Chromatin accessibility profiling methods

AUTHORS[1]

[1]*Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305*

Abstract

Introduction

Chromatin accessibility refers to the level of physical compaction of chromatin, a complex formed by DNA and associated proteins consisting mainly of histones and DNA-binding transcription factors (TF)[81–83]. Although eukaryotic chromatin is generally tightly packed into nucleosomes, which comprise ∼147 bp of DNA wrapped around an octamer of histones[84–86] (Historical review in Trojer and Reinberg[?]), chromatin accessibility is not spatiotemporally stable, nor uniform across the genome. Histones are typically depleted at genomic locations that interact with transcriptional regulators (e.g. TFs), such as at enhancers, promoters and other regulatory elements[88–93]. Therefore, profiling chromatin accessibility on a genome-wide scale can serve as an excellent tool to identify all putative regulatory elements in a cell type or state. Note that not only the compaction but also post-translational modifications of the chromatin, including DNA methylation and histone tail acetylation, are dynamic and change in different cell states, and can reflect specific functions of genomic regions[94,95]. Initial changes in accessibility may be facilitated by chromatin-binding TFs, specifically so-called pioneer factors, that thermodynamically outcompete histones and recruit ATP-dependent chromatin remodelers, allowing other TFs to co-bind and further stabilize the nucleosome depleted region and cooperatively regulate gene expression of target genes[96–98]. Consequently, the analysis of TF binding sites within accessible regions can bring insight into the master regulators and gene regulatory networks of the studied cell type.

Changes in the chromatin landscape as well as mutations in chromatin remodelers and non-coding mutations in accessible regions have been linked to a range of diseases[99–102]. In fact, many causal GWAS variants are located in accessible regulatory elements (Maurano et al., 2012) and TF-bound DNA harbors increased mutation rates since TFs and DNA repair enzymes compete for damaged regulatory regions[103,104]. In order to improve our understanding of chromatin dynamics during development and in disease contexts, since several years, researchers and large consortia such as the ENCODE Consortium and the NIH Roadmap Epigenomics Mapping Consortium have embarked on collecting and comparing chromatin landscapes across cell types and during disease development[88,89,105,106].

Over the past decades, we have witnessed the development and widespread use of several chromatin accessibility profiling methods[77,78,107–110,112,113,122,123]. Generally, these methods are based on the physical accessibility to enzymes that fragment, tagment or methylate DNA in chromatin. Initial screens in the 1970s, showed that regions of active transcription were particularly sensitive to digestion by DNA endonucleases, indicating a more permissive form of the chromatin[76], and that the chromatin digested at regularly spaced sites due to nucleosome phasing[75,81,82]. Note that DNase I was, and still is, the reagent of choice for TF footprinting, which can determine the location of TF binding sites[74,79,80]. Technological advances have led to the first genome-wide assessments of accessible chromatin in 2008 by sequencing genomic DNA fragments following DNase I digestion, a technique referred to as DNase I hypersensitive site sequencing (DNase-seq)[109,124]. This was followed by the development of a handful more genome-wide chromatin accessibility profiling methods[73,110,113,122?] , of which Assay for Transposase-Accessible Chromatin using sequencing ATAC-seq (and variants)[113,122,123], together with DNase-seq are the two most used chromatin accessibility profiling methods today. As these methods are high-throughput sequencing-based, the analysis of the generated omics data heavily relies on bioinformatics, not only for the initial processing but also to biologically interpret chromatin accessibility profiles and to perform more intricate downstream analyses.

Importantly, as regulatory regions co-define a cell type, their accessibility is cell type-dependent, especially for distal regulatory regions[71,72,91]. When investigating heterogeneous samples, it is therefore advisable to measure chromatin accessibility at a single-cell level as bulk methods will yield population-averaged accessibility profiles (Fig. 1).

Currently, the field of single-cell omics, including single-cell epigenomics such as single-cell ATAC-seq (scATAC-seq), is booming due to the unprecedented opportunities to assess genome regulation in complex tissues such as the brain, whole embryos and tumors[69,70,117–120]. Accompanied with the rise of several single-cell chromatin accessibility profiling technologies, bioinformatic tools have been developed that allow analysis of the generated data which is intrinsically sparse XXX.

Although chromatin accessibility profiling methods may serve as an analytic foundation to identify regulatory regions, it is reported that only around 10-26% of accessible or predicted regulatory regions in human are in fact active in[116]. In addition, linking (active) accessible regulatory regions to their target genes solely based on accessibility data remains elusive. Additional data, including transcriptomics, reporter assays and 3D chromatin architecture maps, especially when combined in a multi-omics fashion, may help (in the future) to determine an accessible region's functionality and putative target genes[68] Hafez et al., 2017; Moore et al., 2020; Bravo et al., 2020; Sanyal et al., 2012; Ron et al., 2017.

This Primer on chromatin accessibility profiling methods provides an overview of the most used and latest methods to profile chromatin accessibility, both on bulk and single-cell level. In addition, it provides an in-depth outline of bioinformatic analysis tools and examples of applications in diverse fields. Lastly, the Primer discusses standards for data deposition and examines currently unmet needs and future possibilities to increase our understanding of chromatin accessibility landscapes and their functional role in gene regulation during development, evolution and in disease contexts.

Experimentation

Experimental assays for analyzing bulk cell chromatin accessibility

The emergence of next generation sequencing techniques (NGS) has revolutionized the way that chromatin is investigated. Chromatin accessibility is traditionally probed by assays such as deoxyribonuclease I (DNase) digestion, micrococcal nuclease (MNase), or restriction enzyme digestion, typically at a few selected genomic regions each time. Chromatin modifications and occupancy can be analyzed by immunoprecipitation (ChIP). The combination of these assays with NGS results in the "seq" techniques such as ChIP-seq[64–67], DNase-seq[91,109], ATAC-seq[113], MNase-seq[78], which have enabled the analysis of chromatin states to a whole genome level. We briefly describe the principles, pros and cons of these techniques in this section.

DNase-seq

DNase-seq analysis can and has been applied to both fresh cells or fixed samples[91,109]. First, nuclei are isolated and permeabilized using mild detergent such as 0.1% Triton X-100, so that DNase I can enter the nucleus efficiently. DNase I is an endonuclease that preferentially introduces double-stranded breaks in accessible chromatin[124]. Second, since DNase I digestion is a continuous process, it is necessary to titrate the amount of DNase I to achieve optimal activity when using a new type of cells, or when using DNase I from a different manufacturer or from a different batch. After digestion, the small DNA fragments (50-100 bp) are purified and size-selected for downstream library construction and sequencing.

Major limitations of the traditional DNase-seq include the large number of cells (millions) as input materials and its tedious and lengthy procedures[63]. Recently, a modified DNase-seq assay (scDNase-seq) has been developed to analyze a small number of or even single cells[61,62]. The scDNase-seq requires only hundreds to thousands of either fresh or fixed cells for a bulk cell assay and less than two days for library construction, without the need of fractionation of DNA fragments in the procedure[60].

Caution must be taken when interpreting DNase-seq results because they show some intrinsic bias in cleavage sites[58,59], which should be considered when interpreting the footprint of transcription factor[57]. Besides, the biological significance of potential regulatory elements identified by this assay, they preferably should be interpreted in combination with other assays[63].

ATAC-seq

Assay for Transposase Accessible Chromatin Sequencing (ATAC-seq) emerged as an alternative assay to investigate accessible chromatin profiles[113]. In this assay, a genetically engineered hyperactive DNA transposase (Tn5) transposes preloaded monovalent mosaic end (ME) adapters to accessible or nucleosome-depleted chromatin regions and tags the DNA with the ME sequence simultaneously[55,56,113]. The target DNA fragments are purified, PCR-amplified, and sequenced by NGS platforms. ATAC-seq detected sequences are highly enriched in DHSs and ATAC-seq has been applied to analyze changes of chromatin accessibility in numerous disease or developmental systems[52–54]. Similar to ATAC-seq, a technique termed as TrAC-looping, which utilizes Tn5 and a bivalent ME adaptor, also detects genome-wide chromatin accessibility in addition to providing genome-wide chromatin interaction information on regulatory regions[51].

ATAC-seq and its variants[122,123] is a sensitive assay that works well on low-input samples(for example 500-50,000 cells) and requires a simplified library preparation procedure due to the simultaneous chromatin fragmentation and tagging. Although, the original ATAC-seq protocol works most efficiently with fresh cells and slowly cooled cryopreserved cells, it is possible to generate high signal-to-background profiles from snap-frozen pellets using the improved Omni-ATAC protocol[122].

Some limitations of ATAC-seq are related to the intrin-

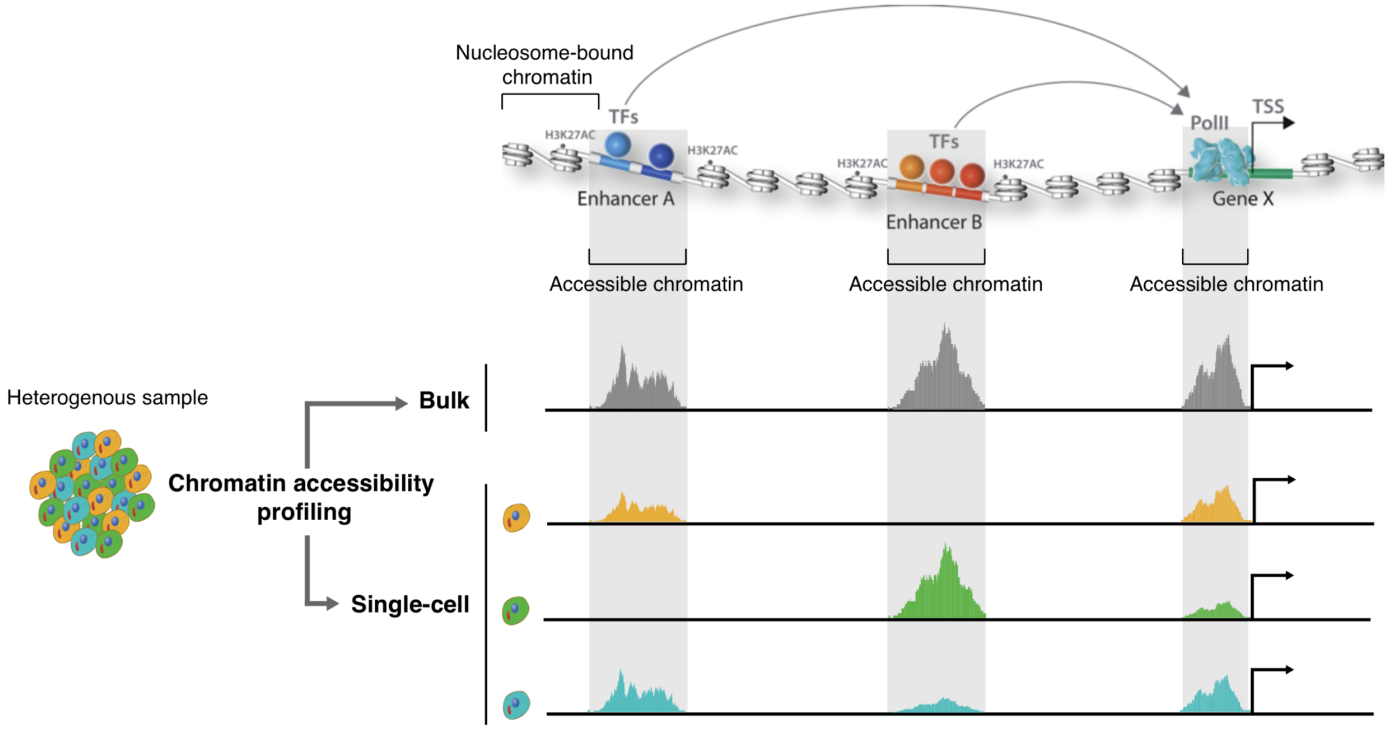

Figure 1: Broad overview of chromatin accessibility profiling approaches. Regulatory elements and linker regions between positioned nucleosomes are characterized by increased susceptibility to enzymatic cleavage/modification, which property is the basis of all chromatin accessibility profiling methods, as it allows for the specific enrichment/labeling of such regions. Chromatin accessibility can be profiled both in bulk as well as at the single-cell level, by applying various strategies to index sequencing library fragments according to their cell of origin. Accessibility can also be profiled at the single-molecule level.

sic properties of Tn5: (1) it shows steric hindrance and sequence bias in chromatin tagmentation [49,50,55], which would be a challenge for the mapping resolution on both chromatin accessibility and transcription factors footprint. (2) The contamination from mitochondrial DNA can also increase the sequencing costs, especially for some cell types which have large amount of mitochondrial DNA. However, this contamination can be significantly reduced either by improved lysis condition (as is the case in Omni-ATAC [122]) or by applying the clustered regularly interspaced short palindromic repeats (CRISPR) technology to cleavage the mitochondrial ribosomal DNA prior to the experiment [48,54].

