Published online 28 November 2016 *Nucleic Acids Research, 2017, Vol. 45, No. 6* ***e41***
doi: 10.1093/nar/gkw1179

Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes

Zefu Lu[1], Brigitte T. Hofmeister[2], Christopher Vollmers[3], Rebecca M. DuBois[3] and Robert J. Schmitz[1,*]

[1]Department of Genetics, University of Georgia, Athens, GA 30602, USA, [2]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA and [3]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA

Received August 05, 2016; Revised November 03, 2016; Editorial Decision November 11, 2016; Accepted November 15, 2016

ABSTRACT

Chromatin structure plays a pivotal role in facilitating proper control of gene expression. Transcription factor (TF) binding of *cis*-elements is often associated with accessible chromatin regions. Therefore, the ability to identify these accessible regions throughout plant genomes will advance understanding of the relationship between TF binding, chromatin status and the regulation of gene expression. Assay for Transposase Accessible Chromatin sequencing (ATAC-seq) is a recently developed technique used to map open chromatin zones in animal genomes. However, in plants, the existence of cell walls, subcellular organelles and the lack of stable cell lines have prevented routine application of this technique. Here, we describe an assay combining ATAC-seq with fluorescence-activated nuclei sorting (FANS) to identify and map open chromatin and TF-binding sites in plant genomes. FANS-ATAC-seq compares favorably with published DNaseI sequencing (DNase-seq) results and it requires less than 50 000 nuclei for accurate identification of accessible genomic regions.

Summary: Application of ATAC-seq to sorted nuclei identifies accessible regions genome-wide.

INTRODUCTION

Eukaryotic genomes are tightly packed, beginning with wrapping short stretches of DNA around nucleosomes in a repeating unit that forms the structural basis of chromatin (1). Chromatin is categorized as either euchromatin or heterochromatin based on its transcriptional competence, its density and its accessibility (2,3). The distribution of nucleosomes along the chromosome provides different levels of accessibility of transcriptional machinery to cis-regulatory elements such as promoters and enhancers (4). These cis-regulatory regions, which are usually targeted by a diverse array of transcription factors (TFs), play pivotal roles in the regulation of gene expression (5,6). Therefore, identification of cis-regulatory sequences in their native chromatin environment is important for understanding how gene expression is coordinated throughout the plant to facilitate growth, development and responses to the environment.

The gold-standard method for identifying *in vivo* protein:DNA interactions and cis-regulatory regions for TFs of interest is chromatin immunoprecipitation-sequencing (ChIP-seq). However, the lack of antibodies for most plant TFs necessitates production of transgenic plants expressing epitope-tagged versions of proteins of interest that has prevented widespread application of this method in plants (7). Therefore, the development of feasible and scalable methods is required to facilitate identification of regulatory elements in plant genomes. Additional methods do exist that systematically reveal the identity of cis-regulatory elements by taking advantage of their location in 'open chromatin' zones, which makes them particularly prone to enzymes that digest exposed DNA such as microccocal nuclease (MNase) and DNase I (8,9). As a result, several methods that combine enzymatic digestion of isolated chromatin with high-throughput sequencing, such as DNase-seq, MNase-seq and FAIRE-seq (formaldehyde-assisted isolation of regulatory elements) have been developed to pinpoint potential accessible regions genome wide (8,10–12). For example, DNase-seq reveals regions of open chromatin (accessible regions) that are a few kilobases in size. Within accessible regions, direct interactions between proteins, such as TFs, and DNA prevent enzymatic digestion by DNase I, leaving protected 'footprints' of these proteins (13). These footprints are often less than 15 base pairs (bp) in size and they can be used to identify potential interacting TFs for which target motifs have been identified from *in vitro* protein:DNA interaction assays such as yeast-one hybrid, protein binding microarrays (PBMs) or DNA affinity purification sequencing (14,15). Finally, the totality of TF-footprint pairs can be used to construct gene regulatory networks (16,17).

[*]To whom correspondence should be addressed. Tel: +1 706 542 188; Fax: +1 706 542 3910; Email: schmitz@uga.edu

A new method, assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) improved the ability to identify accessible regions and DNA footprints (18). In this method, the Tn5 transposase is used instead of DNase I to access open chromatin. The Tn5 transposase integrates sequencing adapters directly into DNA eliminating the need for multiple reactions and purifications typically required for sequencing library construction. As a result, significantly lower amounts of starting nuclei are required for investigation of cis-regulatory elements. In fact, 50 000 nuclei are sufficient for this technique, as opposed to others like MNase-seq or DNAse-seq that often require 20- to 100-fold more nuclei (18). ATAC-seq has even been performed on single cells (19). However, in single cells or samples with reduced input of nuclei the identification of footprints is challenging due to limited Tn5 integration events.

ATAC-seq is now routinely being applied to systematically identify cis-regulatory regions and DNA footprints in animal genomes (20), however, the application of this method to plant species has been challenging. The major hurdle has been contamination of chromatin extracts by the mitochondrial and chloroplast genomes. The organellar genomes are completely accessible to Tn5, which likely depletes Tn5 activity from the nuclear genome (18). To circumvent this impediment we have developed a novel and robust method that combines ATAC-seq with fluorescence-activated nuclei sorting (FANS) for the identification of cis-regulatory regions and DNA footprints in plant genomes.

MATERIALS AND METHODS

Plant growth and extraction of crude nuclei

The *Arabidopsis thaliana* accession Col-0 and *Pro35S:H2AX-GFP* transgenic plants were grown on a vertical plate with $\frac{1}{2}$ LS media and 1% sugar in long day light conditions (16 h light/8 h dark) for 7–8 days. Approximately 0.2 g of whole seedlings or roots were collected and immediately chopped in 2 ml of pre-chilled lysis buffer (15 mM Tris-HCl pH7.5, 20 mM NaCl, 80 mM KCl, 0.5 mM spermine, 5 mM 2-ME, 0.2% TritonX-100). After chopping the total mixture was filtered with miracloth twice and then loaded on the surface of 2 mL dense sucrose buffer (20 mM Tris-HCl Ph8.0, 2 mM $MgCl_2$, 2 mM EDTAl, 15 mM 2-ME, 1.7 M sucrose, 0.2% TritonX-100) in a 15 ml Falcon tube, as described before (21). The nuclei were centrifuged at 2200 *g* at 4C for 20 min and the pellets were resuspended in 500 μl pre-chilled lysis buffer. Key steps and notes are listed in items 1–4 in the Supplementary Methods.

Nuclei sorting and Tn5 integration

Crude nuclei were stained with 4,6-Diamidino-2-phenylindole (DAPI) and loaded into a flow cytometer (Beckman Coulter MoFlo XDP). The exciting light strength of DAPI is set as 600 eV. A total of 50 000 nuclei were sorted based on their size, the strength of the DAPI signal and were subsequently collected in a tube with 500 μl of lysis buffer (Figure 1). The nuclei were pelleted by centrifugation at 1000 *g* at 4C for 10 min. After checking the quality of nuclei under a microscope using a DAPI channel, the nuclei were washed with Tris-Mg buffer (10 mM Tris-HCl pH8.0, 5 mM $MgCl_2$) once and the supernants were removed as clean as possible. To prepare transposomes, 10 μl of Tn5 transposase, which was purified and quantified following a previously published protocol (22), was incubated with 0.143 μl of annealed adapter mixture (25 μM each) at room temperature for 60 min. The purified nuclei were next incubated with 2 μl of transposomes (Tn5 transposase loaded with adapters) in 40 μl of Tagment buffer (10 mM TAPS-NaOH ph 8.0, 5 mM $MgCl_2$) at 37C for 30 min. Tn5 integration can also be carried out using the Illumina kit (Illumina, # FC-121-1031) at 37C for 30 min. The products are purified using a QIAGEN minielute kit and then amplified using Phusion DNA polymerase for 10–15 cycles. The PCR cycles are determined as described in (18). Key steps and notes are listed in items 5–7 in the Supplementary Methods. Amplified libraries were purified with AMPure beads and library concentrations were determined using a Qubit and qPCR prior to sequencing. The libraries were multiplexed and then sequenced using an Illumina NextSeq500 (Supplementary Table S1).

