



HAL
open science

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

► **To cite this version:**

Xuelei Lai, Arnaud Stigliani, Gilles Vachon, Cristel Carles, Cezary Smaczniak, et al.. Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants. *Molecular Plant*, 2019, 12 (6), pp.743-763. 10.1016/j.molp.2018.10.010 . hal-02088235

HAL Id: hal-02088235

<https://hal.science/hal-02088235>

Submitted on 14 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Building transcription factor binding site models to** 2 **understand gene regulation in plants**

3 Xuelei Lai^{1,*}, Arnaud Stigliani¹, Gilles Vachon¹, Cristel Carles¹, Cezary Smaczniak²,
4 Chloe Zubieta¹, Kerstin Kaufmann² and François Parcy^{1,*}

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

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

8 * Corresponding author: François Parcy, Tel: +33 0438784978, Fax: +33
9 0438784091, Email: francois.parcy@cea.fr; Co-corresponding author: Xuelei Lai,
10 Email: xuelei.lai@cea.fr.

11 **Abstract**

12 Transcription factors (TF) are key cellular components that control gene expression.
13 They recognize specific DNA sequences, the TF binding sites (TFBS), and thus are
14 targeted to specific regions of the genome where they can recruit transcriptional
15 cofactors and/or chromatin regulators for fine-tuning spatiotemporal gene regulation.
16 Therefore, the identification of TFBS in genomic sequences and their subsequent
17 quantitative modeling is of crucial importance for understanding and predicting gene
18 expression. Here, we review how TFBS can be determined experimentally, how the
19 TFBS models can be constructed *in silico*, and how they can be optimized by taking
20 into account features such as position interdependence within TFBSs, DNA shape
21 and/or by introducing state-of-the-art computational algorithms such as deep learning
22 methods. In addition, we discuss the integration of context variables into the TFBS
23 modeling, including nucleosome positioning, chromatin states, methylation patterns,
24 3D genome architectures and TF cooperative binding, in order to better predict TF
25 binding under cellular contexts. Finally, we explore the possibilities of combining the
26 optimized TFBS model with technological advances such as targeted TFBS
27 perturbation by CRISPR to better understand gene regulation, evolution and plant
28 diversity.

29 **Running title:** Modeling TF binding sites in plants

30 Introduction

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

48 TFs recognize TFBS in a sequence-specific manner as revealed by structural studies
49 of protein-DNA complexes (Paillard and Lavery, 2004; Rohs et al., 2010) and next
50 generation sequencing (NGS) techniques such as SELEX-seq and ChIP-seq (**Table**
51 **1**). In the last decade, these NGS techniques have revolutionized the exploration of
52 the TF binding landscape both *in vitro* and *in vivo* (Koboldt et al., 2013). This has
53 resulted in many databases for TFBS deposition and profiling, such as TRANSFAC
54 (Matys et al., 2006), JASPAR (Khan et al., 2018), UniPROBE (Hume et al., 2015),
55 HOCOMOCO (Kulakovskiy et al., 2013), CIS-BP (Weirauch et al., 2014) and
56 SwissRegulon (Pachkov et al., 2013). Such efforts have substantially boosted our
57 understanding of interactions between TFs and TFBSs in different species, tissues
58 and different developmental stages. While a great deal of progress has been made to
59 map TFBS, the resulting models are often poorly predictive of actual gene regulation.
60 This can be due to poor modeling and prediction of TFBS, non-productive TF binding
61 (i.e. that does not result in gene regulation) or a combination of the two. Here, we will

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

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

80 **TFBS modeling**

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

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

96 **Experimental methods to identify TFBS**

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

101 ChIP-seq has long been the gold standard for detecting genome-wide TFBSs bound
102 by a given TF *in vivo* (Johnson et al., 2007; Robertson et al., 2007; Kaufmann et al.,
103 2010). In a standard ChIP-seq protocol, sample tissues are treated with a
104 crosslinking reagent and subjected to nuclei purification to isolate chromatin
105 containing TF-DNA complexes. Generally, an additional step of chromatin shearing
106 by sonication is applied before the final step of chromatin immunoprecipitation (IP)
107 using a TF-specific antibody. The IP product containing enriched DNA fragments that
108 are recognized by the TF of interest is then subjected to NGS sequencing. ChIP-seq
109 has been successfully used routinely in many laboratories. However, standard ChIP-
110 seq protocols have intrinsic limitations and technical drawbacks (Park, 2009). One of
111 the limitations comes from sonication, a process that is highly irreproducible and
112 produces variable DNA fragment sizes that are difficult to sequence. The other
113 limitation is crosslinking, which produces low signal to noise ratio and many false
114 positives. To overcome such limitations, many ChIP-seq variant methods have been
115 developed, including ORGANIC (Kasinathan et al., 2014), ChEC-seq (Zentner et al.,
116 2015), CUT&RUN (Skene and Henikoff, 2017) and SLIM-seq (Gutin et al., 2018)
117 (refer to **Table 1** for unique features and details of these methods), all of which use
118 micrococcal nuclease (MNase) to fragment chromatin, therefore avoiding sonication.
119 Due to the mild conditions of DNA fragmentation by MNase, these methods do not
120 denature or disrupt TF-DNA complexes and eliminate the requirement of
121 crosslinking. These protocols also require substantially lower amounts of input
122 materials and are thus feasible for low-input applications. Processing of IP enriched
123 DNA fragments for downstream NGS application poses another challenge in ChIP-
124 seq and is often time-consuming. To simplify the process, ChIPmentation applies
125 Tn5 transposase directly to bead-bound chromatin, allowing single-step NGS
126 compatible DNA library preparation (Schmidl et al., 2015) (**Table 1**).

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

157 **Model TFBS *in silico***

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

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

167 **Position weight matrix**

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

177 **Dependencies**

178 PWMs provide good approximation of TF-DNA interactions in most cases, and can
179 be generated from various datasets, ranging from a small set of known TFBSs to TF-
180 DNA binding data derived from high-throughput assays. However, standard PWM
181 assumes that each position within a TFBS contributes to binding affinity independent
182 of other positions, and is thus unable to represent inter-base dependencies, which
183 have been observed for some TFs (Bulyk, 2002; Tomovic and Oakeley, 2007; Badis
184 et al., 2009). Various models that take into account these dependencies have been
185 shown to outperform standard PWM in *de novo* prediction. For example, the
186 MORPHEUS program allows to introduce di- and tri-nucleotide position
187 dependencies in PWM and has been successfully applied to plant TFs with, in some
188 cases, improved predictive power (Moyroud et al., 2011; Minguet et al., 2015) (**Table**
189 **2**).

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

192 between adjacent and/or distal positions, such as the binding energy model (BEM)
193 (Zhao et al., 2012), dinucleotide weight matrices (DWM) (Siddharthan, 2010) and TF
194 Flexible Model (TFFM) (Mathelier and Wasserman, 2013) (**Table 2**). Others
195 introduce higher-order k -mer features, that take into account all possible sequences
196 with length k , such as the feature motif model (FMM) (Sharon et al., 2008) (**Table 2**).
197 In some cases, model complexity can increase dramatically when arbitrary positions
198 or unconstrained k -mer features are used and become prone to be overfitting.
199 Alternative approaches start from a model without dependencies, and use a greedy
200 algorithm to improve the model by adding dependency features iteratively (Hu et al.,
201 2010; Santolini et al., 2014). Thus, dependency features are iteratively added until no
202 further feature could be found to improve the model. Others use Bayesian Markov
203 models (BaMM) of order k that take into account dependencies between one
204 nucleotide and the k previous positions (Kiesel et al., 2018). Complex models
205 integrating dependency features generally outperform simple PWM models, however,
206 some of these models require more expertise to apply and repeated manual attempts
207 to be trained correctly and are thus not facile to use. This constitutes one of the
208 limiting factors that restricts these models from being used routinely in the community.
209 In **Table 2** we summarize features of some of the most recent models.

210 **Shape features**

211 Sequence-based models provide accurate estimation of base readout, however, it
212 cannot explain why some TFs, which have highly conserved DNA-binding domains,
213 bind different sequences genome-wide. For example, the TF paralogs, androgen and
214 glucocorticoid receptors, which bind similar DNA motifs by a set of identical amino
215 acids (Shaffer et al., 2004; Meijsing et al., 2009), share only a third of their genomic
216 binding sites (Zhang et al., 2018a). It turns out that DNA shape features contribute
217 significantly to distinguish *bona fide* TFBSs from others. In the last decade, many
218 studies have revealed that indeed DNA shape features play an important role for
219 determining TF-DNA binding specificity (Rohs et al., 2009; Abe et al., 2015; Yang et
220 al., 2017). A most recent example is the MADS-box TF, SEPALLATA3 (SEP3), a key
221 regulator of flower organ specification (Muiño et al., 2014; Hugouvieux et al., 2018).
222 MADS-box TFs bind to CArG-boxes with consensus sequence of 5'-CC(A/T)₆GG-3',
223 yet only a fraction of the CArG-boxes available genome-wide is bound by SEP3.
224 Käppel and colleagues showed that SEP3-DNA binding affinity correlates well with

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

238 Introducing shape features into TFBSs modeling requires integrating several distinct
239 shape parameters, including Minor Groove Width, Propeller Twist, Roll and Helix
240 Twist. These features have been shown to be distinguished by different TFs (Yang et
241 al., 2014). Very recently, nine additional shape features were introduced to the
242 repertoire in order to better describe the unique 3D structure encoded in a given DNA
243 sequence (Li et al., 2017). Apart from 'naked' DNA shape features, DNA methylation
244 on cytosine residues also affects DNA structure, making it a unique type of shape
245 feature that could be recognized by many TFs (Lazarovici et al., 2013; Yin et al.,
246 2017; Rao et al., 2018). Several TFBS modeling methods that take into account
247 shape features (some combined with sequence-based features) have been
248 developed, and show improvement compared with only sequence feature-based
249 models (**Table 2**). However, these models use DNA shape information generated
250 from computational simulations, such as Monte Carlo or Molecular Dynamics, and
251 potential biases exist. Improvements have already been obtained by integrating DNA
252 shape information derived from experimental data, such as X-ray crystallography (Li
253 et al., 2017). Thus, one major challenge regarding incorporating shape features into
254 TFBS modeling is to derive unbiased DNA structural data in a high-throughput
255 manner that has been robustly verified experimentally, which currently is a challenge.
256 The other challenge is that both DNA conformation (Azad et al., 2018) and TF
257 conformation (Patel et al., 2018) could be changed in an adaptation mechanism upon

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

260 **Energy- and deep learning-based models**

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

268 Machine learning methods, such as deep learning, are able to leverage very large
269 datasets to discover intricate connections within them and make accurate predictions
270 (Lecun et al., 2015). In the last few years, deep learning has been increasingly
271 applied to resolve complex biological problems, including those from regulatory
272 genomics (Angermueller et al., 2016). Several methods based on deep learning have
273 been developed to model TF-DNA binding specificity or to predict TF *in vivo* binding,
274 including DeepBind (Alipanahi et al., 2015), DeepSEA (Zhou and Troyanskaya,
275 2015), TFImpute (Qin and Feng, 2017), DeFind (Wang et al., 2018a) and DFIM
276 (Greenside et al., 2018) (**Table 2**). Advantages of these models include for example,
277 1) they can be trained from various types of sequencing data in either alone or
278 integrated manner, and can be further combined with other information, such as
279 DNase I hypersensitivity data, for better *in vivo* TFBSs prediction (Zhou and
280 Troyanskaya, 2015); 2) they can tolerate a certain degree of noise stemming from
281 either data acquisition technology or sequencing biases; 3) they can train predictive
282 models fully automatically, alleviating the need for time-consuming manual
283 intervention and expertise; 4) they can accurately identify genomic variants in the
284 regulatory region, and indicate how variations affect TF binding within a specific
285 sequence. However, one of the yet to be tackled difficulties of deep learning models
286 is that they are more difficult to interpret than PWMs given the hidden layers in the
287 networks. More information of these models and their unique properties can be found
288 in **Table 2**. To conclude, it is worth mentioning that, until now, no single model has
289 been identified to be the best for all TFs and the nature of the most adequate model
290 depends on the individual TF.

