins [4,40–42]. See [26,43–45] for focused reviews on the effect of different PTMs on subsets of TFs. TFs can carry several modifications at once and the modifications might be dependent on each other as reviewed by [46, 4, 47].

An early attempt of curating the complexities of PTM regulation in TFs is the PTM-switchboard database [48] which is unfortunately not accessible anymore. Not much progress in understanding TF PTMs has been made since. This is also illustrated by the recently published compendium of TFs, which explicitly refrains from characterizing the PTMs that regulate TFs because only few studies systematically annotated and disentangled the complex combinatorial effects of PTMs on the function of TFs [2].

On the other hand, the rapid advancement of mass spectrometry (MS) in recent years has enabled high-throughput measurement of PTMs on proteins. Specifically for phosphorylation, the number of datasets (phosphoproteomics) increased from 127 annual submissions to PRIDE in 2010 to 2344 in 2019 [49]. These data provide a great resource for investigating the effect of different phosphorylations on the TF function.

To exemplify the potential of this uncharted area, we summarize the phosphosite data available for TFs. We compiled a list of 1967 human TFs using the carefully curated list of 1639 TFs by Lambert et al. [2], supplemented by TFs curated by two commonly used databases that provide TF-gene interactions Dorothea [13] and TRRUST [16] (see Table S1). Most TFs in these two databases are also defined as TFs by Lambert. However, 328 proteins are not considered TFs by the strict definition of Lambert et al, likely because they don’t bind to a specific DNA-sequence but are part of the more general gene regulation machinery. Figure S1 shows the intersection of TFs as defined by the three resources. Databases such as PhosphoSitePlus [50, 51] and PTMcode2 [52, 53] collect and annotate the presence and function of PTMs on proteins in several species. While PTMcode2 specifically curates and predicts functional associations of PTMs between proteins it covers only a few TFs. Hence, we queried the curated list of TFs in PhosphoSitePlus, specifically collecting information about phosphorylation sites (phosphosites) [51]. Of the 1967 TFs, 1857 (94%) have at least one measured phosphosite and 934 (47%) have more than ten. However, only 393 TFs (20%) have a known functional phosphosite (i.e., annotated with a functional effect or known process). Among the functional phosphosites in TFs, the most common effects on protein function are related to regulation of molecular association, intracellular localization, protein degradation, protein stabilization and induced activity, while the most common effects on biological processes are altered transcription (both induced and inhibited), cell cycle regulation and altered cell growth (see Table S2).

When grouping TFs based on their annotated InterPro [54] protein families, expectedly the most common among TFs were DNA-binding Domains (DBDs), such as "Zinc finger C2H2 superfamily", "KRAB domain superfamily", "Homeobox-like domain superfamily", "Winged helix-like DNA-binding domain superfamily" and "Helix-loop-helix DNA-binding domain superfamily" (see Table S1). We found the proportion of TFs with known functional phosphosites varies greatly across TF protein families obtained from InterPro. For example, TFs in the "KRAB domain superfamily" have only two proteins (1%) with functional phosphosites, whereas TFs in the "nuclear hormone receptor-like" and "Zinc finger, NHR/GATA-type" families have 27 proteins (56%) and 31 proteins (55%), respectively (Figure 2A). A very similar distribution was observed when considering only TFs from the curated list of Lambert et al. [2] (Figure S2A). Unsurprisingly, the protein family with the largest proportion of proteins with known functional phosphosites is the "Protein kinase-like domain superfamily", the most studied protein family in signaling (Figure 2B), whereas for a large proportion of human proteins including TFs the effect of phosphosites is understudied despite the availability of phosphoproteomic datasets covering them (Figure 2B).

The functional relevance of measured phosphosites is a matter of ongoing research. PhosphoSitePlus has 29,453 phosphosites recorded for our compiled list of TFs, yet whether all of these phosphosites have actual functional implications remains an open question. A recent study predicts the functional relevance of 119,809 human phosphosites based on 59 features, including protein abundance, protein length, residue conservation and kinase position weight matrix similarity [55]. They provide functional scores for 15,256 phosphosites in 1263 TFs (64% of our curated TFs), most of which have a low functional score and are thus unlikely to regulate TF activity. Still, 275 TFs (14%) possess at least one phosphosite with a high functional score (>0.5) yet no known regulatory annotation in PhosphoSitePlus (Figure 2B), suggesting that further research is needed to understand the effects of individual phosphosites on the activity of TFs.

