Sequencing in High Definition Drives a Changing Worldview of the Epigenome

Emily Hodges

Department of Biochemistry and Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, Tennessee 37232

Correspondence: emily.hodges@vanderbilt.edu

Single-molecule sequencing approaches have transformed the study of the human epigenome, accelerating efforts to describe genome function beyond the sequences that encode proteins. The post-genome era has ignited strong interest in the noncoding genome and profiling epigenetic signatures genome-wide have been critical for the identification and characterization of noncoding gene-regulatory sequences in various cellular and developmental contexts. These technologies enable quantification of epigenetic marks through digital assessment of DNA fragments. With the capacity to probe both the DNA sequence and count DNA molecules at once with unparalleled throughput and sensitivity, deep sequencing has been especially transformative to the study of DNA methylation. This review will discuss advances in epigenome profiling with a particular focus on DNA methylation, highlighting how deep sequencing has generated new insights into the role of DNA methylation in gene regulation. Technical aspects of profiling DNA methylation, remaining challenges, and the future of DNA methylation sequencing are also described.

From locating key mutations driving rare tumor initiating cells to retrieving ancient Neanderthal sequences from highly contaminated bone fossils, next-generation sequencing (NGS) has led to groundbreaking discoveries across a number of biological disciplines. These discoveries are grounded in the ability to perform genome-scale detection and quantification of specific DNA molecules within a complex population of fragments. Techniques developed for mapping the epigenome take advantage of these capabilities as well as the range of detection levels offered by NGS.

By and large, “seq”-based methodologies rely on sequence-based tag counting to quanti-

tatively assess transcriptomic and epigenomic features (Wang et al. 2009). This concept was first introduced more than two decades ago as a way to measure RNA transcript abundance (Velculescu et al. 1995; Brenner et al. 2000). At that time, tag-counting methods offered a number of advantages over the more prolific, hybridization-based microarray platforms, including direct quantification of transcript levels and greater dynamic range of detection, all without prior knowledge of the gene sequence; but the cost associated with Sanger sequencing, among other things, limited their widespread usage. High-density oligonucleotide microarrays, on the other hand, could be programmed to either

Editors: W. Richard McCombie, Elaine R. Mardis, James A. Knowles, and John D. McPherson
Additional Perspectives on Next-Generation Sequencing in Medicine available at www.perspectivesinmedicine.org

Copyright © 2019 Cold Spring Harbor Laboratory Press; all rights reserved; doi: 10.1101/cshperspect.a033076
Cite this article as *Cold Spring Harb Perspect Med* 2019;9:a033076

E. Hodges

selectively probe subsets of the genome, including transcript regions, exons or CpG islands, or tile the entire genome. The relative cost and throughput of the microarray made it an ideal medium for determining relative abundances of a number of functional genomic characteristics, including chromatin modifications and DNA methylation. Still, despite flexibility and throughput, microarrays provide very limited potential for new discoveries because they are predicated on knowing your sequence of interest, not to mention their rigid signal detection range. In contrast, NGS is mostly indifferent to sequence content, and sensitivity is limited only by sequencing depth and library complexity. In other words, the deeper the sequencing, the better the opportunity to detect even the rarest of DNA fragments.

Recognizing the great potential of NGS for epigenome discovery, the appearance of commercially available NGS platforms was soon followed by a series of publications describing post-translational modifications of histones, DNA methylation, transcription factor (TF) binding, and chromatin accessibility at base pair resolution (Albert et al. 2007; Barski et al. 2007; Johnson et al. 2007; Mikkelsen et al. 2007; Boyle et al. 2008; Cokus et al. 2008). These studies were important not only for the new insights revealed, but also because they showed the feasibility of adapting existing molecular techniques to a completely new detection medium. Following more than a decade of NGS use, new discoveries in epigenetics have driven changing worldviews of genome function, shifted paradigms, and even challenged the central dogma of genetics (noncoding RNAs being a good example [Sabin et al. 2013]). This is perhaps best exemplified by DNA methylation, a mark that has often mystified molecular biologists since the mid-1970s, when it was first correlated with transcriptional repression. Described below is how DNA methylomes (genome-wide levels of methylation) are generated and analyzed, how a growing body of genome-wide data has greatly revised the view of DNA methylation within the scope of gene regulation, and how new technologies on the horizon may address current challenges in DNA methylation studies.

