Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants

Xuelei Lai, Arnaud Stigliani, Gilles Vachon, Cristel Carles, Cezary Smaczniak, Chloe Zubieta, Kerstin Kaufmann, François Parcy

Building transcription factor binding site models to understand gene regulation in plants

Xuelei Lai[1,*], Arnaud Stigliani[1], Gilles Vachon[1], Cristel Carles[1], Cezary Smaczniak[2], Chloe Zubieta[1], Kerstin Kaufmann[2] and François Parcy[1,*]

1. CNRS, Univ. Grenoble Alpes, CEA, INRA, BIG-LPCV, 38000 Grenoble, France

2. Department for Plant Cell and Molecular Biology, Institute for Biology, Humboldt-Universität zu Berlin, Germany

* Corresponding author: François Parcy, Tel: +33 0438784978, Fax: +33 0438784091, Email: francois.parcy@cea.fr; Co-corresponding author: Xuelei Lai, Email: xuelei.lai@cea.fr.

Abstract

Transcription factors (TF) are key cellular components that control gene expression. They recognize specific DNA sequences, the TF binding sites (TFBS), and thus are targeted to specific regions of the genome where they can recruit transcriptional cofactors and/or chromatin regulators for fine-tuning spatiotemporal gene regulation. Therefore, the identification of TFBS in genomic sequences and their subsequent quantitative modeling is of crucial importance for understanding and predicting gene expression. Here, we review how TFBS can be determined experimentally, how the TFBS models can be constructed *in silico*, and how they can be optimized by taking into account features such as position interdependence within TFBSs, DNA shape and/or by introducing state-of-the-art computational algorithms such as deep learning methods. In addition, we discuss the integration of context variables into the TFBS modeling, including nucleosome positioning, chromatin states, methylation patterns, 3D genome architectures and TF cooperative binding, in order to better predict TF binding under cellular contexts. Finally, we explore the possibilities of combining the optimized TFBS model with technological advances such as targeted TFBS perturbation by CRISPR to better understand gene regulation, evolution and plant diversity.

Running title: Modeling TF binding sites in plants

Introduction

Transcription factors (TFs) are sequence-specific DNA-binding proteins that regulate gene expression in all organisms (Lelli et al., 2012; Lambert et al., 2018). They constitute a large number of protein-coding genes (between 4% to 10%) in the genomes of all species (Babu et al., 2004). For example, in the model plant *Arabidopsis thaliana*, 2492 genes encode TFs, accounting for more than 9% of its total protein coding genes (Swarbreck et al., 2008; Pruneda-Paz et al., 2014). TFs orchestrate gene regulation by binding to their cognate DNA binding sites (TFBS) that are usually located in *cis*-regulatory regions. Upon binding to a TFBS, some TFs are able to recruit epigenetic factors, such as chromatin remodelers (e.g. BRAHMA and SPLAYED in plants (Bezhani et al., 2007)) or modifiers (e.g. Polycomb Repressive Complexes (PRC) (Xiao and Wagner, 2015)) to alter chromatin states. TFs can also interact with components of transcriptional machineries, such as co-factor (e.g. Mediator and SAGA complexes in animals (Allen and Taatjes, 2015; Baptista et al., 2017)), general transcriptional factors (Müller et al., 2010) and RNA polymerase II for regulation of transcriptional initiation. The interplay between TFs and these factors together leads to robust and dynamic gene expression regulation (Spitz and Furlong, 2012; Voss and Hager, 2013).

TFs recognize TFBS in a sequence-specific manner as revealed by structural studies of protein-DNA complexes (Paillard and Lavery, 2004; Rohs et al., 2010) and next generation sequencing (NGS) techniques such as SELEX-seq and ChIP-seq (**Table 1**). In the last decade, these NGS techniques have revolutionized the exploration of the TF binding landscape both *in vitro* and *in vivo* (Koboldt et al., 2013). This has resulted in many databases for TFBS deposition and profiling, such as TRANSFAC (Matys et al., 2006), JASPAR (Khan et al., 2018), UniPROBE (Hume et al., 2015), HOCOMOCO (Kulakovskiy et al., 2013), CIS-BP (Weirauch et al., 2014) and SwissRegulon (Pachkov et al., 2013). Such efforts have substantially boosted our understanding of interactions between TFs and TFBSs in different species, tissues and different developmental stages. While a great deal of progress has been made to map TFBS, the resulting models are often poorly predictive of actual gene regulation. This can be due to poor modeling and prediction of TFBS, non-productive TF binding (i.e. that does not result in gene regulation) or a combination of the two. Here, we will

focus on the more tractable question of how to model TFBS. As transcriptional regulation is a highly dynamic process that occurs in a cell and tissue-specific manner, to better understand such a complex process unbiased quantitative modeling of TFBS with improved prediction power of TF binding is highly demanded. This includes taking into account of variables such as nucleosome positioning, chromatin states, methylation patterns and the 3D structure of the genome, all of which greatly impact transcription factor binding and, for a subset of these binding events, gene expression. Therefore, these variables need to be incorporated into any model to better describe functional TF binding *in vivo* and the concomitant gene regulation.

In this review, we address how TFBS are identified experimentally, how the TFBS models can be built *in silico*, and their optimization strategies. We further integrate context variables into the TFBS model in order to better understand gene regulation networks, evolution and plant diversity (for an outline of the review, refer to **Figure 1**). Throughout the review, we use examples from case studies of TFs that are involved in flower development, a developmental transition that involves the activation of a wealth of genes that are otherwise silent and the concomitant repression of a subset of genes.

TFBS modeling

TFs read genomic DNA sequences in three fundamental ways, namely base readout, indirect readout and shape readout (Rohs et al., 2010; Slattery et al., 2014). In base readout, TFs recognize a given nucleotide sequence by physical interactions between amino acid side chains and accessible edges of the base pairs of DNA. These interactions include hydrogen bonding, hydrophobic interactions and the formation of salt bridges. Indirect readout involves mostly interactions between the TF and the DNA phosphate backbone, whose position is influenced by the nature of the base but not as strongly as in base readout. In shape readout (Abe et al., 2015; Yang et al., 2017), TFs recognize the structural features of DNA, such as DNA bending, groove width and unwinding (Stella et al., 2010; Chen et al., 2013; Hancock et al., 2013). Although once considered as mutually exclusive driving forces for DNA recognition, recent studies have shown that most TFs likely combine base, indirect and shape readout to recognize their TFBSs. Indeed, the integration of these

features has been shown to improve TFBS prediction (Zhou et al., 2015; Mathelier et al., 2016).

Experimental methods to identify TFBS

With the emergence of NGS technologies, many NGS-based methodologies, both *in vitro* and *in vivo*, have been developed for determining TFBSs. Here we concentrate on some of the most widely used and recently developed methods and discuss their advantages and limitations (**Table 1**).

ChIP-seq has long been the gold standard for detecting genome-wide TFBSs bound by a given TF *in vivo* (Johnson et al., 2007; Robertson et al., 2007; Kaufmann et al., 2010). In a standard ChIP-seq protocol, sample tissues are treated with a crosslinking reagent and subjected to nuclei purification to isolate chromatin containing TF-DNA complexes. Generally, an additional step of chromatin shearing by sonication is applied before the final step of chromatin immunoprecipitation (IP) using a TF-specific antibody. The IP product containing enriched DNA fragments that are recognized by the TF of interest is then subjected to NGS sequencing. ChIP-seq has been successfully used routinely in many laboratories. However, standard ChIP-seq protocols have intrinsic limitations and technical drawbacks (Park, 2009). One of the limitations comes from sonication, a process that is highly irreproducible and produces variable DNA fragment sizes that are difficult to sequence. The other limitation is crosslinking, which produces low signal to noise ratio and many false positives. To overcome such limitations, many ChIP-seq variant methods have been developed, including ORGANIC (Kasinathan et al., 2014), ChEC-seq (Zentner et al., 2015), CUT&RUN (Skene and Henikoff, 2017) and SLIM-seq (Gutin et al., 2018) (refer to **Table 1** for unique features and details of these methods), all of which use micrococcal nuclease (MNase) to fragment chromatin, therefore avoiding sonication. Due to the mild conditions of DNA fragmentation by MNase, these methods do not denature or disrupt TF-DNA complexes and eliminate the requirement of crosslinking. These protocols also require substantially lower amounts of input materials and are thus feasible for low-input applications. Processing of IP enriched DNA fragments for downstream NGS application poses another challenge in ChIP-seq and is often time-consuming. To simplify the process, ChIPmentation applies Tn5 transposase directly to bead-bound chromatin, allowing single-step NGS compatible DNA library preparation (Schmidl et al., 2015) (**Table 1**).

ChIP-seq and its variants not only identify TFBSs *in vivo*, but also provide a wealth of information such as detection of binding sites bound by co-binders of the TF. As such, however, this also poses a challenge in distinguishing the true TFBSs from indirect binding mediated by a TF partner. ChIP-seq can be complemented by DNA binding assays performed *in vitro* using recombinant TFs. Among the most widely used *in vitro* techniques are protein binding microarrays (PBM) (Berger et al., 2006; Berger and Bulyk, 2009) and SELEX-seq (Jolma et al., 2010) (**Table 1**). Both methods allow high-throughput identification of TF binding specificities *in vitro*, with such information useful to predict TFBS in genomic sequences, however, they employ synthetic randomized DNA that lack at least some genomic DNA sequence properties known to impact TF binding, including non-physiological primary sequences, core motif flanking regions, and lack of chemical modifications, such as cytosine methylation. To overcome these biases, DAP-seq (DNA affinity purification sequencing) has been recently developed, which uses fragmented genomic DNA as substrates for IP and recombinant TFs (O'Malley et al., 2016; Bartlett et al., 2017) (**Table 1**). As DNA methylation patterns are conserved in genomic DNA, DAP-seq allows genome-wide mapping of the epicistrome and the discovery of TF binding specificity from genomic DNA. Furthermore, when combined with ampDAP-seq, which uses amplified and thus demethylated genomic DNA as substrates, a comprehensive mapping of both the cistrome and the epicistrome can be derived for a given TF. Compared with ChIP-seq and its variants, DAP-seq can be performed in a high-throughput manner with lower costs, as recently demonstrated (O'Malley et al., 2016). Despite these advantages, DAP-seq has its limitations, for example, some TFs are not stable when recombinantly expressed and thus not compatible with DAP-seq, others require interacting partners for their DNA binding activity, and many TFs have distinct DNA binding properties in the presence of co-factors. These limitations have to be taken into account during experiment design and data analysis. Moreover, it has to be noted that DAP-seq lacks cellular chromatin context, therefore, combination of *in vitro* DAP-seq and *in vivo* ChIP-seq would be an informative approach regarding TFBS modeling and *in vivo* TF binding prediction.

Model TFBS *in silico*

In order to make accurate *de novo* prediction of binding sites of a given TF in the genome, a quantitative TFBS model that is representative of TF-DNA binding affinity

is required. This could be derived from a set of known TFBSs using computational methods. Here we discuss how conventional modeling methods could be improved by integrating complex features, such as sequence position dependencies and DNA shape features, which have been shown to play a role in determining TF-DNA specificity. We focus on the most recent and representative TFBS modeling algorithms (**Table 2**), other algorithms have been extensively reviewed elsewhere (Tompa et al., 2005; Hombach et al., 2016).

Position weight matrix

Position weight matrix (PWM) is the most widely used model to represent TF-DNA binding specificity (Schneider and Stephens, 1990; Stormo and Zhao, 2010). Briefly, from a collection of TFBSs, a matrix is built that gives the frequency of each nucleotide at each position of the motif. Based on these frequencies, a PWM or position specific scoring matric (PSSM) can be computed that gives a log-scale value to each nucleotide at each position. Based on the PWM, a score can be calculated for any sequence corresponding to the sum of all values at each position. The logo representation of a PWM illustrates the information content at each position and represents the four bases depending on their frequency (**Figure 2**).

Dependencies

PWMs provide good approximation of TF-DNA interactions in most cases, and can be generated from various datasets, ranging from a small set of known TFBSs to TF-DNA binding data derived from high-throughput assays. However, standard PWM assumes that each position within a TFBS contributes to binding affinity independent of other positions, and is thus unable to represent inter-base dependencies, which have been observed for some TFs (Bulyk, 2002; Tomovic and Oakeley, 2007; Badis et al., 2009). Various models that take into account these dependencies have been shown to outperform standard PWM in *de novo* prediction. For example, the MORPHEUS program allows to introduce di- and tri-nucleotide position dependencies in PWM and has been successfully applied to plant TFs with, in some cases, improved predictive power (Moyroud et al., 2011; Minguet et al., 2015) (**Table 2**).

Several approaches can be taken with respect to how and what dependencies are to be integrated into the modeling algorithm. Some consider pairwise dependencies

between adjacent and/or distal positions, such as the binding energy model (BEM) (Zhao et al., 2012), dinucleotide weight matrices (DWM) (Siddharthan, 2010) and TF Flexible Model (TFFM) (Mathelier and Wasserman, 2013) (**Table 2**). Others introduce higher-order *k*-mer features, that take into account all possible sequences with length *k*, such as the feature motif model (FMM) (Sharon et al., 2008) (**Table 2**). In some cases, model complexity can increase dramatically when arbitrary positions or unconstrained *k*-mer features are used and become prone to be overfitting. Alternative approaches start from a model without dependencies, and use a greedy algorithm to improve the model by adding dependency features iteratively (Hu et al., 2010; Santolini et al., 2014). Thus, dependency features are iteratively added until no further feature could be found to improve the model. Others use Bayesian Markov models (BaMM) of order k that take into account dependencies between one nucleotide and the k previous positions (Kiesel et al., 2018). Complex models integrating dependency features generally outperform simple PWM models, however, some of these models require more expertise to apply and repeated manual attempts to be trained correctly and are thus not facile to use. This constitutes one of the limiting factors that restricts these models from being used routinely in the community. In **Table 2** we summarize features of some of the most recent models.