3 | REGULATION OF TF-BINDING TO CHROMATIN THROUGH DNA-SEQUENCE AND EPIGENETICS

Once a TF is activated by a signaling cascade, and before it can modulate the expression of its target gene, it has to bind to chromatin. There are multiple ways of how TF binding to chromatin is regulated. We briefly review the major mechanisms below, focusing on
FIGURE 2  Number and classification of phosphosites obtained from PhosphoSitePlus (PSP) across protein families. (A) Comparison across the 13 most common TF protein families. (B) Comparison of the curated list of TFs with the 10 most common non-TF protein families, as annotated in InterPro. Families are sorted by the fraction of proteins with at least one phosphosite (total bar length). The fraction of proteins with at least one functional phosphosite (blue), at least one phosphosite with a high predicted functional score (>0.5) as defined by Ochoa et al [55], (orange), their intersection (shaded area), and the fraction of proteins with no functional annotated phosphosite (grey) are shown (n = total number of proteins; n_TF = number of TFs in family). Our curated list of TFs is highlighted in bold. Some InterPro families are a subset of another family with a similar name, those were merged into one set and called according to the larger family (indicated by an asterisk).

sequence-specific TFs (as opposed to general TFs that are part of the basal transcription machinery and bind less sequence specific).

Sequence-specific TFs typically bind to specific DNA sequences that can be summarized as TF motifs as reviewed in Slattery et al [56]. These motifs are inferred from TF binding assays in vitro (e.g., SELEX [57], protein-binding DNA-arrays [58]) or in vivo (e.g., ChIP-seq, or other, more recent chromatin profiling technologies such as ChIP-exo [59], ChIP-nexus [60], CUT&Tag [61], or CUT&RUN [62]), by finding enriched sequences among the TF-bound DNA fragments. These motifs, which are collected in motif databases, such as JASPAR [63], HOCOMOCO [64], CIS-BP [65], and others [66], can then be used to predict putative TF binding sites in the genome, using tools like
The epigenetic profile of a cell constitutes an additional layer that contributes to context-cell-type specific TF binding [10]. This includes DNA methylation and chromatin modifications, which are PTM of histone tails that correlate with functional properties of chromatin [85]. Chromatin modifications are mostly known for their ability to recruit chromatin remodeling complexes, for example polycomb [86], and parts of the basal transcription machinery, such as TFIIID [87, 88]. Even though a few sequence-specific TFs have also been shown to directly interact with specific histone modifications [89], the main impact of chromatin modifications on TF binding is likely mediated through their effect on DNA accessibility. For example, lysine acetylation neutralizes the positive charge of histone residues and thus decreases nucleosome affinity to DNA [90, 91]. This effect has theoretically been described by a nucleosome-mediated cooperativity model [92], which proposes competition for DNA binding between nucleosomes and a set of TFs as a dynamic equilibrium. A recent study has shown experimental evidence for a slightly updated model of TF-nucleosome cooperativity that includes active nucleosome remodeling [93].

This model also implies that TFs play an important role in modulating chromatin accessibility and thereby define the epigenetic landscape of a cell. This is most evident for the class of so-called pioneer TFs, which are defined based on their ability to bind to closed chromatin and make it accessible for other TFs to bind, for example during cell fate decisions (recently reviewed in Zeitlinger [94]). There is also accumulating evidence that non-pioneer TFs can regulate chromatin. For example, in [95] the authors achieved a reasonably accurate prediction of histone modifications across cell lines based only on TF binding data. More recently a deep-learning framework was able to predict the chromatin accessibility profiles of immune cells based on sequence and thereby discovered the sequence-motifs of cell-type specific TFs ab initio [96]. Furthermore, observations that genetic variants that modulate histone modifications tend to disrupt TF binding sites [97, 98] suggest a causal-direct or indirect-role of TF binding in regulating histone modifications. Thus, while chromatin modifications and accessibility may determine where TFs can bind, and integrating them is useful for inferring context-specific TF binding, they are also actively being modulated by TFs.