Analysis of sequencing data

Sequencing reads and Col-0 wild-type control DNase-seq (22) were mapped to Release 10 of the Arabidopsis Genome (TAIR10) using bowtie1 with parameters '-v 2 -m 3′ (23) (Supplementary Table S1). Duplicated reads were removed using the default parameters of picard. Accessible regions and peaks were identified using the default parameters of HOTSPOT (24). The center of identified peaks was used to define peak overlaps with genomic features using the following criteria. If a center site is located in (1) the promoter of a gene (2000 bp upstream from the transcriptional start site (TSS)), or (2) gene body, the peaks will be assigned to that gene. The distal intergenic regions refer to regions >3 kb from the TSS and >1 kb from the transcriptional end site (TES). The plot and heatmap of regulatory region distribution were obtained using ChIPseeker (25). Footprints were identified with pyDNase (26) using the following parameters '-fp 4,30,1 –dm –A'. The identified footprints were extracted and compared with the PBM database (27) using find individual motif occurrences (28). If the entire sequence of the motif was contained within a footprint it was assigned to that motif.

H2AX GFP transgenic plant

A full-length coding sequence of the histone 2A variant X (H2AX) (At1g08880) was PCR-amplified from cDNA, cloned into pDONR221® and subsequently into the pK7WG2 vector using the following primers (H2AXAt1g 08880_AttF ggggacaagtttgtacaaaaaagcaggctccatgagtacagg cg-caggaa and H2AXAt1g08880_AttRev ggggaccactttgtac aagaaagctgggtcgaactcctgagaa-gcagatcc) and the Gateway technology according to the manufacturer's instructions (Life Technologies^{TM}) (29): Arabidopsis (Col-0) plants were transformed and homozygous transgenic lines were selected as described previously (30).

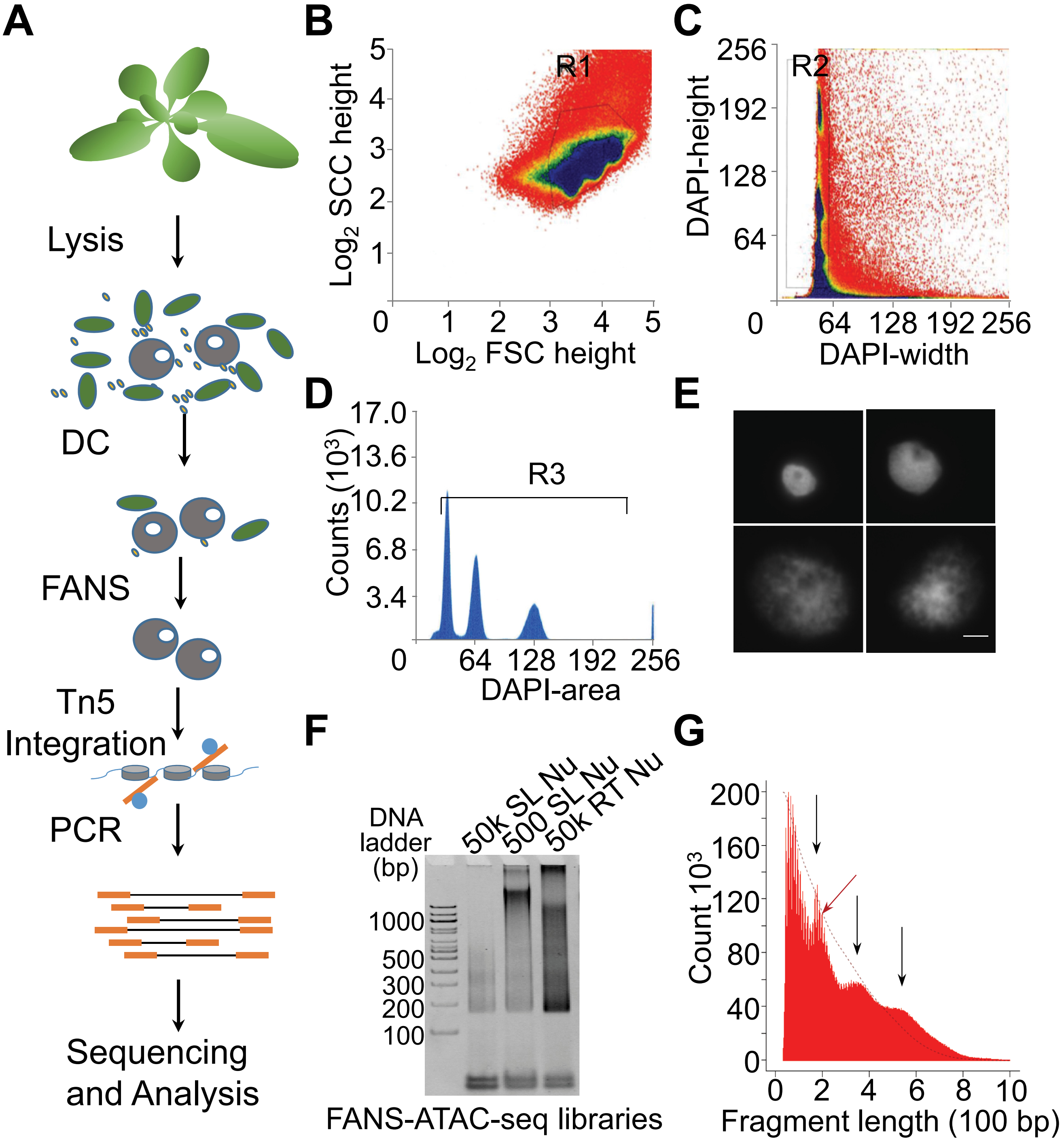

Figure 1. Overview of fluorescence-activated nuclei sorting-assay for transposase-accessible chromatin (FANS-ATAC)-sequencing. (**A**) Workflow of FANS-ATAC-sequencing. Fresh tissue is chopped in lysis buffer to release the nuclei, mitochondria and chloroplasts. Density centrifugation (DC) was used to isolate crude nuclei, which were subsequently stained with DAPI and used for FANS. High quality sorted nuclei were then collected and incubated with Tn5 for integration of adapters, which were then used to construct sequencing libraries. (**B–E**) Strategy of nuclei sorting with flow cytometry. The nuclei are sorted by (**B**) size, (**C**) DAPI width and the strength of (**D**) DAPI signals. R1, R2 and R3 refer to the nuclei selected for ATAC-seq. The number and quality of nuclei were verified using the (**E**) DAPI channel of a microscope. DAPI width refers to the time for the cell/nuclei to pass through the laser-detecting region of the flow cytometer. DAPI height indicates the strongest signal of a single drop. DAPI area refers to the total DAPI signal in each drop. (**F**) A representative gel image of an ATAC-seq library constructed after FANS. Discrete bands indicative of nucleosome fractions are found with high quality FANS-ATAC-seq libraries. 50K SL Nu was amplified for 11 cycles, 50K SL Nu for 15 cycles and 50K RT Nu for 12 cycles. SL, seedlings; RT, roots; Nu, nuclei; 50K, 50 000. (**G**) The insert sizes of ATAC-seq paired-end reads in the sample prepared from 500 nuclei. The dotted line indicates the trendline. Black arrows indicate the fragments containing one or more nucleosomes. The red arrow indicates the helical pitch of DNA that is observed due to steric hindrance between Tn5 transposases (52).

Scatter plots, correlations and signal portion of tags (SPOT) value calculations

To obtain peak intensity files for pairwise comparisons between different FANS-ATAC-seq experiments, accessible regions were identified using HOTSPOT (24). The overlapping accessible regions were then identified using BEDtools (31). Next, an equal number of reads from each sample were used to count the read numbers located in each overlapping regulatory region. These values were then used to create scatter plots and correlation scores between any two investigated samples using linear regression and the lm() function in R (32). To obtain SPOT values, 5 000 000 reads from each sample were randomly selected. The accessible regions were then identified by HOTSPOT and the proportion of reads that align to accessible regions versus the entire genome was calculated to determine the SPOT value as described previously (24,33).

Identification of tissue or assay specific peaks and footprints and gene ontology (GO) analysis

To identify peaks and footprints that are specific to each sample, an equal number of reads were randomly selected from compared samples. A total of 48 million reads were selected for the comparison of the Col-0 50 000 seedling FANS nuclei replicate 1 and Col-0 50 000 root FANS nuclei replicate 1 samples. A total of 59 million reads were selected for the comparison of the Col-0 50 000 seedling FANS nuclei replicate 1 and the Col-0 wild-type control DNase-seq (22) samples. The specific peaks were identified using the findPeaks function in Hypergeometric Optimization of Motif EnRichment (HOMER) (34) with one sample in each sample as a control to compare to the other sample, using the following parameters '-region -size 200 -minDist 50 -tbp 0''. Genes with at least one specific peak within 1 kb upstream of the TSS or gene bodies were labeled as sample specific genes. To identify significantly overrepresented GO terms, Fisher's exact test was performed using the GO Consortium browser tool that identifies biological process, molecular function and cellular component related GO terms (35,36). Bonferroni correction was used to adjust for multiple testing, and categories with a P-value < 0.01 were considered statistically significant.

