REVIEW

Single-cell epigenomics: Recording the past and predicting the future

Gavin Kelsey,[1,2]*† Oliver Stegle,[3,4]*† Wolf Reik[1,2,5]†

Single-cell multi-omics has recently emerged as a powerful technology by which different layers of genomic output—and hence cell identity and function—can be recorded simultaneously. Integrating various components of the epigenome into multi-omics measurements allows for studying cellular heterogeneity at different time scales and for discovering new layers of molecular connectivity between the genome and its functional output. Measurements that are increasingly available range from those that identify transcription factor occupancy and initiation of transcription to long-lasting and heritable epigenetic marks such as DNA methylation. Together with techniques in which cell lineage is recorded, this multilayered information will provide insights into a cell's past history and its future potential. This will allow new levels of understanding of cell fate decisions, identity, and function in normal development, physiology, and disease.

The discovery and description of individual cells in the body has fascinated biologists and pathologists since the cell was discovered (*1*). With the advent of molecular cell biology, methods have been developed for measuring properties and functions of single cells at increasing resolution. This includes, among others, fluorescent protein reporters and single-molecule detection of RNA or DNA. Only recently however, have high-throughput sequencing methods allowed us more comprehensive access to genomic information in single cells. Hence, single-cell RNA sequencing has revealed how heterogeneous the transcriptome of individual cells can be within a seemingly homogeneous cell population or tissue, providing insights into cell identity, fate, and function in the context of both normal biology and pathology [Stubbington *et al.* (*2*) and Lein *et al.* (*3*)]. A few years from now, we likely will have access to total RNA, small and long noncoding RNA, and transcriptional initiation output of the transcriptome (in addition to the stable cytoplasmic component). The development of single-cell RNA sequencing was followed by single-cell genome sequencing, which has provided new insights into genomic stability and genomic variations that occur in physiology and in disease—for example, in cancer, reproductive medicine, or microbial genetics (*4*).

Epigenetics connects the genome with its functional output (Fig. 1). Various epigenetic marks have been described, ranging from DNA (such as DNA methylation) to histone modifications, which can affect the way the cell reads its genome and hence its transcriptional output. Transcription factors that bind to DNA can create or alter epigenetic states (e.g., open or closed chromatin and higher-order chromatin conformation), or their binding can be sensitive to preexisting epigenetic states. Some epigenetic marks can also be heritable from one cell generation to the next (during mitosis) or from one organism generation to the next [intergenerational or transgenerational epigenetic inheritance (*5*)]. However, there are key questions in epigenetics that can only be addressed by determining the epigenome in single cells. For example, how is transcriptional heterogeneity between cells connected with epigenetic heterogeneity (if it is), do changes in transcription precede or follow epigenetic marks when cells change their fate or function, and are epigenetic states better or worse identifiers of rare cell populations and transitional states than the transcriptome? The recent development of single-cell epigenomics methods is beginning to allow us to address these fundamental questions.

> ***"[T]oday we can probe the majority of epigenetic dimensions with single-cell resolution."***

Single-cell epigenome methods can identify open or closed chromatin, including nucleosome positioning (*6*–*11*). From these, one can infer the likelihood of certain transcription factors to bind or not bind to specific DNA sequences within individual cells, and methods are being developed that allow for assaying transcription factor binding directly—for example, single-cell chromatin immunoprecipitation sequencing (ChIP-seq). Thus, one can currently measure (albeit imperfectly) the heterogeneity in a cell population of key histone marks associated with transcriptional states, such as H3K4me3, which indicates active transcription, or H3K27me3, which is found on genes with a repressed transcriptional state (*12*). Functional states (such as transcriptional output) of the genome are also guided by the way the DNA in each cell is organized into higher-order chromatin, which can be determined by single-cell high-throughput chromosome conformation capture (Hi-C) (*13*). Finally, various DNA modifications—such as methylation (5mC), hydroxymethylation (5hmC), and formylcytosine (5fC)—can be located at the single-cell level by sequencing in most areas of the genome, including at single-nucleotide resolution (*14*–*18*). These modifications are part of the biological turnover of DNA methylation and are associated, for example, with transcriptional repression (5mC) or enhancers, including active ones (5hmC and 5fC). Hence, today we can probe the majority of epigenetic dimensions with single-cell resolution.

The techniques described above have been combined into single-cell multi-omics (*19*), which can reveal new connections between regulatory principles that operate in the individual layers (Figs. 1 and 2). Hence, genome sequencing together with transcriptome sequencing can reveal how genetic variation is related to transcriptional variation (*20*, *21*). Furthermore, genome-scale methylome sequencing coupled with the transcriptome (*22*, *23*) has identified widespread associations between epigenetic marks and transcriptional heterogeneity. The latest incarnation, triple-omics, combines genome, methylome, and transcriptome (*24*) assays and can reveal methylome, chromatin accessibility, and the transcriptome (*11*). Together with the development of multidimensional computational methods (*22*, *25*), these techniques are beginning to tease out intricate and unique cell- and locus-specific relationships between, say, methylation and nucleosome accessibility of a gene promoter and the transcriptional output of the gene (*11*).

Single-cell profiling of DNA modifications

Because epigenetic information comes in multiple forms—covalent modifications on DNA, posttranslational modifications of histones, chromatin accessibility and compaction, and higher-order conformation of chromosome domains—each layer of information requires a different biochemical approach to profile it. This has implications for the nature and quality of the information generated from single cells and for the ability to combine multiple measures from the same single cell in multi-omic applications. Depending on the type of question, it will be necessary to determine whether depth or breadth (many, many cells) is required for any specific study (Fig. 2).

Technically, DNA methylation has been the easiest to assay, building on well-established bisulphite chemistry (*26*). However, bisulphite treatment degrades DNA, preventing full-genome coverage and requiring an adaptation of bisulphite sequencing (BS-seq) to the single-cell level (*14*–*16*). BS-seq, by which unmodified cytosine is converted to thymine but 5mC remains unconverted (*26*), yields single-base precision in principle, with the advantage that both modified and unmodified sites are identified (*26*). Therefore, sites without

[1]Epigenetics Programme, Babraham Institute, Cambridge CB22 3AT, UK. [2]Centre for Trophoblast Research, University of Cambridge, Cambridge CB2 3EG, UK. [3]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD Hinxton, Cambridge, UK. [4]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg 69117, Germany. [5]Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK.
*These authors contributed equally to this work. **†Corresponding author. Email: gavin.kelsey@babraham.ac.uk (G.K.); oliver.stegle@ebi.ac.uk (O.S.); wolf.reik@babraham.ac.uk (W.R.)**

