Single-base resolution analysis of DNA epigenome via high-throughput sequencing

Jinying Peng^{1*}, Bo Xia¹ & Chengqi Yi^{1,2*}

¹State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China;

²Synthetic and Functional Biomolecules Center, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

Received January 16, 2015; accepted February 27, 2015; published online January 27, 2016

Epigenetic changes caused by DNA methylation and histone modifications play important roles in the regulation of various cellular processes and development. Recent discoveries of 5-methylcytosine (5mC) oxidation derivatives including 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC) in mammalian genome further expand our understanding of the epigenetic regulation. Analysis of DNA modification patterns relies increasingly on sequencing-based profiling methods. A number of different approaches have been established to map the DNA epigenomes with single-base resolution, as represented by the bisulfite-based methods, such as classical bisulfite sequencing (BS-seq), TAB-seq (TET-assisted bisulfite sequencing), oxBS-seq (oxidative bisulfite sequencing) and etc. These methods have been used to generate base-resolution maps of 5mC and its oxidation derivatives in genomic samples. The focus of this review will be to discuss the chemical methodologies that have been developed to detect the cytosine derivatives in the genomic DNA.

epigenetics, DNA methylation, bisulfite sequencing (BS-Seq), TAB-seq, oxBS-seq, fC-CET

Citation: Peng, J., Xia, B., and Yi, C. (2016). Single-base resolution analysis of DNA epigenome via high-throughput sequencing. *Sci China Life Sci* 59, 219–226. doi: 10.1007/s11427-016-5013-x

INTRODUCTION

The term “Epigenetics” was coined by the British developmental biologist Conrad Waddington in 1939 (Waddington, 1939), and can be defined as heritable changes in traits that occur without alteration of the underlying primary DNA sequence. “Epi” means “upon” or “over”, and the “genetics” indicates its heritability and also the involvement of genes. Epigenetics reflects “the need to study events ‘over’ or beyond the gene” (Jablonka and Lamb, 2002). Genome-wide studies on epigenetic changes are now termed as “epigenomics”. At the molecular level, epigenetic phenomena are regulated by so-called epigenetic marks that include covalent modifications on DNA bases and histones.

Epigenetic changes represented by DNA and histone modifications play essential roles in the regulation of chromatin structure and genome stability, genomic imprinting, development, carcinogenesis and aging etc (Bender, 2004; Bird, 2002; Zhu, 2009). The methylated nucleotide 5-methyldeoxycytidine (5mC) was identified long before DNA was recognized as genetic material (Johnson and Coghill, 1925). Up to 60%–80% of the 28 million CpG dinucleotides in the human genome and approximately 24% of cytosine in higher plants are methylated (Cokus et al., 2008); however, 5mC is undetectable in yeasts or nematodes (Doerfler, 1983; Lister et al., 2009b; Ziller et al., 2013). DNA hypermethylation in promoter regions usually represses gene transcription; the methylation can either directly interfere with the binding of transcriptional factors, or function indirectly by promoting the formation of repres-

*Corresponding author (email: jypengpku@pku.edu.cn; chengqi.yi@pku.edu.cn)

sive chromatin via methyl DNA-binding proteins (Bird, 2002).

DNA methylation is established and maintained by DNA methyltransferases (Wu and Zhang, 2010). DNA methyltransferase catalyzes the transfer of the methyl group of S-adenosyl methionine (SAM) to the 5'-position of cytosines through a reaction involving base flipping (Roberts and Cheng, 1998). In mammals, DNMT3a/b (DNA methyltransferase 3) and DNMT1 are responsible for *de novo* synthesis and maintenance of DNA methylome, respectively (Goll and Bestor, 2005) (Figure 1). Correct establishment of DNA methylation pattern is critical for normal development in animals and plants. For instance, null mutations in mouse *DNMT1* or *DNMT3a/b* result in embryonic lethality; and both *met1* and *drm1 drm2 cmt3* triple mutations in *Arabidopsis* result in developmental abnormalities (Chan et al., 2005; Goll and Bestor, 2005).