Certain modifications, specifically those related to accessible chromatin (e.g., Histone 3 lysine 27 acetylation (H3K27ac)), can therefore even serve as a direct readout of TF activity, which highlights the tight interconnection between signaling and gene regulation [99]. Recent studies have formalized this relationship to quantify differential TF activity by aggregating changes in histone modifications or chromatin accessibility across the predicted binding sites of a TF (differTF [11], chromVar [100]). Together with the development of single cell chromatin accessibility profiling [101] and in particular its recent commercialization, this will dramatically increase our understanding of cell-type specific TF activity profiles in the future.

4 | TFs as part of gene regulatory networks (GRN)

Conceptually, the final result of activating a TF is the modulation of expression in the set of its direct target genes, also referred to as regulon of a TF. The combined activity of a set of TFs connected to their targeted genes is referred to as gene regulatory network (GRN) [102]. These networks are responsible for maintaining cell-type specific transcriptional states and response to signaling. However, the exact nature...
of these networks is unknown and we still lack a global understanding of the impact of TFs on transcriptome changes. This is illustrated by recent attempts to predict the impact of TF perturbations on the transcriptome, which have performed poorly even in yeast [103] that have a much simpler regulatory architecture than mammals. In contrast, models using gene-specific features, such as expression variability across individuals, are highly predictive of transcriptome changes in response to perturbation assays [103, 104]. Partially, this lack of understanding about the direct impact of TFs on gene expression can be ascribed to the lack of a globally accepted (and experimentally measurable) gold standard dataset that can be used to benchmark GRNs. Thus, methods for GRN inference typically rely on strong assumptions and are benchmarked against each other or against small or biased sets of experimentally validated interactions.

The most comprehensive resource for experimentally validated TF-gene interactions is the TRRUST (transcriptional regulatory relationships unravelled by sentence-based text-mining) database [16], which is based on manual curation and currently comprises over 8000 TF-gene interactions. Typically, these links are derived from studies that focus on one TF in one specific context at the time. However, similar to its binding to DNA, the set of genes regulated by a given TF is likely highly context-specific. In fact, most TFs in TRRUST are classified as activator and as repressor almost equally often (Figure 3), suggesting that even the actual function of a TF is highly context-dependent. An alternative explanation for this is that the data-curation underlying the TRRUST database is incomplete. Either way, while it is a great resource for testing individual TF-gene interaction in a given context (i.e., consulting the curated studies) it is not a reliable source for inferring genome-scale GRNs.

One strategy for inferring genome-scale GRNs is based on perturbation studies that alter the activity of a TF (through overexpression, knockdown, knockout or chemical inhibitors) and then measure the resulting changes in DNA binding or target gene expression [105, 106]. A number of these studies have been curated within the KnockTF database covering 308 human TFs [107]. Another set of methods are based on coexpression of TFs and genes (e.g., WGCNA [108]), with some variations that use energy-based or information-based measures instead of correlation (e.g., DPM [109], sdcorGCN [110], PIDC [111, 112]. These approaches (reviewed in [113–115]) are based on the assumption that a change in TF expression level will result in a transcriptional change of its regulon. Despite the significant progress and numerous practical applications of co-expression to GRN inference their direct interpretation in terms of gene regulation is limited due to missing directionality. More recently, the use of co-expression to infer modules of jointly regulated genes (regulons) has been combined with prior knowledge of TF binding sites and/or TF perturbation studies, to define TF-specific regulons [116, 13, 117], in some approaches even integrating TF-mediated enhancer activation [118, 119], which limits the target genes to those co-expressed with and likely bound by a TF.
Figure 4 Assessment of literature bias. Literature bias, defined as number of PubMed-indexed studies that mention a TF's gene name in title or abstract, is compared against curated and predicted knowledge for TFs on the level of signaling (A-D) DNA-binding (E-F) and target genes (G-H). The literature bias (Y-axis) is plotted against (A-B) the number of functionally validated (A) and measured (B) phosphosites in TFs; (C-D) the number of curated TF-protein interactions by IntAct (C) and functional phosphosites predicted by Ochoa et al. [55]; (D-E) TF binding sites measured by ChIP-seq (E) and predicted based on motifs (F); (G-H) the number of TF target genes curated by TRRUST (G) and predicted by DoRothEA (H). Axes are scaled logarithmically. Data points are binned into hexagons to show dense areas of TFs using the openair package [137].

