Single-cell analysis of cis-regulatory elements
Alexandre P. Marand and Robert J. Schmitz

Abstract
Plant tissues and organs are composed of functionally discrete cell types that are all defined by the same genome sequence. Cell-type variation in part arises from differential accessibility of cis-regulatory elements that encode the blueprints for transcriptional programs underlying cell identity and function. Owing to technical limitations, the role of cis-regulatory elements in cell identity maintenance, differentiation, and functional specialization has remained relatively unexplored in plant systems. Single-cell profiling has emerged as a powerful tool to circumvent these past obstacles by enabling unbiased charting of transcriptional and cis-regulatory states at the resolution of individual cells. Here, we review state-of-the-art single-cell approaches and analytical frameworks that have paved the way for establishing the link between cellular phenotypic variation and cis-regulatory mechanisms in plants.

Addresses
Department of Genetics, University of Georgia, Athens, GA 30602, USA
Corresponding author: Schmitz, Robert J (schmitz@uga.edu)

Current Opinion in Plant Biology 2022, 65:102094
This review comes from a themed issue on Growth and Development
Edited by Moritz Nowack and Keiko Sugimoto
For a complete overview see the Issue and the Editorial
Available online xxx
https://doi.org/10.1016/j.pbi.2021.102094
1369-5266/© 2021 Elsevier Ltd. All rights reserved.

Keywords
ATAC-seq, Chromatin, cis-regulatory elements, Development, Single cell, Differentiation, Cell fate.

Introduction
Cis-regulatory elements (CREs) encode the genomic blueprints for cellular gene expression programs underlying differentiation, cell fate acquisition, and responses to environmental stimuli [1–4]. CREs are composed of noncoding clusters of short 4-30bp DNA motifs that act as binding sites for sequence-specific transcription factors (TFs). The cooperative interaction of specific combinations of TFs bound to clusters of CRE sequences interface with various cofactors and components of the transcription machinery that dictate transcriptional outcomes of target genes. Most CRE types (promoters, enhancers, and silencers) directly influence gene expression via physical interaction or proximity. Core promoters, regarded as regions that contain gene transcription start sites (TSSs), are generally insufficient to drive mRNA biosynthesis above basal levels. In contrast, gene-proximal and distal CREs can be located from a few to hundreds of kilobases away from their target genes and regulate transcription through spatial interactions with the core promoter mediated by chromatin looping and higher-order chromatin organization independent of location, orientation, and position of target genes [5,6]. Gene-proximal and distal (and certain intronic) CREs have the capacity to alter the rate and quantity of transcription, including positive (transcriptional enhancers) and negative (silencers) regulation of gene expression [6–13]. However, systematic identification of active CREs within a given cell has remained a major challenge, particularly in plants, owing to a lack of stable cell lines and approaches capable of unbiasedly measuring CRE activity within individual cells.

Efforts to characterize CREs genome-wide from individual cells have been greatly aided by a proliferation of methods that measure chromatin accessibility [14]. Owing to steric requirements of most DNA binding proteins, CREs are often embedded within nucleosome-free chromatin, termed accessible chromatin regions (ACRs) [15]. Chromatin profiling methods such as Assay for Transposase Accessible Chromatin sequencing (ATAC-seq) and deoxyribonuclease (DNase) I hypersensitive sequencing leverage the nucleosome-depleted properties of CRE-associated chromatin to tag genomic locations with potential cis-regulatory functions [16,17]. New approaches for collecting chromatin accessibility information at the level of single cells take advantage of these assays by coupling them with microfluidics or combinatorial indexing to attach cell-specific barcodes to ACRs before sequencing [18–20]. As a result, the development of single-cell approaches has revolutionized the ability to profile chromatin accessibility at the level of individual cells to interrogate cell-type diversity (Figure 1a). In this review, we evaluate the experimental and computational approaches for understanding the role of CREs in cell differentiation and establishment of
cell identity and detail the current state-of-the-art methods for analysis of single-cell ATAC-seq (scATAC-seq) data.

**Profiling chromatin accessibility from individual plant nuclei**

The most widely adopted techniques for profiling accessible chromatin from individual nuclei are (1) droplet-based scATAC-seq (10X Genomics and Bio-Rad scATAC-seq solutions) and (2) single-cell combinatorial indexing ATAC-seq (sciATAC-seq) (Figure 1b, c) [19–21]. Both approaches are amenable to plant nuclei isolated via fluorescence-activated nuclei sorting or traditional wash-and-centrifuge techniques.