Accessible regions of the Col-0 50 000 seedling nuclei replicate 1 and the Col-0 50 000 root nuclei replicate 1 were identified using HOTSPOT (24), and common regions containing all the regulatory regions in both samples were later used to identify footprints with pyDNase (26). To identify tissue-specific motifs *de novo*, the identified tissue specific footprint sequence were first masked with repeatmasker (http://www.repeatmasker.org) and then the resulting sequences were used for analysis using discriminative regular expression motif elicitation (28). To identify known motifs in tissue-specific peaks, find individual motif occurrences was used to search for matching motifs from the Arabidopsis PBM database (28).

RESULTS

Overview of FANS-ATAC-seq

In ATAC-seq experiments of mammalian cells, typically ∼30–70% of the sequenced reads align to the mitochondrial genome (18). This problem is exacerbated in plants that have chloroplasts along with their own genomes. Organellar genomes, such as the mitochondria and chloroplasts, are highly accessible and therefore susceptible to Tn5 integrations that decreases the efficiency of using ATAC-seq to map regulatory elements in the nuclear genome. To reduce the effect of organelle DNA interference on Tn5, we combined FANS followed by treatment of Tn5 transposomes (Tn5 loaded with sequencing adapters) according to the standard ATAC-seq protocol (FANS-ATAC-seq) (Figure 1A).

First, Arabidopsis tissue lysates were prepared by chopping samples in lysis buffer and nuclei were subsequently collected via density centrifugation. Next, crude nuclei were stained with DAPI and further enriched using flow cytometry. Nuclei were isolated according to size and intensity of the DAPI signals (Figure 1B and C). For fast growing young tissues, multiple DAPI signal peaks were observed as a result of endoreplication, which can be used as an important index to further select nuclei of interest (e.g. trichomes) (Figure 1D). DAPI stained nuclei were then screened using a microscope for quantity and quality (Figure 1E). A total of 50 000 or as few as 500 nuclei were collected and then washed with a Tris-Mg buffer to remove ethylenediaminetetraacetic acid, which was present in the lysis buffer prior to treatment with Tn5 transposomes. After incubation with Tn5, PCR and electrophoresis, discrete bands were observed (Figure 1F and G), as integration events are enriched in regions between nucleosomes.

Assessment of the quality of ATAC-seq data

A paired-end sequencing strategy was used to increase the number of nucleotide resolution integration sites per library. As shown in Table 1, samples from purified nuclei contained a significantly higher fraction of reads that mapped to the nuclear genome when compared to nuclei that were not obtained via FANS (crude nuclei – all samples described in this study are further enriched using FANS unless otherwise noted). The percentage of reads that aligned to the organellar genomes decreased from >50% to ∼30%, which is similar to observations from mapping open chromatin using DNase-seq in Arabidopsis (17). In general, duplicate reads and smaller integration products were abundant for those that aligned to the chloroplast and mitochondrial genomes as compared to reads that aligned to the nuclear genome (Figure 1G and Supplementary Figure S1). The percentage of reads that aligned to the nuclear genome for each of the samples generated from plants expressing green fluorescent protein labelled histone 2A.X (H2AX-GFP) replicates was 77.5% and 60.5%, which is comparable to observations from non-GFP experiments. This indicates that using GFP to further sort nuclei only marginally improves the purity of nuclei in Arabidopsis. Auto-fluorescence of organelles as well as their similar size to the Arabidopsis nuclei likely prevents better purity. As

Table 1. Alignment of FANS-ATAC-seq reads to organellar and nuclear genomes

ATAC samples		Reads percentage in rep 1			Reads percentage in rep 2		
		Before removing duplicates	After removing duplicates	Reads remaining if duplicates removed	Before removing duplicates	After removing duplicates	Reads remaining if duplicates removed
FANS	Crude nuclei (50k)			**48.31%**			**49.20%**
	Chr1-5	48.00%	72.43%	72.91%	46.39%	83.37%	79.61%
	ChrC	45.10%	19%	20.30%	48.71%	7.58%	18.79%
	ChrM	6.90%	9%	59.90%	4.89%	9.05%	64.17%
	Seedling nuclei (50k)			**69.73%**			**76.60%**
	Chr1-5	71.66%	83.69%	81.43%	64.90%	76.19%	89.95%
	ChrC	24.57%	12.23%	34.70%	32.13%	20.41%	48.68%
	ChrM	3.77%	4.08%	75.50%	2.97%	3.40%	87.57%
	Root nuclei (50k)			**80.90%**			**79.78%**
	Chr1-5	78.70%	75.03%	85.74%	85.13%	95.92%	80.72%
	ChrC	11.02%	18.59%	55.69%	7.16%	6.19%	68.92%
	ChrM	10.29%	6.38%	71.18%	7.71%	7.68%	79.84%
	Seedling nuclei (500)			**12.72%**			**12.40%**
	Chr1-5	71.70%	72.80%	12.92%	55.08%	59.95%	13.53%
	ChrC	24.40%	21.40%	11.15%	40.83%	33.76%	10.28%
	ChrM	3.90%	5.80%	19.07%	4.09%	6.29%	19.14%
	H2AX-GFP nuclei (50k)			**64.41%**			**58.10%**
	Chr1-5	77.51%	88.48%	73.48%	60.53%	79.34%	76.64%
	ChrC	20.06%	8.84%	28.38%	36.9%	17.46%	27.49%
	ChrM	2.43%	2.68%	70.99%	2.57%	3.20%	72.36%

The percentage of reads aligning to the nuclear, chloroplast and mitochondrial genomes for each sample assayed. Crude nuclei indicate a sample that was prepared without using FANS, whereas FANS was used for all other samples. Using FANS substantially increases the number of nuclear aligned reads. The number of duplicate reads is significantly higher from the ATAC-seq library prepared from 500 nuclei, as more PCR cycles were required to produce this library. 50k indicates 50 000.

expected, the number of reads that result from PCR duplicates increases with the number of PCR cycles, as is evident in data from the 500 nuclei sample (Table 1). Therefore, although purification of nuclei is not essential for successful implementation of ATAC-seq in Arabidopsis, enrichment of nuclei will substantially increase the number of useable reads for identification of integration sites in the nuclear genome as a result of depleted organellar genomes.

Identification of potential regulatory regions

To determine the efficiency of FANS-ATAC-seq, we compared our results with a high-quality published DNase-seq data set generated from nuclei isolated via INTACT (isolation of nuclei tagged in specific cell types) (17). FANS-ATAC-seq resulted in similar patterns of genome sequencing coverage when compared to the published DNase-seq (Figure 2A) (17). To identify potential cis-regulatory regions in our data, the HOTSPOT software (24) was used, which identified more than 20 000 accessible regions from each sample. Additionally, peak intensities of these identified accessible regions were highly reproducible between biological replicates ($R^2 > 0.86$) (Figure 2B and C and Supplementary Figure S2A and B), and were also comparable to a published high-quality DNase-seq data set ($R^2 > 0.77$) (Figure 2D). The identified accessible regions were reproducible as a large number of regions overlapped between replicates (Supplementary Table S2). The accessible regions identified from FANS-ATAC-seq are also highly consistent with those identified from DNase-seq. A total of 85.9% of the accessible regions identified by FANS-ATAC-seq (26 894 accessible regions) overlapped DNase-seq data and 77.2% of the accessible regions identified by DNase-seq (29 905 accessible regions) overlapped with ATAC-seq results (Figure 2E). The differential accessible regions between FANS-ATAC-seq and DNase-seq were identified using HOMER's findPeaks option. A total of 6283 regions were specific to the FANS-ATAC-seq sample and 4979 regions were specific to the DNase-seq sample (Supplementary Figure S3A and B). Many of the FANS-ATAC-seq specific regions were located where accessible regions are expected to reside (Supplementary Figure S3), whereas the DNase-seq specific regions were mostly enriched in gene bodies (Supplementary Figure S3). Lastly, GO analysis of the assay-specific accessible regions revealed that many genes belonged in 'response to environment stimulus' category, which could be attributed to the fact these samples were grown by two different labs in different environments (Supplementary Figure S4).

