Mapping and elucidating the function of modified bases in DNA

Authors

Eun-Ang Raiber[1], Robyn Hardisty[1], Pieter van Delft[1], Shankar Balasubramanian[1,2,3]

[1] Department of Chemistry, University of Cambridge, Lensfield Road Cambridge CB2 1EW, UK.

[2]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

[3]School of Clinical Medicine, University of Cambridge, Hills Road, Cambridge CB2 0SP, UK

Correspondence to: sb10031@cam.ac.uk

Chemically modified bases naturally exist in genomic DNA. Research into these bases has been invigorated by the discovery of several modified bases in the mammalian genome, in particular the oxidised derivatives of 5-methylcytosine, such as 5-hydroxymethylcytosine and 5-formylcytosine, as well as the enzymes that form and process them, such as DNA methyltransferases (DNMTs) and the ten-eleven translocation (TET) enzymes. In this Review we provide an overview of natural, modified bases that have been reported in DNA, our current knowledge of their roles, and the techniques that have enabled us to probe their functions. Analytical methods have been invaluable in helping advance this field. For example, chemical and enzymatic methods have provided the means to detect and decode modified bases, giving rise to an expanding array of sequencing approaches. Advanced liquid chromatography and tandem mass spectrometry have provided the means to detect and quantify modified bases with very high sensitivity, increasing the prospects for the discovery of unknown modifications. It is already evident that natural, modified DNA bases and their associated enzymology are of fundamental importance to normal biology and to disease. The next decade promises to yield more insights, discoveries and impact from this burgeoning field of research.

DNA is a remarkable, functional molecule that has evolved to encode and transmit information through multiple dimensions. The primary code stored within DNA comprises the four nucleobases adenine (A), guanine (G), cytosine (C) and thymine (T) and it's the linear sequence of these four canonical bases that constitutes genetic information. Nature maintains, reads and transfers the genetic code through cognate base-pair recognition, a principle that was originally proposed by Watson and Crick[1] and involves A-T and G-C interactions (Figure 1). These interactions are also largely responsible for the double helical structure of double-stranded DNA. In fact, the primary DNA code can also be read from the major groove and minor groove of the DNA double helix. The major groove presents a distinct pattern of hydrogen bonds, giving rise to a 'major groove code', which reveals the specific sequence of base pairs and can be recognized and interpreted by proteins such as transcription factors **[G]**.[2] This enables the sequence-based activation of DNA functions, such as the transcription of a protein-coding gene to generate mRNA that is subsequently translated into protein. There is also a 'minor groove code', comprising base-pair specific hydrogen bonds in this groove that is narrower that then major groove. Although the major groove is predominantly read by proteins, it has been possible to read the primary sequence of DNA via synthetic polyamide molecules that bind to the minor groove.[3]

5-methylcytosine (5mC) was probably the first variant of a canonical nucleobase to be discovered.[4] Since the discovery of 5mC in 1898 more than 17 modified DNA bases have been reported in the genomes of prokaryotes and eukaryotes, including mammals[5] (Figure 2). A feature of these modifications is that they tend to not interfere with Watson–Crick pairing, but to introduce chemical functionality into the major groove of the double helix. The introduction of added functionality in the major groove can certainly perturb or block protein recognition. This is applied in bacterial warfare whereby methylation of specific sequence sites in the host genome, called restriction methylation [G], provides protection against self-cleavage by special nucleases. However, it is now recognized that site-specific

modifications introduced into the major groove of DNA can recruit specific proteins (known as 'readers') that influence the function of the genome at that locus. Given that base modifications can be dynamically incorporated or removed by specific enzymes, it is possible that a sophisticated and reprogrammable major groove code may contribute to the regulation of the genome and all of its functions. Thus, the pattern of DNA base modifications may constitute an important layer of reprogrammable information in DNA, which is of particular interest in the genomes of higher organisms. In this Review we will discuss a number of classes of naturally occurring DNA modifications and explain how chemistry can be deployed to detect, map and decode these DNA base modifications in the genomic DNA of eukaryotes to elucidate their function.

Known modifications of DNA

The following sections gives an overview on eukaryotic DNA modifications, subdivided by their canonical base precursor.

Cytosine modifications.

The most studied DNA modification is cytosine methylation. It was discovered in 1898[4] and its existence was confirmed by Johnson and Coghill in 1925, who identified it as a hydrolysis product of tuberculinic acid.[6] In eukaryotes, methylation at the C5 position of cytosine is catalyzed by a class of enzymes known as DNA methyltransferases (DNMTs), which use *S*-adenosylmethionine (SAM) as the methyl donor. DNMT1 is mainly responsible for the maintenance of methylation marks during DNA replication, whereas DNMT3a and DNMT3b are involved in the methylation of new sites.[7,8,9] Recently, a new de novo DNA methyltransferase, DNMT3c, was discovered in rodent genomes, and it was shown to methylate the promoters **[G]** of evolutionarily young retrotransposons in the male germ line.[10]

In mammals, DNA methylation is involved in the maintenance of cellular functions and genomic stability, including processes such as X chromosome

inactivation, genomic imprinting **[G]** and transposon silencing **[G]**.[11,12,13,14] Importantly, DNA methylation, does not function in isolation, but works in conjunction with histone modifications **[G]** within specific chromatin contexts. Indeed, it is becoming clear that epigenetic crosstalk exists between DNA marks and key chromatin components; this crosstalk involves various chromatin-remodeling proteins **[G]**.[15]

Cytosine methylation predominantly occurs within the C-G dinucleotide (which is often referred to as cytosine-phosphate-guanine (CpG)) outside CpG islands (CGI) **[G]**.[13] Recent studies, in which defined DNA sequences were inserted into the same locus of the genome, suggest that the overall base sequence composition influences DNA methylation and histone modifications at some CpGs but not at others.[16,17] These studies support the view that the genome shapes its epigenome, although the mechanism underlying how sequence composition imparts methylation status is still not understood.

During early mammalian development, DNA methylation changes throughout the genome to direct pluripotent stem cells **[G]** to differentiate into distinct lineages that form various tissue types. Passive demethylation occurs when the DNA in newly replicated cells is not remethylated. DNMT1 is thought to exhibit a preference towards hemi-methylated sites over hemi-hydroxymethylated sites, thereby inhibiting 5mC maintenance[18]. Active DNA demethylation (that is, enzyme-mediated DNA demethylation) has also been observed, for example within the enhancer regions **[G]** of developmental genes during the phylotypic period (that is, during the development of the basic body shape) in organisms including mouse.[19] The active demethylation pathway involves the iterative oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC) by the ten-eleven translocation (TET) enzymes (namely, TET1, TET2 and TET3), which are 2-oxoglutarate (2-OG) and Fe(II)-dependent dioxygenases. Subsequently, 5fC or 5caC is removed by base excision repair **[G]** (BER), which is mediated by the thymine DNA glycosylase (TDG), and replaced by unmodified cytosine (FIG. 3).[20] The loss of function of enzymes, including TET and TDG,

compromises the differentiation of mouse embryonic stem (mES) cells and is embryonically lethal in mice, illustrating that this DNA metabolism is vital to mammalian development.[21,22]

Although global levels of 5fC in mES cells (0.02% / dG, as measured by liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS) (See Box 2)) are lower than that of 5hmC, 5fC was found to occur at comparable levels to 5-hmC at specific genomic loci.[23,24,25] Recent developments that exploit chemistries for resolving DNA modifications, together with high-throughput (sometimes called 'next generation') sequencing approaches, have allowed the genome-wide mapping of 5hmC, 5fC and 5caC (see Box1 and Table 1). 5hmC, 5fC and 5caC are enriched in poised and active enhancers in ES cells and in mouse embryonic tissues.[26,27,28] Enhancers are distal regulating elements that initiate transcription by delivering protein complexes to promoters. Poised enhancers (that is, enhancers that are inactive but ready for activation) are flanked by bivalent histone marks (that is, histone modifications that combine the activating mark histone H3 lysine 4 monomethylation (H3K4me1) and the repressive mark histone H3 lysine 27 trimethylation (H3K27me3)), whereas active enhancers are marked by histone H3 lysine 27 acetylation (H3K27ac) and H3K4me1. Although oxidized cytosine derivatives were initially regarded as intermediates on the pathway to demethylation (FIG. 3), it is emerging that they may function as epigenetic marks in their own right. Indeed, isotopic labeling experiments have shown that 5hmC and 5fC are largely stable in the genomic DNA of cultured cells and *in vivo*.[29,30] Additionally, 5hmC, 5fC and 5caC-specific binding proteins have been identified from proteomics experiments that support roles for these cytosine modifications in chromatin remodeling and transcriptional regulation.[31,32,33] Another study found that double stranded 5fC-containing DNA could covalently interact with DNMT1; the bond formed between the 6C of 5fC and a catalytically essential cysteine residue of DNMT1 *in vitro*.[34]

Since chromatin is key to all DNA-related processes (that is, replication, recombination, transcription, repair and chromosome segregation), it is essential to comprehend how DNA modifications effect the

regulation and dynamics of chromatin structure to elucidate their biological function. Several groups have analysed the effect of cytosine modifications within CpG repeats on the stability of the DNA double helix.[35,36] 5mC and 5hmC, but not 5fC and 5caC, were found to stabilize the DNA duplex. Additional biophysical experiments and X-ray crystallography on 5fC-containing oligonucleotides showed that 5fC can alter the structure of the DNA double helix through an extensive hydration network, suggesting that 5fC may be involved in chromatin remodeling by causing a change in DNA conformation.[36] In contrast to this work, a recent study[36b] reported no significant alteration of the structure of DNA containing 5fC, which may reflect a capacity for 5fC-DNA to dynamically interconvert between structures. Indeed, a recent molecular dynamics simulation study demonstrated that 5fC and 5hmC enhance the flexibility of the DNA double helix, whereas 5mC reduces DNA flexibility.[37] The increase in the flexibility of DNA caused by 5fC and 5hmC increased the mechanical stability of nucleosomes (that is, the basic packaging unit of DNA in eukaryotes), indicating that these modifications may influence the dynamics of chromatin structure *in vivo*.

