BBA - Gene Regulatory Mechanisms 1863 (2020) 194443

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns

Divyanshi Srivastava, Shaun Mahony*

Center for Eukaryotic Gene Regulation, Department of Biochemistry & Molecular Biology, The Pennsylvania State University, University Park, PA, United States of America

A B S T R A C T

Transcription factors (TFs) selectively bind distinct sets of sites in different cell types. Such cell type-specific binding specificity is expected to result from interplay between the TF's intrinsic sequence preferences, cooperative interactions with other regulatory proteins, and cell type-specific chromatin landscapes. Cell type-specific TF binding events are highly correlated with patterns of chromatin accessibility and active histone modifications in the same cell type. However, since concurrent chromatin may itself be a consequence of TF binding, chromatin landscapes measured prior to TF activation provide more useful insights into how cell type-specific TF binding events became established in the first place. Here, we review the various sequence and chromatin determinants of cell type-specific TF binding specificity. We identify the current challenges and opportunities associated with computational approaches to characterizing, imputing, and predicting cell type-specific TF binding patterns. We further focus on studies that characterize TF binding in dynamic regulatory settings, and we discuss how these studies are leading to a more complex and nuanced understanding of dynamic protein-DNA binding activities. We propose that TF binding activities at individual sites can be viewed along a two-dimensional continuum of local sequence and chromatin context. Under this view, cell type-specific TF binding activities may result from either strongly favorable sequence features or strongly favorable chromatin context.

1. Introduction

Sequence-specific transcription factors (TFs) are the primary drivers of the transcriptional regulatory networks underlying cellular phenotype and behavior. TFs bind to their cognate DNA sequence motifs at particular locations and promote or repress the activities of other TFs, co-factors, chromatin modifiers, and the transcriptional machinery. These regulatory activities are performed at sites both proximal and distal to transcription start sites (TSSs). The primacy of TFs in directing cellular identity is confirmed by numerous transdifferentiation studies, where the expression of particular TF combinations is sufficient to override existing chromatin states and to establish new transcriptional programs [1–3].

Different cohorts of TFs are active in each cell type, and their combinatorial activities define cell-specific gene expression programs. However, the number of TFs encoded by animal genomes is limited (e.g. approx. 1600 in human [4]), and there is a large diversity of cell types to specify. Each individual TF is therefore often reused in multiple distinct cell types and developmental stages. For example, SOX2 is a master regulator of pluripotent embryonic stem cells [5], but is also required throughout neural tube development [6]. Similarly, ISL1 is a key regulator in the development of progenitor motor neurons [7], forebrain cholinergic neurons [8], pancreatic islets [9], and heart cells [10]. Do TFs bind and regulate the same targets in each cell type, or do synergistic interactions with other regulators and the cell-specific chromatin environment lead to context-dependent regulatory activities?

Sequencing-based TF-DNA mapping techniques such as ChIP-seq [11], ChIP-exo [12], and CUT&RUN [13] have by now been applied to hundreds of TFs in order to determine their genome-wide binding profiles in various cell types and cellular conditions [14–23]. Such assays are noisy and are applied to large cell populations, so they are typically thought of as measuring a relative enrichment of the target protein along the genome, as opposed to a binarized bound/unbound label. However, by applying appropriate differential binding analysis techniques [24–26], it is possible to compare ChIP-enrichment signals to identify binding locations where enrichment levels are similar across cell types, or vary significantly in a cell type-specific manner. In the vast majority of cases examined to date, TF binding signals vary substantially from cell type to cell type. For example, only 13% of MYC binding sites and just over half of CTCF binding sites were found to be bound at similar levels by their respective TFs in all of 11 tested human cell types [19].

Comparison of TF ChIP-seq experiments from the ENCODE project demonstrates that cell-specific binding sites are typically located distal to TSSs (Fig. 1). This is consistent with the view that cell-specific expression patterns are controlled by distal enhancer elements [27–30]. Indeed, the enhancer landscape is remarkably dynamic and cell-specific

* Corresponding author.
E-mail address: mahony@psu.edu (S. Mahony).

https://doi.org/10.1016/j.bbagrm.2019.194443
Received 30 June 2019; Received in revised form 21 September 2019; Accepted 6 October 2019
Available online 19 October 2019
1874-9399/

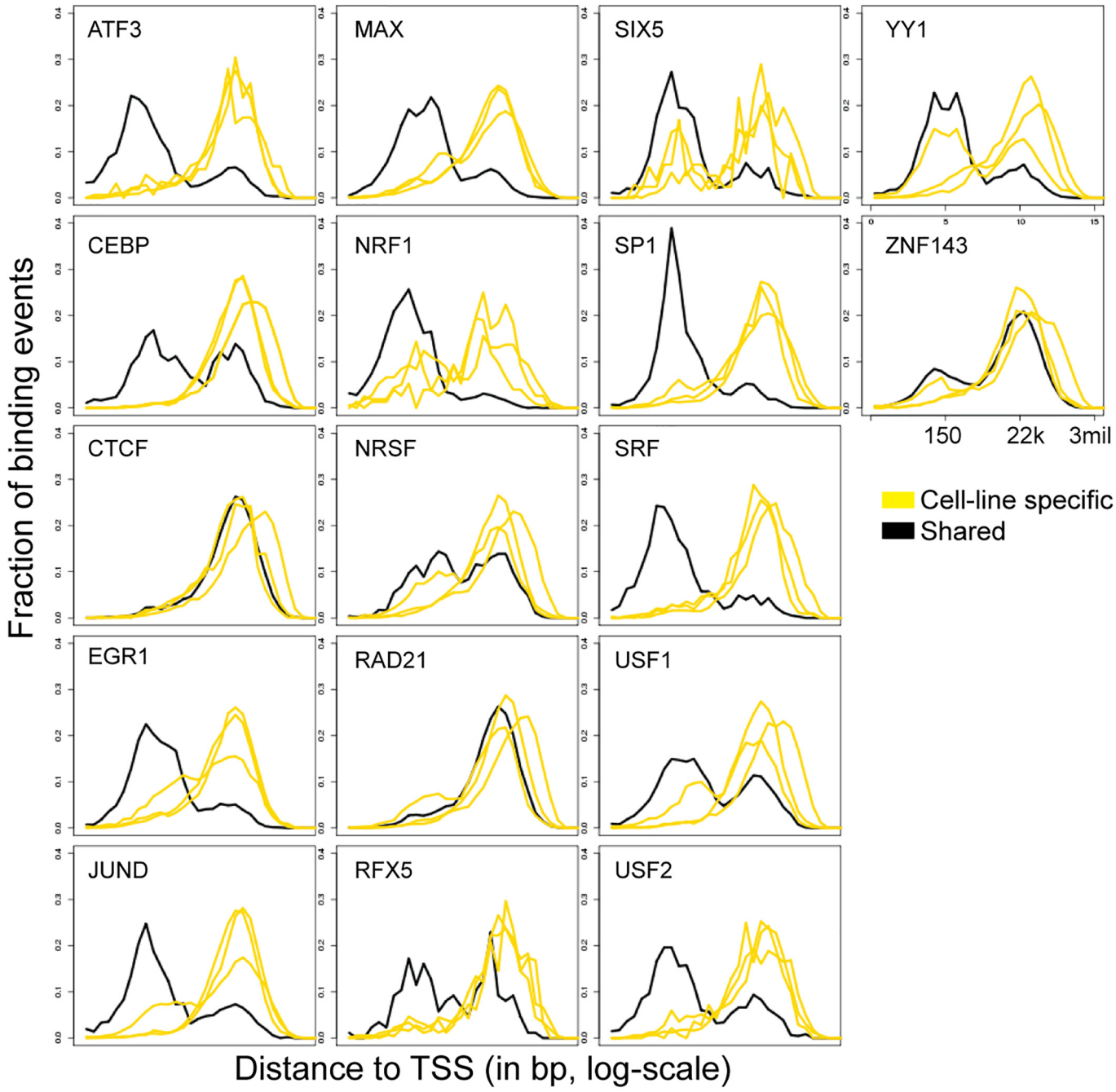

Fig. 1. Cell type-specific TF binding events are typically located at TSS-distal regions. Plots show distance distributions of TF binding events from annotated mRNA TSS for 17 TFs where differential binding analysis was performed on ENCODE ChIP-seq data from K562, GM12878, and H1-hESC cell types. Distributions are stratified based on shared (black) or cell type-specific (gold). The X-axis represents the distance in base pairs plotted according to a log-scale (natural logarithm). Adapted from Figure S3 in reference [33] under the CC BY license.

[29,31,32]. We note that we follow the convention of referring to most TSS-distal TF-bound sites as “enhancers”, although recent results suggest that only a subset of such regions can drive transcription in reporter assays [29].

How do cell type-specific TF binding patterns arise? When selecting binding sites within a given cell type, a TF integrates its intrinsic DNA-binding preferences, interactions with other regulatory proteins active in the same cell type (including other TFs, co-factors, and chromatin modifiers), and more general interactions with the established chromatin landscape. This review examines how a combination of sequence and chromatin features can determine cell type-specific TF binding patterns at individual sites. We begin by focusing on the defining determinant of TF binding specificity - the TF's intrinsic DNA binding preference - before discussing how interactions with other sequence-specific TFs can modify *in vivo* binding patterns. We then describe the chromatin landscape at cell type-specific enhancer elements, and we review computational advances in imputing TF binding patterns from sequence information and concurrent chromatin profiles. Finally, we turn to the causal determinants of cell type-specific TF binding patterns. While concurrent chromatin profiles are predictive of TF binding patterns, they provide limited information with respect to causal determinants of binding, as they are confounded by the chromatin-modifying outcomes of the TF itself. We therefore summarize our current understanding of how TFs find novel cell type-specific binding sites

when they are introduced into a new cellular context in dynamic regulatory settings. As we will discuss, gaps in our current understanding necessarily limit our ability to predict where a TF *would* bind if it were introduced into a characterized chromatin environment. However, progress in pursuit of this predictive goal is necessary in order to reach a mechanistic understanding of how cell type-specific gene regulatory networks are established.

2. Intrinsic DNA-binding preferences incompletely specify TF binding

TFs are characterized by their ability to recognize and bind to specific DNA patterns. In eukaryotes, most TFs bind to short sequences (6-20bp) and do so degenerately (i.e. they bind a range of sequences with similar binding energies) [34,35]. TF binding motifs are most commonly represented using frequency matrices (or derived representations) that record the occurrence of each nucleotide at each position in an alignment of observed binding sequences [34,36]. Advantages of frequency matrix representations include their intuitiveness and the fact that they can be defined using very few observed binding sites. However, they assume that nucleotide positions contribute independently to binding affinity, and thus likely over-simplify protein-DNA interactions [37–42]. Indeed, it has long been recognized that frequency matrix-based motifs are poor predictors of *in vivo* TF-DNA binding, at least in animal genomes. In a phenomenon dubbed the "futility theorem", unbound instances of a TF's cognate motif will typically greatly outnumber those that are occupied by the TF [43].

Over the past decade, the characterization of intrinsic TF-DNA binding preferences has seen significant advances on several fronts (reviewed in [4,44,45]). Many of these advances have been driven by protein binding microarrays [39,46,47], high-throughput SELEX [40,48], and related assays that can comprehensively measure a TF's *in vitro* binding affinity to very large numbers of sequences. These assays have been systematically applied to characterize the intrinsic binding preferences of numerous TFs [39,40], and the resulting motifs are stored in comprehensive databases [49–52]. For example, DNA-binding motifs have now been experimentally characterized for a majority of human TFs [4].