The purity of nuclei is critical to the quality of the ATAC-seq results. Although we were able to, at times, produce useable ATAC-seq data from nuclei not purified using FANS (crude nuclei), comparisons of these data to other samples uncovered some issues. For example, a comparison of the reproducibility between replicates of samples prepared from crude nuclei revealed a much lower correlation ($R^2 = 0.78$)

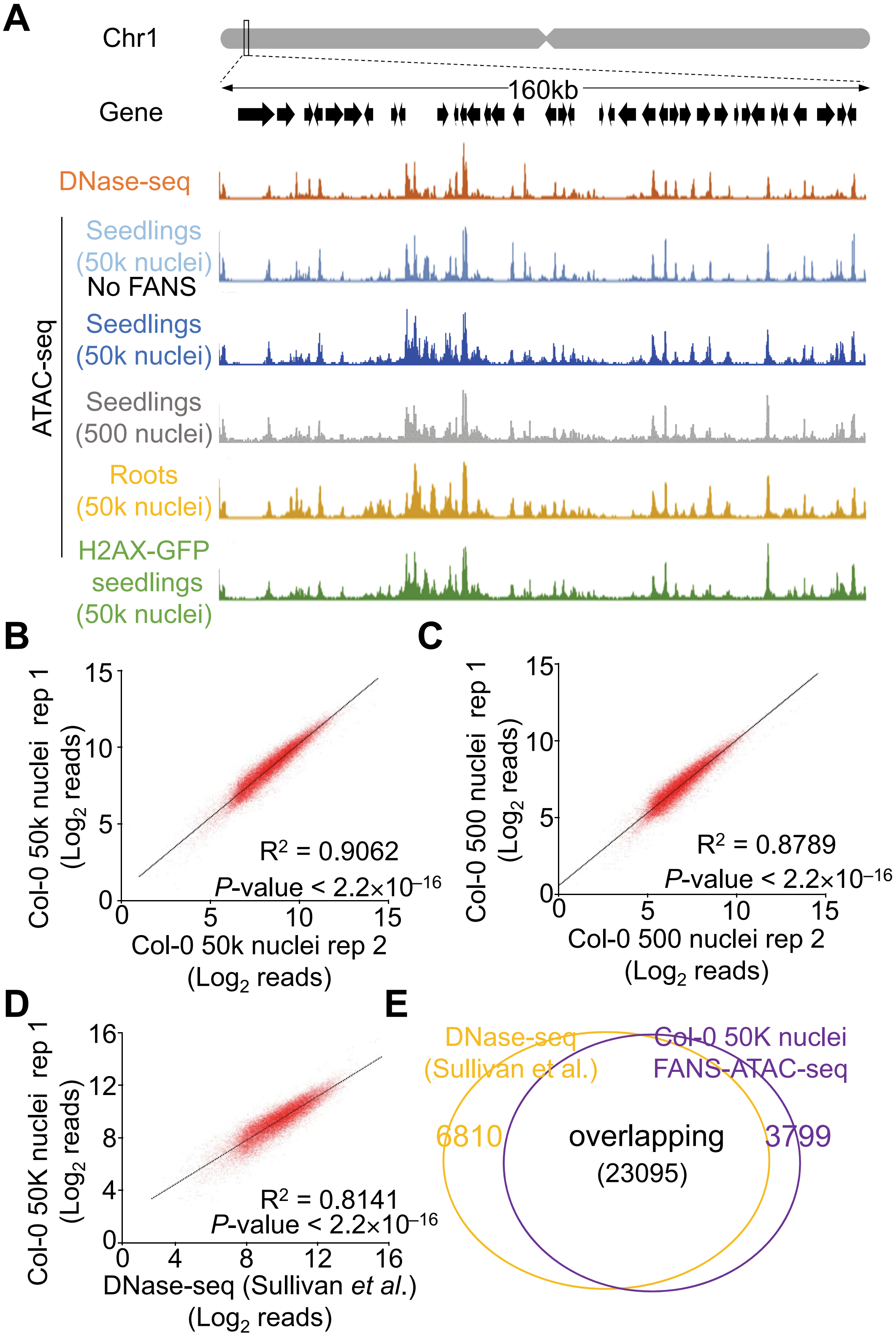

Figure 2. Identification of accessible regions with FANS-ATAC-seq. (**A**) A representative 160 kb region from chromosome 1 showing a comparison between DNase-seq (17), ATAC-seq and FANS-ATAC-seq including a sample sequenced from 500 nuclei and a sample sequenced for sorting a nuclear GFP line. The nuclei were first sorted using the DAPI signal and then both the DAPI and GFP signal were used for isolation of the *Pro35S:H2AX-GFP* nuclei. 50k indicates 50 000. (**B** and **C**) Correlation of peak intensities among different FANS-ATAC-seq biological replicates. (**D**) Correlation of peak intensities between DNase-seq and FANS-ATAC-seq. (**E**) A Venn diagram showing the overlap of accessible regions between FANS-ATAC-seq and DNase-seq. A total of 45M reads were randomly selected from DNase-seq (17) and FANS-ATAC-seq using the Col-0 50 000 seedling nuclei rep1 sample and then HOTSPOT was used to identify accessible regions.

than other samples (Supplementary Figure S2C). Moreover, to evaluate the potential impact of the sorting procedure, we prepared two independent batches of nuclei and for each batch we used half for preparation for a FANS library and the other half for preparation for a non-FANS (crude nuclei) library. A comparison of the peak intensities between the two samples revealed that they were relatively reproducible ($R^2 > 0.76$ and $R^2 > 0.86$ in two replicates, respectively) (Supplementary Figure S5A) and as reproducible as a comparison between the two crude nuclei replicates ($R^2 = 0.78$). The accessible regions identified from the FANS prepared sample and the same non-sorted sample showed more than 87% overlap (Supplementary Figure S5B). These results indicate that the inclusion of FANS results in higher quality data and limits additional effects due to the enrichment of higher purity of intact nuclei.

Next, the SPOT value was calculated, which refers to the percentage of reads that aligned to the identified accessible regions, as this is a good indicator of the signal-to-noise ratio from open chromatin sequencing assays. These results revealed that the signal-to-noise ratio was similar or higher between FANS-ATAC-seq and other published DNase-seq data sets from Arabidopsis (Supplementary Figure S6) (33,37,38). For the FANS-ATAC-seq libraries prepared from 500 nuclei, a similar enrichment of reads aligning to accessible regions was observed when compared to libraries prepared from 50 000 nuclei (Figure 2A and Supplementary Figure S7A and B). Furthermore, the accessible regions identified from the 500 nuclei sample shares substantial overlap of accessible regions when compared to samples prepared from 50 000 nuclei (Supplementary Figure S7C and D). This indicates that FANS-ATAC-seq is a useful methodology for the identification of cis-regulatory regions even from samples that have limited numbers of nuclei such as those found in the stem cell niche.

To determine how many reads are sufficient for identification of Tn5 integration-sensitive regions, a rank test was performed with 10–100 million reads for 10 ranks randomly selected from a FANS-ATAC-seq experiment that started with 50 000 nuclei isolated from a plant expressing a nuclear GFP-tagged protein (*Pro35S:H2AX-GFP*). This analysis revealed that 10 million aligned chromosome reads are sufficient to detect 26 967 accessible regions, which covers more than 92% of the 29 147 accessible regions that were found from alignment of 100 million reads (Supplementary Figure S8A). Furthermore, the identified accessible regions using different numbers of input reads shared a large overlap, although using 100 million reads resulted in the largest number of identified regions (Supplementary Figure S8B). For example, the experiment using 10 million input reads shared all but 114 accessible regions with the 100 million input read sample, however, the 100 million read sample identified an additional 4824 regions (Supplementary Figure S8C). Notably, the quality of accessible regions from using 100 million reads was stronger (Supplementary Figure S8D).

Genomic distribution of accessible regions identified by FANS-ATAC-seq

Of the putative accessible regions identified using FANS-ATAC-seq, more than 90% were located within 3 kb upstream of a TSS whereas less than 10% were located in distal intergenic regions (further than 3 kb from the TSS or further than 1 kb from the transcriptional termination site and gene bodies (Figure 3A). The accessible regions identified are mostly enriched around the TSS (Figure 3B and C), which is consistent with these regions containing cis-regulatory elements in Arabidopsis. Interestingly, we find the accessible regions were also highly enriched in the TES regions (Figure 3B), consistent with some downstream regions being required for transcription.

