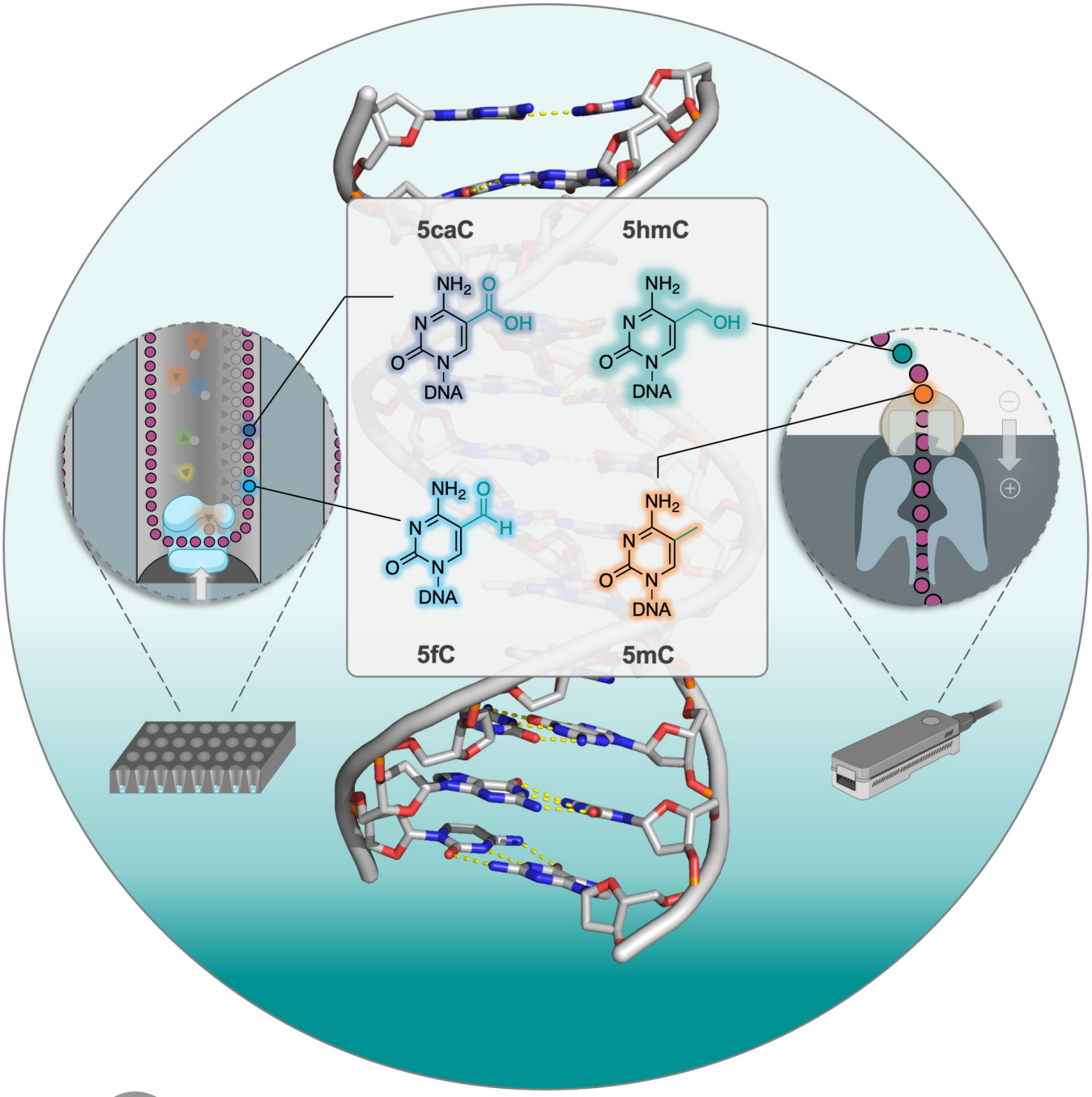
Check for updates

Nucleic Acids

How to cite: *Angew. Chem. Int. Ed.* **2023**, *62*, e202215704
International Edition: doi.org/10.1002/anie.202215704
German Edition: doi.org/10.1002/ange.202215704

Third-Generation Sequencing of Epigenetic DNA

Bethany Searle, Markus Müller, Thomas Carell, and Andrew Kellett**



Angewandte International Edition Chemie

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

Chinese U of HK EX2050
Pacific Biosciences v. Chinese U of HK
IPR2024-00028

Abstract: The discovery of epigenetic bases has revolutionised the understanding of disease and development. Among the most studied epigenetic marks are cytosines covalently modified at the 5 position. In order to gain insight into their biological significance, the ability to determine their spatiotemporal distribution within the genome is essential. Techniques for sequencing on "next-generation" platforms often involve harsh chemical treatments leading to sample degradation. Third-generation sequencing promises to further revolutionise the field by providing long reads, enabling coverage of highly repetitive regions of the genome or structural variants considered unmappable by next generation sequencing technology. While the ability of third-generation platforms to directly detect epigenetic modifications is continuously improving, at present chemical or enzymatic derivatisation presents the most convenient means of enhancing reliability. This Review presents techniques available for the detection of cytosine modifications on third-generation platforms.

1. Introduction

Astonishing progress has been made in DNA sequencing over recent years with significant innovation driven by the Human Genome Project. Consequently, the time and cost associated with sequencing the human genome has decreased dramatically, with the rate in decrease of cost surpassing Moore's Law.[1] Despite this remarkable progress, the four-letter genetic alphabet does not provide the complete mechanism governing gene regulation. The discovery of epigenetic bases has shed further light on disease and development;[2] however, their effective sequencing has presented new challenges. Most epigenetic sequencing methodology has been developed for use in conjunction with short-read "next-generation" sequencing (NGS) platforms; however, the advent of third-generation sequencing and its potential for greater accuracy, direct access and convenience has the potential to expedite research surrounding this secondary information layer in DNA. To reach their full potential in clinical and field research settings, it is vital that the complete, extended genetic alphabet can be read and interpreted by these sequencing platforms.

[*] B. Searle, Prof. Dr. A. Kellett
SSPC, the SFI Research Centre for Pharmaceuticals
School of Chemical Sciences, Dublin City University
Glasnevin, Dublin 9, Dublin (Ireland)
E-mail: andrew.kellett@dcu.ie

Dr. M. Müller, Prof. Dr. T. Carell
Department of Chemistry
Ludwig-Maximilians Universität München
Butenandtstr. 5–13, 81377 Munich (Germany)
E-mail: thomas.carell@lmu.de

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution Non-Commercial NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

2. Discovery of Epigenetic Bases

2.1. Discovery, Origin and Context of Methylation

The existence of methylated cytosine (5mC) was posited in 1925,[3] when it was observed as a product of the hydrolysis of tuberculinic acid, a non-canonical nucleic acid isolated from *Mycobacterium tuberculosis.* Its presence in eukaryotic DNA was confirmed in 1948[4] by paper chromatography of calf thymus DNA and subsequently 5mC has become the most widely studied epigenetic mark. Methylation at the 5-position is catalysed by a family of DNA methyl-transferases (DNMTs), which use *S*-adenosyl methionine as a methyl donor. DNMTs may be divided into two categories:[5] de novo DNMTs which methylate previously unmodified DNA and maintenance DNMTs that act during DNA replication to conserve the methylation patterns of the parent DNA strands. In eukaryotes, methylation of cytosine occurs almost exclusively within the context of CpG dinucleotides, with the human genome containing approximately 28 million of these sites.[6] Less than 10 % of these CpG dinucleotides are clustered together in regions termed CpG islands (CGIs), and their variable methylation status is known to play a key role in gene expression[7] (Figure 1 B) thereby regulating both normal development and disease progression.[8] Genome-wide analysis has linked aberrant methylation to numerous cancers;[9] however, our current ability to identify specific changes in methylation status down to the single-nucleotide resolution has limited the understanding of the role cytosine modifications in tumour progression. Despite this, strategies for cancer therapies targeting known epigenetic functions, including the activity of DNMTs,[10] have proved successful and this highlights the importance in understanding the role of these modifications to aid therapeutic drug design.

