Bioorganic & Medicinal Chemistry 22 (2014) 4356–4370

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Review

Chemical biology on the genome

Shankar Balasubramanian *

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK
Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, UK

ARTICLE INFO

Article history:
Received 29 April 2014
Accepted 29 April 2014
Available online 15 May 2014

Keywords:
DNA sequencing
DNA structure
Chemically modified bases
G-quadruplex
Epigenetics

ABSTRACT

In this article I discuss studies towards understanding the structure and function of DNA in the context of genomes from the perspective of a chemist. The first area I describe concerns the studies that led to the invention and subsequent development of a method for sequencing DNA on a genome scale at high speed and low cost, now known as Solexa/Illumina sequencing. The second theme will feature the four-stranded DNA structure known as a G-quadruplex with a focus on its fundamental properties, its presence in cellular genomic DNA and the prospects for targeting such a structure in cels with small molecules. The final topic for discussion is naturally occurring chemically modified DNA bases with an emphasis on chemistry for decoding (or sequencing) such modifications in genomic DNA. The genome is a fruitful topic to be further elucidated by the creation and application of chemical approaches.

Contents

Chemistry provides an intellectual framework to conceive molecular explanations for potentially all aspects of nature. It fundamentally addresses the reactivity and recognition properties of molecules, leading to an appreciation of the structural and functional characteristics of molecules, great and small, which is central to the behavior of the complex systems of life. Chemistry also enables the design and creation of practical approaches to explore and unlock the enigmas of nature. The study of biology from the viewpoint of a chemist, sometimes referred to as chemical biology, is of course not a recent phenomenon. There are many profoundly important contributions from chemistry that have transformed areas of biology, such as the elucidation of: the chemical structure of nucleic acids by Todd amongst others; the complex architecture of natural products through chemical synthesis exemplified by Woodward and Corey; protein sequencing by Sanger and Edman, DNA sequencing by Sanger and by Gilbert; peptide synthesis by Merrifield; and oligonucleotide synthesis by Caruthers, to name but a few. A dominant subset of the examples I have chosen exist via tools and methods that are exploited very widely in the life sciences, often without much appreciation of the details and origins of the methodology. Given the enormous and growing focus on the life sciences in pure and applied research, the opportunity and need for chemists to make a difference has never been greater than it is today.

DNA is the fundamental molecule of life, and thus the perfect subject thorough which to explore the chemistry of life. As a Cambridge-based scientist, one is constantly reminded of the importance of DNA given the lasting, local influences of Todd, Crick and Watson, Sanger, Brenner and more recently the International Human Genome Project, co-led by Sulston and colleagues. In spite of the major historical landmarks with respect to our understanding

* Corresponding author. Tel.: +44 1223336347.
E-mail address: sb10031@cam.ac.uk

http://dx.doi.org/10.1016/j.bmc.2014.05.016
0968-0896/© 2014 Elsevier Ltd. All rights reserved.

of DNA, we still have a great deal to learn about this extraordinary molecule. In this special article I will discuss ideas and research from our laboratory on three topics concerning DNA. The first is on a method for decoding DNA that is distinct from the Sanger approach and is currently being used for routine human genome sequencing on a population scale. The second topic relates to the existence and consequence of a quadruple helix, non-Watson–Crick structural feature of DNA, called the G-quadruplex. The third topic concerns the dynamic (epi)genetic alphabet, which provides an expansion of DNA through naturally occurring chemical modifications to the DNA bases.

Early Influences

My parents were immigrants from Madras (now called Chennai) in Southern India, and arrived in the UK in 1967 when I was just 9 months old. After moving around the UK in search of work and a place to settle, we eventually ended up living next to a farm in a rural part of Cheshire, somewhere mid-way between Liverpool and Manchester. My childhood was quite relaxed and my parents raised me in a fashion that enabled me to freely explore my interests and discover passions, without prescribing what I should or should not do. I attended a small village school in Daresbury, founded in 1600, baring a weathervane with the Mad Hatter and Alice to remember and celebrate the famous local, Lewis Carroll. I was fortunate in being taught by a somewhat old-fashioned and hugely inspirational headmaster, Mr. (Brian) Leitch. His relaxed style of imparting wisdom and inspiration, from the history of WW2 to trigonometry, continues to profoundly influence the way I think. At high school I developed a strong interest in mathematics and the physical sciences. I did not pursue biology to any great extent, as I found it to be descriptive and imprecise. I recall my high school chemistry teacher was somewhat disappointed that I wanted to study chemistry at university, as he felt I was 'bright enough' to pursue medicine. I lost touch with that teacher and so do not know whether he is still disappointed with my decision. As an undergraduate, I studied natural sciences at Cambridge University. This was a wonderful way to explore science, in a broad sense, whilst ultimately developing a focus in chemistry. I vividly recall my organic chemistry tutorials with the brilliant teacher Stuart Warren, who would typically start by boiling a kettle of water with tripod and Bunsen burner to provide us with a refreshing cup of Earl Grey tea, before quickly moving on to the matter of disconnecting a challenging complex natural product on the chalkboard. It became clear that I wanted to go much further with chemistry. I remained in Cambridge to pursue a PhD during which my mentor, Chris Abell introduced me to mechanistic enzymology and cultivated my first adventures in chemical biology. At that time, Alan Fersht's relocation to Cambridge helped provide ample local excitement at the chemistry-biology interface. I continued to develop my interests as a post-doc in the lab of Stephen Benkovic at Penn State, where I first encountered molecular biology (which is part of chemistry) and learned to think more deeply about interrogating biology through chemistry and physical methods. I was expecting to stay in the USA, until Alan Fersht persuaded me that I should return to Cambridge to start my independent academic career.

