Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Structural Biology

Going beyond five bases in DNA sequencing

Jonas Korlach and Stephen W Turner

DNA sequencing has provided a wealth of information about biological systems, but thus far has focused on the four canonical bases, and 5-methylcytosine through comparison of the genomic DNA sequence to a transformed four-base sequence obtained after treatment with bisulfite. However, numerous other chemical modifications to the nucleotides are known to control fundamental life functions, influence virulence of pathogens, and are associated with many diseases. These modifications cannot be accessed with traditional sequencing methods. In this opinion, we highlight several emerging single-molecule sequencing techniques that have the potential to directly detect many types of DNA modifications as an integral part of the sequencing protocol.

Address
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025, United States

Corresponding authors: Korlach, Jonas (jkorlach@pacificbiosciences.com) and Turner, Stephen W (sturner@pacificbiosciences.com)

Current Opinion in Structural Biology 2012, **22**:251–261

This review comes from a themed issue on
Nucleic acids
Edited by Jamie Williamson and Jody Puglisi

Available online 9th May 2012

0959-440X/$ – see front matter

http://dx.doi.org/10.1016/j.sbi.2012.04.002

Introduction

Nucleic acid sequencing represents one of the most important techniques in the study of biological systems – the order of the bases in genomes and transcriptomes define the genetic blueprints and cellular identities in all organisms. Decoding the sequence of the four canonical bases in a population of identical DNA molecules became technically feasible in the late 1960s (reviewed in [1]), and was developed into a high-throughput and automated technique through Sanger sequencing [2,3]. Subsequently, second-generation sequencing technologies were developed to provide massively greater DNA sequencing throughput at reduced cost, albeit at the expense of sequence readlength (reviewed in [4,5]). Recently, single-molecule sequencing methods have emerged with the promise to provide extremely long DNA sequencing reads, faster intrinsic sequencing speeds, less processing steps before sequencing, and an improved ability to resolve complex heterogeneous DNA mixtures (reviewed in [6–8]).

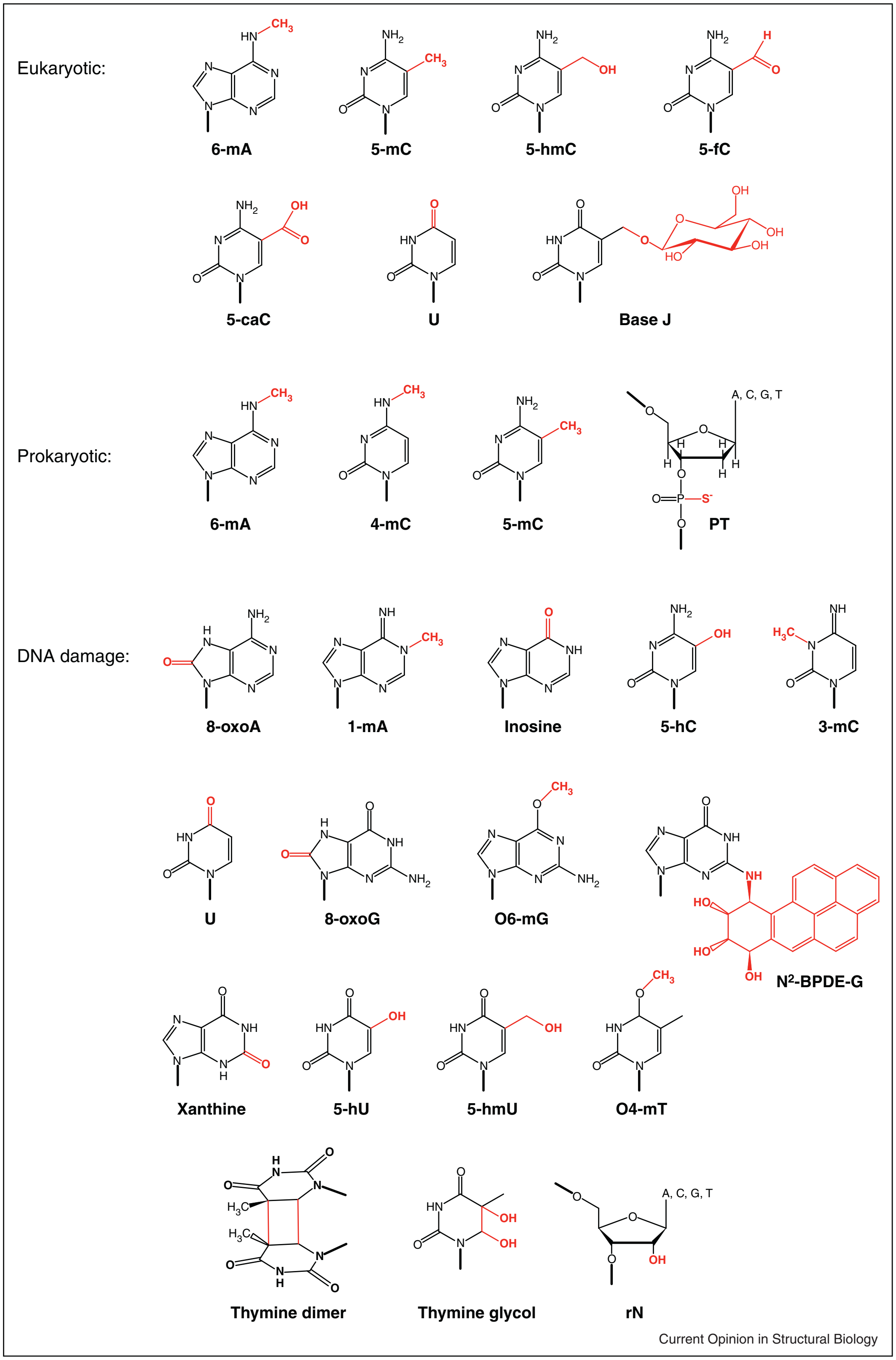
In addition to this basic hereditary genetic information, DNA contains epigenetic modifications that are present in the genomes of virtually all organisms, including viruses and phages. They are variable with respect to space (the organism's different tissues and cell types) and time (different stages in the organism's life cycle), and thereby greatly expand the structural complexity and information depth of DNA. In higher eukaryotes, the most common epigenetic marker is 5-methylcytosine (5-mC) introduced enzymatically by DNA methyltransferases after DNA replication. 5-mC is essential for growth and development, affects gene expression, genomic imprinting, suppression of transposable elements, X chromosome inactivation, and has been implicated in a variety of diseases including autism and colon cancer [9]. The sequencing of 5-mC is commonly achieved through bisulfite chemistry treatment that converts all cytosine residues into uracil, but leaves 5-mC unchanged, followed by amplification of the DNA product, converting uracil into T. Several extensive reviews on this method as well as other array-based, non-sequencing approaches are available [10–12]. Detailed comparisons of the strengths and weaknesses of different variations of the bisulfite method have also been presented [13•,14•].

While the technological progress of sequencing the four canonical bases and by extension 5-mC (through bisulfite sequencing) at ever increasing speed and efficiency has been phenomenal, methods to interrogate the many other important chemical forms of DNA in the context of sequencing have been forthcoming much more slowly. In this opinion, we will summarize the current knowledge of chemical DNA modifications, review existing methods to obtain this additional information, and highlight emerging approaches that enable elucidation of DNA base modifications as an integral part of DNA sequencing.