The direct effects of cytosine modifications on the structure of nucleosomes have been studied using single molecule Förster resonance energy transfer (FRET) analysis. It was reported that CpG methylation resulted in DNA that was more tightly wrapped around the histone octamer **[G]**, suggesting that CpG methylation may contribute to the formation of a repressive chromatin state.[38,39,40] 5hmC-containing DNA also displayed an increased binding affinity for the histone octamer, although a weakened interaction between hydroxymethylated DNA and the H2A-H2B dimer subunit was also observed.

Finally, a nucleosome reconstitution experiment using methylated and unmethylated genomic DNA from two cell lines investigated the effects of DNA methylation on the positioning and stability of nucleosomes.[41] This study showed that the DNMT-mediated methylation of CpG dinucleotides *in vitro* changed the nucleosomal organization. Specifically, unmethylated CpG islands near transcription start sites became enriched in nucleosomes upon

their methylation, suggesting that the impact of DNA methylation on nucleosome positioning *in vitro* can reflect *in vivo* states.

Thymine modifications.

DNA generally contains T as the cognate base to A, rather than uracil (U); U is present in RNA. However, studies using ultra high performance liquid chromatography coupled to tandem mass spectrometry (MS/MS) (UHPLC-MS/MS) demonstrated that the nucleoside 2'-deoxyuridine (dU), which lacks a methyl group at the C5 position of the base, is present at low levels in DNA[42]. A large variance in reported levels of dU in DNA is likely to be an artifact caused by deamination during DNA digestion or extraction. dU may arise due to the aberrant incorporation of 2'-deoxy-uridine-5'-triphosphate (dUTP) or it can be formed by spontaneous or enzyme-mediated hydrolytic cytosine deamination, resulting in a U:G mispair.[43] Several uracil DNA glycosylase repair enzymes can remove uracil from DNA via the BER pathway. The main candidates are the uracil DNA glycosylase (UDG) enzymes, although single monofunctional uracil glycosylase 1 (SMUG1) is thought to be a 'back-up' enzyme and TDG and methyl binding domain-4 (MBD4) can also excise dU, particularly when it is present as U:G mispairs.[43]

Although spontaneous hydrolytic deamination is slow, cytosine deamination can be driven enzymatically by the cytidine deaminases, activation-induced cytidine deaminase (AID) and apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like (APOBEC). They belong to the same family of enzymes, the cytidine deaminases, which is thought to have unique roles for generating specific genetic mutations and C-to-T transitions.[44] AID-mediated deamination in the immunoglobulin genes of B cells **[G]** provides an essential mechanism for somatic hypermutation and class-switch recombination **[G]** of DNA, leading to antibody diversification as part of the immune response. There is also some speculation that U may be an intermediate in the active demethylation of 5mC. [45] Indeed, AID is involved in a TET-independent active demethylation process at promoter regions. However, although AID can deaminate 5mC directly, alternative hypotheses suggest that the AID-driven

deamination of C to U could initiate BER of 5mC containing sequences, leading to overall demethylation.

Eukaryotic DNA contains oxidised thymine derivatives, which are analogous in structure to those observed for cytosine modifications. 5-Hydroxymethyluridine (5hmU) and 5-formyluridine (5fU) occur at a level of 0.00001-0.0001%/dN in mammalian DNA.[46] Both modifications have been traditionally considered to be oxidative lesions caused by radical oxidation of thymine. Both 5hmU and 5fU can be excised by SMUG1 when base-paired with A[47], yet they are repaired by several DNA glycosylases, including TDG and MBD4 when mispaired with G; the pairing of 5hmU and 5fU with G can occur as a consequence of deamination of the corresponding C-derivative.[48]

As well as being the result of oxidative DNA damage, 5hmU may have a functional role in mammalian DNA. For example, some evidence suggests that 5hmU is an intermediate of active demethylation intermediate, formed from the deamination of 5hmC by AID and APOBEC enzymes. However, whereas one study provided evidence to suggest that AID and APOBEC facilitated the deamination of hmC [49], others found such enzymes had little ability to deaminate this modification *in vitro*[50].

To determine the origin of T modifications in mammals, mES cells were grown in the presence of isotopically labeled thymine and methionine, and subsequently analysed by LC-MS/MS (Box 2).[46] Whereas all 5hmU and 5fU bases were found to be derived from isotopically labeled thymidine in wild-type cells, ~7% of 5hmU was found to be derived from 5hmC in cells in which TDG had been knocked down. This indicates that a 5hmC deamination pathway may occur, but that 5hmU resulting from this pathway is rapidly repaired.

Two independent groups have shown that 5hmU can be generated enzymatically from thymine oxidation[46,51] by the TET family of oxidase enzymes. 5hmU levels correlated with TET expression, and subsequent differentiation studies showed that 5hmU appeared to form in processes that also lead to the production of 5hmC and 5fC. Proteomics studies have shown

5hmU to be recognized by chromatin remodeling proteins and transcription factors, suggesting a potential role for it in gene regulation.[46]

T modifications are also prominently observed in the genomes of trypanosomatids **[G]**.[52,53,54] For instance, the hypermodified glucosylated thymine, β-D-glucosyl-hydroxymethyluracil, which is known as Base J (FIG. 2), exists in trypanosomatid species (0.5% J/dT in *Trypanosome brucei*) as do 5hmU and 5fU (0.04% hmU/dT, 0.08% fU/dT *Trypanosome brucei*). Within trypanosomatids, thymine is enzymatically oxidised to 5hmU by the J-binding proteins (JBP)[55], which are Fe(II) and 2-OG dependent dioxygenases homologous to the TET enzymes in mammals. JBP1 is thought to maintain T-modifications, whereas JBP2 is thought to regulate the generation of *de novo* 5hmU biosynthesis. 5hmU is subsequently subject to β-glucosylation by the J-glucosyltransferase (J-GT) enzyme, which results in the production of Base J.[53] Base J exists mainly in telomeric regions **[G]** or repetitive elements **[G]** of DNA, and is enriched at sites of RNA polymerase II **[G]** initiation and termination, consistent with a role for Base J in transcriptional regulation. Depletion of Base J from various trypanosomatids caused increased 'read-through' at transcriptional termination sites or termination defects, leading to altered expression of downstream genes and indicating a direct link between Base J and transcriptional regulation.[56,57] Base J loci are also associated with modified histone H3 variants[58] in certain trypanosomes, providing further evidence of epigenetic crosstalk between DNA and histones in chromatin.

Adenosine modifications: N6-methyladenosine

N6-methyladenosine (6mA) is an adenosine modification that has long been known to exist in the genomes of prokaryotic organisms. By contrast with the DNA modifications discussed so far, 6mA can interfere with Watson–Crick base pairing when the methyl group is in its preferred cis conformation (see FIG. 2). However, the methyl group rotates into the less-favoured trans position when in double stranded DNA, meaning that 6mA destabilizes DNA as compared to A due to the energetic penalty.[59] 6mA in prokaryotes confers resistance against a host immune response as it cannot be broken down by

host endonucleases; however 6mA can also modulate transcription in some bacteria.[60]

6mA is also present in the genomes of simple eukaryotes, such as the green algae Chlamydomonas[61], and there has been a resurgence of interest in 6mA owing to its identification in a number of higher eukaryotic organisms. Specifically, 6mA has been identified in *Caenorhabditis elegans* (nematode worms)[62], *Drosophila melanogaster* (fruit flies)[63] and mammals and vertebrates[64,65], where it is implicated in development.