REVISING THE ROLE OF DNA METHYLATION IN GENE REGULATION

Defining what constitutes “epigenetic” has been a somewhat contentious topic among the field’s most devout. But in recent years, in the wake of NGS and with a more comprehensive view of the noncoding genome, it is clear that many aspects of genome function contribute to the establishment and propagation of an epigenetic state. In 2009, following a meeting of the chromatin community at Cold Spring Harbor Laboratory, a more inclusive “operational” definition of epigenetics was proposed (Berger et al. 2009). By this definition, an epigenetic phenotype is established by an ordered process that involves an “epigenator” (a transient signal such as an environmental cue), followed by an “initiator” (TF, noncoding RNA, or other entity that defines location on a chromosome and persists at that location), and lastly a “maintainer” (DNA methylation, as well as some histone modifications and variants). As a maintainer, DNA methylation sustains an epigenetic state either by transmitting the signal through subsequent cell divisions or by stabilizing the signal in terminally differentiated cells. It is doubtful that anyone would dispute the significance of DNA methylation in the epigenome order, but questions regarding causality and function remain unresolved.

DNA methylation has been described as a paradox (Jones 1999). The idea that DNA methylation blocks transcriptional initiation, but not elongation, or that promoter hypomethylation does not always beget gene activation seems a contradiction. In cancer genomes, a somewhat universal contradiction has been observed—abnormal promoter methylation of a few important genes with widespread loss of methylation in regions depleted of genes. But here is the crux of the problem. DNA methylation is prevalent in vertebrate genomes, occurring at 70%–80% of CpG dinucleotides. The modification, or at least the enzymes responsible for depositing and maintaining it (DNMT3A,B and DNMT1, respectively), is essential for normal development and cellular differentiation as well as stem cell self-renewal, yet other eukaryotes seem perfectly

content to perform gene regulation without it. CpGs occur at one-quarter the frequency of other dinucleotide combinations in the human genome, and this depletion has been attributed to higher mutation frequencies of methylated cytosines. In contrast, many promoter regions and other discrete loci throughout mammalian genomes contain sequences rich in CpG sites. The hypomethylation of such regions, designated CpG islands, is typically stable across cell types and developmental states with the exception of a few promoters of highly cell-type-specific genes. The preservation of CpG sites within these islands, along with the fact that until recently CpGs were the only known target of mammalian cytosine methylation, implicated them as the most important if not only functional references for DNA methylation studies. (For comprehensive reviews on the history of DNA methylation, refer to Bird 1986 and Jones 2012.)

Just as NGS platforms were gaining traction, a seminal study reported high-methylation variability in CpG island “shores” (Doi et al. 2009). This study compared DNA methylation in tumor versus normal colon tissue using a microarray-based technique called CHARM (comprehensive high-throughput arrays for relative methylation) (Irizarry et al. 2008). The array design focused on gene promoters as well as non-promoter CpG islands and their surrounding sequences (shores). Islands were defined using both heuristic criteria and statistical modeling of sequence characteristics (Gardiner-Garden and Frommer 1987; Irizarry et al. 2009). This becomes an important point when we discuss NGS and DNA methylation. The methylation variability in CpG island shores, but not in CpG islands, was a significant observation that challenged the long-held belief that CpG island promoters were the only places where DNA methylation mattered. The differential methylation was also strongly correlated with gene expression differences and appeared to distinguish different cell types more clearly than the methylation states of CpG islands themselves. So, what about the rest of the genome?

Soon after the publication of these studies, the first human methylomes using whole-genome bisulfite sequencing ([WGBS] discussed

in further detail below) were reported (Lister et al. 2009). For the first time, these studies revealed significant non-CG methylation in embryonic and induced pluripotent stem cells. Detecting non-CG methylation would not have been feasible without deep sequencing, given that 85% of mCHG and mCHH sites, in which H stands for any nucleotide except guanine, were in some cases only 10% methylated, as well as the fact that previous approaches focused entirely on CpG sites. An unexpected prevalence of partially methylated domains (PMDs) was also observed on autosomes of immortalized fibroblasts. These regions consistently display lower methylation levels in both immortalized cell lines and in many cancer cell types compared with normal primary cells. Importantly, a high correspondence between PMDs and nuclear lamin-associated domains has been observed, suggesting that methylation states are highly dependent on genome structure (Berman et al. 2011).

WGBS confirmed in very high definition a truth that had already been established by much lower resolution methods—that most of the CpG sites in the genome are methylated. A completely new discovery, however, was that the noncoding genome is scattered with discrete hypomethylated regions (HMRs) that do not correspond to known or predicted CpG islands (Hodges et al. 2011; Stadler et al. 2011). More importantly, DNA methylation levels in intergenic and intronic HMRs (iHMRs) are more variable than other regions impacted by DNA methylation (Fig. 1). These patterns are highly cell-type-specific, frequently co-occurring with nucleosome depletion, posttranslational histone tail modifications, including H3K27 acetylation, and TF occupancy; all of these point to the presence of an enhancer element. So, comparative WGBS can also identify putative cell-type-specific enhancers (Schlesinger et al. 2013).