291 **Link between models and TF 3D structure**

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

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

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

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

339 **Improve the predictive power of TFBS models-genome context**

340 Eukaryotic genomes contain numerous potential binding sites for a given TF,
341 however, only a small fraction is actually bound *in vivo*, and that these sites vary
342 substantially depending on contexts, such as cell types, developmental stages, and
343 environmental or cellular conditions. In addition, only a subset of the bound sites
344 drive transcription (Wasserman and Sandelin, 2004; Hu et al., 2007; Fisher et al.,
345 2012; Whiteld et al., 2012). Therefore, various contexts have to be taken into account
346 to predict functional TFBSs precisely. This includes chromatin states (such as
347 accessibility and epigenetic marks), methylation states, nucleosome positioning and
348 genome 3D structures, and combinatorial binding of TFs.

349 **Nucleosome positioning, chromatin states and 3D genome**

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

356 Zhu et al., 2016) among others. Thus, chromatin structure is intrinsically repressive, a
357 mechanism that helps to establish stable gene expression and prevents unwanted
358 cell fate transitions. For gene activation, eukaryotic cells evolved various counter
359 mechanisms for each of the chromatin compacting factors to create accessible
360 chromatin, such as active histone marks (e.g. H3K4me2/3 and H3K27ac), chromatin
361 remodelers (Ho and Crabtree, 2010) and demethylation machineries (Wu and Zhang,
362 2014). The interplay between all these factors result in a highly dynamic chromatin
363 environment, in which TFs have to find their cognate DNA binding sites.

364 **Nucleosome positioning**

365 In general, TFs preferentially bind to TFBSs in accessible chromatin regions, where
366 nucleosomes are depleted (NDR-nucleosome depleted region). This is evidenced by
367 large scale *cis*-element studies, which showed that the vast majority of the active *cis*-
368 elements reside in the NDR in different species (Thurman et al., 2012; Weber et al.,
369 2016), including Arabidopsis and maize (Zhang et al., 2012; Vera et al., 2014).
370 Therefore, a precise *in vivo* TFBSs prediction model could integrate NDR as its first
371 layer of filter to leave out sites/regions with well-positioned nucleosomes. Indeed,
372 several TFBSs modeling methods that integrate DNase I hypersensitivity datasets,
373 have shown increased prediction power for *in vivo* binding (Zhou and Troyanskaya,
374 2015; Kelley et al., 2016; Wang et al., 2018b). Thus, it is essential to generate
375 datasets representing chromatin accessibility. To address this, recent technological
376 advances are available, such as DNase-seq, MNase-seq, FAIRE-seq and ATAC-seq
377 (Meyer and Liu, 2014). Among them, ATAC-seq is a rising star method as it requires
378 a minimal amount of input sample and even can be used at the single-cell level
379 (Buenrostro et al., 2013; Buenrostro et al., 2015; Corces et al., 2017). This is
380 particularly attractive for the plant biology community, where some plant tissues are
381 extremely scarce, such as flower meristem cells, organ primordia and root tips.
382 Furthermore, when combined with INTACT (isolation of nuclei tagged in specific cell
383 types), which allows isolation of nuclei from individual cell types of a tissue by affinity-
384 based purification, ATAC-seq is able to map chromatin accessibility with high
385 resolution and low noise from a specific cell type (Deal and Henikoff, 2011; Sijacic et
386 al., 2018).

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

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

415 **Chromatin states: histone modifications, histone variants and chromatin** 416 **remodelers**

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

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

435 Cross-talk between TFs and chromatin factors co-regulate chromatin accessibility
436 and exposure of *cis*-elements (Vachon et al., 2018). In these processes, TFs operate
437 either by recruiting chromatin factors or directly competing with them for target sites.
438 There are several examples of TF-mediated recruitment of chromatin factors in
439 plants, such as that of REF6 by NF-Y TFs for H3K27 demethylation at *SOC1*,
440 inducing flowering (Hou et al., 2014), or Polycomb mark reader TFL2/LHP1
441 recruitment by SHORT VEGETATIVE PHASE at *SEP3* for flower patterning (Liu et
442 al., 2009b), or BRAHMA and SPLAYED ATPase recruitment by LFY and SEP3 at
443 flower morphogenetic genes (Wu et al., 2012). Oppositely, several TFs were shown
444 to compete with Polycomb complexes at target genes, such as NF-YC which
445 prevents PRC2 binding to *FLOWERING LOCUS T* for floral transition (Liu et al.,
446 2018) and AG which evicts PRC2 from *KNUCKLES* for flower meristem termination
447 (Sun et al., 2014). Interestingly, at the time of flower termination, AG also has the
448 opposite effect at *WUS*, promoting PRC2 recruitment for deposition of H3K27me3
449 (Liu et al., 2011). Differences in TF behaviour for eviction versus recruitment of PRC2
450 may depend on the distance between the Polycomb recognition element (PRE) and
451 the TFBS. To this regard, large-scale analyses of ChIP-seq data revealed TFBSs in
452 plant PREs, thereby expanding the repertoire of TF-chromatin factor interactions and
453 providing resources for further exploration of the relationship between *cis*-elements
454 and TF/chromatin factor binding (Wang et al., 2016; Xiao et al., 2017). Taken

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

458 **Methylation state**

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

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

495 **3D genome and TF mediated long-range gene interactions**

496 The linear nucleotide sequences are folded into highly organized 3D architectures in
497 the nucleus of higher eukaryotes. Chromosome conformation capture (3C)
498 techniques, such as Hi-C (Eagen, 2018), revealed widespread existence of long-
499 range gene interactions within the so-called topologically associating domains (TAD).
500 Within the TAD, distal and proximal *cis*-elements relative to transcription starting sites
501 (TSS) form cell-type specific long-range interactions that in many cases are
502 established by architectural proteins such as cohesin (Yan et al., 2013; Rao et al.,
503 2017), CTCF (Phillips and Corces, 2009; Ren et al., 2017), Yin Yang 1 (Weintraub et
504 al., 2018) and others (Rada-Iglesias et al., 2018). Such interaction is a highly
505 conserved mechanism for eukaryotes to achieve spatiotemporal gene expression
506 (Sanyal et al., 2012; Harmston and Lenhard, 2013; Dekker and Misteli, 2015). TADs
507 therefore form territories within which more frequent gene interaction occurs,
508 whereas less interaction happens beyond these territories. Disruption of TAD
509 boundaries can lead to ectopic activation of gene expression and eventually to
510 noticeable phenotypes (Lupiáñez et al., 2015; Franke et al., 2016; Lupiáñez et al.,
511 2016).

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

521 role. In general, TDs are enriched with acidic and hydrophobic residues, and
522 residues that are able to form intrinsically disorder structures, such as serine, glycine
523 and proline (Staller et al., 2018). These properties appear to allow TDs to interact
524 with or recruit various factors with modest affinity but high specificity under various
525 contexts. One such factor is Mediator, a mega protein complex that can be recruited
526 by divergent TFs to connect distal and proximal *cis*-elements (Soutourina, 2017).
527 Another factor is the SAGA complex, which has recently been shown to be a general
528 factor that is required for the construction of the pre-initiation complex at the TSS for
529 transcription initiation in animal systems (Baptista et al., 2017). Despite being less
530 well characterized in plants, homolog protein components for both factors are well
531 conserved in plants (Elfving et al., 2011; Mathur et al., 2011; Moraga and Aquea,
532 2015).

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

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

559 **TF cooperative binding**

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

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

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

591 **Incorporating TFBS models in current and future analyses** 592 **of gene regulation**

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

598 **From DNA binding to gene regulation and to regulatory networks**

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

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

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

621 TFBS models are now widely applied not only to characterize gene regulatory
622 networks (GRN), but also to understand mechanisms underlying gene activation or
623 repression. For example, TFBS prediction helped to identify TFs that mediate
624 recruitment of repressive Polycomb protein complexes to specific genomic locations
625 (Xiao et al., 2017; Zhou et al., 2018). A major challenge is still to identify and validate
626 cell type-specific gene regulatory interactions, which can now be addressed by
627 combining cell-type selection by Fluorescence Activated Cell/Nuclei sorting or
628 INTACT with ChIP-seq or other epigenomic technologies (see review (Wang et al.,
629 2012)). Another promising technology is single-cell approaches (see review (Grün
630 and Van Oudenaarden, 2015)), which at the moment are still under development for
631 plant tissues (Brennecke et al., 2013).

632 **Targeted TFBS perturbation**

633 Testing the functional impact of a predicted TFBS usually involves targeted
634 mutagenesis in a transgenic context, e.g. using reporter assays, or in an endogenous
635 context using CRISPR-Cas9-based systems. By using reporter genes (e.g. GFP or
636 luciferase) under control of the target gene regulatory regions with modified TFBSs it
637 is possible to dissect spatiotemporal and quantitative changes in target gene
638 expression depending on the presence or absence of a TFBS (Benn and Dehesh,
639 2016; Díaz-Triviño et al., 2017). For example, the tissue specificity of the *AP3*
640 promoter was altered by replacing native TFBS with the ones of predicted specificity
641 towards SEP3-AG or SEP3-AP1 floral homeotic protein complexes (Smaczniak et al.,
642 2017). Combining results from TFBS prediction and reporter gene assays also helps
643 reveal the mechanisms of TFBS recruitment in a native promoter context. For
644 example, in *Drosophila* the combination of SELEX and reporter gene expression
645 (*lacZ* and GFP) experiments revealed that clusters of low affinity binding sites are
646 maintained and required for the proper tissue-specific expression of the Hox genes,
647 homeotic genes crucial for segment specification (Crocker et al., 2015). To which
648 extent the clusters of TFBSs regulate gene expression in plants is yet to be
649 determined through similar approaches.