Further study of 6mA in these species suggests that the role of this mark varies between organisms. 6mA in green algae (0.4% 6mA/dA) appeared to be associated with active gene expression and was found to be highly enriched in linker DNA, suggesting a role in nucleosome positioning[66]. Work with *C. elegans* (0.3% 6mA/dA) indicates that epigenetic crosstalk occurs between 6mA and histone methylation, with evidence suggesting that the 6mA mark could be inherited across generations. [62]

In D. *melanogaster*, 6mA levels were higher in early embryogenesis (0.07% 6mA/dA) to than at later stages of development (0.0001% 6mA/dA). Furthermore, overexpressing an identified DNA 6mA demethylase (DMAD-1) in D. *melanogaster* was lethal to flies in early development, supporting a requirement for 6mA in genomic DNA during D. *melanogaster* development. [63]

6mA was recently reported in vertebrates and mammals, albeit at low levels (0.00009% 6mA/dA in *Xenopus laevis* (frogs)[64], ~0.0006% 6mA/dA in mouse[65]). Work with mES cells found that the presence of 6mA led to transcriptional silencing.[65] Furthermore this study revealed that 6mA was mainly enriched on the X chromosome and a putative demethylase (ALKBH1) has been identified; knockdown of the gene encoding this protein in mES cells led to elevated levels of 6mA and the downregulation of >500 genes. However, ALKBH1 is also known to function as a m1A demethylase[67] and 5-mC dioxygenase in transfer RNA **[G]** ,which might be important for regulating translational [68], and it has also been shown to possess lyase activity at abasic sites[51].

A link between 6mA and early embryogenesis was also demonstrated in zebrafish and pigs by LC-MS/MS, as 6mA were observed to reach levels as

high as 0.1% 6mA/dA in early embryonic stages, before being drastically reduced.[69] This highlights a potential role for 6mA in early mammalian development, which is in line with a role for 6mA in the development of *D. melanogaster.*

It should be noted that one LCMS/MS study could not replicate the levels of 6mA observed by others in mESC and mouse tissues, cautioning that the presence of 6mA in other studies may have arisen as an artifact due to bacterial contamination.[70] This has raised questions over the potential relevance of 6mA in mammalian DNA, and further work is needed to confirm such studies. The study doesn't explore early embryonic samples in which 6mA levels are reported to be substantially higher.

Over the recent years it has become apparent that eukaryotic genomes contain various DNA modifications. Technical advances have allowed us to detect, quantify and determine their genomic loci. The chemistry associated with these technologies is discussed in the next section.

Chemistry-based sequencing methods

Chemistry has played a central role in the development of innovative tools that contribute to the understanding of the function of modified DNA bases. Herein, we focus on chemical approaches for the detection, mapping and sequencing of modified DNA bases in the genomes of various organisms (TABLE 1).[71]

General sequencing approaches to detect modified bases.

The known modified bases preserve the Watson–Crick base-pairing pattern of canonical bases and consequently cannot be detected or decoded by the most widely used sequencing approaches that read the Watson–Crick primary sequence code. Progress has been made in utilizing single-molecule real-time sequencing (SMRT-seq) (see below), which relies on monitoring single-molecules during real-time temporal fluctuations as a polymerase incorporates nucleotides directed by template DNA. The DNA modifications in the template can lead to a characteristic 'signature' during DNA synthesis, although this approach remains challenging for large genomes such as humans. The nanopore sequencing approach (see below) also has the

potential to differentiate modified bases when decoding DNA and may well have broad potential in due course. A practical and general approach to detecting natural DNA modifications in large genomes (such as the human genome), is to enrich DNA fragments that contain the modified base of interest, before DNA sequencing. This can be done by chemically tagging the modified base with an affinity tag, such as biotin (FIG. 4), followed by enrichment of the tagged base with streptavidin magnetic beads that bind to biotin. Alternatively, affinity enrichment can be carried out using antibodies that are specific for the DNA modification of interest. High throughout sequencing of the enriched fragments results in a build-up of 'reads' at certain genomic locations that reveals where such modifications occur in the genome (BOX 1).

Specific restriction endonucleases, coupled with sequencing, can also be utilized to detect the sites of modified bases. Restriction enzymes cleave the phosphodiester backbone of DNA at a particular sequence, and differentially cut at the site depending on the presence or absence of a modification. The enzyme AbasI, for example, is used for the sequencing of 5hmC sites.[72] AbasI preferentially cuts glucosylated 5hmC (5ghmC) and generates cleavage sites with DNA overhangs. Biotinylated adapters containing randomized sequences are then used to hybridize to the cleavage sites and pullout and sequence 5hmC sites. Although this approach provides a simple and cost effective way of detecting modified bases, target sites are cleaved with varying efficiency.[73]

Detecting modified bases at single-base resolution using bisulfite.

To determine the presence of a modified base at single-base resolution, methods that cause a selective chemical transformation that alters the Watson–Crick base-pairing pattern, such as the bisulfite reaction, can be broadly applied to established sequencing platforms. The bisulfite reaction, which decodes the cytosine modification 5mC, catalyses the hydrolytic deamination of cytosine to uracil, whilst 5mC remains resistant to deamination (FIG. 4).[74] Since U has the same base-pairing properties as T, the resultant U

base pairs with A rather than G. The bisulfite-mediated conversion of DNA followed by sequencing, which is referred to as bisulfite-sequencing (BS-seq), creates a global genome-wide map at single-base resolution on which C and 5mC can be distinguished. A drawback of BS-seq is that it cannot distinguish between 5mC and 5hmC, since both are resistant to deamination[75] and cytosine, 5fC and 5caC all convert to uracil under bisulfite conditions.

Chemical and enzymatic methods have been developed to distinguish 5hmC from 5mC (FIG. 4 and TABLE 1). Oxidative bisulfite sequencing (oxBS-seq)[76] uses potassium perruthenate ($KRuO_4$) to first selectively oxidize 5hmC to 5fC, which deaminates to uracil along with C and 5caC during subsequent bisulfite treatment. This means that, in oxBS-seq, only 5mC is read as "C", giving a direct readout of this C modification. Subtracting an oxBS-seq dataset from a BS-seq dataset of the same DNA sample can reveal the presence of 5hmC at single base resolution.

Chemo-enzymatic TET-assisted bisulfite sequencing (TAB-seq)[77] provides another method to resolve 5mC and 5hmC at single-base resolution. 5hmC is first protected by glucosylation using a T4 β-glucosyltransferase (β-GT). Recombinant mouse TET1 (mTET1) is then used for the iterative oxidation of 5mC to 5caC, which deaminates to uracil during subsequent bisulfite treatment. As glucosylated 5hmC is the only remaining C-derivative that does not deaminate, TAB-seq can give a direct readout of 5hmC. Subtracting a TAB-seq dataset from a BS-seq dataset for the same DNA sample reveals where 5mC is present in the genome at single base resolution.

Other methods have since been developed for the sequencing of modified C bases (FIG. 4 and TABLE1).[71] Reductive bisulfite sequencing (redBS-seq), for example, uses $NaBH_4$ to reduce 5fC to 5hmC.[24] Similar to the procedure for oxBS-seq, subtraction of the redBS-seq dataset from a BS-seq dataset allows the single base resolution sequencing of 5fC. Another bisulfite-based method exploits the use of chemical treatment with $EtONH_2$, which protects 5fC from bisulfite-mediated deamination; this method is called 5fC chemically assisted bisulfite sequencing. (fCAB-seq)[78]. Comparing the fCAB-seq dataset with the

BS-seq dataset allows the genomic location of 5fC to be determined at single base resolution.

While bisulfite based sequencing is regarded as the 'gold standard' for the analysis of C modifications, it has serious deficiencies. Bisulfite treatment induces the loss of pyrimidine bases from the DNA strand, which subsequently facilitates strand cleavage via β-elimination **[G]** and δ-elimination **[G]** of up to 99% of the original DNA.[79] Several rounds of amplification by polymerase chain reaction (PCR) are therefore required to generate enough DNA for sequencing. While the amplification by PCR is a fundamental step, especially when working with low amounts of DNA, it may introduce PCR biases. Two independent approaches have addressed this issue by developing amplification-free bisulfite sequencing methods called recovery after bisulfite treatment (ReBUiLT)[80] and post-bisulfite adaptor tagging (PBAT)[81]. These methods result in increased uniformity of coverage and an adapted version of the latter technique is now the standard method for analyzing DNA methylation in single cells.[82] In this adaptation of PBAT, bisulfite conversion is performed on single cells prior to five rounds of random priming to tag and amplify the material.

Bisulfite-free techniques to detect modified bases at single-base resolution

Bisulfite-free base resolution sequencing methods have recently been developed that enable C-to-T transitions during PCR. One such technique takes advantage of the conversion of hmC to trihydroxylated thymine using peroxotungstate[83], whereas another, 5fC cyclization-enabled C-to-T transition (fC-CET) sequencing, selectively targets 5fC using derivatives of 1,3-indianone [84] (FIG. 4 and TABLE 1).