In 2009, 5-hydroxymethylcytosine (5hmC) was shown to be present in relatively high abundance in certain mammalian cells and tissues (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009). 5-formylcytosine (5fC) and 5-carboxycytosine (5caC) was soon later detected in mouse ES cells and brain cortex by thin layer chromatography (TLC) and liquid chromatography tandem mass spectrometry (LC-MS/MS) (He et al., 2011; Ito et al., 2011; Pfaffender et al., 2011). These cytosine derivatives are produced

Figure 1 (color online) Cytosine modification dynamics and demethylation pathway in mammalian genome. DNA methylation is established and maintained by DNA methyltransferases, while demethylation can be implemented through either passive or active pathways. TET family proteins orderly oxidize 5mC to 5hmC, 5fC, and 5caC. The oxidation products 5fC and 5caC can be recognized and excised by mammalian thymine DNA glycosylase (TDG), and then eventually transformed to normal cytosine through base excision repair (BER) pathway, thus completing an active demethylation process.

from a stepwise oxidation of 5mC by the TET (ten-eleven translocation) family dioxygenase (Ito et al., 2011; Tahiliani et al., 2009) (Figure 1). 5mC hypermodification on promoter region is generally viewed as a “silencing” epigenetic mark of gene expression (Jones, 2012). 5hmC, 5fC and 5caC have been demonstrated to be intermediates in mammalian active demethylation pathway (He et al., 2011; Ito et al., 2011; Tahiliani et al., 2009), and may have more regulatory roles mediated by the interacting proteins (Iurlaro et al., 2013; Kellinger et al., 2012; Spruijt et al., 2013; Yildirim et al., 2011). Mounting evidence has demonstrated that the enzyme-mediated chemical modification of cytosine bases has emerged as an important area of epigenetic regulation and drawn broad attention from various research communities (Song et al., 2012). To better understand the biological roles of these newly-discovered cytosine derivatives, it is important to know the genome distribution patterns, better at single-base resolution.

Thanks to recent advances in high-throughput sequencing technology, which is also termed as next generation sequencing (NGS) and provides massive sequencing data on genome information, a number of strategies for high-throughput detection of cytosine modifications at whole-genome scale have been recently described and have led to important discoveries in many systems, leading to a new paradigm of epigenome studies. As a powerful technology of investigating epigenetic alterations at genome-wide level, high-throughput sequencing could produce thousands or millions of sequences concurrently in one process. Antibody mediated DNA immunoprecipitation sequencing (DIP-seq) is widely used to profile the epigenetic modifications (Pastor et al., 2011; Shen et al., 2013; Wu et al., 2011). Meanwhile, several chemical or enzymatic assisted profiling methods were developed to analyze the 5mC (Zhang et al., 2013), 5hmC (Pastor et al., 2011; Robertson et al., 2011; Song et al., 2011; Sun et al., 2013) and 5fC (Raiber et al., 2012; Song et al., 2013). The affinity-based profiling methods have advanced our understanding of the distribution pattern across the genome, but the resolution (typically several hundred base-pairs) limits its capacity of precisely mapping the modified sequences with single-base resolution. The current methods with single-base resolution mainly relies on the bisulfite sequencing (BS-seq) to distinguish 5mC from C (Lister and Ecker, 2009; Lister et al., 2008). Modified BS-seq include the TAB-seq (TET-assisted bisulfite sequencing) (Yu et al., 2012) and oxBS-seq (oxidative bisulfite sequencing) (Booth et al., 2012) to assess 5hmC, fCAB-seq (5fC-assisted bisulfite sequencing) (Song et al., 2013) and redBS-seq (reduced bisulfite sequencing) (Booth et al., 2014a) to quantitatively sequence 5fC, and CAB-seq (chemical modification-assisted bisulfite sequencing) (Lu et al., 2013) or MAB-seq (M.SssI Methylase-assisted bisulfite sequencing) (Guo et al., 2014; Wu et al., 2014) to detect 5caC or 5fC/5caC simultaneously. Bisulfite-free method fC-CET (cyclization-enabled C-to-T

transition of 5fC) is also discussed. The focus of this review is to discuss these chemical-assisted methodologies for the detection of the cytosine derivatives with single-base resolution at the genome-wide scale.