To test whether the identified accessible regions are associated with biological functions, we searched for tissue-specific accessible regions using nuclei isolated from whole seedlings (Col-0 50 000 seedling replicate 1) and separately from roots (Col-0 50 000 root replicate 1). A total of 10 863 accessible regions were identified as seedling specific and 4246 accessible regions were identified as root specific (Supplementary Figure S9). Of these a total of 1265 seedling-specific genes and 808 root-specific genes were identified by defining genes with peaks within 1 kb upstream of the TSS or within the gene body as tissue-specific genes. GO analysis revealed that many of the seedling-specific genes are involved in photosynthesis whereas root specific genes are mostly involved in response to environment stimulus (Supplementary Figure S10), which is consistent with some of the major functions of these different tissues. Collectively, these results show that FANS-ATAC-seq successfully and efficiently identifies open chromatin regions and by extension likely a large number of biological relevant cis-regulatory elements.

Identification of DNA footprints

In addition to identifying open chromatin regions, assays such as DNase-seq and ATAC-seq also reveal DNA footprints, which are regions of the genome that are bound by a protein, often a TF that protects the DNA from enzymatic digestion or Tn5 integration (17,18). These footprints can be used to identify potential TF-binding motifs *in vivo*. To determine whether FANS-ATAC-seq can likewise detect TF-binding profiles, we first compared published ChIP-seq data from seven transcription factors with FANS-ATAC-seq results, which showed that most of the ChIP-seq peaks are associated with accessible regions (Figure 4A and Supplementary Table S3) (39–45). These results indicate that FANS-ATAC-seq is an appropriate method for genome-wide discovery of DNA-TF interactions. To systematically identify DNA footprints genome-wide we used pyDNase to identify Tn5 integration-insensitive sites from the 50 000 Col-0 seedling sample and for the extraction of footprints from the candidate accessible regions identified by HOTSPOT (Figure 4B). A total of 93 956 DNA footprints were identified from 26 894 accessible regions. Comparing the sequences of these footprints with the protein binding motif database (PBM) (14), identified 29 135 DNA footprints associated with a known footprint from

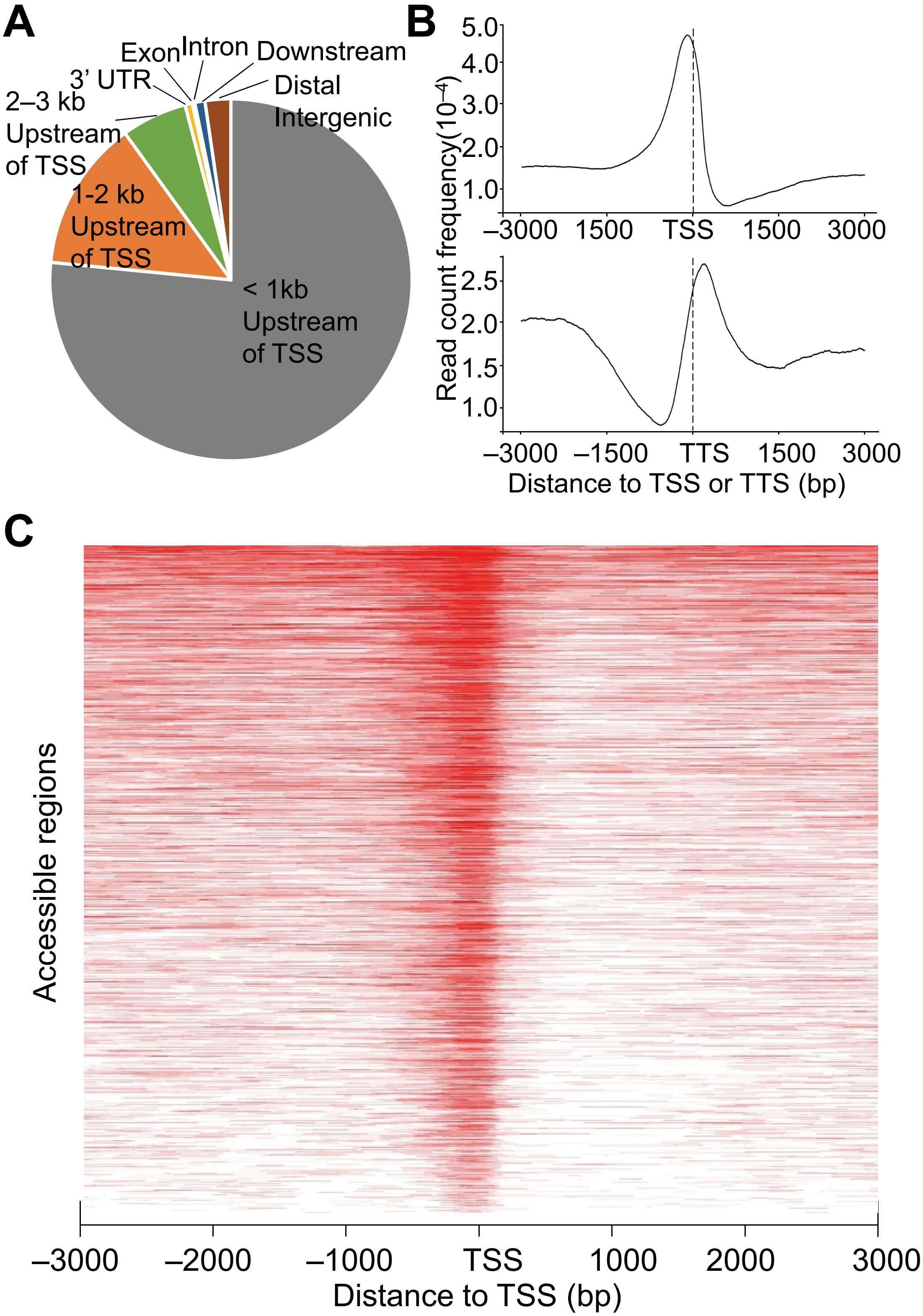

Figure 3. Distribution of accessible regions identified using FANS-ATAC-seq. (**A**) A pie chart showing the distribution of accessible regions identified from FANS-ATAC-seq using the Col-0 50 000 seedling nuclei rep1 sample throughout the Arabidopsis genome. Downstream refers to center of accessible regions located less than 1 kb from the transcriptional termination sites (TTS). Distal intergenic regions refer to accessible regions located more than 3 kb away from the transcriptional start sites (TSS) and more than 1 kb from TTS. (**B**) Distribution of accessible regions around the TSS and TTS identified from FANS-ATAC-seq using the Col-0 50 000 seedling nuclei rep1 sample. The center of accessible regions was used to produce the distribution plots. (**C**) A heatmap showing the distribution of accessible regions around the TSS identified by FANS-ATAC-seq using the Col-0 50 000 seedling nuclei rep1 sample.