Current spectrum of DNA modifications

DNA contains a large variety of functionally important modifications (Figure 1). Beyond 5-mC, 5-hydroxymethylcytosine (5-hmC), resulting from oxidation of 5-mC by the family of Tet enzymes, has been detected in a variety of mammalian cells and has been connected to embryonic stem cell differentiation, cellular development, and carcinogenesis [15–17]. 5-hmC can be detected through bisulfite-based methods, but it cannot be distinguished from 5-mC [18]. Recently, 5-formylcytosine (5-fC), and 5-carboxycytosine (5-caC) were found in mouse embryonic stem cells and mouse organs [19••], bringing

Figure 1



Current Opinion in Structural Biology

Landscape of DNA modifications. The chemical modifications are highlighted in red. Abbreviations: 6-mA: N6-methyladenine; 5-mC: 5-methylcytosine; 5-hmC: 5-hydroxymethylcytosine; 5-fC: 5-formylcytosine; 5-caC: 5-carboxycytosine; U: uracil; base J: β-D-glucopyranosyloxymethyluracil; 4-mC: 4-methylcytosine; PT: phosphorothioate nucleotides; 8-oxoA: 8-oxoadenine; 1-mA: 1-methyladenine; 5-hC: 5-hydroxycytosine; 3-mC: 3-methylcytosine; 8-oxoG: 8-oxoguanine; O6-mG: O6-methylguanine; N^2-BPDE-G: benzo[a]pyrene diol epoxide-guanine; 5-hU: 5-hydroxyuracil; 5-hmU: 5-hydroxymethyluracil; O4-mT: O4-methylthymine; rN: ribonucleotides.

the total number of chemically modified forms of cytosine to four, and making it an intriguing speculation whether additional epigenetic forms of cytosine may exist. Apart from cytosine modifications, N6-methyladenine (6-mA) has been found in protists, plants and mosquitoes, and there is indirect evidence that it may also be present in mammalian cells (reviewed in [20]). In certain cell types, other base modifications have been described to affect specific functions, such as uracil for important roles in innate and adaptive immunology [21], β-D-glucopyranosyloxymethyluracil (base J) that affects gene regulation and telomere function in certain parasitic protozoa such as trypanosomes [22,23•], and ribonucleosides to imprint mating-type switching during the cell cycle in yeast [24].

DNA modifications are also wide-spread in prokaryotes. The majority of known bacterial epigenetic modifications serve as identity markers for protection of the organism's DNA from invading pathogens through methylation of specific sequence contexts that are targeted by restriction enzymes. The three most common base modifications of such restriction modification systems are 4-methylcytosine (4-mC), 5-mC and 6-mA [25], but other types of modifications have been described, such as protection through phosphorothioate modifications, with sulfur replacing a non-bridging phosphate oxygen [26,27•]. The number of recognized restriction modification systems has grown rapidly over the past few years (Figure 2). In addition, DNA adenine methyltransferase (Dam) imparts 6-mA to regulate basic cellular functions such as cell cycle and DNA replication control, post-replicative mismatch repair, regulation of gene expression, phase variation switching, and pathogenicity [28–32]. In *E. coli*, 5-mC catalyzed by DNA cytosine methyltransferase (Dcm) reduces expression of ribosomal protein genes during stationary phase [33].

Apart from desired and enzymatically mediated base modifications, DNA can also be altered as a result of DNA damage. DNA is under constant stress from both endogenous and exogenous sources, and the bases exhibit limited chemical stability and are vulnerable to chemical modifications through different types of damage, including oxidation, alkylation, radiation damage, and hydrolysis. DNA base modifications resulting from these types of DNA damage are wide-spread and play important roles in affecting physiological states and disease phenotypes (reviewed in [34–37]). For example, products of oxidative damage, such as 8-oxoadenine and 8-oxoguanine, particularly prevalent in mitochondrial DNA, have been implicated with aging as well as neurodegenerative diseases, for example, Alzheimer's and Parkinson's [38–42]. The undesired transfer of an alkyl group, for example, in 1-methyladenine, 3-methylcytosine, O6-methylguanine, or O4-methylthymine, has been associated with gliomas and colorectal carcinomas, but is also used in chemotherapy to damage the DNA of cancer cells (reviewed in [35]). Environmental effects from smoking, industrial chemical or UV light exposures can result in large adducts, for example, benzo[a]pyrene diol epoxide (BPDE) and pyrimidine dimers, eliciting lung and skin cancer. Ionizing radiation can confer damage to DNA bases, for example, in the form of 5-hydroxycytosine, 5-hydroxyuracil, 5-hydroxymethyluracil or thymine glycol, either through direct effects or indirectly through the generation of free radicals that elicit damage, resulting in chronic inflammatory diseases, prostate, breast and colorectal cancer [43–45]. Products of spontaneous deamination, such as uracil, xanthine, inosine, and thymine (through deamination of 5-mC), are mutagenic and can lead to various forms of cancer (reviewed in [34,46]). Another form of damage relates to the incorporation of ribonucleotides into genomic DNA during replication that can be as high as 0.1% in yeast [47••] and has been implicated in genome instability [48].

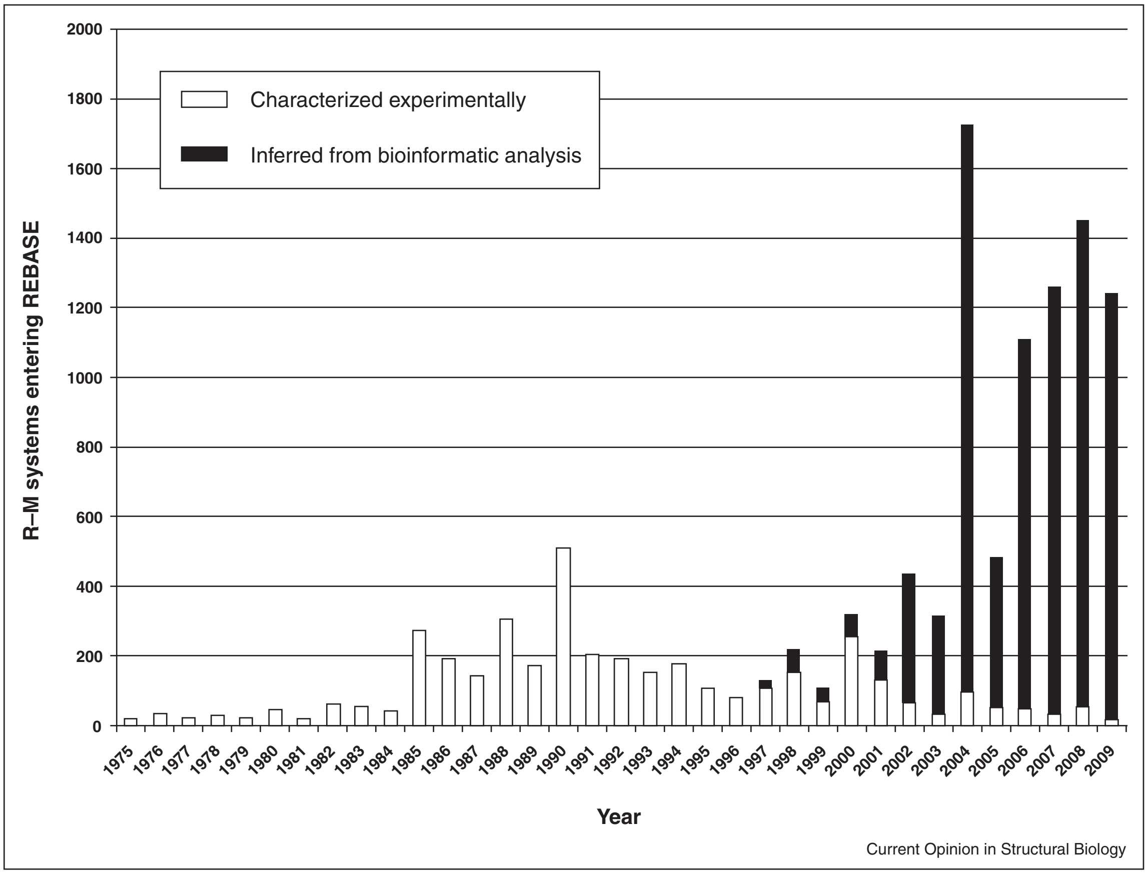
Available technologies for recognizing and characterizing base modifications beyond 5-mC have largely been limited to bulk methods such as chromatography, mass spectroscopy, electrochemistry, radioactive labeling and immunochemical assays, or sensitivity to restriction enzymes (reviewed in [20,49,50]). These methods commonly lack the necessary resolution to determine the exact position and strand identity of base modifications. Therefore, sequencing methods that are capable of directly detecting base modifications are highly desirable.