Enhancers are hubs for TF activity and their regulation lies at the heart of cellular identity. Normal enhancer function involves dynamic interactions between TFs, the genome sequence, and the epigenome, which then trigger the topological reorganization of chromatin through highly ordered enzymatic events. Recent studies

E. Hodges

Figure 1. Global comparison of methylation levels in different cell types. Average methylation was calculated for promoters and intergenic regions (a union of intergenic and intronic hypomethylated regions [iHMRs] from every cell type) for blood cell types (hematopoietic stem and progenitor cells [HSPCs]), brain cell types, and sperm as the outgroup. Pearson correlation and hierarchical clustering of methylation levels recapitulates the relationships between cell types in both promoters and iHMRs. However, differences in methylation are much greater in intergenic regions, indicating these regions harbor cell-type-specific gene regulatory elements.

have suggested that different classes of enhancers are characterized by diverse combinations of epigenetic marks, including DNA methylation (or hypomethylation), and these combinations may represent stages of a stepwise process of gene regulation ultimately dictated by TF interactions with the cell's DNA sequence. Together, these events control patterns of gene expression and guide the specification of cellular function and phenotype. Analysis of steady-state data from cell types representing different stages of development revealed that terminally differentiated cells show four times the number of iHMRs compared with embryonic stem cells. Additionally, a substantial subset of DNase hypersensitive sites identified in stem cells is methylated (Schlesinger et al. 2013). Many of these accessible regions appear to both maintain accessibility and lose methylation in the mature cell, gaining additional markers of enhancer activity. This suggests that loss of methylation is secondary to nucleosome repositioning during enhancer activation.

The order and timescale by which DNA methylation is gained or lost at enhancers during cell fate determination is not well understood. New studies suggest that TFs, especially pioneer factors, play an important role in directing the methylation state of a genomic locus

(Donaghey et al. 2018; Mayran et al. 2018). TF affinity for methylated DNA adds another layer of complexity to the process, as sequential recruitment of TFs occurs at regulatory loci during differentiation, and some TFs bind methylated DNA, whereas others do not. This also implies that demethylating mechanisms, whether active or passive, are also at play. Indeed, oxidation derivatives of 5-methylcytosine, including 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine, have been detected at promoters and enhancers of embryonic stem cells and differentiating neural progenitors (Serandour et al. 2012; Melamed et al. 2018).

How DNA methylation fits into current models of gene regulation is certainly an area of intense research focus. To answer any of these questions requires a look at DNA methylation in the direct context of chromatin, so multidimensional genome-wide assays that integrate genomic information are required. Examples of these types of assays include NOME-seq (Kelly et al. 2012), which measures both nucleosome occupancy and DNA methylation in parallel, or ChIP-BS (Li and Tollefsbol 2011; Statham et al. 2012), which couples chromatin immunoprecipitation with bisulfite sequencing. More importantly, continuous datasets evaluated across time, rather than static and independent data

comparisons, are also required to understand the timescale of DNA methylation changes and how these changes influence the stabilization of cellular phenotypes. Understanding these spatiotemporal relationships will be critical for understanding the precise role of DNA methylation in gene regulatory mechanisms that define normal development and disease.

USING NGS TO DETERMINE DNA METHYLATION GENOME-WIDE

Bisulfite sequencing has long been the gold standard for quantitative assessment of DNA methylation at base-pair resolution. Before NGS, a typical bisulfite sequencing experiment involved genomic DNA extraction from the cell type of interest, strand separation, and cytosine deamination by sodium bisulfite, followed by PCR with oligonucleotide primers that target a specific locus such as a promoter or CpG island. PCR amplicons undergo bacterial cloning and multiple colonies are selected for Sanger sequencing. In this approach, the more colonies evaluated, the more accurate the estimate of methylation at a given CpG site. Like many other applications of the NGS platform, WGBS superseded this approach, cost notwithstanding, because it bypasses the cloning step and does not require the selection of specific genomic loci; however, reimagining the bisulfite sequencing method for the NGS platform demands several important considerations: (1) Efficient deamination of unmethylated cytosines requires single-stranded DNA, which precludes performance of bisulfite conversion before double-stranded adaptor ligation; (2) adaptors added before bisulfite treatment must be methylated to preserve the adaptor sequence for NGS platform compatibility; (3) library amplification is always performed post-bisulfite treatment as PCR performed with unmethylated dNTPs removes the native methylation status of the cytosine; (4) bisulfite treatment is damaging to high-molecular-weight DNA, resulting in a population of fragments with an average size below 500 bp; and (5) amplification of DNA requires a specialized polymerase capable of reading uracil-rich templates.