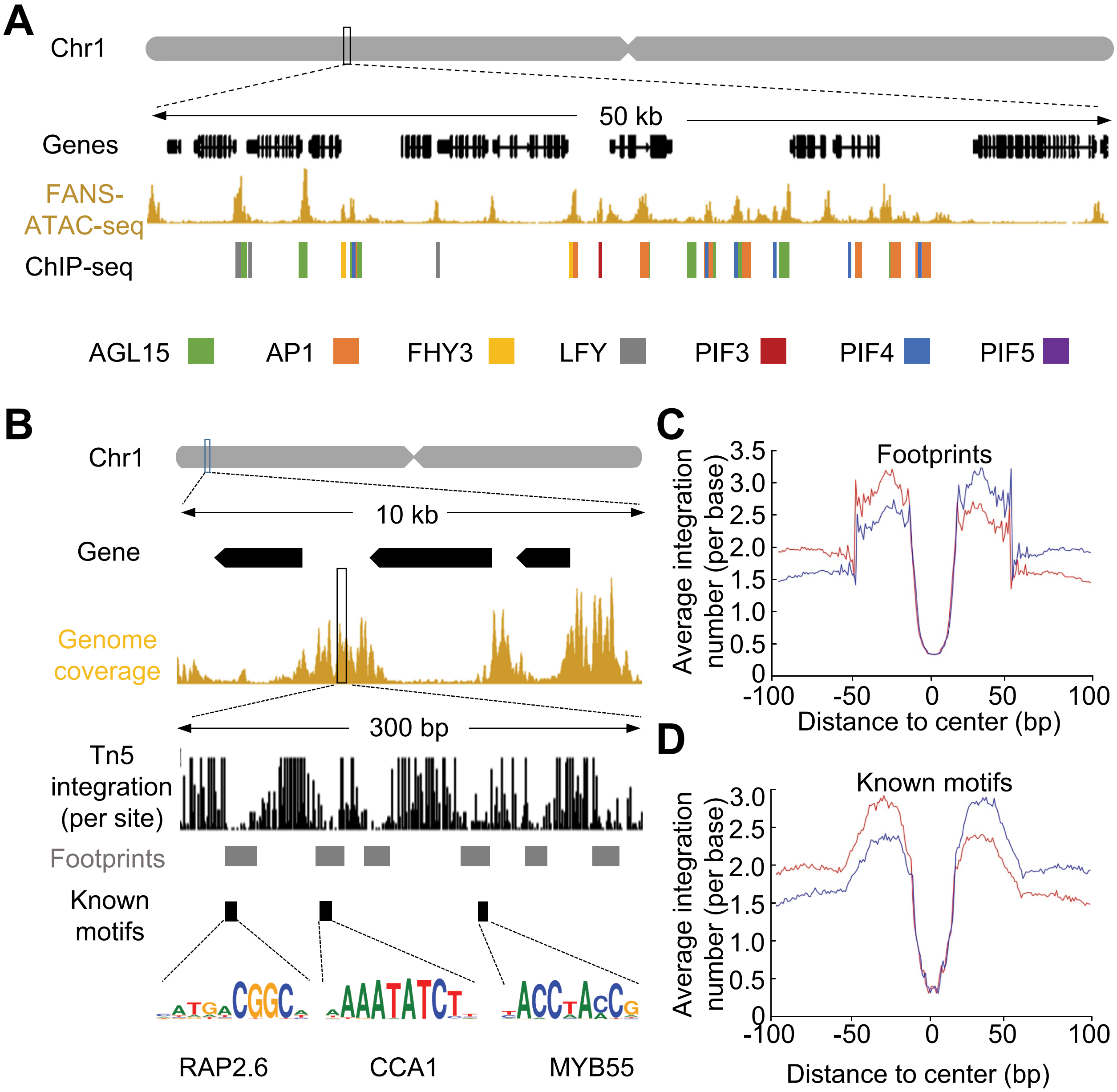

Figure 4. Identification of transcription factors footprints using FANS-ATAC-seq. (**A**) Comparison of published ChIP-seq peaks with FANS-ATAC-seq results using the Col-0 50 000 seedling nuclei rep1 sample. (**B**) Representative footprints identified using the Col-0 50 000 seedling nuclei rep1 sample. The footprints were identified using pyDNase and then labeled with known motifs by comparisons to the protein binding motif database. Tn5 integration per sites are defined as the first base (5′ end) of each sequencing read. (**C** and **D**) Distribution of Tn5 integration sites around (**C**) FANS-ATAC-seq and (**D**) known footprints. High frequencies of Tn5 integrations sites are observed in regions 50 bp around footprints whereas lower Tn5 integrations are observed in the footprints. The blue line indicates the integration sites on anti-sense strand whereas the red line indicates those on positive-sense strand.

the PBM data, demonstrating the ability of FANS-ATAC-seq to identify TF-binding sites *in vivo*. A deeper characterization of the identified footprints showed that Tn5 integration enrichment occurs in a 50 bp region surrounding the footprint (Figure 4B and C). To establish how many reads are sufficient for identification of footprints, the same rank-test approach used to study accessible regions was used to identify how many reads are required for accurate DNA footprint identification. From these data, ∼60 million reads are sufficient to identify ∼80% of the high-quality footprints (96 464 footprints) identified from using 100 million reads (124 627 footprints) (Supplementary Figure S11A). Furthermore, using greater numbers of reads increased the quality of identified footprints (Supplementary Figure S11B). However, there was a substantial number of footprints from the 100 million input read sample that could not be identified using 60 million input reads (Supplementary Figure S11C and D), indicating additional reads are required to obtain more accurate footprints.

To determine if the footprints identified by FANS-ATAC-seq are relevant to biological processes, we identified tissue-specific footprints by comparing FANS-ATAC-seq data from whole seedlings and roots. A total of 54 722 root-specific and 36 624 seedling-specific footprints were found. Next, the DNA sequence in these footprints were masked (see Materials and Methods) and used to identify conserved motifs with discriminative regular expression motif elicitation (28). Several tissue-specific motifs were identified, including TCP and PIF3 binding motifs in seedlings whereas WRKY and some HOMEODOMAIN binding motifs were found in roots (Supplementary Figure S12). We also searched for known motifs in these tissue-specific peaks and found that some tissue-specific binding motifs showed significant variation between tissues. For example, there was significantly more Related to AP2/Ethylene Response Factor (RAP/ERF) protein binding motifs in roots compared to seedlings, whereas there was more TCP/PIF (TEOSINTE BRANCHED1-CYCLOIDEA-PROLIFERATING CELL FACTOR1/PHYTOCHROME INTERACTING FACTOR) protein binding motifs in seedlings (Supplementary Figure S13) that is consistent with some of their functions. These results indicate that FANS-ATAC-seq can be used to identify cis-regulatory regions and transcription factor binding profiles in plant genomes, opening new avenues into studying plant gene regulation.

Tn5 integration bias

Both DNase I and Tn5 transposase show a bias for DNA substrates, as is the case with most enzymes (46,47). We tested the integration bias of Tn5 by using the ATAC-seq protocol on genomic DNA; in this experiment there are no proteins present and therefore DNA footprints should not exist. At a genome-wide view it appears there is no integration bias (Figure 5A), however, zooming in reveals an appreciable amount of bias near some of the identified footprints from FANS-ATAC-seq (Figure 5A). Using the same analysis methods to study footprints from nuclei, a similar distribution of integration sites was observed in genomic DNA (Figure 5B). This indicates that some DNA sequences are prone to Tn5 integrations leading to the presence of 'pseudo DNA footprints' even if a protein does not bind the motif. The bias was further confirmed when only considering pseudo DNA footprints that overlapped with known motifs present in the PBM data (Figure 5C). Lastly, we analyzed the distribution of Tn5 integration events at 18 different known TF-binding motifs that revealed varying degrees of potential bias (Figure 5D and Supplementary Figure S14). For example, the distribution around AHL20 was similar between genomic DNA and FANS-ATAC-seq, although the strength of the footprint signal was not as strong in genomic DNA (Figure 5D). Another TF, WOX13, showed no such bias (Figure 5E). Therefore, some footprints found in ATAC-seq data will result from the integration bias of Tn5 and not necessarily from TF occupancy.

DISCUSSION

ATAC-seq has proven to be a powerful method to advance studies of protein:DNA interactions and to study gene regulatory networks (17,18,48,49). Here, we report the combination of FANS with ATAC-seq to identify cis-regulatory regions in plant genomes. Our initial attempts to perform ATAC-seq in plants repeatedly failed or resulted in data that were not highly reproducible between replicates. For example, although we have generated ATAC-seq data from non-sorted nuclei, in a closer inspection of the nuclei using a microscope it was revealed that many were broken, which leads to a higher background after sequencing. This is further supported by reduced correlations between accessible regions identified in the two samples that were not sorted (Supplementary Figure S2). In our experience the use of FANS minimizes contamination by organellar genomes and it also ensures stable high quality intact nuclei are used in the assay making the performance of FANS-ATAC-seq more robust. The nuclei in Arabidopsis are similar in size to the organelles that result in some contamination of nuclei purification. This limitation is alleviated in plant species with larger genomes, as the larger size and stronger DAPI signal of high-quality intact nuclei decreases the cross contamination of organelles from sorting methods. For example, in FANS-ATAC-seq performed with maize, we found that over 98.77% of the sequenced reads aligned to the nuclear genome (Supplementary Table S4). Other methods to enrich high quality nuclei, such as INTACT, can also be combined with ATAC-seq. However, this requires production of transgenic plants, which is time consuming, expensive and not feasible for all experimental plant systems. By contrast, FANS-ATAC-seq requires the ability to sort nuclei using either DAPI or nuclear florescent proteins. FANS-ATAC-seq combined with laser capture microdissection could be especially impactful for cell-type specific studies as cis-regulatory regions can be identified with as few as 500 nuclei and likely can be performed on single nuclei as has been demonstrated in mammalian systems (19). One challenge of adding FANS in assessing cell-type specific accessibility is the potential impact of the sorting procedure. To evaluate this possibility we compared accessible regions identified via FANS to the non-sorted sample that showed more than 87% overlap (Supplementary Figure S5). Even though the non-sorted samples have a higher background this level of overlap suggests that the sorting procedure does not significantly impact the study of accessible regions. The data presented here show that FANS-ATAC-seq is a robust method and that it will readily translate to other plant species.