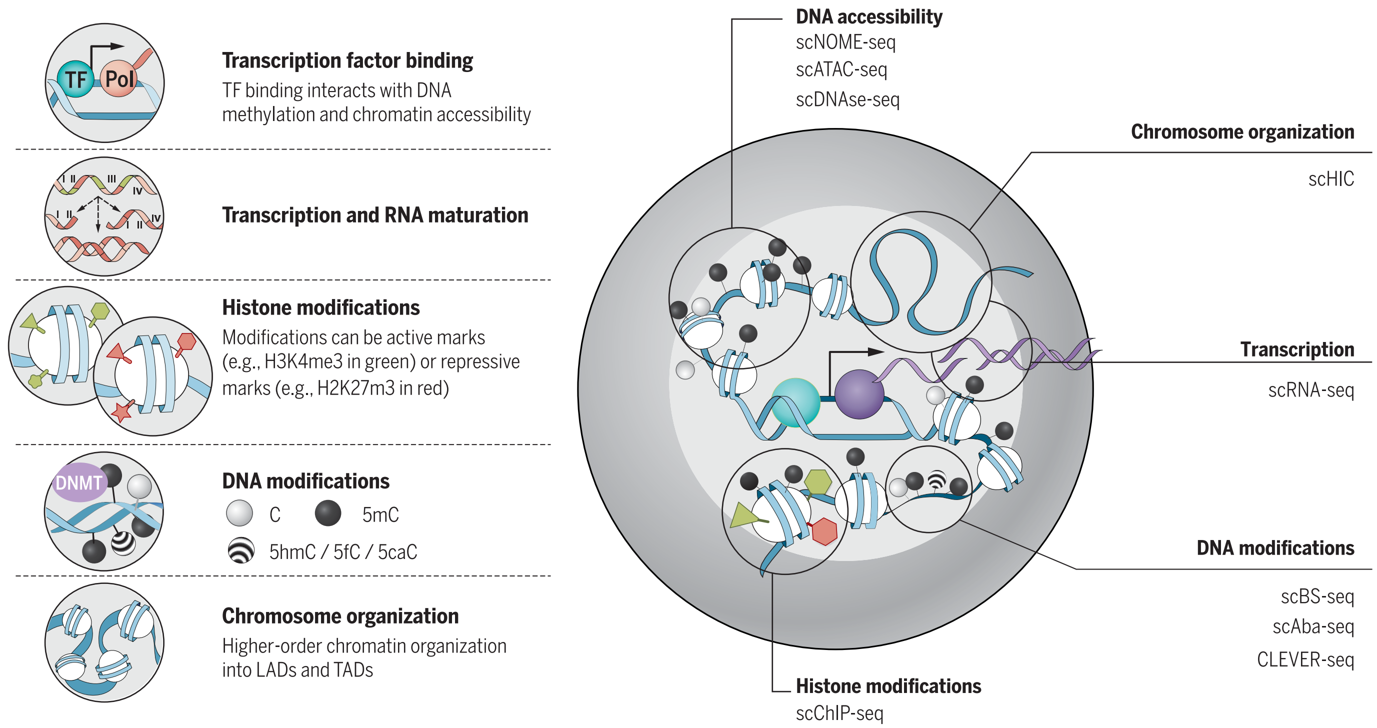

Fig. 1. Single-cell methods and heterogeneity of different molecular layers. (**Left**) Overview of different molecular layers that can be assayed using single-cell protocols. (**Right**) A cell with different layers of multi-omics measurements, as defined on the left. Concordance or heterogeneity respectively may exist between the different layers, and this can be recorded by single-cell sequencing and computationally evaluated.

information are not falsely assigned as unmethylated and, because of the general congruence of methylation over consecutive CpGs in many genomic contexts, missing sites can be imputed from relatively sparse data.

Current single-cell BS-seq (scBS-seq) protocols achieve a coverage up to ~40% (*15*), which means that for most loci the observed sequence reads will originate from only one chromosomal copy. Recent advances in performing single-cell methylation profiling with combinatorial indexing (*27*, *28*) may mitigate some of these limitations while simultaneously offering scalability to thousands of cells in a single experiment (Fig. 2). Alternatively, because methylation state can determine whether particular restriction enzymes cleave their recognition sites, methods that use methylation-sensitive or dependent restriction enzymes could present an alternative to bisulphite-based methods (*29*).

Mapping the derivatives of 5mC in single cells has been particularly useful in preimplantation embryos, in which oxidation of 5mC contributes to the active demethylation of the paternal chromosomes (*30*). The pronounced strand bias in distribution between sister cells of these modifications along the same chromosome has provided high-resolution analysis of sister-chromatin exchange (*31*) and has been used as a lineage reconstruction tool (*17*), as well as mapping active demethylation in advance of expression at the promoters of developmentally important genes (*18*). Such advances have required alternative approaches, because 5mC cannot be discriminated from the less abundant 5hmC after bisulphite treatment, and the rarer derivatives 5fC and 5-carboxycytosine (5caC) are indistinguishable from unmodified cytosine.

Treatment with the CpG methylase M.SssI [methylase-assisted bisulphite sequencing, (MAB-seq)] (*31*) allows indirect detection of 5fC, together with 5caC, due to their retention as the only sites remaining susceptible to C to T conversion after bisulphite treatment. Careful control of the methylation reaction is needed to minimize false-positive calls, particularly for a rare modification such as 5fC, which is present at most at tens of thousands of CpG sites, compared with millions of CpGs modified by 5mC. 5hmC can be profiled in single cells by glucosylating 5hmC positions to generate recognition sites for the restriction endonuclease AbaS1 (scAba-seq) (*17*). This provides a positive readout of 5hmC, but, with the inclusion of multiple enzymatic reactions, there is an unknown false-negative rate, which might contribute to a range in the number of 5hmC positions recorded in single cells. 5fC can be detected in single cells by direct chemical labeling with the specific reactivity of malononitrile [chemical-labeling-enabled C-to-T conversion sequencing (CLEVER-seq)] (*18*). The adduct produced prevents normal pairing with G, such that labeled 5fC sites are read as T during polymerase chain reaction (PCR) amplification. In theory, this approach may allow for robust detection of modified bases on single-molecule sequencing platforms.

Combining methylation profiling into multi-omics approaches

scBS-seq can be combined with scRNA-seq through separation of nuclei from cell cytoplasm, separation of RNA and DNA for separate downstream reactions, or preamplification of RNA and DNA in the same cell lysate before splitting and parallel processing for genomic DNA amplification and cDNA library preparation (*22*–*24*). BS-seq coverage is sufficiently uniform to permit identification of chromosome aneuploidies or large CNVs from regional variations in read depth (*24*). Of note, similar to scRNA-seq protocols that use plate-based methods, scBS-seq can in principle be coupled with profiling of up to tens of cell-surface markers that can be assayed using fluorescence-activated cell sorting, an approach that has been applied in immunology [see Stubbington *et al.*, (*2*)].