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

673 **Evolution of TFBS and plant diversity**

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

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

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

709 **Future perspectives**

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

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

732 **FUNDING**

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

737 **AUTHOR CONTRIBUTIONS**

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

741 **CONFLICT OF INTEREST**

742 We declare no conflict of interest.

743 References

- 744 **Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H. J., Rohs, R.,**
745 **and Mann, R. S.** (2015). Deconvolving the recognition of DNA shape from
746 sequence. *Cell* **161**:307–318.
- 747 **Afek, A., Schipper, J. L., Horton, J., Gordân, R., and Lukatsky, D. B.** (2014).
748 Protein–DNA binding in the absence of specific base-pair recognition.
749 *Proceedings of the National Academy of Sciences* **111**:17140–17145.
- 750 **Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J.** (2015). Predicting the
751 sequence specificities of DNA- and RNA-binding proteins by deep learning.
752 *Nature Biotechnology* **33**:831–838.
- 753 **Allen, B. L., and Taatjes, D. J.** (2015). The Mediator complex: a central integrator of
754 transcription. *Nature Reviews Molecular Cell Biology* **16**:155–166.
- 755 **Allis, C. D., and Jenuwein, T.** (2016). The molecular hallmarks of epigenetic control.
756 *Nature Reviews Genetics* **17**:487–500.
- 757 **Allshire, R. C., and Madhani, H. D.** (2017). Ten Principles of Heterochromatin
758 Formation and Function. *Nature Publishing Group Advance Access published*
759 2017, doi:10.1038/nrm.2017.119.
- 760 **Aloy, P., and Russell, R. B.** (2006). Structural systems biology: Modelling protein
761 interactions. *Nature Reviews Molecular Cell Biology* **7**:188–197.
- 762 **Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O.** (2016). Deep learning
763 for computational biology. *Molecular Systems Biology* **12**:878.
- 764 **Austin, R. S., Hiu, S., Waese, J., Ierullo, M., Pasha, A., Wang, T. T., Fan, J.,**
765 **Foong, C., Breit, R., Desveaux, D., et al.** (2016). New BAR tools for mining
766 expression data and exploring Cis-elements in *Arabidopsis thaliana*. *Plant*
767 *Journal* **88**:490–504.
- 768 **Azad, R. N., Zafiroopoulos, D., Ober, D., Jiang, Y., Chiu, T.-P., Sagendorf, J. M.,**
769 **Rohs, R., and Tullius, T. D.** (2018). Experimental maps of DNA structure at
770 nucleotide resolution distinguish intrinsic from protein-induced DNA
771 deformations. *Nucleic Acids Research* **46**:2636–2647.
- 772 **Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A.**
773 (2004). Structure and evolution of transcriptional regulatory networks. *Current*
774 *Opinion in Structural Biology* **14**:283–291.
- 775 **Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger,**
776 **S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., et al.** (2009).
777 Diversity and complexity in DNA recognition by transcription factors. *Science*
778 **324**:1720–1723.
- 779 **Baptista, T., Grünberg, S., Minoungou, N., Koster, M. J. E., Timmers, H. T. M.,**
780 **Hahn, S., Devys, D., and Tora, L.** (2017). SAGA Is a General Cofactor for

- 781 RNA Polymerase II Transcription. *Molecular Cell* Advance Access published
782 2017, doi:10.1016/j.molcel.2017.08.016.
- 783 **Barakat, T. S., Halbritter, F., Zhang, M., Rendeiro, A. F., Perenthaler, E., Bock,**
784 **C., and Chambers, I.** (2018). Functional Dissection of the Enhancer
785 Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23**:276-288.e8.
- 786 **Bartlett, A., O'Malley, R. C., Huang, S. C., Galli, M., Nery, J. R., Gallavotti, A.,**
787 **and Ecker, J. R.** (2017). Mapping genome-wide transcription-factor binding
788 sites using DAP-seq. *Nature Protocols* **12**:1659–1672.
- 789 **Beltrao, P., Kiel, C., and Serrano, L.** (2007). Structures in systems biology. *Current*
790 *Opinion in Structural Biology* **17**:378–384.
- 791 **Benn, G., and Dehesh, K.** (2016). Quantitative Analysis of Cis -Regulatory Element
792 Activity Using Synthetic Promoters in Transgenic Plants. *Methods in Molecular*
793 *Biology* **1482**:15–30.
- 794 **Berger, M. F., and Bulyk, M. L.** (2009). Universal protein-binding microarrays for the
795 comprehensive characterization of the dna-binding specificities of transcription
796 factors. *Nature Protocols* **4**:393–411.
- 797 **Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and**
798 **Bulyk, M. L.** (2006). Compact, universal DNA microarrays to comprehensively
799 determine transcription-factor binding site specificities. *Nature Biotechnology*
800 **24**:1429–1435.
- 801 **Bezhani, S., Winter, C., Hershman, S., Wagner, J. D., Kennedy, J. F., Kwon, C.**
802 **S., Pfluger, J., Su, Y., and Wagner, D.** (2007). Unique, Shared, and
803 Redundant Roles for the Arabidopsis SWI/SNF Chromatin Remodeling
804 ATPases BRAHMA and SPLAYED. *The Plant Cell* **19**:403–416.
- 805 **Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X.,**
806 **Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., et**
807 **al.** (2013). Accounting for technical noise in single-cell RNA-seq experiments.
808 *Nature Methods* **10**:1093–1098.
- 809 **Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J.**
810 (2013). Transposition of native chromatin for fast and sensitive epigenomic
811 profiling of open chromatin, DNA-binding proteins and nucleosome position.
812 *Nature Methods* **10**:1213–1218.
- 813 **Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder,**
814 **M. P., Chang, H. Y., and Greenleaf, W. J.** (2015). Single-cell chromatin
815 accessibility reveals principles of regulatory variation. *Nature* **523**:486–490.
- 816 **Bulyk, M. L.** (2002). Nucleotides of transcription factor binding sites exert
817 interdependent effects on the binding affinities of transcription factors. *Nucleic*
818 *Acids Research* **30**:1255–1261.
- 819 **Chen, Y., Zhang, X., Dantas Machado, A. C., Ding, Y., Chen, Z., Qin, P. Z., Rohs,**
820 **R., and Chen, L.** (2013). Structure of p53 binding to the BAX response

- 821 element reveals DNA unwinding and compression to accommodate base-pair
822 insertion. *Nucleic Acids Research* **41**:8368–8376.
- 823 **Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-**
824 **Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S.,**
825 **Wu, B., et al.** (2017). An improved ATAC-seq protocol reduces background
826 and enables interrogation of frozen tissues. *Nature Methods Advance Access*
827 published 2017, doi:10.1038/nmeth.4396.
- 828 **Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S.,**
829 **Alsawadi, A., Valenti, P., Plaza, S., Payre, F., et al.** (2015). Low affinity
830 binding site clusters confer HOX specificity and regulatory robustness. *Cell*
831 **160**:191–203.
- 832 **Deal, R. B., and Henikoff, S.** (2011). The INTACT method for cell type-specific
833 gene expression and chromatin profiling in *Arabidopsis thaliana*. *Nature*
834 *Protocols* **6**:56–68.
- 835 **Dekker, J., and Misteli, T.** (2015). Long-Range Chromatin Interactions. *Cold Spring*
836 *Harbor Perspectives in Biology* **7**:1–23.
- 837 **Díaz-Triviño, S., Long, Y., Scheres, B., and Blilou, I.** (2017). Analysis of a Plant
838 Transcriptional Regulatory Network Using Transient Expression Systems.
839 *Methods in Molecular Biology* **1629**:83–104.
- 840 **Dong, P., Tu, X., Chu, P. Y., Lü, P., Zhu, N., Grierson, D., Du, B., Li, P., and**
841 **Zhong, S.** (2017). 3D Chromatin Architecture of Large Plant Genomes
842 Determined by Local A/B Compartments. *Molecular Plant* **10**:1497–1509.
- 843 **Dong, Q., Li, N., Li, X., Yuan, Z., Xie, D., Wang, X., Li, J., Yu, Y., Wang, J., Ding,**
844 **B., et al.** (2018). Genome-wide Hi-C analysis reveals extensive hierarchical
845 chromatin interactions in rice. *The Plant Journal Advance Access* published
846 2018, doi:10.1111/tpj.13925.
- 847 **Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y.** (2015). A
848 widespread role of the motif environment in transcription factor binding across
849 diverse protein families. *Genome Research* **25**:1268–1280.
- 850 **Eagen, K. P.** (2018). Principles of Chromosome Architecture Revealed by Hi-C.
851 *Trends in Biochemical Sciences* **43**:469–478.
- 852 **Elfving, N., Davoine, C., Benlloch, R., Blomberg, J., Brannstrom, K., Muller, D.,**
853 **Nilsson, A., Ulfstedt, M., Ronne, H., Wingsle, G., et al.** (2011). The
854 *Arabidopsis thaliana* Med25 mediator subunit integrates environmental cues to
855 control plant development. *Proceedings of the National Academy of Sciences*
856 **108**:8245–8250.
- 857 **Ezer, D., Jung, J.-H., Lan, H., Biswas, S., Gregoire, L., Box, M. S.,**
858 **Charoensawan, V., Cortijo, S., Lai, X., Stöckle, D., et al.** (2017). The
859 evening complex coordinates environmental and endogenous signals in
860 *Arabidopsis*. *Nature Plants* **3**:17087.

- 861 **Farrel, A., and Guo, J.** (2017). An efficient algorithm for improving structure-based
862 prediction of transcription factor binding sites. *BMC Bioinformatics* **18**:342.
- 863 **Farrel, A., Murphy, J., and Guo, J. T.** (2016). Structure-based prediction of
864 transcription factor binding specificity using an integrative energy function.
865 *Bioinformatics* **32**:i306–i313.
- 866 **Finnegan, E. J., Genger, R. K., Kovac, K., Peacock, W. J., and Dennis, E. S.**
867 (1998). DNA methylation and the promotion of flowering by vernalization.
868 *Proceedings of the National Academy of Sciences* **95**:5824–5829.
- 869 **Fisher, W. W., Li, J. J., Hammonds, A. S., Brown, J. B., Pfeiffer, B. D.,**
870 **Weiszmann, R., MacArthur, S., Thomas, S., Stamatoyannopoulos, J. A.,**
871 **Eisen, M. B., et al.** (2012). DNA regions bound at low occupancy by
872 transcription factors do not drive patterned reporter gene expression in
873 *Drosophila*. *Proceedings of the National Academy of Sciences* **109**:21330–
874 21335.
- 875 **Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin,**
876 **R., Kraft, K., Kempfer, R., Jerković, I., Chan, W. L., et al.** (2016). Formation
877 of new chromatin domains determines pathogenicity of genomic duplications.
878 *Nature* **538**:265–269.
- 879 **Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. Bin, Orlov, Y.**
880 **L., Velkov, S., Ho, A., Mei, P. H., et al.** (2009). An oestrogen-receptor- α -
881 bound human chromatin interactome. *Nature* **462**:58–64.
- 882 **Fyodorov, D. V., Zhou, B.-R., Skoultchi, A. I., and Bai, Y.** (2017). Emerging roles
883 of linker histones in regulating chromatin structure and function. *Nature*
884 *Reviews Molecular Cell Biology* Advance Access published 2017,
885 doi:10.1038/nrm.2017.94.
- 886 **Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M. A.** (2014). Enhanced
887 Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS*
888 *Computational Biology* **10**.
- 889 **Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M. L.**
890 (2013). Genomic Regions Flanking E-Box Binding Sites Influence DNA
891 Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell*
892 *Reports* **3**:1093–1104.
- 893 **Greenham, K., and McClung, C. R.** (2015). Integrating circadian dynamics with
894 physiological processes in plants. *Nature Reviews Genetics* **16**:598–610.
- 895 **Greenside, P. G., Shimko, T., Fordyce, P., and Kundaje, A.** (2018). Discovering
896 epistatic feature interactions from neural network models of regulatory DNA
897 sequences Advance Access published July 26, 2018, doi:10.1101/302711.
- 898 **Grün, D., and Van Oudenaarden, A.** (2015). Design and Analysis of Single-Cell
899 Sequencing Experiments. *Cell* **163**:799–810.