Return to Cambridge

I arrived back in Cambridge just before Christmas in December 1993, at the age of twenty-seven, with considerable energy and motivation, but without a firm idea about what direction I would pursue in my research. It was surreal and also intimidating to be joining so many of my former teachers, now as a colleague. I prepared my first undergraduate lecture course on medicinal chemistry, in which I placed a strong emphasis on DNA structure and function, and DNA interactive drugs. Unbeknown to me at that time, this topic was to become a major focus of my research activities a decade later. As regards initiating a research programme, I was guided by two key principles. My postdoctoral mentor, Steve Benkovic advised that I 'move on' from my previous research and seek out and define a new research direction. I have, since given that same advice to every co-worker from my lab that has pursued an independent research career. And second, the wise words of my former colleague Dudley Williams, who urged me to take time to 'struggle' and define a research problem worthy of spending at least five to ten years on. It actually took me some five to ten years before I managed to define such a research problem (!). I started by exploring a few ideas in the area of combinatorial chemistry and solid phase organic synthesis. This led to a substantial early collaboration with Zeneca (now Astrazeneca) pharmaceuticals together with Chris Abell and was invaluable in establishing my lab. Although I very much enjoyed some of the projects and papers that resulted from the work, I did not stay in this research area for more than five years.

There was a critical turning point in in 1995, when I was elucidating the structural properties of DNA during the process of DNA synthesis on a polymerase, by Förster Resonance Energy Transfer (FRET). I needed access to time-resolved fluorescence spectroscopy. After banging on doors in the Department in search of a suitable laser, I was introduced to David Klenerman, a physical chemist and laser spectroscopy expert, who had joined the department around the same time as I. After some tearoom discussion, we completed the experiment and the work was subsequently published.[1] More importantly, we openly discussed ideas and David stimulated my interest in single molecule fluorescence biophysics, which had just been made possibly at room temperature in solution, and we collectively dreamed up some project ideas. This marked the beginning of what has proven to be a scientifically important relationship that has spanned over fifteen years since. One of my first research grant proposals was a joint one with David Klenerman, which we wrote to the BBSRC in 1995. I found him to be as exploratory as I am in his outlook and sensed that perhaps he too was in search for a ten-year research problem. We proposed to build a single molecule fluorescence system to simply explore fundamental properties of the DNA polymerase. Unbeknown to us at the time, this was to seed our thinking about what was to become an important technology for the future.

Solexa sequencing

In 1952, prior to the publication of the DNA double helix,[2] Todd and Brown made a statement that 'There can be no question of finality about any nucleic acid structure at the present time, since it is clear that there is no available method for determining the nucleotide sequence.'[3] They then proposed a chemical method for controlled, step-wise degradation of RNA from the 3′-end via cycles of oxidation of the 1,2-diol at the terminal ribose to form the di-aldehyde, followed by beta-elimination to release the terminal nucleoside (for characterization) and finally alkaline phosphatase removal of the 3-phosphate. This approach has the potential to identify the nucleotide sequence by step-wise degradation. From discussions I had with Dan Brown many years back, it was evident that in the 1950s there was an attempt to develop this chemistry into a method for polyribonucleic acid sequencing.[4] It was perhaps too far ahead of its time to be fully developed and exploited as the full implications of the double helix and mRNA had not yet been realised and the molecular biology revolution was yet to come. In the late sixties and nineteen seventies, following the elucidation

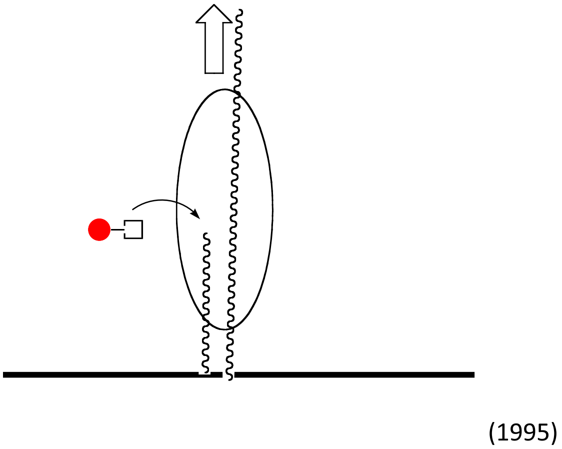


Figure 1. Observing template-directed DNA synthesis by a polymerase by single molecule fluorescence imaging of dye-labeled nucleotides.

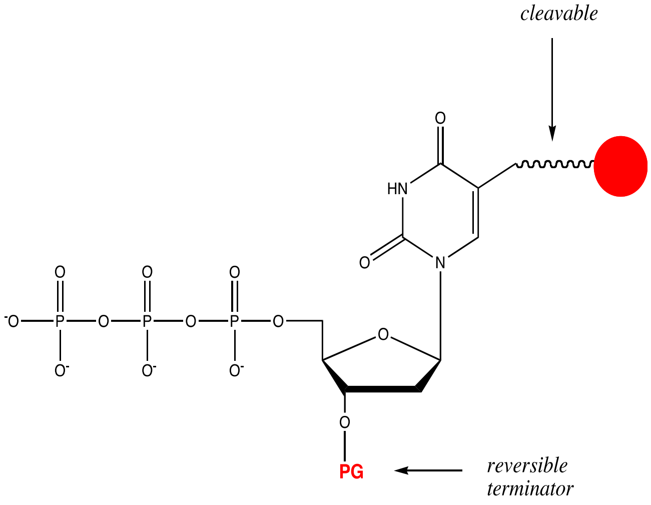


Figure 3. Requirements for a sequencing nucleotide triphosphate.