For droplet-based scATAC-seq, chromatin is tagmented in bulk using Tn5 loaded with universal sequencing adapters. Intact and tagmented nuclei are then loaded onto a microfluidic chip with reagents for polymerase chain reaction (PCR) that collide with an aqueous solution containing barcoded beads, where each bead contains millions of copies of a distinct barcode sequence immobilized to its surface. Isolated droplets containing a barcoded bead, a single nucleus, and PCR reagents are produced as the aqueous stream passes a channel of slow-moving oil. Initial amplification attaches bead-immobilized barcodes to the universal sequencing adapters, ensuring that tagmented nuclei within a droplet receive a barcode sequence distinct from other droplets. Finally, the droplet emulsion is broken, and an Illumina-compatible sequencing library is prepared by bulk PCR [20].

In contrast to droplet-based methods, sciATAC-seq does not require proprietary barcoded beads or a microfluidic device. Instead, nuclei are distributed into a 96-well plate (≈2,500–5,000 nuclei per well) containing distinct combinations of Tn5 loaded with unique barcoded adapters (96 combinations) [18]. After on-plate tagmentation, nuclei are pooled and redistributed into new 96-well plate(s) at low concentrations (10–50 nuclei per well). Each well contains a unique combination of barcoded sequencing adapters that are attached via PCR that result in ≈9, 216 distinct barcode combinations per plate.

**Assessing single-cell data quality**

The success of a scATAC-seq experiment can be gleaned by evaluating several metrics produced from software such as Socrates, snapATAC, or ArchR.
The output of a sequenced scATAC-seq library is millions of paired-end reads associated with cell-specific barcode sequences. Because barcodes can be attached to DNA fragments from broken nuclei or ambient DNA in solution, several computational methods have been used to separate barcodes representing single nuclei from background noise. For example, the number of unique Tn5 integrations can be used to determine cutoffs using mixture models, spline inflection points, or heuristics because barcodes with few associated reads are indicative of empty droplets (Figure 2a). Broken nuclei or fragments in solution can be identified by estimating the enrichment of Tn5 integration at TSSs or by deriving proportions of Tn5 integrations near TSSs or within ACRs (Figure 2b, c). Quality ATAC-seq libraries exhibit nucleosome-free, mono-, di-, and tri-nucleosome-protected fragments (as well as periodicity corresponding to major/minor grooves) and can be evaluated before and after sequencing [24]. Fragment distributions in bulk and for individual nuclei enriched for nucleosome-free fragments are generally desirable, although a lack of any nucleosome-protected fragments could indicate an abundance of broken nuclei or over-saturation of Tn5 (Figure 2d). Owing to a lack of chromatin in nucleoid structures, organellar genomes are highly susceptible to Tn5 integrations, and barcodes associated with high frequency of organellar-derived reads could indicate encapsulation of organelles rather than nuclei [25]. In addition, overloading nuclei in scATAC-seq approaches can result in multiple nuclei with the same barcode, termed multiplets. Low numbers of multiplets are particularly important for trajectory analysis (detailed in the following) as a single barcode derived from two different cell types may appear as a transitional cell state. Empirical identification of multiplet barcodes has been used to validate loading concentrations that consistently produce single-nuclei profiles, such pooled populations of nuclei from different species or genotypes (Figure 2e, f) [2,26]. In contrast to mammalian species, plant cells may undergo extensive endoreduplication, a phenomenon associated with cellular maturation. Thus, removing barcodes associated with read depth should be avoided. Taken together, identification and removal of reads derived from organelles and broken nuclei provide a key advantage of scATAC-seq over traditional bulk-level methods for improving data quality.

Although the following suggestions should be taken as guidelines, establishing robust quality control metrics are essential for determining the utility of a data set. In general, the most important parameters for judging the success of an experiment are the proportion of recovered
cells (>40%, this is the minimal expected recovery noted by 10X Genomics, lower values may indicate extensively degraded nuclei), the number of unique Tn5 insertions per cell (species-dependent, >1,000), the proportion of Tn5 insertions localizing to ACRs per cell (species-dependent, >30%), and a low multiplet rate (<5%, higher values indicate prevalent nuclear aggregation that can negatively affect downstream analysis and lead to erroneous conclusions). In addition, biological replicates are critical for evaluating reproducibility of cellular states and the relative contribution of noise [14].