Modified bases can also be detected at single-base resolution, without any prior chemical transformation, using SMRT-seq.[85] In this single-molecule sequencing technology, fluorescent nucleosides, each of which has a unique kinetic signature, are inserted opposite the template strand by a processing polymerase. Modified bases in the template strand can be differentiated from canonical bases as the polymerase has a greater tendency to pause in the

presence of a modification. Chemical labeling is commonly used in combination with SMRT-seq to improve signal detection, as bulkier modified bases are easier to distinguish from canonical bases.[86] Simultaneous mapping of 5hmC, 5fC and 5caC was achieved using chemical labeling and SMRT-seq at near single base resolution in a fungal model system.[87] SMRT-seq can readily detect Base J[88] and 6mA[65] at single-base resolution without the necessity of chemical tagging. Finally, as well as not requiring the DNA to be chemically transformed, SMRT-seq also enables longer reads (up to 20,000 bp) compared to Illumina sequencing. However, this sequencing method still has reasonably low throughput, meaning it is only feasible for detecting modified bases in smaller genomes or in larger genomes after enrichment.

Another emerging single molecule sequencing technique that does not require chemical or enzymatic transformation or amplification is nanopore sequencing. This technique measures the current as each nucleobase travels through a nanopore electrophoretically. This technology has been successfully used to sequence small viral and microbial genomes and, more recently, to map and quantify 5mC and 5hmC in synthetic and genomic DNA.[89,90,91,92,93,94] Improvement in this technology now allows the detection of 5mC in the human genome with 82% accuracy[92] and we anticipate that nanopore sequencing instruments may be used more routinely in the future.

DNA modifications in disease and therapy

It is emerging that DNA modifications not only play fundamental roles during normal development, but also may contribute to disease progression. Ultimately, understanding and control over these processes may lead to novel therapeutic approaches as outlined in the next section.

DNA base modifications in disease.

The epigenetic profile of a genome reflects its cellular state and can reflect the identity of the cell or tissue. In an evolving cancer genome, the epigenetic state of DNA bases is thought to influence the mutational changes

that occur, such that the mutation rate variation is linked to the epigenomic features of the cell of origin.[95,96,97]

Two independent studies have reported, for example, that 5mC is more likely than 5hmC to undergo C-to-T changes in the context of CpG in human malignancies.[98,99] Also, in disease states, epigenetic marks can become aberrant and methods to detect such changes have the potential to detect, diagnose and monitor disease. For example, atypical methylation patterns at specific CpG islands and in tumor suppressor genes is a hallmark of cancer.[100,101,102] In addition, certain genes are hypermethylated in a number of neurodegenerative disorders, including Alzheimer's disease.[103]

A global reduction in 5hmC levels has been observed in many cancers. This reduction may arise from mutations that result in dysfunctional TET proteins; the increased production of the TET inhibiting metabolite 2-hydroxygluturate, which can be caused by mutations in the genes encoding enzymes such as isocitrate dehydrogenase (IDH)); or the silencing of TET proteins as a result of methylation of the genes that encode them.[104,105,106,98] Alternatively, TET-mediated 5hmC simply may not be maintained in cancerous tissue as there is a correlation between 5hmC levels and cell proliferation, i. e. 5hmC is passively diluted out of the genome.[29] Although global epigenetic changes might be indicative of disease, clinical case studies that identify specific epigenetic changes that mark, or more preferably indicate the cause of, the disease state may lead to clinical tests. Indeed, a 5mC based diagnostic test was recently approved for the non-invasive detection of colorectal cancer, which paves the way for other epigenetics-based diagnostic tests in the future.[107] Also routinely used in the diagnosis of glioblastoma is a test that measures the level of methylation at the promoter of the O6-methylguanine–DNA methyltransferase (*MGMT*) gene. MGMT is involved in the repair of alkylating agent-induced damage in DNA and methylation of the *MGMT* promoter decreases the expression of MGMT protein and, therefore, potentially increases the sensitivity of patients with glioblastoma to therapy.[108]

There is also a potential link between T modifications and disease in mammalian systems. 5fU is a known mutagen in mammalian tissue, likely due to its propensity to mispair with guanine in DNA[109], and elevated blood levels of 5hmU have been linked with breast and renal cancer.[110,111] Levels of uracil are reported to be elevated in cancerous tissue, and enzymatic deamination that is mediated by AID and APOBEC is thought to cause C-to-T transitions in tumour suppressor genes or oncogenes and to be linked with the onset of disease.[112] These cytidine deaminases are overexpressed in certain cancers, although there is currently some debate as to whether the mutations they cause are derived from the deamination of cytosine, or of 5mC, or of both.[113] AID overexpression is associated with global demethylation, with mutational biases located at CpG sites, although AID enzymes demonstrate greater activity towards unmethylated cytosine *in vitro*.[50]

Therapeutic approaches for targeting DNA base modifications in disease.

Due to the link between aberrant epigenetic regulation and disease, it is attractive to consider therapeutics that can reprogramme the epigenome of cells towards a non-disease state. Inhibitors of DNA methylation can reduce methylation and increase the expression of certain genes in disease states. For example, the ribo- and deoxyribonucleoside analogues of 5-azacytidine are currently used in the clinic to treat cancer including leukemia.[114] They are thought to become phosphorylated on three sites before being incorporated into genomic DNA during replicative DNA synthesis. In 5-azacytidine the ring carbon C5 is replaced by a nitrogen, after attack of the DNMT's cysteine at C6, the enzyme is not released causing the formation of an irreversible covalent adduct with, and the inhibition of DNMT enzymes; this inhibition results in the global loss of cytosine methylation, in addition to some DNA damage response. For cancers that exhibit loss of 5hmC due to mutations in the gene encoding IDH, inhibitors of aberrant IDH are currently in clinical trials[115] with the aim of alleviating the inhibition of TET activity to restore a healthy epigenome.

Inhibitor based approaches that cause global epigenetic reprogramming could be cytotoxic as it might affect the expression of many genes. Approaches for

editing the epigenome at specific sites are therefore also being explored. For example, minor-groove polyamides have been shown to specifically target CpG sequences, leading to the inhibition of DNMTs at this particular sequence.[116] Clustered regularly interspaced short palindromic repeats (CRISPR) systems **[G]** or transcription-activator like effectors (TALEs) **[G]** that have been tethered with DNA modifiers such as DNMT and TET have also been shown to specifically alter gene expression, highlighting a potential approach for future therapy.[117,118]

There are also opportunities to develop epigenetic therapies that target the unique regulatory system used in trypanosomatids. Inhibitors could be developed to specifically target JBP or JGT enzymes; targeting these enzymes would globally deplete Base J, which affects transcriptional regulation in these systems. Such therapeutics could be useful to treat a host of tropical diseases that are associated with trypanosomatids, such as Leishmaniasis and African sleeping sickness,[119] potentially without side-effects as these enzymes are absent in humans.

Concluding remarks and future outlook

The discovery and subsequent insights gained into oxidised cytosine modifications in the human genome has stimulated considerable interest in the chemical modification of nucleotides. The discovery of 5hmC in mammalian DNA was partly inspired by knowledge of the homology between JBP enzymes in trypanosomatids and the mammalian TET enzymes.[120,121] The growing knowledge of known DNA base modifications and the huge expansion in our knowledge of genomes and genome function provides information and inspiration for other exciting discoveries in due course.

State-of-the-art chemical analysis techniques, such as LC-MS/MS, now allows the abundance and formation of relatively rare base modifications, such as 5hmU, to be identified as veritable DNA modifications rather than ubiquitous damage in mammals. In the future, the sensitivity of such methods will improve and will enable us to uncover more about other chemical modifications in the genome. It will be important to confirm that observed modifications, especially rare ones, are natural to the genome and rule out the

possibility of artifacts and background contamination. The use of stable isotope tracing experiments in combination with mass spectrometry can help with this and can also provide an understanding of the metabolism and dynamics of DNA base modifications. Understanding the enzymology associated with DNA modifications will be essential in elucidating their roles in biology. Ultimately, sequencing all modified bases in a single experiment, preferably at the single cell level, is a desirable goal and also a major technical challenge. The exploitation of the natural chemo-physical characteristics of base modifications in single molecule approaches such as SMRT-seq and nanopore sequencing is an attractive possibility to consider in the future. DNA comprises a wide, natural and dynamic chemical repertoire that encodes information in living systems. Further advances in chemical methodologies can help drive a more complete understanding of how and why the chemistry of DNA can reprogramme the biology of cells and organisms.

Box 1 Figure:

Box 2 Figure:

BOX 1: Mapping modified DNA bases using DNA immunoprecipitation followed by high throughput sequencing.