2.2. Oxidised Derivatives

Although the most abundant, methylation is not the only covalent modification to occur at the 5-position of cytosine (Figure 1 A). Despite being discovered in bacteriophages in 1952,[11] it was not until 2009 that 5-hydroxymethylcytosine (5hmC) was found to be abundant in the human brain and Purkinje neurons.[12–15] This discovery of 5hmC was precipitated by a search for homologs of the J binding proteins that

 © 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

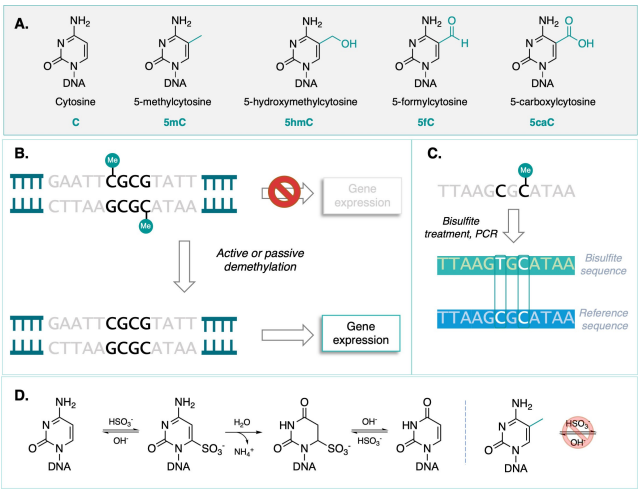


Figure 1. A) Molecular structures of epigenetic DNA bases. B) Effect of methylation on gene expression. C) Interpretation of bisulfite sequencing data for methylation detection. D) Mechanism of bisulfite induced deamination of cytosine.

catalyse the hydroxylation of base J. Base J is generated by the sequential hydroxylation and glucosylation of the methyl group of thymine in kinetoplastids.[16] The search identified TET enzymes and subsequently the authors confirmed the presence of 5hmC in mouse embryonic stem cells. TET enzymes, analogous to J binding proteins, are iron(II)/α-ketoglutarate dependent dioxygenases (Figure 2B), the mechanism of action for which has been extensively studied.[17] Subsequently in 2011, the Carell group identified 5-formylcytosine in embryonic stem cell DNA via HPLC-MS studies.[18] The final stage in the sequential oxidation of 5mC yields 5-carboxylcytosine (5caC), found in mouse genomic DNA in 2011.[19] Oxidation products 5caC and 5fC are recognised by thymine DNA glycosylase (TDG),[20] a DNA repair enzyme that subsequently excises the oxidised bases during base excision repair (BER), constituting an active demethylation pathway. Passive DNA methylation occurs through routine DNA replication, resulting in the dilution and elimination of epigenetic bases.

The first oxidation product, 5hmC, aids in this passive demethylation process by inhibiting methylation by maintenance DNMT1.[21]

3. Generations of Sequencing Technology

3.1. Sanger Sequencing

The first widely used DNA sequencing technique, Sanger sequencing, was developed in 1977[22] and commercialised versions of this technology drove the Human Genome Project.[23] In modern Sanger sequencing, chain-terminating nucleotides lacking a 3′-OH group, dideoxynucleotides (ddNTPs), are coupled to fluorescent labels and incorporated into the sequencing reaction alongside standard deoxynucleotide triphosphates (dNTPs). The fluorescent labelling was a major improvement over the radioactive ddNTPs used by Sanger in the original method.[24] The absence of the 3′-OH group in the ddNTPs prevents the

Beth Searle obtained a Masters degree in Chemistry from the University of St Andrews (UK) in 2020. She is currently an Early Stage Researcher in the Marie Skłodowska-Curie ITN NATURE-ETN under the supervision of Prof. Andrew Kellett. Her work focusses on the development of biomimetic catalysts for the modification of epigenetic bases in DNA.

Andrew Kellett studied chemistry at Maynooth University and received his Ph.D. in 2007 from Technological University Dublin for resarch in bioinorganic chemistry. He is now Associate Professor of Inorganic and Medicinal Chemistry in the School of Chemical Sciences at Dublin City University (Ireland). His group work at the intersection of bioinorganic and nucleic acids research fields. He was coordinator of the Marie Skłodowska-Curie Innovative Training Network (ITN), ClickGene, and the ongoing NATURE-ETN Marie Skłodowska-Curie ITN.

Markus Müller studied Biology at the Philipps Universität Marburg (Germany) and received his Ph.D. in plant biology 2003 under the supervision of Alfred Batschauer. After a postdoc at Umeå Plant Science Center (Sweden) with Marianne Sommarin, he joined Thomas Carell's group in 2006 and has since then worked as Akademischer Rat at Ludwig-Maximilians-Universität München. His research focus is on RNA modification and sequencing methodology.

Thomas Carell studied chemistry at in Münster and Heidelberg (Germany) and received his Ph.D. in 1993 with work at the Max Planck Institute for Medical Research.. After a postdoctoral stay at the MIT in Cambridge (USA), he started independent research at the ETH Zurich (Switzerland). In 2000 he accepted a chair for Organic Chemistry at the University of Marburg (Germany). In 2004 he moved the Ludwig-Maximilians-Universität München (Germany). His research interests center around nucleic acid chemistry and chemical biology.

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

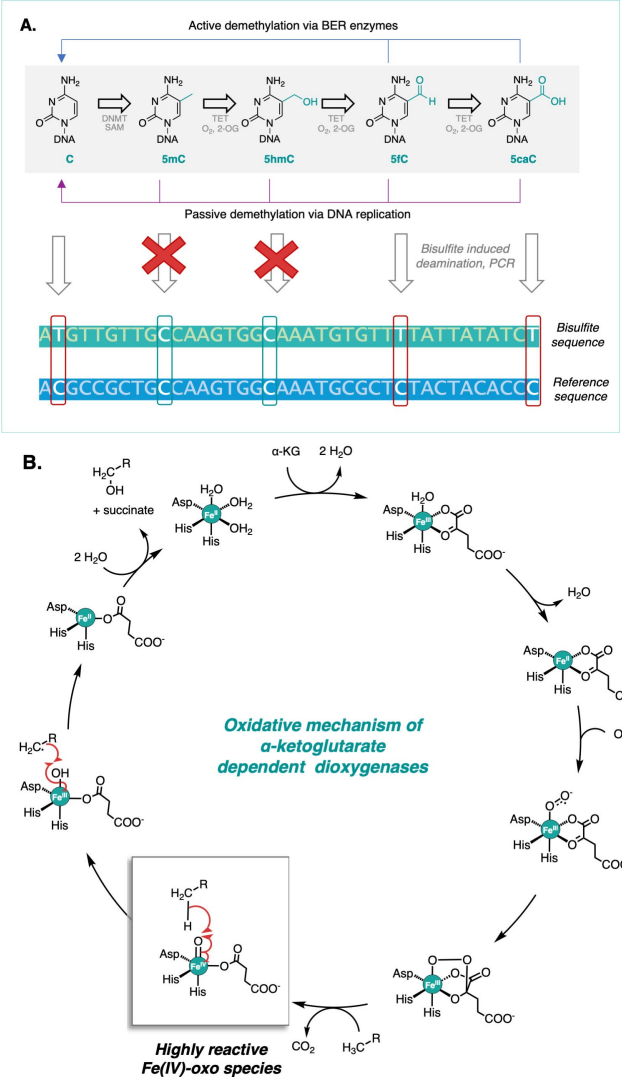


Figure 2. A) Passive vs. active demethylation pathways via sequential oxidation of 5mC by TET enzymes and the effect of these modifications on the bisulfite sequencing protocol. B) Catalytic cycle of α-ketoglutarate dependent dioxygenases, the class of enzymes to which the TET family belong. The Fe^{IV}-oxo species at the TET active site has been recreated synthetically as TET biomimetic shown.

formation of the phosphodiester bond by the polymerase, resulting in cessation of strand elongation upon incorporation. The resulting DNA fragments may then be sorted according to their lengths with single-base precision via electrophoretic methodology. While this technology was the first to enable the sequencing of entire genomes, thus revolutionising the field, its lack of throughput resulted in high costs and slow progress, motivating further innovation.

3.2. Next-Generation Sequencing

The Human Genome Project highlighted the need for improved sequencing technology, precipitating the development of a number of "next-generation" sequencing (NGS) platforms. Although platforms differ in their specification, NGS techniques share the ability to perform massively parallel sequencing reactions, dramatically decreasing associated time and cost relative to Sanger sequencing. NGS involves the amplification of fragmented DNA through PCR, followed by spatial separation and recombination of sequencing reads through mapping to a reference genome. The speed and scalability offered by NGS platforms revolutionised sequencing and genome research, widening the accessibility of sequencing technology and driving discovery and innovation within genetics research. Next-generation sequencing techniques have been reviewed.[25–27]