SINGLE-BASE RESOLUTION SEQUENCING METHODS FOR CYTOSINE DERIVATIVES

Bisulfite sequencing

The bisulfite mediated deamination of cytosine (C) to uridine (U) was first discovered independently by two groups in 1970 (Hayatsu et al., 1970; Shapiro and Weisgras, 1970). High concentrations of bisulfite anion adds to 6'-carbon position of the protonated cytosine, generating a cytosin-sulphonate, which loses aromaticity of the base and undergoes hydrolysis and deamination (Booth et al., 2014b). The resulting bisulfite adduct, uracilsulphonate, rearomatizes in alkaline conditions to form uracil (Figure 2). The methyl group on 5'-carbon of 5mC makes the base stay intact in the bisulfite treatment. In the subsequent PCR amplification and sequencing, all of the cytosines in the DNA convert to thymines and read as T, while 5mC still sequenced as C (Figure 3). Bisulfite conversion of genomic DNA combined with next-generation sequencing (BS-seq), which has been regarded as the golden standard for 5mC detection, is widely used to study genome-wide DNA methylation dynamics with single-base resolution (Lister and Ecker, 2009; Lister et al., 2008; Lister et al., 2009a; Ziller et al., 2013).

When analyzing bulk genomic DNA samples, multiple copies of the same genetic sequence are usually involved due to isolation of DNA from massive cells, thus each cyto-

sine site may contain varied percentages of epigenetic modification. BS-seq also offers the quantitative information of 5mC. With NGS platform which sequencing each genome region with multiple reads, each base can be sequenced several times. By quantifying the percentage of C readout and T readout at anticipated 5mC sites, the quantitative information can be calculated as $n(C)/n(C+T)$ in a genome wide scale.

Oxidative bisulfite sequencing

The presence of 5mC oxidation derivatives in mammalian genome makes bisulfite sequencing more complicated, since 5hmC reads as C, same with the 5mC, while both 5fC and 5caC read as T in BS-seq. To address this problem, Balasubramanian and colleagues developed the oxidative bisulfite sequencing (oxBS-seq) for 5mC and 5hmC base-resolution sequencing in mammalian genomes (Booth et al., 2012). They found that K₂ReO₄ (potassium perruthenate) specifically oxidizes 5hmC to 5fC, which would undergo deamination under repeated bisulfite treatments (Figure 4). Therefore, in a DNA sample sequenced with oxBS-seq, 5hmC would read as thymine, whereas 5mC still reads as cytosine (Table 1, oxBS-seq). Comparing the canonical BS-seq data and oxBS-seq data would generate the 5hmC map quantitatively with single-base resolution.

Booth and colleagues generated the first single-base resolution map of 5mC and 5hmC in the CpG rich regions of mouse embryonic stem cells (mESCs) with reduced representation oxBS-seq (Booth et al., 2012; Booth et al., 2014b). Potential limitations of oxBS-seq include severe damage and degradation of genomic DNA caused by oxidation con-

Figure 2 (color online) Mechanism of bisulfite sequencing. Deamination of C to U is achieved through the addition of sodium bisulfite to the C5-6 double bond, followed by deamination of the cytosinsulphonate. The uracil-sulfite adduct can be converted to be uracil under alkaline conditions and further reads as thymine during PCR amplification.

Table 1 Decoding the epigenomes with single-base resolution sequencing methods^{a)}

Cytosine& derivatives	Standard sequencing	BS-seq	oxBS-Seq	TAB-seq	fCAB-seq	redBS-seq	CAB-seq	MAB-seq	fC-CET
C	C	T	T	T	T	T	T	C*	C
5mC	C	C	C	T	C	C	C	C	C
5hmC	C	C	T	C	C	C	C	C	C
5fC	C	T	T	T	C	C	T	T	T
5caC	C	T	T	T	T	T	C	T	C

a) Sequencing output of DNA epigenome sequencing methods. C* indicated the unmodified cytosine in CpG dinucleotides.

Figure 3 Schematic overview of bisulfite sequencing procedure. Bisulfite conversion of genomic DNA and subsequent PCR amplification produce two PCR products and up to four potentially different DNA fragments for any given locus. Methylated cytosine residues are intact during bisulfite treatment.