DNA modification detection through ensemble sequencing

It has been demonstrated that the three common bacterial epigenetic markers 4-mC, 5-mC and 6-mA can be detected by automated dye-terminator Sanger sequencing [51,52]. The presence of a methyl group in the DNA template alters the efficiency of dideoxyterminator nucleotide incorporation, resulting in variations in the peak heights of the fluorescence trace chromatogram when compared to an unmodified control DNA sample. T signals were higher when 6-mA was present in the template, whereas G signals were lower for 5-mC and higher for 4-mC, respectively, allowing the differentiation of the two different forms of cytosine. The method has been applied to characterizing methyltransferases in *Helicobacter pylori* [51], *E. coli* [53], and *Neisseria meningitidis* [54]. To our knowledge, the method has not found widespread application towards the characterization of microbial methylomes, perhaps owing to the subtle magnitudes of the observed signals and the relatively low-throughput nature of Sanger sequencing. We are not aware of attempts to apply this method to other DNA base modifications.

Second generation DNA sequencing techniques rely on DNA amplification during sample preparation steps,

Figure 2



Current Opinion in Structural Biology

History of restriction-modification (R-M) systems entered into the REBASE database, with open bars denoting systems with accompanying biochemical or genetic characterization, and black bars showing potential R-M systems inferred from bioinformatic sequence analysis. Adapted from reference [25].

resulting in the loss of base modifications and precluding the development of detection strategies for modified bases during the actual sequencing protocol. Consequently, efforts to deduce sites of base modifications have focused on a variety of upfront sample preparation techniques to indirectly infer their presence. Second generation sequencing methods have been used extensively to characterize the abundance and dynamic regulation of 5-mC through the use of bisulfite sequencing (reviewed in [10–12]). In addition, a strategy employing the MspJI family of restriction endonucleases has been described [55•]. These enzymes recognize 5-mC or 5-hmC in DNA and excise a short DNA fragment, 12 bases upstream and 16 bases downstream of the modified base. The extracted fragments can then be subjected to second generation sequencing to determine the positions of the modified bases in the genome. The enzymes favor different nucleotides flanking the modified cytosine that can introduce a bias for establishing a whole-genome representation. Both bisulfite and MspJI-mediated sequencing currently cannot distinguish between 5-mC and 5-hmC [18,56].

To address other base modifications, enrichment techniques that target a specific modification have been employed. Pull-down methods that select DNA fragments containing the modified base recognized by antibodies or modification-specific binding proteins, followed by high-throughput second generation DNA sequencing, allow the regional assignment of the presence of at least one base modification over the length of the sequenced fragment. However, they typically preclude the assignment of the exact position of the modification and which DNA strand was modified. Several different strategies for enrichment of 5-hmC containing DNA have been described (reviewed in [57•,58]), as well as anti-J antibody-based enrichment for characterizing genomic base J distributions [59–61].

Emerging methods for sequencing DNA modifications

There has been great interest in pushing the DNA sequencing sensitivity to the ultimate analytical limit to obtain sequences from individual DNA molecules. One of the driving forces behind these efforts is the enablement of reading DNA base modifications as an integral part of the sequencing method. Several different strategies are pursued, and some have already been commercialized (reviewed in [6,62,63]).

Gated single-molecule sequencing-by-synthesis

Originally at the heart of second generation-based DNA sequencing approaches, gated sequencing-by-synthesis has been adapted to the single-molecule regime in a technology commercialized by Helicos Biosciences. Individual DNA molecules are extended and detected one base at a time through the use of DNA polymerase and fluorescently labeled, terminating nucleotides that carry a cleavable moiety to remove the termination group after the detection step [64]. Although the method does not rely on amplification of the target DNA, the success of the sequencing principle relies on completed nucleotide incorporations on a DNA strand before imaging, so it may be difficult to find solutions that make it directly sensitive to base modifications in the DNA template. It has been applied successfully in conjunction with one of the 5-hmC-specific targeted enrichment strategies for the genome-wide mapping of 5-hmC-containing DNA fragments in embryonic stem cells [65].

Single-molecule, real-time DNA sequencing

Sequencing by monitoring the uninterrupted activity of DNA polymerase is realized in single-molecule, real-time (SMRT®) DNA sequencing, commercialized by Pacific Biosciences [66,67]. Phospholinked nucleotides that carry a fluorescent label attached to the terminal phosphate moiety are employed to allow continuous DNA synthesis, making washing and gating steps unnecessary. As a part of the sequencing process, the detailed dynamics of DNA polymerization is recorded in the form of a train of fluorescent pulses. It was observed that the polymerase dynamics was affected by the presence of modified bases in the DNA template [68•]. The length of time that the polymerase retains a nucleotide bound in its active site (pulse width, PW), and the time interval between successive nucleotide-bound states (interpulse duration, IPD) are the principal pulse metrics used in the analysis to ascertain whether the polymerase kinetics was altered for a modification-containing template when compared to an unmodified control template. Mechanistically, IPDs can be affected by (i) changes in the affinity of binding the incoming nucleotide, or (ii) altered DNA translocation rates following the phospholinked nucleotide incorporation. Variations in PW can be caused from (i) effects on the rates of conformational changes of the enzyme, as well as (ii) the rate of catalysis during the nucleotide incorporation cycle, as the modified base in the template can distort active site geometries. These effects are captured in SMRT sequencing through the real-time monitoring of each nucleotide incorporation event, thereby making the method sensitive to even small changes from relatively subtle chemical modifications, such as 5-mC.

Since SMRT sequencing is compatible with the sequencing of native, unamplified DNA, the direct detection of modified DNA bases as part of this sequencing method is possible. Because the typical template preparation in SMRT sequencing results in a circularly closed DNA molecule [69], it is possible to interrogate the same modified base multiple times for increased statistical power of detection. It also enables sequencing both strands of a DNA molecule in the same sequencing read, directly addressing correlations that may exist between modifications in the sense and antisense strands at a given position (e.g. as in full methylation of CpG contexts).