To prepare DNA for sequencing, it is first fragmented by sonication or broken-down into smaller fragments of around 250 bp (see the figure, step 1). These DNA fragments then undergo a 'library preparation' procedure, which involves the ligation of DNA sequencing adaptors to them using a DNA ligase (see the figure, step 2). DNA fragments can then either be directly hybridized onto the flow cell and bridge amplified for the enhancement of the fluorescent signal (see the figure, arrow pointing down, step 3a) or first enriched for the desired DNA modification (see the figure, arrow pointing to the right, step 3b) followed by hybridization and amplification. [122,123] Complementary dNTPs, which are labeled with a specific fluorophore[124,125], are inserted opposite the template strand *via* Watson–Crick base pairing by a DNA polymerase (see the figure, step 4). Protecting groups present on the 3'-OH group of each nucleoside control the incorporation to a single nucleotide. After each

nucleotide is incorporated, the fluorophore is read by imaging (see the figure, step 5) to reveal the identity of the base. Subsequent deprotection of the 3'-OH protecting group and removal of the fluorescent label *via* Staudinger reduction chemistry allows the cycle to be repeated. This process leads to build up of sequences that are obtained, known as 'reads', and can be bioinformatically aligned to the reference genome of any organism.

Box 2| LC-MS/MS based methods for the global quantitation of modified bases.

Quantitative detection of modified DNA bases is a powerful tool for the discovery and elucidation of such modification in the genomes of organisms.[29,52,46,23,126,127] Liquid chromatography-tandem mass spectrometry (LC-MS/MS) is a chemically discriminating method of nucleotide analysis that is used for quantitatively detecting modified and canonical nucleosides directly from digested samples of DNA (see the figure, part a). During LC-MS/MS, genomic DNA is digested into mononucleosides that are subsequently separated by their polarity *via* liquid chromatography before injection into the mass spectrometer. The unique mass of the parent modified base ion is measured, allowing the accurate mass of nucleoside fragments to be extracted from the total ion count (TIC). This gives a definitive mass signal that is unique to a particular modified base and that can be accurately integrated for quantification. The integration of the mass signal is then compared to the linear fit equation from calibration curves of synthetic standards to determine the concentration of modified base in each sample. The gold-standard of mass quantification employs an internal standard that is isotopically labelled (SIL) to enable robust and accurate quantitation of DNA modifications typically at femtomole quantities.[128] (see the figure, part b). Ideally the SIL has chromatographic and ionisation properties that are virtually the same as the non-labelled analyte but a different mass signal that can be integrated separately (for example, $^{13}CD_3$ is used to label 5mC; $^{15}N_3$ is used to measure dC, note ^{13}C and ^{15}N isotopes are annotated with blue asterisks) The amount of isotope-labeled internal standard is added to the calibration line and to each digested sample in equal amounts. The area ratio between the labeled and non-labeled base can then be used for improved precision in the quantificationThe SIL standard and analyte itself will co-elute, correcting for matrix effects, and will also account for variation in injection volume or ion suppression.

Modern-day mass spectrometers can accurately quantify over a wide-range of concentrations and are sensitive enough to detect femtomol (fmol) levels of (modified) nucleosides.[46] Mass signals can be improved by HPLC or affinity pre-enrichment (for example, biotin can be used to concentrate sample by

affinity enrichment prior to injection of large quantities of digested DNA prior to LC-MS/MS injection[52,23], or *via* the prior chemical derivatization of modified bases with a MS sensitizer tag containing for example a quartenary nitrogen.[129] This has been applied for the detection of low abundance modifications (< 0.001% of bases) such as 5fC[23] and 5fU[129] by hydrazone formation with such a MS sensitizer. LC-MS/MS has been used to look at the difference in modifications between different organisms, tissues, cancer-states and ages of tissue.[130,131,132] Quantification of modified bases can also be used to determine or validate the effect of depletion or overexpression of certain modifying enzymes.[63,46]

LC-MS/MS can also be used to look at the lifetime, origin and dynamics of DNA modifications *via* the use of metabolic isotopic labeling (see the figure, part c). The C modifications can be traced *via* the feeding of cells and mouse models with $^{13}CD_3$ methionine. The SIL labeled methionine is converted to the methyl donor *S*-adenosyl-[$^{13}CD_3$]-methionine (SAM) and through catalysis of the DNMT enzyme, leads to the formation of +4 isotopically-labelled 5-methylcytidine. As a result, the oxidative derivatives, that is, 5hmC and 5fC[29,30] will also become SIL labeled as depicted in figure panel c. Other modified bases have also been studied by this type of isotope tracing experiments. For example, the formation of 5hmU has been discerned by feeding combinations of stable isotope labeled methionine and a labeled thymidine.[46]

Figure 1: **Watson and Crick base pairing and DNA grooves.**

A) The primary genetic code comprises the nucleobases adenine (A), guanine (G), cytosine (C) and thymine (T). Deoxyguanosine (dG) pairs with deoxycytidine (dC), and deoxyadenosine (dA) pairs with deoxythymidine (dT),

in an anti-parallel arrangement to form a secondary structure. B) The major and minor grooves of the AT, TA, CG and GC base pairs are shown. Arrows indicate hydrogen donors (purple arrows) and hydrogen acceptors (cyan arrows), note how this pattern changes between the major but not the minor grooves of AT versus TA and CG versus GC. C) On top of the primary and secondary code (major groove) lies the epigenetic code, which has additional functionality and is depicted here by methylation pointing out into the major groove.

Figure 2: **A) Overview of modified DNA bases**. The four canonical bases are depicted, with the sites of known modification shown in red font or circled. **B)** Chemical space of the modified DNA bases that have been reported in the literature. Names represent the bases depicted, R indicates 2'-deoxy-D-ribose and names in a box are modifications that have been identified in eukaryotic genomes and are discussed in the scope of this Review; other modifications have been reviewed elsewhere[5] *More complex branching of mixtures with up to three furanose residues (Glc and Gal) have also been reported. ** The precise distribution of the two substituents over the 4- and 5-hydroxyl groups has not been determined. Note that 6mA is shown with the N6-Me group in its energetically preferred cis conformation.

Figure 3: **The active demethylation pathway.** In eukaryotes, methylation at the C5 position of cytosine (C) is catalyzed by DNA methyltransferases (DNMTs), which use *S*-adenosylmethionine (SAM) as the methyl donor, to result in the formation of 5mC (orange arrow). The postulated active demethylation pathway involves the iterative oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC) via TET enzymes (pink arrows). 5fC and 5caC then can be excised by thymine DNA glycosylase (TDG, green arrows) to form an apyrimidinic (AP) sites that further undergoes base excision repair (BER, yellow arrow) subsequently reinstalls the cytosine.

Figure 4| Methods for the sequencing modified C bases. A) Bisulfite induced deamination of cytosine to uracil changes the base-pairing and is the

underlying principle of bisulfite sequencing B) Oxidative bisulfite sequencing (OxBS), Reductive bisulfite sequencing (RedBS) and TET-assisted bisulfite (TAB) sequencing add an additional chemical or enzymatic step prior to bisulfite treatment, which enables the discrimination between mC, hmC and fC at single base resolution C) fC-CET (5fC cyclization-enabled C-to-T transition) is a bisulfite-free sequencing method that relies on the selective chemical manipulation of 5fC that changes base-pairing upon PCR amplification D) Two examples of chemical enrichment methods to pulldown modified bases. 5hmC can be chemoenzymatically enriched by coupling to Uridine-5'-(6-deoxy-6-azido-α-D-glucopyranosyl diphosphate) (UDP-6-Azido-Glucose) and subsequent click reaction to a biotin linker to enable pull-down.

Table 1| Approaches for mapping modified DNA bases.