The availability of comprehensive *in vitro* experimental data has enabled greater insights into intrinsic TF-DNA binding mechanisms. Flanking DNA shape features [47,53,54], and higher-order dependencies between nucleotides within the motif [41], can explain dramatic binding affinity differences to sequences that score similarly using frequency matrix motifs. TFs can also have specific affinities for sequences that are not immediately apparent from their overall binding motifs. For example, paralogous DNA binding domains can have similar binding motifs overall, but can differ greatly in their affinities to a subset of specific length *k* (*k*-mer) sequences [39,55,56]. TFs can also display binding preferences that don't neatly fit into a single motif model; distinct binding modes can be due to the TF containing multiple DNA binding domains or binding in multiple multi-meric configurations [44]. In addition, TFs can also have higher or lower affinities to methylated DNA sequences [57].

More complex and subtle characteristics of TF binding preference can be captured by higher-order computational models, including models that account for dependencies between base positions [41,58–60] and *k*-mer based models trained using machine learning approaches such as support vector regression [61,62]. Several higher-order models were systematically evaluated during the DREAM5 TF-DNA Motif Recognition Challenge, where the goal was to represent TF-DNA affinity data from protein-binding microarrays [63]. This project conclusively demonstrated the advantages of *k*-mer models and position-specific models trained using biophysical energy-based frameworks [64,65] in representing intrinsic TF-DNA binding preference. Since then, some notable developments have included the development of *k*-mer based support vector machine approaches [20,66,67], innovations in gapped *k*-mer feature representations [68,69], and the application of convolutional and recurrent neural network architectures [70,71]. TF-DNA binding motif representations are also moving beyond standard DNA base features by incorporating preferences for and against chemically modified bases [72,73], and by incorporating DNA shape features [74–77]. The latter approach provides a dimensionality reduction of *k*-mer sequence space that improves binding mechanism interpretations. However, since DNA shape features are derived from sequence *k*-mers, a higher-order *k*-mer model trained with sufficient data will perform at least as well as simple models that incorporate DNA shape [77].

Importantly, higher-order models that are trained using comprehensive *in vitro* binding datasets more accurately predict *in vivo* TF binding sites [63,70]. However, the futility theorem still holds to a large degree; intrinsic binding preferences can't explain the cell-specific genomic occupancy of a given TF (e.g. [78]). The binding motifs of individual TFs do not typically vary depending on cell type or conditions. Thus, in cases where differential TF binding has been characterized, it is common to see identical copies of the TF's cognate motif being differentially occupied across cell types. In order to explain and predict cell-specific TF binding *in vivo*, we will therefore have to examine how TFs cooperate with one another and how they interact with cell-specific chromatin environments.

3. Cooperativity between TFs

Combinatorial interactions between TFs have long been recognized to increase the robustness and specificity of regulatory systems [79,80]. In terms of increasing or modifying TF-DNA binding specificity, we can consider two broad forms of cooperative interactions: direct cooperativity, where TFs form multimeric complexes via protein-protein interactions [81]; and indirect cooperativity, where TFs assist each other's binding by modifying local chromatin environments [82–84] (Fig. 2).

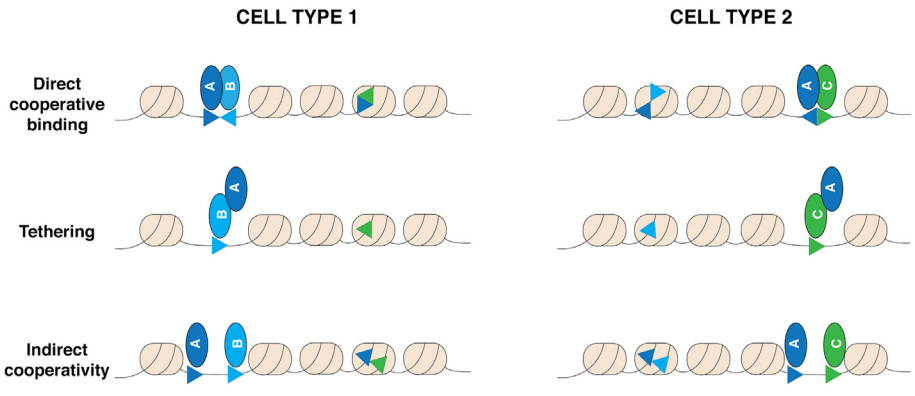

Fig. 2. Modes of cooperative binding for the hypothetical TF A. Direct interactions between TFs may result in the formation of cell type-specific heterodimeric complexes that bind to alternative motifs. The heterodimeric complex of A & B in cell type 1 may therefore bind to distinct sites compared with the heterodimeric complex of A & C in cell type 2. Similarly, tethering interactions may lead to cell type-specific binding patterns even in the absence of A's motif at alternate binding targets. Indirect cooperativity mechanisms (e.g. enhanceosomes or enhancer billboards) may also lead to differential binding of A across cell types by cooperative exclusion of nucleosomes.

Examples of direct cooperativity amongst distinct TFs include the AP1 heterodimer (c-Fos:c-Jun) [85] and the p50:RelA NFKB complex [86]. The formation of heteromeric complexes enables TFs to bind longer motifs with specificities that would not be achievable by individual constituents of the complex. Relatedly, the members of the complex can stabilize each other's binding to DNA to enable longer residence times. For example, OCT4 and SOX2 act in complex as a master regulator of pluripotency, and assist each other's binding to a motif that concatenates the binding preferences of the constituent POU/Homeodomain (OCT4) and HMG (SOX2) family TFs [87,88]. It follows that TFs can change their binding preferences by changing binding partners. For instance, the Retinoid X Receptor (RXR) can heterodimerize with several other nuclear hormone receptors, including RAR, VDR, TR, LXR, PPAR, and FXR. Heterodimers containing RXR bind to direct repeats of the hormone response element motif (HRE: consensus 5′-TGACCT-3′), albeit with variable spacing between half-sites. RAR:RXR prefers direct repeats with 1bp, 2bp, or 5bp spacers [89–91], whereas RXR:VDR prefers direct repeats with a 3bp spacer [92]. Another key example of direct cooperativity is provided by interactions between *Drosophila* Hox TFs and the cofactor Exd. Slattery et al. have demonstrated that interactions with Exd unlock distinct "latent" DNA binding specificities in each of the Hox factors [93], thus partly explaining the longstanding paradox that paralogous Hox factors with similar intrinsic DNA-binding preferences can have drastically different regulatory activities.

Cell type-specific TF binding could thus be explained by the availability of alternate direct binding partners. We previously demonstrated this principle in a system where we can transdifferentiate mouse embryoid body (mEB) cells into motor neurons by over-expressing combinations of TFs [94]. In our system, the TF combination of NEUROG2, ISL1, and LHX3 induces rapid and efficient conversion of mEB cells into spinal motor neurons. Replacing LHX3 with PHOX2A instead leads to the specification of cranial motor neurons. ChIP-seq revealed that ISL1 was alternatively partnering with LHX3 or PHOX2A in these alternate programming systems. The ISL1-LHX3 and ISL1-PHOX2A complexes appear to bind to distinct motif configurations. Even though they were expressed in the same chromatin context (mEB cells), the shift in binding partners leads to widespread differences in the genomic loci that are bound by ISL1, resulting in the establishment of distinct expression programs and ultimately divergent cell fates.

The recognition of higher complexity motifs by directly interacting TFs may be an even more prevalent regulatory strategy than currently assumed. A high-throughput *in vitro* investigation of 9,400 TF-TF pairs revealed 315 cooperative binding interactions at heterodimeric motifs, many of which were novel [95]. The heterodimeric motifs identified were often distinct from what should have been expected from the individual TF's *in vitro* binding preferences. Interestingly, some of these TF-TF interactions appeared to be facilitated by DNA; they were too weak to be stable in the absence of the binding site. This DNA-facilitated mode of binding is complementary to the model that TF-TF interactions can facilitate binding of TFs to weaker instances of cognate motifs [81].

Another consequence of direct protein-protein interactions between TFs is that it's possible for a TF to be recruited by binding partners in the absence of the TF's cognate binding motif [21,96–98]. This process is known as "tethering", and may explain the common observation that a large proportion of TF ChIP-seq peaks do not contain recognizable instances of the TF's cognate binding motif [99]. We and others have demonstrated that the high-resolution ChIP-exo assay can clarify which sites are bound directly vs. tethered [100,101]. In ChIP-exo, a 5′ to 3′ exonuclease is used to digest immunoprecipitated DNA fragments to sites of protein-DNA crosslinking [12]. ChIP-exo thereby more precisely locates protein-DNA binding events to the bound nucleotides, and displays distinct crosslinking patterns depending on the identities of the TFs mediating direct interactions with DNA. For example, we recently used a novel ChIP-exo analysis platform to characterize estrogen receptor alpha (ERα) binding events in MCF7 cells [101]. While most ERα binding events are associated with the cognate nuclear hormone receptor motif, a subset are centered on motifs and ChIP-exo crosslinking patterns that are characteristic of Forkhead domain TFs. Subsequent ChIP-exo experiments found coincident binding of Foxa1 at precisely the same locations, suggesting that ERα is tethered to Foxa1 at these sites.

Moving beyond direct protein-protein interactions between TFs, more subtle forms of indirect cooperativity can also explain cell-specific TF binding patterns. Indirect cooperativity can take the form of collaborative competition with nucleosomes. Under this mass action model, a locus can be maintained in a nucleosome-free state via the competitive binding of multiple TFs, regardless of whether these TFs directly interact with one another [83,84,102–105]. Another mode of indirect cooperativity is provided by DNA-mediated cooperativity, where the binding of a TF to a locus can stabilize DNA, or change its structural shape properties, such that the binding of additional TFs is promoted [82]. Indeed, a classic example of DNA-mediated cooperativity may be provided by the IFNβ enhancer [106,107], which contains a highly constrained array of DNA motifs for several TFs. Modifying the spacing or orientation of these motifs can disrupt the formation of the enhanceosome complex. However, structural analysis reveals few protein-protein interactions between TFs; instead, it appears that enhanceosome stability emerges from TF-induced structural changes to DNA and interactions with co-factors.

All of the modes of cooperativity outlined above are mutually compatible. We can therefore consider the sequence properties of cell type-specific binding sites in the light of cooperative binding mechanisms. Cell type-specific TF binding sites generally contain weaker instances of the TF's cognate motif [26,33,108], consistent with both direct and indirect assisted binding mechanisms. They also contain an enrichment of motifs for other TFs that are active in the same cell type, again consistent with multiple modes of cooperativity [20,33,108]. Indeed, multi-class and multi-label classification approaches typically find that the best discriminative sequence features of cell-specific TF binding sites are related to the motifs of other regulators that are active in the same cell type [20,33,109]. Cell type-specific enhancers contain clusters of TF binding sites [99,110–112]. The organization of motifs within enhancer regions can be highly constrained, like the IFNβ enhancer. For example, the regulatory region upstream of major histocompatibility complex class II genes (MHC-II) contains a series of four DNA motif elements with rigid ordering and spacing constraints [113,114]. Indeed, spacing and ordering constraints have been detected between TF-bound motifs across a broad range of *in vivo* binding sites [115,116]. However, it is far more common to see no constraints (or at least no frequently recurring constraints) in the spacing, ordering, or orientation of motifs for different TFs, consistent with a "billboard" model of enhancer organization [117].

4. Concurrent chromatin landscapes of cell-specific TF-DNA binding events

While the sequence motif features at cell type-specific TF-DNA binding events will vary greatly depending on the cohort of TFs that are active in a given cell type, the general chromatin features surrounding TF binding sites are more consistent across cell types. In this section, we summarize the major chromatin features that are known to co-occur with cell-specific TF binding events at distal enhancers (reviewed in more depth in [118–121]). Naturally, many of these correlated chromatin features can be used to accurately predict the locations of cell-specific TF-DNA binding events within the same cell type (i.e. "concurrent" TF binding). However, it's important to note that the following chromatin features are not necessarily causally associated with the establishment of cell type-specific TF binding sites. Chromatin features at cell type-specific enhancers can be as much the product of cell type-specific TF binding events as binding determinants.