ATAC-seq derivatives

The versatility provided bu the simultaneous fragmentation and adapter attachment in the ATAC procedure has allowed the development of several derivatives of the assay, which employ modified versions of the Tn5 adapter sequences.

One deficiency of the original procedure is that half of all fragments are lost due to the fact that they contain two adapter sequences of the same kind (i.e. AA and BB, while it is only AB fragments that can successfully PCR amplify). The THS-seq version of ATAC-seq attempts to rescue the other half of fragments by utilizing a T7 RNA Polymerase linear amplification protocol [5].

Several protocols now exist for simultaneous profiling of accessibility and methylation (EpiMethylTag [4], methyl-ATAC-seq [3], and ATAC-Me [2]), while repli-ATAC-seq measures accessibility on newly synthesized chromatin fibers [1].

MNase-seq

Nucleosome position and occupancy in the genome play key roles in chromatin accessibility. MNase is an endo-exonuclease that cleaves the DNA regions without nucleosome protection and leaves the nucleosome core particles undigested, which can be purified, ligated to adaptors, PCR-amplified and sequenced by NGS to reveal genome-wide nucleosome positions (MNase-seq [78]).

In MNase-seq, 10,000 to 100,000 of either fresh or formaldehyde crosslinked cells are used for library construction. Digestion of chromatin by MNase typically results in a nucleosome ladder consisting of mononucleosome, dinucleosome, trinucleosome etc., depending on the concentration of MNase in the reaction. The optimal range of digestion usually leads to about 70-80% mononucleosomes and 20-30% higher nucleosome ladders [78].

MNase-seq has been applied to investigate the dynamics of nucleosome landscape and their function in transcrip-

tional regulation[47]. However, since the nucleosome position and occupancy revealed by MNase-seq are based on the average profile of a large number of cells, caution should be taken when interpreting the results, particularly at inactive chromatin regions[46].

Chemical mapping of nucleosome positioning

While MNase is a powerful and straightforward method for mapping nucleosome positions, it suffers in accuracy from the random and imprecise nature of digestion. To address these limitations, chemical approaches for direct mapping of nucleosome positions have been developed. The first such method is based on replacing endogenous histone H4 genes with a H4S47C protein variant. The cysteine in position 47 is located close to the nucleosome center position and can be chemically modified and, using copper and hydrogen peroxide catalysis, used to trigger the cleavage of the DNA backbone close to it[38]. This method was first used to precisely map nucleosome positions in the budding yeast *S. cerevisiae*[37], and more recently in mouse embryonic stem cells[36], though its application is somewhat limited in more complex eukaryotes by the large number of copies of histone genes.

A more recent conceptually similar approach relies on the H3Q85C mutation, which generates cleavage at positions close to the nucleosome flanks[35]

Single-molecule chromatin accessibility profiling

An important emerging class of methods aim to map chromatin accessibility within single molecules. The major advantage of such an approach is that it does not rely on enrichment and provides information about the distribution of accessibility states within the population of chromatin fibers. The assays in this class rely on methyltransferase enzymes that preferentially modify accessible DNA. For years, the only readout that such methods could rely on was bisulfite conversion of DNA followed by Sanger (for localized analysis of particular loci) and later Illumina (for both local and genome-wide coverage) sequencing, which also dictated the enzymes used to modify DNA. The first genome-wide assay of this kind was NOMe-seq[110], which uses a m5C methyltransferase that modifies cytosines in a GpC context.

As mammalian genomes, as well as the genomes of many other eukaryotes contain abundant endogenous methylation in CpG context, and no non-specific m5C methyltransferases are available, this is the only modification that can be used in mammals. This presents a serious limitation, as GpC nucleotides are rare in the genome, only found once every 20 to 30 bp, with much larger stretches of sequence with no informative positions at all being quite common. However, in species such as yeast and *Drosophila*, which lack endogenous methylation, a combination of both a GpC and a CpG methyltransferase can be used, which increases assay resolution down to ∼10 bp, a method termed dSMF[111] (digital Single Molecule Footprinting). This approach has proven to be very powerful in enumerating the distinct functional states of individual regulatory elements[111], down to the ability to footprint the occupancy of individual transcription factors CITE RECENT PAPER.

Still, there are limitations – such footprinting is again only possible where there are sufficiently many informative positions, and these are inherently rare when relying on m5C modification in dinucleotide context, and it only provides information about the state of individual molecules within at most 600 bp (the current limit of combined paired-end read length for Illumina sequencing).

The latter issue has been resolved by the advent of long-read sequencing platforms such as PacBio and Oxford Nanopore, which are capable of reading modified bases directly within individual long molecules, though with significantly decreased accuracy compared to bisulfite sequencing. The nanoNOMe[43]/MeSMLR-seq[42] assay uses GpC methylation and nanopore sequencing to map accessibility on a multikilobase scale, though it is still limited in its resolution by available informative positions.

That limitation has been overcome by taking advantage of the ability of long read platforms to read any modification, and the use of non-specific methyltransferases, such the m6A depositing enzyme EcoGII combined with nanopore or PacBio sequencing, either on total genomic DNA (SMAC-seq[39,41]) or in combination with a phasing MNase digestion step (SAMOSA[40]). The large number of informative positions allows for fine-scale footprinting almost everywhere in the genome, subject to the limitations imposed by the higher error rate of single-molecule sequencing.

Other methods

A variety of other methods have been developed to probe accessibility.

An orthogonal approach to obtain absolute occupancy/protection values along the genome, though not at the single molecule level, is based on the susceptibility of accessible DNA to restriction enzyme cleavage[31,32]. It has been adapted to a high-throughput sequencing format in the form of assays such as NA-seq[28], RED-seq[27], qDA-seq[33], and ORE-seq[34], and has been used to estimate absolute accessibility levels in yeast and mammalian genomes.

Nucleosome positioning has now also been probed using long-read methods too, which allow the mapping of the ends of larger nucleosome arrays rather than the single, di-, or at most trinucleosomes measurable with short reads.

NicE-seq[30] uses a nicking enzyme to probe accessible DNA.

FAIRE-seq[73,112] was based on the preferential release of accessible chromatin during sonication of crosslinked cells. Protect-seq[29] was recently developed to assay the inverse of accessible chromatin, strongly heterochromatinized genomic regions, based on their resistance to nuclease digestion.

DIVA[25,26] utilizes the preferential viral insertion into accessible DNA to map open chromatin regions.

CATaDa[24] labels open chromatin using ectopic expression of the *E. coli* Dam methyltransferase.

Finally, reactive small molecules have also been applied to probe the fine-grained features of accessibility, such as DMS (DMS-seq[22]) and MPE (MPE-seq[21]).

Single-cell chromatin accessibility profiling

Innovation in barcoding and microfluidic strategies have recently enabled high-throughput biochemical profiling of chromatin accessibility at single-cell resolution, including scDNase-seq[62], scMNase-seq[46]) and scATAC-seq XXXXCITEXXX. Of these protocols, scATAC-seq has emerged as a popular and relatively simple approach to profile chromatin accessibility across hundreds to thousands of individual cells. Current scATAC-seq methods rely on either microfluidic or fluorescence cytometrical/plate-based partitioning to uniquely label nuclei in isolation. Procedures characteristic to both flavors of scATAC-seq, as well as consideration for experimental design, are described below:

Microfluidic scATAC-seq

Droplet-based single-cell partitioning via microfluidic devices has emerged as a powerful approach for single-cell data generation owing to its reproducibly and relative ease of use.

The initial version of scATAC-seq employed the Fluidigm platform[20,114], which allowed the profiling a few hundred cells. Higher throughput platforms have emerged since then and become widely used, such from 10X Genomics[18] (Chromium Next Gem Single Cell ATAC-seq Library Kit; 1000176) and BioRad[19] (SureCell ATAC-seq Library Preparation Kit; 17004620), provide all required reagents necessary to produce scATAC-seq libraries. However, these commercial applications require the acquisition of proprietary robotic sample processing devices (Chromium Controller, 10X Genomics and ddSEQ single-cell isolator, BioRad) that are non-standard in most laboratories.

Microfluidic-based scATAC-seq methods are generally composed of four major steps. First, Tn5 adapter integration is performed on the bulk nuclei suspension, similar to traditional ATAC-seq. Second, transposed nuclei are loaded onto an aqueous channel with PCR reagents and proprietary buffers and mixed with gel-beads containing distinct barcodes. To encapsulate individual nuclei in picolitre reactions with a single gel-bead, the aqueous flow is restricted to channels measuring ∼55 uM in width. Droplets are produced by exposing the aqueous flow to a continuous stream of oil. To assure that the vast majority of droplets contain one or no nucleus, nNuclei droplet loading follows a Poisson distribution and nuclei ar thusloaded at low concentrations. Third, barcoded sequences with P5 adapters and tail sequences complementary to Tn5-inserted adapters are released from gel-beads following droplet generation; enabling PCR amplification of accessible chromatin fragments within each droplet in isolation. Finally, the droplet-oil mixture is emulsified, purified with magnetic beads, and subjected to bulk PCR to attach sequencing indices and P7 sequences.

Plate-based scATAC-seq

An alternative to microfluidics approach is to physically separate cells into the wells of plates. Straightforward 96- and 384-well scATAC-seq protocols have been published[44], however, their throughput remains limited by the low number of wells available. The adaptation of scATAC-seq to the ICELL8 Single Cell System (Takara Bio), which has 5084 nanoliter wells, in the form of μATAC-seq[45], increased the throughput of the assay to a few thousand cells.

Combinatorial indexing (sciATAC-seq)

Higher throughput can be achieved using combinatorial indexing, or sciATAC-seq[115,119,120]. In contrast to microfluidic approaches, sciATAC-seq via combinatorial indexing can be performed with access to standard instruments and reagents found on the campuses of most research institutions, with the exception that it requires homemade Tn5. In lieu of micrometer width channels to restrict barcoding and PCR reactions to individual cells, combinatorial indexing leverages fluorescence-activated nuclei sorting (FANS) to dispense low numbers of nuclei to multiplexed well plates. First, nuclei are stained with DAPI and 2500 nuclei are sorted in each well of a 96-well plate. ATAC-seq with 96 uniquely indexed Tn5 transposomes is then performed such that nuclei from each well are tagged by a distinct combination of adapters. Nuclei are then pooled and redispensed onto a new 96-well plate with each well containing up to 25 nuclei and a unique secondary barcode attached to Illumina-compatible adapters. Two rounds of nuclei shuffling and barcode attachment with 96 uniquely-indexed Tn5 transposomes allows for unique tagging of up to 9,216 nuclei.

Multiple rounds of barcoding at also possible, utilizing ligation of barcodes to transposed fragments[15–17], vastly increasing potential throughput.

Another approach for increasing throughput is to combine upstream transposition of barcoded Tn5 with a droplet-based scATAC platfrom such as 10X or BioRad, in the form of droplet combinatorial indexing, or dsciATAC[19].

Single-cell multiomics

The development of single-cell multiomics assays, which probe multiple molecular characteristics of individual cells at the same time, is major emerging area of single cell genomics, and chromatin accessibility is a key component of most such assays.