Shape features

Sequence-based models provide accurate estimation of base readout, however, it cannot explain why some TFs, which have highly conserved DNA-binding domains, bind different sequences genome-wide. For example, the TF paralogs, androgen and glucocorticoid receptors, which bind similar DNA motifs by a set of identical amino acids (Shaffer et al., 2004; Meijsing et al., 2009), share only a third of their genomic binding sites (Zhang et al., 2018a). It turns out that DNA shape features contribute significantly to distinguish *bona fide* TFBSs from others. In the last decade, many studies have revealed that indeed DNA shape features play an important role for determining TF-DNA binding specificity (Rohs et al., 2009; Abe et al., 2015; Yang et al., 2017). A most recent example is the MADS-box TF, SEPALLATA3 (SEP3), a key regulator of flower organ specification (Muiño et al., 2014; Hugouvieux et al., 2018). MADS-box TFs bind to CArG-boxes with consensus sequence of 5'-CC$(A/T)_6$GG-3', yet only a fraction of the CArG-boxes available genome-wide is bound by SEP3. Käppel and colleagues showed that SEP3-DNA binding affinity correlates well with

the width of minor groove of CArG-boxes probes, a shape readout mechanism involves a conserved arginine residue that contact minor groove (Käppel et al., 2018). Although shape features are mainly determined by the TFBS core motif, it can also be affected by flanking regions. In the past, these regions have been overlooked in characterizations of TF binding due to their low sequence information. Now both bioinformatic analysis and biochemical evidence have accumulated pointing towards their importance for TF binding. For example, Dror and colleagues showed a widespread role of the motif environment in TF binding by analyzing some 300 TFs binding data from SELEX-seq and ChIP-seq, and that the preference for a specific environment differs between distinct TF families (Dror et al., 2015). Selective binding of core motifs with different flanking sequences have also been observed by *in vitro* assays for several TFs (Gordân et al., 2013; White et al., 2013; Afek et al., 2014; Levo et al., 2015).

Introducing shape features into TFBSs modeling requires integrating several distinct shape parameters, including Minor Groove Width, Propeller Twist, Roll and Helix Twist. These features have been shown to be distinguished by different TFs (Yang et al., 2014). Very recently, nine additional shape features were introduced to the repertoire in order to better describe the unique 3D structure encoded in a given DNA sequence (Li et al., 2017). Apart from ‘naked’ DNA shape features, DNA methylation on cytosine residues also affects DNA structure, making it a unique type of shape feature that could be recognized by many TFs (Lazarovici et al., 2013; Yin et al., 2017; Rao et al., 2018). Several TFBS modeling methods that take into account shape features (some combined with sequence-based features) have been developed, and show improvement compared with only sequence feature-based models (**Table 2**). However, these models use DNA shape information generated from computational simulations, such as Monte Carlo or Molecular Dynamics, and potential biases exist. Improvements have already been obtained by integrating DNA shape information derived from experimental data, such as X-ray crystallography (Li et al., 2017). Thus, one major challenge regarding incorporating shape features into TFBS modeling is to derive unbiased DNA structural data in a high-throughput manner that has been robustly verified experimentally, which currently is a challenge. The other challenge is that both DNA conformation (Azad et al., 2018) and TF conformation (Patel et al., 2018) could be changed in an adaptation mechanism upon

interacting with each other due to both protein and DNA plasticity. This makes integrating shape feature even more difficult as it changes dynamically.

Energy- and deep learning-based models

Energy based biophysical models are a powerful alternative to probabilistic models such as PWM. They use the action mass law to characterize amino-acid and DNA interaction and are valid on a wider range of protein concentrations than probabilistic models, that in fact represent an approximation of energy-based models. Whenever they can be built, and several methods exist based on PBM or SELEX-seq for example, they should be preferred without disadvantages except PWM are the simplest to build (Zhao et al., 2009; Stormo, 2013; Ruan and Stormo, 2017).

Machine learning methods, such as deep learning, are able to leverage very large datasets to discover intricate connections within them and make accurate predictions (Lecun et al., 2015). In the last few years, deep learning has been increasingly applied to resolve complex biological problems, including those from regulatory genomics (Angermueller et al., 2016). Several methods based on deep learning have been developed to model TF-DNA binding specificity or to predict TF *in vivo* binding, including DeepBind (Alipanahi et al., 2015), DeepSEA (Zhou and Troyanskaya, 2015), TFImpute (Qin and Feng, 2017), DeFind (Wang et al., 2018a) and DFIM (Greenside et al., 2018) (**Table 2**). Advantages of these models include for example, 1) they can be trained from various types of sequencing data in either alone or integrated manner, and can be further combined with other information, such as DNase I hypersensitivity data, for better *in vivo* TFBSs prediction (Zhou and Troyanskaya, 2015); 2) they can tolerate a certain degree of noise stemming from either data acquisition technology or sequencing biases; 3) they can train predictive models fully automatically, alleviating the need for time-consuming manual intervention and expertise; 4) they can accurately identify genomic variants in the regulatory region, and indicate how variations affect TF binding within a specific sequence. However, one of the yet to be tackled difficulties of deep learning models is that they are more difficult to interpret than PWMs given the hidden layers in the networks. More information of these models and their unique properties can be found in **Table 2**. To conclude, it is worth mentioning that, until now, no single model has been identified to be the best for all TFs and the nature of the most adequate model depends on the individual TF.

Link between models and TF 3D structure

TFBS models derived from NGS allow a broad overview of where TFs are able to bind and their sequence specificity. Structures of TF-DNA complexes provide complementary information by identifying the amino acids and specific bases involved in TF-DNA interactions. These structural data not only explain base and shape readout at the residue and even atomic level, but also allow for the prediction of how amino acid mutations and/or changes in a given *cis*-element will affect TF binding. Indeed, many diseases resulting from gene misregulation are due to either mutations in a TF or alterations in its binding site. Combining the “go broad” NGS approach with the “go deep” structural approach provides a powerful tool in refining TFBS and gene regulation models.

Recent modeling tools have attempted to use 3D structural data for improving predictions of TF-DNA binding and structure-based databases for TFBS data are currently available (Turner et al., 2012; Lin and Chen, 2013; Xu et al., 2013). Structure-based TFBS methods rely on different energy functions to score TF-DNA interactions. Such energy functions are used to describe all possible physiochemical interactions such as Van der Waals interactions, hydrogen bonding, electrostatic interactions and solvation energy. Energy functions can be divided into physics-based molecular mechanics force fields (Liu et al., 2009a; Marcovitz and Levy, 2011; Yin et al., 2015) and knowledge-based potentials (Liu et al., 2005; Zhang et al., 2005; Takeda et al., 2013). While physics-based energy functions are able to accurately describe TF-DNA interactions they have a high computational cost and thus are less often applied than knowledge-based potentials. In knowledge-based potentials, statistical analysis is used to describe TF-DNA interactions at the atom or residue scale using known TF-DNA structures. These are simpler and less computationally expensive than physics-based energy models. Recent work combining aspects of both types of models to derive an “integrative energy” function have also been applied to TF-DNA modeling and shown to further improve, in some cases, the predictive power of structure-based TFBS models (Farrel et al., 2016; Farrel and Guo, 2017).

A second way that 3D structural data can be used to help refining TFBS models is through the prediction of protein-protein interactions (PPIs), which may affect TF binding to DNA. Often *in vitro* TFBS models are relatively poor predictors of *in vivo*

TF binding due to the added complexity of interacting proteins *in vivo*. Pull-down assays, mass spectrometry and yeast two-hybrid allow the generation of at least a partial interacting network for a given TF (Yazaki et al., 2016; Trigg et al., 2017). These methods have limitations and often generate incomplete models due to the difficulty in determining true interaction partners and in detecting rare or transient interactions. Structural data can be incorporated to improve PPI models by providing quantitative parameters to determine whether a putative interaction is likely to occur based on energy calculations or homology modeling (Aloy and Russell, 2006; Beltrao et al., 2007). By adding partners to the simple TF-DNA model, differences between *in vitro* and *in vivo* binding are better accounted for and perturbations due to mutations, for example, can be more easily modeled as has been shown for mammalian TFs (Guturu et al., 2013). To our knowledge a full integration of structural data with TFBS models has not been implemented for plant TFs, however as many TF families are conserved across kingdom of life, suggesting these methods are applicable to plant TFs.

Improve the predictive power of TFBS models-genome context

Eukaryotic genomes contain numerous potential binding sites for a given TF, however, only a small fraction is actually bound *in vivo*, and that these sites vary substantially depending on contexts, such as cell types, developmental stages, and environmental or cellular conditions. In addition, only a subset of the bound sites drive transcription (Wasserman and Sandelin, 2004; Hu et al., 2007; Fisher et al., 2012; Whiteld et al., 2012). Therefore, various contexts have to be taken into account to predict functional TFBSs precisely. This includes chromatin states (such as accessibility and epigenetic marks), methylation states, nucleosome positioning and genome 3D structures, and combinatorial binding of TFs.

Nucleosome positioning, chromatin states and 3D genome

In the nucleus of eukaryotic cells, DNA wraps around histone proteins to form nucleosomes (McGinty and Tan, 2015), which can be further compacted into highly condensed structure called heterochromatin by various mechanisms (Allshire and Madhani, 2017). This involves factors like linker histones (Fyodorov et al., 2017), repressive histone marks (Allis and Jenuwein, 2016), such as H3K27me1/3 and H3K9me2, and DNA methylation on cytosine residues (Kim and Zilberman, 2014;

Zhu et al., 2016) among others. Thus, chromatin structure is intrinsically repressive, a mechanism that helps to establish stable gene expression and prevents unwanted cell fate transitions. For gene activation, eukaryotic cells evolved various counter mechanisms for each of the chromatin compacting factors to create accessible chromatin, such as active histone marks (e.g. H3K4me2/3 and H3K27ac), chromatin remodelers (Ho and Crabtree, 2010) and demethylation machineries (Wu and Zhang, 2014). The interplay between all these factors result in a highly dynamic chromatin environment, in which TFs have to find their cognate DNA binding sites.

Nucleosome positioning

In general, TFs preferentially bind to TFBSs in accessible chromatin regions, where nucleosomes are depleted (NDR-nucleosome depleted region). This is evidenced by large scale *cis*-element studies, which showed that the vast majority of the active *cis*-elements reside in the NDR in different species (Thurman et al., 2012; Weber et al., 2016), including Arabidopsis and maize (Zhang et al., 2012; Vera et al., 2014). Therefore, a precise *in vivo* TFBSs prediction model could integrate NDR as its first layer of filter to leave out sites/regions with well-positioned nucleosomes. Indeed, several TFBSs modeling methods that integrate DNase I hypersensitivity datasets, have shown increased prediction power for *in vivo* binding (Zhou and Troyanskaya, 2015; Kelley et al., 2016; Wang et al., 2018b). Thus, it is essential to generate datasets representing chromatin accessibility. To address this, recent technological advances are available, such as DNase-seq, MNase-seq, FAIRE-seq and ATAC-seq (Meyer and Liu, 2014). Among them, ATAC-seq is a rising star method as it requires a minimal amount of input sample and even can be used at the single-cell level (Buenrostro et al., 2013; Buenrostro et al., 2015; Corces et al., 2017). This is particularly attractive for the plant biology community, where some plant tissues are extremely scarce, such as flower meristem cells, organ primordia and root tips. Furthermore, when combined with INTACT (isolation of nuclei tagged in specific cell types), which allows isolation of nuclei from individual cell types of a tissue by affinity-based purification, ATAC-seq is able to map chromatin accessibility with high resolution and low noise from a specific cell type (Deal and Henikoff, 2011; Sijacic et al., 2018).

Although a majority of TFs favor binding in NDRs, exceptions exist. A special group of TFs, so-called pioneer factors, are able to bind TFBSs even when nucleosomes

are present (Iwafuchi-Doi and Zaret, 2016; Zaret and Mango, 2016; Zaret, 2018). As exemplified by FoxA1 and GATA4, pioneer factors are able to outcompete nucleosomes or create NDR through various mechanisms, such as mimicking linker histones, recruiting chromatin remodelers and/or depositing active epigenetic marks (Mayran and Drouin, 2018). Therefore, pioneer factors have to be considered with care while modeling their *in vivo* binding. One of the first reported plant pioneer factor was LEAFY COTYLEDON1 (LEC1), a seed-specific TF and a master regulator of embryogenesis. Tao et al. showed that LEC1 can target mitotically silenced chromatin at the loci of floral repressor *FLOWERING LOCUS C (FLC)* and promote the initial establishment of an active chromatin state (Tao et al., 2017). This activates FLC expression *de novo* in the pro-embryo and leads to the reversal of the silenced chromatin state inherited from gametes. Three TFs, LEAFY (LFY), APETALA1 (AP1) and SEP3, which are key factors in floral development in *Arabidopsis thaliana*, have been shown to be likely pioneer factors. A combination of ChIP-seq and DNase-seq data suggested that LFY is able to bind its TFBSs in closed chromatin, and this activity is highly correlated with its oligomerization activity. This is a potential novel driving force for pioneer activity which has not been reported in other organisms (Sayou et al., 2016). For AP1 and SEP3, it has been shown that upon binding to their TFBSs both TFs are able to confer chromatin accessibility near those sites (Pajoro et al., 2014). Interestingly, both factors are able to form higher order homo and hetero-oligomers, with such activity essential for their function *in vivo*. Therefore, an attracting hypothesis is that oligomerization likely confers high binding affinity in order for them to bind TFBSs that are otherwise inaccessible due to the occupancy of histones at these sites. Although further evidence of pioneer activity of these TFs, including both genome-wide and biochemical studies, are required, modeling their *in vivo* binding requires examination of both closed and open chromatin regions.