Bisulphite sequencing also underlies the nucleosome occupancy and methylome (NOME) sequencing method, which enables information on nucleosome positioning and accessible chromatin to be inferred simultaneously with DNA methylation (*9*–*11*). Individual lysed cells are treated with M.CviPI, which methylates GpC sites in accessible DNA; then, following bisulphite treatment, methylated cytosines in a GpC context demarcate accessible DNA (linker regions and nucleosome-free DNA), while methylation is read from conversion events of CpGs. Because both accessible and nonaccessible states are reported, missing information is not falsely assigned, which provides an advantage over other methods for chromatin accessibility. On the other hand, as a method that sequences the genome with no selectivity for open chromatin, high levels of sequencing may be needed to guarantee coverage of elements of interest.

Another potential limitation is the need to filter out C-C-G and G-C-G positions from the methylation data, which reduces the number of genome-wide cytosines that can be assayed compared with scBS-seq by ~50%. However, despite this filter, a large proportion of the loci in genomic regions with important regulatory roles, such as promoters and enhancers, can still be profiled using scNOME-seq–based methods (*11*). scNOME-seq has identified chromatin remodeling dynamics on the two parental alleles during preimplantation development, discriminating cis-regulatory elements open in all cells and promoters that diverge in accessibility between individual blastomeres, these being relatively enriched in gene ontology (GO) terms related to developmental processes and cell differentiation (*9*). Further enhancements of these data can be provided by incorporating transcriptome information from the same cell (Fig. 2) (*11*) to query the strength of coupling between DNA methylation, open chromatin, and transcriptional output.

Mapping functional chromatin states in single cells

A variety of assays have been adapted to profile chromatin states in single cells; these are predicated on enrichment-based strategies; thus, in principle, they have a lower sequencing overhead than scNOME-seq. Open chromatin can be identified by deoxyribonuclease I (DNase I) sensitivity, which was first adapted to the single-cell level in a low-throughput application able to detect an average of ~40,000 DNase I hypersensitive sites (DHSs) per cell (*6*). However, due to nonspecific signals throughout the genome, the false-discovery rate is high. Thus, previous knowledge of DHSs from bulk experiments is required to identify genuine DHSs, with the confidence of detection of proximal regulatory elements scaling with expression level of associated genes.

Higher-throughput applications have been developed for the assay for transposase-accessible chromatin sequencing (ATAC-seq), in which DNA accessibility is probed by the ability of the prokaryotic Tn5 transposase to insert sequencing adapters into accessible regions of the genome, in contrast to regions that are inaccessible, such as those interacting with a nucleosome. These approaches have used microfluidics to process single cells and introduce cell-identifying barcodes as part of the tagging process (*7*) or by combinatorial-cell barcoding (*8*) (Fig. 2), allowing parallel processing of a large number of samples (>10,000).

Throughput levels face a cost of reduced depth, as typically <10% of known promoters are represented in an individual scATAC-seq library. Sparseness of data limits analysis of cellular variation at individual regulatory elements. This may preclude ab initio identification of open chromatin sites, and the absence of open chromatin at a locus of interest in a single cell may reflect missing data. As well as reporting active regulatory elements governing hematopoietic differentiation, scATAC-seq has identified the evolution of regulatory elements during disease progression in acute myeloid leukemia (*32*). In addition, the ability of scATAC-seq to delineate the cis-regulatory landscapes of constituent cell types from a complex solid tissue has been demonstrated by isolating single nuclei from frozen samples of mouse forebrain (*33*).

"Technological advances for assaying epigenetic diversity at the single-cell level have gone hand-in-hand with computational methods for interpreting the data generated."

Posttranslational modifications of histones that correlate with chromatin activity states are conventionally mapped by ChIP-seq. Adapting ChIP-seq to extract this information from single cells presents additional problems of specificity and sensitivity, because it is dependent on antibody binding to pull down modified histones with associated DNA. Droplet approaches and cellular barcoding to label nuclei individually at the stage of micrococcal nuclease digestion (which fragments chromatin into nucleosomes) with immunoprecipitation on pools of cells and subsequent deconvolution of single-cell data after multiplex library sequencing allow thousands of single cells to be processed in single experiments (*12*) (Fig. 2). Yet, although ~50% of sequencing reads may fall within known peaks of H3K4me3 enrichment (the archetypal mark of active promoters), only ~5% of known peaks are detected per cell, with data too sparse for productive de novo peak calling.

We shall inevitably see technical improvements in each of these chromatin profiling methods, as well as incorporating them into multi-omic approaches. A challenge is to extract RNA from cell lysates in a way that preserves both chromatin state and RNA integrity, but with the sparsity of data from current scATAC-seq, scDNase-seq, or scChIP-seq methods, attainment of parallel data on gene expression and chromatin state at specific loci is challenging, and processing increasing numbers of cells may be necessary to obtain sufficient convergent information. Any of the above methods in theory could be combined with bisulphite sequencing to investigate DNA methylation state, which is not to underestimate the technical challenges that may need to be overcome in adding the chemical steps involved in bisulphite treatment.

Readouts of gross chromatin organization in single cells

Higher orders of chromosome organization in interphase nuclei are represented by a number of configurations: topologically associated domains (TADs) divide the genome into structurally separate segments contained in loops and constrained by boundary elements, and lamin-associated domains (LADs) occupy the nuclear periphery. LADs have been probed at the single-cell level by Dam-ID, in which the Dam adenosine methyltransferase is fused with lamin B1 (a constituent of the nuclear lamina) and expressed in cells so that sites of interaction are mapped from sequence tags after DpnI digestion (*34*). Because LADs are megabase-scale chromosome domains, with 1100 to 1400 domains present in a typical cell, only a low rate of false negatives is expected. The extent of heterogeneity between cells thus allows a good measure of the numbers of constitutive and facultative LADs, as well as cooperativity between LADs; such data are not accessible from population-based approaches. Dam-ID methodology could be applied to any other protein interacting with DNA, such as chromatin remodelers and transcription factors. One caveat is that the false-negative rate will increase as the domain of interaction diminishes, or for proteins with very transient interactions.

Hi-C data measures the proximity of DNA sequences in three-dimensional (3D) space on the basis of ligation events in fixed nuclei. A variety of optimizations have been introduced to increase resolution of the data (*35*), as well as throughput (*36*, *37*), since the first report of a single-cell Hi-C method (*13*). Using haploid cells, single-cell Hi-C has allowed modeling of the 3D organization of all chromosomes in individual cells (*38*) and revealed how bulk-cell data obscures the dynamic reorganization of chromosome compartments during the cell cycle (*36*). Despite recent advances, the resolution of scHi-C methods remains insufficient to interrogate contacts between specific promoters and their enhancers, which awaits progress in miniaturizing approaches to promoter-capture Hi-C or complementation with functional experiments, such as epigenome editing (*39*).