Cell type-specific distal enhancers, like other regulatory elements (e.g. promoters), are typically DNase I hypersensitive [122] and display high nucleosome turnover [123]. These features may be due to the exclusion of nucleosomes from enhancers by competing TF binding events and/or may be a consequence of the enrichment of nucleosomes containing the unstable H2A.Z and H3.3 histone variants [124]. The nucleosomes flanking enhancer regions typically display enrichment of the histone modification H3K4me1 [125]. H3K4me1 is specifically recognized by the TIP60/p400 complex [126], which catalyzes deposition of H2A.Z [127], and the ATP-dependent chromatin remodeler CHD7 [128,129]. Another role of H3K4me1 may be to block binding of the DNA methyltransferase cofactor DNMT3L, which specifically binds to unmodified H3K4 [130]. Indeed, most enhancers have low or intermediate levels of DNA methylation [131,132]. The depletion of 5mC at enhancers (and particularly within TF binding sites) is mirrored by an enrichment of 5-hydroxymethyl-cytosine (5hmC) [133], which is the result of active 5mC demethylation by Tet family proteins [134].

While enhancer-flanking nucleosomes display enrichment of H3K4me1 and H3K4me2, the H3K4me3 mark is relatively depleted. Indeed, the ratio of H3K4me1 and H3K4me3 ChIP enrichment was used as one of the first genome-wide predictors of enhancer potential, as H3K4me3 is more specifically enriched at TSSs [125]. Why do H3K4 residues become monomethylated but not trimethylated at enhancers? H3K4me3 is deposited by the methyltransferases SET1A and SET1B [135], which are recruited to serine-5-phosphorylated Pol II CTDs [136]. In contrast, H3K4me1 is deposited by MLL3 and MLL4 methyltransferases [137,138], which may be recruited to specific enhancer regions via interactions with TFs [139,140].

Distal enhancer elements can be subdivided according to their regulatory output; only a subset of enhancers appears to drive expression when tested in reporter assays [29]. Several histone acetylation features are correlated with enhancer activity, particularly H3K27ac [141–143], but also H3K9ac, H3K18ac and H3K14ac [144]. These marks are dependent on histone acetyltransferases (HATs); for example, p300 & CBP catalyze H3K27ac, while GCN5 & PCAF catalyze H3K9ac [145]. There is some evidence that particular TFs favor binding to regions that are marked with particular histone modification combinations, although it's not known how generalizable these preferences are across cell types [146–148]. Another key feature of active enhancers is the bidirectional production of short non-polyadenylated enhancer RNAs (eRNAs) [29,149–154]. Given that transcription is occurring at enhancers, it should not be surprising to observe components of the transcriptional machinery too. Pol II itself is enriched at enhancers [143], as are TBP and several other GTFs [150]. Some "orphan TAFs" may even associate with enhancers in a cell type-specific manner (e.g. TAF3 in ES cells [155] and TAF7L in adipocytes [156]). Finally, it is thought that active enhancers are brought into physical proximity with their target promoters via chromatin looping. These loops appear to be stabilized by the mediator and cohesin complexes, which display high enrichment levels at both enhancers and promoters [157].

Enhancers that do not drive expression in a given cell type appear instead to be "poised" for future activity [141,142]. Some poised enhancers appear to be directly targeted by repressive protein complexes such as the Polycomb-group proteins [142,143] and display enrichment of repression-associated histone modifications H3K27me3 or H3K9me3 [143]. Other so-called "latent" enhancers have H3K4me1 enrichment and other general enhancer features, but no obvious markers of repression or transcriptional activity. A minority of these enhancers may be in the process of enhancer decommissioning; H3K4me1 tends to persist after the loss of enhancer activation potential [158], and demethylases like LSD1-NuRD are required to complete the decommissioning process [159]. However, many more latent enhancers are truly poised for future activation in a particular cellular context, as evidenced by observations of enhancers that transition from latent to active states in developmental lineages [32,160]. Note that poised enhancers contain similar numbers of cell-specific TF binding events as active enhancers. Indeed, cell-specific TFs may be maintaining some enhancers in a poised state by recruiting histone deacetylases (HDACs). A relevant study along these lines demonstrated that constitutively bound retinoic acid receptors recruit HDAC1/2/3 to latent enhancers in the absence of the retinoic acid ligand [161]. These HDACs are lost after retinoic acid exposure, and H3K27ac is subsequently deposited. Such observations further emphasize the dynamic nature of the chromatin landscape in regulatory systems, and the fact the many of the chromatin features that are associated with TF binding sites might only arise after TF binding activities have occurred.

5. Imputation of cell-specific TF binding events using concurrent chromatin

Intrinsic sequence preferences are poor predictors of cell type-specific TF binding. Instead, the genome-wide profiling of a large number of TFs has demonstrated that TF-DNA occupancy is highly correlated with cell type-specific chromatin accessibility [99,131,162]. A question that then arises is whether cell type-specific chromatin features can help us to predict the locations of cell type-specific TF binding events. In this section, we review computational methods that have been developed to predict cell type-specific TF binding signals using DNA sequence and concurrent chromatin features from the same cell type [163–171]. We again emphasize that chromatin data tracks can indirectly encode TF binding in the same cell type, so the goal of most approaches is to infer or impute unobserved TF binding data as opposed to modeling the DNA-binding activities of a TF in a completely novel biological context. For example, many approaches in this class rely on access to cell type-specific data characterizing chromatin features that are associated with TF binding, and data characterizing the intrinsic DNA-binding preferences of a TF. By combining these data sources, the approaches aim to predict where TF binding is occurring in the same cell type. Conversely, the same methods cannot offer accurate TF binding predictions for cell types in which the chromatin features have not been characterized. Nevertheless, developing methods that accurately solve the concurrent TF binding imputation problem would reduce our reliance on ChIP-based approaches, enabling us to study regulatory mechanisms for TFs and cell types where ChIP-seq is not practical or cost effective.

Methods that aim to impute TF binding can be split into two broad categories. The first category aims to discover TF binding events in a cell type by using TF "footprints" present in chromatin accessibility data from that same cell type [168,169,172,173]. In particular, TFs can protect short DNA segments at their binding sites from cleavage by enzymes such as DNase I and Tn5, resulting in characteristic read depletion patterns around bound motifs in DNase-seq and ATAC-seq experiments. Such footprints in accessibility data have been integrated with sequence information through approaches such as Bayesian mixture models to impute TF binding [167,172,174]. While footprinting is a promising approach to impute binding using only cell-specific accessibility, DNase I and Tn5 cleavage patterns can suffer from significant enzyme-related sequence biases [175]. Therefore, computational models are faced with the challenge of incorporating appropriate bias-correction strategies in order to accurately impute TF binding [176–178]. Furthermore, some suggest that many TFs may not leave detectable cleavage footprints [173,177,179], and therefore the usefulness of footprinting approaches may be limited.

Alternatively, a second category of methods aim to leverage TF binding data from one set of cell types in order to impute TF binding in a new cell type where only the general chromatin features have been profiled [20,163–166,180]. Notably, several methods integrate cross-cell type information with TF footprints to impute TF binding in target cell types. For example, FactorNet uses a deep convolutional neural network to integrate DNA sequence and single nucleotide resolution chromatin accessibility from one or multiple cell types, and the trained network is applied to predict TF binding in a new cell type by using

chromatin accessibility from that cell type as a feature [163,181]. Grau et al. describe Catchit, an iterative training procedure that uses *in vitro* sequence models and summarized DNase-seq read counts to predict TF binding in new cell types [165]. Catchit demonstrates that iteratively including model false positives in the negative training set improves the ability of models to predict TF binding both within the same cell type as well as in new cell types.

Binding models that are trained in one set of cell types in order to predict binding in a different cell type are vulnerable to overfitting; i.e. they tend to learn TF binding related features in training cell types that are not transferable to new cell types. This is especially problematic when differential cofactor interactions influence or correlate with *in vivo* TF sequence preferences in a cell type-dependent fashion [20]. Cell type-specific overfitting is evidenced by the observation that the *in vitro* or intrinsic binding preferences of TFs are often better predictors of TF binding in a new cell type when compared to *in vivo* models derived from one or many cell types [165]. Several techniques have been proposed to overcome the lack of model transferability across unrelated cell types. For example, Anchor, a gradient-boosted tree model for TF binding imputation, attempts to minimize cell type-specific overfitting by using a "criss-cross" validation strategy [164]. Specifically, Anchor trains models in one cell type while validating on a second cell type for which TF binding data is also available. Such a criss-cross validation strategy promotes the learning of generic TF binding features that may be more transferable to new cell types [164]. Alternatively, explicitly computing the similarity between cell types enables the prioritization of binding models that are more likely to impute binding in the target cell types with higher accuracies [166,180]. MOCAP computes the similarity between TF ChIP-seq experiments by comparing genomic features such as chromatin accessibility and CpG islands at TF binding sites. A binding model trained on the most similar TF ChIP-seq experiment is then used for TF imputation in a new cell type [180]. TFBSImpute uses ChIP-seq data to measure the correlation structure between various TF binding signals in the same cell types and between cell types. Missing TF binding data tracks are then imputed through a tensor completion framework [182]. Finally, in an orthogonal approach, Virtual ChIP-seq integrates correlations between chromatin accessibility, TF binding, and gene expression in order to impute TF binding in new cell types using a multi-layer perceptron [183].

The ENCODE DREAM "*in vivo* binding prediction challenge" has recently motivated a standardized evaluation of TF binding imputation accuracies across cell types. While significant progress has been made in imputing binding of TFs in new cell types, there still remains a performance gap between within cell type and cross-cell type imputation models. Some TFs such as CTCF can be imputed in new cell types with high-accuracy. However, for the majority of TFs, even complex models that encompass sequence, chromatin accessibility and expression data are only able to achieve an average true positive rate of approximately 0.5 at a 50% false discovery rate [163–165]. For TFs such as REST and NANOG, our ability to predict binding across cell types is even more limited [163–165]. Further investigation into both the successes and failures of models that impute TF binding will lead to insights into cell type-specific TF binding mechanisms, in turn motivating the development of more accurate imputation models.

6. Chromatin predeterminants of TF binding: pioneers and settlers of nucleosome territories

As we have seen, TF binding events can be imputed using sequence features and concurrent chromatin landscapes that have been collected in the same cell types. However, this does not address the question of whether chromatin environments causally affect TF binding. When a TF encounters a new chromatin environment, how does it determine its binding locations?

The nucleosomal landscape is expected to be a major determinant of TF binding; the steric hindrance created by the interaction of DNA with histone core proteins will exclude most TFs from nucleosomal DNA [83,184,185]. This view was strongly supported by a recent comprehensive exploration of *in vitro* interactions between nucleosomes and 220 human TFs [186]. Zhu et al. confirmed that nucleosomes inhibit the DNA-binding abilities of most TFs, although some such as EN1 and Sox HMG family TFs appear to have the ability to bind nucleosomal DNA directly. Accessing nucleosomal DNA also appears to depend on the position and orientation of cognate motifs on the nucleosome; many TFs can bind to motifs that are on the edges of nucleosomes, while TFs that can access nucleosomal DNA display positional motif preferences along the nucleosome [186,187]. The *in vivo* binding of OCT4, SOX2, and KLF4 to nucleosomal DNA during pluripotency reprogramming has also been observed to display positional dependencies [188]. A recent high-throughput *in vivo* study using yeast that contained systematically modified versions of the nucleosome-occupied HO promoter came to several related conclusions [189]. In that study, inserting certain motifs in single copy into the HO promoter led to nucleosome depletion, but this was true for cognate motifs of only a small subset of yeast TFs. However, several more TFs appeared to be able to deplete nucleosomes if multiple copies of the motif are present within the target locus [189].