Multiple protocols have been published recently from simultaneous single cell ATAC-seq and transcriptome profil-

ing. These include sci-CAR[14], Paired-seq[16], and SHARE-seq[15], which are all based on combinatorial indexing to achieve high-throughput, as well as droplet-based methods, such as SNARE-seq[13].

Single-cell versions of the NOMe-seq assay (scNOMe-seq[11]/COOL-seq[12])allow simutaneous measurement of chromatin accessibility and DNA methylation in single cells, and have also been further extended in combination with transcriptome measurements in the form of assays such as scNMT-seq[10] and snmC2T-seq[9].

Perturb-ATAC has been developed to measure the effect of various CRISPR perturbations on chromatin accessibility at the single-cell level using libraries of sgRNA[8]

Protocols are also being developed for simultaneously measuring protein epitopes and accessibility at the single-cell level (e.g. Pi-ATAC[6])

Finally, spatially resolved single-cell chromatin accessibility measurements are emerging as an important area of method development (e.g. sciMAP-ATAC[7]).

Experimental design

Similar to other sequencing-based profiling methods, scATAC-seq is susceptible to batch effects that can obscure biological variation. Careful attention to experimental design is central to mitigating batch and other sources of technical variation that strongly depend on the goals of the researcher. For example, in atlas and test versus control studies, a common objective is to contrast regulatory patterns among and within cell-types found in different tissues and organs, or between treatments and control samples. For such cases, scATAC-seq libraries should be constructed in parallel from as many sample types as possible, preferably including biological replicates, permitting resources. Prioritizing sample type diversity in preparations from individual batches aids in the mitigation of technical effects and allows researchers to average environmental and genotype influences across replicates. In contrast, comparison of two scATAC-seq libraries produced from separate preparations and from different samples will be confounded by batch effects, resulting in misleading or even erroneous results due to inflated variance between samples (ref). Furthermore, all libraries apart of the intended analysis should be sequenced on the same flow-cell, as sequencing scATAC-seq libraries on separate runs could also lead to increased technical variation (ref). Computational removal of batch effects from single-cell data has been a major focus of many informatics laboratories and shows promise in correcting mistakes stemming from poorly constructed experimental design (see Results). However, there is currently no accepted method to reliably remove all batch effects while preserving biological variation in the absence of true biological replicates. Thus, in cases where generating and sequencing scATAC-seq libraries in different batches is unavoidable, it is pertinent that the researcher takes note of possible sources of variation among samples.

Results

In general, a chromatin accessibility analysis workflow consists of three main steps (1) pre-processing, (2) peak calling and (3) downstream analysis, in which the later can include differential accessibility analysis, annotation, footprinting, motif enrichment and integration with other omics data. We will discuss each of the steps in more detail and mention commonly used bioinformatic tools. Although there is not yet a golden standard in the field, some general pipelines, such as the ENCODE pipeline for ATAC-seq analysis (`https://www.encodeproject.org/pipelines/ENCPL792NWO/`), exist and propose specific tools and a guided workflow for the analysis of chromatin accessibility data. These steps differ for bulk, single-cell and single-molecule analysis and, they are thus discussed separately.

Bulk accessibility analysis

Pre-processing and QC

Like most high-throughput sequencing data, pre-alignment quality control is recommended for chromatin accessibility data and can for instance be done using FastQC (ref) to examine sequencing quality, GC bias and overrepresented sequences. Next, sequencing adaptors should be removed using tools such as cutadapt (ref), trimmonmatic (ref) and fastq-mcf (ref), which require the input of known Illumina adaptor sequences. Depending on the experimental techniques and when paired-end reads are available, a size selection can be performed at this point. For instance, removal of multi-nucleosomal reads is advised for MNase-seq data, and for the double-hit DNase-seq protocol an additional in silico filtering for fragment inserts between 50-100 bp for TFs binding site detection can be performed on top of the gel-based or SPRI-based experimental size selection (Cooper et al., 2017: https://doi.org/10.1038/nprot.2017.099; He et al., 2014). The trimmed and filtered reads are mapped to the organism-specific reference genome. The most widely used aligners for chromatin accessibility data are Bowtie2 (used in the ENCODE ATAC-seq pipeline) (ref), MEM (ref) or STAR (used in CellRanger [check])(ref). Following alignment, some additional filtering steps are advised to discard reads with low mapping quality or multi-mapped reads, PCR-duplicated reads, ENCODE blacklisted regions (Amemiya et al., 2019) and mitochondrial reads (specifically important for ATAC-seq data in which these can make up as high as 75% of the total amount of mapped reads when using the original protocol).

An additional quality control step is recommended at this point by visualizing accumulated read abundance around transcription start sites, which are generally highly accessible (ref). In addition, visually inspecting the distribution of reads across the genome using genome browsers such as IGV (ref) or UCSC (ref) can further increase insight in the quality of the samples.

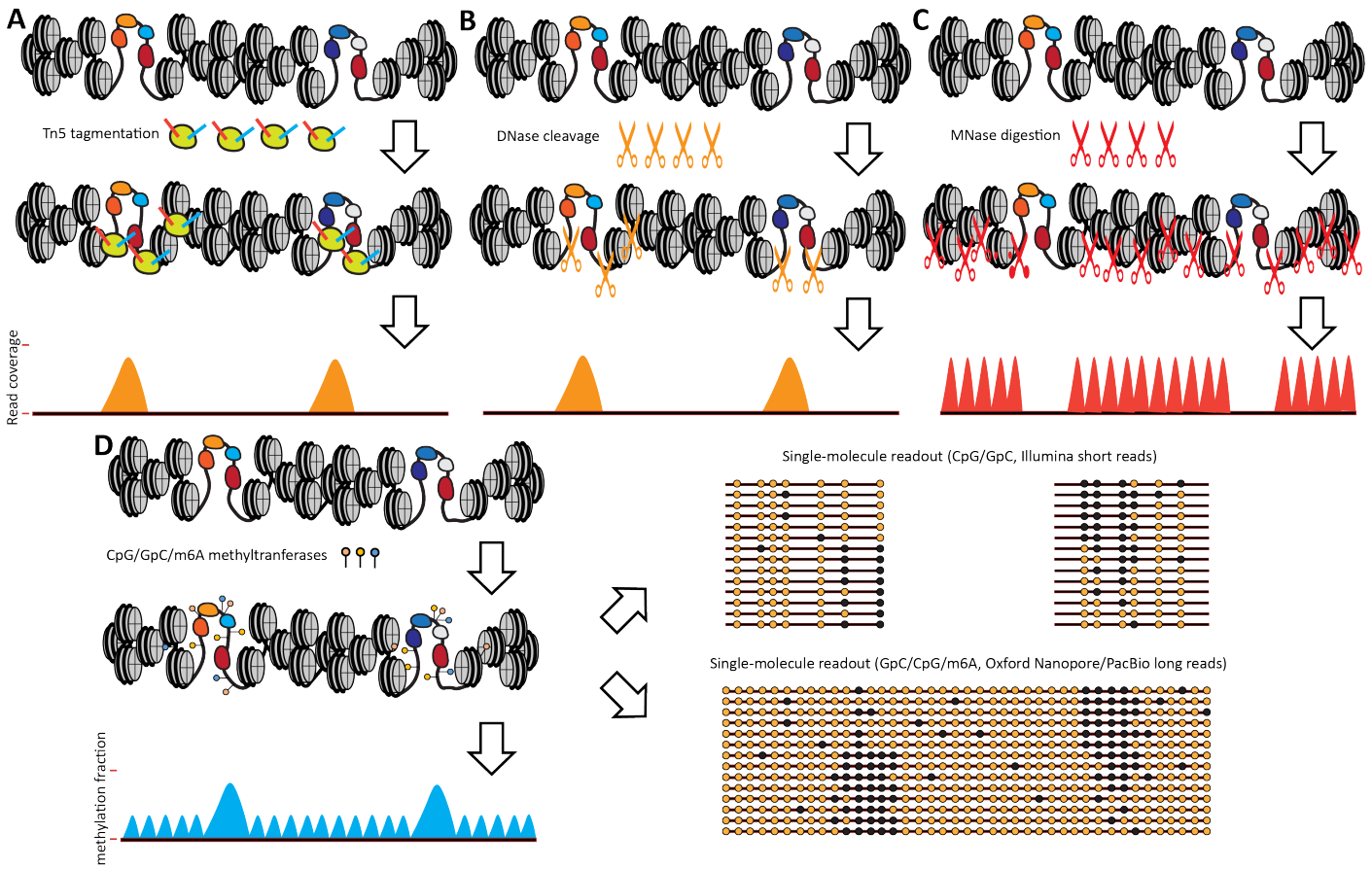

Figure 2: Primary experimental approaches for measuring chromatin accessibility and nucleosome positioning. a) In ATAC-seq, hyperactive version of the Tn5 transposase is used to preferentially insert into accessible chromatin while simultaneously attaching adapters to the resulting fragments that can be used to directly amplify sequencing libraries, b) In DNase-seq, the DNase I enzyme is used to preferentially cleave accessible chromatin, generating fragments that can then be amplified into sequencing libraries. Both ATAC-seq and DNase-seq generate peaks in read coverage over accessible regions in the genome. c) In MNase-seq, the MNAse enzyme is used to digest unprotected DNA, leaving intact fragments protected by protein occupancy (primarily nucleosomes). These fragments are then amplified, resulting in increased read coverage over positioned nucleosomes d) Methyltransferase-based approaches, such as NOMe-seq, dSMF, SMAC-seq, nanoNOMe/MeSLMR-seq and SAMOSA, rely on the labeling of accessible DNA within open chromatin regions and over nucleosome linkers with DNA methylation modifications. These modifications can be m5C methylation in GpC and CpG contexts and also m^6A methylation. Bisulfite conversion or the EM-seq assay can be used to convert fragmented DNA into Illumina-compatible libraries, resulting in short-range and sparse-coverage single-molecule footprints. Alternatively, long-read sequencing, which can also directly read m^6A methylation and take advantage of its much higher density in the genome, can be used, resulting in multikilobase-scale single-molecule footprints. Methyltransferase-based approaches tend to provide a simultaneous readout of both nucleosome positioning and open chromatin regions, appearing as small "bumps" in the methylated fraction of bases over linker regions and larger peaks over regulatory elements, respectively.