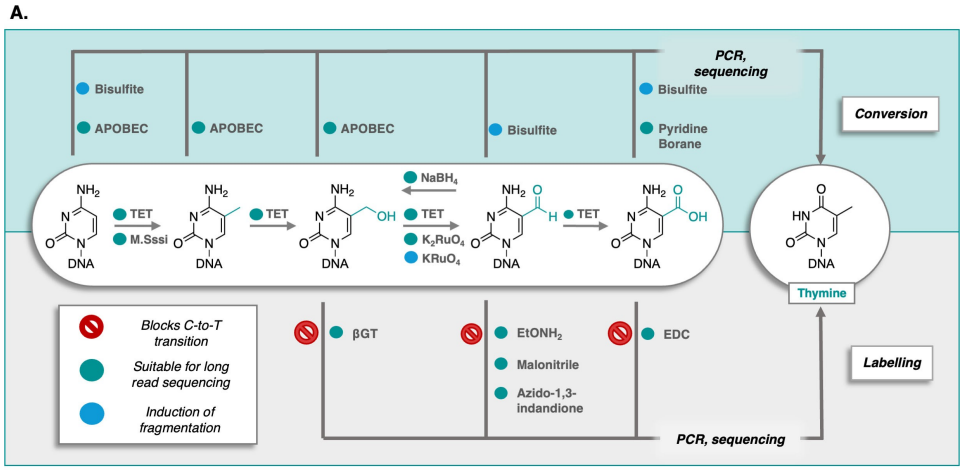
3.3. Epigenetic Base Detection via NGS

Epigenetic marks are not conserved during the PCR-based DNA amplification characteristic of NGS, necessitating pretreatment to convert this secondary information layer into genetic information. This is primarily achieved through endonuclease digestion, affinity enrichment or bisulfite conversion.[28] Currently, bisulfite sequencing is considered to be the gold standard for single base resolution epigenetic sequencing. Developed in 1992,[29] prior to the discovery of the oxidised derivatives, bisulfite sequencing enables the differentiation of 5mC and C by exploiting their differential reactivity with sodium bisulfite where the methyl group is protective against a deamination process (Figure 1 D). As a result, subsequent amplification yields T in place of C for unmodified C, while 5mC is maintained as C, enabling the positive identification of methylated positions (Figure 1 C). Bisulfite sequencing is unable to distinguish 5hmC from 5mC, or C from 5fC and 5caC (Figure 2 A). This conflation of 5mC with 5hmC has likely led to a number of false assumptions, given that both marks can have partially opposing roles in gene regulation.[30] Numerous modifications have been made to the protocol to exploit the varying PCR activity of deamination products or those of labelling reactions, most notably oxidative bisulfite sequencing[31] or oxBS-seq pioneered by Shankar Balasubramanian, which enables profiling of the two most abundant modifications 5mC and 5hmC. Further developments include TAB-seq[32] and oxBS-seq[31] for 5hmC detection, fCAB-seq[33] and redBS-seq[34] for 5fC detection and CAB-seq for 5caC detection[35] as summarised in Figure 3.

This methodology has provided useful insights into these epigenetic marks; however, bisulfite sequencing suffers from inherent limitations; the harsh conditions lead to sample degradation through the formation of abasic sites (>1/200 bases after treatment)[36] which promote strand scission, limiting read length. Sample degradation proves especially challenging in highly repetitive regions of the genome where alignment to a reference is impeded by these short read lengths. This is exacerbated by the loss information induced during bisulfite treatment as C, 5fC and 5caC become T, reducing the complexity of the sequences. In addition to the destructive nature of the treatment, whole-genome bisulfite sequencing (WGBS) is known to overrepresent methylation due to numerous biases introduced during library prepara-

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



B.

Base	Untreated	BS-seq 5mC + 5hmC	oxBS-seq 5mC/5hmC	TAB-seq 5hmC	redBS-seq 5fC	fCAB-seq 5fC	CAB-seq 5caC	MAB-seq 5fC + 5caC
C	C	T	T	T	T	T	T	C
5mC	C	C	C	T	C	C	C	C
5hmC	C	C	T	C	C	C	C	C
5fC	C	T	T	T	C	C	T	T
5caC	C	T	T	T	T	T	C	T

Figure 3. A) Summary of techniques used in library generation for distinction of cytosine and its oxidised derivatives. B) Methodologies developed for epigenetic sequencing and their readouts

tion and sequencing.[37,38] Despite the limitations, the utility of modified sequencing methodology for epigenetic sequencing on NGS is evident and has provided great insights into the methylation profiles of a diverse range of genomes.[39,40] Notably, a number of bisulfite-free methodologies have been developed to improve coverage and eliminate associated biases, with the associated transformations summarised in Figure 3. These techniques, including CLEVER-seq,[41] TAPs,[42] EM-seq[43] and fC-CET[44] use non-destructive chemical or enzymatic treatments generating read inaccessible via bisulfite-based protocols. As such, this makes them well suited for adaptation for use in conjunction with third-generation sequencing platforms.

4. Third-Generation Technologies

4.1. Concepts

Third-generation sequencing (TGS) technologies are characterised by the direct sequencing of single molecules, overcoming the limitations imposed by short read lengths accessible using NGS. By circumventing the need to heavilty fragment DNA, this most recent generation of sequencing technology generates reads multi-kilobases in length which are highly desirable for repetitive or highly variable regions of the genome that cannot be mapped by shorter read lengths, highlighted by the application of TGS in the most recent and highly detailed sequencing of the human genome.[45] Further, mitigating the need for fragmentation and replication reduces the number of stages at which errors or bias are introduced. To date, epigenetic sequencing has been conducted using nanopore technology and single-molecule real-time sequencing, available commercially as Oxford Nanopore and PacBio devices, respectively (Figure 4 A and B).

4.2. Nanopore Sequencing

Conceptualised in the early 1980s by David Deamer[46] and realised by Hagan Bayley through the application of engineered nanopores,[46,48] the first commercially available device using nanopore technology launched in 2015 through Oxford Nanopore MinION.[49] Upon the application of a voltage, single-stranded DNA (ssDNA) passes across a lipid membrane via an embedded protein nanopore. Helicase enzymes impede translocation of ssDNA, facilitating data capture (Figure 4 A). Fluctuation of the ionic current is a function of the DNA sequence, interpreted with single-base

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

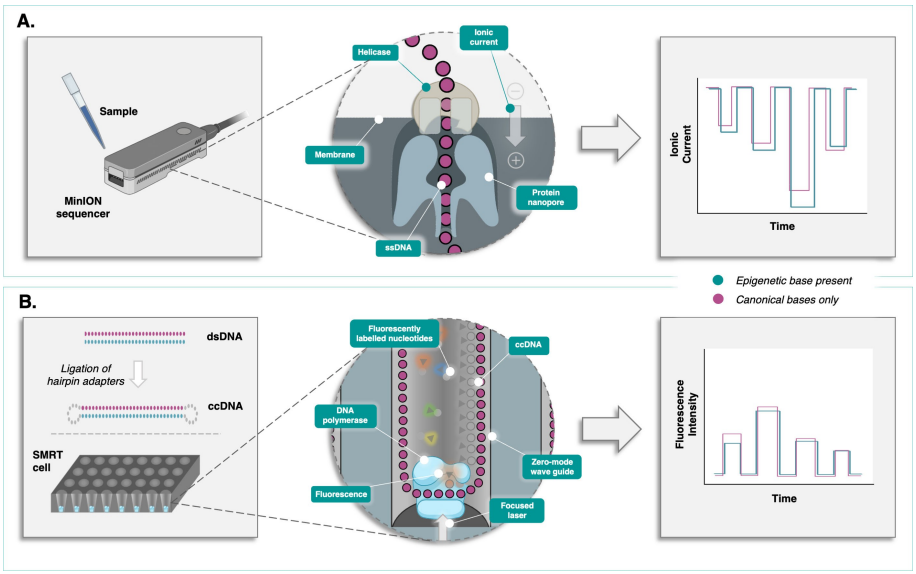


Figure 4. A) Sequencing on the Oxford Nanopore MinION device. Protein nanopores coupled to helicase enzymes are embedded in a membrane. Application of a voltage drives an ionic current across the membrane and through the nanopore. Translocation of ssDNA modulates the ionic current and fluctuations are interpreted algorithmically to generate a sequence. B) SMRT sequencing utilising single polymerases in zero-mode waveguides. Monitoring of fluorescence upon incorporation of labelled nucleotides generates signals with characteristic durations and widths.

precision by "base calling" algorithms. Bases passing through the nanopore exert a unique effect on the ionic current, thereby theoretically enabling the differentiation of all bases and their modifications. As several bases will be present in the nanopore at any time (k-mers), it is the k-mer that causes fluctuations in signal, necessitating the algorithmic treatment of the raw data to deconvolute these fluctuations, generating sequence data with single-base resolution.

The accuracy of nanopore sequencing is therefore dependent upon the nanopore, processive enzyme and bioinformatics methodology.[46] Early nanopore technology suffered from significant error rates; however, optimisation of these components have begun to close the gap between the accuracies of next-generation and nanopore sequencing. This was conceptually demonstrated when nanopore sequencing was shown to successfully distinguish between all five cytosine derivatives,[50] with accuracy ranging from 92–98 % when the bases were inserted into a template strand. The uniformity of the flanking bases present in the template strand inevitably led to overestimations of the accuracy; however, this serves to highlight the potential of nanopore technology in epigenetic sequencing.