In developmental contexts, the TFs that first establish chromatin accessibility at enhancers are termed "pioneer factors" [185,190–192]. The first proposed pioneer factors were FOXA and GATA family TFs, which were recognized to bind in early endodermal development to a chromatin-compacted liver-specific enhancer of the *Alb1* gene [193]. FOXA and GATA binding creates local accessibility, making the enhancer competent for binding by liver-specific activating TFs [194]. At this point, several other TFs have been identified as lineage-specific pioneer TFs, primarily from analysis of temporal TF binding and enhancer accessibility in developmental timecourses. For example, PU.1 in macrophage and B cell lineages [18,195,196], Pax7 in a pituitary lineage [197], ASCL1 in neuronal lineages [147,198], and the pluripotency factors OCT4, SOX2, and KLF4 when induced in fibroblasts [199]. The mechanisms by which pioneer TFs establish chromatin accessibility may be varied. While some may directly bind nucleosomal DNA and exclude histones, others recruit ATP-dependent chromatin remodelers in order to evict nucleosomes [192]. PHA-4, orthologous to the mammalian pioneer FOXA TFs, binds inaccessible promoters in *Caenorhabditis elegans* and recruits RNA Pol II prior to chromatin opening, leading to the hypothesis that transcription may play a role in pioneer TF facilitated chromatin decompaction [200]. Alternatively, pioneer TFs can passively mark silent chromatin as competent for gene activation by enabling co-operative TF binding-mediated nucleosome depletion [185].

The establishment of enhancer accessibility by pioneer TFs has been proposed to be temporally followed by the binding of "settler" and "migrant" TFs [167]. These terms specify overlapping TF binding behaviors; settler TFs are defined to be dependent on chromatin accessibility at their binding sites, while migrant TFs are defined to be dependent on both chromatin accessibility and specific interactions with other regulators [167]. Such dependencies are readily seen in developmental contexts. For example, the pioneer factor PU.1 establishes accessibility and H3K4me1 histone modifications at enhancers in macrophage and B cell lineages, enabling the downstream binding of LXR TFs [18]. Indeed, the preexisting chromatin accessibility and regulatory landscape is known to have a strong influence on determining the binding locations of many TFs. Signaling responsive TFs, including nuclear hormone receptors such as estrogen receptor alpha (ERα) and glucocorticoid receptor (GR), provide convenient models for studying TF binding determinants, as their binding activities can be induced by ligand exposure and assayed in cell types with characterized chromatin environments [17,91,108]. For example, ERα targets distinct sites when DNA binding is ligand-induced in the endometrial cancer cell line ECC-1 versus the breast cancer cell line T-47D [201], and these differentially bound sites are associated with preexisting cell type-specific chromatin accessibility [108]. Consistent with TF binding cooperativity

mechanisms, cell type-specific ERα binding events on average contain weaker cognate binding motifs but are enriched for motifs associated with regulators that were already active in the corresponding cell types (e.g. FOXA and GATA motifs in T-47D cells). Similarly, GR binds to largely different sites when induced via glucocorticoid in mouse mammary and pituitary cell types, but again these cell type-specific sites appear to be predetermined by cell type-specific accessibility patterns [17]. Other examples where signal-induced TF binding appears to be predetermined by the preexisting chromatin landscape include IFNG-induced STAT1 binding in HeLa cells [202], retinoid-induced RAR binding in early neurogenesis [91], and heat shock-induced HSF1 binding in *Drosophila* S2 cells [203].

7. From simple categories to complex interactions: switching roles between opportunists and influencers

The categorization of TFs as pioneers and settlers is appealingly intuitive: pioneers set the chromatin stage and thereby influence the binding of other TFs; meanwhile settler TFs bind opportunistically to sites that have favorable chromatin environments, even if their cognate motif instances are suboptimal. However, the introduction of any TF into a new chromatin environment will lead to complex interactions between the TF, DNA sequences, preexisting chromatin features, and other regulators. These interactions will lead to binding outcomes at individual sites that don't always fit neat TF categories.

Firstly, despite their relative ability to bind to sites that were previously inaccessible, pioneer factors typically bind to distinct target sites in different cell types. The classic pioneer TF FOXA1 binds differential subsets of its target sites in MCF-7 and LNCaP cells, and these differences are correlated with cell type-specific H3K4me1/2 enrichment [204]. Further, overexpression of the H3K4me1/2 demethylase KDM1 in MCF-7 cells impairs FOXA1 binding at some of its target genomic loci, confirming that H3K4me1/2 (or at least a tangentially related determinant of accessibility) is required for FOXA1 binding *in vivo* [204]. Cell type-specific FOXA1 and FOXA2 binding has also been linked to preexisting DNA hypomethylation [205], prior binding of related Forkhead domain TFs in early developmental cell types [206], and priming by low levels of active histone modifications [207]. Similarly, OCT4 binding in embryonic stem (ES) cells is modified by H2A.Z knockdown [208]. The overexpression of CDX2 in mouse developmental cell types [26], and FOXA2, GATA4, and OCT4 in human cell lines [209] leads to binding at some previously inaccessible chromatin sites, as expected of pioneer factors. But each induced TF also binds previously accessible sites in a cell-specific manner, and such cell-specific sites contain weaker cognate motif instances [26]. Thus, pioneer TFs also bind opportunistically to sites of favorable chromatin.

Conversely, signaling-dependent TFs can act as pioneers and influence the binding of other TFs at a subset of their binding sites. For example, a subset of GR binding sites are nucleosomal before GR binding is induced, and GR promotes accessibility at these sites by recruiting Brg1 [210]. When GR and NFKB are activated alongside one another, they co-bind to novel sites that neither can access alone [211]. GR and ERα can even modify the binding of FOXA1 when they are induced in breast cancer cell lines [212]. Several other induced TFs that bind mostly in previously accessible chromatin have been observed to "pioneer" a subset of sites (e.g. [213]).

Taken together, we suggest that pioneer and settler TF categories are not binary; any induced TF might be able to act as a pioneer at a particular inaccessible site, given the right mix of sequence features, nucleosome configurations, and cooperative interactions with existing regulators. However, the relative abilities of different TFs to "pioneer" new sites clearly vary widely. Conversely, and perhaps more obviously, any induced TF can act as an opportunistic settler at sites displaying particularly favorable preexisting chromatin features, even if the sequence features at such sites are suboptimal. A large portion of cell type-specific binding may thus be explained by predetermined chromatin landscapes.

8. TF-induced chromatin dynamics reshape TF binding

In the previous sections, we considered the chromatin predeterminants of TF binding. Of course, regulatory systems are dynamic; any TFs introduced into a given chromatin environment will have regulatory effects that impact chromatin structure and the expression of other regulators. Aside from establishing chromatin accessibility via nucleosome depleting pioneering activities [214], TFs are also the prime recruiters of cofactors and chromatin-modifying enzymes that activate (e.g. [215,216]) or repress (e.g. [159]) enhancer and transcriptional activities.

We might expect that these TF-driven changes in chromatin structure in turn have an impact on the binding of TFs. Evidence for such dynamic interplay between TF binding and chromatin remodeling is clearest in transdifferentiation systems, where one or more "programming TFs" are ectopically introduced into defined cell types with the goal of bringing about a change in cellular identity. While numerous transdifferentiation systems have been described (reviewed in [217–222]), the dynamics of chromatin structure and programming TF binding have been characterized in relatively few. Nevertheless, some common themes are apparent in those transdifferentiation systems that have been examined at the level of TF binding activities.

Firstly, when overexpressed in defined chromatin environments, programming TFs can rapidly remodel chromatin landscapes at their binding sites in a process analogous to pioneering activity [147,198,223–225]. For example, the proneural bHLH TF ASCl1 can convert fibroblasts to neurons when expressed alone or alongside BRN2 and MYT1L [147,226]. ASCL1 can bind to many sites that are inaccessible in fibroblasts, and then rapidly promotes chromatin accessibility at its binding sites [226]. Similarly, the binding of OCT4, SOX2, and KLF4 converts large numbers of regions from inaccessible to accessible during the reprogramming of fibroblasts to induced pluripotent stem cells [199,227].

Secondly, chromatin landscapes that have been remodeled by programming TF binding can impact the binding locations of downstream TFs. We recently demonstrated that ASCL1 and NEUROG2 induce different neuronal subtype expression programs when expressed individually in mouse embryoid body (mEB) cells [198]. These proneural TFs bind to largely distinct sites (guided by distinct E-box motif preferences), but as with other examples, many of their binding sites are inaccessible in the starting cell type and rapidly acquire accessibility and enhancer-associated histone marks after TF binding. However, both ASCL1 and NEUROG2 induce expression of a common set of neuronal identity regulators, including BRN2, EBF2, and ONECUT2. The genomic binding of these downstream TFs is then differentially affected (to varying degrees) by the distinct chromatin landscapes established by ASCL1 or NEUROG2; e.g. up to 40% of BRN2's and EBF2's binding sites are differentially bound, depending on whether they became expressed downstream of ASCL1 or NEUROG2 [198]. This illustrates an epigenetic mechanism (in the Waddingtonian definition of the term [228]), whereby regulatory activities of TFs are modified by chromatin landscapes that were established by prior regulators.

Finally, the regulatory activities initiated by programming TFs can reshape the binding locations of the programming TFs themselves. Pluripotency factor binding undergoes extensive stepwise redistribution during reprogramming [199,227,229]. Early in the pluripotency reprogramming process, OCT4, SOX2, and KLF4 binding appears to recruit endogenous fibroblast expressed TFs (e.g. CEBPA & FRA1) away from their typical binding locations. But later in the process, as additional pluripotency-associated factors (e.g. NANOG & ESRRB) become active, the binding of the reprogramming factors themselves become redistributed to pluripotency-specific enhancers [227]. We have also seen related dynamics when expressing NEUROG2, ISL1, and LHX3 in mEB cells to rapidly program spinal motor neurons (Fig. 3) [230]. The

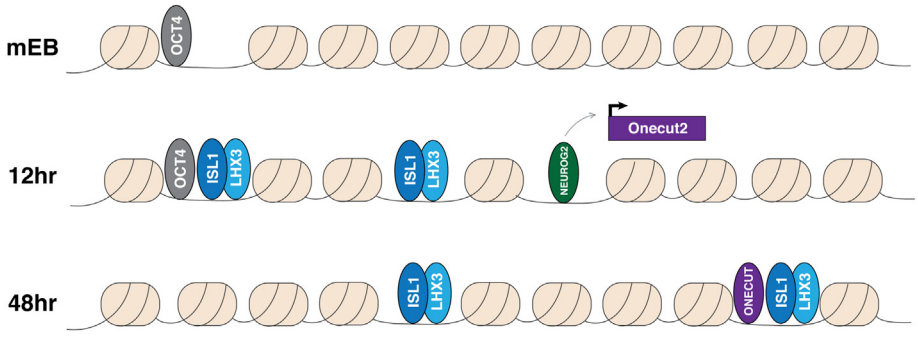

Fig. 3. Dynamics of programming TF during motor neuron programming. NEUROG2, ISL1, and LHX3 are expressed in mEB cells. ISL1:LHX3 can bind numerous sites that were inaccessible in mEB cells within 12 hours of expression, but can also bind opportunistically to sites that are pre-bound by pluripotency TFs (e.g. OCT4). NEUROG2 binds distinct sites and activates Onecut2 gene expression. By 48 hours, ONECUT2 pioneers additional sites and relocalizes ISL1:LHX3 binding.