of the role of mRNA and formulation of the central dogma of molecular biology, methods for sequencing DNA and RNA were being actively explored, and two practical methods were announced in the 1970s. At Harvard, Maxam and Gilbert demonstrated elegant, chemically controlled degradation of DNA in a manner that achieved selective cleavage at sites dependent on the identity of the adjacent base, with the assistance of careful manipulation of the reaction conditions.[5] By radio-labeling DNA, then selective chemical cleavage followed by size dependent separation of the DNA fragments by acrylamide gel electrophoresis it was possible to decode the sequence of DNA. At the same time in Cambridge, England, Fred Sanger and his co-workers demonstrated a method of decoding DNA by systematic termination of DNA-templated synthesis, using a DNA polymerase to incorporate all four building blocks, with any one of the four comprising some terminator monomer, in which the 3′-oxygen, vital for incorporation of the subsequent building block, was absent in the form of a dideoxynucleoside triphosphate.[6] Separating the (labeled) fragments generated by partial termination at G or C or T or A, the sequence of nucleotides could be decoded. I vividly recall using both the Maxam and Gilbert and the Sanger methods manually and from first principles (without using a 'kit') during my post-doctoral research in the early 1990s, when I marveled at how beautifully both methods worked. Ultimately it was the Sanger method that proved more amenable to further improvements and automation into amazing systems that, in due course, would decode the first human genome in the Human Genome Project.

In the mid 1990s, David Klenerman and I, along with our co-workers were making observations to explore how a DNA polymerase carries out template-directed synthesis of DNA, using single molecule fluorescence spectroscopy. We attempted various formats that included free solution,[7] a template immobilized on either a microsphere[8] or on a glass surface.[9] We ultimately employed a total internal reflectance, single molecule microscopy set up to visualise the synthesis of DNA in real time, either by a FRET pair of fluorophores attached to the polymerase and DNA or by incorporation of fluorescently tagged deoxynucleoside triphosphates during extension of an immobilized DNA molecules (Fig. 1). We then came across a most interesting paper by Richard Keller[10] describing a beautiful concept for sequencing a single molecule of DNA by first differentially labeling G, C, A and T during synthesis with a polymerase, followed by the systematic degradation of the DNA from one end using a nuclease enzyme. Whilst this concept appeared to be a challenge to reduce to practice, it inspired us to

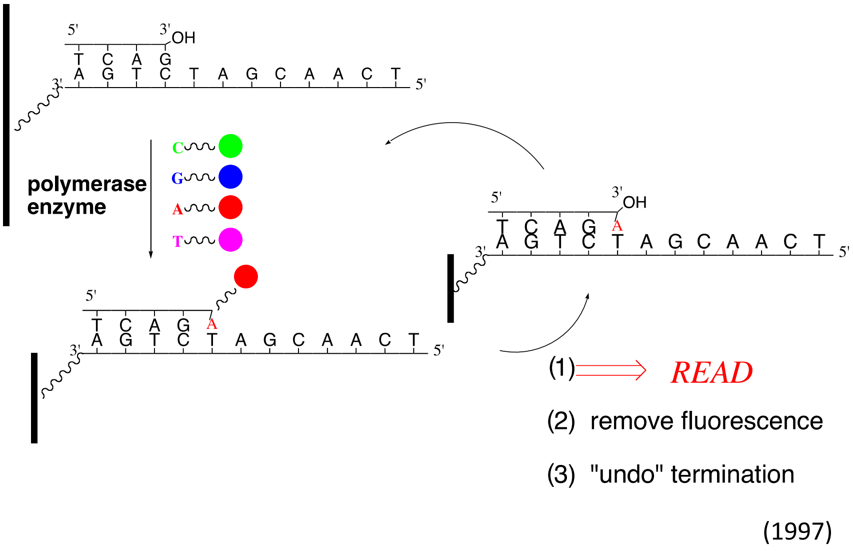
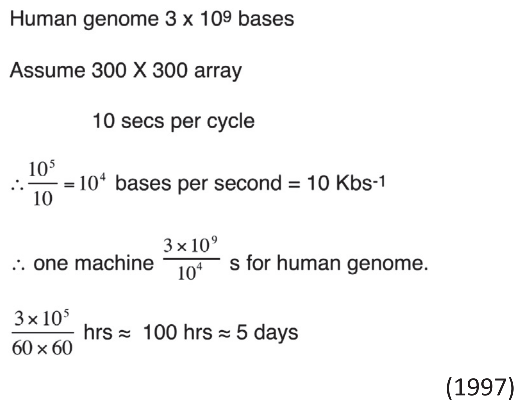


Figure 2. Solid phase DNA sequencing using dye-labeled reversible terminators.



Human genome 3 x 10^9 bases

Assume 300 X 300 array

10 secs per cycle

$\therefore \frac{10^5}{10} = 10^4$ bases per second = 10 Kbs^{-1}

$\therefore$ one machine $\frac{3\times10^9}{10^4}$ s for human genome.