Chromatin states: histone modifications, histone variants and chromatin remodelers

TF binding *in vivo* confronts various chromatin states that are established by various types of histone modifications, histone variants and remodelers. Histone modifications act as either active or repressive marks, corresponding to transcriptionally competent and inactive chromatin, respectively. These marks are deposited by epigenetic writers and removed by epigenetic erasers. For instance, the

PRC2 is a writer responsible for H3K27me3 deposition while the REF6 demethylase erases this mark (Hennig and Derkacheva, 2009; Li et al., 2016). For some, if not all, epigenetic marks there is a corresponding epigenetic reader that reads the specific mark and confers downstream responses. Histone variants are also determinants of chromatin states and affect transcription. For instance, the H3.3 and H3.1 variants differ only four amino acid (Ingouff and Berger, 2010), and while H3.1 is enriched in heterochromatin and preferentially carries repressive H3K27 methylation marks, H3.3 is enriched in transcriptionally active regions and preferentially carries active H3K36 methylation marks (Johnson et al., 2004; Stroud et al., 2012). Chromatin remodelers, which use ATP energy to evict, disassemble or slide nucleosomes, are also landmarks affecting TF binding. The increasing datasets for genome-wide profiling of histone variants, marks, writers, erasers, readers and of chromatin remodelers thus constitutes a highly informative resource to improve TFBS prediction.

Cross-talk between TFs and chromatin factors co-regulate chromatin accessibility and exposure of *cis*-elements (Vachon et al., 2018). In these processes, TFs operate either by recruiting chromatin factors or directly competing with them for target sites. There are several examples of TF-mediated recruitment of chromatin factors in plants, such as that of REF6 by NF-Y TFs for H3K27 demethylation at *SOC1*, inducing flowering (Hou et al., 2014), or Polycomb mark reader TFL2/LHP1 recruitment by SHORT VEGETATIVE PHASE at *SEP3* for flower patterning (Liu et al., 2009b), or BRAHMA and SPLAYED ATPase recruitment by LFY and SEP3 at flower morphogenetic genes (Wu et al., 2012). Oppositely, several TFs were shown to compete with Polycomb complexes at target genes, such as NF-YC which prevents PRC2 binding to *FLOWERING LOCUS T* for floral transition (Liu et al., 2018) and AG which evicts PRC2 from *KNUCKLES* for flower meristem termination (Sun et al., 2014). Interestingly, at the time of flower termination, AG also has the opposite effect at *WUS*, promoting PRC2 recruitment for deposition of H3K27me3 (Liu et al., 2011). Differences in TF behaviour for eviction versus recruitment of PRC2 may depend on the distance between the Polycomb recognition element (PRE) and the TFBS. To this regard, large-scale analyses of ChIP-seq data revealed TFBSs in plant PREs, thereby expanding the repertoire of TF-chromatin factor interactions and providing resources for further exploration of the relationship between *cis*-elements and TF/chromatin factor binding (Wang et al., 2016; Xiao et al., 2017). Taken

together, intricate and dynamic interplays among TFs and chromatin factors have to be carefully examined for TF binding *in vivo* as they define chromatin state of a region, where TFs in turn have to engage with.

Methylation state

DNA methylation at the 5' position of cytosine plays an essential role in gene regulation and genome stability in plants and animals (Zhang et al., 2018b). Precise patterns of DNA methylation are crucial for plant growth and development, including flowering (Finnegan et al., 1998). Unlike animals, in which DNA methylation are predominantly found in the CG context, plant DNA methylation occurs in contexts including CG, CHG and CHH (H represents A, T or C) (Zhang et al., 2006; Lister et al., 2008). Most TFs favor not to bind to methylated TFBSs due to the fact that DNA methylation affects shape features of TFBSs and that methyl groups often clash with residues that form direct interactions with otherwise unmethylated DNA motifs. Interestingly, recent studies have revealed that some TFs preferentially bind to methylated DNA (Zhu et al., 2016; Yin et al., 2017; Zuo et al., 2017). In addition, these TFs seemed to be enriched in embryonic and organismal development, such as homeodomain TFs and pluripotent factors (e.g. OCT4), which are well-characterized pioneer factors. Although proteins that specifically bind to methylated DNA are found in plants as exemplified by Methyl-CpG-binding domain proteins (Zemach and Grafi, 2007), they are not classified as TFs but epigenetic modifiers. To our knowledge, TFs that are insensitive to methylation have not yet been reported in plants, however, it is appealing to investigate whether aforementioned potential plant pioneer factors (i.e. LEC1, LFY, SEP3 and AP1) are insensitive to methylation. Another mechanism that affects TF binding is that widespread DNA methylation promotes repressive histone modifications such as H3K9me2 and inhibits permissive histone modifications such as histone acetylation, resulting in highly compacted heterochromatin (Zhang et al., 2018b), thus inaccessible to vast majority of the TFs, except pioneer factors. Taken together, traditional views suggested that methylation seem to inhibit TF binding to TFBSs, however, there are likely at least a subset of TFs, such as pioneer factors, that can target methylated sites. Therefore, their DNA binding affinity and specificity needs to be carefully examined with regard to prediction of their *in vivo* binding. There are several methods that are available to model the effects of DNA methylation, such as Cytomod (Viner et al., 2016) (Table

2), which uses the classical PWM approach with an extended alphabet (e.g. 5mC representing methylated cytosine). In some practices, multiple PWM logos are given for the same TF, for which the enriched methylated and non-methylated sequences are represented separately (Yin et al., 2017). In addition, apart from aforementioned DAP-seq, two additional experimental approaches are now available to investigate the effects of DNA methylation to TF binding *in vitro*, including Methyl-Spec-seq (Zuo et al., 2017) and methyl-SELEX (Yin et al., 2017) (details refer to Table 1).

3D genome and TF mediated long-range gene interactions

The linear nucleotide sequences are folded into highly organized 3D architectures in the nucleus of higher eukaryotes. Chromosome conformation capture (3C) techniques, such as Hi-C (Eagen, 2018), revealed widespread existence of long-range gene interactions within the so-called topologically associating domains (TAD). Within the TAD, distal and proximal *cis*-elements relative to transcription starting sites (TSS) form cell-type specific long-range interactions that in many cases are established by architectural proteins such as cohesin (Yan et al., 2013; Rao et al., 2017), CTCF (Phillips and Corces, 2009; Ren et al., 2017), Yin Yang 1 (Weintraub et al., 2018) and others (Rada-Iglesias et al., 2018). Such interaction is a highly conserved mechanism for eukaryotes to achieve spatiotemporal gene expression (Sanyal et al., 2012; Harmston and Lenhard, 2013; Dekker and Misteli, 2015). TADs therefore form territories within which more frequent gene interaction occurs, whereas less interaction happens beyond these territories. Disruption of TAD boundaries can lead to ectopic activation of gene expression and eventually to noticeable phenotypes (Lupiáñez et al., 2015; Franke et al., 2016; Lupiáñez et al., 2016).

Although it seems that *Arabidopsis thaliana* does not form TADs likely due to not having the architectural proteins such as CTCF that is important for TAD maintenance (Liu et al., 2017), in many other plant species, such as maize, rice, and tomato, TADs are clearly detected according to Hi-C data (Dong et al., 2017). Nevertheless, long-range gene interactions are still widespread in the Arabidopsis genome but in a less compartmentalized manner compared with other plant species (Liu et al., 2016). Apart from architectural proteins, TFs are usually the links mediating cell-type specific long-range gene interactions, for which the transactivation domains (TD) found in majority of TFs appear to play an essential

role. In general, TDs are enriched with acidic and hydrophobic residues, and residues that are able to form intrinsically disorder structures, such as serine, glycine and proline (Staller et al., 2018). These properties appear to allow TDs to interact with or recruit various factors with modest affinity but high specificity under various contexts. One such factor is Mediator, a mega protein complex that can be recruited by divergent TFs to connect distal and proximal *cis*-elements (Soutourina, 2017). Another factor is the SAGA complex, which has recently been shown to be a general factor that is required for the construction of the pre-initiation complex at the TSS for transcription initiation in animal systems (Baptista et al., 2017). Despite being less well characterized in plants, homolog protein components for both factors are well conserved in plants (Elfving et al., 2011; Mathur et al., 2011; Moraga and Aquea, 2015).

With the accumulation of datasets from Hi-C and related methods, it is now possible to predict spatiotemporal TF binding more precisely. In plants, Hi-C has been carried out from species including *Arabidopsis thaliana* (Liu et al., 2016), rice (*Oryza sativa*) (Dong et al., 2018), barley (*Hordeum vulgare*) (Mascher et al., 2017), tomato (*Solanum lycopersicum*), maize (*Zea mays*), sorghum (*Sorghum bicolor*), foxtail millet (*Setaria italica*) (Dong et al., 2017) and cotton (*Gossypium spp.*) (Wang et al., 2017; Wang et al., 2018c). Considering that chromosome conformation is highly cell-type specific, these Hi-C datasets have to be carefully examined when applied to other cell types. For example, in the process of flowering initiation, it is more relevant to perform Hi-C using flower meristems at a certain stage in order to map long-range gene interactions of the stage. 3C assays can also be performed for a specific TF using chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) (Fullwood et al., 2009; Li et al., 2014) and HiChIP (Mumbach et al., 2016). These methods combine ChIP with 3C to produce a directed view of long-range interactions associated with a TF of interest. To our knowledge, ChIA-PET or HiChIP has not yet been applied in floral TFs despite its high potential to correlate chromatin 3D structure with TF binding. For example, MADS-box TF homo- or hetero-tetramer complex has been shown to bind to two CArG boxes in short linear distance to form loops that are essential for target gene expression (Melzer et al., 2009; Mendes et al., 2013). However, it is not clear if MADS-box TFs (or other oligomeric TFs) also enable long-range looping or even cause 3D chromatin structural rearrangement,

such as breaking TAD boundaries as shown for Yamanaka factors during cell fate reprogramming (Stadhouders et al., 2018). These are potential mechanisms that could explain functional diversity of MADS-box TFs and their potential pioneer activity in flower organ specification, respectively, for which ChIA-PET or HiChIP might provide valuable insight.

TF cooperative binding

Cooperative binding affects TF-DNA affinity and specificity. It is a widespread mechanism in eukaryotes for maximizing TF functional complexity by utilizing the minimum number of TFs. For example, Hox TFs in *Drosophila* bind highly similar sequences as monomers, whereas heterodimerization with the cofactor Extradenticle from the same TF family evokes significant differences in DNA binding affinity and specificity as revealed by SELEX-seq (Slattery et al., 2011). In plants, MADS-box TFs are prominent examples of cooperative binding. They form heterotetrametric complexes, so-called floral quartets, to regulate distinct set of genes in the processes of flower formation and flower organ specification (Ruelens et al., 2017; Hugouvieux and Zubieta, 2018). It has been shown, both *in vivo* and *in vitro*, that different combinations of MADS-box TFs confer unique DNA binding specificity and affinity (Smaczniak et al., 2012; Muiño et al., 2014; Smaczniak et al., 2017; Hugouvieux et al., 2018).

In some cases, a co-factor can be a non-DNA binding protein. For instance, two non-DNA-binding cofactors in yeast, MET4 and MET28, enhance DNA-binding specificity of TF Cbf1 through forming MET4-MET28-Cbf1 complex, which is required for activation of downstream genes (Siggers et al., 2011). In plants, the Evening Complex (EC), consists of ARRHYTHMO (LUX), EARLY FLOWERING 3 (ELF3) and ELF4, is a key component of the circadian clock (Greenham and McClung, 2015; Huang and Nusinow, 2016). While only LUX is a TF, the *in vivo* functioning of the EC in the process of temperature and circadian clock-dependent flowering pathway requires non-DNA binding cofactors ELF3 and ELF4 (Nusinow et al., 2011). Furthermore, ChIP-seq data showed that G-box motifs are enriched adjacent to LUX binding sites, indicating additional cofactors that likely co-bind with EC to obtain further specificity and cooperativity for transcriptional regulation (Ezer et al., 2017). Similar mechanisms have also been proposed for PHYTOCHROME INTERACTING FACTOR 4 (PIF4), a key TF involved in thermoresponsive flowering in Arabidopsis.

Its DNA binding activity can be sequestered by ELF3 (Nieto et al., 2015), or abrogated by DELLA proteins (Lucas et al., 2008), both through direct physical interactions. Taken together, it is crucial to taken into account of presence or absence of TF cofactors for *in vivo* binding prediction.

Incorporating TFBS models in current and future analyses of gene regulation

The capacity to detect TFBS, both *in vitro* and *in vivo*, in increasingly reliable ways offers the opportunity to better answer various types of biological questions. For example, it is now possible to manipulate TFBS with genome editing, study the way how TFBS are evolving, better predict gene regulation and understand the DNA recruitment of chromatin regulators.

From DNA binding to gene regulation and to regulatory networks

Once TFBSs are identified or reliably predicted, the next challenge is to understand whether, how and in which cellular context TF binding results in changes of target gene expression. Here, one can distinguish dedicated analyses of individual binding events and potential target genes (‘bottom up’) or make use of genome-wide expression data followed by mathematical modeling (‘top down’).