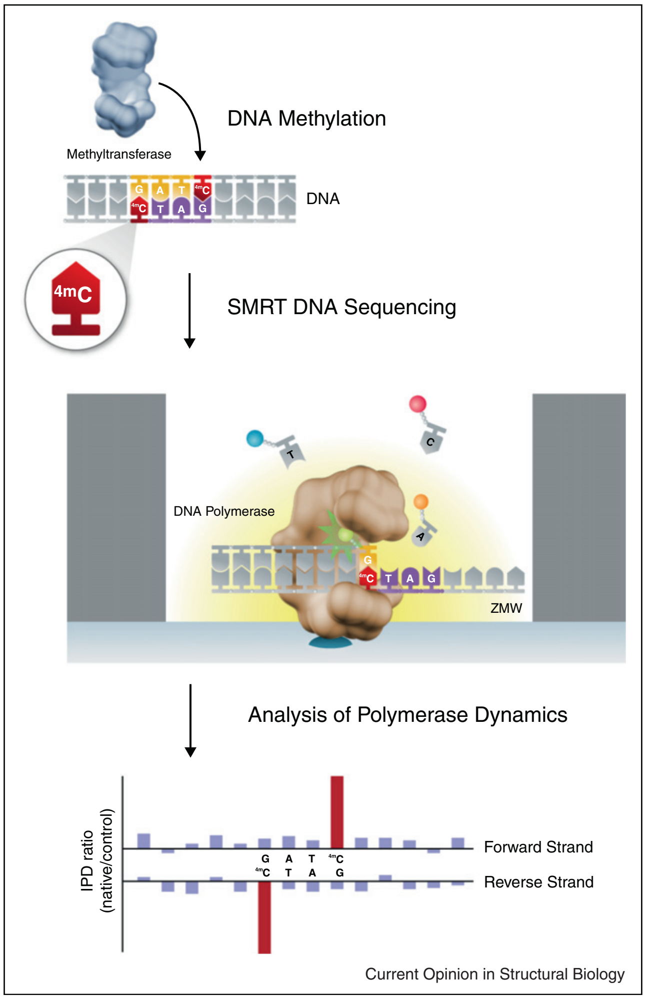
Following the initial description of SMRT sequencing for directly detecting 5-mC, 5-hmC and 6-mA [68•], the approach has thus far been applied in three areas. The first report describes the selective labeling and sequencing of 5-hmC, combining one of the aforementioned 5-hmC targeted enrichment methods with SMRT sequencing [70•]. The addition of chemical groups to 5-hmC as part of the enrichment procedure enhances the kinetic signals, allowing the base-resolved and DNA strand-specific determination of 5-hmC at the level of individual DNA molecules. It was applied to synthetic DNA samples as well as DNA extracted from mouse embryonic stem cells.

In a second study, SMRT sequencing was successfully applied to determine the recognition sequence contexts of 16 methyltransferases, encompassing all three common prokaryotic epigenetic forms of 4-mC, 5-mC and 6-mA, and including three methyltransferases with previously unknown sequence context specificities [71•] (Figure 3). For certain methyltransferases, including the Dam methyltransferase in *E. coli*, unanticipated promiscuities were observed, with methylation of sites that were close but not identical to the cognate sequence context.

Third, the method has been employed to characterize the kinetic signatures of products of DNA damage, including 8-oxoguanine, 8-oxoadenine, O6-methylguanine, 1-methyladenine, O4-methylthymine, 5-hydroxycytosine, 5-hydroxyuracil, 5-hydroxymethyluracil, or thymine dimers, demonstrating that these base modifications can also be readily detected with single-modification resolution and DNA strand specificity [72•]. The distinct kinetic signatures generated by these DNA base modifications were studied on synthetic templates. The method has yet to be applied to identify damaged DNA bases in biological samples.

The kinetic effects are not necessarily restricted to the nucleotide incorporation opposite the modified base, but can span a region surrounding the base modifications, encompassing approximately three nucleotide incorporations prior and seven incorporations after the site of the modification. The magnitudes of kinetic signals over this

Figure 3



Current Opinion in Structural Biology

Characterization of DNA methylation by SMRT sequencing. Base methylation catalyzed by methyltransferases is detected through analysis of the polymerase dynamics during SMRT sequencing, comparing the native DNA sample to an unmodified control of identical sequence. The example shows 4-mC methylation in the 5′-GATC-3′ sequence context by the methyltransferase M.EsaLHCI.
Adapted from reference [71•].

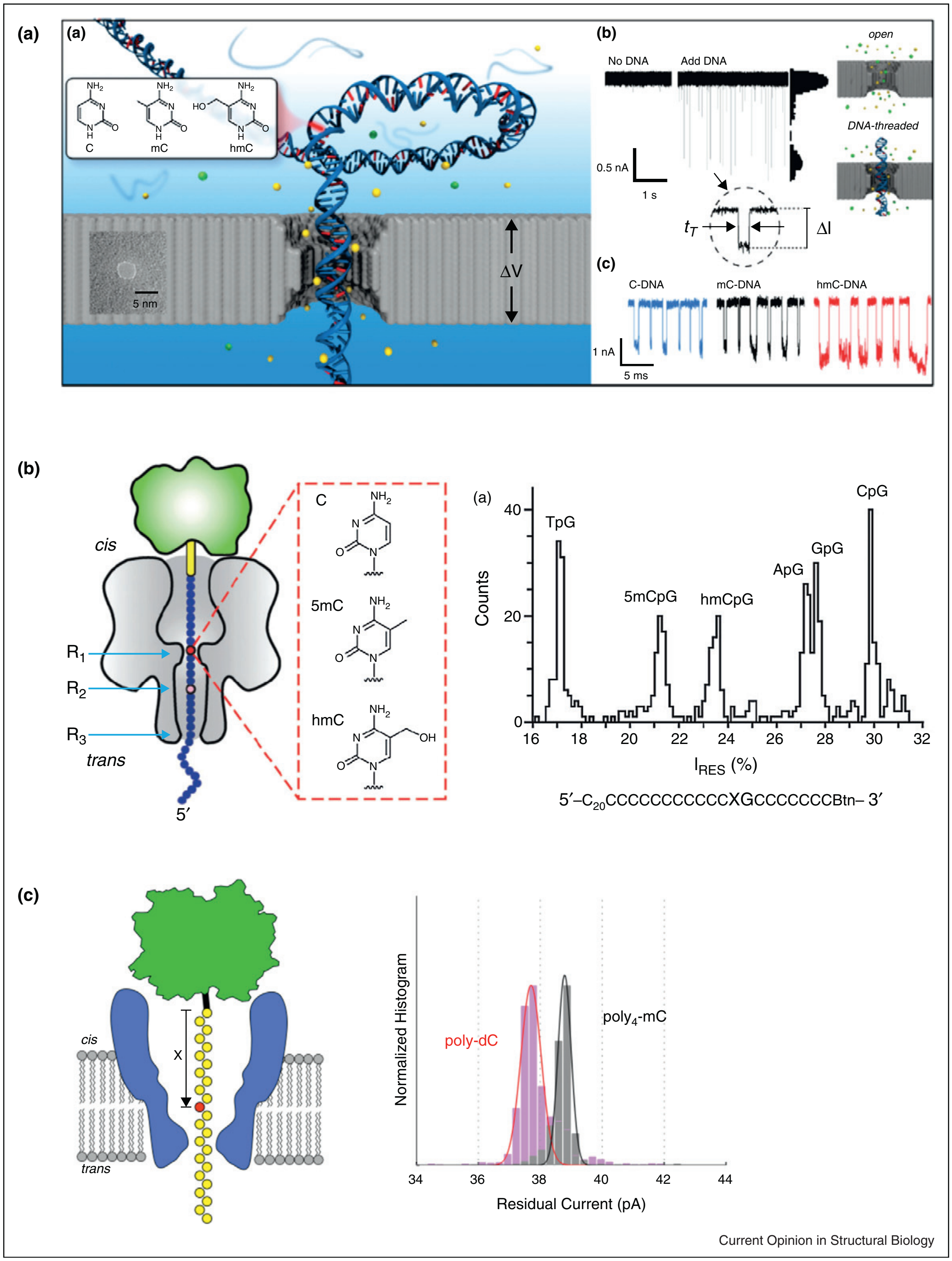
region are dependent on the chemical type of the base modification and the local sequence context. The effects are attributed to the extended contact the polymerase makes with the incoming DNA template and the nascent double-stranded DNA over its DNA binding region of ~11 bases [73]. For misincorporation events, the transmission of the presence of the mismatch back to the enzyme active site through long-range distortions in the DNA has been described [74]. Similarly here, distortions in the DNA geometry by a modified base in the extended DNA binding region can influence the dynamics of nucleotide incorporations at the polymerase active site, giving rise to additional signals that result in a distinct kinetic signature for a given modified DNA base. While the specific DNA–protein interactions and sequence context effects remain to be characterized in more detail, these kinetic signatures can already be harnessed to discriminate between different chemical base modifications. More extensive characterizations of these effects are underway to distinguish modifications that have similar signatures (e.g. 6-mA and 1-mA), and to resolve modifications that are positioned in very close proximity. For modifications with relatively subtle kinetic effects, such as 5-mC, relatively high sequencing fold coverage is currently recommended for confident assignments, and the development of new enzymes with enhanced kinetic signals is desirable.