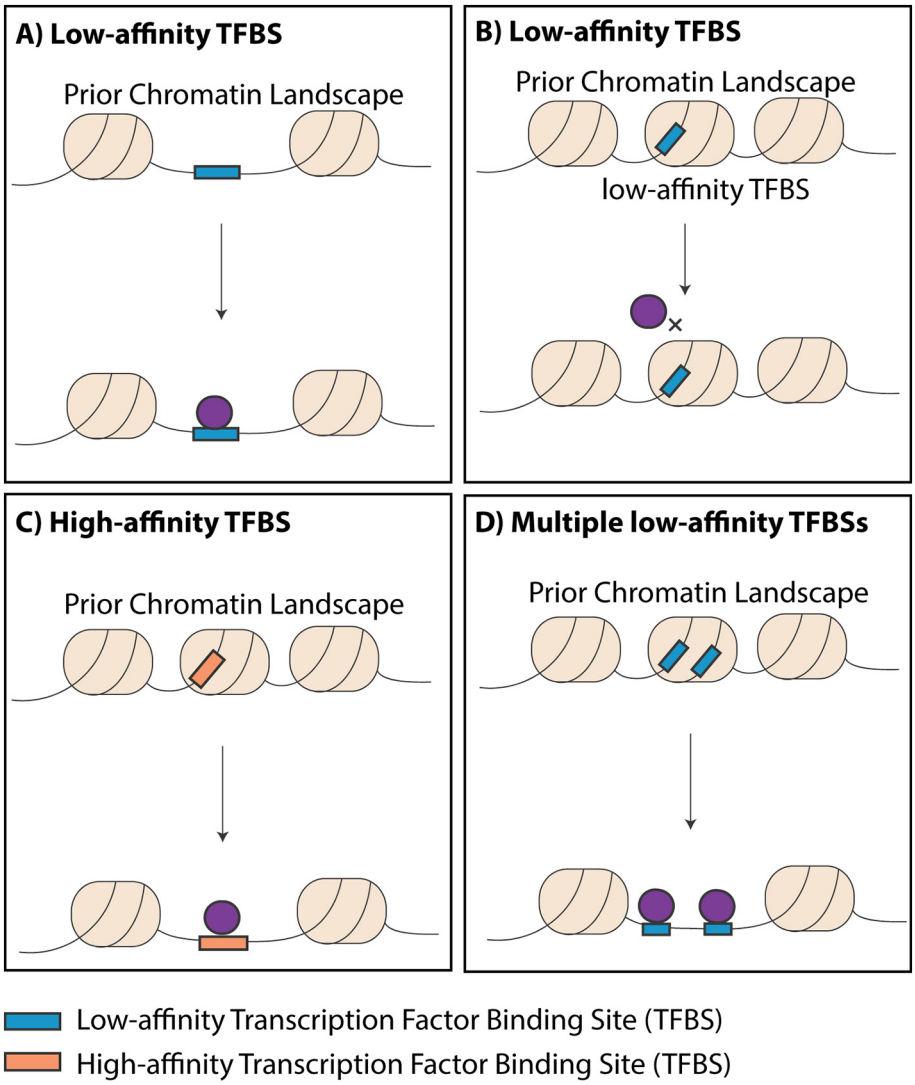

Fig. 4. The ability of TFs to bind potential cognate motifs is characterized by interactions between sequence features, pre-existing chromatin accessibility and other cell type-specific TFs. A) TFs can bind low-affinity cognate motifs in pre-accessible chromatin. B) TFs may not be able to bind low-affinity motifs in pre-inaccessible DNA. C) The presence of high-affinity TFBSs binding motifs support TF binding, even under nucleosomal constraints. D) The presence of multiple motif instances allows co-operative displacement of nucleosomes and promotes TF binding at pre-inaccessible chromatin.

ISL1/LHX3 heterodimer binds to, and rapidly promotes accessibility at, many previously inaccessible sites. But the potent pioneer factor ONECUT2 [225] becomes activated downstream of NEUROG2, leading to a subsequent shift in some ISL1/LHX3 binding towards the sites that ONECUT2 makes accessible. Interestingly, the sites that ISL1/LHX3 co-binds with ONECUT2 contain weaker cognate motif instances compared with sites that ISL1/LHX3 makes accessible itself, suggesting why these sites are dependent on the pioneering activity of ONECUT2.

In summary, the effect that the chromatin environment has on an induced TF's binding cannot be divorced from the dynamic effects the TF itself has on the regulatory system. This necessarily sets a time limit on the degree to which measurements of a given chromatin environment will accurately predict subsequent binding of induced TFs.

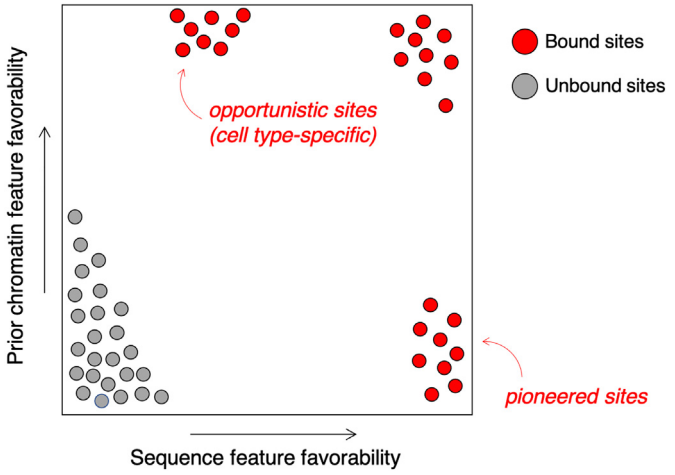

Fig. 5. Cartoon of a joint sequence and chromatin continuum for induced TF binding outcomes. Some sites with strongly favorable sequence features (e.g. high affinity cognate motif instances, or multiple motif copies) might be bound regardless of prior chromatin status. Pioneered sites have low prior chromatin favorability but high sequence favorability. Sites that are opportunistically bound in cell type-specific patterns may have high prior chromatin favorability (e.g. highly accessible DNA, binding of cooperative TFs), which compensates for lower sequence favorability.

9. Discussion

In each cell type, TF binding sites display characteristic sequence and chromatin features. The DNA features associated with a given TF's cell type-specific binding sites contain a mix of the TF's intrinsic binding preferences and motifs associated with other cooperating regulators active in the same cell type. Chromatin features include accessibility, histone modifications associated with enhancer activity, DNA hypomethylation, and transcription. The correlation between TF binding and these sequence and chromatin features has enabled the development of computational approaches that impute TF binding events using information from the same cell state in which TF binding is occurring. Recent advances have greatly improved the performance of methods on this concurrent TF binding imputation task. While there remain challenges in terms of improving false positive rates in a genome-wide setting, and in terms of building better cross-cell type transfer learning approaches, there are reasons to be optimistic that concurrent TF binding imputation is feasible.

However, solving the concurrent TF binding imputation challenge itself provides little insight into the causal mechanisms underlying TF binding specificity. Rather, it is useful to think of an alternative problem; can we predict where a TF *would* bind if it were introduced into a given cellular environment? What types of chromatin information would be most informative from the preexisting chromatin state? Will features such as prior chromatin accessibility and histone modifications be sufficient to enable TF binding predictions, or do we need to characterize more complex features such as the preexisting three dimensional structure of chromatin (e.g. [231])? What types of information might we need to collect about the intrinsic properties of the TF? The TF's DNA binding preferences [40] would certainly be informative - would we additionally require knowledge of its relative pioneering abilities [186] and how it interacts with other regulators on DNA [95]? Integrating these forms of information to predict induced TF binding would not only elucidate causal TF binding determinants, it might also have practical relevance in terms of predicting the regulatory activities of programming TFs in transdifferentiation recipes, or predicting how a given cell type would respond to external signaling events.

Therefore, in order to gain deeper understanding of TF binding determinants, we need to first profile how a wide range of TFs behave in dynamic settings where the chromatin landscape and co-factor repertoires can be assessed prior to TF induction. As described above, the *in vivo* prior chromatin determinants of TF binding have been investigated for several TFs in naturally inducible settings (e.g. nuclear hormone inducible ERα and GR). Analysis of the chromatin pre-determinants of such signalling-dependent TFs' binding demonstrates that cell type-specific binding events tend to be highly correlated with preexisting accessible chromatin. These cell type-specific events typically contain weaker cognate motifs and higher enrichment of other cell type-specific TF motifs, possibly reflecting that their binding may be facilitated through co-operativity with cell-specific regulators that are already bound at target sites prior to TF induction [108]. On the other hand, binding events that show a lower correlation with prior accessibility contain higher affinity cognate motifs and tend to be more conserved across cell types [33,108] (Fig. 4).

Similar observations arise in systems where TFs have been ectopically over-expressed in unnatural cellular settings; e.g. ASCL1, MYOD, and other TFs expressed in ES cells [26,198,224], and OCT4, SOX2, and KLF4 expressed in fibroblasts [199,227]. Such TFs tend to bind larger proportions of sites that are inaccessible in the preexisting chromatin environment, thus displaying greater degrees of pioneering activity than signaling-dependent TFs. However, a similar divergence in sequence features between pre-accessible and pre-inaccessible binding sites is evident. For example, when the bHLH TF ASCL1 is induced in pluripotent cells, pre-inaccessible sites that become bound contain multiple cognate E-box motif instances, while no such enrichment is evident at bound sites that were already accessible [198].

We therefore propose that induced TF binding activities should be viewed in terms of a joint sequence and chromatin continuum (Fig. 5). Individual sites may be bound if they contain favorable sequence features (e.g. stronger cognate binding motifs) or if they contain favorable pre-existing chromatin landscapes (e.g. chromatin accessibility or prior binding by cooperative TFs). Each TF may have different binding site distributions along each axis - pioneer TFs may be more independent of the prior chromatin axis, for example.

With increasing availability of ChIP-seq experiments that profile induced TF binding in cell types with predetermined chromatin landscapes, computational methods can be leveraged to investigate the joint sequence and chromatin determinants of TF binding. For instance, Guertin et al. use a "rules ensemble" linear regression model to integrate *in vitro* DNA binding preferences and preexisting DNase I data, predicting heat shock induced HSF binding with higher accuracy than a sequence-only model [232]. We have also recently presented an interpretable neural network that jointly models sequence and preexisting cell type chromatin data to predict induced TF binding [233]. Application of our model to analyze a range of induced TFs' binding sites demonstrates that, as expected, some TFs are more dependent on the prior chromatin landscape than others. Moreover, our approach evaluates the relative contributions of sequence and prior chromatin features in determining binding at individual sites. Further development and interpretation of joint sequence and preexisting chromatin models will lead to insights into the mechanisms by which TF binding becomes established in cell type-specific patterns.

Transparency document

The https://doi.org/10.1016/j.bbagrm.2019.194443 associated with this article can be found, in online version.

Acknowledgements

This work was supported by a Penn State Academic Computing Fellowship (to DS) and NIGMS R01GM121613 (to SM). We thank Akshay Kakumanu for assistance in creating Fig. 1.