Scalability and limitation of current methods

There are common challenges and limitations that apply to several single-cell epigenome methods. An important bottleneck is the currently limited capture rate (e.g., up to ~40% for scBS-seq), which means that even if libraries are sequenced to saturation, missing values are unavoidable (Fig. 3). Other potential drawbacks are low mappability rates (~20 to 30%) and high levels of PCR duplicates (*15*), in particular for deeply sequenced libraries (*16*), which need to be considered when analyzing the resulting data.

So far, epigenome-based methods tend to offer lower throughput than scRNA-seq, which can already be scaled to tens or hundreds of thousands of cells. Recent advances to perform single-cell methylation profiling, ATAC-seq, and Hi-C using combinatorial indexing (*8*, *28*, *37*) have narrowed this gap. However, in particular, multi-omics methods that require a physical separation step of the RNA and DNA remain limited to medium-throughput analyses of hundreds of cells (Fig. 2). Another current challenge is to estimate and control for technical sources of variation. In single-cell transcriptomics, the level of technical noise can be estimated with spike-in standards, but such normalization strategies are not established for epigenome sequencing. A general strategy that can be useful are negative and positive controls—e.g., diluted bulk material used to create "pseudo cells" or control wells that combine one cell each from different species (*16*), which can be processed alongside each batch of single cells.

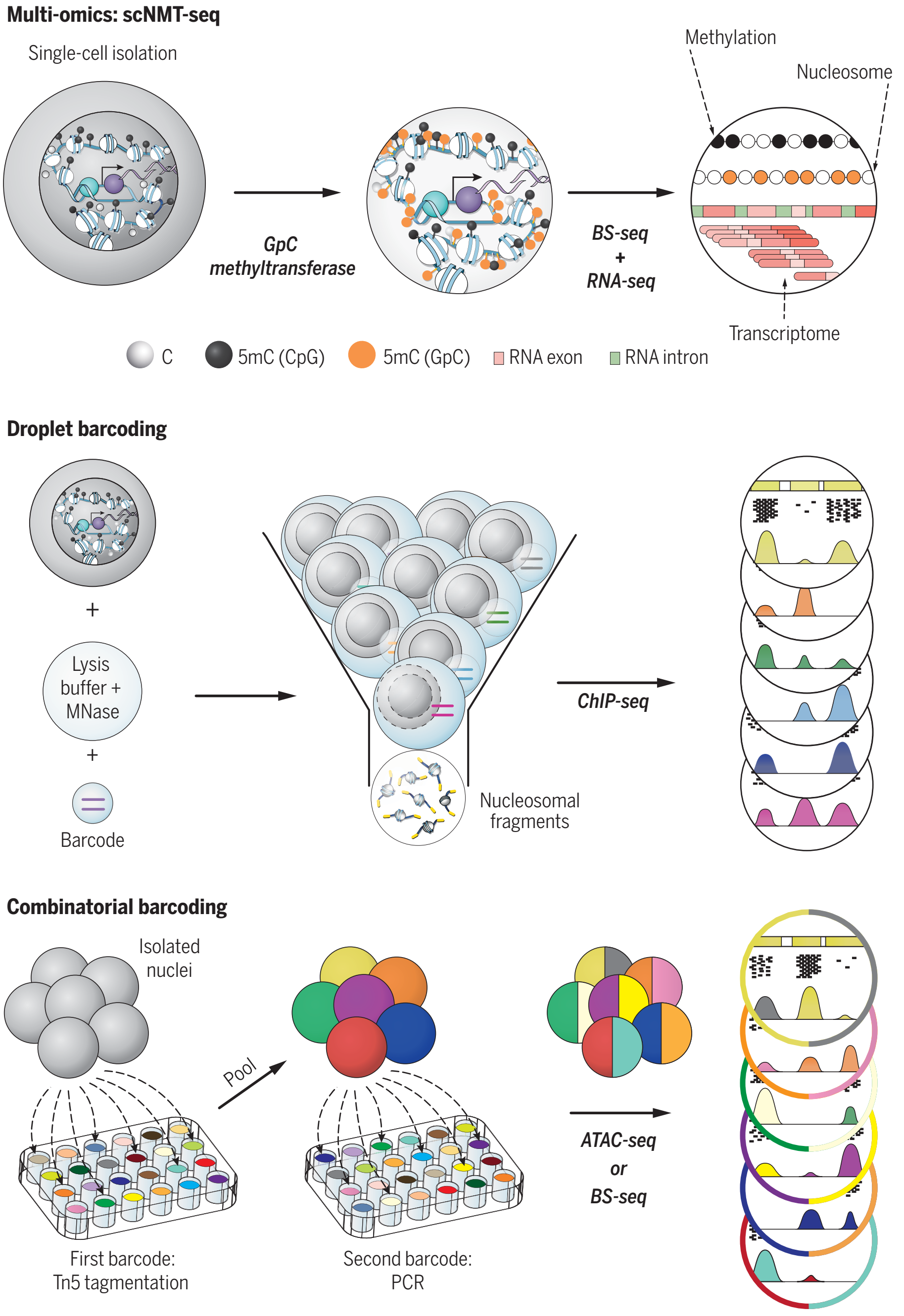

Fig. 2. Depth versus breadth: Multi-omics and cell-barcoding methods. Examples of different technical approaches are shown. (Top) Single-cell nucleosome, methylation, and transcription sequencing (scNMT-seq) (*11*) by which nucleosome accessibility, DNA methylation, and the transcriptome are read simultaneously at considerable depth in each cell; however, with individual cells processed in parallel but separately, cell numbers that can be currently analyzed in this way are limited to hundreds or thousands. (Middle) Barcoding chromatin in individual cells encapsulated in oil droplets, followed by pooling to bulk up material, enables thousands of cells to be processed while seeking to preserve signal-to-noise ratio (*12*). (Bottom) Combinatorial-cell barcoding (*8*, *64*), where readouts can be identified as coming from individual cells by unique combinations of barcodes present in each cell. This approach can be carried out on large numbers of cells (millions), but the depth of information per cell is limited.

Computational analysis to account for missing information using pooling strategies and imputation

Technological advances for assaying epigenetic diversity at the single-cell level have gone hand-in-hand with computational methods for interpreting the data generated (Fig. 3). A first critical step in the computational analysis is the appropriate normalization of the sequencing data while accounting for the typically high levels of noise observed. The sparse coverage of processed single-cell epigenome data sets requires careful consideration in downstream analyses.