Nanopore-based single-molecule sequencing

In nanopore-based single-molecule sequencing approaches, DNA molecules are individually driven through nanoscale pores of dimensions that only permit the passage of single-stranded DNA in strict linear sequence (reviewed in [75]). Nanopores can be of biological origin – membrane protein complexes such as α-hemolysin (αHL) or *Mycobacterium smegmatis* porin A (MspA) protein – or they can be created synthetically from solid state materials, such as silicon nitride, aluminium oxide, multilayer graphene sandwiches, graphene monolayers, or carbon nanotubes. The pore contains a detection region capable of differentiating the different bases, thereby generating a temporal signal corresponding to the sequence of bases as they traverse the pore. Different types of signals are being evaluated, including differential ionic current blockades through the pore, tunnelling currents across the pore, or capacitance changes.

It has been demonstrated that ion current signals can discriminate 5-mC and 5-hmC from the four canonical bases in single-stranded DNA fragments that were immobilized in a αHL nanopore [76•], MspA pore [77•], or driven through a thin silicon nitride nanopore [78•] (Figure 4). Precise control over the DNA transport through the pore is crucial to the success of these methods, and has in some instances been implemented by the use of enzymatic activities, such as exonucleases or polymerases positioned close to the pore's entrance. In conjunction with the exonuclease-based strategy, using an αHL nanopore with a covalently attached cyclodextrin adapter molecule as the detector reading head, 5-mC could be distinguished from the four canonical bases as free nucleotide monophosphates by measuring the residual pore current [79]. Using the polymerase-associated strategy, the real-time detection of sequential nucleotide additions by a DNA polymerase, measured by differential current blockage of an αHL pore as the DNA template strand was drawn through the nanopore lumen during replication, has been reported [80,81•]. Since in this approach the sequencing is based on single-molecule, real-time observations of uninterrupted

Figure 4



Current Opinion in Structural Biology

Detection of 5-mC and 5-hmC through nanopore sequencing. The presence of the modified cytosine is detected through a differential residual pore current. From references [78•] (panel A), [76•] (panel B), and [77•] (panel C).

DNA polymerase activity, it is conceivable that base modifications will be detectable in the same manner as in SMRT sequencing by analyzing the polymerase dynamics, as described above.

To address potential challenges related to achieving large multiplex factors of independently monitoring the small electrical signals from each nanopore, one alternative strategy has been reported that allows for the parallel optical readout from multiple nanopores [82], but it relies on a conversion of each nucleotide of the target DNA sequence to a known oligonucleotide followed by hybridization with molecular beacons, and it may be difficult to implement this sample preparation method to be base-modification sensitive.

Microscopy-based single-molecule sequencing

Advanced electron microscopy techniques are being pursued to directly read genetic and epigenetic information from a DNA template molecule with single-base spatial resolution (reviewed in [6,7]). The different approaches include transmission electron microscopy (TEM) for direct chemical detection of atoms that would uniquely identify the different bases. Other approaches include prior differential chemical labeling of the bases with compounds such as iodine and bromine to enhance detectability and contrast. It is conceivable that in the future, either direct imaging or chemical conversion techniques that are specific to certain base modifications, for example, selective labeling of 5-hmC with glucose containing detectable atoms, or glucose azide followed by addition of a detectable linker (analogous to strategies reviewed in [57•]), can be utilized to detect the presence of certain base modifications in these sequencing strategies.

One of the prerequisites associated with microscopy approaches is the reliable deposition of ordered arrays of individual elongated DNA molecules onto a substrate. This was implemented recently through transfer-printing of DNA molecules onto single-layer graphene substrates (considered the best substrate for high-resolution transmission electron microscopy), followed by high-resolution electron beam imaging and electron energy loss spectroscopy analysis [83•]. This method was then adapted to profile 5-mC distributions along stretched DNA molecules upon binding of fluorescently labeled methyl-CpG binding peptides to allow comparison of DNA molecules with different methylation states, albeit thus far with wide-field fluorescence detection [84]. In the paper, it is noted that the method can be adapted to super-resolution optical microscopy or electron microscopy techniques for enhanced spatial resolution.

Another approach encompasses the imaging of single-stranded DNA molecules by scanning tunnelling microscopy (STM), and detection of the four bases by differential tunnelling current between the atomically sharp STM tips, the base, and a conductive surface. Using this approach, an 'electronic fingerprint' of guanine bases in single-stranded DNA molecules of known sequence has been demonstrated [85]. In the future, it may be possible to collect differential electronic fingerprints for modified bases using this strategy.

Direct RNA sequencing

An even larger number of base modifications have been recognized in RNA molecules. These are essential for determination of structure and regulation of essential metabolic processes, and they have been implicated in disease progression (reviewed in [86•,87]). They have been difficult to study in the context of sequencing owing to the lack of routine, high-throughput techniques that can sequence RNA directly. As a result, RNA base modification research has largely been restricted to cumbersome and labor-intensive non-sequencing assays that often lack base-resolution [88–90]. Bisulfite treatment has been adapted to RNA, followed by PCR-based amplification of cDNA and DNA sequencing, to allow identification and characterization of RNA 5-mC methylation patterns [91•]. Certain RNA base modifications result in a base change of the complementary cDNA strand during cDNA synthesis (e.g. adenine-to-inosine editing, as inosine is recognized as guanine during *in vitro* reverse transcription), and can thereby be inferred from comparing the cDNA sequence with the genomic sequence from the same source [92].

A short-read technology for sequencing RNA directly based on the Helicos approach has been described [93,94•], but analogous to DNA sequencing it may be difficult to adapt this approach to detecting RNA base modifications. In a recent paper [95], SMRT sequencing was used to directly interrogate RNA templates through the replacement of the DNA polymerase with a reverse transcriptase, and the detection of 6-mA in both synthetic RNA templates and cellular RNA expressed from a transfected reporter construct was demonstrated.