References

[1] R.L. Davis, H. Weintraub, A.B. Lassar, Expression of a single transfected cDNA converts fibroblasts to myoblasts, Cell 51 (1987) 987–1000.
[2] K. Takahashi, S. Yamanaka, Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors, Cell 126 (2006) 663–676.
[3] Q. Zhou, J. Brown, A. Kanarek, J. Rajagopal, D.A. Melton, In vivo reprogramming of adult pancreatic exocrine cells to beta-cells, Nature 455 (2008) 627–632.
[4] S.A. Lambert, et al., The Human Transcription Factors, Cell 172 (2018) 650–665.
[5] S. Masui, et al., Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells, Nat. Cell Biol. 9 (2007) 625–635.
[6] V. Graham, J. Khudyakov, P. Ellis, L. Pevny, SOX2 Functions to Maintain Neural Progenitor Identity, Neuron 39 (2003) 749–765.
[7] S.L. Pfaff, M. Mendelsohn, C.L. Stewart, T. Edlund, T.M. Jessell, Requirement for LIM Homeobox Gene Isl1 in motor neuron generation reveals a motor neuron-dependent step in interneuron differentiation, Cell 84 (1996) 309–320.
[8] Fragkouli, A., van Wijk, N. V., Lopes, R., Kessaris, N. & Pachnis, V. LIM homeodomain transcription factor-dependent specification of bipotential MGE progenitors into cholinergic and GABAergic striatal interneurons. Development 136, 3841–3851 (2009).
[9] A. Du, et al., Islet-1 is required for the maturation, proliferation, and survival of the endocrine pancreas, Diabetes 58 (2009) 2059–2069.
[10] L. Bu, et al., Human ISL1 heart progenitors generate diverse multipotent cardiovascular cell lineages, Nature 460 (2009) 113–117.
[11] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-wide mapping of in vivo protein-DNA interactions, Science 316 (2007) 1497–1502.
[12] H.S. Rhee, B.F. Pugh, Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution, Cell 147 (2011) 1408–1419.
[13] P.J. Skene, S. Henikoff, An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites, eLife 6 (2017).
[14] ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (57–74) (2012).
[15] F. Yue, et al., A comparative encyclopedia of DNA elements in the mouse genome, Nature 515 (2014) 355–364.
[16] S. Roy, et al., Identification of functional elements and regulatory circuits by Drosophila modENCODE, Science 330 (2010) 1787–1797.
[17] S. John, et al., Chromatin accessibility pre-determines glucocorticoid receptor binding patterns, Nat. Genet. 43 (2011) 264–268.
[18] S. Heinz, et al., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, Mol. Cell 38 (2010) 576–589.
[19] B.-K. Lee, et al., Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells, Genome Res. 22 (2012) 9–24.
[20] A. Arvey, P. Agius, W.S. Noble, C. Leslie, Sequence and chromatin determinants of cell-type-specific transcription factor binding, Genome Res. 22 (2012) 1723–1734.
[21] Frietze, S. et al. Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. Genome Biol. **13**, R52 (2012).
[22] M.A. Lodato, et al., SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and NPCs to specify cell state, PLoS Genet. 9 (2013) e1003288.
[23] A.M. Tsankov, et al., Transcription factor binding dynamics during human ES cell differentiation, Nature 518 (2015) 344–349.
[24] A.F. Bardet, Q. He, J. Zeitlinger, A. Stark, A computational pipeline for comparative ChIP-seq analyses, Nat. Protoc. 7 (2012) 45–61.
[25] K. Liang, S. Keles, Detecting differential binding of transcription factors with ChIP-seq, Bioinformatics 28 (2012) 121–122.
[26] S. Mahony, et al., An integrated model of multiple-condition ChIP-Seq data reveals predeterminants of Cdx2 binding, PLoS Comput. Biol. 10 (2014) e1003501.
[27] J. Banerji, S. Rusconi, W. Schaffner, Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences, Cell 27 (1981) 299–308.
[28] M. Mercola, X.F. Wang, J. Olsen, K. Calame, Transcriptional enhancer elements in the mouse immunoglobulin heavy chain locus, Science 221 (1983) 663–665.
[29] R. Andersson, et al., An atlas of active enhancers across human cell types and tissues, Nature 507 (2014) 455–461.
[30] S. Heinz, C.E. Romanoski, C. Benner, C.K. Glass, The selection and function of cell type-specific enhancers, Nat. Rev. Mol. Cell Biol. 16 (2015) 144–154.
[31] A.S. Nord, et al., Rapid and pervasive changes in genome-wide enhancer usage during mammalian development, Cell 155 (2013) 1521–1531.
[32] C.A. Gifford, et al., Transcriptional and epigenetic dynamics during specification of human embryonic stem cells, Cell 153 (2013) 1149–1163.
[33] A. Kakumanu, S. Velasco, E. Mazzoni, S. Mahony, Deconvolving sequence features that discriminate between overlapping regulatory annotations, PLoS Comput. Biol. 13 (2017) e1005795.
[34] G.D. Stormo, DNA binding sites: representation and discovery, Bioinformatics 16 (2000) 16–23.
[35] G.D. Stormo, Y. Zhao, Determining the specificity of protein-DNA interactions, Nat. Rev. Genet. 11 (2010) 751–760.
[36] O.G. Berg, P.H. von Hippel, Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters, J. Mol. Biol. 193 (1987) 723–750.
[37] P.V. Benos, M.L. Bulyk, G.D. Stormo, Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res. 30 (2002) 4442–4451.
[38] M.L. Bulyk, P.L.F. Johnson, G.M. Church, Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors, Nucleic Acids Res. 30 (2002) 1255–1261.
[39] G. Badis, et al., Diversity and complexity in DNA recognition by transcription factors, Science 324 (2009) 1720–1723.
[40] A. Jolma, et al., DNA-binding specificities of human transcription factors, Cell 152 (2013) 327–339.
[41] Y. Zhao, S. Ruan, M. Pandey, G.D. Stormo, Improved models for transcription factor binding site identification using non-independent interactions, Genetics 191 (2012) 781–790.
[42] E. Sharon, S. Lubliner, E. Segal, A feature-based approach to modeling protein-DNA interactions, PLoS Comput. Biol. 4 (2008) e1000154.
[43] W.W. Wasserman, A. Sandelin, Applied bioinformatics for the identification of regulatory elements, Nat Rev Genet 5 (2004) 276–287.
[44] T. Siggers, R. Gordân, Protein-DNA binding: complexities and multi-protein codes, Nucleic Acids Res. 42 (2014) 2099–2111.
[45] M. Slattery, et al., Absence of a simple code: how transcription factors read the genome, Trends Biochem. Sci. 39 (2014) 381–399.
[46] M.F. Berger, et al., Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities, Nat. Biotechnol. 24 (2006) 1429–1435.
[47] R. Gordân, et al., Genomic regions flanking E-Box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape, Cell Rep. 3 (2013) 1093–1104.
[48] A. Jolma, et al., Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities, Genome Res. 20 (2010) 861–873.
[49] M.T. Weirauch, et al., Determination and inference of eukaryotic transcription factor sequence specificity, Cell 158 (2014) 1431–1443.
[50] M.A. Hume, L.A. Barrera, S.S. Gisselbrecht, M.L. Bulyk, UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions, Nucleic Acids Res. 43 (2014) D117–D122.
[51] I.V. Kulakovskiy, et al., HOCOMOCO: a comprehensive collection of human transcription factor binding sites models, Nucleic Acids Res. 41 (2013) D195–D202.
[52] A. Khan, et al., JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework, Nucleic Acids Res. 46 (2018) D260–D266.
[53] R. Rohs, et al., The role of DNA shape in protein-DNA recognition, Nature 461 (2009) 1248–1253.
[54] I. Dror, T. Golan, C. Levy, R. Rohs, Y. Mandel-Gutfreund, A widespread role of the motif environment on transcription factor binding across diverse protein families, Genome Res. (2015), https://doi.org/10.1101/gr.184671.114.
[55] A.P. Fong, et al., Genetic and epigenetic determinants of neurogenesis and myogenesis, Dev. Cell 22 (2012) 721–735.
[56] N. Shen, et al., Divergence in DNA specificity among paralogous transcription factors contributes to their differential in vivo binding, Cell Syst. 6 (2018) 470–483.e8.
[57] Y. Yin, et al., Impact of cytosine methylation on DNA binding specificities of human transcription factors, Science 356 (2017).
[58] J.A.L. Gelfond, M. Gupta, J.G. Ibrahim, A Bayesian hidden Markov model for motif discovery through joint modeling of genomic sequence and ChIP-chip data, Biometrics 65 (2009) 1087–1095.
[59] J. Keilwagen, J. Grau, Varying levels of complexity in transcription factor binding motifs, Nucleic Acids Res. 43 (2015) e119.
[60] A. Mathelier, W.W. Wasserman, The next generation of transcription factor binding site prediction, PLoS Comput Biol 9 (2013) e1003214.
[61] P. Agius, A. Arvey, W. Chang, W.S. Noble, C. Leslie, High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions, PLoS Comput. Biol. 6 (2010) e1000916.
[62] F. Mordelet, J. Horton, A.J. Hartemink, B.E. Engelhardt, R. Gordân, Stability selection for regression-based models of transcription factor-DNA binding specificity, Bioinformatics 29 (2013) i117–i125.
[63] M.T. Weirauch, et al., Evaluation of methods for modeling transcription factor sequence specificity, Nat. Biotechnol. 31 (2013) 126–134.
[64] T.R. Riley, A. Lazarovici, R.S. Mann, H.J. Bussemaker, Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE, eLife 4 (2015).
[65] Y. Zhao, G.D. Stormo, Quantitative analysis demonstrates most transcription factors require only simple models of specificity, Nat. Biotechnol. 29 (2011) 480–483.
[66] D. Lee, R. Karchin, M.A. Beer, Discriminative prediction of mammalian enhancers from DNA sequence, Genome Res. 21 (2011) 2167–2180.
[67] C. Fletez-Brant, D. Lee, A.S. McCallion, M.A. Beer, kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets, Nucleic Acids Res. 41 (2013) W544–W556.
[68] M. Ghandi, D. Lee, M. Mohammad-Noori, M.A. Beer, Enhanced regulatory sequence prediction using gapped k-mer features, PLoS Comput. Biol. 10 (2014) e1003711.
[69] M. Ghandi, et al., gkmSVM: an R package for gapped-kmer SVM, Bioinformatics 32 (2016) 2205–2207.
[70] B. Alipanahi, A. Delong, M.T. Weirauch, B.J. Frey, Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning, Nat. Biotechnol. 33 (2015) 831–838.
[71] H. Zeng, M.D. Edwards, G. Liu, D.K. Gifford, Convolutional neural network architectures for predicting DNA-protein binding, Bioinformatics 32 (2016) i121–i127.
[72] C. Viner, et al., Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet, *bioRxiv* 043794, 2016, https://doi.org/10.1101/043794.
[73] A.J. Sood, C. Viner, M.M. Hoffman, DNAmod: the DNA modification database, J.