Peak calling

Following initial read processing and quality control comes one of the crucial steps in chromatin accessibility data analysis, namely defining so-called peaks or locations with a high accumulation of reads. These peaks form the basic units in most of the downstream analyses. The most widely used tool for peak calling is MACS, which is also the default in the ENOCDE ATAC-seq pipeline (Zhang et al., 2008). MACS is a model-based algorithm originally designed for ChIP-seq data analysis, and implements a dynamic Poisson distribution to capture local background biases in the genome and to effectively detect peaks (Zhang et al., 2008). Other general (e.g. ZINBA (ref)) or more technology-specific peak callers exist, e.g. HMMRATAC for ATAC-seq; F-seq and Hotspot for DNase-seq and ATAC-seq. Note that MNase-seq is actually an orthogonal as-

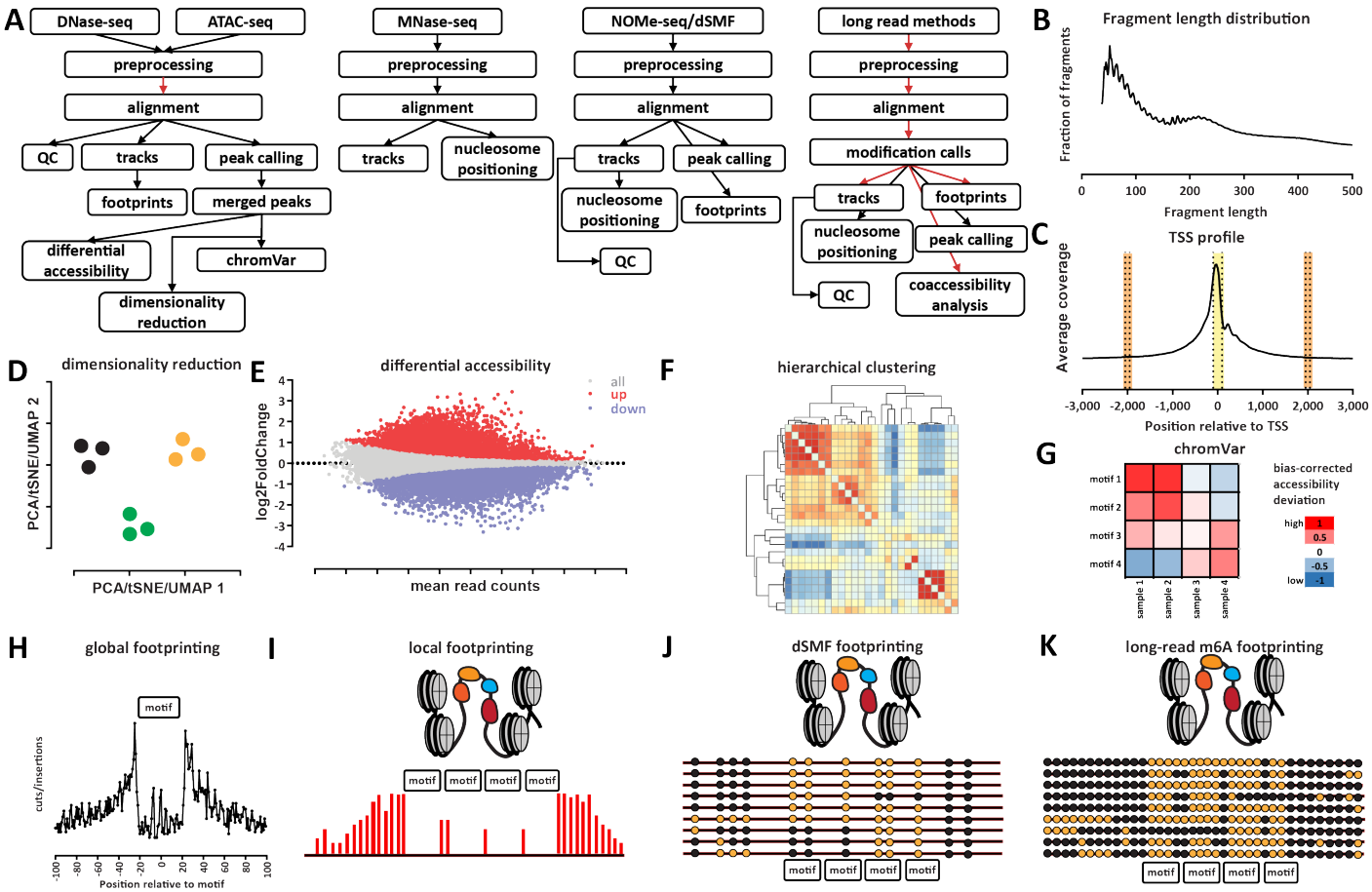

Figure 3: Overview of common bulk chromatin accessibility measurement processing and analysis tasks. a) Outline of key steps in processing ATAC/DNase-seq, MNase-seq, NOMe-seq/dSMF and SMAC-seq and other long-read datasets; b) Fragment distribution is a key informative characteristic of ATAC-seq (shown here), DNase-seq and MNase-seq datasets; c) Enrichment around TSSs is a useful ATAC-seq/DNase-seq quality control metric to assess the degree of global enrichment over open chromatin regions in the genome that is independent of peak calling or other arbitrary thresholds (shown here for an ATAC-seq dataset). It can be formalized in a single number as the ratio of signal immediately around the TSS versus signal in the flanks ±2kbp away. Analogous plots serve the same function for methylation-based methods; d) Dimensionality reduction techniques, such as PCA, *t*-SNE and UMAP, allow for intuitive visualization of the global similarities between datasets; e) Identifying peaks differentially accessible between two conditions is a common task in the analysis of ATAC-seq/DNase-seq datasets; f) Clustering analysis charts the relationships between datasets in a data matrix; g) ChromVar [121] analysis identifies TF motifs with elevated/decreased accessibility across samples; h-k) Transcription factor footprinting. TF footprints are most often measured globally, in the aggregate over a set of motif instances and across conditions (h). Deep DNase-seq datasets also enable local footprint analysis at the level of individual motif instances (i). Short- and long-read single-molecule footprinting techniques enable the quantitative analysis of distinct footprint states (j–k).

say compared to the other discussed chromatin accessibility profiling methods as it measures nucleosome-occupied regions. It is therefore the method of choice to map nucleosome positions genome-wide, for which specific tools have been developed such as GeneTrack (ref) and DANPOS (Chen et al., 2013). Note that ATAC-seq also lends itself for nucleosome-positioning by for instance using the tool NucleoATAC (Schep et al., 2015). An important parameter to consider during the peak calling step is the signal threshold, which influences the sensitivity and specificity of peak retrieval (Koohy et al., 2014). The default minimum FDR cutoff of 0.05 for MACS has been shown to be optimal for a range of DNase-seq datasets (Koohy et al., 2014).

As datasets often comprise different samples, the construction of a common set of features is crucial in order to be able to compare samples to each other in downstream steps. Mostly, a consensus peak file comprising a set of merged peaks across the samples is used. The ENCODE pipeline

provides a possible workflow with merge and filter steps for this purpose (ref), although other tools serve the same purpose (e.g. consensusSeekeR (Samb et al., 2015)) Alternatively, a pre-defined set of regions or a binned genome can be used as features in downstream analyses (Bravo et al., 2020; Cusanovich et al., 2018).

Lastly, to ensure reproducibility in the data, ENCODE guidelines recommend that each ATAC-seq experiment should have two or more biological replicates and that replicate concordance should be checked by calculating Irreproducible Discovery Rate (IDR) values (Qunua et al., 2011). An additional quality control step is to calculate measures that represent the signal-to-noise ratio, namely the fraction of reads in called peaks (FRiP score) for ATAC-seq, which should preferably be greater than 0.3 (or at least greater than 0.2), and the signal proportion of tags (SPOT score) for DNase-seq, which should exceed 0.4 (i.e. 40% of mapped reads within DHSs) (ENCODE projects; Roadmap Epigenomics).

Downstream analysis

Often, the experimental set-up of chromatin accessibility profiling methods includes several samples, such as treatment versus control or cells undergoing differentiation, for one wishes to study the changes in chromatin accessibility profiles. Therefore, defining the set of peaks that is differentially accessible between conditions is often a central question. Most differentially accessibility tools, including DiffBind (ref), HOMER (ref), DBChIP (ref), use as input a read count matrix on a feature file (e.g. consensus peaks) across the different conditions. These methods rely on bioinformatic tools that were designed for differential expression analysis of RNA-seq data, such as DESeq2 (ref) and edgeR (ref). Differential peak calling can also be done using MACS in which mapped bam files of treatment and control samples are provided rather than a count matrix (ref).

These differential accessibility tools are well-suited for pairwise comparison of chromatin accessibility. However, differential accessibility analysis becomes more tricky when comparing three or more samples. In this case, a one-versus-the-rest approach using aforementioned tools can reveal regulatory regions that are specific for a condition or cell type, however, a clustering analysis is more versatile. In general, clustering can both be used to query which samples in a dataset resemble eachother, as well as to define groups of specifically accessible regions per sample class. Both hierarchical clustering and k-means clustering have been applied to a (Pearson or Spearman correlation matrix on) a normalized (reads per million mapped reads or RPM) accessibility count matrix using a consensus peak set (; other refs). Such clustering algorithms are for instance implemented in the deepTools package (ref). The differentially accessible regions are often visualized in a heatmap representing the accessibility of the regions within each cluster across the compared samples. Other researchers have drawn inspiration from tools designed for clustering of regions in single-cell epigenomics data using factor analysis and unsupervised learning (ref). For instance, topic modelling or non-negative matrix factorization, in which a high-dimensional dataset is approximated by a reduced number of representative components, can be applied to bulk datasets after a bootstrapping/subsampling procedure in which simulated single-cells are created from bulk samples (ref).

To gain biological insight in the sets of cell type specific regions identified via differential accessibility analysis, peak annotation tools such as GREAT, ChIPseeker, ChIPpeakAnno, [clusterProfiler?] are used to couple peaks to the nearest gene (and to report associated Gene Ontology terms) as well as to annotated peaks based on their location relative to genes (for instance as promoter, intron or intergenic). In addition, chromatin segmentation approaches such as ChromHMM (Ernst and Kellis 2017), EpicSeg (Mammana and Chung, 2015) and Segway (Hoffman et al., 2012) are used for genome-wide classification of genomic regions based on epigenomic marks (mostly based on histone modification ChIP-seq) into chromatin states, such as active promoter or weak/poised enhancer per cell type. These annotations can be useful to aid interpretation of gained or lost accessible regions in a study.

As combinatorial binding of TFs to accessible regulatory regions forms the basis of gene regulation, one of the major downstream analysis steps is unravelling which TFs are bound to the set of cell type-specific or differentially accessible regions. Since TFs recognize and bind to TF-specific DNA sequences, we can leverage the enrichment of so-called motifs in a set of sequences to detect potential binding of TFs. Two major groups of motif analysis tools exist. The first class of tools, e.g. HOMER, MEME and i-cisTarget, use pre-defined databases (e.g. JASPAR, CIS-BP, TRANSFAC and HOCOMOCO) containing motif sequences, modelled in the form of position-weight matrices (PWMs). These approaches use motif scanner tools to identify enriched cognate TF binding sequences in a provided peak set via the comparison with a tool- or user-provided background set of sequences. On the contrary, other bioinformatic tools, such as RSAT, MEME, Weeder and HOMER, perform de novo motif discovery, allowing an unsupervised identification of TF binding motifs. Next to these classical de novo motif discovery tools, recently, with the advent of deep learning algorithms in regulatory genomics, several convolutional neural network models have been applied to this purpose, including DeepATAC (Hiranuma et al., 2017), DeepLIFT (Shrikumar et al., 2019), DeepMEL (Minnoye et al., 2019). Often, these models not only de novo capture important motifs across the training regions but are also able to predict their importance at single-nucleotide resolution within the sequences. Note that most of the motif discovery tools require a set of regions-of-interest as input that was priorly identified by for instance differential accessibility or clustering analysis. In contrast, MEDEA extracts cell-type-specific peaks from just an in-

put sample using a panel of peaks from reference cell types (e.g. ENCODE-DREAM) prior to a TF motif enrichment analysis (Mariani et al., 2020). Altogether motif detection on a set of specifically accessible regulatory regions allows to decode the genome sequences and may reveal possible master regulators that bind to these regions.

One can also go further and aim to detect the precise location of TF binding events from chromatin accessibility data via TF footprinting. Footprints are small regions (8-30bp) that display relative protection from cleavage due to binding of a TF, and thus correspond to dips in the accessibility peak (Baek and Sung, 2016;.). DNase I has been and is still the preferred footprinting reagent. A study from Schwessinger et al., 2017 showed that ATAC-seq footprinting was less accurate than DNase-seq footprints, which might be attributed to the large size of the Tn5 dimer and Tn5-specific cleavage biases that are not accounted for in DNase-seq-designed footprinting algorithms (Buenrostro et al., 2013; Li et al., 2019). Analytic genomic footprinting approaches either de novo annotate DNase I footprints (e.g. the Wellington algorithm (ref), HINT (ref), DBFP (ref) and DNase2TF (PMID 25242143)); or determine TF occupancy at specific genomic location (e.g. CENTIPEDE (ref) and FLR, (ref)) (Vierstra and Stamatoyannopoulos, 2016). Nevertheless to popularity of DNase-seq data for footprinting, footprinting analysis on ATAC-seq data has also been attempted by several groups with success, for instance in the initial ATAC-seq publication (Buenrostro et al., 2013), using DeFCoM (Quach and Furey, 2017) or ATAC-seq-specific footprinting algorithms such as HINT-ATAC that consider ATAC-seq artefacts (Li et al., 2019). Note that TF footprinting comes with some limitations as it requires extremely deep sequencing, ideally at least 200 million uniquely mapped reads from a DNase-seq experiment (Vierstra and Stamatoyannopoulos, 2016), and it is biased by short residence times for some TFs (ref) and by intrinsic sequence preferences of DNase I (Sung et al., 2016).