Protocols vary in their coverage and whether missing data can be identified directly. For methods that use a bisulphite conversion step, the read coverage is independent of variation in DNA methylation, and hence missing data can be readily identified. For other methods, such as single-cell ATAC-seq, this can be more difficult because the absence or presence of sequence reads is the primary readout of the assay. Different strategies to address the low coverage in these data, such as aggregating read information within regions, by combining reads in consecutive sequence windows (*15*, *16*, *40*) or in annotated genomic contexts, such as promoter regions, enhancers and the like have been proposed. However, there are trade-offs between spatial resolution and coverage, parameters that may greatly affect downstream analyses.

Depending on the question, it may be advantageous to adjust for differences in global methylation, either at the whole-cell level or stratified by genomic context (*16*). A second strategy is to pool cells with similar epigenetic profiles, such as with an initial clustering step to then aggregate read information across cells within each cluster (*27*). These average profiles can offer high spatial resolution, however, at the cost that epigenetic diversity can only be studied at the level of the identified cell clusters (*24*). A third strategy comprises model-based approaches to impute missing information with predictive models. Such strategies have been proposed in the context of bulk epigenome profiles (*41*, *42*) and most recently have been generalised for imputing single-cell DNA methylation data (*25*). Additionally, we note that parallel data from multi-omics experiments will be associated with different patterns of missing data. Because of cost and experimental limitations, not all molecular layers will be assayed in each cell, and hence new computational methods need to handle heterogeneous designs to impute entire molecular layers.

Interrogating single-cell epigenome variation

Depending on the biological question at hand, several downstream analyses can be considered. Caution is required to consider the biological sources of variation that one may expect in a given study. For example, the cell cycle is a dominant driver of gene expression variation in single cells (*43*) but also manifests at other molecular layers, including copy-number states and DNA methylation (*9*). Also, DNA replication dynamics need to be taken into consideration during experimental design and data analysis.

A starting point for many analyses can be tests for differential epigenetic profiles between different cell clusters—for example, to identify differentially methylated regions between cell types or states (*16*). In cell populations without strong substructure, it may be advantageous to quantify the epigenetic diversity of individual loci with the pairwise distance of global methylome (*16*) or estimates of epigenetic variability between cells at individual loci (*15*).

As multi-omics protocols become more widely accessible, there are also exciting opportunities to interrogate associations between different epigenetic layers and to examine associations with the transcriptome. This allows the strength of coupling between different regulatory layers to be probed in great detail. Variation in coupling strength—for example, between DNA methylation and transcription—is known from bulk analyses, comparing pluripotent to somatic cell types (*44*).

However, the variation in coupling strength can be investigated with single-cell techniques for classes of loci or individual loci between cells or between different loci within the same cell. Such variation has already been identified at different levels, including individual loci such as gene promoters and enhancers with epigenetic variation associated with expression levels of individual genes, as well as global genome-wide couplings between different layers (*22*). If multi-omics methods are applied to hybrids or outbred individuals, it may be possible to assess allele-specific methylation and expression, thereby aligning regulatory differences across molecular layers (*23*). For other analyses, it remains an open question how to best integrate data across different molecular layers. Tying together different data modalities will improve cell clustering, and the use of epigenetic information in tandem with transcriptional data will aid in reconstructing pseudotemporal orderings of cells (Fig. 4).

Adding a temporal dimension in single-cell studies

Putting multidimensional information together for each single cell gives insights not only into cell identity and function but also, through the use of different layers of the epigenome, into past history and future potential (Fig. 4A). Imagine that an otherwise stable DNA methylation mark (for example, in an imprinted gene) has changed at a specific developmental time point, which can be recorded through lineage tracing by CRISPR scarring (*45*–*49*) (Fig. 4B). This is an example in which past history is recorded. Conversely, characteristic DNA methylation patterns in induced pluripotent stem cells (iPSCs) can be predictive of the differentiation potential of these iPSCs (*50*), an example of an epigenetic state revealing future potential.

Different epigenetic marks have different stabilities in time, providing the potential to record various biological time scales. An extracellular signal acting via intracellular signaling pathways will affect transcription factor binding and thus transcription. Because transcription factor binding can be highly dynamic and nonprocessed transcripts are usually short-lived, such signals may reflect the shortest possible biological response time scale. Conversely, though, some transcription factors may bind throughout cell division (*51*) and transmit epigenetic information to the next cell generation. Different binding time scales and their functional consequences may be revealed by coupling the analysis to cell cycle state through the transcriptome (*43*). Similarly, nucleosome accessibility in promoters (or other regulatory sequences) may occur before the chromatin opening up (as may be the case with pioneer factors) or, more conventionally, allow access to transcription factors. Within one cell cycle, therefore, we can reconstruct a signaling response at its cognate promoter, giving rise to transcriptional initiation followed by the processed transcript in the cytoplasm. We can discover multiple genomic dimensions in which this signaling response plays out within this single cell. It is currently possible to reconstruct such multidimensional responses in highly synchronized tissue culture systems but not in the natural setting in vivo, let alone in complex disease situations.

The applications with the most fundamental potential for breakthroughs will also consider epigenetic memory in the system. Some epigenetic marks are heritable across cell divisions (more so in somatic cells than in early embryos), including 5mC DNA methylation, where the inheritance is very stable with a well-understood mechanism. Others, such as H3K27me3 and H3K9me2/me3, may also be inherited, although perhaps with less stability and less fidelity. Whether histone marks associated with transcriptional activation could also be heritable is an open question. A key question here is to what extent epigenetic marks are instructive (e.g., imprinting) or follow transcriptional activation or repression to lock in stabilization of cell fate decisions.