Cheminformatics 11 (2019) 30.
[74] J. Yang, S.A. Ramsey, A DNA shape-based regulatory score improves position-weight matrix-based recognition of transcription factor binding sites, Bioinformatics (2015), https://doi.org/10.1093/bioinformatics/btv391.
[75] T. Zhou, et al., Quantitative modeling of transcription factor binding specificities using DNA shape, Proc. Natl. Acad. Sci. 112 (2015) 4654–4659.
[76] A. Mathelier, et al., DNA shape features improve transcription factor binding site predictions in vivo, Cell Syst. 3 (2016) 278–286.e4.
[77] L. Yang, et al., Transcription factor family-specific DNA shape readout revealed by quantitative specificity models, Mol. Syst. Biol. 13 (2017) 910.
[78] J. Guo, et al., Sequence specificity incompletely defines the genome-wide occupancy of Myc, Genome Biol. 15 (2014) 482.
[79] M.A. Ptashne, Genetic switch: phage lambda and Higher organisms, Cell Press (1992).
[80] Johnson, A. A Combinatorial Regulatory Circuit in Budding Yeast. Cold Spring Harb. Monogr, *Arch.* **22B**, 975-1006–1006, 1992.
[81] C. Wolberger, Multiprotein-DNA complexes in transcriptional regulation, Annu. Rev. Biophys. Biomol. Struct. 28 (1999) 29–56.
[82] E. Morgunova, J. Taipale, Structural perspective of cooperative transcription factor binding, Curr. Opin. Struct. Biol. 47 (2017) 1–8.
[83] C.C. Adams, J.L. Workman, Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative, Mol. Cell. Biol. 15 (1995) 1405–1421.
[84] L.A. Mirny, Nucleosome-mediated cooperativity between transcription factors, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 22534–22539.
[85] J.N. Glover, S.C. Harrison, Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA, Nature 373 (1995) 257–261.
[86] F.E. Chen, D.B. Huang, Y.Q. Chen, G. Ghosh, Crystal structure of p50/p65 heterodimer of transcription factor NF-kappaB bound to DNA, Nature 391 (1998) 410–413.
[87] A. Reményi, et al., Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers, Genes Dev. 17 (2003) 2048–2059.
[88] J. Chen, et al., Single-molecule dynamics of enhanceosome assembly in embryonic stem cells, Cell 156 (2014) 1274–1285.
[89] A.M. Näär, et al., The orientation and spacing of core DNA-binding motifs dictate selective transcriptional responses to three nuclear receptors, Cell 65 (1991) 1267–1279.
[90] F. Rastinejad, T. Wagner, Q. Zhao, S. Khorasanizadeh, Structure of the RXR–RAR DNA-binding complex on the retinoic acid response element DR1, EMBO J. 19 (2000) 1045–1054.
[91] S. Mahony, et al., Ligand-dependent dynamics of retinoic acid receptor binding during early neurogenesis, Genome Biol. 12 (2011) R2.
[92] T.L. Towers, B.F. Luisi, A. Asianov, L.P. Freedman, DNA target selectivity by the vitamin D3 receptor: mechanism of dimer binding to an asymmetric repeat element, Proc. Natl. Acad. Sci. U. S. A. 90 (1993) 6310–6314.
[93] M. Slattery, et al., Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins, Cell 147 (2011) 1270–1282.
[94] E.O. Mazzoni, et al., Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity, Nat. Neurosci. 16 (2013) 1219–1227.
[95] A. Jolma, et al., DNA-dependent formation of transcription factor pairs alters their binding specificity, Nature 527 (2015) 384–388.
[96] N. Heldring, et al., Estrogen receptors: how do they signal and what are their targets, Physiol. Rev. 87 (2007) 905–931.
[97] N. Heldring, et al., Multiple Sequence-Specific DNA-Binding Proteins Mediate Estrogen Receptor Signaling through a Tethering Pathway. *Mol. Endocrinol. Baltim*, Md 25 (2011) 564–574.
[98] M. Gheorghe, et al., A map of direct TF-DNA interactions in the human genome, Nucleic Acids Res. e21 (2019) 47.
[99] J. Wang, et al., Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors, Genome Res. 22 (2012) 1798–1812.
[100] S.R. Starick, et al., ChIP-exo signal associated with DNA-binding motifs provide insights into the genomic binding of the glucocorticoid receptor and cooperating transcription factors, Genome Res. 25 (2015) 825–835.
[101] N. Yamada, W.K.M. Lai, N. Farrell, B.F. Pugh, S. Mahony, Characterizing protein-DNA binding event subtypes in ChIP-exo data, Bioinformatics 35 (2018) 903–913.
[102] K.J. Polach, J. Widom, A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites, J. Mol. Biol. 258 (1996) 800–812.
[103] S. Vashee, K. Melcher, W.V. Ding, S.A. Johnston, T. Kodadek, Evidence for two modes of cooperative DNA binding in vivo that do not involve direct protein-protein interactions, Curr. Biol. CB 8 (1998) 452–458.
[104] J.A. Miller, J. Widom, Collaborative competition mechanism for gene activation in vivo, Mol. Cell. Biol. 23 (2003) 1623–1632.
[105] T.C. Voss, et al., Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism, Cell 146 (2011) 544–554.
[106] D. Thanos, T. Maniatis, Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome, Cell 83 (1995) 1091–1100.
[107] D. Panne, T. Maniatis, S.C. Harrison, An atomic model of the interferon-beta enhanceosome, Cell 129 (2007) 1111–1123.
[108] J. Gertz, et al., Distinct properties of cell-type-specific and shared transcription factor binding sites, Mol. Cell 52 (2013) 25–36.
[109] M. Setty, C.S. Leslie, SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps, PLoS Comput. Biol. 11 (2015) e1004271.
[110] V. Gotea, et al., Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers, Genome Res. 20 (2010) 565–577.
[111] J. Crocker, et al., Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness, Cell 160 (2015) 191–203.
[112] J. Yan, et al., Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites, Cell 154 (2013) 801–813.
[113] K. Masternak, et al., CIITA is a transcriptional coactivator that is recruited to MHC class II promoters by multiple synergistic interactions with an enhanceosome complex, Genes Dev. 14 (2000) 1156–1166.
[114] K. Belov, et al., Reconstructing an ancestral mammalian immune supercomplex from a marsupial major histocompatibility complex, PLoS Biol. 4 (2006) e46.
[115] Y. Guo, S. Mahony, D.K. Gifford, High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints, PLoS Comput. Biol. 8 (2012) e1002638.
[116] E.K. Farley, K.M. Olson, W. Zhang, D.S. Rokhsar, M.S. Levine, Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers, Proc. Natl. Acad. Sci. U. S. A. 113 (2016) 6508–6513.
[117] D.N. Arnosti, M.M. Kulkarni, Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J. Cell. Biochem. 94 (2005) 890–898.
[118] C. Buecker, J. Wysocka, Enhancers as information integration hubs in development: lessons from genomics, Trends Genet. 28 (2012) 276–284.
[119] E. Calo, J. Wysocka, Modification of enhancer chromatin: what, how, and why? Mol. Cell 49 (2013) 825–837.
[120] C.-T. Ong, V.G. Corces, Enhancer function: new insights into the regulation of tissue-specific gene expression, Nat. Rev. Genet. 12 (2011) 283–293.
[121] H.K. Long, S.L. Prescott, J. Wysocka, Ever-changing landscapes: transcriptional enhancers in development and evolution, Cell 167 (2016) 1170–1187.
[122] D.S. Gross, W.T. Garrard, Nuclease hypersensitive sites in chromatin, Annu. Rev. Biochem. 57 (1988) 159–197.
[123] Y. Mito, J.G. Henikoff, S. Henikoff, Histone replacement marks the boundaries of cis-regulatory domains, Science 315 (2007) 1408–1411.
[124] C. Jin, et al., H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions, Nat. Genet. 41 (2009) 941–945.
[125] N.D. Heintzman, et al., Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome, Nat. Genet. 39 (2007) 311–318.
[126] K.W. Jeong, et al., Recognition of enhancer element-specific histone methylation by TIP60 in transcriptional activation, Nat. Struct. Mol. Biol. 18 (2011) 1358–1365.
[127] M. Altaf, et al., NuA4-dependent acetylation of nucleosomal histones H4 and H2A directly stimulates incorporation of H2A.Z by the SWR1 complex, J. Biol. Chem. 285 (2010) 15966–15977.
[128] M.P. Schnetz, et al., Genomic distribution of CHD7 on chromatin tracks H3K4 methylation patterns, Genome Res. 19 (2009) 590–601.
[129] M.P. Schnetz, et al., CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression, PLoS Genet. 6 (2010) e1001023.
[130] S.K.T. Ooi, et al., DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA, Nature 448 (2007) 714–717.
[131] R.E. Thurman, et al., The accessible chromatin landscape of the human genome, Nature 489 (2012) 75–82.
[132] G. Elliott, et al., Intermediate DNA methylation is a conserved signature of genome regulation, Nat. Commun. 6 (2015) 6363.
[133] M. Yu, et al., Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome, Cell 149 (2012) 1368–1380.
[134] S. Ito, et al., Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification, Nature 466 (2010) 1129–1133.
[135] M.B. Ardehali, et al., Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcription, EMBO J. 30 (2011) 2817–2828.
[136] H.H. Ng, F. Robert, R.A. Young, K. Struhl, Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity, Mol. Cell 11 (2003) 709–719.
[137] Herz, H.-M. et al. Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. *Genes Dev.* **26**, 2604–2620 (2012).
[138] S.A. Shinsky, K.E. Monteith, S. Viggiano, M.S. Cosgrove, Biochemical reconstitution and phylogenetic comparison of human SET1 family core complexes involved in histone methylation, J. Biol. Chem. 290 (2015) 6361–6375.
[139] S.R. Patel, D. Kim, I. Levitan, G.R. Dressler, The BRCT-domain containing protein PTIP links PAX2 to a histone H3, Lysine 4 Methyltransferase Complex. Dev. Cell 13 (2007) 580–592.
[140] R. Mo, S.M. Rao, Y.-J. Zhu, Identification of the MLL2 complex as a coactivator for estrogen receptor alpha, J. Biol. Chem. 281 (2006) 15714–15720.
[141] M.P. Creyghton, et al., Histone H3K27ac separates active from poised enhancers and predicts developmental state, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 21931–21936.
[142] A. Rada-Iglesias, et al., A unique chromatin signature uncovers early developmental enhancers in humans, Nature 470 (2011) 279–283.
[143] G.E. Zentner, P.J. Tesar, P.C. Scacheri, Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions, Genome Res. 21 (2011) 1273–1283.
[144] Z. Wang, et al., Combinatorial patterns of histone acetylations and methylations in the human genome, Nat. Genet. 40 (2008) 897–903.
[145] Q. Jin, et al., Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation, EMBO J. 30 (2011) 249–262.
[146] K.-J. Won, I. Chepelev, B. Ren, W. Wang, Prediction of regulatory elements in mammalian genomes using chromatin signatures, BMC Bioinformatics 9 (2008) 547.