Single-molecule accessibility analysis

Single-cell chromatin accessibility data analysis

Pre-processing and QC

Single-cell chromatin accessibility data requires similar upstream processing steps as bulk data, including alignment, feature definition and the generation of a count matrix. However due to the substantial scale and sparsity of the feature-by-cell matrix, specialized bioinformatic tools have been developed, mostly for scATAC-seq data, to handle these assay-specific challenges (refs for all tools). One major point in which these tools differ is the way they define features [e.g. peaks from bulk or aggregated single-cell data (chromVar, Cicero, cisTopic, Gene Scoring, scABC, Scasat), pseudo-bulk samples (Cusanovich2018) or fixed size bins (Cusanovich2018, SnapATAC))] and what the count features represent [e.g. counting reads in peaks (cisTopic, Cusanovich2018, scABC, Scasat), counting (gapped) k-mers under peaks or around transposase cut sites (BROCKMAN, chromVAR) or counting reads overlapping TF motifs in peaks or genome-wide (chromVar, SCRAT)] (Chen et al., 2019). Important follow-up steps are transformation (e.g. by binarization) and dimensionality reduction of the feature-by-cell matrix to visualize the cells into a 2D- or 3D-space and to perform further downstream analyses such as clustering to uncover the different populations in the sample and their specifically accessible regions. Recently, 10 computational methods for the analysis of scATAC-seq data have been benchmarked by Chen et al., demonstrating that SnapATAC (ref), Cusanovich2018 (ref), and cisTopic (ref) performed best in distinguishing cell populations in both synthetic and real datasets (Chen et al., 2019). Note that compared to scRNA-seq frameworks, there are no designated tools that correct for batch effects in scATAC-seq data, but batch correction is performed inexplicitly during the processing steps such as during feature selection or dimensionality reduction (Baek and Lee, 2020). Batch correction tools designed for scRNA-seq data can be used with precautions to not remove biological variance. Batch effect removal becomes especially important when combining multiple runs into atlases or when integration with scRNA-seq data, for which BBKNN, Scanorama and scVI performed best in a recent benchmark (Luecken et al., 2020).

As the complexity of a system or disease exists across all molecular layers, computationally integrating multiple omics modalities on the system of interest holds the promise to gain a systems biology view and to reconstruct gene regulatory networks. Especially the integration of chromatin accessibility profiles with ChIP-seq and RNA-seq data are of interest. As TFBS enrichment within regulatory regions may elude to TF binding, correlation with TF ChIP-seq tracks (cisTarget/ other methods??); or enrichment/overlap of TF ChIP-seq signal/peaks within accessible regions can validate the predicted target sites. For the reconstruction of regulatory networks, specifically the integration of epigenomics and transcriptomics is of interest as this may predict links between accessible regulatory regions and target genes (Ackermann et al., 2016). An example from the single-cell field is the study by Cao et al. where they used a least absolute shrinkage and selection operator (LASSO) model to correlate a genes expression level with the accessibility of all peak within 100kB around its TSS, linking 1,260 distal regions to 321 potential target genes, which improved predictions of gene expression based on accessibility profiles by a fourfold as compared to only using chromatin accessibility at promoters (Cao et al., 2018).

Applications

Chromatin accessibility profiling is widely useful for applications in biology and biomedicine, ranging from the analysis of gene regulation and cellular states (section 1 and 2 below) over the dissection of healthy and diseased tissues and organs (section 3 and 4) to the investigation of pop-

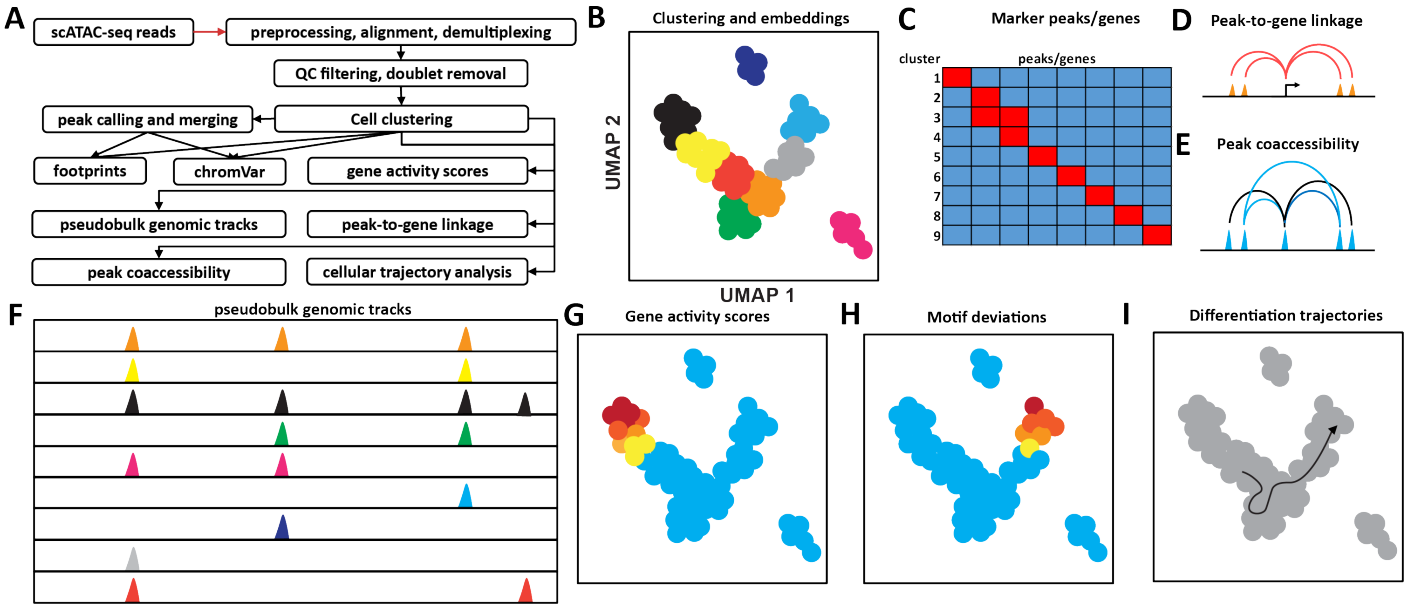

Figure 4: Overview of common scATAC-seq processing and analysis tasks. a) Outline of key steps in processing scATAC-seq datasets; b) Clustering of cell types and UMAP embedding of single cells; c) Identification of marker genes and/or peaks; d) Identification of peak-to-gene links; e) Assessing peak coaccessibility; f) Generation of pseudobulk genome browser tracks for each cell type; g-h) Analysis of gene activity scores (g) and motif deviations (h) within embeddings; i) Differentiation trajectories analysis.

ulations and species (section 5 and 6). These applications profit from the high genomic resolution of chromatin accessibility profiling and from it relative ease and throughput of these assay.

Regulation of chromatin accessibility

¡Enhancers/super-enhancer, 3D structure, GRNs etc., Georgi to write a draft?¿ Chromatin accessibility at regulatory regions such as promoters and enhancers plays an important role for gene regulation, and it is itself controlled by a complex interplay of transcription factors, chromatin remodeling complexes, non-coding RNAs and other factors. Chromatin accessibility profiling has provided important insights into ¡...¿

Cellular dynamics of chromatin accessibility

¡Embryogenesis, stem cells, etc., Georgi to write a draft?¿ Chromatin accessibility captures developmental cell states such as pluripotency and controls how cells respond to stimuli. Changes in chromatin accessibility often concur with changes in the expression of the corresponding genes, although the correlation is far from perfect. For example, pluripotent stem cells retain a high level of chromatin accessibility at key regulatory regions ¡...¿

Chromatin accessibility across cell types and organs

Chromatin accessibility at gene-regulatory regions is highly dynamic during cellular differentiation and organ development (PMID: 21116306, 20110991). Chromatin accessibility profiling has contributed to our understanding of chromatin regulation across a broad range of cells in human and mouse (PMID: 25693563, 30078704) and in specific organs and cell types. The hematopoietic lineage in particular has served as a blueprint for deciphering the role of chromatin accessibility and epigenetic changes in cellular differentiation (PMID: 29364285, 22398613). Application of ATAC-seq and/or ChIP-seq to FACS-purified hematopoietic cell populations established comprehensive maps of regulatory regions and their dynamic changes in the hematopoietic lineage of human and mouse (PMID: 25103404, 30686579, 27526324, 29706549). Detailed investigations of macrophages connected the regulation of these immune cells to their tissue environment (PMID: 25480296, 25480297), while analyses of CD4+ T cells (19144320, 24097267, 29686426) and innate lymphoid cells (PMID: 27156451, 27156452) uncovered a striking degree of plasticity in these immune cell populations. Chromatin regulation in immune cells also contributes to the generation of memory T cells (PMID: 29236683), which are poised to implement effector functions upon re-exposure to pathogens, and to the more limited memory of inflammation in regulatory T cells (PMID: 27499023). Importantly, immune cell memory is not restricted to B cells and T cells, but also

includes monocytes and NK cells (PMID: 32132681), and the regulation of such trained immunity appears to involve tightly regulated changes in the epigenomes of the affected cells (PMID: 25258085, 27863248).

Reaching beyond the hematopoietic lineage, RNA-seq, ATAC-seq and ChIPmentation profiling of epithelial cells, endothelial cells and fibroblasts from 12 different organs uncovered widespread immune gene regulation in these non-hematopoietic, structural cells, as well as an epigenetic potential that appears to pre-program these cells for contributing to pathogen response (DOI: 10.1038/s41586-020-2424-4). Chromatin accessibility has also been studied in neural development (PMID: 31974223, 29307494, 26365491, 29434377) as well as in brain samples of humans (PMID: 31727856, 29945882, 29227469) and non-human primates (PMID: 31980617). Notable applications of chromatin accessibility profiling to other cell types and organs have included the analysis of cardiac development (PMID: 31271750, 30451828), epidermal progenitor cells in skin (30220568), and mammary gland development (PMID: 30174241). Finally, initial single-cell atlases of chromatin accessibility across tissues and organs are emerging (PMID: 30078704, 32231307), which have the potential to discover new cell types and to define the chromatin states of cell types that are difficult to purify or enrich using FACS. In summary, chromatin accessibility profiling has uncovered a transcription-regulatory landscape that is cell-type-specific and organ-specific, and dynamically changed over the course of cellular differentiation and organ development.

Chromatin accessibility in human diseases

Changes in chromatin accessibility have been implicated in multiple diseases, where they reflect disease-linked changes in cell composition, gene regulation and/or epigenetic cell states. Aberrant epigenetic states are ubiquitous in cancer and often linked to the developmental abnormalities of cancer cells (PMID: 30361341). In blood cancers, chromatin accessibility patterns have been shown to reflect the cancers cell-of-origin as well as regulatory changes that appear to contribute to the process of malignant transformation and cancer evolution (PMID: 27526324, 31996669, 29785028, 30503705, 30673601). Changes in chromatin accessibility have been investigated over the course of targeted therapy in patients with chronic lymphocytic leukemia (PMID: 31996669) and combined with chemosensitivity screening to identify promising drug combination therapies (PMID: 30692684). Chromatin accessibility landscapes have also been mapped in solid tumors, including breast cancer (PMID: 31152164), colon cancer (PMID: 22499810, 28169291), glioblastoma (PMID: 30275445, DOI: 10.1101/370726), gastric cancer (27677335), and lung cancer (PMID: 31209061, 27374332). Pediatric cancers tend to carry particularly pronounced epigenetic changes, contrasting with their comparatively low rate of somatic mutations. For example, the EWS-FLI1 fusion oncogene in Ewing sarcoma has been shown to impose de novo enhancers and super-enhancers on the tumor cells (PMID: 25704812, 25453903); and epigenome profiling has uncovered subtype-specific regulatory mechanisms in atypical teratoid rhabdoid tumors (PMID: 27960086) and in Langerhans cell histiocytosis (31345789).