Lineage marking via single-cell sequencing methods will allow us to follow the timing of particular epigenetic changes with regard to the states before the initiation of, during, and post transcription. Furthermore, hairpin bisulphite sequencing (*52*, *53*) (in which methylation information is obtained from both DNA strands) in single cells will identify how heritable methylation is at individual loci and how heterogeneous or homogeneous such heritability is within a cell population. Measurements of 5hmC, 5fC, and 5caC across cell populations, together with mechanistic modeling approaches (*54*, *55*), will allow insights into the generation of epigenetic heterogeneity versus stable inheritance in early development, aging, and disease. The exciting prospect of single-cell epigenome editing (*39*) suggests that detailed

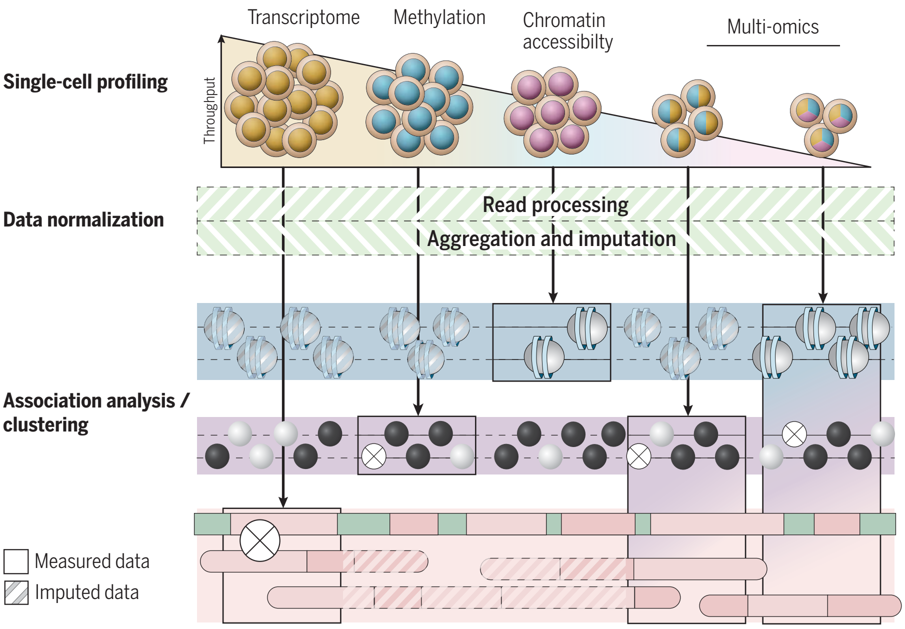

Fig. 3. Multi-omics and computational methods. Shown are typical trade-offs between single-cell RNA-seq, single-cell epigenome protocols, and multi-omics methods that provide readouts from multiple molecular layers in parallel. Consequently, it is commonly required to integrate data from different sequencing protocols. Raw sequence reads from these methods are deduplicated and aggregated into locus-specific readouts, with an optional imputation step to complete missing information. Associations between molecular layers can be used for completing missing data and allow for discovering regulatory associations.

functional testing of epigenetic marks in their various roles may soon become a reality too.

Epigenetic information may also be used to measure cell lineages (Fig. 4B). Lineage-tracing methods using CRISPR scarring have been devised (*45*–*49*), but it is not clear how accurately and reliably they work in different biological settings. Thus, DNA modifications may allow us to trace lineages by marking a particular chromosome or DNA strand, which is segregated into a particular cell type (*17*). This will be especially useful for DNA modifications that are not normally heritable (such as 5hmC, 5fC, or 5caC).

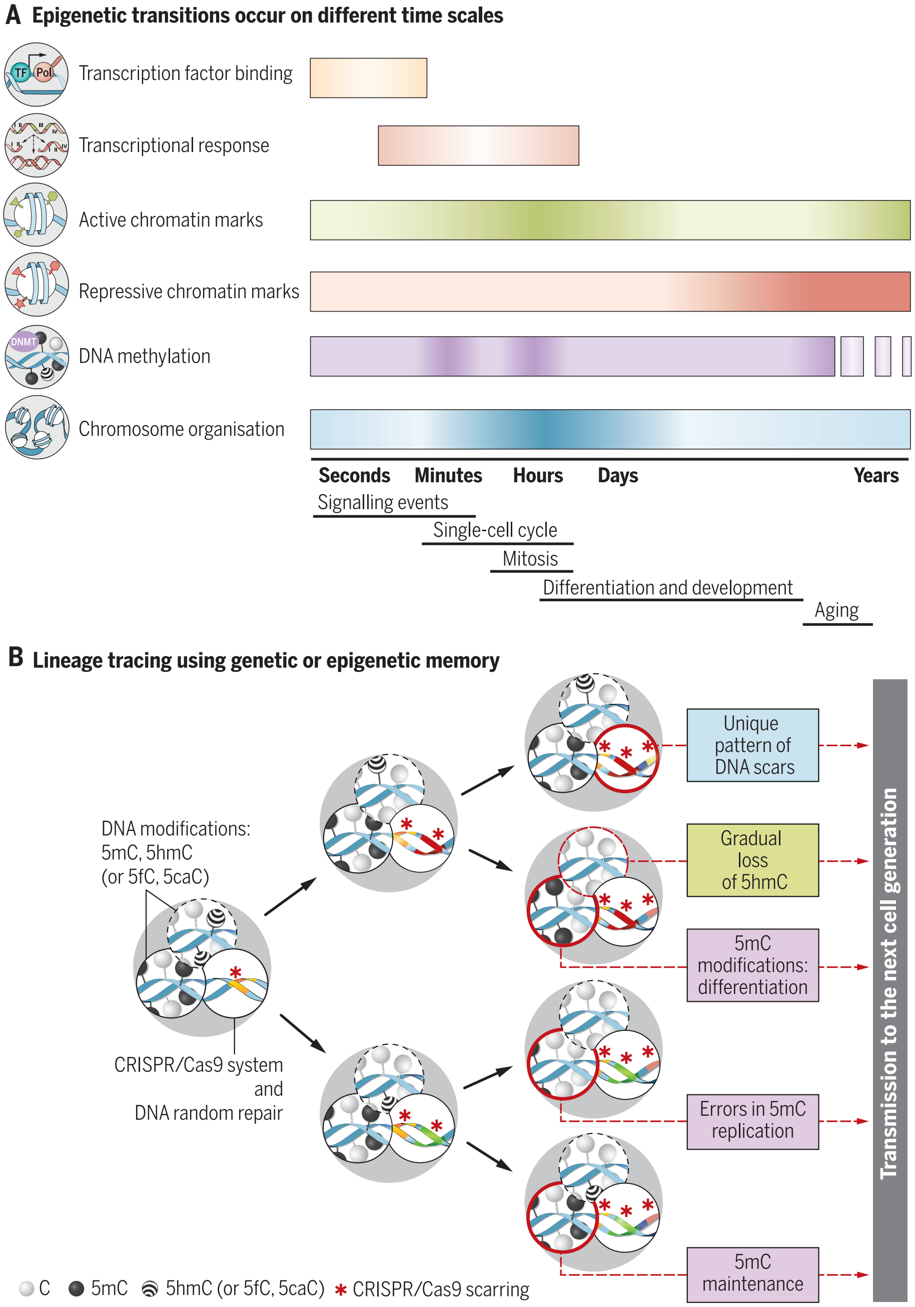
A Epigenetic transitions occur on different time scales
TF
Pol
Transcription factor binding
Transcriptional response
Active chromatin marks
Repressive chromatin marks
DNMT
DNA methylation
Chromosome organisation
Seconds
Minutes
Hours
Days
Years
Signalling events
Single-cell cycle
Mitosis
Differentiation and development
Aging
B Lineage tracing using genetic or epigenetic memory
DNA modifications:
5mC, 5hmC
(or 5fC, 5caC)
CRISPR/Cas9 system
and
DNA random repair
Unique pattern of DNA scars
Gradual loss of 5hmC
5mC modifications: differentiation
Errors in 5mC replication
5mC maintenance
Transmission to the next cell generation
○ C ● 5mC ◎ 5hmC (or 5fC, 5caC) * CRISPR/Cas9 scarring