A classical and powerful way to identify regulators of a given biological process or developmental transition consists of building lists of co-regulated genes and identifying *cis*-elements overrepresented in their promoters. Once identified, these motifs can be compared to motifs in TFBS databases to identify TFs or TF families that are candidate regulators. Combined with detailed TF expression data, this represents a way to identify regulators. Several bioinformatics tools were developed based on this approach, such as Cistome (Austin et al., 2016), PlantRegMap (Jin et al., 2017) and TF2Network (Kulkarni et al., 2017). These tools take as an input a set of genes for which predicted regulators are searched. As a result, a set of potential regulators is identified and can be further validated using experimental approaches. In the example of TF2Network, using a standard dataset based on experimental TF binding data revealed that it recovers 92% of the true regulators using the long region promoter definition and the overall of 56% of the correct regulators when fed with a set of differentially expressed genes (Kulkarni et al., 2017). In a related approach,

mathematical modeling using gene expression data can reveal gene network modules, and knowledge from known TF binding preferences can be used to validate predicted key gene-regulatory interactions (Ichihashi et al., 2014).

TFBS models are now widely applied not only to characterize gene regulatory networks (GRN), but also to understand mechanisms underlying gene activation or repression. For example, TFBS prediction helped to identify TFs that mediate recruitment of repressive Polycomb protein complexes to specific genomic locations (Xiao et al., 2017; Zhou et al., 2018). A major challenge is still to identify and validate cell type-specific gene regulatory interactions, which can now be addressed by combining cell-type selection by Fluorescence Activated Cell/Nuclei sorting or INTACT with ChIP-seq or other epigenomic technologies (see review (Wang et al., 2012)). Another promising technology is single-cell approaches (see review (Grün and Van Oudenaarden, 2015)), which at the moment are still under development for plant tissues (Brennecke et al., 2013).

Targeted TFBS perturbation

Testing the functional impact of a predicted TFBS usually involves targeted mutagenesis in a transgenic context, e.g. using reporter assays, or in an endogenous context using CRISPR-Cas9-based systems. By using reporter genes (e.g. GFP or luciferase) under control of the target gene regulatory regions with modified TFBSs it is possible to dissect spatiotemporal and quantitative changes in target gene expression depending on the presence or absence of a TFBS (Benn and Dehesh, 2016; Díaz-Triviño et al., 2017). For example, the tissue specificity of the *AP3* promoter was altered by replacing native TFBS with the ones of predicted specificity towards SEP3-AG or SEP3-AP1 floral homeotic protein complexes (Smaczniak et al., 2017). Combining results from TFBS prediction and reporter gene assays also helps reveal the mechanisms of TFBS recruitment in a native promoter context. For example, in Drosophila the combination of SELEX and reporter gene expression (lacZ and GFP) experiments revealed that clusters of low affinity binding sites are maintained and required for the proper tissue-specific expression of the Hox genes, homeotic genes crucial for segment specification (Crocker et al., 2015). To which extent the clusters of TFBSs regulate gene expression in plants is yet to be determined through similar approaches.

With the advent of new technologies such as CRISPR-Cas9, mutations can now be introduced in endogenous genomic locations. For example, when a regulatory region of *AG* gene located in the second intron was deleted by the CRISPR-Cas9, mutant plants show partial homeotic transformations of stamens to petals, supporting an important role of this regulatory region (Yan et al., 2016). One of the challenges for CRISPR-Cas9 strategy is that it requires TFBSs that contain a PAM motif 5'-NGG-3' for efficient cleavage. Thus, the generation of new versions of Cas9 with different sequence requirement or the possibility to perform directed mutagenesis with a template DNA will open new avenues for precise TFBS perturbation. Moreover, the modifications of the CRISPR-Cas9 system by fusing cytidine deaminases to catalytically inactive Cas9 allow for the targeted, programmable single nucleotide changes within a TFBS of interest (Yan et al., 2016). This is a paradigm shift as genetics until now has mainly challenged regulatory networks by modifying their nodes-the protein coding genes, and TFBS mutations will allow challenging the links without compromising all functions of a potentially pleiotropic TF. Recently, Barakat et al., reported an assay that combines ChIP and a massively parallel reporter assay (ChIP-STARR-seq) to identify functional TFBS in primed and naive human embryonic stem cells (Barakat et al., 2018) (Table 1). The resulting functional TFBSs of a given TF were further validated by CRISPR-Cas9 followed by a transient expression assay, proving the robustness of such method. This method is potentially applicable in plants, but with a limitation that maintaining and transfection of plant cells in culture (e.g. leaf protoplast) of a stage of interest is more challenging than that of mammalian system.

Evolution of TFBS and plant diversity

Studying the conservation of *cis*-elements containing TFBSs between different species or between promoters of closely related paralogous genes in a genome can shed light on the evolution of a GRN. How changes in *cis*-elements relate to alterations in the expression pattern of a gene and subsequently lead to novel gene functions is not yet fully understood. At another level, understanding the evolutionary dynamics of gene-regulatory interactions can provide deeper insights into how developmental programs evolve. The first step into this direction is to develop experimental approaches to study TFBSs in different species. In a genome-wide comparative ChIP-seq study, Muino et al. studied binding of SEP3 in two closely

related Arabidopsis species with similar flower morphology (Muino et al., 2016). They found that TF binding conservation was associated with sequence conservation of CArG-box motifs and with the relative position of the TFBS to its potential target gene, and that loss/gain of binding sites tended to be associated with changes in gene expression. Their study revealed clear differences in SEP3-bound regions between the two species. A high level of binding divergence (13 % overlap) was also reported for two orthologous MADS-box TFs, FLC in Arabidopsis and PERPETUAL FLOWERING1 in *Arabis alpine* (Mateos et al., 2017). Therefore, comparative ChIP-seq studies can indicate conserved core target gene networks of developmental TFs in plants and distinguish plant lineage-specific functions and potentially less relevant binding sites. Interestingly, similar observation was also reported in animals, as revealed by ChIP-seq on livers of five vertebrates (Schmidt et al., 2010). The authors observed highly conserved TFBS motif for two TFs, but highly divergent binding events on conserved genes of different species.

Besides genome-wide approaches, targeted analysis of individual promoters or regulatory regions can elucidate regulatory divergence after speciation or gene duplication. For example, absence/presence of a single CArG-box in the promoter regions of the two MADS-box TF paralogs, *AP1* and *CAL,* determines spatiotemporal and quantitative differences in gene activity (Ye et al., 2016). Studying the TFBSs of homologues from different species can also reveal the evolution of their molecular function. For example, analyzing TF binding specificity of the LFY homologues from different plant species, land plants, mosses and algae, has revealed subtle changes in their preferred TFBSs motifs, suggesting that LFY DNA binding specificity changed during land plant evolution (Sayou et al., 2014). Thus, the combination of numerous TFBS models with novel genome sequences could ultimately unlock mechanisms of GRN evolution.

Future perspectives

Technological advances such as NGS has revolutionized TFBS identification. It is now possible to identify TFBSs for hundreds of TFs for any organism in a short time with limited costs. However, accurate quantitative TFBS modeling is not a trivial process and seems to lag behind the pace of NGS dataset generation. Innovative analyses will be required to better extract valuable information hidden in datasets and

compensate for the biases and drawbacks inherent to each particular method. As an example, by combining DAP-seq data for auxin response factor 5 (ARF5), list of auxin induced genes and PWM model for dimeric ARF binding, we have uncovered a new ARF binding site configuration (inverted repeat 13) that seems to favor regulation and was not noticed before ((Stigliani et al., 2018); co-submitted paper). This experience highlights that re-examination of the publicly available datasets could lead to novel findings, and that TFBS modeling requires careful planning and implementation. Therefore, an important future goal is full automation for TFBS modeling. In this regard, emerging artificial intelligence and machine learning are projected to make important contributions. Indeed, machine learning approaches have already played an important role in TFBS modeling and shown to be more powerful than conventional algorithms in several aspects as discussed earlier. Finally, integrating datasets from chromatin environments, such as chromatin accessibility and 3D genome maps, is of great importance to better predict TF binding and the concomitant transcriptional events. A dedicated and accessible tool that could integrate such complex datasets in a combinatorial way is still lacking and will likely be an important focus of future investigations.

FUNDING

This work was supported by Agence Nationale de la Recherche (project FloPiNet to C.Z., K.K., and F.P.) the Grenoble Alliance for Cell and Structural Biology [ANR-10-LABX-49-01 to FP, CZ and AS], and Action Thématique et Incitative sur Programme (ATIP)-Avenir (to C.Z.).

AUTHOR CONTRIBUTIONS

F.P., C.Z., and X.L. planned the review outline and content. X.L. wrote the manuscript with inputs from all authors. All authors contributed to the reviewing and editing of the manuscript.

CONFLICT OF INTEREST

We declare no conflict of interest.

References

Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H. J., Rohs, R., and Mann, R. S. (2015). Deconvolving the recognition of DNA shape from sequence. *Cell* **161**:307–318.

Afek, A., Schipper, J. L., Horton, J., Gordân, R., and Lukatsky, D. B. (2014). Protein−DNA binding in the absence of specific base-pair recognition. *Proceedings of the National Academy of Sciences* **111**:17140–17145.

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* **33**:831–838.

Allen, B. L., and Taatjes, D. J. (2015). The Mediator complex: a central integrator of transcription. *Nature Reviews Molecular Cell Biology* **16**:155–166.

Allis, C. D., and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nature Reviews Genetics* **17**:487–500.

Allshire, R. C., and Madhani, H. D. (2017). Ten Principles of Heterochromatin Formation and Function. *Nature Publishing Group* Advance Access published 2017, doi:10.1038/nrm.2017.119.

Aloy, P., and Russell, R. B. (2006). Structural systems biology: Modelling protein interactions. *Nature Reviews Molecular Cell Biology* **7**:188–197.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology* **12**:878.

Austin, R. S., Hiu, S., Waese, J., Ierullo, M., Pasha, A., Wang, T. T., Fan, J., Foong, C., Breit, R., Desveaux, D., et al. (2016). New BAR tools for mining expression data and exploring Cis-elements in Arabidopsis thaliana. *Plant Journal* **88**:490–504.

Azad, R. N., Zafiropoulos, D., Ober, D., Jiang, Y., Chiu, T.-P., Sagendorf, J. M., Rohs, R., and Tullius, T. D. (2018). Experimental maps of DNA structure at nucleotide resolution distinguish intrinsic from protein-induced DNA deformations. *Nucleic Acids Research* **46**:2636–2647.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology* **14**:283–291.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science* **324**:1720–1723.

Baptista, T., Grünberg, S., Minoungou, N., Koster, M. J. E., Timmers, H. T. M., Hahn, S., Devys, D., and Tora, L. (2017). SAGA Is a General Cofactor for

RNA Polymerase II Transcription. *Molecular Cell* Advance Access published 2017, doi:10.1016/j.molcel.2017.08.016.

Barakat, T. S., Halbritter, F., Zhang, M., Rendeiro, A. F., Perenthaler, E., Bock, C., and Chambers, I. (2018). Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23**:276-288.e8.

Bartlett, A., O'Malley, R. C., Huang, S. C., Galli, M., Nery, J. R., Gallavotti, A., and Ecker, J. R. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nature Protocols* **12**:1659–1672.

Beltrao, P., Kiel, C., and Serrano, L. (2007). Structures in systems biology. *Current Opinion in Structural Biology* **17**:378–384.

Benn, G., and Dehesh, K. (2016). Quantitative Analysis of Cis -Regulatory Element Activity Using Synthetic Promoters in Transgenic Plants. *Methods in Molecular Biology* **1482**:15–30.

Berger, M. F., and Bulyk, M. L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the dna-binding specificities of transcription factors. *Nature Protocols* **4**:393–411.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology* **24**:1429–1435.

Bezhani, S., Winter, C., Hershman, S., Wagner, J. D., Kennedy, J. F., Kwon, C. S., Pfluger, J., Su, Y., and Wagner, D. (2007). Unique, Shared, and Redundant Roles for the Arabidopsis SWI/SNF Chromatin Remodeling ATPases BRAHMA and SPLAYED. *The Plant Cell* **19**:403–416.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* **10**:1093–1098.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**:1213–1218.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**:486–490.

Bulyk, M. L. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research* **30**:1255–1261.

Chen, Y., Zhang, X., Dantas Machado, A. C., Ding, Y., Chen, Z., Qin, P. Z., Rohs, R., and Chen, L. (2013). Structure of p53 binding to the BAX response

element reveals DNA unwinding and compression to accommodate base-pair insertion. *Nucleic Acids Research* **41**:8368–8376.

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods* Advance Access published 2017, doi:10.1038/nmeth.4396.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., et al. (2015). Low affinity binding site clusters confer HOX specificity and regulatory robustness. *Cell* **160**:191–203.

Deal, R. B., and Henikoff, S. (2011). The INTACT method for cell type–specific gene expression and chromatin profiling in Arabidopsis thaliana. *Nature Protocols* **6**:56–68.

Dekker, J., and Misteli, T. (2015). Long-Range Chromatin Interactions. *Cold Spring Harbor Perspectives in Biology* **7**:1–23.

Díaz-Triviño, S., Long, Y., Scheres, B., and Blilou, I. (2017). Analysis of a Plant Transcriptional Regulatory Network Using Transient Expression Systems. *Methods in Molecular Biology* **1629**:83–104.

Dong, P., Tu, X., Chu, P. Y., Lü, P., Zhu, N., Grierson, D., Du, B., Li, P., and Zhong, S. (2017). 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. *Molecular Plant* **10**:1497–1509.

Dong, Q., Li, N., Li, X., Yuan, Z., Xie, D., Wang, X., Li, J., Yu, Y., Wang, J., Ding, B., et al. (2018). Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice. *The Plant Journal* Advance Access published 2018, doi:10.1111/tpj.13925.

Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Research* **25**:1268–1280.

Eagen, K. P. (2018). Principles of Chromosome Architecture Revealed by Hi-C. *Trends in Biochemical Sciences* **43:469-478**.

Elfving, N., Davoine, C., Benlloch, R., Blomberg, J., Brannstrom, K., Muller, D., Nilsson, A., Ulfstedt, M., Ronne, H., Wingsle, G., et al. (2011). The Arabidopsis thaliana Med25 mediator subunit integrates environmental cues to control plant development. *Proceedings of the National Academy of Sciences* **108**:8245–8250.

Ezer, D., Jung, J.-H., Lan, H., Biswas, S., Gregoire, L., Box, M. S., Charoensawan, V., Cortijo, S., Lai, X., Stöckle, D., et al. (2017). The evening complex coordinates environmental and endogenous signals in Arabidopsis. *Nature Plants* **3**:17087.

Farrel, A., and Guo, J. (2017). An efficient algorithm for improving structure-based prediction of transcription factor binding sites. *BMC Bioinformatics* **18**:342.

Farrel, A., Murphy, J., and Guo, J. T. (2016). Structure-based prediction of transcription factor binding specificity using an integrative energy function. *Bioinformatics* **32**:i306–i313.

Finnegan, E. J., Genger, R. K., Kovac, K., Peacock, W. J., and Dennis, E. S. (1998). DNA methylation and the promotion of flowering by vernalization. *Proceedings of the National Academy of Sciences* **95**:5824–5829.

Fisher, W. W., Li, J. J., Hammonds, A. S., Brown, J. B., Pfeiffer, B. D., Weiszmann, R., MacArthur, S., Thomas, S., Stamatoyannopoulos, J. A., Eisen, M. B., et al. (2012). DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proceedings of the National Academy of Sciences* **109**:21330–21335.

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W. L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**:265–269.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. Bin, Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., et al. (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**:58–64.

Fyodorov, D. V., Zhou, B.-R., Skoultchi, A. I., and Bai, Y. (2017). Emerging roles of linker histones in regulating chromatin structure and function. *Nature Reviews Molecular Cell Biology* Advance Access published 2017, doi:10.1038/nrm.2017.94.

Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M. A. (2014). Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS Computational Biology* **10**.

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M. L. (2013). Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Reports* **3**:1093–1104.

Greenham, K., and McClung, C. R. (2015). Integrating circadian dynamics with physiological processes in plants. *Nature Reviews Genetics* **16**:598–610.

Greenside, P. G., Shimko, T., Fordyce, P., and Kundaje, A. (2018). Discovering epistatic feature interactions from neural network models of regulatory DNA sequences Advance Access published July 26, 2018, doi:10.1101/302711.

Grün, D., and Van Oudenaarden, A. (2015). Design and Analysis of Single-Cell Sequencing Experiments. *Cell* **163**:799–810.

Guo, Y., Tian, K., Zeng, H., Guo, X., and Gifford, D. K. (2018). A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. *Genome Research* Advance Access published 2018, doi:10.1101/130815.

Gutin, J., Sadeh, R., Bodenheimer, N., Joseph-Strauss, D., Klein-Brill, A., Alajem, A., Ram, O., and Friedman, N. (2018). Fine-Resolution Mapping of TF Binding and Chromatin Interactions. *Cell Reports* **22**:2601–2614.

Guturu, H., Doxey, A. C., Wenger, A. M., and Bejerano, G. (2013). Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**:20130029.

Hancock, S. P., Ghane, T., Cascio, D., Rohs, R., Di Felice, R., and Johnson, R. C. (2013). Control of DNA minor groove width and Fis protein binding by the purine 2-amino group. *Nucleic Acids Research* **41**:6750–6760.

Harmston, N., and Lenhard, B. (2013). Chromatin and epigenetic features of long-range gene regulation. *Nucleic Acids Research* **41**:7185–7199.

Hennig, L., and Derkacheva, M. (2009). Diversity of Polycomb group complexes in plants: same rules, different players? *Trends in Genetics* **25**:414–423.

Ho, L., and Crabtree, G. R. (2010). Chromatin remodelling during development. *Nature* **463**:474–484.

Hombach, D., Schwarz, J. M., Robinson, P. N., Schuelke, M., and Seelow, D. (2016). A systematic, large-scale comparison of transcription factor binding site models. *BMC Genomics* **17**:1–10.

Hou, X., Zhou, J., Liu, C., Liu, L., Shen, L., and Yu, H. (2014). Nuclear factor Y-mediated H3K27me3 demethylation of the SOC1 locus orchestrates flowering responses of Arabidopsis. *Nature Communications* **5**:1–14.

Hu, Z., Killion, P. J., and Iyer, V. R. (2007). Genetic reconstruction of a functional transcriptional regulatory network. *Nature Genetics* **39**:683–687.

Hu, M., Yu, J., Taylor, J. M. G., Chinnaiyan, A. M., and Qin, Z. S. (2010). On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Research* **38**:2154–2167.

Huang, H., and Nusinow, D. A. (2016). Into the Evening: Complex Interactions in the Arabidopsis Circadian Clock. *Trends in Genetics* **32**:674–686.

Hugouvieux, V., and Zubieta, C. (2018). MADS transcription factors cooperate : complexities of complex formation. *Journal of Experimental Botany* Advance Access published 2018, doi:10.1093/jxb/ery099.

Hugouvieux, V., Silva, C. S., Jourdain, A., Stigliani, A., Charras, Q., Conn, V., Conn, S. J., Carles, C. C., Parcy, F., and Zubieta, C. (2018). Tetramerization of MADS family transcription factors SEPALLATA3 and AGAMOUS is required

for floral meristem determinacy in Arabidopsis. *Nucleic Acids Research* Advance Access published 2018, doi:10.1093/nar/gky205.

Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., and Bulyk, M. L. (2015). UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research* **43**:D117–D122.

Ichihashi, Y., Aguilar-Martinez, J. A., Farhi, M., Chitwood, D. H., Kumar, R., Millon, L. V., Peng, J., Maloof, J. N., and Sinha, N. R. (2014). Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape. *Proceedings of the National Academy of Sciences* **111**:E2616–E2621.

Ingouff, M., and Berger, F. (2010). Histone3 variants in plants. *Chromosoma* **119**:27–33.

Isakova, A., Groux, R., Imbeault, M., Rainer, P., Alpern, D., Dainese, R., Ambrosini, G., Trono, D., Bucher, P., and Deplancke, B. (2017). SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nature Methods* **14**:316–322.

Iwafuchi-Doi, M., and Zaret, K. S. (2016). Cell fate control by pioneer transcription factors. *Development* **143**:1833–1837.

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., and Gao, G. (2017). PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research* **45**:D1040–D1045.

Johnson, L., Mollah, S., Garcia, B. A., Muratore, T. L., Shabanowitz, J., Hunt, D. F., and Jacobsen, S. E. (2004). Mass spectrometry analysis of Arabidopsis histone H3 reveals distinct combinations of post-translational modifications. *Nucleic Acids Research* **32**:6511–6518.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**:1497–1502.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpa, M. J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research* Advance Access published 2010, doi:10.1101/gr.100552.109.

Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**:384–388.

Käppel, S., Melzer, R., Rümpler, F., Gafert, C., and Theißen, G. (2018). The Floral Homeotic Protein SEPALLATA3 Recognizes Target DNA Sequences By Shape Readout Involving A Conserved Arginine Residue In The MADS-

Domain. *The Plant Journal* Advance Access published 2018, doi:10.1101/133678.

Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K., and Henikoff, S. (2014). High-resolution mapping of transcription factor binding sites on native chromatin. *Nature Methods* **11**:203–209.

Kaufmann, K., Muiño, J. M., Østerås, M., Farinelli, L., Krajewski, P., and Angenent, G. C. (2010). Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nature Protocols* **5**:457–472.

Kelley, D. R., Snoek, J., and Rinn, J. L. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research* **26**:990–999.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., Van Der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., et al. (2018). JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research* **46**:D260–D266.

Kiesel, A., Roth, C., Ge, W., Wess, M., Meier, M., and Söding, J. (2018). The BaMM web server for de-novo motif discovery and regulatory sequence analysis. *Nucleic Acids Research* **46**:W215–W220.

Kim, M. Y., and Zilberman, D. (2014). DNA methylation as a system of plant genomic immunity. *Trends in Plant Science* **19**:320–326.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell* **155**:27–38.

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., and Makeev, V. J. (2013). HOCOMOCO: A comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Research* **41**:195–202.

Kulkarni, S. R., Vaneechoutte, D., Van de Velde, J., and Vandepoele, K. (2017). TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Research* **46**.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The Human Transcription Factors. *Cell* **172**:650–665.

Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A. C., Riley, T. R., Sandstrom, R., Sabo, P. J., Lu, Y., Rohs, R., Stamatoyannopoulos, J. A., et al. (2013). Probing DNA shape and methylation state on a genomic scale with DNase I. *Proceedings of the National Academy of Sciences* **110**:6376–6381.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* **521**:436–444.

Lelli, K. M., Slattery, M., and Mann, R. S. (2012). Disentangling the Many Layers of Eukaryotic Transcriptional Regulation. *Annual Review of Genetics* **46**:43–68.

Levo, M., Zalckvar, E., Sharon, E., Machado, A. C. D., Kalma, Y., Lotam-Pompan, M., Weinberger, A., Yakhini, Z., Rohs, R., and Segal, E. (2015). Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research* Advance Access published 2015, doi:10.1101/gr.185033.114.6.

Li, G., Cai, L., Chang, H., Hong, P., Zhou, Q., Kulakova, E. V., Kolchanov, N. A., and Ruan, Y. (2014). Chromatin interaction analysis with paired-end tag (ChIA-PET) sequencing technology and application. *BMC Genomics* **15**:1–10.

Li, C., Gu, L., Gao, L., Chen, C., Wei, C. Q., Qiu, Q., Chien, C. W., Wang, S., Jiang, L., Ai, L. F., et al. (2016). Concerted genomic targeting of H3K27 demethylase REF6 and chromatin-remodeling ATPase BRM in Arabidopsis. *Nature Genetics* **48**:687–693.

Li, J., Sagendorf, J. M., Chiu, T. P., Pasi, M., Perez, A., and Rohs, R. (2017). Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic acids research* **45**:12877–12887.

Lin, C. K., and Chen, C. Y. (2013). PiDNA: Predicting protein-DNA interactions with structural models. *Nucleic acids research* **41**:523–530.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133**:523–536.

Liu, Z., Mao, F., Guo, J., Yan, B., Wang, P., Qu, Y., and Xu, Y. (2005). Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. *Nucleic Acids Research* **33**:546–558.

Liu, H., Shi, Y., Chen, X. S., and Warshel, A. (2009a). Simulating the electrostatic guidance of the vectorial translocations in hexameric helicases and translocases. *Proceedings of the National Academy of Sciences* **106**:7449–7454.

Liu, C., Xi, W., Shen, L., Tan, C., and Yu, H. (2009b). Regulation of Floral Patterning by Flowering Time Genes. *Developmental Cell* **16**:711–722.

Liu, X., Kim, Y. J., Muller, R., Yumul, R. E., Liu, C., Pan, Y., Cao, X., Goodrich, J., and Chen, X. (2011). AGAMOUS Terminates Floral Stem Cell Maintenance in Arabidopsis by Directly Repressing WUSCHEL through Recruitment of Polycomb Group Proteins. *The Plant Cell* **23**:3654–3670.

Liu, C., Wang, C., Wang, G., Becker, C., Zaidem, M., and Weigel, D. (2016). Genome-wide analysis of chromatin packing in *Arabidopsis thaliana* at single-gene resolution. *Genome Research* Advance Access published 2016, doi:10.1101/gr.204032.116.

Liu, C., Cheng, Y. J., Wang, J. W., and Weigel, D. (2017). Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nature Plants* **3**:742–748.

Liu, X., Yang, Y., Hu, Y., Zhou, L., Li, Y., and Hou, X. (2018). Temporal-Specific Interaction of NF-YC and CURLY LEAF during the Floral Transition Regulates Flowering. *Plant Physiology* **177**:pp.00296.2018.

Lucas, M., Davière, J.-M., Rodríguez-Falcón, M., Pontin, M., Iglesias-Pedraz, J. M., Lorrain, S., Fankhauser, C., Blázquez, M. A., Titarenko, E., Prat, S., et al. (2008). A molecular framework for light and gibberellin control of cell elongation. *Nature* **451**:480–484.

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**:1012–1025.

Lupiáñez, D. G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics* **32**:225–237.

Marcovitz, A., and Levy, Y. (2011). Frustration in protein-DNA binding influences conformational switching and target search kinetics. *Proceedings of the National Academy of Sciences* **108**:17957–17962.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., Radchuk, V., Dockter, C., Hedley, P. E., Russell, J., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**:427–433.

Mateos, J. L., Tilmes, V., Madrigal, P., Severing, E., Richter, R., Rijkenberg, C. W. M., Krajewski, P., and Coupland, G. (2017). Divergence of regulatory networks governed by the orthologous transcription factors FLC and PEP1 in Brassicaceae species. *Proceedings of the National Academy of Sciences* Advance Access published 2017, doi:10.1073/pnas.1618075114.

Mathelier, A., and Wasserman, W. W. (2013). The Next Generation of Transcription Factor Binding Site Prediction. *PLoS Computational Biology* **9**.

Mathelier, A., Xin, B., Chiu, T. P., Yang, L., Rohs, R., and Wasserman, W. W. (2016). DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Systems* **3**:278-286.e4.