An interesting line of research has investigated the role of the tumour-infiltrating/tumour-associated cells in solid tumors. Epigenetic changes have been implicated in T cell exhaustion in the context of chronic inflammation and the tumor microenvironment (PMID: 27789799, 30778252), which compromises the ability of these T cells to fight the tumor. Immunotherapy, most notably blocking of the PD1/PD-L1 pathway, has been shown to revert some of the epigenetic changes associated with T cell exhaustion (PMID: 27789795, 31375813, 28648661) and is widely useful for the treatment of those solid tumors that have a high degree of immunogenicity (32433532). However, not all exhausted T cells can be rejuvenated by immune checkpoint blockade, as some T cells appear to transition to a fixed epigenetic state that renders them resistant to reprogramming (30778252). In addition to immunotherapy, selective epigenetic reprogramming can be used to alter the T-cell landscape resulting in enhanced treatment efficiency (PMID: 26503055, PMID: 28625481).

Beyond cancer, where chromatin accessibility has been studied most extensively, changes in chromatin accessibility have also been observed in immune diseases such as inflammatory bowel disease (29695774) and rheumatoid arthritis (29765031). Moreover, changes in epigenome and chromatin accessibility profiles have been observed in post-mortem brain tissue from patients with Alzheimers disease (PMID: 30559478), schizophrenia (PMID: 30087329) and autism spectrum disorder (PMID: 27863250). In summary, chromatin accessibility profiling of primary patient samples is already widely used for identifying disease-linked changes in chromatin structure and transcription regulation, and there is substantial scope for new discoveries as researchers move beyond cancer and are investigating regulatory mechanisms in many diseases that have yet received little attention.

Chromatin accessibility variation across populations

Extension of chromatin accessibility assays to populations of diverse genetic backgrounds has proven valuable for advancing our understanding of how sequence variation impacts cis-regulation within a species. A striking 90% of disease-associated variants in humans identified via GWAS localize to gene-distal non-coding loci, obfuscating functional predictions (PMID: 22955828, 22955986, 28039028). Mounting evidence has implicated alteration of gene regulation as a key driver of phenotypic evolution and disease proliferation. Quantitative trait loci (QTL) mapping of molecular traits, such as expression variation (eQTL), provides an attractive approach for deciphering the gene regulatory potential of genetic variants within a popula-

tion. Leveraging a molecular QTL framework, a large-scale DNase-seq panel of 70 lymphoblastoid cell lines from the Yoruba HapMap showed that approximately 50% of chromatin accessibility associated variants coincide with variants associated with expression variation, with the allele conferring increased accessibility generally associated with increased gene expression (PMID: 22307276). This study also provided evidence that sequence alterations underlying cis-elements perturb transcription factor binding affinities, leading to weakened or ablated binding. An analysis of CD4+ T cell chromatin accessibility from 105 healthy donors revealed that only 15% of genetic variants embedded within accessible chromatin regions affect the relative accessibility of the cognate regions (PMID: 29988122). Thus, the majority of genetic variants located within accessible chromatin appear to lack functional consequences. The same study further demonstrated that pairwise correlations of accessible regions (co-accessible regions) readily recapitulates three-dimensional higher-order chromatin interactions as defined by in situ HiC data, suggesting that local chromatin accessibility among pairs of regions are coordinated with higher-order genome structure, particularly within the same topologically-associated domains (TADs). In line with these findings, local chromatin accessibility in a subset of regions were associated with variants located 10s to 100s of kilobases away, reflecting putative interactions. Importantly, integration of population-scale accessibility data captured 10-30% of previously reported autoimmune-associated variants and explained 1-7% of disease heritability. Taken together, population-based analysis of chromatin accessibility provides a powerful approach for dissecting the regulatory potential of genetic variants associated with a trait of interest. Additional studies in other tissues and disease states leveraging single-cell technologies have the potential to systematically map all chromatin accessibility modifying variants in a cell-type specific fashion.

Evolution of chromatin accessibility

The use of chromatin accessibility data has greatly facilitated the identification of causal genetic variants underlying disease and trait variation; however, it is also proving useful to study the evolution of gene regulation and morphological evolution between species. A major advantage of incorporating chromatin accessibility data into these studies is that DNA sequence variation is often too high in intergenic regions to identify cis-regulatory elements using sequence-based alignments alone (ref?). This is especially problematic for studies in plants, where sequence turnover between related plant species is much greater than what is observed between related animal species (ref?). Regardless, sequence-based methods can be used to identify conserved non-coding sequences, which are highly enriched in accessible chromatin. Comparative epigenomics is revealing important clues about the evolution of gene regulation. For example, major morphological transitions, such as the loss of limbs in snakes and eye degeneration in subterranean mammals, have been linked to loss of CREs (ref?). These CREs were discovered using a combination of tissue-specific ATAC-seq and comparative genomics. In another study, chromatin accessibility data in combination with H3K27ac and H3K4me3 was used to identify promoters and enhancers in liver tissue of 20 mammalian species (Villar et al., 2015). It was determined that the rate of sequence variation is much greater for enhancers in comparison to promoters. This was reflected by a lower conservation of enhancers between species, yet, newly evolved enhancers were more likely to be under positive selection in a lineage specific manner. Rapid evolution of cis-regulatory regions has also identified in a comparative epigenomics study of numerous flowering plant species ranging in genome size from 150 Mb to 5,000 Mb (ref?). The frequency of distal chromatin accessible regions was correlated with genome size and their distal location from genes was mostly likely due to transposon and repeat expansion in these plants. The lack of distal CREs in Capsaspora owczarzaki, a unicellular organism sister to other animal species, has led to the hypothesis that distal regulation is a feature of animal multicellularity, however, with the increase in profiles of chromatin accessibility across taxa it seems more likely distal regulation is a consequence of genome size (ref?). Additional comparative epigenomic studies of chromatin accessibility across diverse taxa and of species that represent key nodes in the tree-of-life will further unveil diverse mechanisms in the evolution of gene regulatory mechanisms.

Reproducibility and Resources

The genomics community has been leading the way in creating standards for data information, data quality and data deposition for decades. This reflects that many genome-wide datasets serve as community resources and, as a result, they are repeatedly used and incorporated into future studies by individual labs. These efforts also improve reproducibility, especially for data analyses. To increase the usability of epigenomics data, it must be available in public data repositories and it should include the original unprocessed sequencing data, in addition to processed output files from analyzing the data. For example, it is common practice to include files that contain regions enriched for chromatin modification and or accessibility for ChIP-seq and chromatin accessibility mapping assays, respectively. There are numerous ways to publicly distribute raw and processed epigenome datasets ranging from custom websites from consortia or individual labs to national data archives. At a minimum, epigenome sequencing datasets should be replicated and deposited to a well-funded and stable data archive facility such as the National Center for Biotechnology Information (i.e. Sequence Read Archive or Gene Expression Omnibus), European Nucleotide Archive, DNA Data Bank of Japan, or Dryad Digital Repository to name a few examples. Although not required, it is also useful if data are hosted in publicly accessible genome browsers. This

increases data dissemination and provides a user-friendly tool for scientists not as familiar with computational methods for analyzing data. The deposited data should include supplementary information to facilitate interpretation and reproducibility. These data entry requirements will range from information regarding the sample used to technical aspects of experimental design to the type of instrument used for sequencing. For example, data entry requirements that are useful to addresses issues associated with reproducibility could include sources of possible biological variation (i.e. genotype, sex of samples, age, tissue/organ/cell type) and technical variation (i.e. antibodies lot number, nucleases/integrases lot number, sequencing library procedure, instrument used for sequencing and type of sequencing run). They are also important variables that can be incorporated into data analyses as covariates or to correct for batch effects. Genome assembly and genome annotation versions used in data analyses should also be provided. Altogether, this information is useful to assess reproducibility. Lastly, distribution of custom code and descriptions of computational methods are also paramount to reproducibility. As one example, the ENCODE Consortia has developed extensive open source software that is accompanied with best practices and descriptive details on the rationale for data processing steps, thresholds and quality metrics for data evaluation. In general, software used for data analyses should include the software version and parameter options applied. Custom code should be disseminated through public hosts such as GitHub, as opposed to personal communication with the developer. Efforts to address the biological, experimental and computational variables described above will increase reproducibility in addition to the usability of these data for years to come.

Limitations and optimizations

Outlook

The past decade has seen an explosion in studies examining chromatin accessibility and its variation in different cell types, tissues, organs and organisms. The current and future challenge is to dissect the function of these regulatory regions in relation to other regulatory layers and gene expression. Accessibility alone does not reveal the activity state or the functional properties of the region (whether it acts as a promoter, enhancer, silencer), or which factors are bound to the region or its potential role in other functions such as 3D genome topology or replication origins. Moreover, information on the identity of the target genes, and whether a regulatory region is functionally required for gene expression, is also missing.

Many of these challenges can be overcome by a more holistic multi-omics approach, by profiling multiple molecular layers from the same sample, such as the transcriptome, chromatin modifications and transcription factor occupancy, in addition to chromatin accessibility. A common approach is to run multiple omics methods on fractions of the same sample, using protocols optimized separately for each assay, thus generating comparable datasets. However, running separate assays can introduce batch effects that are difficult to mitigate computationally, which can be a drawback of this strategy.

Chromatin accessibility profiling in single cells has surged dramatically in recent years, in part due to combinatorial indexing (sciATAC-seq) and the recent availability of commercial kits for droplet-based scATAC-seq. We expect further improvements to the assay in the coming years as this trend keeps increasing. In contrast to RNA, which has a high dynamic range, there are only two loci that can be measured simultaneously in a diploid genome by single cell regulatory genomics based methods. As a result, the data is mostly binary and still very sparse due to the low coverage per cell, making the analysis of accessibility and other regulatory features at the single cell level extremely challenging and a certain degree of data aggregation across cells or features is usually required. It is also difficult to estimate the sensitivity of scATAC-seq. Roughly 10-15

Recent advances in single cell methods are pushing technologies to perform multi-omic measurements from the same cell. Multiple methods already exist for the joint profiling of chromatin accessibility and DNA methylation (scNOMe-seq (Pott. 2017, PMID: 28653622)), gene expression (sci-CAR (Cao et al. 2018, PMID: 30166440)) and protein levels (Pi-ATAC (Chen et al. 2018, PMID: 30389926)). Several technical challenges have so far limited the widespread application of these methods. Sample fixation, reaction conditions and other experimental parameters are often not compatible for multiple assays, complicating the optimization of joint protocols. Moreover, the resulting data is limited by the combined sensitivity of the methods, for example running two assays each having a 10

Functionality of accessible chromatin regions can also be probed by perturbation, for example by mutation of key transcription factors. The high degree of cellular heterogeneity in complex systems, such as developing embryos, has limited the usefulness of this approach. However, single cell accessibility profiling could solve this issue, by identifying the impact of the mutations directly in the affected cell types, revealing both changes in regulation as well as alterations in cell fate decisions. Large-scale perturbation and profiling of regulatory networks has been performed in cell culture models by coupling CRISPR screening with scATAC-seq readout (Perturb-ATAC (Rubin et al. 2019, PMID: 30580963)). In more complex systems, where high-throughput targeted mutagenesis is not feasible, natural sequence variation could be exploited as a large-scale perturbation tool. In this context, profiling accessibility both intra- and inter- species can give insights into regulatory variation and functionality, as discussed above.

Finally, a particularly exciting area of future development is the integration of chromatin accessibility profiling with imaging-based approaches. Current protocols involve

tissue dissociation to extract cells or nuclei, which leads to the loss of the native spatial context. ATAC-see (Chen et al. 2016, PMID: 27749837) mitigates this problem by performing the Tn5 reaction in-situ on microscopy slides and using fluorescent adaptors that are compatible with both imaging and sequencing. Further integration of ATAC-seq with high-throughput FISH and other imaging-based methods will lead to new ways of interrogating the genome of complex systems in situ after stimuli and perturbations.