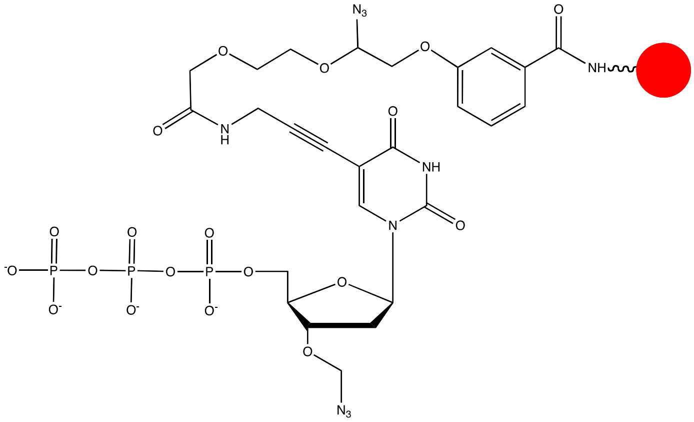
$\frac{3\times10^5}{60\times60}$ hrs ≈ 100 hrs ≈ 5 days

(1997)

Figure 4. An early calculation that predicted a pathway to a system capable of sequencing several billion bases (Gigabases) of DNA

think about whether we could decode DNA. Indeed, it struck us that by watching single nucleotides incorporated by a polymerase, templated by a pre-existing copy of DNA, we were in a sense decoding the template strand of DNA. In a series of discussion at our local pub, the Panton Arms, we began to see practical ways of adapting our experiment to introduce four-colour coding of the DNA bases whereby incorporation of a given base could be marked by a fluorophore during each cycle of decoding (Fig. 2). The scheme necessitated absolute chemical control for the stepwise incorporation of labeled nucleotides by the polymerase. By mobilizing the sample DNA template to be sequenced and hybridizing a DNA primer, one could therefore carry out cycles of stepwise synthesis whereby after each incorporation one could stop the reaction and read-off the identity of the incorporated base by imaging the colour of the fluorescence. This would be achieved, not by the use of dideoxy-terminators—à la Sanger—but by installing a protecting group at the 3′-oxygen of the incoming deoxynucleotide that could be subsequently removed using orthogonal cleavage chemistry. At the same time, a readable fluorophore tag could be attached in a benign way, via the non-Watson/Crick edges of the bases via a chemically cleavable linker (Fig. 3). We anticipated the need to re-engineer part of the polymerase active site, in particular to accommodate the 3′-O-protecting group. Overall, whilst technically challenging, this seemed eminently tractable. So why sequence DNA on solid phase? By creating an immobilized array of DNA sample molecules one could decode huge numbers of DNA samples on a chip in parallel. An array of single molecules of DNA sample could be generated by fragmenting genomic DNA then immobilizing the fragments on a surface, at suitably high dilution, such that fragments each occupied optically resolvable space. Thus, one could relatively simply format a huge number of DNA fragments on a surface for decoding by massively parallel solid phase sequencing. An early, 'back of the envelope' calculation (Fig. 4) suggested such a scheme had the potential to sequence about a billion bases of DNA per experiment (a target that was ultimately exceeded by a large margin—see later). Why would one want to decode a billion bases of DNA? In early 1998 we paid a visit to the nearby Wellcome Trust Sanger Institute where, under the direction of John Sulston, there was an enormous effort to decode the first human genome, as part of the international human genome project, using state-of-the-art capillary sequencers and Sanger sequencing. There we met three leaders of the project, David Bentley, Richard Durbin and Jane Rogers, in a small tearoom. During the discussions we learned about the scale of the International Human Genome Project that comprised a number of major genome centers and thousands of people working with hundreds of sequencing machines. In the Sanger Institute, there was a digital notice board in the main entrance constantly punching out sequence data as it was being generated, in real time. It was quite evident that this ambitious project would one day come to a completion, thereby providing a single human genome reference sequence of just over 3 billion bases. It was also apparent that this would constitute merely the beginning, given that every human and indeed every organism has a unique genome. Many more human genomes would need to be decoded to begin to truly understand the genetic basis for human variation, predispositions to disease and indeed genetically caused diseases. Clearly, it was out of the question to re-run a human genome project scale experiment for each subsequent human genome—there would be a need for a complete change in the way we decode DNA. Our hosts voiced considerable enthusiasm for our vision to create a new method of DNA sequencing, to routinely decode whole genomes for the post-human genome project phase that was yet to come. Armed with the certainty that such an invention would be 'useful' and the confidence that we could technically reduce the approach to practice, we put our minds to the driving the research phase of this project. Proof of concept experiments continued in our lab with our existing expertise in organic and physical chemistry. We ultimately set up a small biotech company as a vehicle to mobilize the resource and interdisciplinary expertise (including chemists, molecular biologists, physicists, engineers and bioinformations) that we needed to fully reduce this idea to practice in the form of a commercial system that could be put into the hands of geneticists. We raised some initial investment from a bold venture capital company called Abingworth management who saw it as a high-risk project, but recognized that it could lead to a paradigm shift. A company, which we named 'Solexa' was created in the summer of 1998, however all the technical work continued in the University Chemical Laboratory at Lensfield Road for the next two years, where we completed the proof of concept work necessary to justify scaling up via an external facility towards producing a full commercial system.

There were many technical challenges that needed to be overcome, which I do not have space to describe in detail here.[11–13] The chemistry that prevailed with the first generation sequencing nucleotides exploited the pioneering work of Hermann Staudinger,[14,15] by adapting the reduction of azides by phosphines. The 3′-oxygen was reversibly protected via an azido-methyl group (Fig. 5) which upon reaction with a water-soluble phosphine, such as tricarboxyethyl phosphine (TCEP), to reduces the azide to unmask the hemiaminal which is spontaneously hydrolyzed in water to reveal the 3′-oxygen. Similarly the fluorescent tag is attached via C-5 of the pyrimidine C and T, and via an N-7 deaza substituted position on the purines G and A, without perturbing the Watson and Crick hydrogen bonding required for DNA-templated synthesis. The linker to the dye comprises an azido alkyl ether linkage enabling the simultaneous removal of fluorophore and 3′-protecting groups, after each cycle of incorporation. The polymerase required some protein engineering of the active site to accommodate the 3′-blocking group, which was helped by the relatively small size of the azido-methyl moiety. An important addition to the original concept was to amplify each DNA molecule on the single molecule sample array to form hundreds of identical copies, at the same site on the array, thereby enhancing the signal at each cycle to enable detection with lower costs cameras and to also eliminate stochastic modes of (single molecule) error generation leading to improved sequencing accuracy. Thus, although the DNA sample array was formed from single molecules, the sequencing reaction was ultimately carried out on amplified copies of each molecule. The amplification method was based on an elegant idea from Kawashima and co-workers[16] for copying immobilized DNA molecules whilst preserving spatial integrity. The very first genome to be sequenced by the method was the relatively small (5386 base pairs) genome of bacteriophage Phi-X-174 in 2005,