Mathur, S., Vyas, S., Kapoor, S., and Tyagi, A. K. (2011). The Mediator Complex in Plants: Structure, Phylogeny, and Expression Profiling of Representative Genes in a Dicot (Arabidopsis) and a Monocot (Rice) during Reproduction and Abiotic Stress. *Plant Physiology* **157**:1609–1627.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006).

TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* **34**:D108–D110.

Mayran, A., and Drouin, J. (2018). Pioneer transcription factors shape the epigenetic landscape. *Journal of Biological Chemistry* Advance Access published 2018, doi:10.1074/jbc.R117.001232.

McGinty, R. K., and Tan, S. (2015). Nucleosome structure and function. *Chemical Reviews* **115**:2255–2273.

Meijsing, S. H., Pufall, M. A., So, A. Y., Bates, D. L., Chen, L., and Yamamoto, K. R. (2009). DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science* **324**:407–410.

Melzer, R., Verelst, W., and Theißen, G. (2009). The class E floral homeotic protein SEPALLATA3 is sufficient to loop DNA in 'floral quartet'-like complexes in vitro. *Nucleic Acids Research* **37**:144–157.

Mendes, M. A., Guerra, R. F., Berns, M. C., Manzo, C., Masiero, S., Finzi, L., Kater, M. M., and Colombo, L. (2013). MADS Domain Transcription Factors Mediate Short-Range DNA Looping That Is Essential for Target Gene Expression in Arabidopsis. *The Plant Cell* **25**:2560–2572.

Meyer, C. A., and Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics* **15**:709–721.

Minguet, E. G., Segard, S., Charavay, C., and Parcy, F. (2015). MORPHEUS, a webtool for transcription factor binding analysis using position weight matrices with dependency. *PLoS ONE* **10**:1–12.

Moraga, F., and Aquea, F. (2015). Composition of the SAGA complex in plants and its role in controlling gene expression in response to abiotic stresses. *Frontiers in Plant Science* **6**:1–9.

Moyroud, E., Minguet, E. G., Ott, F., Yant, L., Posé, D., Monniaux, M., Blanchet, S., Bastien, O., Thévenon, E., Weigel, D., et al. (2011). Prediction of Regulatory Interactions from Genome Sequences Using a Biophysical Model for the *Arabidopsis* LEAFY Transcription Factor. *The Plant Cell* **23**:1293–1306.

Muiño, J. M., Smaczniak, C., Angenent, G. C., Kaufmann, K., and Van Dijk, A. D. J. (2014). Structural determinants of DNA recognition by plant MADS-domain transcription factors. *Nucleic Acids Research* **42**:2138–2146.

Muino, J. M., De Bruijn, S., Pajoro, A., Geuten, K., Vingron, M., Angenent, G. C., and Kaufmann, K. (2016). Evolution of DNA-binding sites of a floral master regulatory transcription factor. *Molecular Biology and Evolution* **33**:185–200.

Müller, F., Zaucker, A., and Tora, L. (2010). Developmental regulation of transcription initiation: More than just changing the actors. *Current Opinion in Genetics and Development* **20**:533–540.

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., and Chang, H. Y. (2016). HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods* **13**:919–922.

Nieto, C., López-Salmerón, V., Davière, J. M., and Prat, S. (2015). ELF3-PIF4 interaction regulates plant growth independently of the evening complex. *Current Biology* **25**:187–193.

Nusinow, D. A., Helfer, A., Hamilton, E. E., King, J. J., Imaizumi, T., Schultz, T. F., Farré, E. M., and Kay, S. A. (2011). The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature* **475**:398–402.

O'Malley, R. C., Huang, S. shan C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A., and Ecker, J. R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **166**:1598.

Omidi, S., Zavolan, M., Pachkov, M., Breda, J., Berger, S., and van Nimwegen, E. (2017). Automated incorporation of pairwise dependency in transcription factor binding site prediction using dinucleotide weight tensors. *PLoS Computational Biology* **13**:1–22.

Pachkov, M., Balwierz, P. J., Arnold, P., Ozonov, E., and Van Nimwegen, E. (2013). SwissRegulon, a database of genome-wide annotations of regulatory sites: Recent updates. *Nucleic Acids Research* **41**:214–220.

Paillard, G., and Lavery, R. (2004). Analyzing Protein-DNA Recognition Mechanisms. *Structure* **12**:113–122.

Pajoro, A., Madrigal, P., Muiño, J. M., Matus, J. T., Jin, J., Mecchia, M. A., Debernardi, J. M., Palatnik, J. F., Balazadeh, S., Arif, M., et al. (2014). Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biology* **15**:R41.

Park, P. J. (2009). ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10**:669–680.

Patel, A., Yang, P., Tinkham, M., Pradhan, M., Sun, M. A., Wang, Y., Hoang, D., Wolf, G., Horton, J. R., Zhang, X., et al. (2018). DNA Conformation Induces Adaptable Binding by Tandem Zinc Finger Proteins. *Cell* **173**:221-233.e12.

Phillips, J. E., and Corces, V. G. (2009). CTCF: Master Weaver of the Genome. *Cell* **137**:1194–1211.

Pruneda-Paz, J. L., Breton, G., Nagel, D. H., Kang, S. E., Bonaldi, K., Doherty, C. J., Ravelo, S., Galli, M., Ecker, J. R., and Kay, S. A. (2014). A Genome-Scale Resource for the Functional Characterization of Arabidopsis Transcription Factors. *Cell Reports* **8**:622–632.

Qin, Q., and Feng, J. (2017). Imputation for transcription factor binding predictions based on deep learning. *PLoS Computational Biology* Advance Access published 2017.

Rada-Iglesias, A., Grosveld, F. G., and Papantonis, A. (2018). Forces driving the three-dimensional folding of eukaryotic genomes. *Molecular Systems Biology* **14**:e8214.

Rao, S. S. P., Huang, S. C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K. R., Sanborn, A. L., Johnstone, S. E., Bascom, G. D., Bochkov, I. D., et al. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell* **171**:305-320.e24.

Rao, S., Chiu, T. P., Kribelbauer, J. F., Mann, R. S., Bussemaker, H. J., and Rohs, R. (2018). Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein-DNA binding. *Epigenetics and Chromatin* **11**:1–11.

Rastogi, C., Rube, H. T., Kribelbauer, J. F., Crocker, J., Loker, R. E., Martini, G. D., Laptenko, O., Freed-Pastor, W. A., Prives, C., Stern, D. L., et al. (2018). Accurate and sensitive quantification of protein-DNA binding affinity. *Proceedings of the National Academy of Sciences* Advance Access published 2018, doi:10.1073/pnas.1714376115.

Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., Larson, D. R., and Zhao, K. (2017). CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Molecular Cell* **67**:1049-1058.e6.

Rhee, H. S., and Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**:1408–1419.

Riley, T. R., Lazarovici, A., Mann, R. S., and Bussemaker, H. J. (2015). Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using featureREDUCE. *eLife* **4**:1–14.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**:651–657.

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. *Nature* **461**:1248–1253.

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., and Mann, R. S. (2010). Origins of Specificity in Protein-DNA Recognition. *Annual Review of Biochemistry* **79**:233–269.

Ruan, S., and Stormo, G. D. (2017). Inherent limitations of probabilistic models for protein-DNA binding specificity. *PLOS Computational Biology* **13**:e1005638.

Ruan, S., Swamidass, S. J., and Stormo, G. D. (2017). BEESEM: Estimation of binding energy models using HT-SELEX data. *Bioinformatics* **33**:2288–2295.

Ruelens, P., Zhang, Z., van Mourik, H., Maere, S., Kaufmann, K., and Geuten, K. (2017). The Origin of Floral Organ Identity Quartets. *The Plant Cell* **29**:229–242.

Santolini, M., Mora, T., and Hakim, V. (2014). A general pairwise interaction model provides an accurate description of in Vivo transcription factor binding sites. *PLoS ONE* **9**.

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature* **489**:109–113.

Sayou, C., Monniaux, M., Nanao, M. H., Moyroud, E., Brockington, S. F., Thévenon, E., Chahtane, H., Warthmann, N., Melkonian, M., Zhang, Y., et al. (2014). A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science* **343**:645–648.

Sayou, C., Nanao, M. H., Jamin, M., Pose, D., Thevenon, E., Gregoire, L., Tichtinsky, G., Denay, G., Ott, F., Llobet, M. P., et al. (2016). A SAM oligomerization domain shapes the genomic binding landscape of the LEAFY transcription factor. *Nature Communications* **48**:829–834.

Schmidl, C., Rendeiro, A. F., Sheffield, N. C., and Bock, C. (2015). ChIPmentation: Fast, robust, low-input ChIP-seq for histones and transcription factors. *Nature Methods* **12**:963–965.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., et al. (2010). Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science* **328**:6.

Schneider, T. D., and Stephens, R. M. (1990). Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research* **18**:6097–6100.

Setty, M., and Leslie, C. S. (2015). SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLoS Computational Biology* **11**:1–22.

Shaffer, P. L., Jivan, A., Dollins, D. E., Claessens, F., and Gewirth, D. T. (2004). Structural basis of androgen receptor binding to selective androgen response elements. *Proceedings of the National Academy of Sciences of the United States of America* **101**:4758–4763.

Sharon, E., Lubliner, S., and Segal, E. (2008). A feature-based approach to modeling protein-DNA interactions. *PLoS Computational Biology* **4**.

Siddharthan, R. (2010). Dinucleotide weight matrices for predicting transcription factor binding sites: Generalizing the position weight matrix. *PLoS ONE* **5**.

Siggers, T., Duyzend, M. H., Reddy, J., Khan, S., and Bulyk, M. L. (2011). Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Molecular Systems Biology* **7**:1–14.

Sijacic, P., Bajic, M., McKinney, E. C., Meagher, R. B., and Deal, R. B. (2018). Changes in chromatin accessibility between Arabidopsis stem cells and mesophyll cells illuminate cell type-specific transcription factor networks. *Plant Journal* **94**:215–231.

Skene, P. J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6**:1–35.

Skene, J. P., and Henikoff S (2018). CUT& RUN: Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nature Protocols* **13**:1–28.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H. J., et al. (2011). Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. *Cell* **147**:1270–1282.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: How transcription factors read the genome. *Trends in Biochemical Sciences* **39**:381–399.

Smaczniak, C., Immink, R. G. H., Muiño, J. M., Blanvillain, R., Busscher, M., Busscher-Lange, J., Dinh, Q. D. P., Liu, S., Westphal, A. H., Boeren, S., et al. (2012). Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. *Proceedings of the National Academy of Sciences* **109**:1560–1565.

Smaczniak, C., Muiño, J. M., Chen, D., Angenent, G. C., and Kaufmann, K. (2017). Differences in DNA-binding specificity of floral homeotic protein complexes predict organ-specific target genes. *The Plant Cell* Advance Access published 2017, doi:10.1105/tpc.17.00145.

Soutourina, J. (2017). Transcription regulation by the Mediator complex. *Nature Reviews Molecular Cell Biology* Advance Access published 2017, doi:10.1038/nrm.2017.115.

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

Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., Gomez, A., Collombet, S., Berenguer, C., Cuartero, Y., et al. (2018). Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics* **50**:238–249.

Staller, M. V., Holehouse, A. S., Swain-Lenz, D., Das, R. K., Pappu, R. V., and Cohen, B. A. (2018). A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Systems* Advance Access published 2018, doi:10.1016/j.cels.2018.01.015.

Stella, S., Cascio, D., and Johnson, R. C. (2010). The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. *Genes and Development* **24**:814–826.

Stigliani, A., Martin-Arevalillo, R., Lucas, J., Bessy, A., Vinos-Poyo, T., Mironova, V., Vernoux, T., Dumas, R., and Parcy, F. (2018). Capturing auxin response factors syntax using DNA binding models Advance Access published 2018.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics* **16**:16–23.

Stormo, G. D. (2013). Modeling the specificity of protein-DNA interactions. *Quantitative Biology* **1**:115–130.

Stormo, G. D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. *Nature Reviews Genetics* **11**:751–760.

Stormo, G. D., Schneider, T. D., and Gold, L. M. (1982). Characterization of translational initiation sites in E. coli. *Nucleic acids research* **10**:2971–2996.

Stroud, H., Otero, S., Desvoyes, B. B., Ramírez-Parra, E., Jacobsen, S. E., Gutierrez, C., Ramirez-Parra, E., Jacobsen, S. E., and Gutierrez, C. (2012). Genome-wide analysis of histone H3. 1 and H3. 3 variants in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* **109**:5370–5.

Sun, B., Looi, L.-S., Guo, S., He, Z., Gan, E.-S., Huang, J., Xu, Y., Wee, W.-Y., and Ito, T. (2014). Timing Mechanism Dependent on Cell Division Is Invoked by Polycomb Eviction in Plant Stem Cells. *Science* **343**:1248559–1248559.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., et al. (2008). The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Research* **36**:1009–1014.

Takeda, T., Corona, R. I., and Guo, J. T. (2013). A knowledge-based orientation potential for transcription factor-DNA docking. *Bioinformatics* **29**:322–330.

Tao, Z., Shen, L., Gu, X., Wang, Y., Yu, H., and He, Y. (2017). Embryonic epigenetic reprogramming by a pioneer transcription factor in plants. *Nature* **551**:124–128.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. *Nature* **489**:75–82.

Tomovic, A., and Oakeley, E. J. (2007). Position dependencies in transcription factor binding sites. *Bioinformatics* **23**:933–941.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology* **23**:137–144.

Trigg, S. A., Garza, R. M., MacWilliams, A., Nery, J. R., Bartlett, A., Castanon, R., Goubil, A., Feeney, J., O'Malley, R., Huang, S. S. C., et al. (2017). CrY2H-seq: A massively multiplexed assay for deep-coverage interactome mapping. *Nature Methods* **14**:819–825.