Epigenetics and Next-Generation Sequencing

In early versions of the WGBS protocol, sonicated DNA would undergo end-repair, adenylation, and adaptor ligation before bisulfite conversion (Cokus et al. 2008). Although the protocol is fairly reproducible and robust, loss of input at every step reduces fragment library complexity, necessitating higher amounts of starting material on the front end. An alternative approach using Tn5 transposase accomplishes fragmentation and ligation in the same reaction (“tagmentation”), reducing the number of enzymatic steps performed, and requiring only 10% of the input needed in the standard protocol (Adey and Shendure 2012; Wang et al. 2013). However, a major caveat to this approach is the 9 bp gap created by the Tn5 insertion on the “bottom” strand of the DNA fragment. The gap must be enzymatically repaired with dNTP replacements before bisulfite conversion and PCR, resulting in loss of methylation status for any replaced CpG site, which eventually must be dealt with during bioinformatic analysis. In both of the aforementioned protocols, adaptors are introduced before bisulfite treatment, so any damage incurred during treatment could compromise the integrity of fragment ends, thus biasing the size distribution of sequenceable fragments and reducing the complexity of fragments sampled. More recently, WGBS protocols perform bisulfite treatment directly on high-molecular-weight genomic DNA, which results in a fragmented pool of single-stranded DNA. Random primed synthesis and primer tailing of the complement strand creates the first “adapted” end of the fragment (Miura et al. 2012). In some instances of the protocol, a ligation step adds a 3′ adaptor to the newly synthesized strand to create a “di-tagged” fragment, followed by PCR. In other protocols, a second round of extension is performed with random primers to add the second adaptor. Because adaptation of fragment ends is performed post-bisulfite, loss of sequenceable fragments is significantly lower, thereby reducing the amount of input required. Strand specificity is also preserved, which can be an important advantage for determining allele-specific differences in methylation.

E. Hodges

Bioinformatic Considerations for Cost, Coverage, and Sample Number

To determine sequence coverage requirements for WGBS, it is important to understand the analysis workflow. During bisulfite treatment, unmethylated cytosines are deaminated to a uracil base, and during library amplification the uracil is replaced by a thymine. This means that an unmethylated “C” will be reported as “T” in the sequence. The process of mapping the converted sequence back to a reference genome can be computationally intensive and time consuming because of the reduced nucleotide complexity of the sequence, and existing mapping algorithms attempt to mitigate this issue in different ways. Sequenced reads are typically mapped back to a reference genome by one of two strategies, either mapping to an “unconverted” reference genome and allowing for C to T

mismatches, or mapping to a “converted” reference genome, which accounts for the conversion of most cytosines outside the CpG context (Krueger and Andrews 2011; Chen et al. 2016). Once the genome mapping locations of the reads are determined, the fraction of Cs mapping to Ts is calculated and is depicted as a ratio from 0 to 1 (Fig. 2). A segmentation algorithm will identify regions of low methylation (hypomethylation) or high methylation (hypermethylation), or differentially methylated regions (DMRs) will be determined between samples (Song et al. 2013). Given that DNA methylation levels are continuous and measured on an absolute scale, coverage is an important consideration. Coverage recommendations range anywhere from 10×, which is the level typically targeted, to 30×, which is the coverage level recommended by the Roadmap Epigenome Project (see www.roadmapepigenomics.org/protocols).

Figure 2. Mapping DNA methylation across the genome. Genomic locus and data browser tracks depict DNA methylation levels for various cell types, including B cells (GM12878), liver, H1 embryonic stem cells, and sperm. Tracks were generated with the University of California at Santa Cruz (UCSC) genome browser (see genome.ucsc.edu). Methylation levels are calculated as the number of Cs detected, divided by the read coverage at a given CpG site (yellow bars). Horizontal blue bars are hypomethylated regions identified by MethPipe (Song et al. 2013). Other tracks represent overlapping acetylated H3K27 signals from ENCODE ChIP-seq data (see www.encodeproject.org) and Refseq annotated gene transcripts. Stable hypomethylation is observed across multiple cell types at the ARGEF1 promoter, whereas CD79 hypomethylation is restricted to B cells. Putative enhancer regions with H3K27ac are also hypomethylated exclusively in B cells.

In effect, setting a coverage target level really depends on the relationships between the samples to be compared (i.e., two highly related cell types such as CD4⁺ or CD8⁺ T cells or two distantly related cell types such as neurons and hepatocytes). A recent study reported diminishing returns of true-positive DMRs above 10× coverage for unrelated cell types, whereas 15× was recommended for comparisons of highly similar cell types caused by higher false discovery rates of DMRs at lower coverage (Ziller et al. 2015). Overall, DMR discovery was improved by having two biological replicates at 10× coverage.