Fig. 4. Time scales of epigenetic heterogeneity at different layers and lineage tracing. (**A**) Shown are different layers of information that can be recorded at least in principle by single-cell multi-omics, from transcription factor binding and transcriptional responses to long-term epigenetic memory such as is possible with DNA methylation. Rough time scales are indicated by colored bars—with shading indicating transitions in information—and may range from seconds to years. With aging, fidelity of epigenetic information such as DNA methylation may degrade, leading to increased cell-to-cell heterogeneity. (**B**) Lineage tracing using genetic or epigenetic memory. Cell lineage can be traced by CRISPR scarring approaches in which each cell and its descendants within a lineage are linked by unique mutations or barcodes. DNA modifications may also be used to track lineage based on their inheritance and on errors in their maintenance at DNA replication. Nonheritable modifications (5hmC, 5fC, and 5caC) have a short-term lineaging potential, whereas heritable modifications (5mC) have long-term noninvasive lineaging potential.

Some heritable epigenetic marks may be functionally neutral—i.e., set up in early development but simply mechanically copied at each cell division. Because the maintenance methylation machinery has a finite error rate [1 in 25 cell divisions per CpG, although this has only been measured in certain contexts (*56*)], every cell may harbor a unique code of methylation sites that would allow tracking of its developmental trajectory. This acts as if lineage were marked by DNA mutations (either natural ones or induced) (Fig. 4B). This may allow noninvasive lineaging in the future without genetic manipulation, which might be particularly useful in human studies.

We have highlighted the different time scales of variation of these different layers of the epigenome, as well as their interdependencies. It is important to recognize that most of these are from indirect measurements or inferences. In due course, we may connect epigenome dimensions by pseudotime measurements, allowing us to formulate temporal connections and dependencies. However, what is yet to materialize are real-time in vivo recording systems of epigenetic states, ideally at a single-locus level. Hence the single-cell epigenomics revolution has additional challenges to overcome. Our existing methods are already allowing us to zoom in on new concepts of "cell fate"—for example, in developmental systems where cell history can be recorded in epigenetic marks. Yet their actions at key decision points require yet unknown mechanisms (*57*, *58*). This presumably requires new epigenomic codes for cell plasticity and future potential. Deeper insights into these rules will provide not only a better understanding of living biological systems but also new tools and new ways of thinking about changing cell fate experimentally.

At the other end of the spectrum, we anticipate information regarding the presumed degradation of cell fate during aging. Models involve either clonal competition or exhaustion and hence a potential loss of cell heterogeneity in an aging tissue. Conversely, an increase in heterogeneity may occur with a concomitant loss of coherence of transcriptional networks (*59*). Interestingly, programmed changes of the epigenome during aging, particularly of the DNA methylome, accurately record chronological age. However, this "methylation aging clock" can be accelerated or decelerated by biological interventions that shorten or lengthen life span, respectively (*60*–*62*). It remains to be seen how this methylation clock plays out at the single-cell level. As many human adult diseases, including cancer, are associated with altered epigenome patterns, individual cells may gradually and in a potentially programmed way acquire disease risk via changes in epigenetic marks during aging. Conversely, single-cell multi-omics methods may identify hidden cell states with potential for tissue repair or rejuvenation.

As large-scale efforts are mapping all human cells transcriptionally and spatially [e.g., the

Human Cell Atlas (*63*)], there is the prospect in the future that epigenomics measurements, in particular, will add unprecedented layers of information about memory of past experiences and about future potential of cells in the human body.

Outlook

Imagine that we had at our disposal the techniques for single-cell multi-omics, including the ability to identify all key epigenetic modalities, robustly and at an affordable cost. Imagine similarly that we had the computational tools to unravel and visualize connections between the different molecular layers within and between cells. From such advances, we anticipate answering many questions in embryonic development (including comparisons of various organisms). We would like to know any epigenetic determinants of cell fate and lineage decisions and their timing and/or memory of such decisions.

Travelling back in time (i.e., generating iPSCs) or across tissues (via transdifferentiation), we will be able to see how each cell responds in terms of erasing epigenetic memory and acquiring new cell fate trajectories, especially those not part of the normal developmental repertoire. We also anticipate unraveling tissue-level heterogeneity. Highly multiplexed methylome sequencing can already identify cell types in a complex tissue such as the brain with similar accuracy as transcriptome sequencing (*27*).

Finally, we aim to discover links between epigenetic and genetic heterogeneity, showing to what extent epigenetic change (particularly in disease) is driven by underlying changes in DNA sequence such as copy-number variation, mutations, and rearrangements in cancer, or the mobility of selfish DNA elements. Conversely, primary epimutations may underlie the initiation of some diseases but may subsequently elicit more permanent genetic change that stabilizes the disease phenotype.

These advances have implications for diagnosing and understanding disease progression. We envision that precancerous cell states may be detected at an early stage in tissues by their single-cell epigenome signatures, and other chronic diseases may also reveal unique signatures of progression. Single-cell epigenomic analyses might allow for a biopsy of only a few cells or by capturing small amounts of cell-free DNA in circulation. Such tools may also reveal cell populations in tissues with the greatest potential for regeneration and tissue repair.