Base derivative	Chemical enrichment techniques	Antibody Enrichment techniques	Single-base resolution techniques
5mC	No technique published	5Me-DIP[133]	Bisulfite ox-BS[76]
5hmC	GLIB-seq[134] hMeSeal-seq[135]	5hme-DIP[136] CMS-DIP[134] JBP1-DIP[137]	TAB-seq[77] SCL-exo[138] AbaSI-seq[72] PvuRts1[139]
5fC	Aldehyde reactive probe[140,27] fCSeal-seq[78]	5fC-DIP[28]	fC-CET[84] red-BS[24] fCAB-seq[78] MAB-seq[141] CLEVER-seq[142]
5caC	No technique published	5caC-DIP[28]	DIP-Cab-seq[143] MAB-seq[144]
5hmU	$KRuO_4$ / Aldehyde reactive probe[54]	5hmU-DIP[54]	No technique published
5fU	Aldehyde reactive probe[54]	No technique published	No technique published
Base J	$NaIO_4$ / Aldehyde reactive probe[54]	Base J-DIP[56]	No technique published

6mA	No technique published	6mA-DIP[63]	6mA-RE-seq[66]

GLIB-seq: Glucosylation, periodate oxidation and biotinylation sequencing

hMe-Seal: 5hmC selective chemical labeling

DIP: DNA immunoprecipitation

SCL-exo: Selective chemical labeling-exonuclease digestion

fCAB seq: 5fC chemically assisted bisulfite sequencing

MAB-seq: Methylase assisted bisulfite sequencing

CLEVER-seq: Chemical labeling enabled C-to-T conversion sequencing

6mA-RE seq: 6mA restriction enzyme guided sequencing

References

1. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171,** 737–738 (1953).
2. Klug, A. The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation. *Q. Rev. Biophys.* **43,** 1–21 (2010).
3. Dervan, P. B. Molecular recognition of DNA by small molecules. *Bioorg. Med. Chem.* **9,** 2215–2235 (2001).
4. GW., R. Zur chemie der tuberkelbacillen. *Z Physiol Chem* **26,** 218–232 (1898).
5. Gommers-Ampt, J. H. & Borst, P. Hypermodified bases in DNA. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **9,** 1034–1042 (1995).
6. Johnson, T. B. & Coghill, R. D. Researches on Pyrimidines. C111. The Discovery of 5-Methyl-Cytosine in Tuberculinic acid, the nucleic acid of the Tubercle bacillus. *J. Am. Chem. Soc.* **47,** 2838–2844 (1925).
7. Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* **19,**

219–220 (1998).

8. Jeltsch, A. Molecular enzymology of mammalian DNA methyltransferases. *Curr. Top. Microbiol. Immunol.* **301,** 203–225 (2006).
9. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell* **99,** 247–257 (1999).
10. Barau, J. *et al.* The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science (80-.).* **354,** 909–912 (2016).
11. Wyatt, G. R. Occurrence of 5-Methyl-Cytosine in Nucleic Acids. *Nature* **166,** 237–238 (1950).
12. Bird, A. DNA methylation patterns and epigenetic memory DNA methylation patterns and epigenetic memory. *Genes Dev.* 6–21 (2002). doi:10.1101/gad.947102
13. Deaton, A. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
14. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13,** 484–492 (2012).
15. Du, J., Johnson, L. M., Jacobsen, S. E. & Patel, D. J. DNA methylation pathways and their crosstalk with histone methylation. *Nat Rev Mol Cell Biol* **16,** 519–532 (2015).
16. Wachter, E. *et al.* Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *Elife* **3,** e03397 (2014).
17. Krebs, A. R., Dessus-Babus, S., Burger, L. & Schubeler, D. High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *Elife* **3,** e04094 (2014).
18. Hahn, M. A., Szabo, P. E. & Pfeifer, G. P. 5-Hydroxymethylcytosine: a stable or transient DNA modification? *Genomics* **104,** 314–323 (2014).
19. Bogdanović, O. *et al.* Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat Genet.* **48,** (2016).
20. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502,** 472–479 (2013).

21. Cortázar, D. *et al.* Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature.* **470,** (2011).
22. Dawlaty, M. M. *et al.* Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Dev. Cell* **29,** 102–111 (2014).
23. Pfaffeneder, T. *et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew Chem Int Ed Engl.* **50,** (2011).
24. Booth, M. J., Marsico, G., Bachman, M., Beraldi, D. & Balasubramanian, S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat Chem.* **6,** (2014).
25. Su, M. *et al.* 5-Formylcytosine Could Be a Semipermanent Base in Specific Genome Sites. *Angew. Chemie Int. Ed.* **55,** 11797–11800 (2016).
26. McInroy, G. R. *et al.* in *Epigenetic Mechanisms in Cellular Reprogramming* (eds. Meissner, A. & Walter, J.) 167–191 (Springer Berlin Heidelberg, 2015). doi:10.1007/978-3-642-31974-7_8
27. Iurlaro, M. *et al.* In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome Biol.* **17,** 141 (2016).
28. Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell.* **153,** (2013).
29. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem.* **6,** (2014).
30. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat Chem Biol.* **11,** (2015).
31. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14,** (2013).
32. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell.* **152,** (2013).
33. Wang, D. *et al.* MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw1184
34. Sato, K., Kawamoto, K., Shimamura, S., Ichikawa, S. & Matsuda, A. An oligodeoxyribonucleotide containing 5-formyl-2′-deoxycytidine (fC) at the CpG site forms a covalent complex with DNA cytosine-5

methyltransferases (DNMTs). *Bioorg. Med. Chem. Lett.* **26,** 5395–5398 (2016).

35. Thalhammer, A., Hansen, A. S., El-Sagheer, A. H., Brown, T. & Schofield, C. J. Hydroxylation of methylated CpG dinucleotides reverses stabilisation of DNA duplexes by cytosine 5-methylation. *Chem. Commun.* **47,** 5325–5327 (2011).
36. Raiber, E.-. A. *et al.* 5-Formylcytosine alters the structure of the DNA double helix. *Nat Struct Mol Biol.* **22,** (2015).
37. Ngo, T. T. M. *et al.* Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.* **7,** 10813 (2016).
38. Choy, J. S. *et al.* DNA Methylation Increases Nucleosome Compaction and Rigidity. *J. Am. Chem. Soc.* **132,** 1782–1783 (2010).
39. Lee, J. Y. & Lee, T.-H. Effects of DNA Methylation on the Structure of Nucleosomes. *J. Am. Chem. Soc.* **134,** 173–175 (2012).
40. Mendonca, A., Chang, E. H., Liu, W. & Yuan, C. Hydroxymethylation of DNA influences nucleosomal conformation and stability in vitro. *Biochim. Biophys. Acta* **1839,** 1323–1329 (2014).
41. Collings, C. K., Waddell, P. J. & Anderson, J. N. Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res.* **41,** 2918–2931 (2013).
42. Galashevskaya, A. *et al.* A robust, sensitive assay for genomic uracil determination by LC/MS/MS reveals lower levels than previously reported. *DNA Repair (Amst).* **12,** 699–706 (2013).
43. Krokan, H. E., Drabløs, F. & Slupphaug, G. Uracil in DNA--occurrence, consequences and repair. *Oncogene* **21,** 8935–8948 (2002).
44. Conticello, S. G. The AID/APOBEC family of nucleic acid mutators. *Genome Biol.* **9,** 229 (2008).
45. Santos, F. *et al.* Active demethylation in mouse zygotes involves cytosine deamination and base excision repair. *Epigenetics Chromatin* **6,** 39 (2013).
46. Pfaffeneder, T. *et al.* Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat Chem Biol* **10,** 574–581 (2014).
47. Masaoka, A. *et al.* Mammalian 5-formyluracil-DNA glycosylase. 2. Role of SMUG1 uracil-DNA glycosylase in repair of 5-formyluracil and other

oxidized and deaminated base lesions. *Biochemistry* **42,** 5003–5012 (2003).

48. Bauer, N. C., Corbett, A. H. & Doetsch, P. W. The current state of eukaryotic DNA base damage and repair. *Nucleic Acids Res.* **43,** 10083–10101 (2015).
49. Guo, J. U., Su, Y., Zhong, C., Ming, G. & Song, H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* **145,** 423–434 (2011).
50. Nabel, C. S. *et al.* AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nat. Chem. Biol.* **8,** 751–758 (2012).
51. Pais, J. E. *et al.* Biochemical characterization of a Naegleria TET-like oxygenase and its application in single molecule sequencing of 5-methylcytosine. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 4316–21 (2015).
52. Liu, S. *et al.* Quantitative Mass Spectrometry-Based Analysis of β-D-Glucosyl-5-hydroxymethyluracil in Genomic DNA of Trypanosoma brucei. *J. Am. Soc. Mass Spectrom.* **25,** 1763–1770 (2014).
53. Bullard, W., Lopes Da Rosa-Spiegler, J., Liu, S., Wang, Y. & Sabatini, R. Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome. *J. Biol. Chem.* **289,** 20273–20282 (2014).
54. Kawasaki, F. *et al.* Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite Leishmania. *Genome Biol.* **18,** 23 (2017).
55. Cliffe, L. J., Siegel, T. N., Marshall, M., Cross, G. A. M. & Sabatini, R. Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. *Nucleic Acids Res.* **38,** 3923–3935 (2010).
56. van Luenen, H. G. A. M. *et al.* Glucosylated Hydroxymethyluracil, DNA Base J, Prevents Transcriptional Readthrough in Leishmania. *Cell* **150,** 909–921 (2012).
57. Hazelbaker, D. Z. & Buratowski, S. Base J: Blocking RNA Polymerase's Way. *Curr. Biol.* **22,** R960–R962 (2012).
58. Reynolds, D. *et al.* Histone H3 Variant Regulates RNA Polymerase II

Transcription Termination and Dual Strand Transcription of siRNA Loci in Trypanosoma brucei. *PLOS Genet.* **12,** e1005758 (2016).