Overall, the biggest practical challenge in linking TFs to their target genes is the lack of a ground-truth dataset and thus GRN reconstructions are not uniformly and globally validated. The second set of challenges are the conceptual limitations of the individual approaches to infer GRNs. Co-expression inferred networks assume TFs are regulated on the transcriptional level, which may be particularly misleading for TFs involved in signaling pathways, since they are regulated by PTMs. TF-perturbation-based GRNs on the other hand don't account for gene-specific features that may affect their overall responsiveness independent of the perturbation and don't easily distinguish direct from indirect effects. Thus, understanding the role of TFs on their target genes requires a range of complementary GRN inference methods that are able to cover each other’s blind spots.

5 THE SAME TFs ARE UNDERSTUDIED AT ALL MOLECULAR LAYERS

For each layer of regulating TF activity there are literature curated and large-scale measured or inferred data. For example, the collection of phosphosites in PhosphoSitePlus incorporates high-throughput mass-spectrometry screens [51]. In contrast to functional studies that focus on a few proteins at a time, these screens are not biased a priori towards specific sets of proteins. Similarly, TF binding to chromatin as measured by ChIP-seq data requires experiments in a specific cell type and context, whereas motif-based predictions of TF binding sites are data-independent. Finally, genes regulated by TFs can be curated in small, functional studies, or inferred based on high-throughput data.

To quantify a potential literature bias in functional annotation of these different measures of TF activity, we defined a measure of how well a TF is studied as the number of PubMed-indexed studies that mention its gene name in their titles or abstracts (query on 09.03.2021, see Table S3). This revealed between 0 and 1,120,174 studies per TF with 50% of TFs having less than 44. Hence, a few TFs are studied very intensively, while most TFs gather little attention. This bias towards a small set of well-studied TFs was already observed over ten years ago by Vaquerizas et al. [9]. Notably, most of the least-cited TFs belong to the Zinc finger C2H2 family. Hence the largest family of TFs (716, Figure 2A) is greatly understudied compared with other families. This is further reflected by the relatively low percentage of Zinc finger C2H2 TFs with known functional phosphosites (Figure 2A).

Overall, the number of unbiasedly measured phosphosites per TF is independent from the number of studies citing the TF (Figure 4A), whereas, as expected, functional annotations of phosphosites show a clear bias towards well studied TFs (Figure 4B). Along the same lines,
the number of functional phosphosites proposed by the machine learning model of Ochoa et al. [55], which included several non-literature based features, shows little literature bias (Figure 4C), whereas IntAct [120], which relies mainly on interactions curated from literature, shows a clear relationship between the number of publications and the number of annotated interaction partners (Figure 4D). For TF binding to chromatin, as measured by ChIP-seq data and collected by ReMap [75], the number of TF-bound regions from ChIP-seq experiments increases with the number of studies citing the TF (Figure 4F), thus indicating a strong literature bias. In contrast, no strong bias is observed for predicted TF binding sites in the human genome (assembly GRCh38) based on the binding models from HOCOMOCOv11 [64], except where predictions are not possible due to less-studied TFs often lacking motif annotations (Figure 4E). Curated TF targets in TRRUST [16] seem mostly available for highly studied TFs, as illustrated by the strong relationship between the number of studies citing a TF and the number of its target genes reported in TRRUST (Figure 4H). A similar relationship between literature bias and number of predicted targets is not observed for more data-driven approaches to link TFs to their targets, such as DoRothEA [13] (Figure 4G), which, in addition to literature curation also includes ChIP-seq peaks, TF binding site motifs and gene co-expression.

Thus, many of the measured phosphosites in TFs, their predicted binding sites and inferred target genes await further functional studies (Figure 4). To assess whether the same TFs are well-studied for their role in signaling (i.e., PTM regulation) and their role in gene regulation (i.e., effect on chromatin binding or gene regulation), we compared their literature-curated and predicted/inferred measures of TF activity. As expected we observe a strong relationship between the number of literature curated functional phosphosites in PhosphoSitePlus [51] and curated target genes of a TF from TRRUST [16] (Figure 5A). This relationship is less strong — yet still visible when comparing functional phosphosites with the number of measured TF binding sites by ChIP-seq data [75] (Figure 5B). In contrast, comparing the unbiased measures of phosphosites versus inferred targets from DoRothEA [13] reveals an inverse relationship (Figure 5C), and no relationship is observed with predicted binding sites from HOCOMOCO [64] (Figure 5D).
Overall, this indicates that a small subset of TFs has been studied extensively both in terms of their involvement in signaling pathways (approximated by functionally annotated phosphosites) and in gene regulation (approximated in the number of curated target genes). The inverse relationship between the corresponding literature-independent measures of signaling and inferred targets may represent an interesting biological observation that TFs highly regulated by signaling have smaller regulons to allow a more focused response. Alternatively, it could stem from a technical bias since DoRoTHEA partially relies on co-expression patterns, and TFs heavily regulated by PTMs may not be captured as efficiently. Either way, this observation suggests further studies and new approaches are needed to jointly investigate the role of TFs in signaling and transcriptional regulation.