- 900 **Guo, Y., Tian, K., Zeng, H., Guo, X., and Gifford, D. K.** (2018). A novel k-mer set
901 memory (KSM) motif representation improves regulatory variant prediction.
902 *Genome Research* Advance Access published 2018, doi:10.1101/130815.
- 903 **Gutin, J., Sadeh, R., Bodenheimer, N., Joseph-Strauss, D., Klein-Brill, A.,**
904 **Alajem, A., Ram, O., and Friedman, N.** (2018). Fine-Resolution Mapping of
905 TF Binding and Chromatin Interactions. *Cell Reports* **22**:2601–2614.
- 906 **Guturu, H., Doxey, A. C., Wenger, A. M., and Bejerano, G.** (2013). Structure-aided
907 prediction of mammalian transcription factor complexes in conserved non-
908 coding elements. *Philosophical transactions of the Royal Society of London.*
909 *Series B, Biological sciences* **368**:20130029.
- 910 **Hancock, S. P., Ghane, T., Cascio, D., Rohs, R., Di Felice, R., and Johnson, R.**
911 **C.** (2013). Control of DNA minor groove width and Fis protein binding by the
912 purine 2-amino group. *Nucleic Acids Research* **41**:6750–6760.
- 913 **Harmston, N., and Lenhard, B.** (2013). Chromatin and epigenetic features of long-
914 range gene regulation. *Nucleic Acids Research* **41**:7185–7199.
- 915 **Hennig, L., and Derkacheva, M.** (2009). Diversity of Polycomb group complexes in
916 plants: same rules, different players? *Trends in Genetics* **25**:414–423.
- 917 **Ho, L., and Crabtree, G. R.** (2010). Chromatin remodelling during development.
918 *Nature* **463**:474–484.
- 919 **Hombach, D., Schwarz, J. M., Robinson, P. N., Schuelke, M., and Seelow, D.**
920 (2016). A systematic, large-scale comparison of transcription factor binding
921 site models. *BMC Genomics* **17**:1–10.
- 922 **Hou, X., Zhou, J., Liu, C., Liu, L., Shen, L., and Yu, H.** (2014). Nuclear factor Y-
923 mediated H3K27me3 demethylation of the SOC1 locus orchestrates flowering
924 responses of Arabidopsis. *Nature Communications* **5**:1–14.
- 925 **Hu, Z., Killion, P. J., and Iyer, V. R.** (2007). Genetic reconstruction of a functional
926 transcriptional regulatory network. *Nature Genetics* **39**:683–687.
- 927 **Hu, M., Yu, J., Taylor, J. M. G., Chinnaiyan, A. M., and Qin, Z. S.** (2010). On the
928 detection and refinement of transcription factor binding sites using ChIP-Seq
929 data. *Nucleic Acids Research* **38**:2154–2167.
- 930 **Huang, H., and Nusinow, D. A.** (2016). Into the Evening: Complex Interactions in
931 the Arabidopsis Circadian Clock. *Trends in Genetics* **32**:674–686.
- 932 **Hugouvieux, V., and Zubieta, C.** (2018). MADS transcription factors cooperate :
933 complexities of complex formation. *Journal of Experimental Botany* Advance
934 Access published 2018, doi:10.1093/jxb/ery099.
- 935 **Hugouvieux, V., Silva, C. S., Jourdain, A., Stigliani, A., Charras, Q., Conn, V.,**
936 **Conn, S. J., Carles, C. C., Parcy, F., and Zubieta, C.** (2018). Tetramerization
937 of MADS family transcription factors SEPALLATA3 and AGAMOUS is required

- 938 for floral meristem determinacy in Arabidopsis. *Nucleic Acids Research*
939 Advance Access published 2018, doi:10.1093/nar/gky205.
- 940 **Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., and Bulyk, M. L.** (2015).
941 UniPROBE, update 2015: New tools and content for the online database of
942 protein-binding microarray data on protein-DNA interactions. *Nucleic Acids*
943 *Research* **43**:D117–D122.
- 944 **Ichihashi, Y., Aguilar-Martinez, J. A., Farhi, M., Chitwood, D. H., Kumar, R.,**
945 **Millon, L. V., Peng, J., Maloof, J. N., and Sinha, N. R.** (2014). Evolutionary
946 developmental transcriptomics reveals a gene network module regulating
947 interspecific diversity in plant leaf shape. *Proceedings of the National*
948 *Academy of Sciences* **111**:E2616–E2621.
- 949 **Ingouff, M., and Berger, F.** (2010). Histone3 variants in plants. *Chromosoma*
950 **119**:27–33.
- 951 **Isakova, A., Groux, R., Imbeault, M., Rainer, P., Alpern, D., Dainese, R.,**
952 **Ambrosini, G., Trono, D., Bucher, P., and Deplancke, B.** (2017). SMiLE-
953 seq identifies binding motifs of single and dimeric transcription factors. *Nature*
954 *Methods* **14**:316–322.
- 955 **Iwafuchi-Doi, M., and Zaret, K. S.** (2016). Cell fate control by pioneer transcription
956 factors. *Development* **143**:1833–1837.
- 957 **Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., and Gao, G.** (2017).
958 PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory
959 interactions in plants. *Nucleic Acids Research* **45**:D1040–D1045.
- 960 **Johnson, L., Mollah, S., Garcia, B. A., Muratore, T. L., Shabanowitz, J., Hunt, D.**
961 **F., and Jacobsen, S. E.** (2004). Mass spectrometry analysis of Arabidopsis
962 histone H3 reveals distinct combinations of post-translational modifications.
963 *Nucleic Acids Research* **32**:6511–6518.
- 964 **Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B.** (2007). Genome-wide
965 mapping of in vivo protein-DNA interactions. *Science* **316**:1497–1502.
- 966 **Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M.,**
967 **Vaquerizas, J. M., Yan, J., Sillanpa, M. J., et al.** (2010). Multiplexed
968 massively parallel SELEX for characterization of human transcription factor
969 binding specificities. *Genome Research* Advance Access published 2010,
970 doi:10.1101/gr.100552.109.
- 971 **Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M.,**
972 **Kivioja, T., Morgunova, E., and Taipale, J.** (2015). DNA-dependent
973 formation of transcription factor pairs alters their binding specificity. *Nature*
974 **527**:384–388.
- 975 **Käppel, S., Melzer, R., Rümpler, F., Gafert, C., and Theißen, G.** (2018). The Floral
976 Homeotic Protein SEPALLATA3 Recognizes Target DNA Sequences By
977 Shape Readout Involving A Conserved Arginine Residue In The MADS-

- 978 Domain. *The Plant Journal* Advance Access published 2018,
979 doi:10.1101/133678.
- 980 **Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K., and Henikoff, S.** (2014).
981 High-resolution mapping of transcription factor binding sites on native
982 chromatin. *Nature Methods* **11**:203–209.
- 983 **Kaufmann, K., Muiño, J. M., Østerås, M., Farinelli, L., Krajewski, P., and**
984 **Angenent, G. C.** (2010). Chromatin immunoprecipitation (ChIP) of plant
985 transcription factors followed by sequencing (ChIP-SEQ) or hybridization to
986 whole genome arrays (ChIP-CHIP). *Nature Protocols* **5**:457–472.
- 987 **Kelley, D. R., Snoek, J., and Rinn, J. L.** (2016). Basset: Learning the regulatory
988 code of the accessible genome with deep convolutional neural networks.
989 *Genome Research* **26**:990–999.
- 990 **Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., Van**
991 **Der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., et al.** (2018).
992 JASPAR 2018: Update of the open-access database of transcription factor
993 binding profiles and its web framework. *Nucleic Acids Research* **46**:D260–
994 D266.
- 995 **Kiesel, A., Roth, C., Ge, W., Wess, M., Meier, M., and Söding, J.** (2018). The
996 BaMM web server for de-novo motif discovery and regulatory sequence
997 analysis. *Nucleic Acids Research* **46**:W215–W220.
- 998 **Kim, M. Y., and Zilberman, D.** (2014). DNA methylation as a system of plant
999 genomic immunity. *Trends in Plant Science* **19**:320–326.
- 1000 **Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R.**
1001 (2013). The next-generation sequencing revolution and its impact on
1002 genomics. *Cell* **155**:27–38.
- 1003 **Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I.**
1004 **E., Bajic, V. B., and Makeev, V. J.** (2013). HOCOMOCO: A comprehensive
1005 collection of human transcription factor binding sites models. *Nucleic Acids*
1006 *Research* **41**:195–202.
- 1007 **Kulkarni, S. R., Vanechoutte, D., Van de Velde, J., and Vandepoele, K.** (2017).
1008 TF2Network: predicting transcription factor regulators and gene regulatory
1009 networks in Arabidopsis using publicly available binding site information.
1010 *Nucleic Acids Research* **46**.
- 1011 **Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X.,**
1012 **Taipale, J., Hughes, T. R., and Weirauch, M. T.** (2018). The Human
1013 Transcription Factors. *Cell* **172**:650–665.
- 1014 **Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A. C., Riley, T. R.,**
1015 **Sandstrom, R., Sabo, P. J., Lu, Y., Rohs, R., Stamatoyannopoulos, J. A.,**
1016 **et al.** (2013). Probing DNA shape and methylation state on a genomic scale
1017 with DNase I. *Proceedings of the National Academy of Sciences* **110**:6376–
1018 6381.

- 1019 **Lecun, Y., Bengio, Y., and Hinton, G.** (2015). Deep learning. *Nature* **521**:436–444.
- 1020 **Lelli, K. M., Slattery, M., and Mann, R. S.** (2012). Disentangling the Many Layers of
1021 Eukaryotic Transcriptional Regulation. *Annual Review of Genetics* **46**:43–68.
- 1022 **Levo, M., Zalckvar, E., Sharon, E., Machado, A. C. D., Kalma, Y., Lotam-
1023 Pompan, M., Weinberger, A., Yakhini, Z., Rohs, R., and Segal, E.** (2015).
1024 Unraveling determinants of transcription factor binding outside the core
1025 binding site. *Genome Research Advance Access* published 2015,
1026 doi:10.1101/gr.185033.114.6.
- 1027 **Li, G., Cai, L., Chang, H., Hong, P., Zhou, Q., Kulakova, E. V., Kolchanov, N. A.,
1028 and Ruan, Y.** (2014). Chromatin interaction analysis with paired-end tag
1029 (ChIA-PET) sequencing technology and application. *BMC Genomics* **15**:1–10.
- 1030 **Li, C., Gu, L., Gao, L., Chen, C., Wei, C. Q., Qiu, Q., Chien, C. W., Wang, S.,
1031 Jiang, L., Ai, L. F., et al.** (2016). Concerted genomic targeting of H3K27
1032 demethylase REF6 and chromatin-remodeling ATPase BRM in Arabidopsis.
1033 *Nature Genetics* **48**:687–693.
- 1034 **Li, J., Sagendorf, J. M., Chiu, T. P., Pasi, M., Perez, A., and Rohs, R.** (2017).
1035 Expanding the repertoire of DNA shape features for genome-scale studies of
1036 transcription factor binding. *Nucleic acids research* **45**:12877–12887.
- 1037 **Lin, C. K., and Chen, C. Y.** (2013). PiDNA: Predicting protein-DNA interactions with
1038 structural models. *Nucleic acids research* **41**:523–530.
- 1039 **Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar,
1040 A. H., and Ecker, J. R.** (2008). Highly Integrated Single-Base Resolution
1041 Maps of the Epigenome in Arabidopsis. *Cell* **133**:523–536.
- 1042 **Liu, Z., Mao, F., Guo, J., Yan, B., Wang, P., Qu, Y., and Xu, Y.** (2005). Quantitative
1043 evaluation of protein-DNA interactions using an optimized knowledge-based
1044 potential. *Nucleic Acids Research* **33**:546–558.
- 1045 **Liu, H., Shi, Y., Chen, X. S., and Warshel, A.** (2009a). Simulating the electrostatic
1046 guidance of the vectorial translocations in hexameric helicases and
1047 translocases. *Proceedings of the National Academy of Sciences* **106**:7449–
1048 7454.
- 1049 **Liu, C., Xi, W., Shen, L., Tan, C., and Yu, H.** (2009b). Regulation of Floral
1050 Patterning by Flowering Time Genes. *Developmental Cell* **16**:711–722.
- 1051 **Liu, X., Kim, Y. J., Muller, R., Yumul, R. E., Liu, C., Pan, Y., Cao, X., Goodrich, J.,
1052 and Chen, X.** (2011). AGAMOUS Terminates Floral Stem Cell Maintenance in
1053 Arabidopsis by Directly Repressing WUSCHEL through Recruitment of
1054 Polycomb Group Proteins. *The Plant Cell* **23**:3654–3670.
- 1055 **Liu, C., Wang, C., Wang, G., Becker, C., Zaidem, M., and Weigel, D.** (2016).
1056 Genome-wide analysis of chromatin packing in *Arabidopsis thaliana* at single-
1057 gene resolution. *Genome Research Advance Access* published 2016,
1058 doi:10.1101/gr.204032.116.

