Current Plant Biology 3-4 (2015) 40–47

Contents lists available at ScienceDirect

Current Plant Biology

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

DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants

Alessandra M. Sullivan, Kerry L. Bubb, Richard Sandstrom, John A. Stamatoyannopoulos, Christine Queitsch*

Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

ARTICLE INFO

Article history:
Received 18 March 2015
Received in revised form 2 October 2015
Accepted 7 October 2015

Keywords:
DNase I hypersensitivity
Transcription factor (TF) networks
DHSs
Gene regulation
Genomic Footprinting
Chromatin accessibility
Plants

ABSTRACT

Understanding gene regulatory networks in plants requires knowledge of *cis*-regulatory DNA, *trans*-acting factors, and their dynamics across development and in response to stimuli. Active *cis*-regulatory elements are hypersensitive to cleavage by the endonuclease DNase I. Motifs within DNase I hypersensitive sites indicate potential *trans*-acting factor occupancy and, when combined with DNase I cleavage data, can be used to construct provisional regulatory networks. Several recent studies have applied genome-wide DNase I hypersensitivity mapping to *Arabidopsis thaliana* and rice, generating chromatin accessibility landscapes for an array of tissues, cell types, and treatments. Here, we discuss these studies, with an emphasis on building regulatory networks and possible directions for improvements.

1. Introduction

While *cis*-regulatory elements and their dynamics constitute a fundamental part of plant biology, efforts to identify and characterize them at large scale have historically been limited to yeast and animal models. It is only recently that methods for the genome-wide identification of *cis*-regulatory elements have been applied to plants [2–6].

1.1. Detecting cis-regulatory elements by their sensitivity to DNase I

Active *cis*-regulatory elements, including promoters [7], enhancers [8], insulators [9], silencers [10], and locus control regions [11], can be detected by their characteristic hypersensitivity to the endonuclease DNase I [12,13]. DNase I hypersensitive sites (DHSs) can be mapped genome-wide using DNase I-seq. DNase I-seq relies on the treatment of isolated nuclei with DNase I. DNase I preferentially cleaves DNA at "open" or "accessible" chromatin, releasing low molecular weight DNA fragments, which can be purified, sequenced, and mapped back to the genome. Within DHSs, atypical cleavage patterns ("footprints") suggest protein occupancy [14–17]. DNase I-seq and similar, accessibility-based methods [18], have been employed in many organisms, including humans, to (i) delineate the *cis*-regulatory landscape in tissues, cell types, and in response to treatments [1–6,16,17,19–23], (ii) identify potential TF occupancy with genomic footprinting [16,17,24–28], and (iii) to construct TF regulatory networks based on TF occupancy information [2,29,30] (Fig. 1).

1.2. Review objectives

This review covers emerging approaches and topics in plant chromatin biology, including cell-type-specific DNase I-seq profiling and using DNase I cleavage to deduce TF occupancy and construct TF regulatory networks. We discuss the potential and limitations of DNase I-seq and point out differences in methods, sample quality, and nomenclature among recent studies. We highlight promising future directions for the growing number of investigators interested in employing DNase I-seq. For general information on plant chromatin landscapes, see these recent reviews on DNase I-seq in *Arabidopsis thaliana* and rice [31,32].

2. Detecting *cis*-regulatory elements genome-wide in plants

To date, there are four sets of published DNase I-seq data for either *A. thaliana* [2,4–6] or rice [3]. Tissues, cell types, and

* Corresponding author.
E-mail address: queitsch@uw.edu (C. Queitsch).

http://dx.doi.org/10.1016/j.cpb.2015.10.001

Definitions

Chromatin accessibility the characteristic of being accessible to DNA endonucleases, transposons, and DNA-binding proteins.

DNase I an endonuclease that cleaves both single-stranded and double-stranded DNA. In chromatin, the enzyme preferentially cuts in regions that are accessible.

DHS a DNase I hypersensitive site is a region that is characterized by elevated DNase I cleavage. DHSs are considered hallmarks of regulatory DNA, based on their location at transcription start sites, and their overlap with ChIP peaks and known regulatory elements such as enhancers and silencers.

Hotspot a program that uses a sliding-window statistic to identify significantly DNase I-sensitive regions with respect to background levels of DNase I cleavage in the genome. Two types of regions are defined by this program: hotspots and DHSs. DHSs are regions of maximum sensitivity (150 bp peaks) within hotspots [1].

DNase I-seq a technique for measuring regions of DNase I hypersensitivity genome-wide. This technique involves digestion of chromatin within intact nuclei with DNase I. Cleavage products are size-selected and sequenced, such that the 5′-most bp of a sequencing read indicates the site at which cleavage by DNase I occurred. Various peak-calling algorithms are used to identify DHSs.

Genomic footprinting a systematic approach to identify TF-occupied DNA from DNase I-seq data. Algorithms vary, but all rely on the detection of atypical cleavage patterns. Typically, TF-occupied DNA is protected from DNase I cleavage leaving a paucity of cleavages compared to flanking regions.

Footprint a short region of DNA within a DNase I hypersensitive site with a cleavage pattern consistent with protein occupancy. This pattern is classically-defined as a paucity of cleavage because protein-bound DNA is typically protected from DNase I cleavage.

Aggregate plot a plot displaying DNase I cleavage within a window centered on some commonly-occurring entity (*e.g.* instances of a TF motif).

Transcription factor (TF) motif a short (usually 5–20 bp) sequence that is the recognized binding site for a TF. Motifs are often represented as a Position Weight Matrix (PWM), with the frequency of each nucleotide indicated for each position in the motif.

Cis-regulatory element a regulatory element that resides on the same DNA molecule as its target gene. Frequently used to describe a regulatory element in close proximity to its target gene.

Promoter a *cis*-regulatory element that coordinates the initiation of transcription of the adjacent gene.

Enhancer a *cis*-regulatory element that enhances transcription of its target gene. Enhancers may act independently of distance and orientation, and play important roles in development.

Footprint-derived TF network a network displaying a regulatory cascade of TFs by drawing edges between TFs when a footprinted motif of a source TF occurs in the regulatory region or body of a target TF gene.