6  |  MULTOMICS STUDIES CAN EXPLOIT TFs AS BRIDGE BETWEEN SIGNALING AND GENE REGULATION

Both signaling and TF-motif based inference of TF activity have their blind spots. While signaling studies typically do not focus on TFs, chromatin-based inference of TF activity only measures the effect of TF binding on chromatin, which may or may not result in transcriptional changes of its target genes, while TF activity inferred from target gene expression may be confounded by indirect effects. Nevertheless, TFs represent a natural bridge between the field of signaling and the field of gene regulation, since their activity—in principle—can be measured in both.

Over the past years several studies appreciated the need for integration of the different layers to derive a more detailed understanding of cell functions and phenotypes in the conditions under study. In the following we summarize different approaches that have extracted information that sheds light on both the signaling and gene regulation layers to gain insights that would have been missed by looking at each of the layers separately.

One group of approaches uses causal reasoning to integrate transcriptomics or on occasion phosphoproteomics data with prior knowledge-based protein interaction networks and pathways to infer upstream signaling regulation of the gene networks that are represented in the transcriptomics data [121–127]. Common to the cited studies is the dependence on a reliable annotation of TFs with their target genes, highlighting once more the pivotal role of TFs in understanding the connection of signaling and gene regulation. CARNIVAL [126], in particular, explicitly estimates TF- and pathway-activities [128, 13] from the gene expression data using prior knowledge, which improves the performance of the method compared to other causal reasoning approaches. Another recent approach, named KPNN (knowledge-primed neural network), uses prior knowledge about signaling and regulatory interactions as a constraint on the architecture of the hidden layers in a neural network, which marginally improved the predictions yet greatly increased stability and biological interpretability of the neural networks [125]. Extending beyond signaling, NicheNet, focuses on understanding cell-cell interactions by combining various data sources for creating a ligand-to-target gene network, and integrating prior knowledge on signaling and GRNs [122]. It calculates a ligand-target gene interaction score, and based on this, predicts cell-cell interactions.

A complementary approach relying on chromatin accessibility-inferred TF activity to approximate signaling has used a combination of RNA-seq and chromatin accessibility to generate a cell-type specific regulatory network [118] and projected TF activity onto the network. This revealed an enhancer-driven remodeling, which primed patient-derived cells to an aberrant response to TGFβ signaling in pulmonary arterial hypertension. A combination of both causal reasoning and direct use of accessibility data is used in methods like CellOracle [129] and Inferelator3.0 [130] which use chromatin accessibility data to define the cell-type specific prior regulatory network that is then refined by training a regularized linear regression model on gene expression data.

Other approaches explicitly integrate (phospho)proteomics and transcriptomics datasets and/or other omics data to identify multi-layer components that can explain the observed differences across datasets or provide integrated signaling-gene regulation networks that represent the cell function in the studied condition or patient. For example [131], adapted the TieDIE signaling network diffusion algorithm [132] to phosphoproteomics and transcriptomics data to find druggable kinase pathways. Based on the two datasets they inferred differential TF activity and kinase activity in metastatic castration-resistant prostate cancer patients and proposed new therapeutic targets, which would have been missed in each dataset on its own. Another study investigated the differentiation of mouse embryonic stem cells into neurons using RNA-seq, ATAC-seq, proteomics and protein-interaction experiments [78]. They established the relative importance of each regulatory layer along neuronal differentiation and could show that chromatin accessibility preceded changes in RNA and protein abundance. Furthermore, the authors uncovered a new role of the canonical pluripotency TF SOX2 as a regulator in differentiated neurons, which was only possible by integrating information from protein-interaction, chromatin accessibility and gene expression layer.