In practice, achieving a 10× sequencing depth would require roughly 250 million paired-end 150 bp reads with a 70% mapping rate (65%–75% mapped bisulfite sequencing reads is considered a good mapping rate with most of the algorithms currently in use). Newer platforms with higher throughput capabilities, such as the Novaseq (see Illumina, www.illumina.com), are driving down the cost of WGBS, but performance of large-scale studies with many patient samples is still somewhat cost-prohibitive. Motivated to reduce the cost of increasing sample size, a number of groups have opted to profile a subset of the genome at higher coverage. Methodologies that enable selective methylome sequencing include methylated DNA immunoprecipitation (MeDIP), reduced representation bisulfite sequencing (RRBS), and targeted enrichment techniques that involve sequence capture with oligonucleotide probes. In MeDIP, methylated DNA is immunoprecipitated with an antibody recognizing methylated cytosine (Down et al. 2008; Taiwo et al. 2012). Because the approach is limited by antibody specificity, and given the majority of the genome is methylated, MeDIP is not as sensitive as other methods at identifying methylation differences between samples.

RRBS uses the methylation insensitive restriction enzyme MspI, which has a short recognition sequence high in GC content and contains a CpG site (C⁺CGG) (Meissner et al. 2005). MspI sites are heavily enriched at promoters and CpG islands, but less so in noncoding regions where CpGs are less frequent. To improve coverage of CpG island shores, enhanc-

Epigenetics and Next-Generation Sequencing

ers, and other regions of gene regulatory importance, an enhanced RRBS (ERRBS) protocol was developed (Akalın et al. 2012). Typically, ERRBS experiments can determine methylation levels for around 10%–12% of the 28 million CpG sites in the human genome and have been used to identify DNA methylation patterns that distinguish patients with genetically distinct cancer subtypes or treatment outcomes.

Unlike enzyme-based approaches, NGS-coupled sequence capture methods use libraries of synthesized oligonucleotides programmed to selectively enrich DNA fragments from genomic regions of interest followed by NGS (Ball et al. 2009; Deng et al. 2009; Hodges et al. 2009). A number of variations on this technique have been reported, several of which are commercially available with off-the-shelf reagents, including pooled oligos targeting a curated set of genomic loci (Li et al. 2015). Hybridization protocols are also compatible with sample multiplexing, keeping the relative cost of sequence enrichment low, provided a minimal number of samples/reactions are met. These approaches are currently capable of interrogating twice as many CpG sites compared with ERRBS; however, as with any predetermined set of loci, the potential for discovery is limited to the genomic space surveyed and may not include unknown regions of biological significance. Of course, with accumulating functional genomic data from a growing list of cellular and developmental states, the target space is evolving.

DIRECT ACCESS TO DNA METHYLATION WITH THIRD-GENERATION TECHNOLOGY

A majority of the applications developed for epigenetic questions have been adapted for short-read platforms, where paired read-lengths currently do not exceed 300 bp because of the combined limitations of the chemistry and hardware that comprise the technology. Third-generation or long-read platforms, on the other hand, permit direct ascertainment of certain epigenetic marks in real time from multi-kilobase continuous sequences. In PacBio (see Pacific Biosciences of California, www.pacb.com) sequencing, hairpin adaptors are ligated to dou-

E. Hodges

ble-stranded DNA to create circular templates. These templates are loaded, without amplification, into nanoscale chambers called “zero mode waveguides,” which record polymerase incorporation of fluorescently labeled dNTPs in real time. DNA methylation can be detected from delays in the “interpulse duration,” or time between nucleotide incorporations (Schadt et al. 2013). With circular templates, a sequence can be read multiple times, and the number of revolutions varies according to polymerase duration and fragment insert length. A greater number of passes around the sequence is correlated with higher quality base determination. Therefore, shorter inserts have fewer error rates than longer inserts. This is especially important for profiling nucleic acid modifications because changes in nucleotide incorporation kinetics can be subtle and therefore difficult to detect without multiple passes of the template.

Nanopore technology, for example, Nanopore Technologies platforms (see www.nanoporetech.com), in contrast with other single-molecule sequencing platforms, is based on the behavior of ion currents rather than fluorescent nucleotide incorporation. As a DNA template is driven through a tiny pore, the ion current is modulated by the different nucleotides that pass through the narrowest part of the pore. Nucleotide sequences can be determined in real time as the current is recorded, and fluctuations in current (i.e., size and shape of the current) may indicate a nucleic acid modification (Rand et al. 2017; Simpson et al. 2017). Although DNA modifications are better distinguished by ion current differences than by changes in polymerase kinetics, ion current differences are strongly influenced by characteristics of the surrounding sequence, and comparison to an unmodified reference sequence is required. Still in its early stages, computational methods are being developed to model the behaviors of the ion current differences that characterize nucleic acid modifications to better detect them with nanopore sequencing.