The refinement of base calling algorithms is an active area of research[51] and methodology has been specifically developed for nanopore sequencing of epigenetic modifications. A deep learning method, DeepSignal,[52] was developed for genome-wide methylation calling. 90 % accuracy was achieved using 2× coverage of reads, with accuracy comparable to bisulfite sequencing achieved with 20× coverage. This provides greater accuracy with lower coverage compared to previously developed statistical models, however, remains below methodology developed for NGS protocols. Notably DeepSignal was able to identify the methylation status of 5 % more CpGs than bisulfite sequencing, owing to the inherently less destructive nature of Nanopore sequencing.

Simultaneously, the physical refinement of the technology through the development of increasingly sensitive nanopores provides another route to decrease errors. Biological nanopores present opportunities for refinement via coupling to processive enzymes and protein engineering. A recent iteration of the protein nanopore features a dual constriction site, demonstrated to improve accuracy by 25–70 % for homonucleotide sequences up to 9 bases long; these are regions known to generate errors during nanopore sequencing.[53] Improved sensitivity of the channels through which the bases pass naturally allows for increasingly subtle modifications to be detected.

4.3. SMRT Sequencing

Single-molecule real-time (SMRT) technology exploits the real-time detection of fluorescent dNTPs incorporated by a DNA polymerase into a complementary strand in a sequencing-by-synthesis approach. (Figure 4B). Polymerases are embedded into zero-mode waveguides (ZMWs) which are wells on a silicon chip. These ZMWs are narrower than the wavelength of light emitted by the excitation laser, allowing the excitation of fluorescence only of nucleotides that are actively being incorporated (Figure 3). The ability to monitor this process at single-base resolution allows the primary DNA sequence to be determined by monitoring fluorescence emissions; however, data generated from the incorporation dynamics, including pulse width and interpulse duration (IPD), provides insight into the presence of base modifications.[54]

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

As with Nanopore technology, early error rates in base calling proved to be significant for SMRT sequencing, with an overall error rate of up to 13 % reported with no single reads being error free.[55] While single-read error rates have remained consistent, improvements in sequencing chemistry and library preparation, including the ligation of hairpin adapters to double dsDNA to form circular templates, have reduced errors by enabling greater sequencing depth.[56] The circular template allows the polymerase to sequence the same fragment multiple times in a process termed "circular consensus sequencing" (CCS). Additionally, this protocol allows the identification of hemi and symmetrically modified positions. However, SMRT sequencing is reliant upon the quality of the polymerase; the depth of sequencing of DNA is limited by its longevity.[57] Likewise, longevity determines maximum read length. Sequencing of longer sequences at higher depths requires a robust polymerase.

Modifications to the bases of the template strand affect polymerase kinetics, and these unique kinetic signatures can theoretically be used to identify epigenetic modifications. Since the 4 position of cytosine and the 6 position of adenine are involved in complementary hydrogen bonding, it can be expected that modifications here will directly impact polymerase kinetics. In contrast, modifications at the 5 position of cytosine are positioned in the major grove and thus have little contact with polymerase enzymes, resulting a in less significant modulation of incorporation dynamics.[58] However, when located within an identical sequence context, characteristic variations in these polymerase kinetics have allowed distinction between C, 5hmC and 5mC.[59] An investigation into the kinetic signatures of all oxidised modifications was conducted as part of an effort to overcome the limitations of direct SMRT sequencing of methylated cytosine by oxidation of 5mC containing DNA with TET.[60] A template with no modifications was sequenced and from this the ratios of IPDs at each position featuring the modifications can be calculated. The size of the modification was shown to elicit a proportional interference with the polymerase, with 5caC generating the greatest response; however, differences are small and context dependent—de novo identification of the full suite of modifications has proven challenging. Recently, a neural network was used to identify 5mC in native DNA, which significantly improved detection generating 99 % agreement with bisulfite sequencing data.[61] Uniquely, the neural network is able to simultaneously account for multiple variables of polymerase kinetics, including IPDs, pulse width and sequence context simultaneously, leading to greater accuracy. At present, this does not extend beyond the direct detection of 5mC and 5hmC, analogous to bisulfite sequencing. The detection of further oxidation of cytosine residues has yet to be achieved.

Inaccuracies in detecting epigenetic bases in native DNA remain a barrier for the application of third-generation platforms to epigenetic research. Despite this, portable and inexpensive sequencing devices are highly desirable, irrespective of the innovation anticipated from long-reads, as demonstrated by their usage in the on-going SARS-CoV-2 pandemic.[62–65] To circumvent the currently inherent limitations in directly detecting base modifications, methodologies for epigenetic sequencing have been developed in conjunction with the devices, including the adaptation of existing techniques developed with NGS for TGS platforms. The capability of base conversion and modification protocols that exploit the differential reactivity of modified cytosine bases alongside the long-reads enabled by third-generation platforms promises increasingly accurate mapping of epigenetic variation across entire genomes.

4.4. Labelling Techniques for Epigenetic Detection

The chemistry of the oxidised cytosine derivatives makes them well-suited to selective labelling by both chemical and enzymatic means, as demonstrated by the numerous techniques developed with NGS. Labelling techniques present a facile method for increasing the bulk of a given base, enhancing the associated kinetic signatures and making it a promising avenue for epigenetic detection on third-generation platforms (Figure 5).

The first example of derivatisation for TGS aimed to enhance kinetic signatures of oxidised cytosines for SMRT sequencing. Sequential glucosylation by T-even bacteriophage enzymes, generating a diglucosylated adduct of 5hmC, enabled the simultaneous mapping of all three oxidised derivatives of 5mC.[66] The authors demonstrate that while 5fC and 5caC produce sufficiently distinct modulation in polymerase dynamics without modification, 5hmC requires additional modification for detection—sequencing of a template featuring 2 5hmC sites showed the IPD ratio of 5hmC to C was only 2. Diglucosylation enhanced this ratio to 29; however, the effect was shown to be sequence-dependent for all modifications. IPD modulations were observed up to 6 bases from 5fC, with multiple 5mCs also shown to have a cumulative effect. Sequencing was performed on ≈6 kb fragments of native DNA with ≈120× coverage and validated using 5-methylenesulfonate (CMS) immunoprecipitation, in which 5hmC is treated with bisulfite to yield CMS. Strong correlations were found between techniques; however, the authors concede that due to the sequence-dependence of the kinetic effects, some modified sites are likely to be missed.

This methodology was applied to profile the oxidised 5mC species present in the genome of fungal *Coprinopsis cinerea.* The data indicates that oxidation of 5mC at paralogous gene arrays and repetitive elements limits gene expression, differing from previous studies that found high levels of 5hmC at the gene bodies of highly expressed genes in mammalian cells, suggesting oxidation pathways in *Coprinopsis cinerea* and mammalian cells play different roles. While the methodology is unable to distinguish 5hmC from 5fC and 5caC when only treated DNA is present, an untreated control would enable this differentiation. Despite this, the utility of mapping exclusively oxidised derivatives provides valuable insight into the function of the oxidative process, as demonstrated by the data generated in this study.

Using a bisulfite-mediate thiol substitution, coupling to peptides, fluorescein and biotin can be achieved and

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

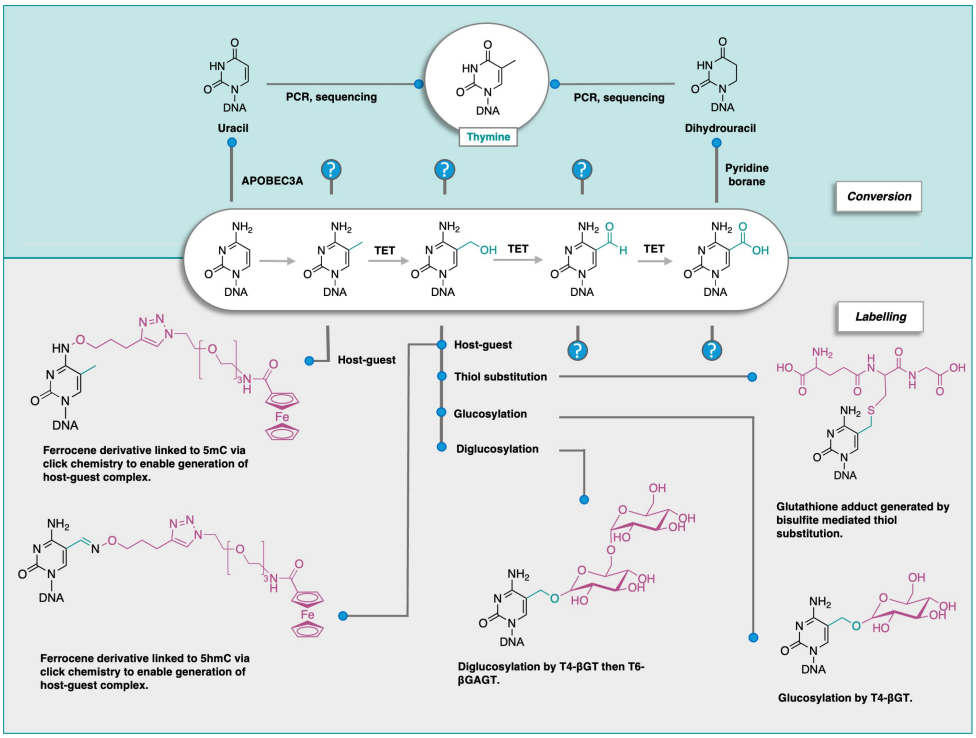


Figure 5. Species generated during conversion and labelling methodology used with third-generation sequencing platforms.