- 1059 **Liu, C., Cheng, Y. J., Wang, J. W., and Weigel, D.** (2017). Prominent topologically
1060 associated domains differentiate global chromatin packing in rice from
1061 *Arabidopsis*. *Nature Plants* **3**:742–748.
- 1062 **Liu, X., Yang, Y., Hu, Y., Zhou, L., Li, Y., and Hou, X.** (2018). Temporal-Specific
1063 Interaction of NF-YC and CURLY LEAF during the Floral Transition Regulates
1064 Flowering. *Plant Physiology* **177**:pp.00296.2018.
- 1065 **Lucas, M., Davière, J.-M., Rodríguez-Falcón, M., Pontin, M., Iglesias-Pedraz, J.**
1066 **M., Lorrain, S., Fankhauser, C., Blázquez, M. A., Titarenko, E., Prat, S., et**
1067 **al.** (2008). A molecular framework for light and gibberellin control of cell
1068 elongation. *Nature* **451**:480–484.
- 1069 **Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E.,**
1070 **Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., et al.** (2015). Disruptions of
1071 topological chromatin domains cause pathogenic rewiring of gene-enhancer
1072 interactions. *Cell* **161**:1012–1025.
- 1073 **Lupiáñez, D. G., Spielmann, M., and Mundlos, S.** (2016). Breaking TADs: How
1074 Alterations of Chromatin Domains Result in Disease. *Trends in Genetics*
1075 **32**:225–237.
- 1076 **Marcovitz, A., and Levy, Y.** (2011). Frustration in protein-DNA binding influences
1077 conformational switching and target search kinetics. *Proceedings of the*
1078 *National Academy of Sciences* **108**:17957–17962.
- 1079 **Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker,**
1080 **T., Radchuk, V., Dockter, C., Hedley, P. E., Russell, J., et al.** (2017). A
1081 chromosome conformation capture ordered sequence of the barley genome.
1082 *Nature* **544**:427–433.
- 1083 **Mateos, J. L., Tilmes, V., Madrigal, P., Severing, E., Richter, R., Rijkenberg, C.**
1084 **W. M., Krajewski, P., and Coupland, G.** (2017). Divergence of regulatory
1085 networks governed by the orthologous transcription factors FLC and PEP1 in
1086 Brassicaceae species. *Proceedings of the National Academy of Sciences*
1087 Advance Access published 2017, doi:10.1073/pnas.1618075114.
- 1088 **Mathelier, A., and Wasserman, W. W.** (2013). The Next Generation of Transcription
1089 Factor Binding Site Prediction. *PLoS Computational Biology* **9**.
- 1090 **Mathelier, A., Xin, B., Chiu, T. P., Yang, L., Rohs, R., and Wasserman, W. W.**
1091 (2016). DNA Shape Features Improve Transcription Factor Binding Site
1092 Predictions In Vivo. *Cell Systems* **3**:278-286.e4.
- 1093 **Mathur, S., Vyas, S., Kapoor, S., and Tyagi, A. K.** (2011). The Mediator Complex in
1094 Plants: Structure, Phylogeny, and Expression Profiling of Representative
1095 Genes in a Dicot (*Arabidopsis*) and a Monocot (Rice) during Reproduction and
1096 Abiotic Stress. *Plant Physiology* **157**:1609–1627.
- 1097 **Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A.,**
1098 **Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al.** (2006).

- 1099 TRANSFAC and its module TRANSCompel: transcriptional gene regulation in
1100 eukaryotes. *Nucleic Acids Research* **34**:D108–D110.
- 1101 **Mayran, A., and Drouin, J.** (2018). Pioneer transcription factors shape the
1102 epigenetic landscape. *Journal of Biological Chemistry Advance Access*
1103 published 2018, doi:10.1074/jbc.R117.001232.
- 1104 **McGinty, R. K., and Tan, S.** (2015). Nucleosome structure and function. *Chemical*
1105 *Reviews* **115**:2255–2273.
- 1106 **Meijsing, S. H., Pufall, M. A., So, A. Y., Bates, D. L., Chen, L., and Yamamoto, K.**
1107 **R.** (2009). DNA binding site sequence directs glucocorticoid receptor structure
1108 and activity. *Science* **324**:407–410.
- 1109 **Melzer, R., Verelst, W., and Theißen, G.** (2009). The class E floral homeotic protein
1110 SEPALLATA3 is sufficient to loop DNA in 'floral quartet'-like complexes in
1111 vitro. *Nucleic Acids Research* **37**:144–157.
- 1112 **Mendes, M. A., Guerra, R. F., Berns, M. C., Manzo, C., Masiero, S., Finzi, L.,**
1113 **Kater, M. M., and Colombo, L.** (2013). MADS Domain Transcription Factors
1114 Mediate Short-Range DNA Looping That Is Essential for Target Gene
1115 Expression in Arabidopsis. *The Plant Cell* **25**:2560–2572.
- 1116 **Meyer, C. A., and Liu, X. S.** (2014). Identifying and mitigating bias in next-generation
1117 sequencing methods for chromatin biology. *Nature Reviews Genetics* **15**:709–
1118 721.
- 1119 **Minguet, E. G., Segard, S., Charavay, C., and Parcy, F.** (2015). MORPHEUS, a
1120 webtool for transcription factor binding analysis using position weight matrices
1121 with dependency. *PLoS ONE* **10**:1–12.
- 1122 **Moraga, F., and Aquea, F.** (2015). Composition of the SAGA complex in plants and
1123 its role in controlling gene expression in response to abiotic stresses. *Frontiers*
1124 *in Plant Science* **6**:1–9.
- 1125 **Moyroud, E., Minguet, E. G., Ott, F., Yant, L., Posé, D., Monniaux, M., Blanchet,**
1126 **S., Bastien, O., Thévenon, E., Weigel, D., et al.** (2011). Prediction of
1127 Regulatory Interactions from Genome Sequences Using a Biophysical Model
1128 for the *Arabidopsis* LEAFY Transcription Factor. *The Plant Cell* **23**:1293–1306.
- 1129 **Muiño, J. M., Smaczniak, C., Angenent, G. C., Kaufmann, K., and Van Dijk, A. D.**
1130 **J.** (2014). Structural determinants of DNA recognition by plant MADS-domain
1131 transcription factors. *Nucleic Acids Research* **42**:2138–2146.
- 1132 **Muino, J. M., De Bruijn, S., Pajoro, A., Geuten, K., Vingron, M., Angenent, G. C.,**
1133 **and Kaufmann, K.** (2016). Evolution of DNA-binding sites of a floral master
1134 regulatory transcription factor. *Molecular Biology and Evolution* **33**:185–200.
- 1135 **Müller, F., Zaucker, A., and Tora, L.** (2010). Developmental regulation of
1136 transcription initiation: More than just changing the actors. *Current Opinion in*
1137 *Genetics and Development* **20**:533–540.

- 1138 **Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W.**
1139 **J., and Chang, H. Y.** (2016). HiChIP: Efficient and sensitive analysis of
1140 protein-directed genome architecture. *Nature Methods* **13**:919–922.
- 1141 **Nieto, C., López-Salmerón, V., Davière, J. M., and Prat, S.** (2015). ELF3-PIF4
1142 interaction regulates plant growth independently of the evening complex.
1143 *Current Biology* **25**:187–193.
- 1144 **Nusinow, D. A., Helfer, A., Hamilton, E. E., King, J. J., Imaizumi, T., Schultz, T.**
1145 **F., Farré, E. M., and Kay, S. A.** (2011). The ELF4-ELF3-LUX complex links
1146 the circadian clock to diurnal control of hypocotyl growth. *Nature* **475**:398–402.
- 1147 **O’Malley, R. C., Huang, S. shan C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J.**
1148 **R., Galli, M., Gallavotti, A., and Ecker, J. R.** (2016). Cistrome and
1149 Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **166**:1598.
- 1150 **Omidi, S., Zavolan, M., Pachkov, M., Breda, J., Berger, S., and van Nimwegen,**
1151 **E.** (2017). Automated incorporation of pairwise dependency in transcription
1152 factor binding site prediction using dinucleotide weight tensors. *PLoS*
1153 *Computational Biology* **13**:1–22.
- 1154 **Pachkov, M., Balwierz, P. J., Arnold, P., Ozonov, E., and Van Nimwegen, E.**
1155 (2013). SwissRegulon, a database of genome-wide annotations of regulatory
1156 sites: Recent updates. *Nucleic Acids Research* **41**:214–220.
- 1157 **Paillard, G., and Lavery, R.** (2004). Analyzing Protein-DNA Recognition
1158 Mechanisms. *Structure* **12**:113–122.
- 1159 **Pajoro, A., Madrigal, P., Muiño, J. M., Matus, J. T., Jin, J., Mecchia, M. A.,**
1160 **Debernardi, J. M., Palatnik, J. F., Balazadeh, S., Arif, M., et al.** (2014).
1161 Dynamics of chromatin accessibility and gene regulation by MADS-domain
1162 transcription factors in flower development. *Genome Biology* **15**:R41.
- 1163 **Park, P. J.** (2009). ChIP-seq: Advantages and challenges of a maturing technology.
1164 *Nature Reviews Genetics* **10**:669–680.
- 1165 **Patel, A., Yang, P., Tinkham, M., Pradhan, M., Sun, M. A., Wang, Y., Hoang, D.,**
1166 **Wolf, G., Horton, J. R., Zhang, X., et al.** (2018). DNA Conformation Induces
1167 Adaptable Binding by Tandem Zinc Finger Proteins. *Cell* **173**:221-233.e12.
- 1168 **Phillips, J. E., and Corces, V. G.** (2009). CTCF: Master Weaver of the Genome.
1169 *Cell* **137**:1194–1211.
- 1170 **Pruneda-Paz, J. L., Breton, G., Nagel, D. H., Kang, S. E., Bonaldi, K., Doherty, C.**
1171 **J., Ravelo, S., Galli, M., Ecker, J. R., and Kay, S. A.** (2014). A Genome-
1172 Scale Resource for the Functional Characterization of Arabidopsis
1173 Transcription Factors. *Cell Reports* **8**:622–632.
- 1174 **Qin, Q., and Feng, J.** (2017). Imputation for transcription factor binding predictions
1175 based on deep learning. *PLoS Computational Biology* Advance Access
1176 published 2017.