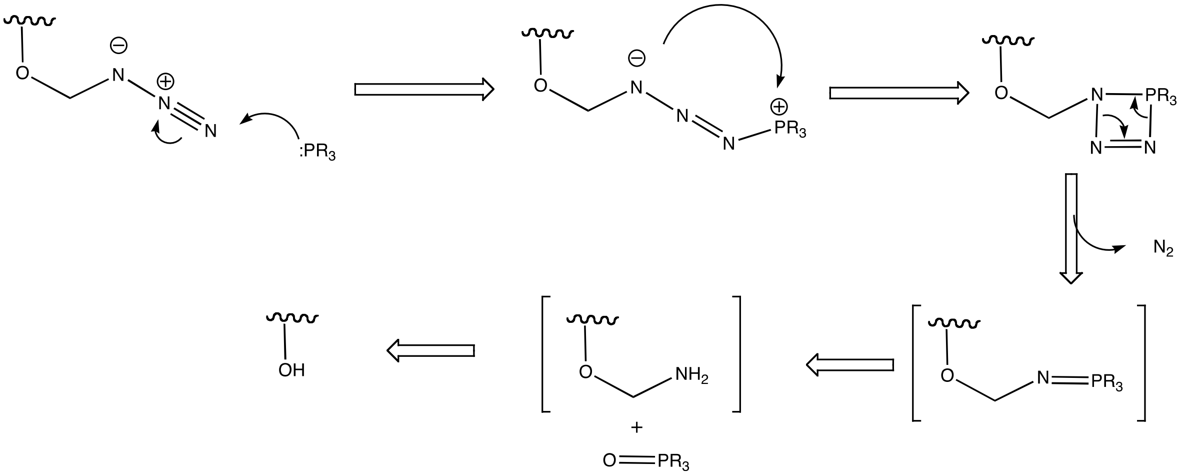


Figure 5. Adaptation of Staudinger reduction for orthogonal protection and cleavage.

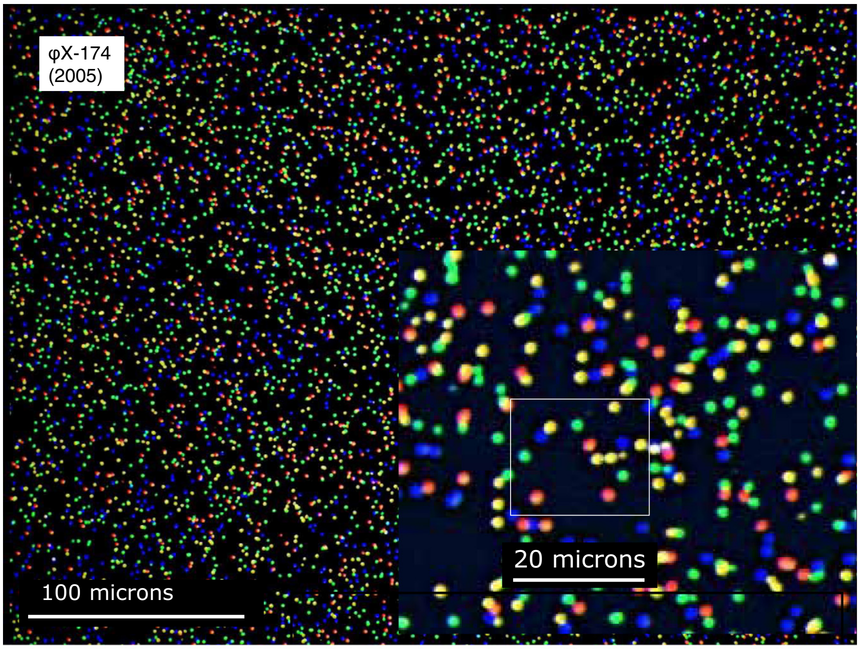


Figure 6. A partial image of the surface during one step of the cycle. Each spot is a cluster of DNA fragments of identical sequence and the colour indicates the identity of the base at that step.

which was also the first whole genome to ever be sequenced by Sanger in 1977[17] (Fig. 6). In 2006, the first instrument to employ this method, called the Genome Analyzer was launched by Solexa and delivered to genome centers for early adoption. In January 2007 the system sequenced one billion bases of human DNA accurately in a single experiment realizing the aspirational target set in 1997 (Fig. 4). Illumina acquired the Solexa sequencing technology and company later that month, and the methodology has been subsequently subjected to continued innovations and improvements within Illumina. Currently the technology is being routinely exploited in a suite of sequencing systems that range from table-top sequencers, capable of delivering billions of bases of DNA in just a few hours, to the latest system of large scale sequencers, sequencing tens of thousands of human genomes a year at below a $1000 price tag per genome.[18,19] The overall increase in capacity and speed per sequencing system, and cost reduction (per genome), in the period since 1997 has been about a million-fold. Today a small lab, such as my own, can have a bench-top sequencer with a capacity considerably greater than the global sequencing capacity as it stood in 1997 when we started the project.