biotinylation of 5hmC-containing ssDNA was demonstrated to be detectable using an α-hemolysin protein nanopore.[67] Treatment with sodium bisulfite and a nucleophilic thiolate can be used to access a number of 5-thiomethyl derivatives. When used at a concentration of 0.05–0.06 M at 42 °C, cytosine was shown to be unaffected in contrast to the deamination reaction that occurs during bisulfite sequencing at 4 M and 72 °C, thereby maintaining genetic complexity and eliminating the fragmentation associated with the harsh conditions used in BS-seq. Conversion of 5hmC-containing DNA to yield a biotinylated adduct over two steps was 30–65 % and a single-step approach yielded a 35–55 % conversion. As well as enabling direct sequencing, biotinylation enables enrichment of DNA containing 5hmC via immobilised streptavidin, facilitating increased sequencing depth. An α-hemolysin nanopore was used to detect a 40-mer ssDNA containing a 5hmC site. Sequencing data shows that when the site was modified to generate a glutathione-ssDNA conjugate via bisulfite treatment, modulation of the current amplitude was indicative of the presence of the modified 5hmC strand. In control experiments in which the modified 5hmC site was replaced with an unmodified 5hmC in addition to G, A T and C, the histograms of the electrical recordings showed no significant deviation therefore rendering them indistinguishable. Incomplete conversion rates limit the applicability of this labelling strategy for sequencing given the low abundance of 5hmC within the genome. Despite this, this early example demonstrates the utility of a labelling technique for single-molecule epigenetic sequencing and the authors show the potential of exploiting the reactivity of bisulfite using milder conditions to generate a range of useful modified 5hmC species. They also note the potential of labelling techniques to extend to higher-oxidise derivatives. The use of chemical labelling avoids the selectivity issues associated with enzymatic labelling, while being less expensive and allowing for labelling of ssDNA, in addition to maintaining genetic complexity by acting only upon epigenetic positions.

Later, a method for profiling 5mC and 5hmC via the generation of a host–guest complex was developed in conjunction with nanopore sequencing.[68] Chemical labelling generates structures too sterically demanding to pass through the nanopore, causing the detachment of a non-covalently bound element. This generates highly characteristic fluctuations in ionic current. Modification of 5mC was achieved via treatment with bisulfite followed by condensation with aminooxy-alkyne *O*-(pent-4-yn-1-yl)hydroxylamine to generate an alkyne-modified base, which can be attached to a derivative of ferrocene or adamantyl via click chemistry and non-covalently coupled to curcubit[7]-uril. Following this protocol, 5mC and 5hmC were indistinguishable and a further labelling reaction was developed in which 5hmC was first oxidised to 5fC using $KRuO_4$ before reaction with the aminooxy-alkyne at a lower temperature. This reaction is highly selective for 5hmC, leaving 5mC and C intact. Notably, this method successfully identified 5mC and 5hmC loci independent of sequence context; however, authors note current signature generated by a doubly methylated strand did not vary significantly from DNA containing only one methylation site, thus the method may not be applicable for quantification of densely methylated regions. The authors show that selective labelling strategies can be employed to

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

significantly disrupt translocation kinetics, thereby overcoming issues regarding sensitivity in detection of modified bases. Although this methodology does not currently facilitate base-resolution sequencing, it provides a facile way of screening of 5mC- and 5hmC-modified DNA using TGS without the need for very high sequencing depths.

4.5. Milder Chemistry for Long-Read Epigenetic Sequencing

While labelling techniques are useful for identifying DNA containing the selected modification, a reliable and highly sensitive method that enables identification of epigenetic bases at single-base resolution has not yet emerged. The conversion of epigenetic signals into coding that can be read by sequencing technology, analogous to the C-to-T transition observed in bisulfite sequencing, is highly desirable. However, bisulfite conversion and oxidative treatment lead to a substantial amount of fragmentation. Such fragmentation may be less consequential for short-read workflows as they rely on DNA fragmentation. To harvest the full potential of TGS methodology, gentler chemistry to avoid this degradation is needed. Two techniques are currently available for base conversion (Figure 5).

EMseq (enzymatic methyl sequencing)[69] is the first exclusively enzymatic method for mapping 5mC and 5hmC. EMseq exploits the same C-to-T transition used in bisulfite sequencing but using milder conditions. During EMseq, 5mC and 5hmC are oxidised using TET-2 to 5caC, which is not a substrate for deamination by APOBEC3A.[70] Pre-existing or residual 5hmC generated from the oxidation of 5mC is protected from deamination through glucosylation using T4-βGT. As a result, deamination of DNA with ABOPEC3A affects only unmodified cytosines, resulting in cytosine reading as thymine after PCR. EMseq was shown to outperform bisulfite sequencing in the detection of CpG sites with equivalent input quantities of DNA, as a result the absence of the fragmentation bias introduced under bisulfite conditions. This protocol was developed using NGS Illumina platforms; however, to take advantage of the long reads facilitated by enzymatic discrimination of bases, the technique has been adapted for TGS. Additional epigenetic information can be obtained through an additional step to differentiate 5mC from 5hmC, with the modified techniques termed LR-EMseq (long-read enzymatic modification sequencing).[71] Using exclusively T4-βGT in the absence of TET2-catalysed oxidation selectively protects pre-existing 5hmC against deamination by APOBEC3A, resulting in only 5hmC being read as C upon PCR. LR-EMseq protocols involve the preparation of multi-kilobase DNA amplicons for long-range phasing by third-generation platforms. In this proof-of-concept for the application of EMseq for long-read sequencing, the authors showed that while the size of DNA fragments dropped from 15 kB to 0.8 kB upon bisulfite treatment, no observable degradation occurred following EMseq protocols. Sequencing the same amplicons after treatment across Illumina, Nanopore and SMRT platforms all yielded similar methylation profiles. Nanopore sequencing showed the highest incorrect calls across C, 5hmC and 5mC, as expected given the platform's inherently higher error rate. EMseq was shown to outperform bisulfite protocols adapted for low input amounts of DNA.[72]

EMseq was further adapted for long-read whole-genome sequencing, designated nanoEM, for combined 5hmC and 5mC identification.[73] Verification of the sequencing technology was conducted on two breast cancer cell lines and three clinical samples and compared with direct methylation calling on unamplified DNA using Nanopolish, a computational approach to direct methylation calling, WGBS and EMseq using short-read sequencing. NanoEM showed higher correlation in CpG methylation states with WGBS than Nanopolish ($R = 0.91$–0.87 for breast cancer cell lines) but correlation between NanoEM and Nanopolish was also high ($R = 0.89$–0.84). Given that all three methodologies are susceptible to biases, the generally high agreement is a good indicator of performance. While read length using Nanopolish was greater than for NanoEM, (17–32 kb compared to 3.4–7.6 kb), the quantity of input DNA for NanoEM at 1–100 ng is significantly lower than the 500 ng–100 μg required for direct sequencing using Nanopolish. This higher requirement for input DNA limits the applicability of direct methylation analysis for clinical samples; tumour cell enrichment of a clinical specimen by microdissection typically leads to DNA yields ranging from 50 to 300 ng. Surgically dissected samples however are not guaranteed to provide 1 μg of genomic DNA. For the analysis of the early stages of tumour development, this often renders direct methylation analysis impossible, thus a base-conversion method that is compatible with DNA amplification is desirable. In addition to enabling low input quantities, NanoEM was able to detect allele-specific methylation patterns that could not be resolved using short-read technologies and the methylation status of structural variations in the relevant genes was assigned; the inherently inconsistent sequences make aligning short reads of these regions challenging.

While clearly having great potential in long-read TGS, EMseq features still one major drawback in common with bisulfite sequencing. Due to the conversion of all unmodified cytosines to uracil, the sequence complexity decreases, generating mapping problem that only in part can be alleviated by the longer read-length. Conversion of only the modified positions in the genome with a mild conversion chemistry would thus be highly beneficial.