Author contributions

XXX

Acknowledgments

XXX

References

1. Stewart-Morgan KR, Reverón-Gómez N, Groth A. 2019. Transcription Restart Establishes Chromatin Accessibility after DNA Replication. *Mol Cell* **75**(2):284–297.e6.
2. Barnett KR, Decato BE, Scott TJ, Hansen TJ, Chen B, Attalla J, Smith AD, Hodges E. 2020. ATAC-Me Captures Prolonged DNA Methylation of Dynamic Chromatin Accessibility Loci during Cell Fate Transitions. *Mol Cell* **77**(6):1350–1364.e6.
3. Spektor R, Tippens ND, Mimoso CA, Soloway PD. 2019. methyl-ATAC-seq measures DNA methylation at accessible chromatin. *Genome Res* **29**(6):969–977.
4. Lhoumaud P, Sethia G, Izzo F, Sakellaropoulos T, Snetkova V, Vidal S, Badri S, Cornwell M, Di Giammartino DC, Kim KT, Apostolou E, Stadtfeld M, Landau DA, Skok J. 2019. EpiMethylTag: simultaneous detection of ATAC-seq or ChIP-seq signals with DNA methylation. *Genome Biol* **20**(1):248.
5. Sos BC, Fung HL, Gao DR, Osothprarop TF, Kia A, He MM, Zhang K. 2016. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol* **17**:20.
6. Chen X, Litzenburger UM, Wei Y, Schep AN, LaGory EL, Choudhry H, Giaccia AJ, Greenleaf WJ, Chang HY. 2018. Joint single-cell DNA accessibility and protein epitope profiling reveals environmental regulation of epigenomic heterogeneity. *Nat Commun* **9**(1):4590.
7. Thornton CA, Mulqueen RM, Torkenczy KA, Lowenstein EG, Fields AJ, Steemers FJ, Wright KM, Adey AC. 2020. Spatially-mapped single-cell chromatin accessibility *bioRxiv* 815720
8. Rubin AJ, Parker KR, Satpathy AT, Qi Y, Wu B, Ong AJ, Mumbach MR, Ji AL, Kim DS, Cho SW, Zarnegar BJ, Greenleaf WJ, Chang HY, Khavari PA. 2019. Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* **176**(1–2):361–376.e17.
9. Luo C, Liu H, Xie F, Armand EJ, Siletti K, Bakken TE, Fang B, Doyle WI, Hodge RD, Hu L, Wang BA, Zhang Z, Preissl S, Lee DS, Zhou J, Niu SY, Castanon R, Bartlett A, Rivkin A, Wang X, Lucero J, Nery JR, Davis DA, Mash DC, Dixon JR, Linnarsson S, Lein E, Behrens MM, Ren B, Mukamel EA, Ecker JR. 2019. Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants *bioRxiv* 2019.12.11.873398
10. Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda–Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC, Stegle O, Reik W. 2018. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* **9**(1):781.
11. Pott S. 2017. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *Elife* **6**. pii: e23203.
12. Guo F, Li L, Li J, Wu X, Hu B, Zhu P, Wen L, Tang F. 2017. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res* **27**(8):967–988.
13. Chen S, Lake BB, Zhang K. 2019. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* **37**(12):1452–1457.
14. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, Daza RM, McFaline-Figueroa JL, Packer JS, Christiansen L, Steemers FJ, Adey AC, Trapnell C, Shendure J. 2018. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**(6409):1380–1385.
15. Ma S, Zhang B, LaFave L, Chiang Z, Hu Y, Ding J, Brack A, Kartha VK, Law T, Lareau C, Hsu YCRegev A, Buenrostro JD. Chromatin potential identified by shared single cell profiling of RNA and chromatin *bioRxiv* 2020.06.17.156943
16. Zhu C, Yu M, Huang H, Juric I, Abnousi A, Hu R, Lucero J, Behrens MM, Hu M, Ren B. 2019. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol* **26**(11):1063–1070.
17. Yin Y, Jiang Y, Lam KG, Berletch JB, Disteche CM, Noble WS, Steemers FJ, Camerini-Otero RD, Adey AC, Shendure J. 2019. High-Throughput Single-Cell Sequencing with Linear Amplification. *Mol Cell* **76**(4):676–690.e10.
18. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, Shah P, Bell JC, Jhutty D, Nemec CM, Wang J, Wang L, Yin Y, Giresi PG, Chang ALS, Zheng GXY, Greenleaf WJ, Chang HY. 2019. Massively parallel single-cell chromatin landscapes of hu-

man immune cell development and intratumoral T cell exhaustion.. *Nat Biotechnol* **37**(8):925–936.

19. Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, Pokholok D, Aryee MJ, Steemers FJ, Lebofsky R, Buenrostro JD. 2019. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol* **37**(8):916–924.
20. Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, Majeti R, Chang HY, Greenleaf WJ. 2018. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**(6):1535–1548.e16.
21. Ishii H, Kadonaga JT, Ren B. 2015. MPE-seq, a new method for the genome-wide analysis of chromatin structure. *Proc Natl Acad Sci U S A* **112**(27):E3457–3465.
22. Umeyama T, Ito T. 2017. DMS-Seq for In Vivo Genome-wide Mapping of Protein-DNA Interactions and Nucleosome Centers. *Cell Rep* **21**(1):289–300.
23. Baldi S, Krebs S, Blum H, Becker PB. 2018. Genome-wide measurement of local nucleosome array regularity and spacing by nanopore sequencing. *Nat Struct Mol Biol* **25**(9):894–901.
24. Aughey GN, Estacio Gomez A, Thomson J, Yin H, Southall TD. 2018. CATaDa reveals global remodelling of chromatin accessibility during stem cell differentiation in vivo. *Elife* **7**. pii: e32341.
25. Tchasovnikarova IA, Timms RT, Douse CH, Roberts RC, Dougan G, Kingston RE, Modis Y, Lehner PJ. 2017. Hyperactivation of HUSH complex function by Charcot-Marie-Tooth disease mutation in MORC2. *Nat Genet* **49**(7):1035–1044.
26. Timms RT, Tchasovnikarova IA, Lehner PJ. 2019. Differential viral accessibility (DIVA) identifies alterations in chromatin architecture through large-scale mapping of lentiviral integration sites. *Nat Protoc* **14**(1):153–170.
27. Chen PB, Zhu LJ, Hainer SJ, McCannell KN, Fazzio TG1. 2014. Unbiased chromatin accessibility profiling by RED-seq uncovers unique features of nucleosome variants in vivo. *BMC Genomics* **15**:1104.
28. Gargiulo G, Levy S, Bucci G, Romanenghi M, Fornasari L, Beeson KY, Goldberg SM, Cesaroni M, Ballarini M, Santoro F, Bezman N, Frigé G, Gregory PD, Holmes MC, Strausberg RL, Pelicci PG, Urnov FD, Minucci S. 2009. NA-Seq: a discovery tool for the analysis of chromatin structure and dynamics during differentiation. *Dev Cell* **16**(3):466–481.
29. Spracklin G, Pradhan S. 2020. Protect-seq: genome-wide profiling of nuclease inaccessible domains reveals physical properties of chromatin. *Nucleic Acids Res* **48**(3):e16.
30. Ponnaluri VKC, Zhang G, Estéve PO, Spracklin G, Sian S, Xu SY, Benoukraf T, Pradhan S. 2017. NicE-seq: high resolution open chromatin profiling. *Genome Biol* **18**(1):122.
31. Gregory PD, Barbaric S, Hörz W. 1999. Restriction nucleases as probes for chromatin structure. *Methods Mol Biol* **119**:417–425.
32. Almer A , Rudolph H, Hinnen A , Hörz W. 1986. Removal of Positioned Nucleosomes From the Yeast PHO5 Promoter Upon PHO5 Induction Releases Additional Upstream Activating DNA Elements *EMBO J* **5**(10):2689–2696.
33. Chereji RV, Eriksson PR, Ocampo J, Prajapati HK, Clark DJ. 2019. Accessibility of promoter DNA is not the primary determinant of chromatin-mediated gene regulation. *Genome Res* **29**(12):1985–1995.
34. Oberbeckmann E, Wolff M, Krietenstein N, Heron M, Ellins JL, Schmid A, Krebs S, Blum H, Gerland U, Korber P. 2019. Absolute nucleosome occupancy map for the *Saccharomyces cerevisiae* genome. *Genome Res* **29**(12):1996–2009.
35. Chereji RV, Ramachandran S, Bryson TD, Henikoff S. 2018. Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biol* **19**(1):19.
36. Voong LN, Xi L, Sebeson AC, Xiong B, Wang JP, Wang X. 2016. Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. *Cell* **167**(6):1555–1570.e15.
37. Brogaard K, Xi L, Wang JP, Widom J. 2012. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**(7404):496–501.
38. Flaus A, Luger K, Tan S, Richmond TJ. 1996. Mapping Nucleosome Position at Single Base-Pair Resolution by Using Site-Directed Hydroxyl Radicals. *Proc Natl Acad Sci U S A* **93**(4):1370–1375.
39. Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. 2020. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**(6498):1449–1454.
40. Abdulhay NJ, McNally CP, Hsieh LJ, Kasinathan S, Keith A, Estes LS, Karimzadeh M, Underwood JG, Goodarzi H, Narlikar GJ, Ramani V. 2020. Massively multiplex single-molecule oligonucleosome footprinting *bioRxiv* 2020.05.20.105379
41. Shipony Z, Marinov GK, Swaffer MP, Sinnott-Armstrong NA, Skotheim JM, Kundaje A, Greenleaf WJ. 2020. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat Methods* **17**(3):319–327.
42. Wang Y, Wang A, Liu Z, Thurman AL, Powers LS, Zou M, Zhao Y, Hefel A, Li Y, Zabner J, Au KF. 2019. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res* **29**(8):1329–1342.
43. Lee I, Razaghi R, Gilpatrick TSadowski N, Sedlazeck FJ, Timp W. 2018. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing *bioRxiv* 504993