Long-read platforms have substantially lower throughput capabilities than short-read platforms, which limit their practical usage in large-scale, genome-wide studies. Nonetheless,

there are a number of scenarios related to DNA methylation in which these technologies are ideally suited. For example, DNA methylation is critical for the suppression of transposons and other repetitive elements. Failure to suppress the activities of these elements can lead to infertility and disease, and numerous studies have sought to identify DNA methylation differences that distinguish disease versus normal states within these genomic loci. Longer reads have the advantage of higher, more accurate mapping rates in repetitive areas of the genome, which can be problematic for short bisulfite converted sequences. In another example, because DNA methylation is depicted as a ratio of methylated and unmethylated cytosines from a population of fragments, the methylation call at a given CpG site is not always binary. In fact, regions of partial or intermediate methylation are found throughout the genome. Mixed methylation states may result from allele-specific methylation, heterogeneity within the cell population, or possibly stochastic loss of methylation maintenance during DNA replication (discussed in more detail below). Long-read platforms provide a contiguous profile of DNA methylation over a longer stretch of the same DNA molecule establishing an epitype akin to a haplotype, which may help to resolve the origin of DNA fragments contributing to a methylation state.

Single-Cell Methylome Sequencing

Single-cell techniques attempt to characterize molecular signatures of individual cells to understand the diversity of signals that make up the collective behavior of a tissue. Single-cell RNA-seq (scRNA-seq) studies have shown transcriptional heterogeneity among ostensibly identical cell populations, which may be a reflection of underlying epigenetic heterogeneity (Angermueller et al. 2016). Most transcripts within a cell have multiple copies, and only a small fraction of the genome is transcribed. As a result, analysis of scRNA-seq data is less problematic in terms of coverage than data derived from a two-copy genome. Based on these and other technical challenges, a limited number of epigenetic features have been probed at single-cell resolu-

tion, including DNA methylation (Kelsey et al. 2017). To generate DNA methylation profiles from single cells, a modified version of the post-bisulfite adaptor tagging method, which involves random primer extension of bisulfite converted templates (mentioned earlier), has been described (Clark et al. 2017). Depending on the sequencing depth, coverage of up to 48% of CpG sites has been reported using this approach (reaching a saturation level at 50 million reads), which is a significant improvement on the scBS-seq protocols that were first reported.

Bisulfite exposure is damaging to DNA, restricting the potential of any protocol involving bisulfite treatment to achieve full genome coverage from a single cell. Bioinformatic analysis of scBS-seq data attempts to compensate for the lack of full genome coverage to reconstruct a single-cell methylome using some or all of the following strategies: imputation of methylation states for missing data based on the assumption that adjacent CpG sites are congruent, averaging methylation across a multi-kilobase sliding window of covered CpG sites, or merging single-cell data and comparing the merged data with levels obtained from bulk samples. The limitation with any approach is that single-nucleotide resolution, as well as single-cell resolution may be forfeited in correcting for missing data. Allele-specific information is also lost, as coverage is mostly limited to one chromosome.

Non-allelic methylation heterogeneity, which gives rise to mixed or partial methylation measurements, is a confounding factor in DNA methylation profiles from bulk cells. Noisy data can obscure the identification of biologically meaningful DMRs, and scBS-seq may help to disentangle signal from noise along with understanding the factors that contribute to both. These factors include cell heterogeneity, or cells undergoing asynchronous (or asymmetric) cell fate transitions. For example, scBS-seq has so far revealed significant epigenetic heterogeneity in mouse liver and embryonic stem cells and has been used to construct the methylome of mouse oocytes (Smallwood et al. 2014; Gravina et al. 2016). Other factors contributing to methylation heterogeneity can include deficiencies in faithful transmission and maintenance of

DNA methylation states during replication. Indeed, stochastic methylation variation has been observed in regions of the genome outside of promoters and CpG islands where methylation patterns are more stable (Gravina et al. 2016). Single-cell studies suggest that as much as 3% of CpG sites are variable between cells originating from the same tissue, and this “epivariation” may be attributed to errors in methylation maintenance, frequently occurring in transposons and other repetitive elements. Single-cell methylome profiling is still in the early stages of development, and improvements to methodologies will likely focus on increasing scalability of the assay to thousands of cells (current scales are 100–1000 cells) as well as widening breadth of coverage. Regardless, single-cell epigenomics holds enormous potential for understanding the epigenetic basis of developmental trajectories within lineages, for defining cell type of origin among rare tumor cells, or studying physiological variation within a complex tissue.