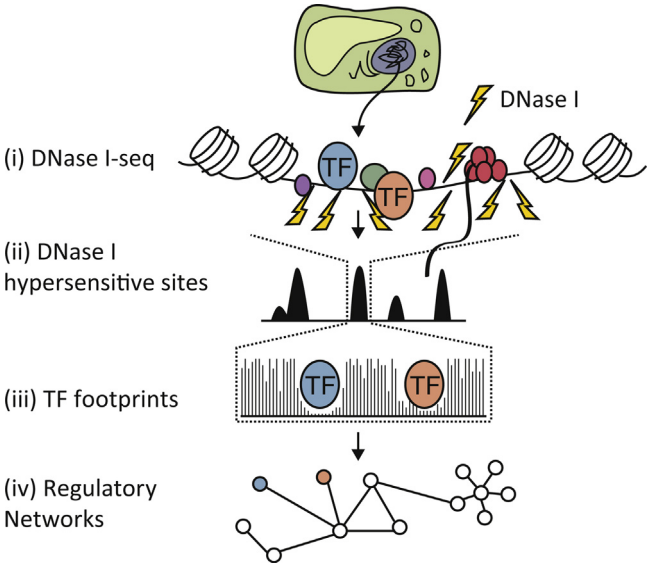

Fig. 1. Schematic of DNase I-seq-derived data.
(i) Nuclei are harvested from plant tissues and treated with the endonuclease DNase I. (ii) Regulatory regions are hypersensitive to cleavage by DNase I. (iii) Protein-bound regions within DNase I hypersensitive sites are protected from DNase I cleavage leaving detectable "footprints." (iv) Footprint and TF motif information can be integrated to generate TF-to-TF regulatory networks.

treatment conditions vary (Table 1). Several unpublished data sets for *A. thaliana* cell types and accessions are also available for browsing and downloading at www.plantregulome.org Below, we summarize briefly the major conclusions of each DNase I-seq study, and their relevance to our discussion on genomic footprinting and TF networks.

The first plant DNase I-seq experiments were conducted in rice seedlings and calli demonstrating that (i) some DHSs are tissue-specific, (ii) DHSs primarily reside in promoters and intergenic regions, (iii) gene expression level and promoter accessibility are associated, and (iv) epigenetic modifications are correlated (H3K27me3) or anti-correlated (DNA methylation) with DHSs [3]. These findings are generally conserved across multicellular eukaryotes [21–23]. In a subsequent paper using this DNase I-seq data, these authors show that DHSs are flanked by strongly phased nucleosomes, consistent with evidence that the binding of regulatory proteins is a determinant of nucleosome positioning [33,34]. To our knowledge, there has been no attempt to determine footprints and build footprint-derived TF networks in rice.

The first *A. thaliana* DNase I-seq paper [5] was authored by the same team that conducted the rice studies. The authors performed DNase I-seq on leaves and flowers of wild type and the methylation mutant *ddm1*. Conclusions similar to the rice study are reported with regard to DHS tissue specificity, genomic distribution, and relationship to DNA methylation. The authors also integrated existing ChIP-seq data (AP1 and SEP3) [35,36] and 63 TF motifs (CCA1 and others) [37] with DNase I accessibility to predict TF occupancy [27]. Cleavage patterns coinciding with TF motifs are displayed in aggregate. This approach revealed paucity of DNase I cleavage for less than half of the 63 analyzed motifs, identifying approximately 20,000 footprinted motifs per sample. The authors also show a differentially accessible MADS box motif in the promoter of the *SUP* gene. Overall ~90% of the ChIP-seq binding sites for AP1 and SEP3 overlap with DHSs, validating DHSs as putative regulatory elements. This high overlap is also consistent with previous data on DHSs and ChIP-seq in humans [24].

The subsequent *A. thaliana* DNase I-seq paper, authored by a different team [4], integrates ChIP-seq, expression, and DNase I-seq data to interrogate the regulatory landscape during flower develop-