**Analysis of single-cell cis-regulatory elements enables discovery**

The cornerstone of single-cell analysis is the identification of cell types from an input tissue or organ sample. As scATAC-seq data are highly sparse and binary, current

---

**Figure 3**

Analysis of cis-regulatory elements from single-cell chromatin accessibility data.
methods for revealing cellular identities have relied on aggregating data from many nuclei with similar genome-wide chromatin accessibility profiles. Specifically, a binary matrix of nuclei scored by the presence/absence of Tn5 insertions for a list of features (such as 1 kb windows or ACRs) is generated, normalized, and then compressed (dimensionality reduction) to enable graph-based clustering and visualization of nuclei-to-nuclei relationships (Figure 3a). Once nuclei with similar genome-wide chromatin accessibility patterns are clustered into groups, cellular identities can then be revealed by assessing the relative chromatin accessibility of marker genes or by transferring cell-type labels from annotated single-cell RNA-seq data (Figure 3b) [2,27,28].

Aggregation of nuclei by cell type has proven useful for identifying differentially accessible cell-type-specific ACRs and TF motifs across cell types in Zea mays (maize) and Arabidopsis thaliana (Figure 3c) [2,27,28]. Approximately, 1/3 of all ACRs within maize and A. thaliana exhibit some degree of cell-type specificity [2,27,28]. However, TF motifs may be even more discriminative when evaluated combinatorially; a model trained solely on relative TF motif accessibilities in maize was able to predict cell-type annotations with high accuracy (0.94), sensitivity (0.93), and specificity (0.99) [2]. Evaluation of chromatin accessibility dynamics among TF motifs and their cognate locus across cell types has even been used to reveal TFs with putative autonomous and non-cell autonomous activity [2]. Thus, aggregating individual nuclei based on cell identity has the potential to uncover dynamic cell-type-specific regulatory mechanisms and reveal gene regulatory networks underlying cell identity and function.

One advantage of single-cell approaches over cell lines and purification techniques is the ability to study native transitional cell states. Analysis of pseudotime trajectories — computational ordering of cells by their differentiation status via supervised and/or unsupervised approaches — has enabled remarkable insights into the fine-scale regulation of cellular development (Figure 3d) [29]. For example, Dorrity et al. [28] demonstrated that endodermal subtype cells in A. thaliana roots exhibit increasing levels of endoreduplication, progressive loss of transcriptional diversity, and distinct chromatin landscapes associated with different stages of development that were not detected in single-cell RNA-seq data alone. Similarly, phloem companion cell differentiation in maize was associated with a decrease in total chromatin accessibility, suggesting that acquisition of specialized cell functions may be underlined by coordinated transcriptional silencing through unknown mechanisms of chromatin regulation [2].

In addition to chromatin accessibility information at individual loci, pairwise analysis of co-accessible chromatin regions (coACRs) in maize has been successfully demonstrated to capture known chromatin interactions, such as TEOSINTE BRANCHED 1 (tb1), RELATED TO AP2.7 (ZmRAP2.7), and BENZOXAZINLESS 1 (bx1) from chromatin conformation capture sequencing techniques (Figure 3e) [2,6,30]. A recent study in maize identified three million coACRs, around 1/3 of which were associated with cell-type specificity (Figure 3f) [2]. ACRs underlying cell-type-specific coACRs were associated with known cell-type-specific marker genes, enhancer activity, and genetic polymorphisms implicated in phenotypic variation [2].

Plant species lack a homolog to CCCTC-binding factor, a key chromatin architectural regulator in metazoans [31]. In search of a CCCTC-binding factor-like factor in plants, an analysis of motif enrichment in the anchors of coACRs revealed several TF candidates with reciprocal-binding sites that were consistently enriched across cell types. In addition to new candidates, many of the enriched motifs belonged to the plant-specific TF family TEOSINTE BRANCHED 1/CYCLOIDEA/PROLIFERATING CELL NUCLEAR ANTIGEN FACTOR, which had been previously implicated in chromatin interactions based on proximity ligation methods in several plant species [32–34]. Thus, analysis of coACRs offers an indirect approach to study the regulation of three-dimensional chromatin structure at cell-type resolution.
Conclusion and future directions