While our primary motivation was to provide a method capable of fast, low-cost, accurate human genome sequencing to help advance the understanding of the genome for human health, the number of basic research applications of high-throughput sequencing that users subsequently developed came as something of a surprise. Such applications are a numerous and growing family of sequencing 'apps' that typically end in '...-seq', such as high resolution mapping of chromatin proteins, or Chromatin Immunoprecipitation (ChIP-seq), or quantitative RNA sequencing (RNA-seq) for identifying transcripts and measuring their levels. The main advantage being the ability to ask open questions without presuming (or limiting) the possible answers, whereby the answers can be ascertained from an output in the form of digitally quantified DNA sequence identifiers. This contrasts with the use of PCR or DNA microarrays, where one must pre-define a set of solutions (sequences) prior to carrying out the experiment.

Marra and co-workers described an early example of the potential clinical benefits of whole human genome sequencing.[20] Here, the detailed characterisation of the mutating cancer genome of a patient provided the insights necessary to identify the dominant cancer-driver pathways as the cancer 'evolved'. This enabled clinicians to make judicious choice of effective therapeutic agents as the evolving tumour became resistant to the previous treatment. Whilst the improved management exhibited in this challenging early case did not lead to a 'cure', the case revealed the genetic evolution of a cancer under pressure from selected drugs. Accumulated knowledge from further such studies may ultimately provide intelligence on commonly emerging pathways of cancer evolution that suggest combination therapies, analogous to what has been achieved for the treatment of HIV. The Cancer Genome Project is sequencing tens of thousands of cancers and has already determined important major genetic determinants/signatures of cancers.[21] Another area of medicine where genome sequence analysis is having an impact is in so-called rare diseases that collectively affect one in seventeen of us, primarily manifest in young children and are predominantly genetic in origin.[22] There are early examples of the positive medical impact of whole genome sequencing of newborns/infants (and their parents for comparative analysis) to provide rapid diagnosis in clinics that have started to operate genome sequencing routinely.[23,24] The UK government and National Health Service (NHS) is leading a pioneering initiative to sequence the genomes of 100,000 patients and link the information to the clinical patient data.[25] This will serve to provide a substantial, systematic curated infrastructure to stimulate further advances in the practical application of genome sequencing to medicine. The application of genomic science to medicine will be vigorously explored during the next decade during which time we will start to observe the extent to which it can influence the practice of medicine and the development of new therapeutic agents.

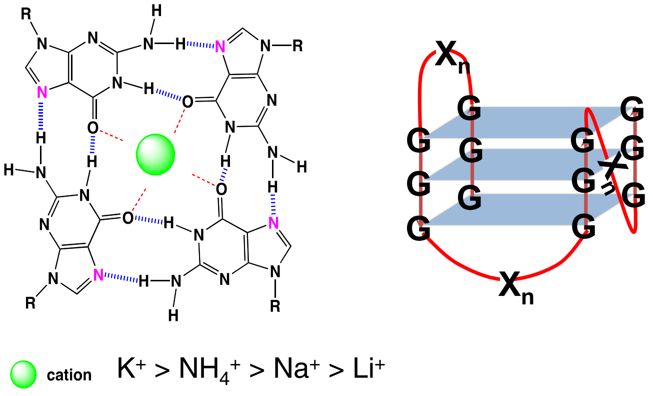


Figure 7. The G-tetrad motif stabilized by cations (in order of preference) and a schematic of an intramolecular G-quadruplex.

The advancement from concept to a working sequencing technology and the onset of clinical implementation has taken less than two decades, and has already exceeded some of the expectations I had set in 1997. In closing this section, I would like to express that the inventions and ideas that led to Solexa/Illumina sequencing came as an unintended consequence of blue skies research funded by the BBSRC of the UK. The subsequent commercial technology development was financed by venture capital investment. It is absolutely essential that curiosity-driven basic research be strongly supported for science and for innovation.

G-quadruplexes in DNA

Subsequent to the discovery of the folded B-form DNA double helix structure it has become evident that DNA is also predisposed to non-B-DNA conformations and may be structurally dynamic. Alternative structures that have been investigated include the reverse left-handed helix Z-DNA and the triple helix (triplex) in which a third DNA strand binds to the major groove of the double helix via Hoogsteen hydrogen bonds. While the physical evidence for the formation of such alternative structures in vitro has been undisputable, their existence and roles in the context of living systems are topics for continued research and in some cases controversy.