Conclusions

DNA sequencing has already revolutionized our understanding of biology, and has fundamentally altered the way biological research is undertaken. Extensive and cost-effective DNA and cDNA sequencing, when combined through powerful computational methods with proteomics, biomolecular dynamics and a host of other phenotypic information, are expected to precipitate a radical transformation to improve human health, food and energy supplies, and environmental conservation. However to date, sequencing has focused on the canonical four bases and 5-mC, and it is increasingly apparent that this provides an insufficient view of the native structure of DNA. The multitude of known nucleotide modifications play defining roles in a large variety of important biological processes, including gene expression, immunity, disease, and pathogenicity. In addition, judging from the recent discovery rate of new base modifications, it is probable that the full complement of all functionally relevant chemical DNA modifications has not yet been recognized. Therefore, a pressing need exists to develop automated, high-throughput sequencing technologies that can interrogate the DNA beyond A, C, G, T and 5-mC to reveal the full structural and functional complexity of DNA. With several emerging new sequencing approaches, we are poised to enter a next phase in the sequencing revolution that enables such a comprehensive epigenomic viewpoint.

Acknowledgement

This work was supported partly by National Institutes of Health grant 1RC2HG005618-01 (NHGRI).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Wu R: **Development of enzyme-based methods for DNA sequence analysis and their applications in the genome projects**. *Adv Enzymol Relat Areas Mol Biol* 1993, **67**:431-468.

2. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci USA* 1977, **74**:5463-5467.

3. Shendure JA, Porreca GJ, Church GM, Gardner AF, Hendrickson CL, Kieleczawa J, Slatko BE: **Overview of DNA sequencing strategies**. *Curr Protoc Mol Biol* 2011, **Chapter7** Unit 7.1.

4. Mardis ER: **Next-generation DNA sequencing methods**. *Annu Rev Genomics Hum Genet* 2008, **9**:387-402.

5. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nat Biotechnol* 2008, **26**:1135-1145.

6. Xu M, Fujita D, Hanagata N: **Perspectives and challenges of emerging single-molecule DNA sequencing technologies**. *Small* 2009, **5**:2638-2649.

7. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE: **Landscape of next-generation sequencing technologies**. *Anal Chem* 2011, **83**:4327-4341.

8. Efcavitch JW, Thompson JF: **Single-molecule DNA analysis**. *Annu Rev Anal Chem (Palo Alto Calif)* 2010, **3**:109-128.

9. Ferguson-Smith AC, Greally JM, Martienssen RA: *Epigenomics*. Springer; 2009.

10. Hirst M, Marra MA: **Next generation sequencing based approaches to epigenomics**. *Brief Funct Genomics* 2010, **9**:455-465.

11. Fouse SD, Nagarajan RP, Costello JF: **Genome-scale DNA methylation analysis**. *Epigenomics* 2010, **2**:105-117.

12. Estecio MR, Issa JP: **Tackling the methylome: recent methodological advances in genome-wide methylation profiling**. *Genome Med* 2009, **1**:106.

13. • Robinson MD, Statham AL, Speed TP, Clark SJ: **Protocol matters: which methylome are you actually studying?** *Epigenomics* 2010, **2**:587-598.
Comparison of different next-generation sequencing methods to study 5-mC. Evaluation of the impact of the DNA sequence and bias effects introduced to datasets by genome-wide DNA methylation technologies.

14. • Bock C, Tomazou EM, Brinkman AB, Muller F, Simmer F, Gu H, Jager N, Gnirke A, Stunnenberg HG, Meissner A: **Quantitative comparison of genome-wide DNA methylation mapping technologies**. *Nat Biotechnol* 2010, **28**:1106-1114.
Detailed comparison of methylated DNA immunoprecipitation sequencing (MeDIP-seq), methylated DNA capture by affinity purification (MethylCap-seq), reduced representation bisulfite sequencing (RRBS) and the Infinium HumanMethylation assay for genome-wide mapping of 5-mC.

15. Kriaucionis S, Heintz N: **The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain**. *Science* 2009, **324**:929-930.

16. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L *et al.*: **Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1**. *Science* 2009, **324**:930-935.

17. Munzel M, Globisch D, Carell T: **5-Hydroxymethylcytosine, the sixth base of the genome**. *Angew Chem Int Ed Engl* 2011, **50**:6460-6468.

18. Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A: **The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing**. *PLoS ONE* 2010, **5**:e8888.

19. •• Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y: **Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine**. *Science* 2011, **333**:1300-1303.
Demonstration that 5-fC and 5-caC are present in genomic DNA of mouse ES cells and mouse organs. The two new forms of cytosine can be generated by Tet proteins in an enzymatic activity-dependent manner.

20. Ratel D, Ravanat JL, Berger F, Wion D: **N6-methyladenine: the other methylated base of DNA**. *Bioessays* 2006, **28**:309-315.

21. Neuberger MS, Harris RS, Di Noia J, Petersen-Mahrt SK: **Immunity through DNA deamination**. *Trends Biochem Sci* 2003, **28**:305-312.

22. Borst P, Sabatini R: **Base J: discovery, biosynthesis, and possible functions**. *Annu Rev Microbiol* 2008, **62**:235-251.

23. • Ekanayake DK, Minning T, Weatherly B, Gunasekera K, Nilsson D, Tarleton R, Ochsenreiter T, Sabatini R: **Epigenetic regulation of transcription and virulence in Trypanosoma cruzi by O-linked thymine glucosylation of DNA**. *Mol Cell Biol* 2011, **31**:1690-1700.
Demonstrates the presence of base J near polycistronic units, loss of base J correlates with genome-wide increased gene expression and deficiencies in virulence.

24. Vengrova S, Dalgaard JZ: **The wild-type Schizosaccharomyces pombe mat1 imprint consists of two ribonucleotides**. *EMBO Rep* 2006, **7**:59-65.

25. Roberts RJ, Vincze T, Posfai J, Macelis D: **REBASE – a database for DNA restriction and modification: enzymes, genes and genomes**. *Nucleic Acids Res* 2010, **38**:D234-D236.

26. Wang L, Chen S, Xu T, Taghizadeh K, Wishnok JS, Zhou X, You D, Deng Z, Dedon PC: **Phosphorothioation of DNA in bacteria by dnd genes**. *Nat Chem Biol* 2007, **3**:709-710.

27. • Wang L, Chen S, Vergin KL, Giovannoni SJ, Chan SW, DeMott MS, Taghizadeh K, Cordero OX, Cutler M, Timberlake S *et al.*: **DNA phosphorothioation is widespread and quantized in bacterial genomes**. *Proc Natl Acad Sci USA* 2011, **108**:2963-2968.
Analysis of phosphorothioate modifications in a wide variety of bacteria, consistent with the involvement of PT modifications in a type of restriction-modification system.

28. Marinus MG, Casadesus J: **Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more**. *FEMS Microbiol Rev* 2009, **33**:488-503.

29. Giacomodonato MN, Sarnacki SH, Llana MN, Cerquetti MC: **Dam and its role in pathogenicity of Salmonella enterica**. *J Infect Dev Ctries* 2009, **3**:484-490.

30. Collier J: **Epigenetic regulation of the bacterial cell cycle**. *Curr Opin Microbiol* 2009, **12**:722-729.

31. Collier J, McAdams HH, Shapiro L: **A DNA methylation ratchet governs progression through a bacterial cell cycle**. *Proc Natl Acad Sci USA* 2007, **104**:17111-17116.

32. Srikhanta YN, Fox KL, Jennings MP: **The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes**. *Nat Rev Microbiol* 2010, **8**:196-206.