Single cell—based methods have the potential to address several outstanding questions in plant biology (Figure 4). Owing to a continuously replenished pool of undifferentiated meristematic stem cells, most plants contain cells at various stages of development at any given static time point, making it possible to capture distinct cell states along the continuum of development [35]. However, by profiling cells across a dense time series, the effects of physiological maturity toward differentiation programs and heterogeneity of cell states can be detangled. Furthermore, it has become clear from studies in mammals that a significant proportion of genome-wide association study variants are associated with cell-type-specific effects in vivo [36,37]. Accumulating evidence indicates that phenotypetype-associated variants co-localize with expression Quantitative Trait Loci and chromatin-accessible Quantitative Trait Loci for genes and genomic regions with marked cell-type specificity [38,39]. Application of scATAC-seq to large diversity panels and mutant stocks offers an opportunity to uncover the genetic determinants of cis-regulatory variation within and across cell types and to quantify cell-type-specific effects underlying phenotypic variation [40–44]. In addition to questions of allelic variation, it is unknown whether spatial localization and cell identity influence perception of environmental cues. Comparison of scATAC-seq data between stress and control conditions has the potential to uncover CREs with cell-type-specific biases toward stress responses [45,46]. Coupling such case—control designs with dense time course sampling or in large diversity panels and/or mutant backgrounds would be highly informative for decoupling the effects of developmental stage and cis-regulatory genetic variation for how plant cells interpret and respond to a host of biotic and abiotic signals. We envision that the increase in throughput from thousands to millions of cells per experiment, the adoption of multi-modal techniques, such as characterization of chromatin accessibility and gene expression from the same cell, and refinement of spatial chromatin accessibility profiling (such as spatial-ATAC-seq) will result in new paradigms of transcriptional regulation in the context of development [47–53]. Furthermore, coupling scATAC-seq/sciATAC-seq methods with orthogonal histone modification and TF-binding site assays at single-cell resolution will be key to elucidating fine-scale regulatory mechanisms [54,55]. Through these innovations, analysis of multiple modalities from the same cell has the unprecedented potential to decouple chromatin signatures with gene expression. A future goal will be to produce reference cell-type atlases for plant species similar to reference genome sequencing efforts. High-quality reference atlases will provide a single framework for multiple plant species on which various environmental, genetic, and other treatments can be measured against. Taken together, single-cell methods have and will continue to revolutionize our ability to precisely model the molecular control of cell identity and will enable the next frontier of discovery in plant sciences.

Declaration of competing interest

RJS is a co-founder of REquest Genomics, a company that provides epigenomic services. APM declares no competing interests.

Acknowledgements

This study was funded with support from the NSF (IOS-1856627 and IOS-2026534) and the UGA Office of Research to RJS and an NSF postdoctoral fellowship in biology (DBI-1905869) to APM.

References

Papers of particular interest, published within the period of review, have been highlighted as: * of special interest ** of outstanding interest


Multi-organ scATAC-seq study in maize that moved beyond proof-of-principle. scATAC-seq was optimized for use in plants and revealed cell-type-specific (1) accessible chromatin regions, (2) TF binding sites, (3) putative chromatin interactions, (3) autonomous and non-autonomous TFs, (4) signatures of selection, (5) differentiation pseudo-time trajectories and (6) evolution of CREs between maize and Arabidopsis thaliana.


Identification and validation of CREs within intronic enhancers in Arabidopsis thaliana. Genome editing of discrete CREs within intronic DNAse-I hypersensitive sites revealed modular CREs enhancer activity with marked effects on transcription levels and developmental/cell-type-specific expression.


Proof-of-principle integration of scATAC-seq and snRNA-seq in Arabidopsis thaliana root. This study highlighted the utility of scATAC-seq to uncover cell identities by differential chromatin accessibility through comparison of single-cell chromatin accessibility and nuclear transcription profiles.


Characterization of cellular endoreduplication and endoreduplication via integration of scATAC-seq and scRNA-seq from Arabidopsis thaliana root. This study also demonstrated that cell-type-specific enrichment of TF motifs could be linked to expression of individual TFs and dynamic chromatin accessibility at distinct genomic loci — providing a detailed gene regulatory network underlying cellular development of Arabidopsis thaliana root cell types.


Demonstration of single-cell co-accessibility as a proxy for chromatin interactions. This study highlighted a computational method to link distal-CREs to their target genes, uncovered chromatin hubs regulated by similar sets of TFs, and illustrated dynamic chromatin interactions through pseudotime analysis of myoblast differentiation.


Demonstration of joint chromatin accessibility and gene transcription from the same cell using mixed mouse and human cell lines. This study established proof-of-principle for collecting multomic data from the same cell and revealed that chromatin accessibility underlies a significant proportion of transcriptional variance among and within discrete cell types in the mouse kidney.