Turner, D., Kim, R. G., and Guo, J. tao (2012). TFinDit: transcription factor-DNA interaction data depository. *BMC Bioinformatics* **13**.

Vachon, G., Engelhorn, J., and Carles, C. C. (2018). Interactions between transcription factors and chromatin regulators in the control of flower development. *Journal of Experimental Botany* Advance Access published 2018, doi:10.1093/jxb/ery079.

Vera, D. L., Madzima, T. F., Labonne, J. D., Alam, M. P., Hoffman, G. G., Girimurugan, S. B., Zhang, J., McGinnis, K. M., Dennis, J. H., and Bass, H. W. (2014). Differential Nuclease Sensitivity Profiling of Chromatin Reveals Biochemical Footprints Coupled to Gene Expression and Functional DNA Elements in Maize. *The Plant Cell* **26**:3883–3893.

Viner, C., Johnson, J., Walker, N., Shi, H., Sjöberg, M., Adams, D. J., Ferguson-Smith, A. C., Bailey, T. L., and Hoffman, M. M. (2016). Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet Advance Access published March 15, 2016, doi:10.1101/043794.

Voss, T. C., and Hager, G. L. (2013). Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics* **15**:69–81.

Wang, D., Mills, E. S., and Deal, R. B. (2012). Technologies for systems-level analysis of specific cell types in plants. *Plant Science* **197**:21–29.

Wang, H., Liu, C., Cheng, J., Liu, J., Zhang, L., He, C., Shen, W. H., Jin, H., Xu, L., and Zhang, Y. (2016). Arabidopsis Flower and Embryo Developmental Genes are Repressed in Seedlings by Different Combinations of Polycomb Group Proteins in Association with Distinct Sets of Cis-regulatory Elements. *PLoS Genetics* **12**:1–25.

Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., Ye, Z., Shen, C., Li, J., Zhang, L., et al. (2017). Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nature Genetics* **49**:579–587.

Wang, M., Tai, C., E, W., and Wei, L. (2018a). DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Research* Advance Access published 2018, doi:10.1093/nar/gky215.

Wang, Z., Civelek, M., Miller, C., Sheffield, N., Guertin, M. J., and Zang, C. (2018b). BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics* Advance Access published 2018, doi:10.1101/280982.

Wang, M., Wang, P., Lin, M., Ye, Z., Li, G., Tu, L., Shen, C., Li, J., Yang, Q., and Zhang, X. (2018c). Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nature Plants* **4**:90–97.

Wasserman, W. W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics* **5**:276–287.

Weber, B., Zicola, J., Oka, R., and Stam, M. (2016). Plant Enhancers: A Call for Discovery. *Trends in Plant Science* **21**:974–987.

Weintraub, A. S., Li, C. H., Zamudio, A. V, Sigova, A. A., Hannet, N. M., Day, D. S., Abraham, B. J., Cohen, M. A., Nabet, B., Buckley, D. L., et al. (2018). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* Advance Access published 2018, doi:10.1016/j.cell.2017.11.008.

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., et al. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**:1431–1443.

White, M. A., Myers, C. A., Corbo, J. C., and Cohen, B. A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences* **110**:11952–11957.

Whiteld, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., Myers, R. M., and Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biology* **13**:R50.

Wu, H., and Zhang, Y. (2014). Reversing DNA methylation: Mechanisms, genomics, and biological functions. *Cell* **156**:45–68.

Wu, M., Sang, Y., Bezhani, S., Yamaguchi, N., Han, S., Li, Z., Su, Y., Slewinski, T. L., and Wagner, D. (2012). SWI2/SNF2 chromatin remodeling ATPases overcome polycomb repression and control floral organ identity with the LEAFY and SEPALLATA3 transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **109**:3576–3581.

Xiao, J., and Wagner, D. (2015). Polycomb repression in the regulation of growth and development in Arabidopsis. *Current Opinion in Plant Biology* **23**:15–24.

Xiao, J., Jin, R., Yu, X., Shen, M., Wagner, J. D., Pai, A., Song, C., Zhuang, M., Klasfeld, S., He, C., et al. (2017). Cis and trans determinants of epigenetic silencing by Polycomb repressive complex 2 in Arabidopsis. *Nature Genetics* Advance Access published 2017, doi:10.1038/ng.3937.

Xu, B., Schones, D. E., Wang, Y., Liang, H., and Li, G. (2013). A Structural-Based Strategy for Recognition of Transcription Factor Binding Sites. *PLoS ONE* **8**.

Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M., et al. (2013). Transcription factor binding in

human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**:801–813.

Yan, W., Chen, D., and Kaufmann, K. (2016). Efficient multiplex mutagenesis by RNA-guided Cas9 and its use in the characterization of regulatory elements in the AGAMOUS gene. *Plant Methods* **12**:1–9.

Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W. W., Gordân, R., and Rohs, R. (2014). TFBSshape: A motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Research* **42**:148–155.

Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R., and Rohs, R. (2017). Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Molecular Systems Biology* **13**:910.

Yazaki, J., Galli, M., Kim, A. Y., Nito, K., Aleman, F., Chang, K. N., Carvunis, A.-R., Quan, R., Nguyen, H., Song, L., et al. (2016). Mapping transcription factor interactome networks using HaloTag protein arrays. *Proceedings of the National Academy of Sciences* **113**:E4238–E4247.

Ye, L., Wang, B., Zhang, W.-G., Shan, H., and Kong, H. (2016). Gain of An Auto-regulatory Site Led to Divergence of the Arabidopsis APETALA1 and CAULIFLOWER Duplicate Genes in the Time, Space and Level of Expression and Regulation of One Paralog by the Other. *Plant Physiology* **171**:pp.00320.2016.

Yin, Y., Sieradzan, A. K., Liwo, A., He, Y., and Scheraga, H. A. (2015). Physics-based potentials for coarse-grained modeling of protein-DNA interactions. *Journal of Chemical Theory and Computation* **11**:1792–1808.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**:eaaj2239.

Zaret, K. S. (2018). Pioneering the chromatin landscape. *Nature Genetics* Advance Access published 2018, doi:10.1038/s41588-017-0038-z.

Zaret, K. S., and Mango, S. E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. *Current Opinion in Genetics & Development* **37**:76–81.

Zemach, A., and Grafi, G. (2007). Methyl-CpG-binding domain proteins in plants: interpreters of DNA methylation. *Trends in Plant Science* **12**:80–85.

Zentner, G. E., Kasinathan, S., Xin, B., Rohs, R., and Henikoff, S. (2015). ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. *Nature Communications* **6**:8733.

Zhang, C., Liu, S., Zhu, Q., and Zhou, Y. (2005). A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *Journal of Medicinal Chemistry* **48**:2325–2335.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W. L., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E., et al. (2006). Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell* **126**:1189–1201.

Zhang, W., Zhang, T., Wu, Y., and Jiang, J. (2012). Genome-Wide Identification of Regulatory DNA Elements and Protein-Binding Footprints Using Signatures of Open Chromatin in Arabidopsis. *The Plant Cell* **24**:2719–2731.

Zhang, L., Martini, G. D., Tomas Rube, H., Kribelbauer, J. F., Rastogi, C., FitzPatrick, V. D., Houtman, J. C., Bussemaker, H. J., and Pufall, M. A. (2018a). SelexGLM differentiates androgen and glucocorticoid receptor DNA-binding preference over an extended binding site. *Genome Research* **28**:111–121.

Zhang, H., Lang, Z., and Zhu, J. (2018b). Dynamics and function of DNA methylation in plants. *Nature Reviews Molecular Cell Biology* Advance Access published 2018, doi:10.1038/s41580-018-0016-z.

Zhao, Y., Granas, D., and Stormo, G. D. (2009). Inferring binding energies from selected binding sites. *PLoS Computational Biology* **5**.

Zhao, Y., Ruan, S., Pandey, M., and Stormo, G. D. (2012). Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* **191**:781–790.

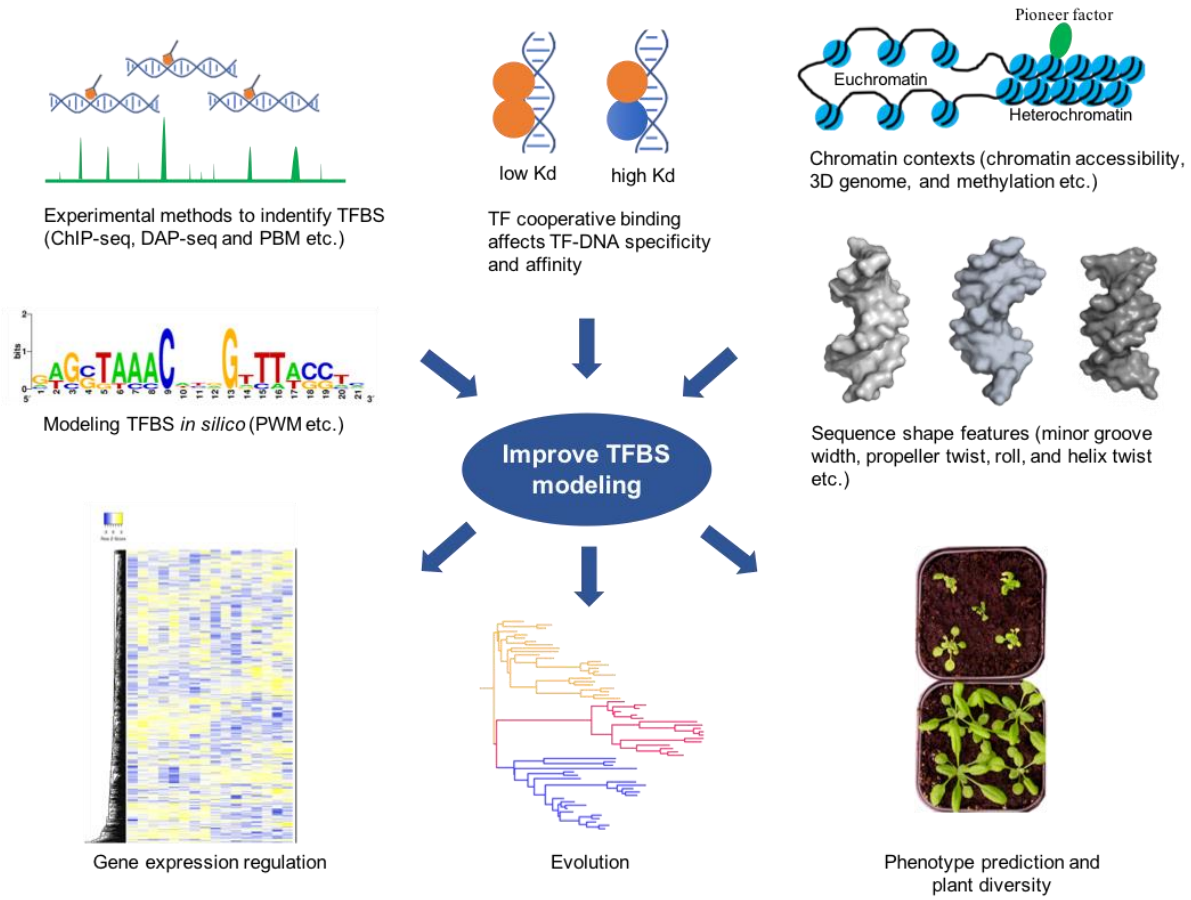
Zhou, J., and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* **12**:931–934.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., Bussemaker, H. J., Gordân, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. *Proceedings of the National Academy of Sciences* **112**:4654–4659.

Zhou, Y., Wang, Y., Krause, K., Yang, T., Dongus, J. A., Zhang, Y., and Turck, F. (2018). Telobox motifs recruit CLF / SWN – PRC2 for H3K27me3 deposition via TRB factors in Arabidopsis. *Nature Genetics* Advance Access published 2018, doi:10.1038/s41588-018-0109-9.

Zhu, H., Wang, G., and Qian, J. (2016). Transcription factors as readers and effectors of DNA methylation. *Nature Reviews Genetics* **17**:551–565.

Zuo, Z., Roy, B., Chang, Y. K., Granas, D., and Stormo, G. D. (2017). Measuring quantitative effects of methylation on transcription factor–DNA binding affinity. *Science Advances* **3**:eaao1799.