- 1177 **Rada-Iglesias, A., Grosveld, F. G., and Papantonis, A.** (2018). Forces driving the
1178 three-dimensional folding of eukaryotic genomes. *Molecular Systems Biology*
1179 **14:e8214.**
- 1180 **Rao, S. S. P., Huang, S. C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M.,**
1181 **Kieffer-Kwon, K. R., Sanborn, A. L., Johnstone, S. E., Bascom, G. D.,**
1182 **Bochkov, I. D., et al.** (2017). Cohesin Loss Eliminates All Loop Domains. *Cell*
1183 **171:305-320.e24.**
- 1184 **Rao, S., Chiu, T. P., Kribelbauer, J. F., Mann, R. S., Bussemaker, H. J., and**
1185 **Rohs, R.** (2018). Systematic prediction of DNA shape changes due to CpG
1186 methylation explains epigenetic effects on protein-DNA binding. *Epigenetics*
1187 *and Chromatin* **11:1–11.**
- 1188 **Rastogi, C., Rube, H. T., Kribelbauer, J. F., Crocker, J., Loker, R. E., Martini, G.**
1189 **D., Laptenko, O., Freed-Pastor, W. A., Prives, C., Stern, D. L., et al.** (2018).
1190 Accurate and sensitive quantification of protein-DNA binding affinity.
1191 *Proceedings of the National Academy of Sciences Advance Access published*
1192 *2018, doi:10.1073/pnas.1714376115.*
- 1193 **Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., Larson, D. R., and Zhao,**
1194 **K.** (2017). CTCF-Mediated Enhancer-Promoter Interaction Is a Critical
1195 Regulator of Cell-to-Cell Variation of Gene Expression. *Molecular Cell*
1196 **67:1049-1058.e6.**
- 1197 **Rhee, H. S., and Pugh, B. F.** (2011). Comprehensive genome-wide protein-DNA
1198 interactions detected at single-nucleotide resolution. *Cell* **147:1408–1419.**
- 1199 **Riley, T. R., Lazarovici, A., Mann, R. S., and Bussemaker, H. J.** (2015). Building
1200 accurate sequence-to-affinity models from high-throughput in vitro protein-
1201 DNA binding data using featureREDUCE. *eLife* **4:1–14.**
- 1202 **Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T.,**
1203 **Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al.** (2007). Genome-
1204 wide profiles of STAT1 DNA association using chromatin immunoprecipitation
1205 and massively parallel sequencing. *Nature Methods* **4:651–657.**
- 1206 **Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B.** (2009).
1207 The role of DNA shape in protein-DNA recognition. *Nature* **461:1248–1253.**
- 1208 **Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., and Mann, R. S.** (2010).
1209 Origins of Specificity in Protein-DNA Recognition. *Annual Review of*
1210 *Biochemistry* **79:233–269.**
- 1211 **Ruan, S., and Stormo, G. D.** (2017). Inherent limitations of probabilistic models for
1212 protein-DNA binding specificity. *PLOS Computational Biology* **13:e1005638.**
- 1213 **Ruan, S., Swamidass, S. J., and Stormo, G. D.** (2017). BEESEM: Estimation of
1214 binding energy models using HT-SELEX data. *Bioinformatics* **33:2288–2295.**

- 1215 **Ruelens, P., Zhang, Z., van Mourik, H., Maere, S., Kaufmann, K., and Geuten, K.**
 1216 (2017). The Origin of Floral Organ Identity Quartets. *The Plant Cell* **29**:229–
 1217 242.
- 1218 **Santolini, M., Mora, T., and Hakim, V.** (2014). A general pairwise interaction model
 1219 provides an accurate description of in Vivo transcription factor binding sites.
 1220 *PLoS ONE* **9**.
- 1221 **Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J.** (2012). The long-range
 1222 interaction landscape of gene promoters. *Nature* **489**:109–113.
- 1223 **Sayou, C., Monniaux, M., Nanao, M. H., Moyroud, E., Brockington, S. F.,**
 1224 **Thévenon, E., Chahtane, H., Warthmann, N., Melkonian, M., Zhang, Y., et**
 1225 **al.** (2014). A promiscuous intermediate underlies the evolution of LEAFY DNA
 1226 binding specificity. *Science* **343**:645–648.
- 1227 **Sayou, C., Nanao, M. H., Jamin, M., Pose, D., Thevenon, E., Gregoire, L.,**
 1228 **Tichtinsky, G., Denay, G., Ott, F., Llobet, M. P., et al.** (2016). A SAM
 1229 oligomerization domain shapes the genomic binding landscape of the LEAFY
 1230 transcription factor. *Nature Communications* **48**:829–834.
- 1231 **Schmidl, C., Rendeiro, A. F., Sheffield, N. C., and Bock, C.** (2015).
 1232 ChIPmentation: Fast, robust, low-input ChIP-seq for histones and transcription
 1233 factors. *Nature Methods* **12**:963–965.
- 1234 **Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall,**
 1235 **A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., et al.** (2010).
 1236 Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription
 1237 Factor Binding. *Science* **328**:6.
- 1238 **Schneider, T. D., and Stephens, R. M.** (1990). Sequence logos: A new way to
 1239 display consensus sequences. *Nucleic Acids Research* **18**:6097–6100.
- 1240 **Setty, M., and Leslie, C. S.** (2015). SeqGL Identifies Context-Dependent Binding
 1241 Signals in Genome-Wide Regulatory Element Maps. *PLoS Computational*
 1242 *Biology* **11**:1–22.
- 1243 **Shaffer, P. L., Jivan, A., Dollins, D. E., Claessens, F., and Gewirth, D. T.** (2004).
 1244 Structural basis of androgen receptor binding to selective androgen response
 1245 elements. *Proceedings of the National Academy of Sciences of the United*
 1246 *States of America* **101**:4758–4763.
- 1247 **Sharon, E., Lubliner, S., and Segal, E.** (2008). A feature-based approach to
 1248 modeling protein-DNA interactions. *PLoS Computational Biology* **4**.
- 1249 **Siddharthan, R.** (2010). Dinucleotide weight matrices for predicting transcription
 1250 factor binding sites: Generalizing the position weight matrix. *PLoS ONE* **5**.
- 1251 **Siggers, T., Duyzend, M. H., Reddy, J., Khan, S., and Bulyk, M. L.** (2011). Non-
 1252 DNA-binding cofactors enhance DNA-binding specificity of a transcriptional
 1253 regulatory complex. *Molecular Systems Biology* **7**:1–14.

- 1254 **Sijacic, P., Bajic, M., McKinney, E. C., Meagher, R. B., and Deal, R. B.** (2018).
 1255 Changes in chromatin accessibility between Arabidopsis stem cells and
 1256 mesophyll cells illuminate cell type-specific transcription factor networks. *Plant*
 1257 *Journal* **94**:215–231.
- 1258 **Skene, P. J., and Henikoff, S.** (2017). An efficient targeted nuclease strategy for
 1259 high-resolution mapping of DNA binding sites. *eLife* **6**:1–35.
- 1260 **Skene, J. P., and Henikoff S** (2018). CUT& RUN: Targeted in situ genome-wide
 1261 profiling with high efficiency for low cell numbers. *Nature Protocols* **13**:1–28.
- 1262 **Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs,**
 1263 **R., Honig, B., Bussemaker, H. J., et al.** (2011). Cofactor binding evokes
 1264 latent differences in DNA binding specificity between hox proteins. *Cell*
 1265 **147**:1270–1282.
- 1266 **Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., and Rohs,**
 1267 **R.** (2014). Absence of a simple code: How transcription factors read the
 1268 genome. *Trends in Biochemical Sciences* **39**:381–399.
- 1269 **Smaczniak, C., Immink, R. G. H., Muiño, J. M., Blanvillain, R., Busscher, M.,**
 1270 **Busscher-Lange, J., Dinh, Q. D. P., Liu, S., Westphal, A. H., Boeren, S., et**
 1271 **al.** (2012). Characterization of MADS-domain transcription factor complexes in
 1272 Arabidopsis flower development. *Proceedings of the National Academy of*
 1273 *Sciences* **109**:1560–1565.
- 1274 **Smaczniak, C., Muiño, J. M., Chen, D., Angenent, G. C., and Kaufmann, K.**
 1275 (2017). Differences in DNA-binding specificity of floral homeotic protein
 1276 complexes predict organ-specific target genes. *The Plant Cell Advance*
 1277 Access published 2017, doi:10.1105/tpc.17.00145.
- 1278 **Soutourina, J.** (2017). Transcription regulation by the Mediator complex. *Nature*
 1279 *Reviews Molecular Cell Biology Advance Access published 2017,*
 1280 doi:10.1038/nrm.2017.115.
- 1281 **Spitz, F., and Furlong, E. E. M.** (2012). Transcription factors: From enhancer
 1282 binding to developmental control. *Nature Reviews Genetics* **13**:613–626.
- 1283 **Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J.,**
 1284 **Gomez, A., Collombet, S., Berenguer, C., Cuartero, Y., et al.** (2018).
 1285 Transcription factors orchestrate dynamic interplay between genome topology
 1286 and gene regulation during cell reprogramming. *Nature Genetics* **50**:238–249.
- 1287 **Staller, M. V., Holehouse, A. S., Swain-Lenz, D., Das, R. K., Pappu, R. V., and**
 1288 **Cohen, B. A.** (2018). A High-Throughput Mutational Scan of an Intrinsically
 1289 Disordered Acidic Transcriptional Activation Domain. *Cell Systems Advance*
 1290 Access published 2018, doi:10.1016/j.cels.2018.01.015.
- 1291 **Stella, S., Cascio, D., and Johnson, R. C.** (2010). The shape of the DNA minor
 1292 groove directs binding by the DNA-bending protein Fis. *Genes and*
 1293 *Development* **24**:814–826.

- 1294 **Stigliani, A., Martin-Arevalillo, R., Lucas, J., Bessy, A., Vinos-Poyo, T.,**
1295 **Mironova, V., Vernoux, T., Dumas, R., and Parcy, F.** (2018). Capturing
1296 auxin response factors syntax using DNA binding models Advance Access
1297 published 2018.
- 1298 **Stormo, G. D.** (2000). DNA binding sites: representation and discovery.
1299 *Bioinformatics* **16**:16–23.
- 1300 **Stormo, G. D.** (2013). Modeling the specificity of protein-DNA interactions.
1301 *Quantitative Biology* **1**:115–130.
- 1302 **Stormo, G. D., and Zhao, Y.** (2010). Determining the specificity of protein-DNA
1303 interactions. *Nature Reviews Genetics* **11**:751–760.
- 1304 **Stormo, G. D., Schneider, T. D., and Gold, L. M.** (1982). Characterization of
1305 translational initiation sites in *E. coli*. *Nucleic acids research* **10**:2971–2996.
- 1306 **Stroud, H., Otero, S., Desvoyes, B. B., Ramírez-Parra, E., Jacobsen, S. E.,**
1307 **Gutierrez, C., Ramirez-Parra, E., Jacobsen, S. E., and Gutierrez, C.** (2012).
1308 Genome-wide analysis of histone H3. 1 and H3. 3 variants in *Arabidopsis*
1309 *thaliana*. *Proceedings of the National Academy of Sciences of the United*
1310 *States of America* **109**:5370–5.
- 1311 **Sun, B., Looi, L.-S., Guo, S., He, Z., Gan, E.-S., Huang, J., Xu, Y., Wee, W.-Y.,**
1312 **and Ito, T.** (2014). Timing Mechanism Dependent on Cell Division Is Invoked
1313 by Polycomb Eviction in Plant Stem Cells. *Science* **343**:1248559–1248559.
- 1314 **Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M.,**
1315 **Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., et al.** (2008). The
1316 *Arabidopsis* Information Resource (TAIR): Gene structure and function
1317 annotation. *Nucleic Acids Research* **36**:1009–1014.
- 1318 **Takeda, T., Corona, R. I., and Guo, J. T.** (2013). A knowledge-based orientation
1319 potential for transcription factor-DNA docking. *Bioinformatics* **29**:322–330.
- 1320 **Tao, Z., Shen, L., Gu, X., Wang, Y., Yu, H., and He, Y.** (2017). Embryonic
1321 epigenetic reprogramming by a pioneer transcription factor in plants. *Nature*
1322 **551**:124–128.
- 1323 **Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E.,**
1324 **Sheffield, N. C., Stergachis, A. B., Wang, H., Vernet, B., et al.** (2012). The
1325 accessible chromatin landscape of the human genome. *Nature* **489**:75–82.
- 1326 **Tomovic, A., and Oakeley, E. J.** (2007). Position dependencies in transcription
1327 factor binding sites. *Bioinformatics* **23**:933–941.
- 1328 **Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov,**
1329 **A. V., Frith, M. C., Fu, Y., Kent, W. J., et al.** (2005). Assessing computational
1330 tools for the discovery of transcription factor binding sites. *Nature*
1331 *Biotechnology* **23**:137–144.