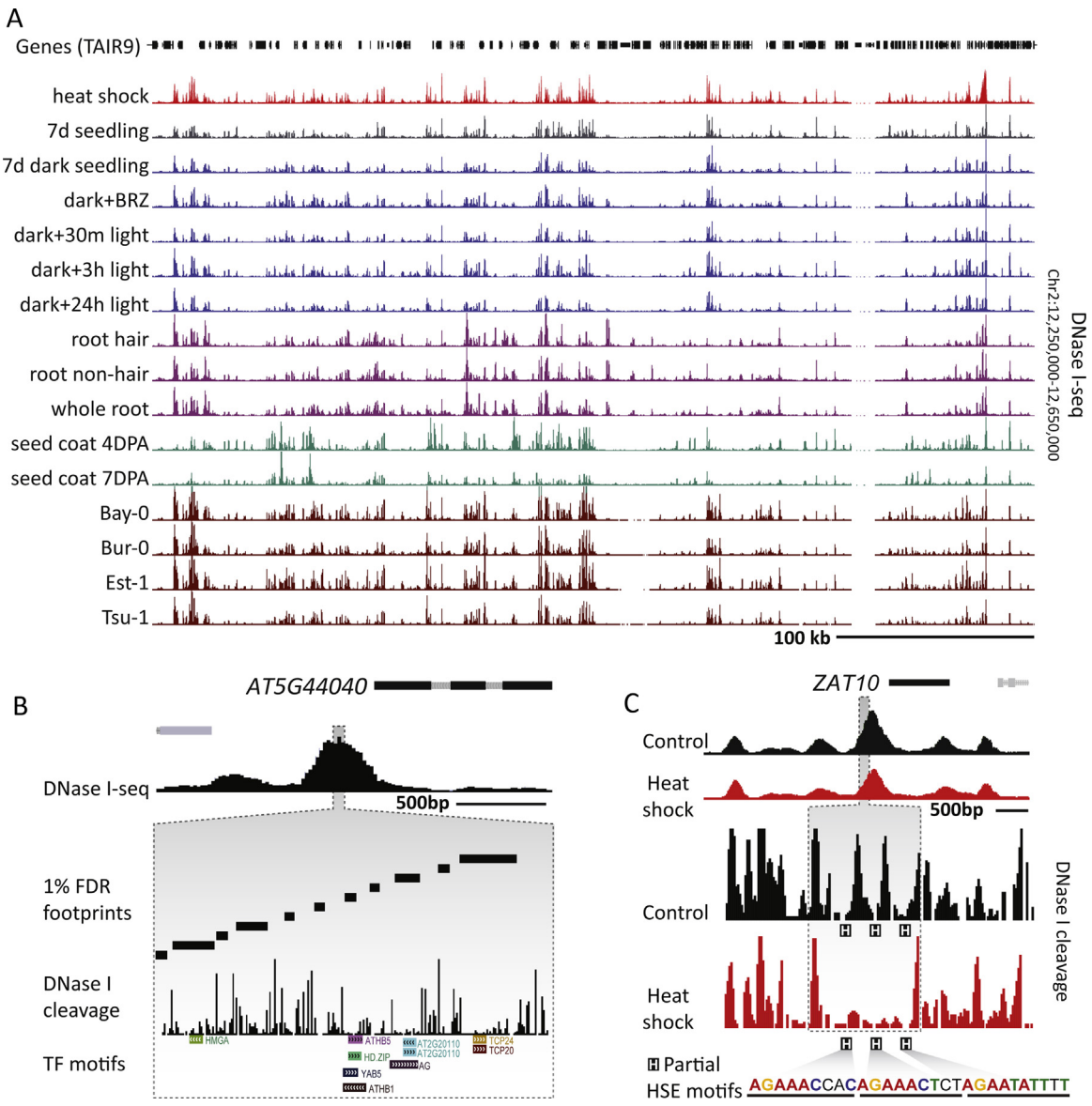

Fig. 2. DHS landscapes and genomic footprinting.
(A) Browser view of a large genomic region of DNase I-seq data in seedlings, tissues, cell types, and *A. thaliana* accessions taken from www.plantregulome.org. This region displays different patterns of chromatin accessibility across the sample types. Many DNase I hypersensitive regions are also shared among samples. (B) A DNase I hypersensitive site in 7 day old light grown seedling in the promoter region of *AT5G44040*, a gene of unknown function. Footprints have been systematically detected and TF motifs from TRANSFAC [83] and a recent protein binding microarray study [39] are shown. (C) A differentially accessible region in the promoter of *ZAT10* during heat shock (red). Partial heat shock elements (HSEs) are observed within the differentially accessible region. Control sample is shown in black.

ment. Though the number of DHSs identified in this study is small, less than 10,000 per sample, the authors identify 1370 developmentally regulated DHSs. Overlap of ChIP-seq binding sites for AP1 and SEP3 and DHSs is weak early in development (day two after flowering induction, 27% and 32% overlap respectively), but reaches approximately 75% by day eight. The detection of ChIP-seq binding sites in the absence of chromatin accessibility suggests AP1 and SEP3 may act as pioneer TFs during flower development. Alternatively, the low levels of ChIP-seq and DHS overlap may be explained by the author's highly stringent approach to calling DHSs in these samples. TF occupancy for three CArG-box motifs is predicted by CENTIPEDE [27]. On average, occupancy increases from day two to day eight for instances of the CArG-box 2 motif that is predominantly occupied by SEP3, but not for the CArG-box 1 motif that is occupied by both SEP3 and AP1, implicating SEP3 complexes lacking AP1 in later flower development. TF networks are also reported; however they are created from ChIP-seq and expression data rather than DNase I-seq data.

In our recent *A. thaliana* DNase I-seq paper, we adapt methods that were previously developed for the human ENCODE project to identify heat, light, and developmentally dynamic DHSs [2]. In a single *A. thaliana* seedling sample, we identify ~34,000 DHSs and delineate ~700,000 footprints at nucleotide resolution (Fig. 2A–C). As previously reported [21–24], we find that ChIP-seq peaks and DHSs largely overlap: over 90% of reproducible PIF3ChIP-seq peaks [38] overlap a DHS in dark-grown seedlings. When we combine all available regulatory information – reproducible ChIP-seq peaks, DHSs and footprints – the PIF3 motif density increases about 400-fold over that found genome-wide and about 4-fold over motif density in reproducible ChIP-seq peaks. This result suggests that DNaseI-seq is a valuable orthogonal method to facilitate TF motif discovery.

Table 1
Recent plant DNase I-seq studies. Published DNase I-seq samples are listed. Additional unpublished samples, including *A. thaliana* accessions and seed coat cell samples are available for browsing at www.plantregulome.org.

Background	Tissue	Conditions	Replicates	Publication
Rice "Nipponbare"	Leaf, stem	2 week old seedling	3	[3]
Rice "Nipponbare"	Callus	3 week old calli	2	[3]
Col-0; *ddm*-1-2	Flower	LD	1	[5]
Col-0; *ddm*-1-2	Leaf	2 week old seedling, LD	2	[5]
Col-0	Flower	LD	2	[5]
Col-0	Leaf	2 week old seedling, LD	2	[5]
Landsberg; pAP1:AP1:GR ap1 cal	Inflorescence	no flowering induction	2	[4]
Landsberg; pAP1:AP1:GR ap1 cal	Inflorescence	2 day flowering induction	2	[4]
Landsberg; pAP1:AP1:GR ap1 cal	Inflorescence	4 day flowering induction	2	[4]
Landsberg; pAP1:AP1:GR ap1 cal	Inflorescence	8 day flowering induction	2	[4]
Col-0; UBQ10:NTF::ATC2:BirA	Whole seedling	7 day old dark + 24 h LD	2	[2]
Col-0; UBQ10:NTF::ATC2:BirA	Whole seedling	7 day old dark + 30 min light	2	[2]
Col-0; UBQ10:NTF::ATC2:BirA	Whole seedling	7 day old dark + 3hr light	3	[2]
Col-0; UBQ10:NTF::ATC2:BirA	Whole seedling	7 day old, Dark	2	[2]
Col-0; UBQ10:NTF::ATC2:BirA	Whole seedling	7 day old, LD	1	[2]
Col-0; UBQ10:NTF::ATC2:BirA	Whole seedling	7 day old, LD, 30 min 45 °C HS	2	[2]
Col-0; UBQ10:NTF::ATC2:BirA	Whole seedling	7 day old, LD, spray control	4	[2]
Col-0; ADF8:NTF::ACT2:BirA	Root hair cells	7 day old, LD	3	[2]
Col-0; GL2:NTF::ACT2:BirA	Root-non hair cells	7 day old, LD	1	[2]
Col-0; UBQ10:NTF::ATC2:BirA	Root maturation zone	7 day old, LD	1	[2]

TF footprints encode an extensive *cis*-regulatory lexicon of 636 *de novo* motifs. Ninety six percent of motif models derived from protein-binding microarrays [39] are close matches to at least one of our 636 *de novo* footprint-derived motifs. The nucleotide diversity of the remaining ~100 novel footprint-derived motifs is indistinguishable from known TF motifs, consistent with recent purifying selection and hence functional importance. Leveraging our ability to call footprints, we find evidence for widespread TF binding within exons. The preferentially-protected codons differ between *A. thaliana* and human; these differences correlate with the known codon bias in both organisms, suggesting that TF binding may have shaped codon usage. Using footprints, we build TF networks and interrogate their dynamics in response to heat shock and light. We also compare the topology of the *A. thaliana* TF regulatory network to that of animals and reveal that they are strikingly similar. Many novel aspects of this study critically rely on systematically identified individual footprints, which are based on DNase I cleavage alone.