Figure 4 Mechanism of oxBS-Seq. Upon treatment with KRuO_4 , 5hmC is oxidized to 5fC, which is then deaminated by sodium bisulfite.

ditions, as well as by the repeated bisulfite treatments in order to fully deaminate 5fC. With further optimizing the oxidation step to reduce DNA degradation, oxBS-seq could be applied to as low as 100 ng input gDNA (Booth et al., 2013).

TET-assisted bisulfite sequencing

Another method to detect 5hmC, TET-assisted bisulfite sequencing (TAB-Seq) was independently developed by He and collaborators (Yu et al., 2012). TAB-Seq uses the β -glucosyltransferase (β GT) to modify 5hmC, and a recombinant mouse TET1 to convert the 5mC to 5caC. While the previously glucosylated 5hmC stays intact during oxidation by the TET1, subsequent bisulfite treatment causes decarboxylation and deamination of 5caC to form uracil, hence “wiping out” signals of 5mC (Figure 5). Therefore, TAB-Seq gives a direct read-out of 5hmC (Table 1, TAB-seq). To obtain base-resolution information of 5mC, the results of TAB-Seq can be compared with those of traditional bisulfite sequencing, which reveals the sum of 5mC+5hmC (Yu et al., 2012). A subtraction yields the base-resolution map of 5mC (Table 1, TAB-seq).

TAB-Seq has been used to generate whole-genome 5hmC maps of mESC (Yu et al., 2012). By this means, sites that contained significant levels of 5hmC throughout the entire genome and new distribution features have been detected. 5hmC is enriched at distal functional regulatory elements near but not on transcription factor-binding sites. Also, distinct from the symmetric pattern of 5mC on the CpG dinucleotide, most of the 5hmC sites show strand asymmetry (Song et al., 2012; Yu et al., 2012). A minor limitation is that β GT enzyme was shown to exhibit inefficiencies when glucosylating 5hmCpGp5hmCpGs (Yu et al., 2012), which may cause difficulties for sites with multiple 5hmCs in close proximity. However, such 5hmC pattern is rare in genome, thus this method has been widely used in various samples (Lister et al., 2013; Wang et al., 2014).

Figure 5 Mechanism and sequencing output of TAB-Seq. 5hmC is protected from further bisulfite conversion by glucosylation of β GT. Upon treatment with TET1, 5mC is oxidized to 5caC, which is then deaminated by sodium bisulfite. In TAB-Seq, 5hmC is the only base to read as a C.

Chemical-assisted bisulfite sequencing and reduced bisulfite sequencing for 5fC

Recently, two kinds of chemical methods, fCAB-Seq (chemical-assisted bisulfite sequencing for 5fC) (Song et al., 2013) and redBS-Seq (reduced bisulfite sequencing) (Booth et al., 2014a) have been developed to detect 5fC at single base resolution. The basic principle of these methods is to block the conversion of 5fC to U during bisulfite treatment.

fCAB-Seq is developed based on the findings that hydroxylamine-protected 5fC can be resistant to the bisulfite treatment (Song et al., 2013). He and colleagues used the O-ethylhydroxylamine (EtONH₂) to label 5fC ahead of the bisulfite treatment. The labeled 5fC product stays intact in the bisulfite treatment and reads as C in BS-seq (Figure 6). Therefore, by subtracting the data obtained from BS-Seq and fCAB-Seq, 5fC can be identified as the difference. In the former case, C, 5fC and 5caC read as T, while in the latter case, only C and 5caC reads as a T while 5fC reads as a C (Table 1, fCAB-seq). So far, fCAB-Seq has been applied successfully to detect 5fC at base-resolution within several targeted regions in mammalian genomic DNAs (Song et al., 2013; Wang et al., 2014).