- 1332 **Trigg, S. A., Garza, R. M., MacWilliams, A., Nery, J. R., Bartlett, A., Castanon, R.,**
1333 **Goubil, A., Feeney, J., O'Malley, R., Huang, S. S. C., et al. (2017).** CrY2H-
1334 seq: A massively multiplexed assay for deep-coverage interactome mapping.
1335 *Nature Methods* **14**:819–825.
- 1336 **Turner, D., Kim, R. G., and Guo, J. tao (2012).** TFinDit: transcription factor-DNA
1337 interaction data depository. *BMC Bioinformatics* **13**.
- 1338 **Vachon, G., Engelhorn, J., and Carles, C. C. (2018).** Interactions between
1339 transcription factors and chromatin regulators in the control of flower
1340 development. *Journal of Experimental Botany* Advance Access published
1341 2018, doi:10.1093/jxb/ery079.
- 1342 **Vera, D. L., Madzima, T. F., Labonne, J. D., Alam, M. P., Hoffman, G. G.,**
1343 **Girimurugan, S. B., Zhang, J., McGinnis, K. M., Dennis, J. H., and Bass,**
1344 **H. W. (2014).** Differential Nuclease Sensitivity Profiling of Chromatin Reveals
1345 Biochemical Footprints Coupled to Gene Expression and Functional DNA
1346 Elements in Maize. *The Plant Cell* **26**:3883–3893.
- 1347 **Viner, C., Johnson, J., Walker, N., Shi, H., Sjöberg, M., Adams, D. J., Ferguson-**
1348 **Smith, A. C., Bailey, T. L., and Hoffman, M. M. (2016).** Modeling methyl-
1349 sensitive transcription factor motifs with an expanded epigenetic alphabet
1350 Advance Access published March 15, 2016, doi:10.1101/043794.
- 1351 **Voss, T. C., and Hager, G. L. (2013).** Dynamic regulation of transcriptional states by
1352 chromatin and transcription factors. *Nature Reviews Genetics* **15**:69–81.
- 1353 **Wang, D., Mills, E. S., and Deal, R. B. (2012).** Technologies for systems-level
1354 analysis of specific cell types in plants. *Plant Science* **197**:21–29.
- 1355 **Wang, H., Liu, C., Cheng, J., Liu, J., Zhang, L., He, C., Shen, W. H., Jin, H., Xu,**
1356 **L., and Zhang, Y. (2016).** Arabidopsis Flower and Embryo Developmental
1357 Genes are Repressed in Seedlings by Different Combinations of Polycomb
1358 Group Proteins in Association with Distinct Sets of Cis-regulatory Elements.
1359 *PLoS Genetics* **12**:1–25.
- 1360 **Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., Ye, Z., Shen, C., Li, J.,**
1361 **Zhang, L., et al. (2017).** Asymmetric subgenome selection and cis-regulatory
1362 divergence during cotton domestication. *Nature Genetics* **49**:579–587.
- 1363 **Wang, M., Tai, C., E, W., and Wei, L. (2018a).** DeFine: deep convolutional neural
1364 networks accurately quantify intensities of transcription factor-DNA binding
1365 and facilitate evaluation of functional non-coding variants. *Nucleic Acids*
1366 *Research* Advance Access published 2018, doi:10.1093/nar/gky215.
- 1367 **Wang, Z., Civelek, M., Miller, C., Sheffield, N., Guertin, M. J., and Zang, C.**
1368 **(2018b).** BART: a transcription factor prediction tool with query gene sets or
1369 epigenomic profiles. *Bioinformatics* Advance Access published 2018,
1370 doi:10.1101/280982.

- 1371 **Wang, M., Wang, P., Lin, M., Ye, Z., Li, G., Tu, L., Shen, C., Li, J., Yang, Q., and**
1372 **Zhang, X.** (2018c). Evolutionary dynamics of 3D genome architecture
1373 following polyploidization in cotton. *Nature Plants* **4**:90–97.
- 1374 **Wasserman, W. W., and Sandelin, A.** (2004). Applied bioinformatics for the
1375 identification of regulatory elements. *Nature Reviews Genetics* **5**:276–287.
- 1376 **Weber, B., Zicola, J., Oka, R., and Stam, M.** (2016). Plant Enhancers: A Call for
1377 Discovery. *Trends in Plant Science* **21**:974–987.
- 1378 **Weintraub, A. S., Li, C. H., Zamudio, A. V, Sigova, A. A., Hanne, N. M., Day, D.**
1379 **S., Abraham, B. J., Cohen, M. A., Nabet, B., Buckley, D. L., et al.** (2018).
1380 YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell Advance*
1381 Access published 2018, doi:10.1016/j.cell.2017.11.008.
- 1382 **Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A.,**
1383 **Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., et al.**
1384 (2014). Determination and Inference of Eukaryotic Transcription Factor
1385 Sequence Specificity. *Cell* **158**:1431–1443.
- 1386 **White, M. A., Myers, C. A., Corbo, J. C., and Cohen, B. A.** (2013). Massively
1387 parallel in vivo enhancer assay reveals that highly local features determine the
1388 cis-regulatory function of ChIP-seq peaks. *Proceedings of the National*
1389 *Academy of Sciences* **110**:11952–11957.
- 1390 **Whiteld, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein,**
1391 **N. D., Myers, R. M., and Weng, Z.** (2012). Functional analysis of transcription
1392 factor binding sites in human promoters. *Genome Biology* **13**:R50.
- 1393 **Wu, H., and Zhang, Y.** (2014). Reversing DNA methylation: Mechanisms, genomics,
1394 and biological functions. *Cell* **156**:45–68.
- 1395 **Wu, M., Sang, Y., Bezhani, S., Yamaguchi, N., Han, S., Li, Z., Su, Y., Slewinski,**
1396 **T. L., and Wagner, D.** (2012). SWI2/SNF2 chromatin remodeling ATPases
1397 overcome polycomb repression and control floral organ identity with the
1398 LEAFY and SEPALLATA3 transcription factors. *Proceedings of the National*
1399 *Academy of Sciences of the United States of America* **109**:3576–3581.
- 1400 **Xiao, J., and Wagner, D.** (2015). Polycomb repression in the regulation of growth
1401 and development in Arabidopsis. *Current Opinion in Plant Biology* **23**:15–24.
- 1402 **Xiao, J., Jin, R., Yu, X., Shen, M., Wagner, J. D., Pai, A., Song, C., Zhuang, M.,**
1403 **Klasfeld, S., He, C., et al.** (2017). Cis and trans determinants of epigenetic
1404 silencing by Polycomb repressive complex 2 in Arabidopsis. *Nature Genetics*
1405 Advance Access published 2017, doi:10.1038/ng.3937.
- 1406 **Xu, B., Schones, D. E., Wang, Y., Liang, H., and Li, G.** (2013). A Structural-Based
1407 Strategy for Recognition of Transcription Factor Binding Sites. *PLoS ONE* **8**.
- 1408 **Yan, J., Enge, M., Whittington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma,**
1409 **A., Kivioja, T., Taipale, M., et al.** (2013). Transcription factor binding in

- 1410 human cells occurs in dense clusters formed around cohesin anchor sites.
1411 *Cell* **154**:801–813.
- 1412 **Yan, W., Chen, D., and Kaufmann, K.** (2016). Efficient multiplex mutagenesis by
1413 RNA-guided Cas9 and its use in the characterization of regulatory elements in
1414 the AGAMOUS gene. *Plant Methods* **12**:1–9.
- 1415 **Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W. W., Gordân, R., and**
1416 **Rohs, R.** (2014). TFBSshape: A motif database for DNA shape features of
1417 transcription factor binding sites. *Nucleic Acids Research* **42**:148–155.
- 1418 **Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R., and Rohs, R.**
1419 (2017). Transcription factor family-specific DNA shape readout revealed by
1420 quantitative specificity models. *Molecular Systems Biology* **13**:910.
- 1421 **Yazaki, J., Galli, M., Kim, A. Y., Nito, K., Aleman, F., Chang, K. N., Carvunis, A.-**
1422 **R., Quan, R., Nguyen, H., Song, L., et al.** (2016). Mapping transcription
1423 factor interactome networks using HaloTag protein arrays. *Proceedings of the*
1424 *National Academy of Sciences* **113**:E4238–E4247.
- 1425 **Ye, L., Wang, B., Zhang, W.-G., Shan, H., and Kong, H.** (2016). Gain of An Auto-
1426 regulatory Site Led to Divergence of the Arabidopsis APETALA1 and
1427 CAULIFLOWER Duplicate Genes in the Time, Space and Level of Expression
1428 and Regulation of One Paralog by the Other. *Plant Physiology*
1429 **171**:pp.00320.2016.
- 1430 **Yin, Y., Sieradzan, A. K., Liwo, A., He, Y., and Scheraga, H. A.** (2015). Physics-
1431 based potentials for coarse-grained modeling of protein-DNA interactions.
1432 *Journal of Chemical Theory and Computation* **11**:1792–1808.
- 1433 **Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S.,**
1434 **Das, P. K., Kivioja, T., Dave, K., Zhong, F., et al.** (2017). Impact of cytosine
1435 methylation on DNA binding specificities of human transcription factors.
1436 *Science* **356**:eaaj2239.
- 1437 **Zaret, K. S.** (2018). Pioneering the chromatin landscape. *Nature Genetics Advance*
1438 Access published 2018, doi:10.1038/s41588-017-0038-z.
- 1439 **Zaret, K. S., and Mango, S. E.** (2016). Pioneer transcription factors, chromatin
1440 dynamics, and cell fate control. *Current Opinion in Genetics & Development*
1441 **37**:76–81.
- 1442 **Zemach, A., and Grafi, G.** (2007). Methyl-CpG-binding domain proteins in plants:
1443 interpreters of DNA methylation. *Trends in Plant Science* **12**:80–85.
- 1444 **Zentner, G. E., Kasinathan, S., Xin, B., Rohs, R., and Henikoff, S.** (2015). ChEC-
1445 seq kinetics discriminates transcription factor binding sites by DNA sequence
1446 and shape in vivo. *Nature Communications* **6**:8733.
- 1447 **Zhang, C., Liu, S., Zhu, Q., and Zhou, Y.** (2005). A knowledge-based energy
1448 function for protein-ligand, protein-protein, and protein-DNA complexes.
1449 *Journal of Medicinal Chemistry* **48**:2325–2335.