44. Chen X, Miragaia RJ, Natarajan KN, Teichmann SA. 2018. A rapid and robust method for single cell chromatin accessibility profiling. *Nat Commun* **9**(1):5345.
45. Mezger A, Klemm S, Mann I, Brower K, Mir A, Bostick M, Farmer A, Fordyce P, Linnarsson S, Greenleaf W. 2018. High-throughput chromatin accessibility profiling at single-cell resolution. *Nat Commun* **9**(1):3647.
46. Lai B, Gao W, Cui K, Xie W, Tang Q, Jin W, Hu G, Ni B, Zhao K. 2018. Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature* **562**(7726):281–285.
47. Lai WKM, Pugh BF. 2017. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat Rev Mol Cell Biol* **18**(9):548–562.
48. Montefiori L, Hernandez L, Zhang Z, Gilad Y, Ober C, Crawford G, Nobrega M, Jo Sakabe N. 2017. Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. *Sci Rep* **7**(1):2451.
49. Meyer CA, Liu XS. 2014. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* **15**(11):709–721.
50. Sato S, Arimura Y, Kujirai T, Harada A, Maehara K, Nogami J, Ohkawa Y, Kurumizaka H. 2019. Biochemical analysis of nucleosome targeting by Tn5 transposase. *Open Biol* **9**(8):190116.
51. Lai B, Tang Q, Jin W, Hu G, Wangsa D, Cui K, Stanton BZ, Ren G, Ding Y, Zhao M, Liu S, Song J, Ried T, Zhao K. 2018. Trac-looping measures genome structure and chromatin accessibility. *Nat Methods* **15**(9):741–747.
52. Qu K, Zaba LC, Satpathy AT, Giresi PG, Li R, Jin Y, Armstrong R, Jin C, Schmitt N, Rahbar Z, Ueno H, Greenleaf WJ, Kim YH, Chang HY. 2017. Chromatin Accessibility Landscape of Cutaneous T Cell Lymphoma and Dynamic Response to HDAC Inhibitors. *Cancer Cell* **32**(1):27–41.e4.
53. Wu J, Xu J, Liu B, Yao G, Wang P, Lin Z, Huang B, Wang X, Li T, Shi S, Zhang N, Duan F, Ming J, Zhang X, Niu W, Song W, Jin H, Guo Y, Dai S, Hu L, Fang L, Wang Q, Li Y, Li W, Na J, Xie W, Sun Y. 2018. Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature* **557**(7704):256–260
54. Wu J, Huang B, Chen H, Yin Q, Liu Y, Xiang Y, Zhang B, Liu B, Wang Q, Xia W, Li W, Li Y, Ma J, Peng X, Zheng H, Ming J, Zhang W, Zhang J, Tian G, Xu F, Chang Z, Na J, Yang X, Xie W. 2016. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**(7609):652–657.
55. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, Shendure J. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* **11**(12):R119.
56. Goryshin IY, Reznikoff W. 1998. Tn5 in vitro transposition. *J Biol Chem* **273**(13):7367–7374.
57. He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, Rao PK, Fei T, Xu H, Long H, Liu XS, Brown M. 2014. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods* **11**(1):73–78.
58. Lazarovici A, Zhou T, Shafer A, Dantas Machado AC, Riley TR, Sandstrom R, Sabo PJ, Lu Y, Rohs R, Stamatoyannopoulos JA, Bussemaker HJ. 2013. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc Natl Acad Sci U S A* **110**(16):6376–6381.
59. Suck D, Lahm A, Oefner C. 1988. Structure refined to 2A of a nicked DNA octanucleotide complex with DNase I. *Nature* **332**(6163):464–468.
60. Cooper J, Ding Y, Song J, Zhao K. 2017. Genome-wide mapping of DNase I hypersensitive sites in rare cell populations using single-cell DNase sequencing. *Nat Protoc* **12**(11):2342–2354.
61. Lu F, Liu Y, Inoue A, Suzuki T, Zhao K, Zhang Y. 2016. Establishing Chromatin Regulatory Landscape during Mouse Preimplantation Development. *Cell* **165**(6):1375–1388.
62. Jin W, Tang Q, Wan M, Cui K, Zhang Y, Ren G, Ni B, Sklar J, Przytycka TM, Childs R, Levens D, Zhao K. 2015. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**(7580):142–146.
63. Song L, Crawford GE. 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* **2010**(2):pdb.prot5384.
64. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**(8):651–657.
65. Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830):1497–1502.
66. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**(7153):553–560.
67. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**(4):823–837.
68. Kempfer R, Pombo A. 2020. Methods for mapping 3D

chromosome architecture. *Nat Rev Genet* **21**(4):207–226.

69. Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, Zhang Y, Sos BC, Afzal V, Dickel DE, Kuan S, Visel A, Pennacchio LA, Zhang K, Ren B. 2018. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci* **21**(3):432–439.
70. Fullard JF, Hauberg ME, Bendl J, Egervari G, Cirnaru MD, Reach SM, Motl J, Ehrlich ME, Hurd YL, Roussos P. 2018. An atlas of chromatin accessibility in the adult human brain. *Genome Res* **28**(8):1243–1252.
71. Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**(2 Pt 1):299–308.
72. West JA, Cook A, Alver BH, Stadtfeld M, Deaton AM, Hochedlinger K, Park PJ, Tolstorukov MY, Kingston RE. 2014. Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. *Nat Commun* **5**:4719.
73. Giresi PG, Lieb JD. 2009. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods* **48**(3):233–239.
74. Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen E, Rynes E, Reynolds A, Nelson J, Johnson A, Frerker M, Buckley M, Kaul R, Meuleman W, Stamatoyannopoulos JA. 2020. Global reference mapping and dynamics of human transcription factor footprints. *bioRxiv* 2020.01.31.927798
75. Hewish DR, Burgoyne LA. 1973. Chromatin substructure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. *Biochem Biophys Res Commun* **52**(2):504–510.
76. Weintraub H, Groudine M. 1976. Chromosomal subunits in active genes have an altered conformation. *Science* **193**(4256):848–856.
77. Taberlay PC, Kelly TK, Liu CC, You JS, De Carvalho DD, Miranda TB, Zhou XJ, Liang G, Jones PA. 2011. Polycomb-repressed genes have permissive enhancers that initiate reprogramming. *Cell* **147**(6):1283–1294.
78. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**(5):887–898.
79. Galas DJ, Schmitz A. 1978. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **5**(9):3157–3170.
80. Jackson PD, Felsenfeld G. 1985. A method for mapping intranuclear protein-DNA interactions and its application to a nuclease hypersensitive site. *Proc Natl Acad Sci U S A* **82**(8):2296–2300.
81. Kornberg RD. 1974. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**(4139):868–871.
82. Kornberg RD, Thomas JO. 1974. Chromatin structure; oligomers of the histones. *Science* **184**(4139):865–868.
83. Klemm SL, Shipony Z, Greenleaf WJ. 2019. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**(4):207–220.
84. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**(6648):251–260
85. Woodcock CL, Sweetman HE, Frado LL. 1976. Structural repeating units in chromatin. II. Their isolation and partial characterization. *Exp Cell Res* **97**:111–119.
86. Woodcock CL, Safer JP, Stanchfield JE. 1976. Structural repeating units in chromatin. I. Evidence for their general occurrence. *Exp Cell Res* **97**:101–110.
87. Trojer P, Reinberg D. 2007. Facultative heterochromatin: is there a distinctive molecular signature? *Mol Cell* **28**(1):1–13.
88. ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146):799–816.
89. ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414):57–74.
90. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD. 2004. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet* **36**(8):900–905.
91. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**(7414):75-82.
92. Ozsolak F, Song JS, Liu XS, Fisher DE. 2007. High-throughput mapping of the chromatin structure of human promoters. *Nat Biotechnol* **25**(2):244–248.
93. Sheffield NC, Furey TS. 2012. Identifying and characterizing regulatory sequences in the human genome with chromatin accessibility assays. *Genes (Basel)* **3**(4):651–670. doi: 10.3390/genes3040651.
94. Turner BM. 2007. Defining an epigenetic code. *Nat Cell Biol* **9**(1):2–6.

95. Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**(6):465–476.
96. Zaret KS, Carroll JS. 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**(21):2227–2241.
97. Cirillo LA, Zaret KS. 1999. An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA. *Mol Cell* **4**(6):961–969.
98. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* **32**(2):171–178.
99. Schwartzentruber J, Korshunov A, Liu XY, Jones DT, Pfaff E, Jacob K, Sturm D, Fontebasso AM, Quang DA, Tönjes M, Hovestadt V, Albrecht S, Kool M, Nantel A, Konermann C, Lindroth A, Jäger N, Rausch T, Ryzhova M, Korbel JO, Hielscher T, Hauser P, Garami M, Klekner A, Bognar L, Ebinger M, Schuhmann MU, Scheurlen W, Pekrun A, Frühwald MC, Roggendorf W, Kramm C, Dürken M, Atkinson J, Lepage P, Montpetit A, Zakrzewska M, Zakrzewski K, Liberski PP, Dong Z, Siegel P, Kulozik AE, Zapatka M, Guha A, Malkin D, Felsberg J, Reifenberger G, von Deimling A, Ichimura K, Collins VP, Witt H, Milde T, Witt O, Zhang C, Castelo–Branco P, Lichter P, Faury D, Tabori U, Plass C, Majewski J, Pfister SM, Jabado N. 2012. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* **482**(7384):226–231.
100. Hendrich B, Bickmore W. 2001. Human diseases with underlying defects in chromatin structure and modification. *Hum Mol Genet* **10**(20):2233–2242.
101. Vinagre J, Almeida A, Pópulo H, Batista R, Lyra J, Pinto V, Coelho R, Celestino R, Prazeres H, Lima L, Melo M, da Rocha AG, Preto A, Castro P, Castro L, Pardal F, Lopes JM, Santos LL, Reis RM, Cameselle-Teijeiro J, Sobrinho-Simes M, Lima J, Mximo V, Soares P. 2013. Frequency of TERT promoter mutations in human cancers. *Nat Commun* **4**:2185.
102. Matsumoto L, Takuma H, Tamaoka A, Kurisaki H, Date H, Tsuji S, Iwata A. 2010. CpG demethylation enhances alpha-synuclein expression and affects the pathogenesis of Parkinson's disease. *PLoS One* **5**(11):e15522.
103. Moore SP, Kruchten J, Toomire KJ, Strauss PR. 2016. Transcription Factors and DNA Repair Enzymes Compete for Damaged Promoter Sites. *J Biol Chem* **291**(11):5452–5460.
104. Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. 2016. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**(7598):264–267.
105. Roadmap Epigenomics Consortium. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**(7539):317–330.
106. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, Amrhein H, Fisher-Aylor K, Antoshechkin I, DeSalvo G, See LH, Fastuca M, Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu YC, Rasmussen MD, Bansal MS, Kellis M, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N, Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, Kent WJ, Ramalho Santos M, Herrero J, Notredame C, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Canfield T, Sabo PJ, Wilken MS, Reh TA, Giste E, Shafer A, Kutyavin T, Haugen E, Dunn D, Reynolds AP, Neph S, Humbert R, Hansen RS, De Bruijn M, Selleri L, Rudensky A, Josefowicz S, Samstein R, Eichler EE, Orkin SH, Levasseur D, Papayannopoulou T, Chang KH, Skoultchi A, Gosh S, Disteche C, Treuting P, Wang Y, Weiss MJ, Blobel GA, Cao X, Zhong S, Wang T, Good PJ, Lowdon RF, Adams LB, Zhou XQ, Pazin MJ, Feingold EA, Wold B, Taylor J, Mortazavi A, Weissman SM, Stamatoyannopoulos JA, Snyder MP, Guigo R, Gingeras TR, Gilbert DM, Hardison RC, Beer MA, Ren B; Mouse ENCODE Consortium. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**(7527):355–364.
107. Wu C, Wong YC, Elgin SC. 1979. The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell* **16**(4):807–814.
108. Wu C, Bingham PM, Livak KJ, Holmgren R, Elgin SC. 1979. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell* **16**(4):797–806.
109. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**(2):311–322.
110. Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. 2012. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* **22**(12):2497–2506.
111. Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schübeler D. 2017. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol Cell* **67**(3):411–422.e4.

112. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**(6):877–885.
113. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**(12):1213–1218.
114. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**(7561):486–490.
115. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. 2015. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**(6237):910–914.
116. Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **24**(10):1595–1602.
117. Pijuan-Sala B, Wilson NK, Xia J, Hou X, Hannah RL, Kinston S, Calero-Nieto FJ, Poirion O, Preissl S, Liu F, Göttgens B. 2020. Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. *Nat Cell Biol* **22**(4):487–497.
118. Al-Ali R, Bauer K, Park JW, Al Abdulla R, Fermi V, von Deimling A, Herold-Mende C, Mallm JP, Herrmann C, Wick W, Turcan S. 2019. Single-nucleus chromatin accessibility reveals intratumoral epigenetic heterogeneity in IDH1 mutant gliomas. *Acta Neuropathol Commun* **7**(1):201.
119. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, Lee C, Regalado SG, Read DF, Steemers FJ, Disteche CM, Trapnell C, Shendure J. 2018. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**(5):1309–1324.e18.
120. Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco–Ferreres R, Pliner HA, Christiansen L, Qiu X, Steemers FJ, Trapnell C, Shendure J, Furlong EEM. 2018. The *cis*-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**(7697):538–542.
121. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. 2017 chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**:975–978.
122. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B, Kathiria A, Cho SW, Mumbach MR, Carter AC, Kasowski M, Orloff LA, Risca VI, Kundaje A, Khavari PA, Montine TJ, Greenleaf WJ, Chang HY. 2017. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**(10):959–962.
123. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ, Majeti R, Chang HY. 2016. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**(10):1193–1203.
124. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, Fields S, Stamatoyannopoulos JA. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* **6**(4):283–289.