3. Nomenclature, methods, and data quality

3.1. Multiple methods for determining DHSs

Each of the plant DNase I-seq studies use a different method to call DHSs. Therefore, it is not surprising that the numbers of DHSs identified varies considerably between studies. In *A. thaliana*, Zhang et al. [5] identify between 38,000 and 42,000 DHSs per sample using F-seq [http://fureylab.web.unc.edu/software/fseq/] [40], Pajoro et al. [4] identify between 5000 and 9000 DHSs (a conservative set that overlaps with DHSs found by Zhang and co-authors [5]) using MACS 2.0.10 [http://liulab.dfci.harvard.edu/MACS/] [41], and Sullivan et al. [2] identify between 25,000 and 35,000 DHSs using Hotspot [http://www.uwencode.org/proj/hotspot/] [1]. In addition to DHS-calling method, both read depth and sample quality affect the number of identifiable DHSs. We use a quality metric that calculates the proportion of reads that fall into hotspots [1]; the Signal Portion of Tags, or SPOT. SPOT is a measure of signal over background noise. SPOT scores for each of the published *A. thaliana* data sets are in Fig. 3. We found it imperative to normalize for both read depth and sample quality (SPOT scores) before quantitative DHS comparisons across samples. This normalization may be particularly important for comparing existing and future data from different laboratories.

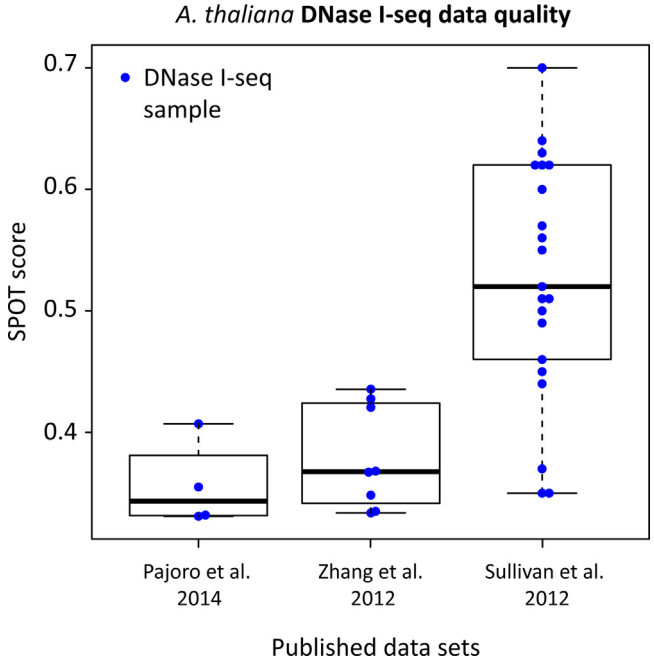

Fig. 3. *A. thaliana* data quality.
The signal portion of tags, or SPOT score, was calculated for each published DNase I-seq library by subsampling 5000,000 reads and calculating the proportion that fall in hotspots [http://www.uwencode.org/proj/hotspot/].

3.2. Multiple methods of determining TF occupancy-what is a DNase I footprint?

As the plant chromatin field evolves rapidly, new methods and terms are being introduced at a rapid pace, leading to some confusion about nomenclature. A good example is the usage of the term "footprint". The term DNase I footprint derives from classic *in vitro* studies, in which investigators added purified TF protein to labeled DNA containing a putative TF binding site. Increasing amounts of DNA-binding protein were used to identify short, protected stretches of nucleotides; evidence of protection was equated with protein occupancy. This classic definition of protection-based TF occupancy is the basis of the footprint-detection algorithm [24] for genome-wide data that we and others applied [2,16,17,24,26] (Fig. 2B–C). This footprint-detection algorithm relies on DNase I

cleavage signal alone, and is agnostic to the presence of underlying motifs. The value of this approach is two-fold: (i) it allows deriving *de novo* TF motifs which is particularly important in non-model plant species and crops, and (ii) it allows deriving hypotheses on the function of individual regulatory elements for future experimental exploration.

The other two *A. thaliana* DNase I-seq studies [4,5] use CENTIPEDE [27] to infer TF occupancy at motifs. The CENTIPEDE algorithm integrates several inputs such as histone modifications, proximity to TSS, evolutionary conservation as well as DNAse I cleavage data to make predictions as to whether a given motif is occupied. The authors of CENTIPEDE [27] report that for some TFs footprint data improved occupancy predictions, for others the presence of DHSs sufficed. Another study reports that DNase I cleavage data – *i.e.*, DHSs – hold the majority of predictive power for occupancy compared to various histone modifications, even if those are combined [42].

Currently, efforts are being made to determine whether an alternative definition of footprints – based on a cleavage pattern that differs from that found with naked DNA – can improve binding predictions. Defining footprints exclusively based on protection from DNase I cleavage appears too narrow as TFs likely differ in binding affinity, occupancy time, and ability to bend DNA which may produce enhanced accessibility rather than protection. Comparing new analysis methods with data from experimental disruption of footprinted motifs [43] is the most promising way forward to determine TF occupancy. Although such experiments are by necessity low-throughput, time and labor intensive, they are essential for determining the rules of gene regulation [43].

3.3. Multiple methods of nuclei isolation and treatment

Each of the plant DNase I-seq studies follows a different protocol for nuclei isolation and DNase I treatment. The protocol by Zhang et al. is the most straight-forward and likely the fastest (the authors kindly provide their protocol upon request). Pajoro et al. isolate nuclei using a Percoll gradient. Recently, a modified, faster protocol for DNase I seq has been described [6]. We use a more complex and time-consuming protocol that involves streptavidin-capture of biotinylated nuclei [2]. The major drawback of our method is that it requires transgenic INTACT lines [44,45]. The major advantage of INTACT lines is that they enable cell-and tissue-type-specific capture of nuclei or even TF-specific nuclei capture. Several INTACT lines are already available through ABRC (CS68649; CS68650; CS68651; CS68652; CS68653; CS68654; CS68655; CS68656), more could be easily generated in a community-wide effort. It is noteworthy, that in our hands the INTACT-based method produces high quality data, even for cell-type specific capture (Fig. 2A, 3). This may, however, not apply to all cell types, especially very rare ones; optimization of our protocol will be required in such instances.