In redBS-Seq, NaBH₄ is used to reduce 5fC to 5hmC in genomic samples (Booth et al., 2014a). 5fC deformylates then deaminates to U in the bisulfite treatment, while 5hmC converts to cytosine-5-methylsulfonate (CMS) and read as C. Upon reduction with NaBH₄, 5fC is converted to 5hmC in DNA, leading to all 5fC being read as C in BS-seq (Figure 6). By comparing the output of redBS-Seq where 5fC reads as C with the BS-Seq data where 5fC reads as T, would enable the detection of 5fC quantitatively with single-base resolution (Table 1, redBS-seq). This method was also employed in parallel with oxBS-Seq to generate a first map of 5mC, 5hmC and 5fCs simultaneously. It should be mentioned that He and colleagues compared the protection

Figure 6 (color online) Mechanism of fCAB-Seq and redBS-Seq. In fCAB-Seq, 5fC is protected by EtONH₂ and can be resistant to bisulfite treatment, thus reads as C in the subsequent BS-seq. In redBS-seq, 5fC is reduced to 5hmC by NaBH₄ treatment, and reads as C in the BS-seq. By comparing the output of redBS-Seq where 5fC reads as C with the BS-Seq data where 5fC reads as T, would enable the detection of 5fC quantitatively with single-base resolution.

efficiency of EtONH₂-labeled 5fC (fCAB-seq) and NaBH₄-reduced 5fC (redBS-seq) under bisulfite treatment, and found that the former method works a bit better (Song et al., 2013).

Chemical modification-assisted bisulfite sequencing

5caC was found at levels ten times lower than 5fC in genomic DNA (Ito et al., 2011). Under typical bisulfite conditions, both 5fC and 5caC are deaminated and read as T, making it difficult to distinguish them from unmodified cytosine. One method has been published to detect 5caC in DNA, termed CAB-Seq (chemical modification-assisted bisulfite sequencing) (Lu et al., 2013). In CAB-Seq, 5caC is converted to an amide by reaction with EDC (1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide hydrochloride) and a primary amine. The amide-derivatives of 5caC inhibit its conversion to U during bisulfite treatment, thus read as C (Figure 7). This method could potentially be used to detect 5caC by subtracting the BS-Seq data (where C, 5fC and 5caC read as T) from this CAB-Seq data (where 5caC reads as a C) (Table 1, CAB-seq). Currently, the CAB-Seq method has been used on synthetic DNA; it would be of great interest if CAB-Seq or other methods could be used for genome-wide detection of 5caC.

M.SssI Methylase-assisted bisulfite sequencing

The above chemical modification-assisted bisulfite (BS)-seq methods all require subtraction of BS-seq signals to indirectly map 5mC, 5hmC, 5fC or 5caC. A recently reported method named MAB-seq (M.SssI methylase-assisted bisulfite sequencing) provided a direct base-resolution mapping method for 5fC/5caC simultaneously (Guo et al., 2014; Wu et al., 2014). In the first step of MAB-seq, genomic DNA is treated with the bacterial DNA CpG methyltransferase M.SssI, an enzyme that was isolated from *Spiroplasma* sp. strain MQ1 and is known to methylate cytosines within the context of CpG dinucleotides. In the second step, when the M.SssI-treated DNA is subjected to bisulfite treatment, only 5fC and 5caC and non-CpG cytosines would be deaminated; originally unmodified C within CpGs is protected as 5mC (Figure 8). Subsequent sequencing reveals 5fC and 5caC as T, while C/5mC/5hmC would be read as C (Table 1, MAB-seq). Considering that most cytosine modifications present on the CpG samples, the MAB-seq is practical in the samples regarding embryonic development to directly determine the location and abundance of 5fC and

Figure 7 (color online) Mechanism of CAB-Seq. In CAB-Seq, 5caC is converted to an amide through reacting with EDC and a primary amine, thus avoiding the conversion to U during bisulfite treatment, and are therefore read as a C.

Figure 8 Schematic overview of M.SssI methylase-assisted bisulfite sequencing procedure. Genomic DNA is treated with methyltransferase M.SssI, and then subjected to bisulfite treatment. Only 5fC, 5caC and non-CpG cytosines is deaminated; the original C within CpGs is protected as 5mC. Subsequent sequencing reveal 5fC and 5caC as T, whereas C/5mC/5hmC would be read as C.

5caC as active demethylation intermediates (Guo et al., 2014; Wu et al., 2014). However, when it comes to the samples where non-CpG modification is in higher levels like neuronal samples (Lister et al., 2013; Wang et al., 2014), MAB-seq may not be useful as well.