Recently, a method combining both enzymatic and chemical treatment of DNA for sequencing 5hmC and 5mC, TET-assisted pyridine borane sequencing (TAPS[74]) was adapted for long-read sequencing—lrTAPS.[42] The first step in sequencing is the TET-assisted oxidation of 5mC and 5hmC to 5caC, which are then reduced by pyridine borane to dihydrouracil (DHU). Upon PCR, DHU is recognised as thymine resulting in the 5mC/5hmC-to-T transition analogous to bisulfite sequencing; however, pyridine borane treatment does not affect the unmodified cytosines, resulting in a higher complexity of the resulting sequence. Both nanopore and SMRT platforms were used during development. For optimising TAPS methodology for long-read sequencing, the authors first developed a single-tube TAPS procedure, thereby minimising the loss of DNA fragments

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and the required quantity of input DNA. Verification of Nano-TAPS and SMRT-TAPS was conducted by sequencing a 4-kb model DNA methylated using HpaII enzymes, which methylate the internal cytosine in the CCGG sequence. Nano-TAPS and SMRT-TAPS showed good agreement with BS-seq data generated on the Illumina platform with Pearson correlation coefficients of 0.992 and 0.999, respectively. Non-amplified TAPS-treated DNA containing DHU was subjected to SMRT and Nanopore sequencing. SMRT sequencing was not possible due to the presence of DHU stalling the polymerase, while Nanopore sequencing using Nanopolish and Tombo produced correlation coefficients of only 0.650 and 0.808, respectively. Phage lambda DNA was used to assess the potential read lengths of lrTAPS. The longest amplicon generated, 10 kb, was sequenced with Nanopore and SMRT platforms with both showing good agreement with bisulfite sequencing data. Having confirmed the sensitivity and specificity of the methodology, the authors sought to sequence an amplicon of mouse embryonic stem cell (mESC) DNA including a 500-base-pair sequence featuring a gene previously considered unmappable. Outside of this gap, the NanoTAPS and SMRT-TAPS showed good agreement with Illumina-TAPS in assigning methylation status at CpG sites with a sequencing depth of greater than 8, providing confidence in the reading of the previously unmapped sequence. To further demonstrate the utility of lrTAPS, the authors applied the technique to the study of the variability of methylation status of the hepatitis B virus (HBV) across its life cycle. HBV replicates via a covalently closed circular DNA (cccDNA) and linearised HBV DNA can be generated and integrated into the host DNA. The authors showed using lrTAPS that in de novo infected engineered HepG2 cells, this cccDNA is unmethylated, consistent with active transcription. Upon integration within human hepatoma cells, CpGs within CGIs and gene bodies were found to be methylated. Notably, when individual reads were examined, distinct methylation events that were correlated or anti-correlated across long distance could be identified. Observing this heterogeneity is only possible with long-read sequencing technology.

TAPS has been applied for whole-genome long-read sequencing (wglrTAPS) using the SMRT platform. The authors used wglrTAPS for complete 5hmC and 5mC profiling of mESCs, alongside short-read TAPS, finding that of the nearly 21 million CpG sites in mESCs, wglrTAPS was able to identify over 19 million, compared to the 10.7 million found using short-read TAPS. Importantly, the authors note that CpG sites covered by wglrTAPS tended to be those located in repeat regions which cannot be resolved using short-read platforms. wglrTAPS showed a lower global level of methylation than short-read sequencing; however, this is attributed to this coverage of regions not accessible using short reads. For equivalent sequencing depth, wglrTAPS outperformed short-read TAPS in the detection of structural variants, with a number of deletion and insertion events covered only by wglrTAPS found in repetitive regions, deciphering two genetic elements that remain elusive using short-read technology. TGS promises to provide novel insights into the genetic features of disease and development notwithstanding its potential in epigenetics research.

The original TAPS methodology has been expanded and two sister methods, TAPSSβ, for exclusively 5mC sequencing, and CAPS for 5hmC specific sequencing, are available. The authors further demonstrate that exclusively pyridine borane treatment can be used for sequencing 5fC and 5caC. Future adaptation of these techniques for long-read platforms offers the potential for complete sequencing of 5mC and its oxidised derivatives. The novel insights gained through the application of long-read TAPS is a promising indication of the potential for TGS platforms to generate further discovery and innovation.

4.6. Emerging Chemistry to Facilitate Long-Read Sequencing

Under the correct conditions, chemical methodologies offer greater reliability than the use of enzymes which may suffer inherent biases, while being cheaper and more accessible. In TAB-seq,[32] (TET-assisted bisulfite sequencing), TET enzymes are used to oxidise 5mC to distinguish it from 5hmC; however, a TET biomimetic catalyst was recently developed featuring a high-valent iron centre found analogous to the TET active site[75] (Figure 6). This Fe^{IV}-oxo species was shown to effectively and selectively oxidise a 5hmC residue within 10mer oligonucleotide context.[76] Although yet to be applied in a sequencing context, the ability to perform this transformation synthetically will no doubt decrease the cost associated with the transformation, driving data collection.

As demonstrated by lrTAPS, chemical transformations can provide a convenient means of converting epigenetic bases into coding that can be readily interpreted using long-read TGS platforms. A number of approaches to epigenetic sequencing, including the most widely used BS-seq, are based on deamination reactions, generating species which through PCR are converted into genetic information which can be unambiguously interpreted by sequencing platforms. The utility and efficacy of biomimetic species is evident as demonstrated by the Fe^{IV}-oxo complex. The development of a catalyst capable of performing the deamination function of APOBEC enzymes in the absence of the harsh conditions used in bisulfite sequencing would further enable the potential of long-read sequencing to be realised. Clearly, biomimetic species can present a cheap and accessible way to manipulate epigenetic bases, increasing the ease of sequencing.

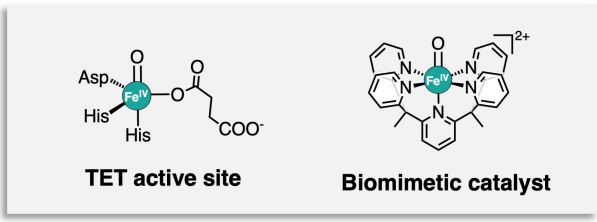


Figure 6. Active site of TET enzyme and inspired Fe^{IV}-oxo catalyst developed for cytosine oxidation.

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5. Alternative Single-Molecule Methodology

In addition to the widely available Oxford Nanopore and PacBio SMRT platforms, optical mapping provides a direct route to high-resolution, long-read sequencing data—BioNano Genomics technology uses nanochannels arrays to detect fluorescently labelled DNA.[77] By using a unique label for each base, both genetic and epigenetic features can be observed simultaneously. The technique was applied to generate the 5hmC profile of human peripheral blood mononuclear cells, with 5hmC quantification verified by LC–MS/MS and sequencing data compared with hMeDIP-seq data. Remarkably, optical mapping was able to provide epigenetic information regarding the human leukocyte antigen which is among the most heterogeneous regions in the human genome, spanning 3.6 Mb.[78] This polymorphism renders alignment to a reference genome very challenging; however, the authors demonstrate this can be overcome using ultralong reads. While hMeDIP-seq was not able to align any reads, the long optical reads could be aligned unambiguously, enabling identification of 5hmC around this region in the absence of amplification or targeting methodology. Given the gene's association with over 100 diseases,[79] generating its epigenetic profile would undoubtedly drive therapeutic and diagnostic innovation.

6. Summary and Outlook

The potential of long-read technology to decipher previously uncharted regions of the genome will continue to drive further discovery and innovation across the field. In conjunction with improved methods for epigenetic sequencing, unlocking the secondary information layer in DNA promises to present a host of opportunities in developing diagnostics and therapeutics, as well as understanding the fundamentals of disease and development. The continuous improvement in the sequencing science is compounded by the affordability and accessibility of TGS devices, making feasible their application in a clinical setting. DNA methylation has been identified as a cancer biomarker,[80] both in biopsied tumour samples and cell-free DNA (cfDNA). Liquid biopsy assays using cfDNA allow for facile monitoring of patients, in turn generating the potential to improve patient outcomes, from facilitating early diagnosis to monitoring for recurrence. Fluctuations in methylation may also inform the choice of therapy in addition aiding the assessment of a patient's response to treatment. Notably, 5hmC profiling in was recently used to identify genetics signatures in cfDNA associated with acute myeloid leukaemia (AML).[81] Here, the authors used data generated from nano-5hmC-Seal[82] on NGS platforms to develop diagnostic and prognostic models which were able to accurately categorise AML patients and controls while identifying genes and pathways associated with the disease dependent upon 5hmC prevalence. These examples of diagnostic and prognostic developments are indicative of the utility of identifying epigenetic signatures in informing clinical practice. TGS platforms have the potential to expedite this progress, bridging the gap between research and clinical applications to improve patient outcomes.