4. Footprint-derived TF networks

DNase I mapping and genomic footprinting [16] allow the generation of provisional TF-to-TF regulatory networks. For the generation of TF networks, footprinted TF motifs within the regulatory region and gene body of a target TF are used to define regulatory edges between TFs (Fig. 4A). By systematically applying this approach to all TFs for which high quality motifs were available ($n = 251$), we map TF-to-TF network dynamics in response to light and heat, and analyze the higher-order architecture of the network. TFs with a high number of input and output edges, or degree, are thought to be functionally important for a given condition. Large gains in a TF's regulatory edges across conditions imply a key role in network re-wiring; losses imply decreased functional importance. This approach has been previously employed in 41 human cell and tissue types, revealing dramatic differences in network wiring across different cell types [29].

However, these analyses have several caveats. For example, this approach of building TF networks does not distinguish between positive and negative regulation or TF identity when TF motifs are highly similar. As the occurrence of several TF motifs within a single footprint implies high regulatory potential (*i.e.*, multiple TF may occupy this regulatory element across time and space), all possible regulatory connections are reported (Fig. 4A). Footprint-derived TF networks could be improved with complementary expression and ChIP-seq data. For example, output-edges from TFs that are not expressed in a particular condition or cell type could be removed from networks; ChIP-seq experiments could identify the specific TF occupying a footprint with multiple motifs. Increasing the collection of high-quality TF motifs in plants will also improve footprint-derived TF networks.

4.1. Higher order network architecture is similar in plants and animals

Information processing networks, including TF networks, in uni- and multicellular organisms have characteristic patterns of simple network motifs (Fig. 4B) [46–48]. Because these network motifs are universal, their frequencies can be used to compare the architecture of diverse networks [46–48]. We find that the *A. thaliana* TF network topology converges on the established *Caenorhabditis elegans* neuronal network topology, and on the TF network topology of all previously analyzed multicellular organisms, including flies, mouse, sea urchin, and human [2,29,30,46]. Like these previously analyzed animals, *A. thaliana* shows evidence of non-rate-limited information processing [2]. This conservation is found despite the vast divergence in non-coding genomic regions and TFs between plants and animals. Moreover, plants are acutely sensitive to environmental stimuli. Other sessile organisms with similarly acute environmental responses such as the unicellular eukaryote yeast and prokaryotic bacteria show rate-limited, sensory networks for quick responses to transient environmental signals. Thus, multicellularity and hence development rather than environmental responsiveness is reflected in the network topology of *A. thaliana.* As the ancestors of plants and animals were unicellular, our results are consistent with convergent multicellularity as a major driver of information processing architecture.

5. Cell-type-specific DNase I-seq

To explore cell-type-specific regulatory landscapes, we employed cell-type-specific nuclear capture (INTACT) [44,45] followed by DNase I-seq [2]. We used existing INTACT lines [44,45] to capture chromatin landscapes of root hair, root non-hair, and seed coat epidermal cells. We found evidence of cell-type specific DHSs and footprints among these three epidermal cell-types (Fig. 2A) (manuscript in review). Root hair and non-hair cell regulatory landscapes were distinct from the regulatory landscape of whole roots and seed coat cells; however, they were very similar to each other. This similarity is not surprising; both cell types are closely related, terminally differentiated root epidermis cell types arising from a common progenitor cell type. To understand the different developmental trajectories of both cell types at the level of chromatin, a better strategy may have been to compare progenitor cells to the differentiated cells. Indeed, probing the regulatory landscape of seed coat cells at different developmental stages revealed many dynamic DHSs (manuscript in review). Studies in human demonstrate that cell-type specific DHS landscapes hold rich information about developmental fate and cell lineages

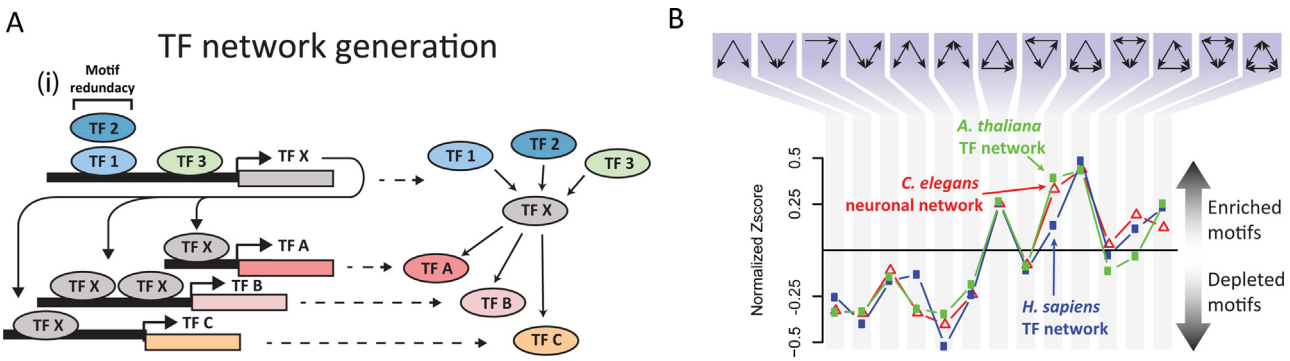

Fig. 4. TF networks in *A. thaliana*.
(A) TF-to-TF network edges are generated when a footprinted motif of a TF overlaps a target TF gene, including 500 bp upstream of the transcription start site. Motif redundancy and its consequence on the network is also illustrated (i) (B) *A. thaliana* network architecture (green), as described by the frequency of TF trios relative to a shuffled version of the network, converges on the architecture of other developmental networks, including human TF networks (blue) and *C. elegans* neuronal networks (red). The *x*-axis lists (unordered) the 13 possible types of TF trios (three-node architectural "network motifs") in a regulatory network. The *y*-axis is the relative enrichment or depletion of those 13 possible "network motifs" within our regulatory network as compared to a network in which both the set of source nodes is shuffled and the set of sink nodes is shuffled, thus preserving the degree distribution of the nodes (*z*-score). Network motifs are displayed above the graph.

such that related cell types have more similar DHS landscapes [49]. Although plant cell lineages are perhaps less well defined than animal cell lineages [50], we find that DHSs are important markers of lineage and cell fate in plants (manuscript in review). Of the over 55,000 DHSs we have identified across several samples to date www.plantregulome.org, approximately 5000 are unique to seed coat cells, suggesting that cell-type-specific DHS profiling adds substantial information and presumably identifies the most relevant regulatory elements for a particular developmental process. Chromatin accessibility mapping may even give us clues about why plants display increased plasticity in cell fate, including their ability to generate totipotent callus tissue from adult tissues with relative ease [51].