33. Militello KT, Simon RD, Qureshi M, Maines R, Van Horne ML, Hennick SM, Jayakar SK, Pounder S: **Conservation of Dcm-mediated cytosine DNA methylation in Escherichia coli**. *FEMS Microbiol Lett* 2011 http://dx.doi.org/10.1111/j.1574-6968.2011.02482.

34. Geacintov NE, Broyde S: *The Chemical Biology of DNA Damage*. Wiley-VCH Verlag GmbH & Co. KGaA; 2010.

35. Kelley MR: *DNA Repair in Cancer Therapy: Molecular Targets and Clinical Applications*. Elsevier Science; 2011.

36. Preston BD, Albertson TM, Herr AJ: **DNA replication fidelity and cancer**. *Semin Cancer Biol* 2011, **20**:281-293.

37. Lindahl T, Barnes DE: **Repair of endogenous DNA damage**. *Cold Spring Harb Symp Quant Biol* 2000, **65**:127-133.

38. Beal MF: **Mitochondria take center stage in aging and neurodegeneration**. *Ann Neurol* 2005, **58**:495-505.

39. De Bont R, van Larebeke N: **Endogenous DNA damage in humans: a review of quantitative data**. *Mutagenesis* 2004, **19**:169-185.

40. Maynard S, Schurman SH, Harboe C, de Souza-Pinto NC, Bohr VA: **Base excision repair of oxidative DNA damage and association with cancer and aging**. *Carcinogenesis* 2009, **30**:2-10.

41. Rao KS: **Free radical induced oxidative damage to DNA: relation to brain aging and neurological disorders**. *Indian J Biochem Biophys* 2009, **46**:9-15.

42. Lindahl T: **Instability and decay of the primary structure of DNA**. *Nature* 1993, **362**:709-715.

43. Ward JF: **DNA damage produced by ionizing radiation in mammalian cells: identities, mechanisms of formation, and reparability**. *Prog Nucleic Acid Res Mol Biol* 1988, **35**:95-125.

44. Boorstein RJ, Cummings A Jr, Marenstein DR, Chan MK, Ma Y, Neubert TA, Brown SM, Teebor GW: **Definitive identification of mammalian 5-hydroxymethyluracil DNA N-glycosylase activity as SMUG1**. *J Biol Chem* 2001, **276**:41991-41997.

45. Trzeciak AR, Nyaga SG, Jaruga P, Lohani A, Dizdaroglu M, Evans MK: **Cellular repair of oxidatively induced DNA base lesions is defective in prostate cancer cell lines, PC-3 and DU-145**. *Carcinogenesis* 2004, **25**:1359-1370.

46. Laird PW, Jaenisch R: **DNA methylation and cancer**. *Hum Mol Genet* 1994, **3 Spec No**:1487-1495.

47. Nick McElhinny SA, Watts BE, Kumar D, Watt DL, Lundstrom EB, •• Burgers PM, Johansson E, Chabes A, Kunkel TA: **Abundant ribonucleotide incorporation into DNA by yeast replicative polymerases**. *Proc Natl Acad Sci USA* 2010, **107**:4949-4954.

Characterization of the content of ribonucleotides in yeast genomic DNA, as a result of polymerase-mediated misincorporations during replication. As much as 0.1% of gDNA may consist of rNMPs, perhaps making it the most common noncanonical nucleotides introduced into eukaryotic genomes.

48. Nick McElhinny SA, Kumar D, Clark AB, Watt DL, Watts BE, Lundstrom EB, Johansson E, Chabes A, Kunkel TA: **Genome instability due to ribonucleotide incorporation into DNA**. *Nat Chem Biol* 2010, **6**:774-781.

49. Kumari S, Rastogi RP, Singh KL, Singh SP, Sinha RP: **DNA damage: detection strategies**. *EXCLI J* 2008, **7**:44-62.

50. Nelson M, Raschke E, McClelland M: **Effect of site-specific methylation on restriction endonucleases and DNA modification methyltransferases**. *Nucleic Acids Res* 1993, **21**:3139-3154.

51. Bart A, van Passel MW, van Amsterdam K, van der Ende A: **Direct detection of methylation in genomic DNA**. *Nucleic Acids Res* 2005, **33**:e124.

52. Rao BS, Buckler-White A: **Direct visualization of site-specific and strand-specific DNA methylation patterns in automated DNA sequencing data**. *Nucleic Acids Res* 1998, **26**:2505-2507.

53. Broadbent SE, Balbontin R, Casadesus J, Marinus MG, van der Woude M: **YhdJ, a nonessential CcrM-like DNA methyltransferase of Escherichia coli and Salmonella enterica**. *J Bacteriol* 2007, **189**:4325-4327.

54. Bart A, Pannekoek Y, Dankert J, van der Ende A: **NmeSI restriction-modification system identified by representational difference analysis of a hypervirulent Neisseria meningitidis strain**. *Infect Immun* 2001, **69**:1816-1820.

55. Cohen-Karni D, Xu D, Apone L, Fomenkov A, Sun Z, Davis PJ, • Kinney SR, Yamada-Mabuchi M, Xu SY, Davis T *et al.*: **The MspJI family of modification-dependent restriction endonucleases for epigenetic studies**. *Proc Natl Acad Sci USA* 2011, **108**:11040-11045.

Characterization of a family of restriction enzymes specific for cleaving a short DNA fragment containing a central 5-mC or 5-hmC, followed by second generation sequencing.

56. Zheng Y, Cohen-Karni D, Xu D, Chin HG, Wilson G, Pradhan S, Roberts RJ: **A unique family of Mrr-like modification-dependent restriction endonucleases**. *Nucleic Acids Res* 2010, **38**:5527-5534.

57. Song CX, He C: **The hunt for 5-hydroxymethylcytosine: the** • **sixth base**. *Epigenomics* 2011, **3**:521-523.

Review of targeted enrichment strategies for 5-hmC and subsequent analysis.

58. Matarese F, Carrillo-de Santa Pau E, Stunnenberg HG: **5-Hydroxymethylcytosine: a new kid on the epigenetic block?** *Mol Syst Biol* 2011, **7**:562.

59. Ekanayake DK, Cipriano MJ, Sabatini R: **Telomeric co-localization of the modified base J and contingency genes in the protozoan parasite Trypanosoma cruzi**. *Nucleic Acids Res* 2007, **35**:6367-6377.

60. Genest PA, Ter Riet B, Cijsouw T, van Luenen HG, Borst P: **Telomeric localization of the modified DNA base J in the genome of the protozoan parasite Leishmania**. *Nucleic Acids Res* 2007, **35**:2116-2124.

61. van Leeuwen F, Wijsman ER, Kieft R, van der Marel GA, van Boom JH, Borst P: **Localization of the modified base J in telomeric VSG gene expression sites of Trypanosoma brucei**. *Genes Dev* 1997, **11**:3232-3241.

62. Schadt EE, Turner S, Kasarskis A: **A window into third-generation sequencing**. *Hum Mol Genet* 2010, **19**:R227-R240.

63. Treffer R, Deckert V: **Recent advances in single-molecule sequencing**. *Curr Opin Biotechnol* 2010, **21**:4-11.

64. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW *et al.*: **Single-molecule DNA sequencing of a viral genome**. *Science* 2008, **320**:106-109.

65. Pastor WA, Pape UJ, Huang Y, Henderson HR, Lister R, Ko M, McLoughlin EM, Brudno Y, Mahapatra S, Kapranov P *et al.*: **Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells**. *Nature* 2011, **473**:394-397.

66. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al.*: **Real-time DNA sequencing from single polymerase molecules**. *Science* 2009, **323**:133-138.

67. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW: **Real-time DNA sequencing from single polymerase molecules**. *Methods Enzymol* 2010, **472**:431-455.

68. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, • Clark TA, Korlach J, Turner SW: **Direct detection of DNA methylation during single-molecule, real-time sequencing**. *Nat Methods* 2010, **7**:461-465.

Demonstrates detection of 5-mC, 5-hmC and 6-mA by analysis of the polymerase dynamics during SMRT sequencing.

69. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW: **A flexible and efficient template format for circular consensus sequencing and SNP detection**. *Nucleic Acids Res* 2010, **38**:e159.

70. Song XS, Clark TA, Lu XY, Kislyuk A, Dai Q, Turner SW, He C, • Korlach J: **Sensitive and specific single-molecule sequencing**

of 5-hydroxymethylcytosine. *Nat Methods* 2011 http://dx.doi.org/10.1038/nmeth.1779.
Adaptation of a selective labeling method for 5-hmC, followed by SMRT DNA sequencing, to enable base-resolution and DNA strand specific detection of 5-hmC.

71. • Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, Fomenkov A, Roberts RJ, Korlach J: **Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing**. *Nucleic Acids Res* 2011 http://dx.doi.org/10.1093/nar/gkr1146.
Characterization of the specificities of 16 methyltransferases by SMRT sequencing. Detection of 4-mC, 5-mC, and 6-mA, including three methyltransferases of previously unknown target contexts. Certain methyltransferases are found to display off-target activities.

72. • Clark TA, Spittle KE, Turner SW, Korlach J: **Direct detection and sequencing of damaged DNA bases**. *Genome Integr* 2011, **2**:10.
Survey a different types of DNA damage products for detection by SMRT sequencing using synthetic DNA templates, including the characterization of lesion-specific kinetic signatures.

73. Berman AJ, Kamtekar S, Goodman JL, Lazaro JM, de Vega M, Blanco L, Salas M, Steitz TA: **Structures of phi29 DNA polymerase complexed with substrate: the mechanism of translocation in B-family polymerases**. *EMBO J* 2007, **26**:3494-3505.

74. Johnson SJ, Beese LS: **Structures of mismatch replication errors observed in a DNA polymerase**. *Cell* 2004, **116**:803-816.

75. Venkatesan BM, Bashir R: **Nanopore sensors for nucleic acid analysis**. *Nat Nanotechnol* 2011, **6**:615-624.

76. • Wallace EV, Stoddart D, Heron AJ, Mikhailova E, Maglia G, Donohoe TJ, Bayley H: **Identification of epigenetic DNA modifications with a protein nanopore**. *Chem Commun (Camb)* 2010, **46**:8195-8197.
Discrimination of 5-mC and 5-hmC containing oligonucleotides immobilized in an alpha-hemolysin protein nanopore.

77. • Manrao EA, Derrington IM, Pavlenok M, Niederweis M, Gundlach JH: **Nucleotide discrimination with DNA immobilized in the MspA nanopore**. *PLoS ONE* 2011, **6**:e25723.
Differentiation of cytosine and 5-mC containing DNA using a MspA nanopore.

78. • Wanunu M, Cohen-Karni D, Johnson RR, Fields L, Benner J, Peterman N, Zheng Y, Klein ML, Drndic M: **Discrimination of Methylcytosine from Hydroxymethylcytosine in DNA Molecules**. *J Am Chem Soc* 2011, **133**:486-492.
Molecular dynamics simulations of different forms of cytosine in DNA, and discrimination of 5-mC and 5-hmC containing DNA fragments in solid-state nanopores.

79. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H: **Continuous base identification for single-molecule nanopore DNA sequencing**. *Nat Nanotechnol* 2009, **4**:265-270.

80. Chu J, Gonzalez-Lopez M, Cockroft SL, Amorin M, Ghadiri MR: **Real-time monitoring of DNA polymerase function and stepwise single-nucleotide DNA strand translocation through a protein nanopore**. *Angew Chem Int Ed Engl* 2010, **49**:10106-10109.

81. • Lieberman KR, Cherf GM, Doody MJ, Olasagasti F, Kolodji Y, Akeson M: **Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase**. *J Am Chem Soc* 2010, **132**:17961-17972.
Single-molecule observation of processive replication by phi29 DNA polymerase immobilized at the entrance of an alpha-hemolysin nanopore.

82. McNally B, Singer A, Yu Z, Sun Y, Weng Z, Meller A: **Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays**. *Nano Lett* 2010, **10**:2237-2244.

83. • Cerf A, Alava T, Barton RA, Craighead HG: **Transfer-printing of single DNA molecule arrays on graphene for high-resolution electron imaging and analysis**. *Nano Lett* 2011, **11**:4232-4238.
Description of a method for depositing ordered arrays of individual elongated DNA molecules on single-layer graphene substrates for high-resolution electron beam imaging and electron energy loss spectroscopy analysis, addressing a previous bottleneck towards electron microscopy based sequencing.

84. Cerf A, Cipriany BR, Benitez JJ, Craighead HG: **Single DNA molecule patterning for high-throughput epigenetic mapping**. *Anal Chem* 2011, **83**:8073-8077.

85. Tanaka H, Kawai T: **Partial sequencing of a single DNA molecule with a scanning tunnelling microscope**. *Nat Nanotechnol* 2009, **4**:518-522.

86. • Yi C, Pan T: **Cellular dynamics of RNA modification**. *Acc Chem Res* 2011, **44**:1380-1388.
Review of several RNA modifications that are dynamically controlled in cells, as well as discussion of recently developed methods to study the cellular dynamics of RNA modifications.

87. Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, Zhang X, Vendeix FA, Fabris D, Agris PF: **The RNA modification database, RNAMDB: 2011 update**. *Nucleic Acids Res* 2011, **39**:D195-D201.

88. Motorin Y, Lyko F, Helm M: **5-methylcytosine in RNA: detection, enzymatic formation and biological functions**. *Nucleic Acids Res* 2010, **38**:1415-1430.

89. Saikia M, Dai Q, Decatur WA, Fournier MJ, Piccirilli JA, Pan T: **A systematic, ligation-based approach to study RNA modifications**. *RNA* 2006, **12**:2025-2033.

90. Zhao X, Yu YT: **Detection and quantitation of RNA base modifications**. *RNA* 2004, **10**:996-1002.

91. • Schaefer M, Pollex T, Hanna K, Lyko F: **RNA cytosine methylation analysis by bisulfite sequencing**. *Nucleic Acids Res* 2009, **37**:e12.
Protocol for direct bisulfite conversion of RNA, followed by cDNA conversion and DNA sequencing.

92. Eisenberg E, Li JB, Levanon EY: **Sequence based identification of RNA editing sites**. *RNA Biol* 2010, **7**:248-252.

93. Ozsolak F, Milos PM: **Single-molecule direct RNA sequencing without cDNA synthesis**. *Wiley Interdiscip Rev RNA* 2011, **2**:565-570.

94. • Ozsolak F, Milos PM: **Transcriptome profiling using single-molecule direct RNA sequencing**. *Methods Mol Biol* 2011, **733**:51-61.
Direct RNA sequencing approach without prior conversion to cDNA using the Helicos platform.

95. Vilfan ID, Tsai YC, Clark TA, Wegener J, Dai Q, Yi C, Pan T, Turner SW, Korlach J: **Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription**. *RNA* 2012, submitted for publication.