Cyclization enabled C-to-T transition of 5fC

Recently, we have developed a “bisulfite-free” method, fC-CET, to sequence genome-wide 5fC at single-base resolution (Xia et al., 2015). Based on a cyclization labeling of Friedländer reaction, which involves both formyl group and amino group, we identified several chemicals which labeled 5fC with superior selectivity. Though much of the chemicals do not affect its base-pairing with guanine (G), one chemical, 1,3-indandione, is unique as its labeling product is read as thymine (T) instead of cytosine (C) during PCR amplification (Figure 9A). The “C-to-T” transition property of this chemical labeling thus directly detects 5fC sites (Table 1).

Considering the limited abundance of 5fC in genome, we further modified the chemical to allow enrichment of 5fC-modified genomic fragments (Figure 9B). We synthesized an azido derivative of 1,3-indandione (AI), which has an azide “tail” but does not affect its reactivity with 5fC. A biotin can be labeled to azide group through well-established “click chemistry” to facilitate pulling-down 5fC-modified fragments. Combining the “C-to-T” transition property and the affinity-enriched 5fC region, we figured out the comprehensive map of 5fC in mouse embryonic stem cells. The differential modified 5fC sites indicates that 5fC-modified marked genomic regions are more active than 5hmC-marked ones (Xia et al., 2015). Meanwhile, it is worth noting that fC-CET based chemical treatment causes little DNA degradation and detects modifications directly

Figure 9 Mechanism of fC-CET. A, Chemical (1,3-indandione) treatment of 5fC and PCR amplification output. B, AI mediated pulldown of 5fC modified DNA fragments. AI, azido derivative of 1,3-indandione.

(Xia et al., 2015), suggesting potential applications of analyzing rare samples even at single-cell level.

FUTURE PROSPECTS

So far, the bisulfite sequencing based chemical methods have been developed to decode 5mC, 5hmC, 5fC, and 5caC at single base resolution in bulk genomic DNA, which greatly advanced our understanding of the epigenetic regulation in development, diseases and even brain function. Most of the methods have been successfully used to analyze the genome wide distribution pattern of modified cytosines. However, analyzing rare bases like 5hmC requires significant sequencing depth (Yu et al., 2012), let alone 5fC or 5caC which only occupies hundredth or thousandth of the 5hmC level in genome (Ito et al., 2011), respectively. Meanwhile, because of the extremely low abundance of 5caC, genome wide sequencing of 5caC with single-base resolution is still challenging. Future methods should address the problem to economically analyzing the epigenome with single-base resolution.

Since the bisulfite treatment causes significant degradation of DNA samples, it is biased to analyze the methylome in single-cell samples like zygote or early embryos (Guo et al., 2013; Smallwood et al., 2014). Moreover, analyzing the epigenome of 5mC oxidation derivatives at single-cell level is desired but not approachable currently because of the severe degradation of DNA in bisulfite treatment. Developing some alternative methods even bisulfite-free methods for single-cell epigenome sequencing remains a big challenge for both biologists and chemists, as well as engineers. Recent development of single-molecule, re-

al-time sequencing (SMRT-seq) platform (Flusberg et al., 2010) and nanopore-based sequencing (Wescoe et al., 2014) provide possible solutions for these questions, but much remains to be refined like input quantity, error rate, sequencing depth etc. Novel chemical methods for epigenome research are still appealing.

Compliance and ethics The author(s) declare that they have no conflict of interest.

Acknowledgements This work was supported by the National Basic Research Foundation of China (2014CB964900 to Yi Chengqi) and the National Natural Science Foundation of China (31270838, 21472009 to Yi Chengqi).