6. Conclusions

The collections of plant DNase I-seq data (atlas of *cis*-elements, dynamic elements, TF footprints, and TF networks) generated by recent studies are an essential resource for the plant research community. Taken together, the available data support the notion that plants have complex, dynamic, and developmentally regulated patterns of *cis*-regulatory elements similar to animals. In contrast, in the unicellular yeast, few distal regulatory elements exist and promoters tend to follow a more simple pattern of upstream activating or silencing elements [52]. This similarity to multicellular animals is perhaps not surprising given that plants are multicellular and face complex developmental decisions which require regulatory complexity.

DNase I-seq, genomic footprinting, and footprint-derived regulatory networks are powerful tools to generate hypotheses on regulatory relationships, gene function and promising candidate polymorphisms for functional studies. However, there are limits to what DNase I-seq can reveal about the underlying biology. As we show [2], many DHSs are static, even in response to drastic environmental change (Fig. 2A). Promoters involved in development and environmental response are often poised, resulting in DNase I accessibility before and after stimulus. TF motifs are redundant as many TF families are highly expanded in plants and motifs are short. As we discuss in more detail below, assignment of regulatory elements to target genes is not trivial. Like all genome-wide analyses, DNase I-seq and genomic footprinting are subject to experimental bias; these have been reviewed previously [53]. Nevertheless, DNase I-seq is a very powerful complementary approach to expression analysis, ChIP-seq, and most importantly functional analysis [53]. In particular, footprint-derived, ChIP-seq-supported regulatory networks are ripe for functional exploration with genetic and biochemical approaches.

7. Future directions

7.1. Assigning function to individual DHSs

While DNase I-seq in *A. thaliana* and rice improved our understanding of plant gene regulation, we are still missing half the picture: there is no easy way of connecting individual putative regulatory elements, especially distal ones, to the genes they regulate. Our current analytical methods assume regulatory elements are near their target genes. This assumption is passable because (i) more than a third of DHSs fall within 400 bp of the transcription start site suggesting they are part of the proximal promoter region [2,5], (ii) genes nearest to *dynamic* DHSs implicate known biological processes and/or are differentially expressed [2–5], and (iii) the organization of *cis*-regulatory elements in the similarly compact genome of *Drosophila* indicates that the greatest proportions of developmental enhancers are gene-proximal or intragenic [54]. However, assignments by proximity can fail. For example, in vertebrate development, the enhancer controlling *Shh* expression in the developing limb bud is located in the intron of another gene *Lmbr1* [55] which resides more than a megabase from *Shh* [56]. Thus far, very few long-range regulatory elements have been identified in plants, including the *tb1* enhancer in maize, which acts at a distance greater than ~60 kb [57], and the *b1* enhancer in maize, ~100 kb upstream of the transcription start site [58]. Such long-range elements have not been identified in the smaller genomes of rice or *A. thaliana*. The average distance between genes in *A. thaliana* is ~1 kb, therefore, even at a distance of 1 or 2 kb from the TSS, proximity assignments can be problematic.

7.2. Strategies for connecting regulatory elements to their target genes

Connecting regulatory elements to their target genes is a difficult problem. Reporter gene assays testing the regulatory potential of a given DHSs do not identify the endogenous target of the candidate regulatory element. Labor-intensive strategies have been successfully employed in other systems to link the expression patterns of reporter assays to the *in situ* hybridizations patterns of endogenous genes [54,59,60]. To assign function, classical genetics uses the experimental paradigm of disrupting genes or other genetic elements and assaying molecular and morphological conse-

quences. Using the same logic, perturbing DHSs (and the footprints inside them) using CRISPR/Cas9 [61–64] or other genome editing techniques, will reveal their regulatory function [43]. A major advantage of this targeted strategy is that candidate regulatory elements are studied in their native chromatin context rather than on a transiently expressed plasmid or in newly integrated transgene.

An alternative approach that does not require reporter assays, targeted perturbations, or *in situ* hybridizations are chromatin conformation capture (3-C)-based assays such as Hi-C and ChIA-PET [65,66]. 3-C based-assays can capture physical interactions between distal elements and promoters. 3-C has been used to measure the long-range (40–60 kb) interactions between enhancer elements of the locus control region and the promoters of active globin genes [67]. However, genome-wide 3-C-based technologies are challenging, do not distinguish between functional and non-functional interactions, do not capture dynamics, are muddled by the heterogeneity in chromatin architecture among cells, and with the exception of ChIA-PET cannot identify the proteins responsible for the observed association [66]. Despite these limitations, large-scale chromatin looping between enhancers and promoters has been demonstrated in human cell lines with high resolution Hi-C [68]. Recent Hi-C plant studies [69–71] approach this high resolution and show the enormous potential of this method.

7.3. Mapping variation in DHSs to phenotype

Modifications of gene expression patterns (*i.e.*, regulatory changes) are thought to be a major driving force for evolution of morphological diversity across living organisms [72,73]. Extensive genetic and phenotypic variation exists among *A. thaliana* accessions [74,75] and *cis*-variation associated with expression change are common [76–78]. In both human [79] and *A. thaliana* [2], disease and trait-associated variants are enriched in DHSs. To characterize variation in the regulatory landscape of *A. thaliana,* we have performed DNase I-seq of five geographically and genetically diverse *A. thaliana* accessions: Bay-0, Bur-0, Est-1, Tsu-1 and Col-0. These data are available at www.plantregulome.org and a full analysis is forthcoming. Even quick glance through the browser, reveals that thousands of DHSs are variable among the accessions. This variability may arise through accession-specific structural variation or regulatory variation; nevertheless, it's likely to be of functional consequence.

7.4. systematic PlantENCODE effort?

The human and model animal communities mustered large resources and extensive collaborative efforts (ENCODE [80] and ModENCODE [81,82] to identify regulatory DNA, regulatory networks, chromatin modifications and other features, and integrate this knowledge in a series of recent and forthcoming publications. These discoveries have critically advanced our knowledge of human and animal biology; more importantly, they have provided vast resources for future studies and propelled humans to the status of the new tractable model organism of choice. The centralized resources of ENCODE and ModENCODE enable smaller, less-well funded laboratories to conduct cutting-edge hypothesis-driven research without the need to conduct these expensive genome-scale experiments. The absence of a similar collaborative and sustained effort from both the plant research community and the relevant funding agencies puts plant research at a profound disadvantage, especially in the current funding climate. Plant biology in general will be impeded as young scientists will enter fields with deeper resources.