Despite their promise, a number of limitations need to be addressed before this technology reaches its full potential. While complete profiling of the epigenome remains unfeasible, sequencing approaches that exploit the differential reactivity of bases is proving an effective means of epigenetic profiling. Analogous to NGS epigenetic sequencing techniques, these suffer from their own biases and limitations. Ongoing work to improve these techniques will continue to improve their accuracy and efficiency. While many of the techniques discussed in this Review were initially developed for use on NGS platforms, novel techniques developed specifically for third-generation sequencing will work synergistically with the strengths of the platforms. Notably, at present no techniques are available for third-generation profiling of the 5fC and 5caC exclusively; characterisation of the full suite of cytosine modifications is required for a complete epigenetic profiling. In principle, labelling and base conversion techniques could be developed for the sequencing of any non-canonical base on TGS. While great progress is being made in NGS of DNA damage,[83] this methodology has yet to be demonstrated on TGS platforms. In addition to DNA, third-generation sequencing has been used for canonical RNA sequencing.[84] While covalent base modification in eukaryotic DNA is limited to the five covalent cytosine modifications, RNA modifications are extensive and significant innovation will be required to incorporate their identification into third-generation sequencing.

Beyond improvements to their epigenetic sequencing capabilities, these challenges present the opportunity to expand the scope of these devices, which will undoubtedly help innovation in the coming years.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement No 741912 (ERC-ADG, EpiR) and from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 861381 (NATURE-ETN). AK acknowledges support from Science Foundation Ireland (12/RC/2275_P2) and the Irish Research Council (IRCLA/2022/3815). Open Access funding provided by IReL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Keywords: DNA Methylation · Epigenetics · Nanopores · SMRT · Sequence Determination

[1] "The Cost of Sequencing a Human Genome," can be found under https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost, (accessed 11th August 2022).
[2] T. Carell, M. Q. Kurz, M. Müller, M. Rossa, F. Spada, *Angew. Chem. Int. Ed.* **2018**, *57*, 4296–4312; *Angew. Chem.* **2018**, *130*, 4377–4394.
[3] T. B. Johnson, R. D. Coghill, *J. Am. Chem. Soc.* **1925**, *47*, 7.
[4] R. D. Hotchkiss, *J. Biol. Chem.* **1948**, *175*, 315–332.
[5] J. R. Edwards, O. Yarychkivska, M. Boulard, T. H. Bestor, *Epigenet. Chromatin* **2017**, *10*, 23.
[6] Z. D. Smith, A. Meissner, *Nat. Rev. Genet.* **2013**, *14*, 204–220.
[7] P. A. Jones, *Nat. Rev. Genet.* **2012**, *13*, 484–492.
[8] M. V. C. Greenberg, D. Bourc'his, *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 590–607.
[9] M. Ando, Y. Saito, G. Xu, N. Q. Bui, K. Medetgul-Ernar, M. Pu, K. Fisch, S. Ren, A. Sakai, T. Fukusumi, C. Liu, S. Haft, J. Pang, A. Mark, D. A. Gaykalova, T. Guo, A. V. Favorov, S. Yegnasubramanian, E. J. Fertig, P. Ha, P. Tamayo, T. Yamasoba, T. Ideker, K. Messer, J. A. Califano, *Nat. Commun.* **2019**, *10*, 2188.
[10] J. Zhang, C. Yang, C. Wu, W. Cui, L. Wang, *Cancers* **2020**, *12*, 2123.
[11] G. R. Wyatt, S. S. Cohen, *Nature* **1952**, *170*, 1072–1073.
[12] M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, A. Rao, *Science* **2009**, *324*, 930–935.
[13] S. Kriaucionis, N. Heintz, *Science* **2009**, *324*, 929–930.
[14] M. Münzel, D. Globisch, T. Brückl, M. Wagner, V. Welzmiller, S. Michalakis, M. Müller, M. Biel, T. Carell, *Angew. Chem. Int. Ed.* **2010**, *49*, 5375–5377.
[15] D. Globisch, M. Münzel, M. Müller, S. Michalakis, M. Wagner, S. Koch, T. Brückl, M. Biel, T. Carell, *PLoS ONE* **2010**, *5*, e15367.
[16] H. G. A. M. van Luenen, C. Farris, S. Jan, P.-A. Genest, P. Tripathi, A. Velds, R. M. Kerkhoven, M. Nieuwland, A. Haydock, G. Ramasamy, S. Vainio, T. Heidebrecht, A. Perrakis, L. Pagie, B. van Steensel, P. J. Myler, P. Borst, *Cell* **2012**, *150*, 909–921.
[17] X. Wu, Y. Zhang, *Nat. Rev. Genet.* **2017**, *18*, 517–534.
[18] T. Pfaffeneder, B. Hackner, M. Truß, M. Münzel, M. Müller, C. A. Deiml, C. Hagemeier, T. Carell, *Angew. Chem. Int. Ed.* **2011**, *50*, 7008–7012; *Angew. Chem.* **2011**, *123*, 7146–7150.
[19] S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He, Y. Zhang, *Science* **2011**, *333*, 1300–1303.
[20] R. M. Kohli, Y. Zhang, *Nature* **2013**, *502*, 472–479.
[21] H. Hashimoto, Y. Liu, A. K. Upadhyay, Y. Chang, S. B. Howerton, P. M. Vertino, X. Zhang, X. Cheng, *Nucleic Acids Res.* **2012**, *40*, 4841–4849.
[22] F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463–5467.
[23] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, et al., *Science* **2001**, *291*, 1304–1351.
[24] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H. Kent, L. E. Hood, *Nature* **1986**, *321*, 674–679.
[25] E. R. Mardis, *Annu. Rev. Anal. Chem.* **2013**, *6*, 287–303.
[26] S. Goodwin, J. D. McPherson, W. R. McCombie, *Nat. Rev. Genet.* **2016**, *17*, 333–351.
[27] B. E. Slatko, A. F. Gardner, F. M. Ausubel, *Curr. Protoc. Mol. Biol.* **2018**, *122*, e59.
[28] W.-S. Yong, F.-M. Hsu, P.-Y. Chen, *Epigenet. Chromatin* **2016**, *9*, 26.
[29] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, C. L. Paul, *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 1827–1831.
[30] M. Mellén, P. Ayata, N. Heintz, *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E7812–E7821.
[31] M. J. Booth, T. W. B. Ost, D. Beraldi, N. M. Bell, M. R. Branco, W. Reik, S. Balasubramanian, *Nat. Protoc.* **2013**, *8*, 1841–1851.
[32] M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, P. Jin, B. Ren, C. He, *Nat. Protoc.* **2012**, *7*, 2159–2170.
[33] C.-X. Song, K. E. Szulwach, Q. Dai, Y. Fu, S.-Q. Mao, L. Lin, C. Street, Y. Li, M. Poidevin, H. Wu, J. Gao, P. Liu, L. Li, G.-L. Xu, P. Jin, C. He, *Cell* **2013**, *153*, 678–691.
[34] M. J. Booth, G. Marsico, M. Bachman, D. Beraldi, S. Balasubramanian, *Nat. Chem.* **2014**, *6*, 435–440.
[35] X. Lu, C.-X. Song, K. Szulwach, Z. Wang, P. Weidenbacher, P. Jin, C. He, *J. Am. Chem. Soc.* **2013**, *135*, 9315–9317.
[36] K. Tanaka, A. Okamoto, *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1912–1915.
[37] L. Ji, T. Sasaki, X. Sun, P. Ma, Z. A. Lewis, R. J. Schmitz, *Front. Genet.* **2014**, *5*, 341.
[38] N. Olova, F. Krueger, S. Andrews, D. Oxley, R. V. Berrens, M. R. Branco, W. Reik, *Genome Biol.* **2018**, *19*, 33.
[39] E. Meaburn, R. Schulz, *Semin. Cell Dev. Biol.* **2012**, *23*, 192–199.
[40] H. Wu, Y. Zhang, *Nat. Struct. Mol. Biol.* **2015**, *22*, 656–661.
[41] C. Zhu, Y. Gao, H. Guo, J. Song, X. Wu, H. Zeng, K. Kee, F. Tang, C. Yi, *Cell Stem Cell* **2017**, *20*, 720–731.
[42] Y. Liu, J. Cheng, P. Siejka-Zielińska, C. Weldon, H. Roberts, M. Lopopolo, A. Magri, V. D'Arienzo, J. M. Harris, J. A. McKeating, C.-X. Song, *Genome Biol.* **2020**, *21*, 54.
[43] R. Vaisvila, V. K. C. Ponnaluri, Z. Sun, B. W. Langhorst, L. Saleh, S. Guan, N. Dai, M. A. Campbell, B. S. Sexton, K. Marks, M. Samaranayake, J. C. Samuelson, H. E. Church, E. Tamanaha, I. R. Corrêa, S. Pradhan, E. T. Dimalanta, T. C. Evans, L. Williams, T. B. Davis, *Genome Res.* **2019**, *7*, 1280–1289.
[44] B. Xia, D. Han, X. Lu, Z. Sun, A. Zhou, Q. Yin, H. Zeng, M. Liu, X. Jiang, W. Xie, C. He, C. Yi, *Nat. Methods* **2015**, *12*, 1047–1050.
[45] S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, et al., *Science* **2022**, *376*, 44–53.
[46] D. Deamer, M. Akeson, D. Branton, *Nat. Biotechnol.* **2016**, *34*, 518–524.
[47] L.-Q. Gu, S. Cheley, H. Bayley, *Science* **2001**, *291*, 636–640
[48] S. Howorka, S. Cheley, H. Bayley, *Nat. Biotechnol* **2001**, *19*, 636–639..
[49] M. Jain, H. E. Olsen, B. Paten, M. Akeson, *Genome Biol.* **2016**, *17*, 239.
[50] Z. L. Wescoe, J. Schreiber, M. Akeson, *J. Am. Chem. Soc.* **2014**, *136*, 16582–16587.
[51] L. Xu, M. Seki, *J. Hum. Genet.* **2020**, *65*, 25–33.
[52] P. Ni, N. Huang, Z. Zhang, D.-P. Wang, F. Liang, Y. Miao, C.-L. Xiao, F. Luo, J. Wang, *Bioinformatics* **2019**, *35*, 4586–4595.
[53] S. E. Van der Verren, N. Van Gerven, W. Jonckheere, R. Hambley, P. Singh, J. Kilgour, M. Jordan, E. J. Wallace, L. Jayasinghe, H. Remaut, *Nat. Biotechnol.* **2020**, *38*, 1415–1420.
[54] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, *Science* **2009**, *323*, 133.
[55] M. Quail, M. E. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, Y. Gu, *BMC Genomics* **2012**, *13*, 341.
[56] K. J. Travers, C.-S. Chin, D. R. Rank, J. S. Eid, S. W. Turner, *Nucleic Acids Res.* **2010**, *38*, e159.