There are, however, some important considerations to take into account prior to using FANS-ATAC-seq. For example, using DAPI to enrich nuclei will result in complex cell populations making it more difficult to identify certain cis-regulatory elements compared to cell-type specific enrichment of nuclei. However, FANS-ATAC-seq can be applied to nuclear-labeled fluorescent protein tagged transgenic plants, which are widely available for a number of distinct cell types in numerous plant species. Another limitation is the use of the Tn5 transposase, which similar to DNase I, shows bias for certain DNA sequences in the genome (Figure 5 and Supplementary Figure S14). This will affect the ability to accurately identify DNA footprints for certain TFs that may be associated with pseudo DNA footprints created by Tn5. The careful dissection of these Tn5 prone integration sites will be an important area of future investigation for all species ATAC-seq. The identification of

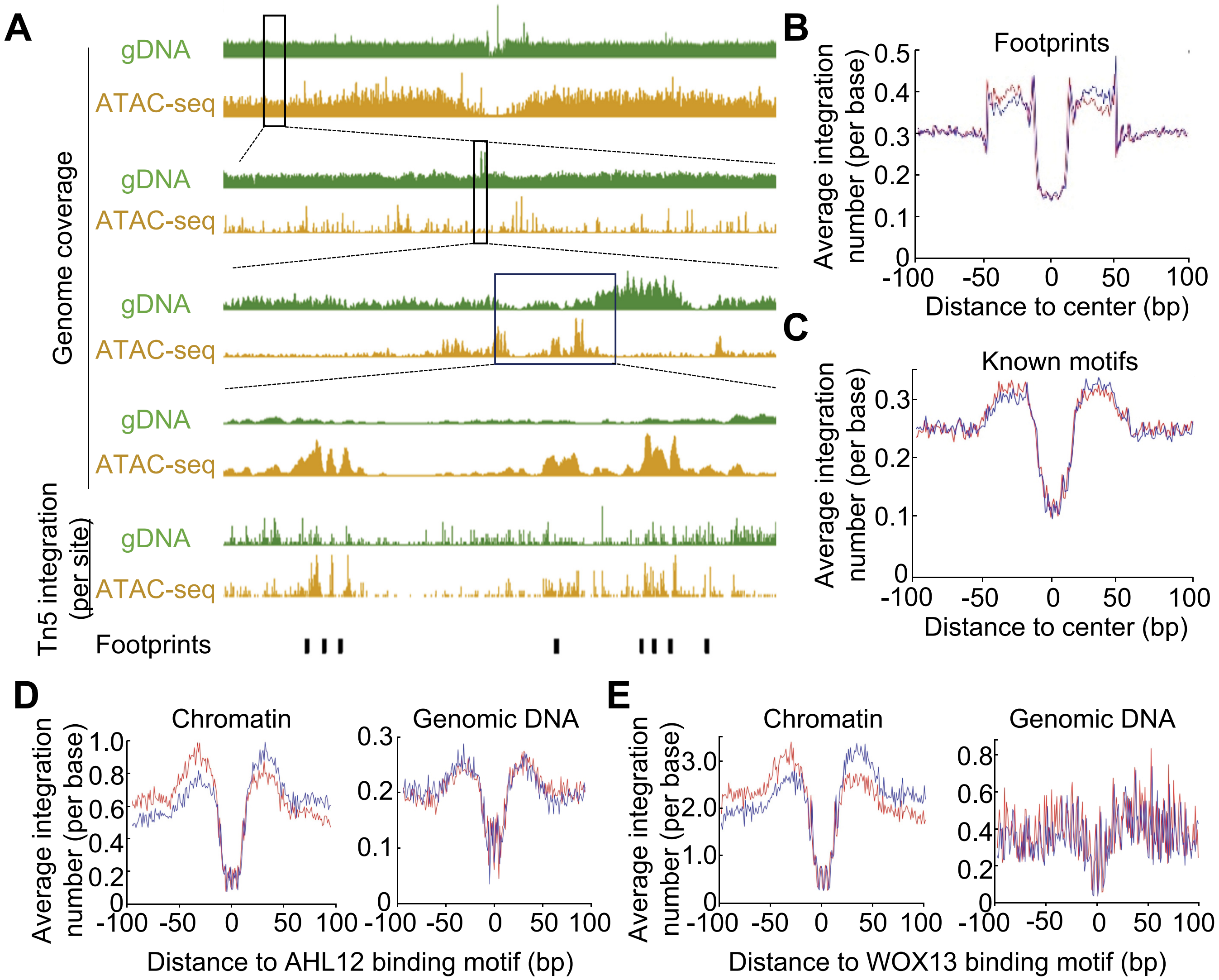

Figure 5. Bias of Tn5 integration in Arabidopsis. (**A**) Representative Tn5 integration plots from genomic DNA for chromosome 1. The integration-sensitive regions are uniformly distributed along the entire chromosome (first two lines) with few biased regions (the middle 6 lines), whereas a strong integration bias is observed when zooming in (bottom line). Genome coverage indicates the read coverage; Tn5 integrations per sites indicates the position of the 5′ end of each read that reflects the direct integration sites of Tn5. (**B**) Distribution of Tn5 integration sites from genomic DNA around footprints identified using FANS-ATAC-seq. Tn5 integration sites showed a similar pattern to that of FANS-ATAC-seq, however, the signal is not as strong in genomic DNA. Furthermore, the distributions of integration sites are symmetrical around footprints in genomic DNA whereas they are asymmetrical in FANS-ATAC-seq. (**C**) Distribution of Tn5 integration sites in genomic DNA around known motifs identified from the PBM database. (**D** and **E**) Distribution of Tn5 integration sites around (**D**) AHL12 and (**E**) WOX13 in data sets from FANS-ATAC-seq (left) and genomic DNA (right). Tn5 integrations sites showed similar patterns around the AHL12 binding motif in FANS-ATAC-seq and genomic DNA whereas the patterns are different for WOX13. The blue line indicates the integration sites on anti-sense strand whereas the red line indicates those on positive-sense strand.

these biased integrations can be overcome with each additional ATAC-seq data set that is generated similar to background signals that result from ChIP-seq.

The ability to rapidly and efficiently identify cis-regulatory regions and DNA footprints in plant genomes will accelerate the ability to link genotype to phenotype. Studies in mammalian systems have already demonstrated that combining epigenome maps, including open chromatin maps, increases the power to identify signals from genome-wide association studies (GWAS) (50). Recently, a similar approach of combining open chromatin maps with GWAS data in maize revealed that genetic variation in accessible regions significantly improves the ability to identify candidate causal mutations associated with quantitative traits (51). Creating maps of accessible regions for additional accessions and for other plant species will likely have similar impacts on GWAS reducing the time and effort to identification of causal genetic variants.

The study of gene regulatory networks is an exciting area of development in the plant sciences that has been mostly limited to the study of DNA sequence in combination with transcriptomics. However, the recent development and application of DNA affinity purification sequencing to Arabidopsis TFs will undoubtedly accelerate this research area

(15). However, a full understanding of the biology of gene regulatory networks will require integration of data from *in vivo* methods that can detect protein:DNA interactions, such as FANS-ATAC-seq.

DATA DEPOSITION

Raw and processed data files have been deposited to the National Center for Biotechnology Information Gene Expression Omnibus under accession GSE85203.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Nathan Springer and Rhiannon Macrae for comments on this manuscript as well as Eirini Kaiserli for sharing *Pro35S:H2AX-GFP* seeds. The Georgia Genomics Facility provided assistance with sequencing and the Cytometry Shared Resource Laboratory at UGA provided assistance with nuclei sorting. This work was supported by the National Science Foundation (IOS – 1546867) and by the Office of the Vice President of Research at UGA. BTH is supported by a National Institutes of Health Training Grant (T32GM007103) awarded to the Department of Genetics at UGA.

Author Contributions. R.J.S. and Z.L. conceived and designed the study. Z.L. performed experiments and data analysis. CV and RMD contributed reagents. R.J.S. and Z.L. wrote the manuscript.

FUNDING

National Science Foundation [IOS – 1546867]; Office of the Vice President of Research at UGA; National Institutes of Health [Training Grant T32GM007103 awarded to the Department of Genetics at UGA to B.T.H.]. Funding for open access charge: National Science Foundation [IOS – 1546867].

Conflict of interest statement. None declared.