1486

1487 **Figure 1. Schematic overview of the TFBS modeling and application.**

1488

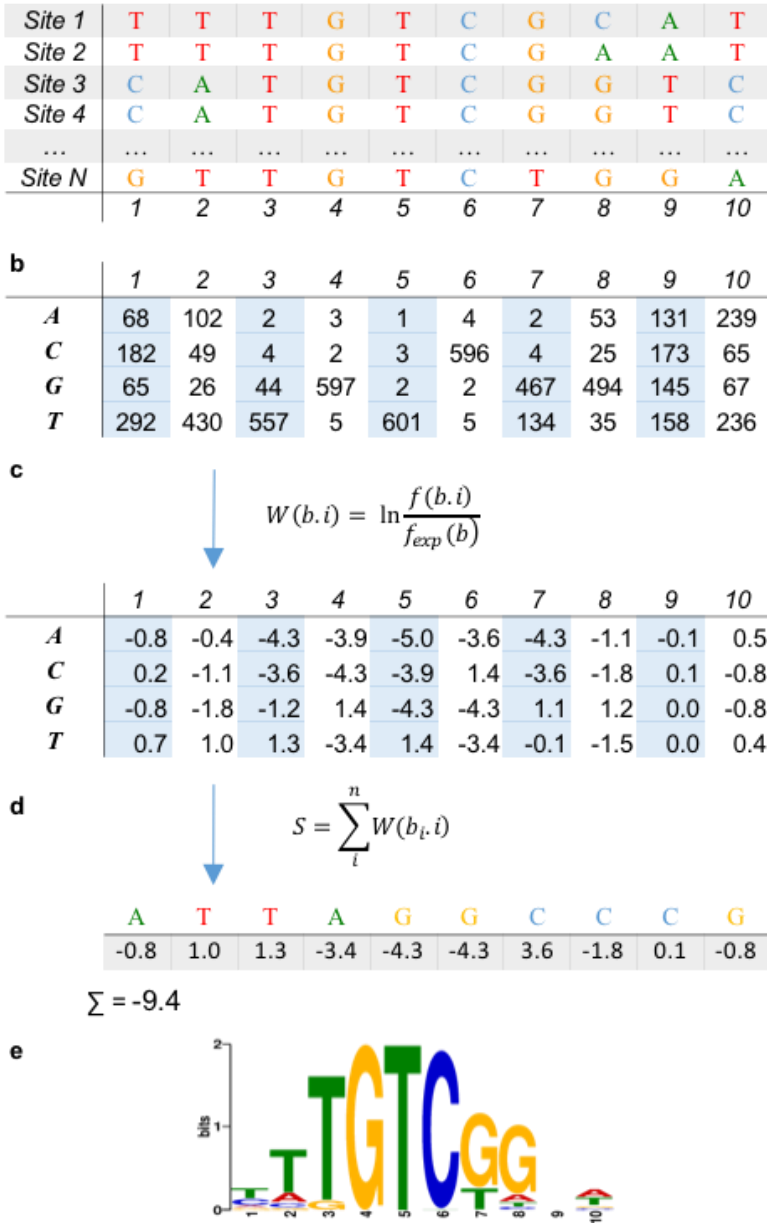


Figure 2. Workflow to generate PWM for a set of known TFBS. Here, ARF5 bound sequences are retrieved from DAP-seq (O'Malley et al., 2016). The sequences of N different binding sites are aligned **(a)**. The nucleotide frequency is computed at each position of the binding site to yield the position frequency matrix **(b)**, which is then converted to a PWM **(c)**. In the formula, *W(b, i)* stands for the weight of a nucleotide *b* at position *i*, *f(b,i)* is the frequency of this nucleotide, and f_{exp} is the expected background frequency of the given nucleotide. If each nucleotide appearance is equal, one can take $f_{exp}=N/4$ **(c)**. One can score a given sequence by summing the corresponding PWM weights **(d)**. The TFBS logo represents the preference of the TF at each position of the binding site **(e).** This calculation represents one possible method among several to calculate a sequence score (Stormo, 2000; Wasserman and Sandelin, 2004; Stormo, 2013).

Table1. Experimental methods to identify TFBS.

Experimental methods	Description	*in vivo or in vitro*	DNA ligand	Unique features	TF source	References
PBM (Protein binding microarrays) and variants	PBM uses microarrays of randomized DNA to which TF binding can be assayed by fluorescent antibodies to the TF.	*in vitro*	synthetic and randomized	High-throughput	Recombinant, usually fused with tags, such as GST	(Berger et al., 2006; Berger and Bulyk, 2009)
ChIP-seq (Chromatin immunoprecipitation followed by sequencing) and variants	ChIP-seq couples chromatin immunoprecipitation with massively parallel sequencing, is capable of mapping genome-wide TFBSs in vivo	*in vivo*	genomic	High-throughput and most widely used protocol for TFBSs mapping *in vivo*	native protein or fused with an epitope tag	(Johnson et al., 2007; Robertson et al., 2007; Kaufmann et al., 2010; Rhee and Pugh, 2011)
SELEX-seq (Systematic evolution of ligands by exponential enrichment followed by sequencing) and variants.	SELEX-seq uses recombinant TF to IP randomized DNA sequence in one or more cycles. The enriched DNA are sequenced by NGS and used to infer a model of specificity, typically a PWM, for a TF.	*in vitro*	synthetic and randomized	High-throughput, widely used, and allows to detect effects of DNA methylation (methyl-SELEX) and TF cooperative binding (CAP-SELEX).	Recombinant or *in vitro* translated TF	(Jolma et al., 2010; Jolma et al., 2015; Yin et al., 2017)
ORGANIC (Occupied regions of genomes from affinity-purified naturally isolated chromatin)	ORGANIC applies MNase to digest non-cross-linked chromatin then perform affinity purification by TF followed by NGS sequencing.	*in vivo*	genomic	Avoiding sonication and cross-linking	Endogenous TF or fused with an epitope tag	(Kasinathan et al., 2014)
BunDLE-seq (Binding to Designed Library, Extracting, and sequencing)	BunDLE-seq provides quantitative measurements of TF binding to thousands of fully designed sequences of 200 bp in length within a single experiment.	*in vitro*	synthetic and randomized	Allows comprehensive characterization of TF binding determinants within and outside of core binding sites	Recombinant TF	(Levo et al., 2015)

ChEC-seq (Chromatin endogenous cleavage followed by sequencing)	ChEC-seq uses fusion of a TF to MNase to target calcium-dependent cleavage to specific genomic loci *in vivo*.	*in vivo*	genomic	Rapidly inducible nature of ChEC-seq allows separation of TFBSs based on their recognition by DNA sequence and shape or shape alone.	Fused with MNase, produced *in vivo* under native or high inducible promotor	(Zentner et al., 2015)
ChIPmentation	ChIPmentation introduces sequencing-compatible adaptors in a single-step reaction directly on bead-bound chromatin, which reduces time, cost and input requirements, thus providing a convenient and broadly useful alternative to existing ChIP-seq protocols.	*in vivo*	genomic	Avoids sequencing adaptor dimers which are common in standard ChIP-seq protocol, and requires only a single DNA purification step before library amplification.	Endogenous TF	(Schmidl et al., 2015)
(amp)DAP-seq (DNA affinity purification followed by sequencing)	DAP-seq uses recombinant TF to affinity-purify genomic DNA fragments followed by NGS sequencing, capable of derive cistrome; ampDAP-seq applies PCR amplification to remove methylation patterns of fragmented genomic DNA before affinity purification, capable of deriving epi-cistrome.	*in vitro*	genomic	Allows low-cost and high-throughput generation of cistrome and epicistrome maps for hundreds of TFs of an organism	Recombinant or *in vitro* translated TF	(O'Malley et al., 2016; Bartlett et al., 2017)
SMiLE-seq (selective microfluidics-based ligand enrichment followed by sequencing)	SMiLE-seq applies microfluidics-based technology to perform a rigorous on-chip isolation of interacting TF-DNA complexes, allows robust identification of DNA-binding specificities of TF monomers, homodimers and heterodimers.	*in vitro*	synthetic and randomized	Distinguish TF binding specificity from TF monomers and dimers of a TF (or hetero-/oligo-dimers of TFs) by microfluidics.	Recombinant or *in vitro* translated TF	(Isakova et al., 2017)
SLIM-ChIP (short-fragment-enriched, low-input, indexed MNase ChIP)	SLIM-ChIP combines enzymatic fragmentation of chromatin and on-bead indexing to map high-resolution binding landscape of a TF.	*in vivo*	genomic	Low material input and allows mapping DNA binding proteins and charting the surrounding chromatin	Endogenous TF	(Gutin et al., 2018)

				occupancy landscape at a single-cell level		
CUT&RUN (Cleavage under targets and release using nuclease)	CUT&RUN is an epigenomic profiling strategy in which antibody-targeted controlled cleavage by MNase releases specific protein–DNA complexes into the supernatant for NGS sequencing.	*in vivo*	genomic	Avoids crosslinking and solubilization issues, and requires less sequencing depth.	Endogenous TF	(Skene and Henikoff, 2017; Skene and Henikoff S, 2018)
Methyl-Spec-seq	Methyl-Spec-seq measures the effects of CpG methylation (mCPG) on TF binding affinity, allowing quantitative assessment of the effects at every position in a binding site.	*In vitro*	synthetic and randomized	Facilitates the quantitative modeling of mCpG effects on gene regulation.	Recombinant TF	(Zuo et al., 2017)
ChIP-STARR-seq	ChIP-STARR-seq combines ChIP with a massively parallel reporter assay to identify functional enhancers genome-wide in a quantitative manner. This method is potentially applicable in plant system.	*In vivo*	genomic	ChIP-STARR-seq allows high-throughput identification of functional enhancer *in vivo*.	Endogenous TF	(Barakat et al., 2018)

Table 2. TFBS modeling methods.

TFBS modelling methods	Description	Features integrated	Web server or source code	Motif representation	References
PWM (position weight matrix)	PWMs are normalized representations of the position-specific log-likelihoods of a nucleotide's probability to occur at each position in a sequence	N.A (Not applicable)	N.A	PWM logo	(Stormo et al., 1982; Schneider and Stephens, 1990)
DWM (dinucleotide weight matrix)	DWM considers the 16 combinations of dinucleotide instead of the 4 nucleotides used for PWM.	Dinucleotides	N.A	DWM logo	(Siddharthan, 2010)
BEM (binding energy model)	BEM introduces energy parameters of adjacent nucleotides to the binding affinity quantification.	Dependencies (adjacent positions) and binding affinity data	http://stormo.wustl.edu/TF-BEMs/	Binding energy logo	(Zhao et al., 2012)
TFFM (TF Flexible Model)	TFFMs model integrates a markov model to take dependencies and background into account.	Dependencies (adjacent position) and background	http://cisreg.cmmt.ubc.ca/cgi-bin/TFFM/TFFM_webapp.py?rm=start	TFFM logo	(Mathelier and Wasserman, 2013)
PIM (pairwise interaction model)	PIM is based on the principle of maximum entropy and describes pairwise correlations between nucleotides at different positions.	Dependencies between all positions	https://github.com/msantolini/PIM	PWM mixture model	(Santolini et al., 2014)
gkm-SVM (gapped k-mer support vector machine)	gkm-SVM predicts regulatory sequence using gapped k-mer features.	k-mers supporting gaps	http://www.beerlab.org/gkmsvm/	N.A	(Ghandi et al., 2014)
SeqGL	SeqGL is a *de novo* motif discovery algorithm to identify multiple TF sequence signals from ChIP-seq, DNase-seq, and ATAC-seq profiles.	K-mer, chromatin accessibility	http://cbio.mskcc.org/public/Leslie/SeqGL/	N.A	(Setty and Leslie, 2015)
MORPHEUS	MORPHEUS is a webtool for TF binding analysis using PWM	Dependencies between all	http://biodev.cea.fr/morpheus/	PWM logo with	(Minguet et al., 2015)

	with dependencies.	positions		dependancies	
FeatureREDUCE	FeatureREDUCE provides a flexible framework for building sequence-to-affinity models from PBM data.	Dependencies between all positions	http://software.bussemakerlab.org	N.A	(Riley et al., 2015)
Cytomod	Cytomod models methyl-sensitive TF motifs with an expanded epigenetic alphabet.	DNA methylation	N.A	PWM logo with an extended alphabet (e.g. 5mC stands for 5-Methylcytosine)	(Viner et al., 2016)
DeepBind	DeepBind can learn several motifs to predict binding sites of DNA and RNA binding proteins.	N.A	http://tools.genes.toronto.edu/deepbind/	Weighted ensemble of PWM logos	(Alipanahi et al., 2015)
DeepSEA (deep learning-based sequence analyzer)	DeepSEA predicts effects of noncoding variants with deep learning–based sequence model	Integrate DNase I hypersensitivity data and histone-mark profiles	http://deepsea.princeton.edu/job/analysis/create/	N.A	(Zhou and Troyanskaya, 2015)
DWT (dinucleotide weight tensor)	DWT is a regulatory motif model that incorporates arbitrary pairwise dependencies for TFBS prediction.	Dependencies between all positions	http://dwt.unibas.ch/fcgi/dwt	DWT 'dilogo' motifs	(Omidi et al., 2017)
TFImpute	TFImpute predict cell-specific TF binding trained by deep learning.	N.A	https://bitbucket.org/feeldead/tfimpute	N.A	(Qin and Feng, 2017)
BEESEM (short for Binding Energy Estimation on SELEX with Expectation Maximization)	BEESEM estimates binding energy models using SELEX-seq data based on expectation maximization method.	N.A	http://stormo.wustl.edu/beeml/	N.A	(Ruan et al., 2017)
DeFine	DeFine quantifies TF-DNA binding affinity and facilitate evaluation of functional non-coding variants in the genome	Integrate Hi-C data	http://define.cbi.pku.edu.cn	PWM logo	(Wang et al., 2018a)

	based on deep learning algorithms.				
DFIM (Deep Feature Interaction Maps)	DFIM estimates pairwise interactions between features (such as nucleotides or subsequences) in any input DNA sequences by a neural network.	Dependencies between all positions, interaction between motifs, core motif flanking region and chromatin accessibility	https://github.com/kundajelab/dm.	DFIM with feature importance scores	(Greenside et al., 2018)
NRLB (No Read Left Behind)	NRLB provides scalable and quantitative method to identify functional in vivo binding sites of TF and to define relative binding affinities for any TF-DNA complex.	N.A	N.A	Energy logo representation	(Rastogi et al., 2018)
KSM model (k-mer set memory)	A k-mer finder (KMAC) finds k-mers that are over-represented in TFBSs, and KSM allow accurates regulatory variant prediction.	k-mers	https://github.com/gifford-lab/GEM3)	KSM (k-mer set memory)	(Guo et al., 2018)
SelexGLM	SelexGLM incorporates core motif flanking region for TFBS binding quantification.	Core motif flanking region	https://www.bioconductor.org	Energy logo representation	(Zhang et al., 2018a)