- Bender, J. (2004). DNA methylation and epigenetics. *Annu Rev Plant Biol* 55, 41–68.
- Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev* 16, 6–21.
- Booth, M.J., Branco, M.R., Ficz, G., Oxley, D., Krueger, F., Reik, W., and Balasubramanian, S. (2012). Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* 336, 934–937.
- Booth, M.J., Marsico, G., Bachman, M., Beraldi, D., and Balasubramanian, S. (2014). Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat Chem* 6, 435–440.
- Booth, M.J., Ost, T.W., Beraldi, D., Bell, N.M., Branco, M.R., Reik, W., and Balasubramanian, S. (2013). Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc* 8, 1841–1851.
- Booth, M.J., Raiber, E.A., and Balasubramanian, S. (2015). Chemical methods for decoding cytosine modifications in DNA. *Chem Rev* 115, 2240–2254.
- Chan, S.W., Henderson, I.R., and Jacobsen, S.E. (2005). Gardening the genome: DNA methylation in *Arabidopsis thaliana*. *Nat Rev Genet* 6, 351–360.
- Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulfite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452, 215–219.
- Doerfler, W. (1983). DNA methylation and gene activity. *Annu Rev Biochem* 52, 93–124.
- Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korch, J., and Turner, S.W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* 7, 461–465.
- Goll, M.G., and Bestor, T.H. (2005). Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* 74, 481–514.
- Guo, F., Li, X., Liang, D., Li, T., Zhu, P., Guo, H., Wu, X., Wen, L., Gu, T.P., Hu, B., Walsh, C.P., Li, J., Tang, F., and Xu, G.L. (2014). Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell* 15, 447–458.
- Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* 23, 2126–2135.
- Hayatsu, H., Wataya, Y., Kai, K., and Iida, S. (1970). Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry* 9, 2858–2865.
- He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.X., Zhang, K., He, C., and Xu, G.L. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* 333, 1303–1307.
- Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* 333, 1300–1303.
- Iurlaro, M., Ficz, G., Oxley, D., Raiber, E.A., Bachman, M., Booth, M.J., Andrews, S., Balasubramanian, S., and Reik, W. (2013). A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol* 14, R119.
- Jablonska, E., and Lamb, M.J. (2002). The changing concept of epigenetics. *Ann N Y Acad Sci* 981, 82–96.
- Johnson, T.B., and Coghill, R.D. (1925). Research on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the *Tubercle bacillus*. *J Am Chem Soc* 47, 2844.
- Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 13, 484–492.
- Kellinger, M.W., Song, C.X., Chong, J., Lu, X.Y., He, C., and Wang, D. (2012). 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nat Struct Mol Biol* 19, 831–833.
- Kriaucionis, S., and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* 324, 929–930.
- Lister, R., and Ecker, J.R. (2009). Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res* 19, 959–966.
- Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J.C., Rao, A., Esteller, M., He, C., Haghghi, F.G., Sejnowski, T.J., Behrens, M.M., and Ecker, J.R. (2013). Global epigenomic reconfiguration during mammalian brain development. *Science* 341, 1237905.
- Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133, 523–536.
- Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., and Ecker, J.R. (2009a). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315–322.
- Lu, X., Song, C.X., Szulwach, K., Wang, Z., Weidenbacher, P., Jin, P., and He, C. (2013). Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. *J Am Chem Soc* 135, 9315–9317.
- Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P., Tahiliani, M., Daley, G.Q., Liu, X.S., Ecker, J.R., Milos, P.M., Agarwal, S., and Rao, A. (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* 473, 394–397.
- Pfaffeneder, T., Hackner, B., Truss, M., Munzel, M., Muller, M., Deiml, C.A., Hagemeyer, C., and Carell, T. (2011). The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew Chem Int Ed Engl* 50, 7008–7012.
- Raiber, E.A., Beraldi, D., Ficz, G., Burgess, H.E., Branco, M.R., Murat, P., Oxley, D., Booth, M.J., Reik, W., and Balasubramanian, S. (2012). Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol* 13, R69.
- Roberts, R.J., and Cheng, X. (1998). Base flipping. *Annu Rev Biochem* 67, 181–198.
- Robertson, A.B., Dahl, J.A., Vagbo, C.B., Tripathi, P., Krokan, H.E., and Klungland, A. (2011). A novel method for the efficient and selective identification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic Acids Res* 39, e55.
- Shapiro, R., and Weisgras, J.M. (1970). Bisulfite-catalyzed transamination of cytosine and cytidine. *Biochem Biophys Res Commun* 40, 839–843.
- Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, A.C., Fung, H.L., Zhang, K., and Zhang, Y. (2013). Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* 153, 692–706.

- Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 11, 817–820.
- Song, C.X., Szulwach, K.E., Dai, Q., Fu, Y., Mao, S.Q., Lin, L., Street, C., Li, Y., Poidevin, M., Wu, H., Gao, J., Liu, P., Li, L., Xu, G.L., Jin, P., and He, C. (2013). Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* 153, 678–691.
- Song, C.X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.H., Zhang, W., Jian, X., Wang, J., Zhang, L., Looney, T.J., Zhang, B., Godley, L.A., Hicks, L.M., Lahn, B.T., Jin, P., and He, C. (2011). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* 29, 68–72.
- Song, C.X., Yi, C., and He, C. (2012). Mapping recently identified nucleotide variants in the genome and transcriptome. *Nat Biotechnol* 30, 1107–1116.
- Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W., Bauer, C., Munzel, M., Wagner, M., Muller, M., Khan, F., Eberl, H.C., Mensinga, A., Brinkman, A.B., Lephikov, K., Müller, U., Walter, J., Boelens, R., van Ingen, H., Leonhardt, H., Carell, T., and Vermeulen, M. (2013). Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* 152, 1146–1159.
- Sun, Z., Terragni, J., Borgaro, J.G., Liu, Y., Yu, L., Guan, S., Wang, H., Sun, D., Cheng, X., Zhu, Z., Pradhan, S., and Zheng, Y. (2013). High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells. *Cell Rep* 3, 567–576.
- Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., and Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 324, 930–935.
- Waddington, C.H. (1939). Preliminary Notes on the development of the wings in normal and mutant strains of *Drosophila*. *Proc Natl Acad Sci USA* 25, 299–307.
- Wang, L., Zhang, J., Duan, J., Gao, X., Zhu, W., Lu, X., Yang, L., Li, G., Ci, W., Li, W., Zhou, Q., Aluru, N., Tang, F., He, C., Huang, X., and Liu, J. (2014). Programming and inheritance of parental DNA methylomes in mammals. *Cell* 157, 979–991.
- Wescoe, Z.L., Schreiber, J., and Akeson, M. (2014). Nanopores discriminate among five C5-cytosine variants in DNA. *J Am Chem Soc* 136, 16582–16587.
- Wu, H., D'Alessio, A.C., Ito, S., Wang, Z., Cui, K., Zhao, K., Sun, Y.E., and Zhang, Y. (2011). Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev* 25, 679–684.
- Wu, H., Wu, X., Shen, L., and Zhang, Y. (2014). Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat Biotechnol* 32, 1231–1240.
- Wu, S.C., and Zhang, Y. (2010). Active DNA demethylation: many roads lead to Rome. *Nat Rev Mol Cell Biol* 11, 607–620.
- Xia, B., Han, D., Lu, X., Sun, Z., Zhou, A., Yin, Q., Zeng, H., Liu, M., Jiang, X., Xie, W., He, C., and Yi, C. (2015). Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat Methods* 12, 1047–1050.
- Yildirim, O., Li, R., Hung, J.H., Chen, P.B., Dong, X., Ee, L.S., Weng, Z., Rando, O.J., and Fazzio, T.G. (2011). Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* 147, 1498–1510.
- Yu, M., Hon, G.C., Szulwach, K.E., Song, C.X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., Min, J.H., Jin, P., Ren, B., and He, C. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 149, 1368–1380.
- Zhang, L., Szulwach, K.E., Hon, G.C., Song, C.X., Park, B., Yu, M., Lu, X., Dai, Q., Wang, X., Street, C.R., Tan, H., Min, J.H., Ren, B., Jin, P., and He, C. (2013). Tet-mediated covalent labelling of 5-methylcytosine for its genome-wide detection and sequencing. *Nat Commun* 4, 1517.
- Zhu, J.K. (2009). Active DNA demethylation mediated by DNA glycosylases. *Annu Rev Genet* 43, 143–166.
- Ziller, M.J., Gu, H., Muller, F., Donaghey, J., Tsai, L.T., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., Gnirke, A., and Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature* 500, 477–481.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.