These studies together with several others highlight the additional insights that can be obtained by integrating data of the signaling, epigenetic and gene regulatory layer.

7  |  OUTLOOK/DISCUSSION

Signaling and gene regulation are interlinked and one can’t be studied independently from the other if we really want to understand cell responses to changes in the environment and their link to resulting phenotypes. We highlighted that TFs are a strong candidate for linking the two processes (end point of signaling, starting point of gene regulation) and are ideally suited to do so also practically, since their activity can be inferred by both epigenetic and transcriptomic data as well as phospho/proteomic data.

TFs are regulated on many levels, with the most understudied being their regulation by PTMs. This might be partly due to their relatively low abundance compared to other signaling proteins, making them
harder to detect by MS. However, given the fact that we identified very high numbers of putatively functional phosphosites on TFs, it is more likely that the bottleneck is formed by literature bias: Well-studied signaling pathways converge to a handful of well-studied TFs and their regulation in the context of these pathways becomes the focus of most studies. TFs that have not been already annotated in specific signaling pathways are harder to study and therefore often ignored.

Indeed, we found that literature-based annotations, which are vital for gathering detailed and robust evidence for TF function, are lacking for a large proportion of TFs, and even for well-studied TFs they seem contradictory, as illustrated by the fact that well-studied TFs are equally often classified as activators or repressors by curated studies. When comparing unbiased studies of both phosphorylation and other functional indications for TFs, we find no relationship with the number of studies citing a TF, strongly suggesting that shedding light on the functions of understudied TFs has the potential to be transformative for our understanding of cell functions. For example, investigating understudied TFs (“dark TFs”) could turn the focus from already well-studied canonical signaling pathways to more context-dependent functions of TFs leading to novel insights about TF regulation, activation and interaction with other proteins and chromatin. Prioritization strategies for the functional studies of the “dark TFs” could include intersecting those that have both predicted functional phosphosites and large inferred regulons or particularly interesting target genes. Integrating the different layers of TF involvement, such as signaling and chromatin, will be vital to derive a comprehensive understanding of the mechanisms involved in defining transcriptional changes upon an extracellular signal.

Multi-omics data integration has become standard practice and is applied in many studies. However, most studies still focus on either genomics technologies based on sequencing (i.e., transcriptomics and epigenetic) or (phospho-)/proteome studies based on MS. Partially, this may represent an experimental challenge, since many of the MS-based technologies require large amounts of cellular material, which is not easily obtained from primary samples, while sequencing-based technologies are available for single cells, which allows studies on a much more detailed cellular resolution. While some recent technologies allow joint interrogation of surface proteins (based on barcoded antibodies against epitopes) and transcriptome, they are currently still limited to a small number of surface proteins that can be interrogated. We propose that future multi-omics studies should routinely go beyond their immediate neighbor fields to integrate truly complementary data and gain a comprehensive view on the interplay between cell signaling and gene regulation.

The employment of single-molecule imaging techniques adds another layer of insight into the dynamics, regulation and binding properties of TFs. Imaging of TFs, interaction partners, chromatin and gene expression has the power to observe and quantify the spatial and temporal processes underlying the gene regulatory machinery in vivo. Methods such as fluorescence recovery after photobleaching (FRAP), single molecule tracking (SMT) and fluorescence correlation spectroscopy (FCS) helped to investigate the kinetics and timescales of TF-chromatin interaction, uncovered the presence of fast and transient TF binding and aid in studying nuclear-cytoplasmic shuttling of TFs upon signaling activation. For more detailed reviews we refer to [133–135] and [136].

TFs provide an excellent anchor for linking signaling studies with gene regulation. Given the importance of the currently well-studied TFs for cell function, the large number of understudied TFs and our poor understanding of the interplay of these regulatory layers, a focus on functionally annotating TFs in the context of cell signaling and gene regulation is likely to prove transformational for our understanding of cell functions.

ACKNOWLEDGMENTS
Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT
Data for generating the figures is given in supplementary tables.

ORCID
Paula Weidemüller https://orcid.org/0000-0003-3867-3131
Maksim Kholmatov https://orcid.org/0000-0002-4201-8124
Evangelia Petsalaki https://orcid.org/0000-0002-8294-2995
Judith B. Zaug https://orcid.org/0000-0001-8324-4040

REFERENCES
1. Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: From enhancer binding to developmental control. Nature Reviews Genetics, 13, 613–626.