59. Engel, J. D. & von Hippel, P. H. Effects of methylation on the stability of nucleic acid conformations. Studies at the polymer level. *J. Biol. Chem.* **253,** 927–934 (1978).
60. Ratel, D., Ravanat, J.-L., Berger, F. & Wion, D. N6-methyladenine: the other methylated base of DNA. *Bioessays* **28,** 309–315 (2006).
61. Rae, P. M. M. & Steele, R. E. Modified bases in the DNAs of unicellular eukaryotes: an examination of distributions and possible roles, with emphasis on hydroxymethyluracil in dinoflagellates. *Biosystems* **10,** 37–53 (1978).
62. Greer, E. L. *et al.* DNA Methylation on N6-Adenine in C. elegans. *Cell* **161,** 868–878 (2015).
63. Zhang, G. *et al.* N^6-Methyladenine DNA Modification in *Drosophila*. *Cell* **161,** 893–906 (2017).
64. Koziol, M. J. *et al.* Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nat Struct Mol Biol* **23,** 24–30 (2016).
65. Wu, T. P. *et al.* DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature* **532,** 329–333 (2016).
66. Fu, Y. *et al. N*6-Methyldeoxyadenosine Marks Active Transcription Start Sites in *Chlamydomonas*. *Cell* **161,** 879–892 (2017).
67. Liu, F. *et al.* ALKBH1-Mediated tRNA Demethylation Regulates Translation. *Cell* **167**(3), 816-828 (2016).
68. Kawarada, L. *et al.* ALKBH1 is an RNA dioxygenase responsible for cytoplasmic and mitochondrial tRNA modifications. *Nucleic Acids Res.* **45**, 7401-7415 (2017)
69. Liu, J. *et al.* Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nat. Commun.* **7,** 13052 (2016).
70. Schiffers, S. *et al.* Quantitative LC–MS Provides No Evidence for m6dA or m4dC in the Genome of Mouse Embryonic Stem Cells and Tissues. *Angew. Chemie Int. Ed.* **56**, 1-5 (2017)
71. Booth, M. J., Raiber, E.-A. & Balasubramanian, S. Chemical Methods for Decoding Cytosine Modifications in DNA. *Chem. Rev.* **115,** 2240–

2254 (2015).

72. Sun, Z. *et al.* High-Resolution Enzymatic Mapping of Genomic 5-Hydroxymethylcytosine in Mouse Embryonic Stem Cells. *Cell Rep.* **3,** 567–576 (2013).
73. Horton, J. R. *et al.* Structure of 5-hydroxymethylcytosine-specific restriction enzyme, AbaSI, in complex with DNA. *Nucleic Acids Res.* **42,** 7947–7959 (2014).
74. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 1827–31 (1992).
75. Huang, Y. *et al.* The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing. *PLoS One* **5,** e8888 (2010).
76. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336,** 934–937 (2012).
77. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149,** 1368–1380 (2012).
78. Song, C.-. X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell.* **153,** (2013).
79. Tanaka, K. & Okamoto, A. Degradation of DNA by bisulfite treatment. *Bioorg. Med. Chem. Lett.* **17,** 1912–1915 (2007).
80. McInroy, G. R. *et al.* Enhanced Methylation Analysis by Recovery of Unsequenceable Fragments. *PLoS One* **11,** e0152322 (2016).
81. Miura, F., Enomoto, Y., Dairiki, R. & Ito, T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res.* **40,** (2012).
82. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Meth* **11,** 817–820 (2014).
83. Hayashi, G. *et al.* Base-Resolution Analysis of 5-Hydroxymethylcytosine by One-Pot Bisulfite-Free Chemical Conversion with Peroxotungstate. *J. Am. Chem. Soc.* **138,** 14178–14181 (2016).
84. Xia, B. *et al.* Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat Methods.* **12,** (2015).
85. Feng, Z. *et al.* Detecting DNA Modifications from SMRT Sequencing

Data by Modeling Sequence Context Dependence of Polymerase Kinetic. *PLOS Comput. Biol.* **9,** e1002935 (2013).

86. Song, C.-X. *et al.* Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. *Nat Meth* **9,** 75–77 (2012).
87. Chavez, L. *et al.* Simultaneous sequencing of oxidized methylcytosines produced by TET/JBP dioxygenases in Coprinopsis cinerea. *Proc. Natl. Acad. Sci.* **111,** E5149–E5158 (2014).
88. Genest, P. A. *et al.* Defining the sequence requirements for the positioning of base J in DNA using SMRT sequencing. *Nucleic Acids Res.* **43,** 2102–2115 (2015).
89. Wanunu, M. *et al.* Discrimination of Methylcytosine from Hydroxymethylcytosine in DNA Molecules. *J. Am. Chem. Soc.* **133,** 486–492 (2011).
90. Laszlo, A. H. *et al.* Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 18904–9 (2013).
91. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14,** R51 (2013).
92. Simpson, J. T. *et al.* Detecting DNA Methylation using the Oxford Nanopore Technologies MinION sequencer. *bioRxiv* (2016). doi:10.1101/047142
93. Hoenen, T. *et al.* Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. *Emerg. Infect. Dis.* **22,** 331–334 (2016).
94. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–232 (2016).
95. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518,** 360–364 (2015).
96. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504–507 (2012).
97. Liu, L., De, S. & Michor, F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.* **4,** 1502 (2013).
98. Raiber, E.-A. *et al.* Base resolution maps reveal the importance of 5-

hydroxymethylcytosine in a human glioblastoma. *npj Genomic Med.* **2,** 6 (2017).

99. Tomkova, M., McClellan, M., Kriaucionis, S. & Schuster-Boeckler, B. 5-hydroxymethylcytosine marks regions with reduced mutation frequency in human DNA. *Elife* **5,** (2016).
100. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* **8,** 286–298 (2007).
101. Baylin, S. B. *et al.* Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum. Mol. Genet.* **10,** 687–692 (2001).
102. Yoo, C. B. & Jones, P. A. Epigenetic therapy of cancer: past, present and future. *Nat Rev Drug Discov* **5,** 37–50 (2006).
103. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nat Biotech* **28,** 1057–1068 (2010).
104. Pfeifer, G. P., Xiong, W., Hahn, M. A. & Jin, S.-G. The role of 5-hydroxymethylcytosine in human cancer. *Cell Tissue Res.* **356,** 631–641 (2014).
105. Mou, H., Kennedy, Z., Anderson, D. G., Yin, H. & Xue, W. Precision cancer mouse models through genome editing with CRISPR-Cas9. *Genome Med.* **7,** 53 (2015).
106. Kim, Y.-H. *et al.* TET2 promoter methylation in low-grade diffuse gliomas lacking IDH1/2 mutations. *J. Clin. Pathol.* **64,** 850–852 (2011).
107. Imperiale, T. F. *et al.* Multitarget Stool DNA Testing for Colorectal-Cancer Screening. *N. Engl. J. Med.* **370,** 1287–1297 (2014).
108. E., H. M. *et al.* MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma. *N. Engl. J. Med.* **352,** 997–1003 (2005).
109. Klungland, A. *et al.* 5-Formyluracil and its nucleoside derivatives confer toxicity and mutagenicity to mammalian cells by interfering with normal RNA and DNA metabolism. *Toxicol. Lett.* **119,** 71–78 (2001).
110. Djuric, Z. *et al.* Levels of 5-hydroxymethyl-2′-deoxyuridine in DNA from blood as a marker of breast cancer. *Cancer* **77,** 691–696 (1996).
111. Djuric, Z. *et al.* Levels of 5-hydroxymethyl-2’-deoxyuridine in DNA from blood of women scheduled for breast biopsy. *Cancer Epidemiol. Biomarkers Prev.* **10,** 147–149 (2001).

112. Rebhandl, S., Huemer, M., Greil, R. & Geisberger, R. AID/APOBEC deaminases and cancer. *Oncoscience* **2,** 320–333 (2015).
113. Pfeifer, G. P. Mutagenesis at methylated CpG sequences. *Curr. Top. Microbiol. Immunol.* **301,** 259–281 (2006).
114. Stresemann, C. & Lyko, F. Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine. *Int. J. cancer* **123,** 8–13 (2008).
115. Stein, E. *et al.* Abstract CT103: Clinical safety and activity in a phase I trial of AG-221, a first in class, potent inhibitor of the IDH2-mutant protein, in patients with IDH2 mutant positive advanced hematologic malignancies. *Cancer Res.* **74,** CT103--CT103 (2014).
116. Kang, J. S., Meier, J. L. & Dervan, P. B. Design of Sequence-Specific DNA Binding Molecules for DNA Methyltransferase Inhibition. *J. Am. Chem. Soc.* **136,** 3687–3694 (2014).
117. Kubik, G. & Summerer, D. TALEored Epigenetics: A DNA-Binding Scaffold for Programmable Epigenome Editing and Analysis. *ChemBioChem* **17,** 975–980 (2016).
118. Xu, X. *et al.* A CRISPR-based approach for targeted DNA demethylation. *Cell Discov.* **2,** 16009 (2016).
119. Borst, P. & Sabatini, R. Base J: discovery, biosynthesis, and possible functions. *Annu. Rev. Microbiol.* **62,** 235–251 (2008).
120. Iyer, L. M., Tahiliani, M., Rao, A. & Aravind, L. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle* **8,** 1698–1710 (2009).
121. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science.* **324,** (2009).
122. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107,** 1–8 (2016).
123. Head, S. R. *et al.* Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques* **56,** (2014).
124. Balasubramanian, S. Sequencing nucleic acids: from chemistry to medicine. *Chem. Commun.* **47,** 7281 (2011).
125. Bentley, D. R. *et al.* Accurate whole human genome sequencing using

reversible terminator chemistry. *Nature* **456,** (2008).