Acknowledgements

This work was supported by grants from the National Science Foundation (MCB1243627; J.A.S., C.Q., and J.L.N.), Graduate Research Fellowship (DGE-0718124; to A.M.S.). We thank Jeff Vierstra and Roger Deal for thoughtful discussions.

References

[1] S. John, et al., Chromatin accessibility pre-determines glucocorticoid receptor binding patterns, Nat. Genet. 43 (3) (2011) 264–268.
[2] A.M. Sullivan, et al., Mapping and dynamics of regulatory dna and transcription factor networks in *A. thaliana*, Cell Rep. (2014).
[3] W. Zhang, et al., High-resolution mapping of open chromatin in the rice genome, Genome Res. 22 (1) (2012) 151–162.
[4] A. Pajoro, et al., Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development, Genome Biol. 15 (3) (2014) R41.
[5] W. Zhang, et al., Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in *Arabidopsis*, Plant Cell 24 (7) (2012) 2719–2731.
[6] J.S. Cumbie, S.A. Filichkin, M. Megraw, Improved DNase-seq protocol facilitates high resolution mapping of DNase I hypersensitive sites in roots in Arabidopsis thaliana, Plant Methods 11 (2015) 42.
[7] C. Wu, The 5′ ends of *Drosophila* heat-shock genes in chromatin are hypersensitive to Dnase-I, Nature 286 (5776) (1980) 854–860.
[8] A. Banerji, L. Olson, W. Schaffner, A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes, Cell 33 (3) (1983) 729–740.
[9] J.H. Chung, A.C. Bell, G. Felsenfeld, Characterization of the chicken beta-globin insulator, Proc. Natl. Acad. Sci. U. S. A. 94 (2) (1997) 575–580.
[10] M. Antoniou, et al., The human beta-globin gene contains multiple regulatory regions—identification of one promoter and 2 downstream enhancers, EMBO J. 7 (2) (1988) 377–384.
[11] D. Talbot, et al., A dominant control region from the human beta-globin locus conferring integration site-independent gene-expression, Nature 338 (6213) (1989) 352–355.
[12] C. Wu, et al., The chromatin structure of specific genes. I. Evidence for higher order domains of defined DNA sequence, Cell 16 (4) (1979) 797–806.
[13] C. Wu, Y.-C. Wong, S.C.R. Elgin, The chromatin structure of specific genes. II. Disruption of chromatin structure during gene activity, Cell 16 (4) (1979) 807–814.
[14] C. Wu, Two protein-binding sites in chromatin implicated in the activation of heat-shock genes, Nature 309 (5965) (1984) 229–234.
[15] D.J. Galas, A. Schmitz, DNAse footprinting: a simple method for the detection of protein-DNA binding specificity, Nucleic Acids Res. 5 (9) (1978) 3157–3170.
[16] S. Neph, et al., An expansive human regulatory lexicon encoded in transcription factor footprints, Nature 489 (7414) (2012) 83–90.
[17] J.R. Hesselberth, et al., Global mapping of protein–DNA interactions in vivo by digital genomic footprinting, Nat. Methods 6 (4) (2009) 283–289.
[18] J.D. Buenrostro, et al., ATAC-seq: A method for assaying chromatin accessibility genome-wide, Curr. Protoc. Mol. Biol. 109 (21–29) (2015) 1–9.
[19] P.V. Kharchenko, et al., Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*, Nature 471 (7339) (2011) 480–485.
[20] J.A. Stamatoyannopoulos, et al., An encyclopedia of mouse DNA elements (Mouse ENCODE), Genome Biol. 13 (8) (2012).
[21] S. Thomas, et al., Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development, Genome Biol. 12 (5) (2011) R43.
[22] R.E. Thurman, et al., The accessible chromatin landscape of the human genome, Nature 489 (7414) (2012) 75–82.
[23] F. Yue, et al., A comparative encyclopedia of DNA elements in the mouse genome, Nature 515 (7527) (2014) 355–364.
[24] S. Neph, et al., BEDOPS: high-performance genomic feature operations, Bioinformatics 28 (14) (2012) 1919–1920.
[25] R.M. Samstein, et al., Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification, Cell 151 (1) (2012) 153–166.
[26] J. Vierstra, et al., Mouse regulatory DNA landscapes reveal global principles of *cis*-regulatory evolution, Science 346 (6212) (2014) 1007–1012.
[27] R. Pique-Regi, et al., Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data, Genome Res. 21 (3) (2011) 447–455.
[28] J. Piper, et al., Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data, Nucleic Acids Res. 41 (21) (2013) e201.
[29] S. Neph, et al., Circuitry and dynamics of human transcription factor regulatory networks, Cell 150 (6) (2012) 1274–1286.
[30] A.B. Stergachis, et al., Conservation of trans-acting circuitry during mammalian regulatory evolution, Nature 515 (7527) (2014) 365–370.
[31] W.L. Zhang, et al., Open chromatin in plant genomes, Cytogenet. Genome Res. 143 (1–3) (2014) 18–27.
[32] J. Jiang, The 'dark matter' in the plant genomes: non-coding and unannotated DNA sequences associated with open chromatin, Curr. Opin. Plant Biol. 24C (2015) 17–23.