2. 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, 650–665.

3. Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell, 126, 663–676.

4. Filtz, T. M., Vogel, W. K., & Leid, M. (2014). Regulation of transcription factor activity by interconnected post-translational modifications. Trends in Pharmacological Sciences, 35, 76–85.

5. Karin, M., & Hunter, T. (1995). Transcriptional control by protein phosphorylation: Signal transmission from the cell surface to the nucleus. Current Biology, 5, 747–757.

6. Deshmukh, A. S., Murgia, M., Nagaraj, N., Treebak, J. T., Cox, J., & Mann, M. (2015). Deep proteomics of mouse skeletal muscle enables quantitation of protein isoforms, metabolic pathways, and transcription factors. Mol Cell Proteomics MCP, 14, 841–853.

7. Li, J. J., Bickel, P. J., & Biggin, M. D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ, 2, E270.
immune cell differentiation. Proceedings of the National Academy of Sciences of the United States of America, 117, 25655–25666.
97. Grubert, F., Zaug, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., Greenside, P., Srivast, R., Phanstiel, D. H., Pekowska, A., Heidari, N., Euskirchen, G., Huber, W., Pritchard, J. K., Bustamante, C. D., Steinmetz, L. M., Kundaje, A., & Snyder, M. (2015). Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell, 162, 1051–1065.
98. Waszak, S. M., Delaneau, O., Goechw, A. R., Kilpinen, H., Raghav, S. K., Witwicki, R. M., Orioli, A., Wiedekerh, M., Panousis, N. I., Yurovsky, A., Romano-Paluimo, L., Planchoh, A., Bielser, D., Padioleu, I., Udin, G., Thurnheer, S., Hacker, D., Hernandez, N., Reymond, A., ... Dermitzakis, E. T. (2015). Population variation and genetic control of modular chromatin architecture in humans. Cell, 162, 1039–1050.
99. Peng, Y., Li, S., Landsman, D., & Panchenko, A. R. (2021). Histone tails as signaling antennas of chromatin. Current Opinion in Structural Biology, 67, 153–160, https://doi.org/10.1016/j.sbi.2020.10.018.
100. 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, 975–978.
101. Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature, 523(7561), 486–490, https://doi.org/10.1038/nature14590.
102. Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Himman, V., Arenas-Mena, C., Otim, O., Brown, C. T., Livis, C. B., Lee, P. Y., Revilla, R., Rust, A. G., Pan, Zj, Schilstra, M. J., Clarke, P. J., ... Boulo, H. (2002). A genomic regulatory network for development. Science, 295, 1669–1678.
103. Yang, Y. & Brent, M. R. (2020). Predicting which genes will respond to perturbations of a TF: TF-independent properties of genes are major determinants of their responsiveness. bioRxiv, https://doi.org/10.1101/2020.12.15.422864.
104. Sigalova, O. M., Sheai, A., Forneris, M., Furlong, E. E., & Zaug, J. B. (2020). Predictive features of gene expression variation reveal mechanistic link with differentiation expression. Molecular Systems Biology, 16, E9539.
105. Alda-Catalinas, C., Bredikhin, D., Hernandez-Herraez, I., Santos, F., Kubinyecz, O., Eckersley-Isalyn, 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, 25–41.e9.
106. Freimer, J. W., Shaked, O., Naqvi, S., Sottom-Arron, N., Kathiriya, A., Chen, A. F., Cortez, J. T., Greenleaf, W. J., Pritchard, J. K., & Marson, A. (2021). Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks. bioRxiv, https://doi.org/10.1101/2021.04.18.440363.
107. 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 expression profile database with knockdown/knockout of transcription factors. Nucleic Acids Research, 48, D93–D100.
108. Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559.
109. Ghanbari, M., Lasserre, J., & Vingron, M. (2019). The distance precision matrix: Computing networks from non-linear relationships. Bioinformatics, 35, 1009–1017.
110. Pardo-Diaz, J., Bozhilova, L. V., Beguerisse-Diaz, M., Poole, P. S., Deane, C. M., & Reinert, G. (2021). Robust gene coexpression networks using signed distance correlation. Bioinformatics.
111. Chan, T. E., Stumpf, M. P. H., & Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate
information measures. Cell Systems, 5(3), 251–267.e3, http://doi.org/10.1016/j.cels.2017.08.014.