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[57] A. M. Wenger, P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C.-S. Chin, A. M. Phillippy, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, M. W. Hunkapiller, *Nat. Biotechnol.* **2019**, *37*, 1155–1162.
[58] T. A. Clark, I. A. Murray, R. D. Morgan, A. O. Kislyuk, K. E. Spittle, M. Boitano, A. Fomenkov, R. J. Roberts, J. Korlach, *Nucleic Acids Res.* **2012**, *40*, e29.
[59] B. A. Flusberg, D. R. Webster, J. H. Lee, K. J. Travers, E. C. Olivares, T. A. Clark, J. Korlach, S. W. Turner, *Nat. Methods* **2010**, *7*, 461–465.
[60] T. A. Clark, X. Lu, K. Luong, Q. Dai, M. Boitano, S. W. Turner, C. He, J. Korlach, *BMC Biol.* **2013**, *11*, 4.
[61] O. Y. O. Tse, P. Jiang, S. H. Cheng, W. Peng, H. Shang, J. Wong, S. L. Chan, L. C. Y. Poon, T. Y. Leung, K. C. A. Chan, R. W. K. Chiu, Y. M. D. Lo, *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2019768118.
[62] Y. Yan, K. Wu, J. Chen, H. Liu, Y. Huang, Y. Zhang, J. Xiong, W. Quan, X. Wu, Y. Liang, K. He, Z. Jia, D. Wang, D. Liu, H. Wei, J. Chen, *Virol. Sin.* **2021**, *36*, 901–912.
[63] M. Wang, A. Fu, B. Hu, Y. Tong, R. Liu, Z. Liu, J. Gu, B. Xiang, J. Liu, W. Jiang, G. Shen, W. Zhao, D. Men, Z. Deng, L. Yu, W. Wei, Y. Li, T. Liu, *Small* **2020**, *16*, 2002169.
[64] S. Lhomme, J. Latour, N. Jeanne, P. Trémeaux, N. Ranger, M. Migueres, G. Salin, C. Donnadieu, J. Izopet, *Viruses* **2021**, *13*, 2544.
[65] O. González-Recio, M. Gutiérrez-Rivas, R. Peiró-Pastor, P. Aguilera-Sepúlveda, C. Cano-Gómez, M. Á. Jiménez-Clavero, J. Fernández-Pinero, *Appl. Microbiol. Biotechnol.* **2021**, *105*, 3225–3234.
[66] L. Chavez, Y. Huang, K. Luong, S. Agarwal, L. M. Iyer, W. A. Pastor, V. K. Hench, S. A. Frazier-Bowers, E. Korol, S. Liu, M. Tahiliani, Y. Wang, T. A. Clark, J. Korlach, P. J. Pukkila, L. Aravind, A. Rao, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E5149–E5158.
[67] W.-W. Li, L. Gong, H. Bayley, *Angew. Chem. Int. Ed.* **2013**, *52*, 4350–4355; *Angew. Chem.* **2013**, *125*, 4446–4451.
[68] T. Zeng, L. Liu, T. Li, Y. Li, J. Gao, Y. Zhao, H.-C. Wu, *Chem. Sci.* **2015**, *6*, 5628–5634.
[69] R. Vaisvila, V. K. C. Ponnaluri, Z. Sun, B. W. Langhorst, L. Saleh, S. Guan, N. Dai, M. A. Campbell, B. S. Sexton, K. Marks, M. Samaranayake, J. C. Samuelson, H. E. Church, E. Tamanaha, I. R. Corrêa, S. Pradhan, E. T. Dimalanta, T. C. Evans, L. Williams, T. B. Davis, *Genome Res.* **2021**, *31*, 1280–1289.
[70] C. S. Nabel, H. Jia, Y. Ye, L. Shen, H. L. Goldschmidt, J. T. Stivers, Y. Zhang, R. M. Kohli, *Nat. Chem. Biol.* **2012**, *8*, 751–758.
[71] Z. Sun, R. Vaisvila, L.-M. Hussong, B. Yan, C. Baum, L. Saleh, M. Samaranayake, S. Guan, N. Dai, I. R. Corrêa, S. Pradhan, T. B. Davis, T. C. Evans, L. M. Ettwiller, *Genome Res.* **2021**, *31*, 291–300.
[72] Y. Han, G. Y. Zheleznyakova, Y. Marincevic-Zuniga, M. P. Kakhki, A. Raine, M. Needhamsen, M. Jagodic, *Epigenetics* **2021**, 1–10.
[73] Y. Sakamoto, S. Zaha, S. Nagasawa, S. Miyake, Y. Kojima, A. Suzuki, Y. Suzuki, M. Seki, *Nucleic Acids Res.* **2021**, *49*, 14.
[74] Y. Liu, P. Siejka-Zielińska, G. Velikova, Y. Bi, F. Yuan, M. Tomkova, C. Bai, L. Chen, B. Schuster-Böckler, C.-X. Song, *Nat. Biotechnol.* **2019**, *37*, 424–429.
[75] N. S. W. Jonasson, L. J. Daumann, *Chem. Eur. J.* **2019**, *25*, 12091–12097.
[76] D. Schmidl, N. Jonasson, E. Korytiakova, T. Carell, L. Daumann, *Angew. Chem. Int. Ed.* **2021**, *60*, 21457–21463; *Angew. Chem.* **2021**, *133*, 21627–21633.
[77] T. Gabrieli, H. Sharim, G. Nifker, J. Jeffet, T. Shahal, R. Arielly, M. Levi-Sakin, L. Hoch, N. Arbib, Y. Michaeli, Y. Ebenstein, *ACS Nano* **2018**, *12*, 7148–7158.
[78] C. Vandiedonck, J. C. Knight, *Briefings Funct. Genomics Proteomics* **2009**, *8*, 379–394.
[79] O. Galm, S. Wilop, J. Reichelt, E. Jost, G. Gehbauer, J. G. Herman, R. Osieka, *Leukemia* **2004**, *18*, 1687–1692.
[80] W. J. Locke, D. Guanzon, C. Ma, Y. J. Liew, K. R. Duesing, K. Y. C. Fung, J. P. Ross, *Front. Genet.* **2019**, *10*, 1150.
[81] J. Shao, S. Wang, D. West-Szymanski, J. Karpus, S. Shah, S. Ganguly, J. Smith, Y. Zu, C. He, Z. Li, *Sci. Rep.* **2022**, *12*, 12410.
[82] D. Han, X. Lu, A. H. Shih, J. Nie, Q. You, M. M. Xu, A. M. Melnick, R. L. Levine, C. He, *Mol. Cell* **2016**, *63*, 711–719.
[83] C. Mingard, J. Wu, M. McKeague, S. J. Sturla, *Chem. Soc. Rev.* **2020**, *49*, 7354–7377.
[84] R. Stark, M. Grzelak, J. Hadfield, *Nat. Rev. Genet.* **2019**, *20*, 631–656.

Manuscript received: October 25, 2022
Accepted manuscript online: December 16, 2022
Version of record online: January 25, 2023

© 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2023, 14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202215704 by Perkins Coie LLP, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License