A G-quadruplex is a structural motif whose core comprises four guanines arranged in a tetragonal fashion via hydrogen-bonded interactions at the Watson Crick and Hoogsteen edges of the G base, as shown in Figure 7. The lone pair of O-6 of each guanine points to the interior of the G-tetrad motif, generating a central region of negative electrostatic potential that nicely forms a binding site for cations. The details of the G-tetrad structural motif were first elucidated using X-ray fibre diffraction studies by Davis and co-workers[26] on higher-order structures formed by guanylic acid derivatives, that a paper by Bang first alluded to at the beginning of the last century.[27] These tetrads stack on top of each other to form defined three-dimensional structures, further stabilised by cations. Sen and Gilbert demonstrated that the repeat sequence found in telomeres could self assemble to form stable four-stranded G-quadruplex structures.[28] This was also confirmed by other groups that included Klug[29] and Blackburn,[30] leading to the suggestion that the G-tetrad motif might actually exist in nature and play some kind of role in biology. My interest in the G-quadruplex structural motif started in the late 1990s, as a result of our studies on the molecular mechanism of the enzyme human telomerase, responsible for the synthesis of telomeres at the chromosome ends. Given that the G-rich telomeric strand of DNA synthesized by telomerase was naturally predisposed to forming G-quadruplex structures at physiological potassium concentration, there was much speculation about the potential for this structure to have some role associated with the mechanism of telomere extension. My coworkers and I were intrigued by a series of elegant studies from Hurley's group, in which they seemed to have some biophysical and mechanistic evidence that G-quadruplexes were somehow involved in the telomerase mechanism, at least in vitro.[31,32] Another paper from that time that stimulated my interest came from the laboratories of Neidle and Hurley, demonstrating that a small molecule anthraquinone derivative recognized the G-quadruplex structural motif and was able to inhibit the extension mechanism of human telomerase by trapping the DNA substrate.[33] Given the interest in telomerase as a potential anti-cancer target, as it is up regulated in almost every cancer type studied, this study linked G-quadruplex structures with potential strategies for anti-cancer therapeutics. My own interest was focused on the potential role of G-quadruplex structure in the mechanism of telomerase. In collaboration with the lab of Yen Choo, we employed phage display and engineered zinc finger libraries to generate a zinc finger protein (Gq1), that recognised G-quadruplex folded DNA, rather than the DNA double helix.[34] Armed with this new structure-specific probe molecule, we demonstrated inhibition of the extension mechanism for human telomerase supporting the link between G-quadruplex structure telomeres and telomerase.[35] A central question to address was whether this DNA structure existed in a human cell, which we first attempted to address in 2000 by expressing a green florescent protein (GFP) fusion with GQ1 protein in human HeLa cells to see whether fluorescence imaging might reveal an accumulation of Gq1 within nuclear genomic DNA. Whilst this was without doubt an important question our experimental design lacked the sensitivity to draw any clear conclusions. We were to revisit this same important question more than a decade later (see later). Our efforts then focused on the design, synthesis and evaluation of small organic molecules that target the G-quadruplex structural motif. Structural insights in the field have been largely built on a relatively small number of high-resolution G-quadruplex structures and co-structures with small molecules, mostly generated from the laboratories of Stephen Neidle and Dinshaw Patel.[36,37] All intramolecular G-quadruplexes comprise a core of stacked G-tetrads around which single-stranded loops are folded (Fig. 7). The polarity, length, topology and sequence composition of the three loops can vary substantially from one G-quadruplex to another, providing scope for differential molecular recognition by proteins and by small organic molecules. The paradigm for small molecule recognition of G-quadruplexes has generally been based on end-stacking of the core scaffold onto one of the terminal G-tetrads, whilst targeting specific interactions with the loop residues and also the cavities created between the loops and the grooves. Using these design principles, my laboratory and many others have generated families of scaffolds that showed molecular selectivity for G-quadruplexes as compared to the double helix. Examples of some of the G-quadruplex-selective scaffolds we have described are shown in the Figure 8.[38–45] Such molecules do not intercalate into the base stack of a double helix, due to a high kinetic barrier to threading into the base stack of a double helix as compared to docking onto the terminal exposed tetrad of the G-quadruplex. Whilst the formation of co-crystal structures has