CONCLUDING REMARKS

It is becoming increasingly clear that a majority of disease- and trait-associated sequence variants are located in noncoding regions of the genome, a substantial portion of which is enriched in functional regulatory elements. Moreover, genome-wide study variants localize primarily within cell-type-specific regulatory DNA of disease-relevant tissues (Maurano et al. 2012). Recent studies have also revealed that cell-type-specific enhancers show higher DNA methylation variability between individuals than other genomic elements, which may be relevant to interindividual differences in gene regulation and to disease susceptibility. Central to these observations is the proposition that genotype and methylation state are strongly linked, and methylation changes that accompany genetic modifications could be used to infer the activity status of the enhancer. The strong connection between genotype and “epitype” is reflected in the observation that, for many TFs, stable occupancy at target sites relies on a DNA sequence free of cytosine methylation and nucleosome interference (Jones 2012). Furthermore, TF binding ap-

E. Hodges

pears to be a criterion for loss of methylation upon enhancer activation. The implication of this possibility is that sequence variation within TF binding sites may alter binding affinity and consequently disrupt demethylation of the enhancer, potentially deregulating long-term enhancer stability. Ultimately, pairing sequence information with methylation levels across individuals will be an important step in understanding the relationship between sequence variation and methylation state, providing initial clues toward the impact of this noncoding sequence variability on gene regulatory activity.

ACKNOWLEDGMENTS

I thank my colleagues, and Kelly Barnett in particular, for helpful discussions in preparing this manuscript. E.H. is supported by funding from the National Institutes of Health (NIH) (NCI K22 CA184308).

REFERENCES

- Adey A, Shendure J. 2012. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res* 22: 1139–1143.
- Akalin A, Garrett-Bakelman FE, Kormaksson M, Busuttill J, Zhang L, Khrebtkova I, Milne TA, Huang Y, Biswas D, Hess JL, et al. 2012. Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genet* 8: e1002781.
- Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF. 2007. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* 446: 572–576.
- Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood S, Ponting CP, Voet T, et al. 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* 13: 229–232.
- Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM. 2009. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* 27: 361–368.
- 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: 823–837.
- Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. 2009. An operational definition of epigenetics. *Genes Dev* 23: 781–783.
- Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange CP, van Dijk CM, Tollenaar RA, et al. 2011. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet* 44: 40–46.
- Bird AP. 1986. CpG-rich islands and the function of DNA methylation. *Nature* 321: 209–213.
- 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: 311–322.
- Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 18: 630–634.
- Chen H, Smith AD, Chen T. 2016. WALT: Fast and accurate read mapping for bisulfite sequencing. *Bioinformatics* 32: 3507–3509.
- Clark SJ, Smallwood SA, Lee HJ, Krueger F, Reik W, Kelsey G. 2017. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat Protoc* 12: 534–547.
- Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulfite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452: 215–219.
- Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J, et al. 2009. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* 27: 353–360.
- Doi A, Park IH, Wen B, Murakami P, Aryee MJ, Irizarry R, Herb B, Ladd-Acosta C, Rho J, Loewer S, et al. 2009. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet* 41: 1350–1353.
- Donaghey J, Thakurela S, Charlton J, Chen JS, Smith ZD, Gu H, Pop R, Clement K, Stamenova EK, Karnik R, et al. 2018. Genetic determinants and epigenetic effects of pioneer-factor occupancy. *Nat Genet* 50: 250–258.
- Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM, et al. 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* 26: 779–785.
- Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. *J Mol Biol* 196: 261–282.
- Gravina S, Dong X, Yu B, Vijg J. 2016. Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome. *Genome Biol* 17: 150.
- Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M, Zhang MQ, Ye K, Bhattacharjee A, Brizuela L, et al. 2009. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res* 19: 1593–1605.
- Hodges E, Molaro A, Dos Santos CO, Thekkat P, Song Q, Uren PJ, Park J, Butler J, Rafii S, McCombie WR, et al. 2011. Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol Cell* 44: 17–28.

- Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, Jeddeloh JA, Wen B, Feinberg AP. 2008. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res* **18**: 780–790.
- Irizarry RA, Wu H, Feinberg AP. 2009. A species-generalized probabilistic model-based definition of CpG islands. *Mamm Genome* **20**: 674–680.
- Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**: 1497–1502.
- Jones PA. 1999. The DNA methylation paradox. *Trends Genet* **15**: 34–37.
- Jones PA. 2012. Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**: 484–492.
- 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**: 2497–2506.
- Kelsey G, Stegle O, Reik W. 2017. Single-cell epigenomics: Recording the past and predicting the future. *Science* **358**: 69–75.
- Krueger F, Andrews SR. 2011. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571–1572.
- Li Y, Tollefsbol TO. 2011. Combined chromatin immunoprecipitation and bisulfite methylation sequencing analysis. *Methods Mol Biol* **791**: 239–251.
- Li Q, Suzuki M, Wendt J, Patterson N, Eichten SR, Hermanson PJ, Green D, Jeddeloh J, Richmond T, Rosenbaum H, et al. 2015. Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. *Nucleic Acids Res* **43**: e81.
- Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315–322.
- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**: 1190–1195.
- Mayran A, Khetchoumian K, Hariri F, Pastinen T, Gauthier Y, Balsalobre A, Drouin J. 2018. Pioneer factor Pax7 deploys a stable enhancer repertoire for specification of cell fate. *Nat Genet* **50**: 259–269.
- Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**: 5868–5877.
- Melamed P, Yosefzon Y, David C, Tsukerman A, Pnueli L. 2018. Tet enzymes, variants, and differential effects on function. *Front Cell Dev Biol* **6**: 22.
- Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Gianoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**: 553–560.
- Miura F, Enomoto Y, Dairiki R, Ito T. 2012. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res* **40**: e136.
- Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B. 2017. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods* **14**: 411–413.
- Sabin LR, Delas MJ, Hannon GJ. 2013. Dogma derailed: The many influences of RNA on the genome. *Mol Cell* **49**: 783–794.
- Schadt EE, Banerjee O, Fang G, Feng Z, Wong WH, Zhang X, Kislyuk A, Clark TA, Luong K, Keren-Paz A, et al. 2013. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Res* **23**: 129–141.
- Schlesinger F, Smith AD, Gingeras TR, Hannon GJ, Hodges E. 2013. De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. *Genome Res* **23**: 1601–1614.
- Serandour AA, Avner S, Oger F, Bizot M, Percevault F, Lucchetti-Miganeh C, Palierne G, Gheeraert C, Barloy-Hubler F, Peron CL, et al. 2012. Dynamic hydroxymethylation of deoxyribonucleic acid marks differentiation-associated enhancers. *Nucleic Acids Res* **40**: 8255–8265.
- Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407–410.
- Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G. 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* **11**: 817–820.
- Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD. 2013. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS ONE* **8**: e81148.
- Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, et al. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**: 490–495.
- Statham AL, Robinson MD, Song JZ, Coolen MW, Stirzaker C, Clark SJ. 2012. Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Res* **22**: 1120–1127.
- Taiwo O, Wilson GA, Morris T, Seisenberger S, Reik W, Pearce D, Beck S, Butcher LM. 2012. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc* **7**: 617–636.
- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science* **270**: 484–487.
- Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57–63.
- Wang Q, Gu L, Adey A, Radlwimmer B, Wang W, Hovestadt V, Bahr M, Wolf S, Shendure J, Eils R, et al. 2013. Tagmentation-based whole-genome bisulfite sequencing. *Nat Protoc* **8**: 2022–2032.
- Ziller MJ, Hansen KD, Meissner A, Aryee MJ. 2015. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nat Methods* **12**: 230–232.

Cold Spring Harbor Perspectives in Medicine

Sequencing in High Definition Drives a Changing Worldview of the Epigenome

Emily Hodges

Cold Spring Harb Perspect Med 2019; doi: 10.1101/cshperspect.a033076 originally published online September 10, 2018

Subject Collection [Next-Generation Sequencing in Medicine](#)

Next-Generation Sequencing Technologies

W. Richard McCombie, John D. McPherson and Elaine R. Mardis

The Impact of Next-Generation Sequencing on Cancer Genomics: From Discovery to Clinic

Elaine R. Mardis

Next-Generation Sequencing in Autism Spectrum Disorder

Stephan J. Sanders

Sequencing in High Definition Drives a Changing Worldview of the Epigenome

Emily Hodges

Whole-Genome Sequencing in Cancer

Eric Y. Zhao, Martin Jones and Steven J.M. Jones

High Throughput Sequencing and Assessing Disease Risk

Shannon M. Rego and Michael P. Snyder

Clinical Versus Research Sequencing

Yuriy Shevchenko and Sherri Bale

Single-Cell Applications of Next-Generation Sequencing

Naishitha Anaparthi, Yu-Jui Ho, Luciano Martelotto, et al.

Future Promises and Concerns of Ubiquitous Next-Generation Sequencing

W. Richard McCombie and John D. McPherson

Next-Generation Sequencing Strategies

Shawn E. Levy and Braden E. Boone

Characterizing the Cancer Genome in Blood

Sarah-Jane Dawson

The Role of Next-Generation Sequencing in Pharmacogenetics and Pharmacogenomics

Ute I. Schwarz, Markus Gulilat and Richard B. Kim

The Use of Next-Generation Sequencing for Research and Diagnostics for Intellectual Disability

Ricardo Harripaul, Abdul Noor, Muhammad Ayub, et al.

Next-Generation Sequencing and the Return of Results

Bartha Maria Knoppers, Minh Thu Nguyen, Karine Sénécal, et al.

For additional articles in this collection, see <http://perspectivesinmedicine.cshlp.org/cgi/collection/>