[147] O.L. Wapinski, et al., Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons, Cell 155 (2013) 621–635.
[148] B. Xin, R. Rohs, Relationship between histone modifications and transcription factor binding is protein family specific, Genome Res. 28 (2018) 321–333.
[149] T.-K. Kim, et al., Widespread transcription at neuronal activity-regulated enhancers, Nature 465 (2010) 182–187.
[150] F. Koch, et al., Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters, Nat. Struct. Mol. Biol. 18 (2011) 956–963.
[151] K. Pulakanti, et al., Enhancer transcribed RNAs arise from hypomethylated, Tet-occupied genomic regions, Epigenetics 8 (2013) 1303–1320.
[152] K. Mousavi, et al., eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci, Mol. Cell 51 (2013) 606–617.
[153] N. Hah, S. Murakami, A. Nagari, C. Danko, W.L. Kraus, Enhancer transcripts mark active estrogen receptor binding sites, Genome Res. 23 (2013) 1210–1223.
[154] J.G. Azofeifa, et al., Enhancer RNA profiling predicts transcription factor activity, Genome Res. 28 (2018) 334–344.
[155] Z. Liu, D.R. Scannell, M.B. Eisen, R. Tjian, Control of embryonic stem cell lineage commitment by core promoter factor, TAF3, Cell 146 (2011) 720–731.
[156] H. Zhou, et al., Dual functions of TAF7L in adipocyte differentiation, eLife 2 (2013) e00170.
[157] M.H. Kagey, et al., Mediator and cohesin connect gene expression and chromatin architecture, Nature 467 (2010) 430–435.
[158] S. Bonn, et al., Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development, Nat. Genet. 44 (2012) 148–156.
[159] W.A. Whyte, et al., Enhancer decommissioning by LSD1 during embryonic stem cell differentiation, Nature 482 (2012) 221–225.
[160] M.J. Ziller, et al., Dissecting neural differentiation regulatory networks through epigenetic footprinting, Nature 518 (2015) 355–359.
[161] A.M. Urvalek, L.J. Gudas, Retinoic acid and histone deacetylases regulate epigenetic changes in embryonic stem cells, J. Biol. Chem. 289 (2014) 19519–19530.
[162] X.-Y. Li, et al., The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding, Genome Biol. 12 (2011) R34.
[163] D. Quang, X. Xie, FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data, Methods (2019), https://doi.org/10.1016/j.ymeth.2019.03.020.
[164] H. Li, D. Quang, Y. Guan, Anchor: Trans-cell type prediction of transcription factor binding sites, Genome Res. 29 (2018) 281–292.
[165] J. Keilwagen, S. Posch, J. Grau, Accurate prediction of cell type-specific transcription factor binding, Genome Biol. 20 (2019) 9.
[166] Q. Qin, J. Feng, Imputation for transcription factor binding predictions based on deep learning, PLoS Comput. Biol. 13 (2017) e1005403.
[167] R.I. Sherwood, et al., Discovery of non-directional and directional pioneer transcription factors by modeling DNase profile magnitude and shape, Nat. Biotechnol. 32 (2014) 171–178.
[168] S. Neph, et al., An expansive human regulatory lexicon encoded in transcription factor footprints, Nature 489 (2012) 83–90.
[169] J.R. Hesselberth, et al., Global mapping of protein-DNA interactions in vivo by digital genomic footprinting, Nat. Methods 6 (2009) 283–289.
[170] T. Kaplan, et al., Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development, PLOS Genet. 7 (2011) e1001290.
[171] L. Narlikar, R. Gordân, A.J.A. Hartemink, Nucleosome-guided map of transcription factor binding sites in yeast, PLOS Comput. Biol. 3 (2007) e215.
[172] R. Pique-Regi, et al., Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data, Genome Res. 21 (2011) 447–455.
[173] A.P. Boyle, et al., High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells, Genome Res. 21 (2011) 456–464.
[174] B. Quach, T.S.D.C. Furey, analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter, Bioinformatics 33 (2017) 956–963.
[175] H.H. He, et al., Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification, Nat. Methods 11 (2014) 73–78.
[176] A. Karabacak Calviello, A. Hirsekorn, R. Wurmus, D. Yusuf, U. Ohler, Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling, Genome Biol. 20 (2019) 42.
[177] G.G. Yardımcı, C.L. Frank, G.E. Crawford, U. Ohler, Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection, Nucleic Acids Res. 42 (2014) 11865–11878.
[178] A. Youn, E.J. Marquez, N. Lawlor, M.L. Stitzel, D. Ucar, BiFET: sequencing bias-free transcription factor Footprint Enrichment Test, Nucleic Acids Res. 47 (2019) e11.
[179] M.-H. Sung, S. Baek, G.L. Hager, Genome-wide footprinting: ready for prime time? Nat. Methods 13 (2016) 222–228.
[180] X. Chen, B. Yu, N. Carriero, C. Silva, R. Bonneau, Mocap: large-scale inference of transcription factor binding sites from chromatin accessibility, Nucleic Acids Res. 45 (2017) 4315–4329.
[181] D. Quang, X. DanQ Xie, a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences, Nucleic Acids Res. 44 (2016) e107.
[182] W.-L. Guo, D.-S. Huang, An efficient method to transcription factor binding sites imputation via simultaneous completion of multiple matrices with positional consistency, Mol. Biosyst. 13 (2017) 1827–1837.
[183] M. Karimzadeh, M.M. Hoffman, Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome, *bioRxiv* 168419, 2018, https://doi.org/10.1101/168419.
[184] K. Luger, T.J. Rechsteiner, A.J. Flaus, M.M. Waye, T.J. Richmond, Characterization of nucleosome core particles containing histone proteins made in bacteria, J. Mol. Biol. 272 (1997) 301–311.
[185] K.S. Zaret, J.S. Carroll, Pioneer transcription factors: establishing competence for gene expression, Genes Dev. 25 (2011) 2227–2241.
[186] F. Zhu, et al., The interaction landscape between transcription factors and the nucleosome, Nature 562 (2018) 76–81.
[187] X. Yu, M.J. Buck, Defining TP53 pioneering capabilities with competitive nucleosome binding assays, Genome Res. 29 (2019) 107–115.
[188] A. Soufi, et al., Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming, Cell 161 (2015) 555–568.
[189] C. Yan, H. Chen, L. Bai, Systematic study of nucleosome-displacing factors in budding yeast, Mol. Cell. 71 (2018) 294–305.e4.
[190] K.S. Zaret, S.E. Mango, Pioneer transcription factors, chromatin dynamics, and cell fate control, Curr. Opin. Genet. Dev. 37 (2016) 76–81.
[191] J. Drouin, Pioneer transcription factors in cell fate specification, Mol. Endocrinol. 28 (2014) 989–998.
[192] E.E. Swinstead, V. Paakinaho, D.M. Presman, G.L. Hager, Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: a new perspective: Multiple transcription factors can effect chromatin pioneer functions through dynamic interactions with ATP-dependent chromatin remodeling factors, BioEssays 38 (2016) 1150–1157.
[193] R. Gualdi, et al., Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control, Genes Dev. 10 (1996) 1670–1682.
[194] L.A. Cirillo, et al., Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4, Mol. Cell 9 (2002) 279–289.
[195] S. Ghisletti, et al., Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages, Immunity 32 (2010) 317–328.
[196] I. Barozzi, et al., Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers, Mol. Cell 54 (2014) 844–857.
[197] L. Budry, et al., The selector gene Pax7 dictates alternate pituitary cell fates through its pioneer action on chromatin remodeling, Genes Dev. 26 (2012) 2299–2310.
[198] B. Aydin, et al., Proneural factors Ascl1 and Neurog2 contribute to neuronal subtype identities by establishing distinct chromatin landscapes. *Nat*, Neurosci. 22 (2019) 897–908.
[199] A. Soufi, G. Donahue, K.S. Zaret, Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome, Cell 151 (2012) 994–1004.
[200] H.-T. Hsu, et al., Recruitment of RNA polymerase II by the pioneer transcription factor PHA-4, Science 348 (2015) 1372–1376.
[201] J. Gertz, T.E. Reddy, K.E. Varley, M.J. Garabedian, R.M. Myers, Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner, Genome Res. 22 (2012) 2153–2162.
[202] A.G. Robertson, et al., Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding, Genome Res. 18 (2008) 1906–1917.
[203] M.J. Guertin, J.T. Lis, Chromatin landscape dictates HSF binding to target DNA elements, PLoS Genet. 6 (2010) e1001114.
[204] M. Lupien, et al., FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription, Cell 132 (2008) 958–970.
[205] A.A. Sérandour, et al., Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers, Genome Res. 21 (2011) 555–565.
[206] J. Xu, et al., Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells, Genes Dev. 23 (2009) 2824–2838.
[207] F.M. Cernilogar, et al., Pre-marked chromatin and transcription factor co-binding shape the pioneering activity of Foxa2, Nucleic Acids Res, 2019, https://doi.org/10.1093/nar/gkz627.
[208] G. Hu, et al., H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation, Cell Stem Cell 12 (2013) 180–192.
[209] J. Donaghey, et al., Genetic determinants and epigenetic effects of pioneer-factor occupancy, Nat. Genet. 50 (2018) 250–258.
[210] T.A. Johnson, et al., Conventional and pioneer modes of glucocorticoid receptor interaction with enhancer chromatin in vivo, Nucleic Acids Res. 46 (2018) 203–214.
[211] N.A.S. Rao, et al., Coactivation of GR and NFKB alters the repertoire of their binding sites and target genes, Genome Res. 21 (2011) 1404–1416.
[212] E.E. Swinstead, et al., Steroid receptors reprogram FoxA1 occupancy through dynamic chromatin transitions, Cell 165 (2016) 593–605.
[213] Y. Ding, et al., Ikaros tumor suppressor function includes induction of active enhancers and super-enhancers along with pioneering activity, Leukemia, 2019, https://doi.org/10.1038/s41375-019-0474-0.
[214] Z. Li, et al., Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation, Cell 151 (2012) 1608–1616.
[215] M.U. Kaikkonen, et al., Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription, Mol. Cell 51 (2013) 310–325.
[216] E. Ortega, et al., Transcription factor dimerization activates the p300 acetyltransferase, Nature 562 (2018) 538.
[217] Q. Zhou, D.A. Melton, Extreme Makeover: Converting One Cell into Another, Cell Stem Cell 3 (2008) 382–388.
[218] D.E. Cohen, D. Melton, Turning straw into gold: directing cell fate for regenerative medicine, Nat. Rev. Genet. 12 (2011) 243–252.
[219] T. Graf, Historical origins of transdifferentiation and reprogramming, Cell Stem

Cell 9 (2011) 504–516.
[220] T. Vierbuchen, M. Wernig, Molecular roadblocks for cellular reprogramming, Mol. Cell 47 (2012) 827–838.
[221] J. Ladewig, P. Koch, O. Brüstle, Leveling Waddington: the emergence of direct programming and the loss of cell fate hierarchies, Nat. Rev. Mol. Cell Biol. 14 (2013) 225–236.
[222] B. Aydin, E.O. Mazzoni, Cell reprogramming: the many roads to success, Annu. Rev. Cell Dev. Biol. (2019), https://doi.org/10.1146/annurev-cellbio-100818-125127.
[223] A. Pataskar, et al., NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program, EMBO J. 35 (2016) 24–45.
[224] B.H. Casey, R.K. Kollipara, K. Pozo, J.E. Johnson, Intrinsic DNA binding properties demonstrated for lineage-specifying basic helix-loop-helix transcription factors, Genome Res. 28 (2018) 484–496.
[225] van der Raadt, J., van Gestel, S. H. C., Nadif Kasri, N. & Albers, C. A. ONECUT transcription factors induce neuronal characteristics and remodel chromatin accessibility. Nucleic Acids Res. 47, 5587–5602 (2019).
[226] O.L. Wapinski, et al., Rapid chromatin switch in the direct reprogramming of fibroblasts to neurons, Cell Rep. 20 (2017) 3236–3247.
[227] C. Chronis, et al., Cooperative Binding of Transcription Factors Orchestrates Reprogramming, *Cell* **168** e20 (2017) 442–459.
[228] C.H. Waddington, Organisers & Genes, The University Press, 1940.
[229] A.S. Knaupp, et al., Transient and permanent reconfiguration of chromatin and transcription factor occupancy drive reprogramming, Cell Stem Cell 21 e6 (2017) 834–845.
[230] S. Velasco, et al., A multi-step transcriptional and chromatin state cascade underlies motor neuron programming from embryonic stem cells, *Cell Stem Cell* **20** e8 (2017) 205–217.
[231] H. Liu, et al., Visualizing long-term single-molecule dynamics in vivo by stochastic protein labeling, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) 343–348.
[232] M.J. Guertin, A.L. Martins, A. Siepel, J.T. Lis, Accurate prediction of inducible transcription factor binding intensities in vivo, PLoS Genet. 8 (2012) e1002610.
[233] D. Srivastava, B. Aydin, E.O. Mazzoni, S. Mahony, Characterizing the sequence and prior chromatin determinants of induced TF binding with bimodal neural networks, *bioRxiv* 672790, 2019, https://doi.org/10.1101/672790.