REFERENCES AND NOTES

1. R. Hooke, *Micrographia: Or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses, with Observations and Inquiries Thereupon* (Courier Corporation, 2003).
2. M. J. T. Stubbington, O. Rozenblatt-Rosen, A. Regev, S. A. Teichmann, *Science* **358**, 58–63 (2017).
3. E. Lein, L. E. Borm, S. Linnarsson, *Science* **358**, 64–69 (2017).
4. C. Gawad, W. Koh, S. R. Quake, *Nat. Rev. Genet.* **17**, 175–188 (2016).
5. C. D. Allis, T. Jenuwein, D. Reinberg, *Epigenetics* (CSHL Press, 2007).
6. W. Jin *et al.*, *Nature* **528**, 142–146 (2015).
7. J. D. Buenrostro *et al.*, *Nature* **523**, 486–490 (2015).
8. D. A. Cusanovich *et al.*, *Science* **348**, 910–914 (2015).
9. F. Guo *et al.*, *Cell Res.* **27**, 967–988 (2017).
10. S. Pott, *eLife* **6**, e23203 (2017).
11. S. J. Clark *et al.*, *bioRxiv* 138685 [Preprint] (17 May 2017).
12. A. Rotem *et al.*, *Nat. Biotechnol.* **33**, 1165–1172 (2015).
13. T. Nagano *et al.*, *Nature* **502**, 59–64 (2013).
14. H. Guo *et al.*, *Genome Res.* **23**, 2126–2135 (2013).
15. S. A. Smallwood *et al.*, *Nat. Methods* **11**, 817–820 (2014).
16. M. Farlik *et al.*, *Cell Reports* **10**, 1386–1397 (2015).
17. D. Mooijman, S. S. Dey, J. C. Boisset, N. Crosetto, A. van Oudenaarden, *Nat. Biotechnol.* **34**, 852–856 (2016).
18. C. Zhu *et al.*, *Cell Stem Cell* **20**, 720–731.e5 (2017).
19. I. C. Macaulay, C. P. Ponting, T. Voet, *Trends Genet.* **33**, 155–168 (2017).
20. I. C. Macaulay *et al.*, *Nat. Methods* **12**, 519–522 (2015).
21. S. S. Dey, L. Kester, B. Spanjaard, M. Bienko, A. van Oudenaarden, *Nat. Biotechnol.* **33**, 285–289 (2015).
22. C. Angermueller *et al.*, *Nat. Methods* **13**, 229–232 (2016).
23. Y. Hu *et al.*, *Genome Biol.* **17**, 88 (2016).
24. Y. Hou *et al.*, *Cell Res.* **26**, 304–319 (2016).
25. C. Angermueller, H. J. Lee, W. Reik, O. Stegle, *Genome Biol.* **18**, 67 (2017).
26. M. Frommer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1827–1831 (1992).
27. C. Luo *et al.*, *Science* **357**, 600–604 (2017).
28. R. M. Mulqueen *et al.*, *bioRxiv* 157230 [Preprint] (2 June 2017).
29. L. F. Cheow, S. R. Quake, W. F. Burkholder, D. M. Messerschmidt, *Nat. Protoc.* **10**, 619–631 (2015).
30. J. R. Peat *et al.*, *Cell Reports* **9**, 1990–2000 (2014).
31. X. Wu, A. Inoue, T. Suzuki, Y. Zhang, *Genes Dev.* **31**, 511–523 (2017).
32. M. R. Corces *et al.*, *Nat. Genet.* **48**, 1193–1203 (2016).
33. S. Preissl *et al.*, *bioRxiv* 159137 [Preprint] (6 July 2017).
34. J. Kind *et al.*, *Cell* **163**, 134–147 (2015).
35. I. M. Flyamer *et al.*, *Nature* **544**, 110–114 (2017).
36. T. Nagano *et al.*, *Nature* **547**, 61–67 (2017).
37. V. Ramani *et al.*, *Nat. Methods* **14**, 263–266 (2017).
38. T. J. Stevens *et al.*, *Nature* **544**, 59–64 (2017).
39. J. van Arensbergen, B. van Steensel, *Mol. Cell* **66**, 167–168 (2017).
40. S. Gravina, X. Dong, B. Yu, J. Vijg, *Genome Biol.* **17**, 150 (2016).
41. W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell, B. E. Engelhardt, *Genome Biol.* **16**, 14 (2015).
42. J. Ernst, M. Kellis, *Nat. Biotechnol.* **33**, 364–376 (2015).
43. F. Buettner *et al.*, *Nat. Biotechnol.* **33**, 155–160 (2015).
44. G. Ficz *et al.*, *Cell Stem Cell* **13**, 351–359 (2013).
45. A. McKenna *et al.*, *Science* **353**, aaf7907 (2016).
46. J. P. Junker *et al.*, *bioRxiv* 056499 [Preprint] (1 June 2016).
47. S. D. Perli, C. H. Cui, T. K. Lu, *Science* **353**, aag0511 (2016).
48. R. Kalhor, P. Mali, G. M. Church, *Nat. Methods* **14**, 195–200 (2017).
49. K. L. Frieda *et al.*, *Nature* **541**, 107–111 (2017).
50. M. Nishizawa *et al.*, *Cell Stem Cell* **19**, 341–354 (2016).
51. X. Huang, J. Wang, *Cell Stem Cell* **20**, 741–742 (2017).
52. C. D. Laird *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 204–209 (2004).
53. L. Zhao *et al.*, *Genome Res.* **24**, 1296–1307 (2014).
54. F. von Meyenn *et al.*, *Mol. Cell* **62**, 983 (2016).
55. P. Giehr, C. Kyriakopoulos, G. Ficz, V. Wolf, J. Walter, *PLOS Comput. Biol.* **12**, e1004905 (2016).
56. T. Ushijima *et al.*, *Genome Res.* **13**, 868–874 (2003).
57. H. J. Lee, T. A. Hore, W. Reik, *Cell Stem Cell* **14**, 710–719 (2014).
58. H. Mohammed *et al.*, *Cell Reports* **20**, 1215–1228 (2017).
59. C. P. Martinez-Jimenez *et al.*, *Science* **355**, 1433–1436 (2017).
60. S. Horvath, *Genome Biol.* **14**, R115 (2013).
61. G. Hannum *et al.*, *Mol. Cell* **49**, 359–367 (2013).
62. T. M. Stubbs *et al.*, *Genome Biol.* **18**, 68 (2017).
63. A. Regev *et al.*, *bioRxiv* 121202 [Preprint] (8 May 2017).
64. B. Lake *et al.*, *bioRxiv* 128520 [Preprint] (19 April 2017).

ACKNOWLEDGMENTS

W.R. thanks I. Herraez, T. Stubbs, S. Clark, C. Alda, H. Mohammed, M. Eckersley-Maslin, S. Rulands, W. Dean, J. Marioni, and B. Simons for discussions or comments on the manuscript. Thank you to V. Juvin (SciArtWork) for artwork. Work in W.R.'s laboratory is supported by the Wellcome Trust, the Biotechnology and Biological Sciences Research Council (BBSRC), and the Medical Research Council (MRC): G.K. is supported by the BBSRC and the MRC; O.S. is supported by European Molecular Biology Laboratory core funding, the Wellcome Trust, and the European Research Council. W.R. is a consultant and shareholder of Cambridge Epigenetix.

10.1126/science.aan6826

Science

Single-cell epigenomics: Recording the past and predicting the future

Gavin Kelsey, Oliver Stegle and Wolf Reik

Science **358** (6359), 69-75.
DOI: 10.1126/science.aan6826

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.