[33] Y. Wu, W. Zhang, J. Jiang, Genome-wide nucleosome positioning is orchestrated by genomic regions associated with DNase I hypersensitivity in rice, PLoS Genet. 10 (5) (2014) e1004378.
[34] J. Vierstra, et al., Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH, Nat. Methods 11 (1) (2014) 66–72.
[35] K. Kaufmann, et al., Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the *Arabidopsis* flower, PLoS Biol. 7 (4) (2009) e1000090.
[36] K. Kaufmann, et al., Orchestration of floral initiation by APETALA1, Science 328 (5974) (2010) 85–89.
[37] R.V. Davuluri, et al., AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis cis*-regulatory elements and transcription factors, BMC Bioinformatics 4 (2003).
[38] Y. Zhang, et al., A quartet of PIF bHLH factors provides a transcriptionally centered signaling hub that regulates seedling morphogenesis through differential expression-patterning of shared target genes in *Arabidopsis*, PLoS Genet. 9 (1) (2013) e1003244.
[39] M.T. Weirauch, et al., Determination and inference of eukaryotic transcription factor sequence specificity, Cell 158 (6) (2014) 1431–1443.
[40] A.P. Boyle, et al., F-Seq: a feature density estimator for high-throughput sequence tags, Bioinformatics 24 (21) (2008) 2537–2538.
[41] Y. Zhang, et al., Model-based analysis of ChIP-Seq (MACS), Genome Biol. 9 (9) (2008) R137.
[42] G. Cuellar-Partida, et al., Epigenetic priors for identifying active transcription factor binding sites, Bioinformatics 28 (1) (2012) 56–62.
[43] J. Vierstra, et al., Functional footprinting of regulatory DNA, Nat. Methods 12 (10) (2015) 927–930.
[44] R.B. Deal, S. Henikoff, A simple method for gene expression and chromatin profiling of individual cell types within a tissue, Dev. Cell 18 (6) (2010) 1030–1040.
[45] R.B. Deal, S. Henikoff, The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana*, Nat. Protoc. 6 (1) (2011) 56–68.
[46] R. Milo, et al., Superfamilies of evolved and designed networks, Science 303 (5663) (2004) 1538–1542.
[47] R. Milo, et al., Network motifs: simple building blocks of complex networks, Science 298 (5594) (2002) 824–827.
[48] S.S. Shen-Orr, et al., Network motifs in the transcriptional regulation network of *E. coli*, Nat. Genet. 31 (1) (2002) 64–68.
[49] A.B. Stergachis, et al., Developmental fate and cellular maturity encoded in human regulatory DNA landscapes, Cell 154 (4) (2013) 888–903.
[50] P. Sieber, et al., Pattern formation during early ovule development in *Arabidopsis thaliana*, Dev. Biol. 273 (2) (2004) 321–334.
[51] K. Sugimoto, S.P. Gordon, E.M. Meyerowitz, Regeneration in plants and animals: dedifferentiation, transdifferentiation, or just differentiation? Trends Cell Biol. 21 (4) (2011) 212–218.
[52] M. Levine, R. Tjian, Transcription regulation and animal diversity, Nature 424 (6945) (2003) 147–151.
[53] C.A. Meyer, X.S. Liu, Identifying and mitigating bias in next-generation sequencing methods for chromatin biology, Nat. Rev. Genet. 15 (11) (2014) 709–721.
[54] E.Z. Kvon, et al., Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo, Nature 512 (7512) (2014) 91–95.
[55] T. Amano, et al., Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription, Dev. Cell 16 (1) (2009) 47–57.
[56] V.C. Calhoun, M. Levine, Long-range enhancer-promoter interactions in the Scr-Antp interval of the *Drosophila Antennapedia* complex, Proc. Natl. Acad. Sci. U. S. A. 100 (17) (2003) 9878–9883.
[57] R.M. Clark, et al., A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture, Nat. Genet. 38 (5) (2006) 594–597.
[58] M. Stam, et al., Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation, Genes Dev. 16 (15) (2002) 1906–1918.
[59] A.S. Nord, et al., Rapid and pervasive changes in genome-wide enhancer usage during mammalian development, Cell 155 (7) (2013) 1521–1531.
[60] A. Visel, et al., A high-resolution enhancer atlas of the developing telencephalon, Cell 152 (4) (2013) 895–908.
[61] W. Jiang, et al., Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in *Arabidopsis*, tobacco, sorghum and rice, Nucleic Acids Res. 41 (20) (2013) e188.
[62] J. Miao, et al., Targeted mutagenesis in rice using CRISPR-Cas system, Cell Res. 23 (10) (2013) 1233–1236.
[63] Q. Shan, et al., Targeted genome modification of crop plants using a CRISPR-Cas system, Nat. Biotechnol. 31 (8) (2013) 686–688.
[64] Z.Y. Feng, et al., Multigeneration analysis reveals the inheritance, specificity, and patterns of CRISPR/Cas-induced gene modifications in *Arabidopsis*, Proc. Natl. Acad. Sci. U. S. A. 111 (12) (2014) 4632–4637.
[65] B. van Steensel, J. Dekker, Genomics tools for unraveling chromosome architecture, Nat. Biotechnol. 28 (10) (2010) 1089–1095.
[66] J. Dekker, M.A. Marti-Renom, L.A. Mirny, Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data, Nat. Rev. Genet. 14 (6) (2013) 390–403.
[67] B. Tolhuis, et al., Looping and interaction between hypersensitive sites in the active beta-globin locus, Mol. Cell 10 (6) (2002) 1453–1465.
[68] S.S. Rao, et al., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, Cell 159 (7) (2014) 1665–1680.
[69] S. Feng, et al., Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in *Arabidopsis*, Mol. Cell 55 (5) (2014) 694–707.
[70] C. Wang, et al., Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*, Genome Res. 25 (2) (2015) 246–256.
[71] S. Grob, M.W. Schmid, U. Grossniklaus, Hi-C analysis in *Arabidopsis* identifies the KNOT, a structure with similarities to the flamenco locus of *Drosophila*, Mol. Cell 55 (5) (2014) 678–693.
[72] M.C. King, A.C. Wilson, Evolution at two levels in humans and chimpanzees, Science 188 (4184) (1975) 107–116.
[73] S.B. Carroll, Evolution and development of the insect body plan, Mol. Biol. Cell 6 (1995) 5.
[74] M. Koornneef, C. Alonso-Blanco, D. Vreugdenhil, Naturally occurring genetic variation in *Arabidopsis thaliana*, Ann. Rev. Plant Biol. 55 (2004) 141–172.
[75] T. Mitchell-Olds, J. Schmitt, Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis*, Nature 441 (7096) (2006) 947–952.
[76] R. DeCook, et al., Genetic regulation of gene expression during shoot development in *Arabidopsis*, Genetics 172 (2) (2006) 1155–1164.
[77] J. de Meaux, A. Pop, T. Mitchell-Olds, *Cis*-regulatory evolution of chalcone-synthase expression in the genus *Arabidopsis*, Genetics 174 (4) (2006) 2181–2202.
[78] X. Gan, et al., Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*, Nature 477 (7365) (2011) 419–423.
[79] M.T. Maurano, et al., Systematic localization of common disease-associated variation in regulatory DNA, Science 337 (6099) (2012) 1190–1195.
[80] ENCODE, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (7414) (2012) 57–74.
[81] M.B. Gerstein, et al., Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project, Science 330 (6012) (2010) 1775–1787.
[82] S. Roy, et al., Identification of functional elements and regulatory circuits by *Drosophila* modENCODE, Science 330 (6012) (2010) 1787–1797.
[83] V. Matys, et al., TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, Nucleic Acids Res. 34 (Database issue) (2006) D108–110.