REFERENCES

1. Kornberg,R.D. and Lorch,Y. (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, **98**, 285–294.
2. Simpson,R.T. (1990) Nucleosome positioning can affect the function of a cis-acting DNA element *in vivo*. *Nature*, **343**, 387–389.
3. Schones,D.E., Cui,K., Cuddapah,S., Roh,T.Y., Barski,A., Wang,Z., Wei,G. and Zhao,K. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.
4. He,H.H., Meyer,C.A., Shin,H., Bailey,S.T., Wei,G., Wang,Q., Zhang,Y., Xu,K., Ni,M., Lupien,M. *et al.* (2010) Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.*, **42**, 343–347.
5. Felsenfeld,G. (1992) Chromatin as an Essential Part of the Transcriptional Mechanism. *Nature*, **355**, 219–224.
6. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
7. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
8. Zaret,K. (2005) Micrococcal nuclease analysis of chromatin structure. *Curr. Protoc. Mol. Biol.*, doi:10.1002/0471142727.mb2101s69.
9. Lu,Q. and Richardson,B. (2004) DNaseI hypersensitivity analysis of chromatin structure. *Methods Mol. Biol.*, **287**, 77–86.
10. Carey,M. and Smale,S.T. (2007) Micrococcal nuclease-southern blot assay: I. MNase and restriction digestions. *CSH Protoc.*, **2007**, doi:10.1101/pdb.prot4890.
11. Giresi,P.G., Kim,J., McDaniell,R.M., Iyer,V.R. and Lieb,J.D. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**, 877–885.
12. Song,L. and Crawford,G.E. (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *CSH Protoc.*, **2010**, doi:10.1101/pdb.prot5384.
13. Brenowitz,M., Senear,D.F. and Kingston,R.E. (2001) DNase I footprint analysis of protein-DNA binding. *Curr. Protoc. Mol. Biol.*, doi:10.1002/0471142727.mb1204s07.
14. Franco-Zorrilla,J.M., Lopez-Vidriero,I., Carrasco,J.L., Godoy,M., Vera,P. and Solano,R. (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 2367–2372.
15. O'Malley,R.C., Huang,S.S., Song,L., Lewsey,M.G., Bartlett,A., Nery,J.R., Galli,M., Gallavotti,A. and Ecker,J.R. (2016) Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, **165**, 1280–1292.
16. Neph,S., Stergachis,A.B., Reynolds,A., Sandstrom,R., Borenstein,E. and Stamatoyannopoulos,J.A. (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, **150**, 1274–1286.
17. Sullivan,A.M., Arsovski,A.A., Lempe,J., Bubb,K.L., Weirauch,M.T., Sabo,P.J., Sandstrom,R., Thurman,R.E., Neph,S., Reynolds,A.P. *et al.* (2014) Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. *Cell Rep.*, **8**, 2015–2030.
18. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
19. Buenrostro,J.D., Wu,B., Litzenburger,U.M., Ruff,D., Gonzales,M.L., Snyder,M.P., Chang,H.Y. and Greenleaf,W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486–490.
20. Meyer,C.A. and Liu,X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*, **15**, 709–721.
21. Yu,P., McKinney,E.C., Kandasamy,M.M., Albert,A.L. and Meagher,R.B. (2015) Characterization of brain cell nuclei with decondensed chromatin. *Dev. Neurobiol.*, **75**, 738–756.
22. Picelli,S., Bjorklund,A.K., Reinius,B., Sagasser,S., Winberg,G. and Sandberg,R. (2014) Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.*, **24**, 2033–2040.
23. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
24. John,S., Sabo,P.J., Thurman,R.E., Sung,M.H., Biddie,S.C., Johnson,T.A., Hager,G.L. and Stamatoyannopoulos,J.A. (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.*, **43**, 264–268.
25. Yu,G., Wang,L.G. and He,Q.Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.
26. Piper,J., Elze,M.C., Cauchy,P., Cockerill,P.N., Bonifer,C. and Ott,S. (2013) Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.*, **41**, e201.
27. Franco-Zorrilla,J.M., Lopez-Vidriero,I., Carrasco,J.L., Godoy,M., Vera,P. and Solano,R. (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 2367–2372.
28. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.

29. Karimi,M., Inze,D. and Depicker,A. (2002) GATEWAY vectors for Agrobacterium-mediated plant transformation. *Trends Plant Sci.*, **7**, 193–195.
30. Kaiserli,E. and Jenkins,G.I. (2007) UV-B promotes rapid nuclear translocation of the Arabidopsis UV-B specific signaling component UVR8 and activates its function in the nucleus. *Plant Cell*, **19**, 2662–2673.
31. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
32. Ito,K. and Murphy,D. (2013) Application of ggplot2 to Pharmacometric Graphics. *CPT Pharmacometrics Syst. Pharmacol.*, **2**, e79.
33. Sullivan,A.M., Bubb,K.L., Sandstrom,R., Stamatoyannopoulos,J.A. and Queitsch,C. (2015) DNase I hypersensitivity mapping, genomic footprinting and transcription factor networks in plants. *Curr. Plant Biol.*, **3–4**, 40–47.
34. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
35. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
36. Gene Ontology, C. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
37. Pajoro,A., Madrigal,P., Muino,J.M., Matus,J.T., Jin,J., Mecchia,M.A., Debernardi,J.M., Palatnik,J.F., Balazadeh,S., Arif,M. *et al.* (2014) Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biol.*, **15**, R41.
38. Zhang,W., Zhang,T., Wu,Y. and Jiang,J. (2012) Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *Plant Cell*, **24**, 2719–2731.
39. Biddie,S.C., John,S., Sabo,P.J., Thurman,R.E., Johnson,T.A., Schiltz,R.L., Miranda,T.B., Sung,M.H., Trump,S., Lightman,S.L. *et al.* (2011) Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol. Cell*, **43**, 145–155.
40. Zheng,Y., Ren,N., Wang,H., Stromberg,A.J. and Perry,S.E. (2009) Global identification of targets of the Arabidopsis MADS domain protein AGAMOUS-Like15. *Plant Cell*, **21**, 2563–2577.
41. Ouyang,X., Li,J., Li,G., Li,B., Chen,B., Shen,H., Huang,X., Mo,X., Wan,X., Lin,R. *et al.* (2011) Genome-wide binding site analysis of FAR-RED ELONGATED HYPOCOTYL3 reveals its novel function in Arabidopsis development. *Plant Cell*, **23**, 2514–2535.
42. Winter,C.M., Austin,R.S., Blanvillain-Baufume,S., Reback,M.A., Monniaux,M., Wu,M.F., Sang,Y., Yamaguchi,A., Yamaguchi,N., Parker,J.E. *et al.* (2011) LEAFY target genes reveal floral regulatory logic, cis motifs, and a link to biotic stimulus response. *Dev. Cell*, **20**, 430–443.
43. Zhang,Y., Mayba,O., Pfeiffer,A., Shi,H., Tepperman,J.M., Speed,T.P. and Quail,P.H. (2013) 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**, e1003244.
44. Oh,E., Zhu,J.Y. and Wang,Z.Y. (2012) Interaction between BZR1 and PIF4 integrates brassinosteroid and environmental responses. *Nat. Cell Biol.*, **14**, U802–U864.
45. Hornitschek,P., Kohnen,M.V., Lorrain,S., Rougemont,J., Ljung,K., Lopez-Vidriero,I., Franco-Zorrilla,J.M., Solano,R., Trevisan,M., Pradervand,S. *et al.* (2012) Phytochrome interacting factors 4 and 5 control seedling growth in changing light conditions by directly controlling auxin signaling. *Plant J.*, **71**, 699–711.
46. He,H.H., Meyer,C.A., Hu,S.S., Chen,M.W., Zang,C., Liu,Y., Rao,P.K., Fei,T., Xu,H., Long,H. *et al.* (2014) Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods*, **11**, 73–78.
47. Green,B., Bouchier,C., Fairhead,C., Craig,N.L. and Cormack,B.P. (2012) Insertion site preference of Mu, Tn5 and Tn7 transposons. *Mob. DNA*, **3**, 3.
48. Giorgetti,L., Lajoie,B.R., Carter,A.C., Attia,M., Zhan,Y., Xu,J., Chen,C.J., Kaplan,N., Chang,H.Y., Heard,E. *et al.* (2016) Structural organization of the inactive X chromosome in the mouse. *Nature*, **535**, 575–579.
49. Davie,K., Jacobs,J., Atkins,M., Potier,D., Christiaens,V., Halder,G. and Aerts,S. (2015) Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. *PLoS Genet.*, **11**, e1004994.
50. Pickrell,J.K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, **94**, 559–573.
51. Rodgers-Melnick,E., Vera,D.L., Bass,H.W. and Buckler,E.S. (2016) Open chromatin reveals the functional maize genome. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E3177–E3184.
52. Adey,A., Morrison,H.G., Asan, Xun,X., Kitzman,J.O., Turner,E.H., Stackhouse,B., MacKenzie,A.P., Caruccio,N.C., Zhang,X. *et al.* (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.*, **11**, R119.