1450 **Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W. L., Chen, H.,**
1451 **Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E., et al. (2006).**
1452 **Genome-wide High-Resolution Mapping and Functional Analysis of DNA**
1453 **Methylation in Arabidopsis. *Cell* **126**:1189–1201.**

1454 **Zhang, W., Zhang, T., Wu, Y., and Jiang, J. (2012).** Genome-Wide Identification of
1455 **Regulatory DNA Elements and Protein-Binding Footprints Using Signatures of**
1456 **Open Chromatin in Arabidopsis. *The Plant Cell* **24**:2719–2731.**

1457 **Zhang, L., Martini, G. D., Tomas Rube, H., Kribelbauer, J. F., Rastogi, C.,**
1458 **FitzPatrick, V. D., Houtman, J. C., Bussemaker, H. J., and Pufall, M. A.**
1459 **(2018a). SelexGLM differentiates androgen and glucocorticoid receptor DNA-**
1460 **binding preference over an extended binding site. *Genome Research* **28**:111–**
1461 **121.**

1462 **Zhang, H., Lang, Z., and Zhu, J. (2018b).** Dynamics and function of DNA
1463 **methylation in plants. *Nature Reviews Molecular Cell Biology* Advance Access**
1464 **published 2018, doi:10.1038/s41580-018-0016-z.**

1465 **Zhao, Y., Granas, D., and Stormo, G. D. (2009).** Inferring binding energies from
1466 **selected binding sites. *PLoS Computational Biology* **5**.**

1467 **Zhao, Y., Ruan, S., Pandey, M., and Stormo, G. D. (2012).** Improved models for
1468 **transcription factor binding site identification using nonindependent**
1469 **interactions. *Genetics* **191**:781–790.**

1470 **Zhou, J., and Troyanskaya, O. G. (2015).** Predicting effects of noncoding variants
1471 **with deep learning-based sequence model. *Nature Methods* **12**:931–934.**

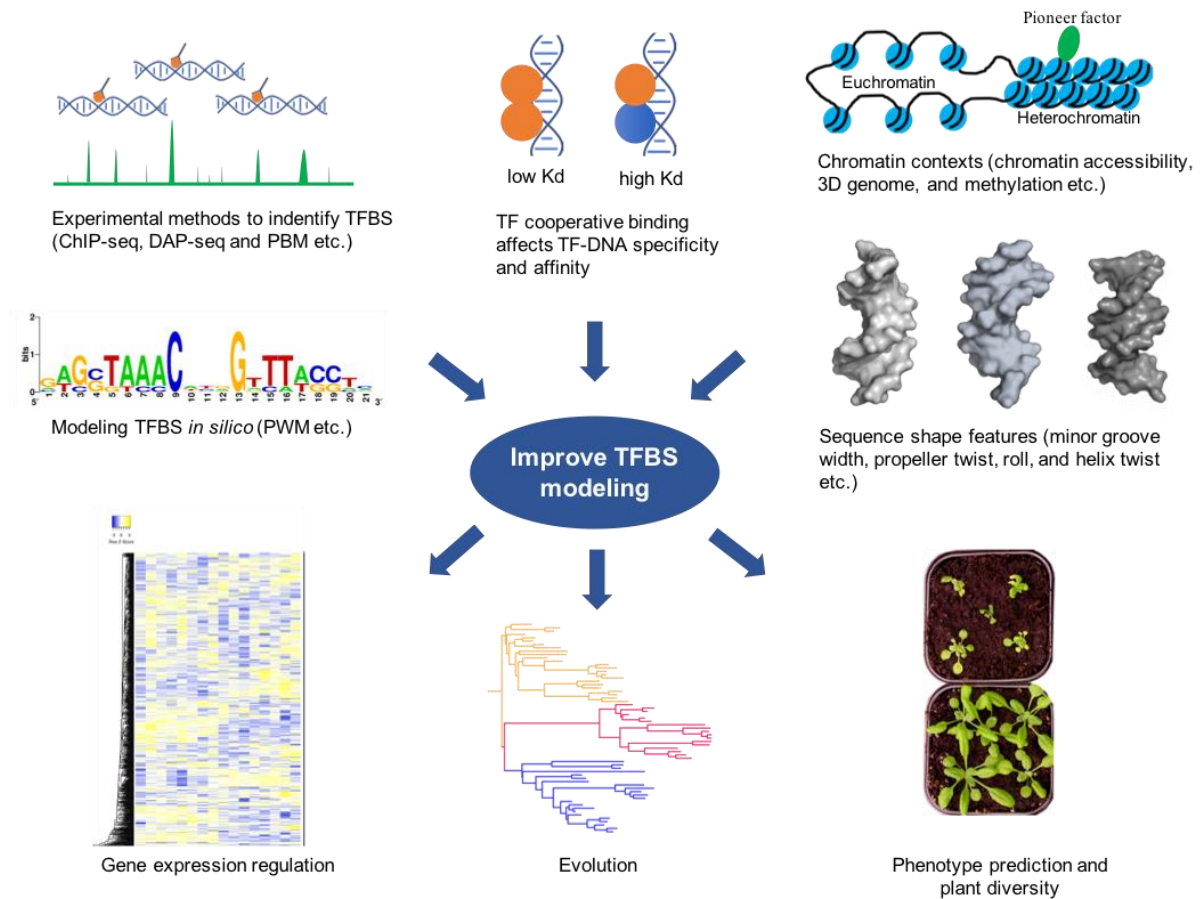
1472 **Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., Bussemaker, H. J.,**
1473 **Gordân, R., and Rohs, R. (2015).** Quantitative modeling of transcription
1474 **factor binding specificities using DNA shape. *Proceedings of the National***
1475 ***Academy of Sciences* **112**:4654–4659.**

1476 **Zhou, Y., Wang, Y., Krause, K., Yang, T., Dongus, J. A., Zhang, Y., and Turck, F.**
1477 **(2018). Telobox motifs recruit CLF / SWN – PRC2 for H3K27me3 deposition**
1478 **via TRB factors in Arabidopsis. *Nature Genetics* Advance Access published**
1479 **2018, doi:10.1038/s41588-018-0109-9.**

1480 **Zhu, H., Wang, G., and Qian, J. (2016).** Transcription factors as readers and
1481 **effectors of DNA methylation. *Nature Reviews Genetics* **17**:551–565.**

1482 **Zuo, Z., Roy, B., Chang, Y. K., Granas, D., and Stormo, G. D. (2017).** Measuring
1483 **quantitative effects of methylation on transcription factor–DNA binding affinity.**
1484 ***Science Advances* **3**:eaao1799.**

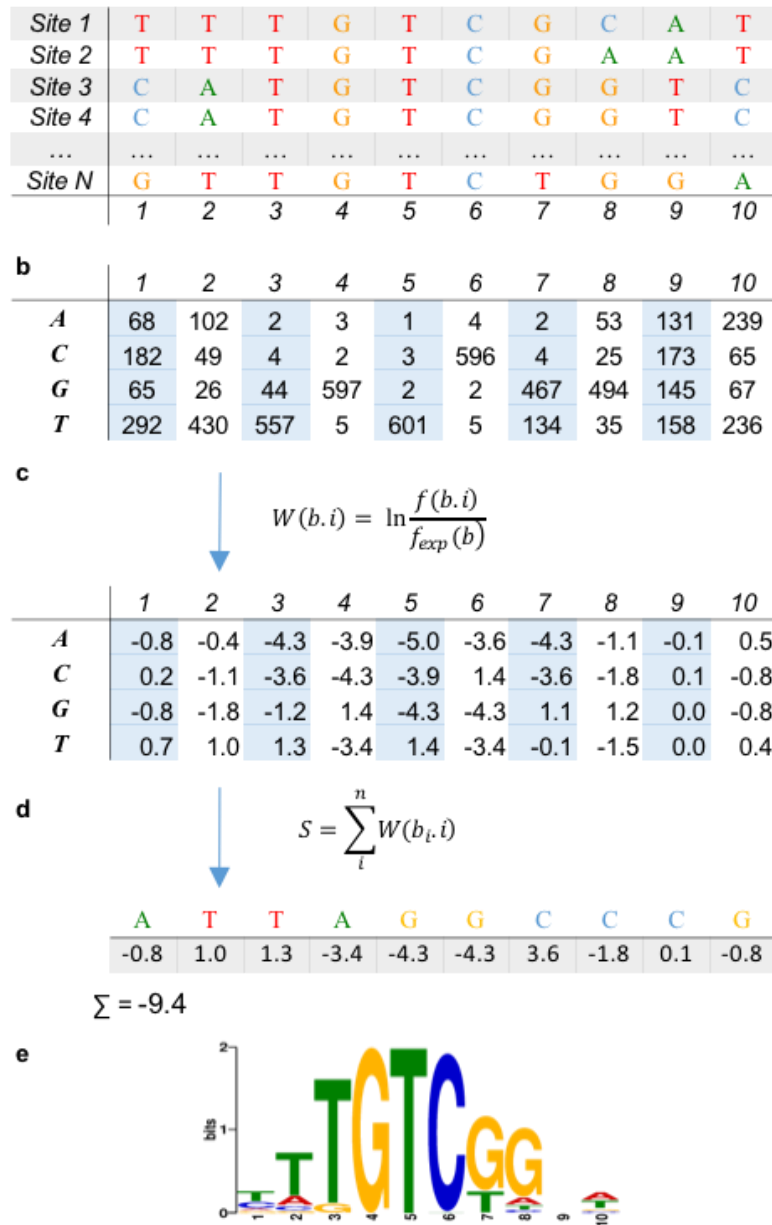
1485



1486

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

1488



1490

1491

1492 **Figure 2. Workflow to generate PWM for a set of known TFBS.** Here, ARF5
 1493 bound sequences are retrieved from DAP-seq (O'Malley et al., 2016). The
 1494 sequences of N different binding sites are aligned (a). The nucleotide frequency is
 1495 computed at each position of the binding site to yield the position frequency matrix
 1496 (b), which is then converted to a PWM (c). In the formula, $W(b, i)$ stands for the
 1497 weight of a nucleotide b at position i , $f(b,i)$ is the frequency of this nucleotide, and f_{exp}
 1498 is the expected background frequency of the given nucleotide. If each nucleotide
 1499 appearance is equal, one can take $f_{exp}=N/4$ (c). One can score a given sequence by
 1500 summing the corresponding PWM weights (d). The TFBS logo represents the
 1501 preference of the TF at each position of the binding site (e). This calculation
 1502 represents one possible method among several to calculate a sequence score
 1503 (Stormo, 2000; Wasserman and Sandelin, 2004; Stormo, 2013).

Table1. Experimental methods to identify TFBS.

Experimental methods	Description	<i>in vivo or in vitro</i>	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.	<i>in vitro</i>	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 <i>in vivo</i>	<i>in vivo</i>	genomic	High-throughput and most widely used protocol for TFBSs mapping <i>in vivo</i>	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.	<i>in vitro</i>	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 <i>in vitro</i> 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.	<i>in vivo</i>	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.	<i>in vitro</i>	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 <i>in vivo</i> .	<i>in vivo</i>	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 <i>in vivo</i> 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.	<i>in vivo</i>	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 epicistrome.	<i>in vitro</i>	genomic	Allows low-cost and high-throughput generation of cistrome and epicistrome maps for hundreds of TFs of an organism	Recombinant or <i>in vitro</i> 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.	<i>in vitro</i>	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 <i>in vitro</i> 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.	<i>in vivo</i>	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.	<i>in vivo</i>	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.	<i>In vitro</i>	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.	<i>In vivo</i>	genomic	ChIP-STARR-seq allows high-throughput identification of functional enhancer <i>in vivo</i> .	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/PI-M	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 <i>de novo</i> 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		dependencies	
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 'dilogos' motifs	(Omidi et al., 2017)
TFImpute	TFImpute predict cell-specific TF binding trained by deep learning.	N.A	https://bitbucket.org/feeldead/tf-impute	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/beem/	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)