126. Taghizadeh, K. *et al.* Quantification of DNA damage products resulting from deamination, oxidation and reaction with products of lipid peroxidation by liquid chromatography isotope dilution tandem mass spectrometry. *Nat. Protoc.* **3,** 1287–1298 (2008).
127. Globisch, D. *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* **5,** e15367 (2010).
128. Gackowski, D. *et al.* Accurate, Direct, and High-Throughput Analyses of a Broad Spectrum of Endogenously Generated DNA Base Modifications with Isotope-Dilution Two-Dimensional Ultraperformance Liquid Chromatography with Tandem Mass Spectrometry: Possible Clinical Implication. *Anal. Chem.* **88,** 12128–12136 (2016).
129. Hong, H. & Wang, Y. Derivatization with Girard Reagent T Combined with LC-MS/MS for the Sensitive Detection of 5-Formyl-2′-deoxyuridine in Cellular DNA. *Anal. Chem.* **79,** 322–326 (2007).
130. Munzel, M. *et al.* Quantification of the sixth DNA base hydroxymethylcytosine in the brain. *Angew. Chem. Int. Ed. Engl.* **49,** 5375–5377 (2010).
131. Kraus, T. F. J. *et al.* Low values of 5-hydroxymethylcytosine (5hmC), the ‘sixth base,’ are associated with anaplasia in human brain tumors. *Int. J. cancer* **131,** 1577–1590 (2012).
132. Wagner, M. *et al.* Age-dependent levels of 5-methyl-, 5-hydroxymethyl-, and 5-formylcytosine in human and mouse brain tissues. *Angew. Chem. Int. Ed. Engl.* **54,** 12511–12514 (2015).
133. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **37,** 853–862 (2005).
134. Pastor, W. A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473,** 394–397 (2011).
135. Song, C.-X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.* **29,** 68–72 (2011).
136. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse

ES cells and during differentiation. *Nature* **473,** 398–402 (2011).

137. Robertson, A. B., Dahl, J. A., Ougland, R. & Klungland, A. Pull-down of 5-hydroxymethylcytosine DNA using JBP1-coated magnetic beads. *Nat. Protoc.* **7,** 340–350 (2012).
138. Sérandour, A. A. *et al.* Single-CpG resolution mapping of 5-hydroxymethylcytosine by chemical labeling and exonuclease digestion identifies evolutionarily unconserved CpGs as TET targets. *Genome Biol.* **17,** 56 (2016).
139. Sun, Z. *et al.* A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. *Mol Cell.* **57,** (2015).
140. Raiber, E. A. *et al.* Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol.* **13,** (2012).
141. Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.* **32,** 1231–1240 (2014).
142. Zhu, C. *et al.* Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell Stem Cell* **20,** 720–731.e5 (2017).
143. Lu, X. *et al.* Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Res.* **25,** 386–389 (2015).
144. Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.* **32,** 1231–1240 (2014).

Acknowledgements

The Balasubramanian laboratory is supported by core funding from Cancer Research UK (C14303/A17197). S.B. is a Senior Investigator of the Wellcome Trust (grant no. 099232/z/12/z).

Author contribution statement

E-A. R., R. H., P. VD. and S.B. wrote the article and reviewed and/or edited the manuscript before submission.

Competing interests statement

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Transcription factors: A protein that binds to a specific DNA sequence and thus controls the transcription of the genetic information from DNA to RNA.

Restriction methylation: In bacteria or other prokaryotic systems, methylation to the DNA protects from restriction endonuclease enzymes providing a defense mechanism against invasions from bacteriophages or virus.

Restriction endonucleases: Enzymes that cut DNA at endogenous phosphodiester bonds

Promoters: Region of DNA that is located near to the transcription start site and controls transcription initiation

Retrotransposons: Genetic elements that are transcribed into RNA, then reverse transcribed back into DNA and inserted into the genome.

Genomic imprinting: Epigenetic marking of one copy of the gene (from mother or father) that ensures gene expression in a parent-of-origin specific manner.

Transposon silencing: Gene silencing of transposons by epigenetic mechanisms, including DNA methylation and small non-coding RNA, prevents transcription and ensures genome stability.

Histone modifications: Post-translational chemical modification to amino acid residues on histone.

Chromatin-remodeling proteins: Proteins that control access to the genetic information by either affecting histone modifications or using energy to alter histone-DNA interactions.

CpG islands: A region with high CpG dinucleotide density

Pluripotent stem cells: Cell that is able to differentiate into any other tissue of the body.

Enhancer regions: Regulatory regions of the genome that are marked by histone modifications and enhance the transcription of their associated genes when bound to transcription factors

Base excision repair: Cellular mechanism that removes small base lesions from the DNA caused by mismatched or modified DNA bases.

Histone octamer: Consist of histone H2A, H2B, H3, H4 dimers that together form the core of the nucleosome

B cells: Type of white blood cell that is fundamental to the adaptive immune system

Class-switch recombination: A process whereby B cells rearrange parts of the immunoglobin heavy chain locus to generate antibodies with different properties

Telomeric regions: Repetitive nucleotide sequences that protect the ends of chromosomes

Repetitive elements: Repeat sequences that occur multiple times throughout the genome

RNA polymerase II: Polymerase that catalyses the transcription of DNA to RNA

Transfer RNA: An adapter RNA and amino acid carrier that helps decode messenger RNA for translation into the synthesis of proteins

β-elimination: DNA cleavage at the phosphodiester bond resulting in the elimination of the 3'-phosphate residue

δ-elimination: DNA cleavage at the phosphodiester bond resulting in the elimination of the 5'-phosphate residue

Clustered regularly interspaced short palindromic repeats (CRISPR) systems: Repetitive base sequence that form the basis of the genome editing system known as CRISP/Cas9

Transcription-activator like effectors (TALEs): Proteins that can be programmed to target specific DNA sequences in the genome.

TOC

Mapping and elucidating the function of modified bases in DNA

Eun-Ang Raiber, Robyn Hardisty, Pieter van Delft and Shankar Balasubramanian

Research into naturally occurring chemically modified DNA bases has been invigorated by new chemical and enzymatic methods that, when coupled with sequencing approaches, enable us to detect and decode them. These techniques will enable a better understanding of the role of chemically modified DNA bases in normal physiology and disease.

Subject categories

Physical sciences / Chemistry / Chemical biology / Nucleic acids
[URI /639/638/92/610]
Physical sciences / Chemistry / Biochemistry / DNA
[URI /639/638/45/147]

1) The first report on the double helical structure of DNA and the pairing of G-C and A-T bases.

13. Deaton, A. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).

This is a comprehensive review explaining the concept and importance of CGIs as regulatory features of the genome.

23. Pfaffeneder, T. *et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew Chem Int Ed Engl.* **50,** (2011)

This work provides the first evidence of the existence of 5fC in embryonic stem cell.

29) Work describing the timing of formation and metabolism of 5mC and 5hmC in genomic DNA using LC-MS/MS and stable isotopes

46) Work describing the use of stable isotopes and LC-MS/MS to elucidate the mechanism for 5-hmU formation in DNA

62) The discovery, quantitation and mapping of *N*6-methyladenosine in the model animal *C. elegans*

63) The discovery, quantitation and mapping of *N*6-methyladenosine in the model animal *D. melanogaster*

76. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-

hydroxymethylcytosine at single-base resolution. *Science* **336,** 934–937 (2012)

This article describes the development of a sequencing method that enabled the study of the dynamics and sequence context of 5hmC in mouse embryonic stem cell at single-base resolution for the first time

82. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Meth* **11,** 817–820 (2014).

This work describes the adaptation of the bisulfite sequencing method to single cells revealing 5mC heterogeneity within cell populations.

121) The first report on the formation of 5hmC in mammalian genomes from 5mC by the TET family of enzymes

140. Raiber, E. A. *et al.* Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol.* **13,** (2012).

This work provides the first genome-wide map of 5fC in embryonic stem cell demonstrating that the 5fC pattern was TDG-dependent.