112. Zhao, J., Zhou, Y., Zhang, X., & Chen, L. (2016). Part mutual information for quantifying direct associations in networks. Proceedings of the National Academy of Sciences, 113(18), 5130–5135, http://doi.org/10.1073/pnas.1522586113.

113. Emmert-Streib, F., Glazko, G., Gloor, M., & De Matos Simoes, R. (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Frontiers in Genetics, 3, 8.

114. Reyes-Palomares, A., Gu, M., Grubert, F., Berest, I., Sa, S., Kasowski, M., Arnold, C., Shuai, M., Srivas, R., Miao, S., Snyder, M. P., Rabinovitch, M., & Zaugg, J. B. (2020). Remodeling of active endothelial enhancers is associated with aberrant gene-regulatory networks in pulmonary arterial hypertension. Nature Communications, 11, 1673.

115. Zhao, J., Zhou, Y., Zhang, X., & Chen, L. (2016). Perturbation-response genes reveal signaling footprints in cancer gene expression. Nature Communications, 9, 20.

116. Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imri, S., Jimenez, R., Khadake, J., Lagreid, A. ... Hermjakob, H. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Database, 2014, D358–D363.

117. Malis, J. N., Sakellaropoulos, T., Iorio, F., Alexopoulos, L. G., Loh, W.-Y., Laubenberger, D. A., Saiz-Rodriguez, J., & Bai, J. P. F. (2015). Identification of drug-specific pathways based on gene expression data: Application to drug induced lung injury. Integrative Biology, 7, 904–920.

118. Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Waters, A., De Veaux, N., Tchourine, K., Yi, R., Hamamsy, T., Castro, D. M., Carrierio, N., Gresham, D., Mira/aldi, E. R., & Bonnaire, R. (2021). Single-cell gene regulatory network inference at scale: The inferelator 3.0. bioRxiv, https://doi.org/10.1101/2021.05.03.442499.

119. Drake, J. M., Pauli, E. O., Graham, N. A., Lee, J. K., Smith, B. A., Titz, B., Stoyanova, T., Faltermeyer, C. M., Uzunangelov, V., Carlin, D. E., Flemming, D. T., Wong, C. K., Newton, Y., Sudha, S., Vashisht, A. A., Huang, J., Wohlschlegel, J. A., Graeber, T. G., & Witte, O. N. (2016). Phosphoproteome integration reveals patient-specific networks in prostate cancer. Cell, 166, 1041–1054.

120. 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., Jaganathan, 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, D358–D363.

121. Bradley, G., & Barrett, S. J. (2017). CausalR: Extracting mechanistic sense from genome scale data. Bioinformatics, 33, 3670–3672.

122. Browaeys, R., Glazko, G., Gloor, M., & De Matos Simoes, R. (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Frontiers in Genetics, 3, 8.

123. Chaisson, D., Ho, Y.-H., Berry, D. B., Nemec, C. M., MacGillivray, M. E., Hose, J., Merrill, A. E., Lee, M. V., Will, J. L., Coon, J. J., Coon, J. J., Ansari, A. Z., Craven, M., & Gasch, A. P. (2014). Pathway connectivity and signaling coordination in the yeast stress-activated signaling network. Molecular Systems Biology, 10, 759.

124. Chindelevitch, L., Ziemek, D., Enayetallah, A., Randhawa, R., Sidders, B., Brockel, C., & Huang, E. S. (2012). Causal reasoning on biological networks: Interpreting transcriptional changes. Bioinformatics, 28, 1114–1121.

125. Fortelny, N., & Bock, C. (2020). Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. Genome Biology, 21, 190.

126. Liu, A., Trairatphisan, P., Gjerde, E., Didangelos, A., Barratt, J., & Saiz-Rodriguez, J. (2019). From expression footprints to causal pathways: Contextualizing large signaling networks with CARNIVAL. bioRxiv, 541888, https://doi.org/10.1101/541888.

SUPPORTING INFORMATION
Additional supporting information may be found online https://doi.org/10.1002/pmic.202000034 in the Supporting Information section at the end of the article.

How to cite this article: Weidemüller, P., Kholmatov, M., Petsalaki, E., & Zaugg, J. B. (2021). Transcription factors: Bridge between cell signaling and gene regulation. Proteomics, 21:e2000034. https://doi.org/10.1002/pmic.202000034