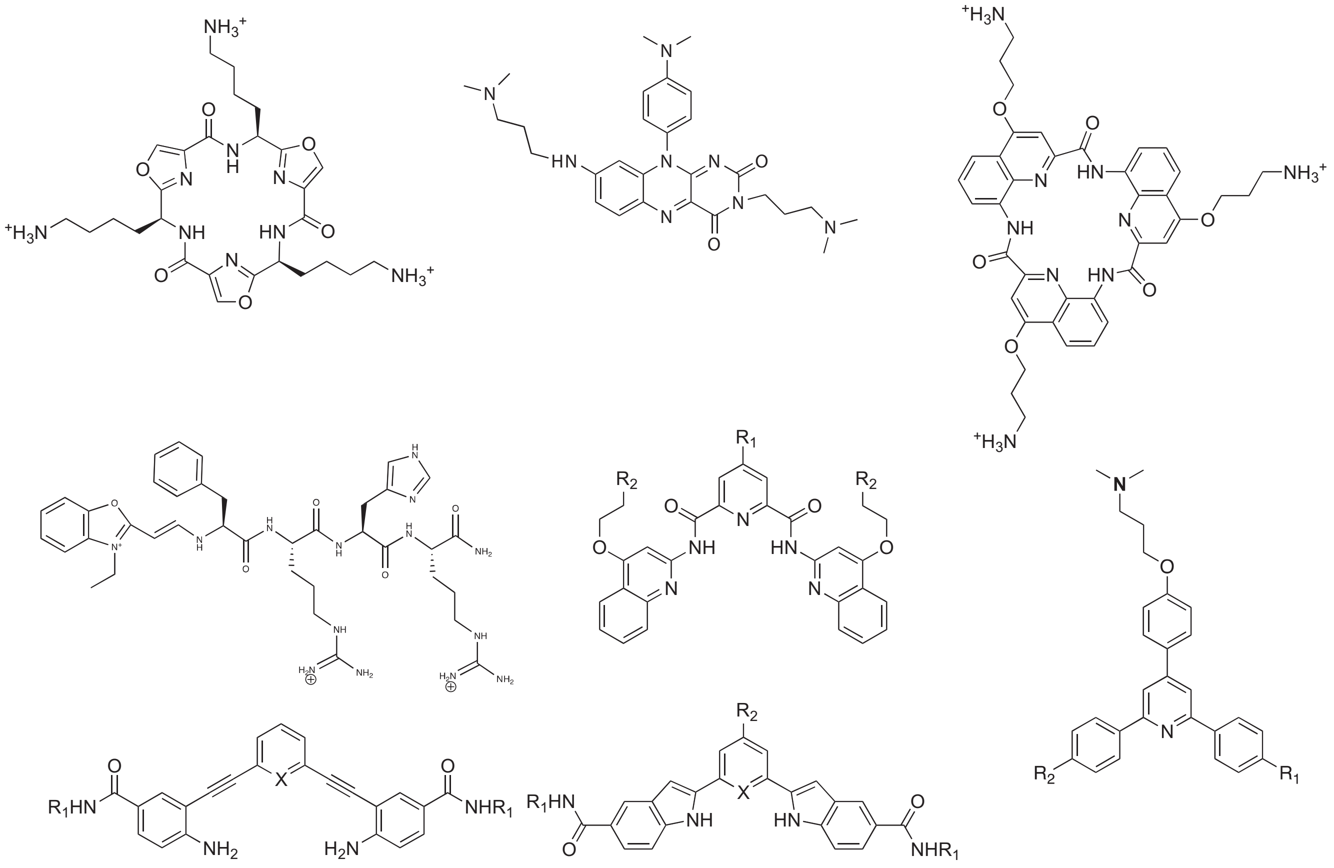


Figure 8. Examples of G-quadruplex ligands we have studied.

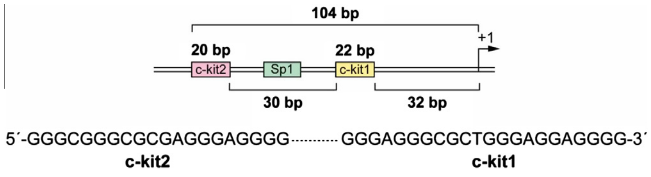


Figure 9. Two quadruplex motifs in the c-kit promoter.

been challenging, with Stephen Neidle we have exploited molecular modeling and simulations to help refine some of these molecules and gain insights into the key interactions of constituents with loops and grooves. These ligands showed very high selectivity for G-quadruplexes versus duplex, and discriminated between G-quadruplexes by more than 10-fold, with low nanomolar binding affinities by surface plasmon resonance measurement. Such studies from our lab and others in the field have generated excellent chemical probe molecules to explore G-quadruplexes in living systems.

In addition to the links between telomeres and G-quadruplexes, several papers had noted that G-rich sequences capable of forming G-quadruplexes were found in the promoter regions (that regulate transcription) hinting at their potential involvement in regulating genes.[46] In particular, an oligonucleotide comprising a G-rich region in the promoter of the proto-oncogene *c-MYC* was shown to fold into a stable G-quadruplex[47] and subsequently a provocative paper from Lawrence Hurley's lab suggested that the G-quadruplex forming sequence in the promoter of *c-MYC* could be targeted by a small molecule to down regulate the expression of *c-MYC*.[48] This hypothesis stimulated my interest in thinking beyond G-quadruplexes in telomeres. In collaboration with Stephen Neidle, we identified two G-quadruplex motifs in the promoter of the protooncogene *c-KIT*, having the sequences shown in Figure 9. The structures of the G-quadruplexes formed by these individual sequences were ultimately solved by 2-D NMR spectroscopy and by X-ray crystallography[49–51] and we later even found a third G-quadruplex forming sequence in the intervening region,[52] clearly the situation was more complex than we first thought! We were able to demonstrate that small molecules shown to bind to the c-kit G-quadruplex(es) in biophysical experiments could lower the expression of *c-KIT* in cancer cell lines; a correlation that supported the working hypothesis. There have subsequently been a good number of papers in the field that describe experiments on various genes to explore the G-quadruplex promoter transcription hypothesis, and the topic has been recently reviewed.[53,54] This remains to be a compelling and interesting hypothesis, however as much of the data has been correlative, there is room for further experimentation to provide more explicit mechanistic proof in a cellular (or in vivo) context.

Having considered the increasing number of G-quadruplex motifs from specific genes and locations in the genome that were appearing in the literature, we started to consider systematic ways of revealing the full potential of genomes to form these structures. On consideration of the available biophysical information on primary sequences of DNA that could fold into stable G-quadruplexes, we formulated a simple algorithm describing of G-quadruplex-forming sequences, which could be employed to computationally search the vastly increasing genome sequence data that is available. The sequence motif we employed was $G_{3-6}N_{1-7}$ $G_{3-6}N_{1-7}$ $G_{3-6}N_{1-7}$ G_{3-6} and the search algorithm was named *Quadparser*.[55] It was clear at that time, and even more so now, that there are sequences that fall outside this definition that also form stable G-quadruplexes. We found close to four hundred thousand independent putative G-quadruplex forming sequence motifs in the human genome. In a parallel study conducted independently, Stephen Neidle came up with a very similar figure[56] confirming that there was considerable potential for such structures to form throughout the human genome. Such approaches from our lab and others have subsequently been used extensively to mine genome sequence data for G-quadruplexes across many organisms. We carried out some focused studies to reveal enrichment in G-quadruplex motifs in gene promoters (transcription)[57] and also in the 5′-untranslated (non-coding) regions of mRNAs[58] consistent with a potential role in gene regulation.

Do quadruplexes really exist in cells?

Despite all the high-quality chemical and biophysical data on G-quadruplexes and their complexes with ligands that the field had collectively accumulated, we remained critical of the lack of explicit evidence for such structures in the nucleus of mammalian cells. It was noteworthy that Plückthun and co-workers had used a G-quadruplex specific antibody to 'light up' telomeres at the chromosome ends of the ciliate *Stylonychia lemnae*[59] and that subsequently Rhodes, Lipps and co-workers employed the same antibody to show that in the ciliate, G-quadruplex formation at telomeres was directly coupled to the biochemistry of cell cycle.[60,61] However, there was no such evidence in the DNA of mammalian cells and a lack of evidence for G-quadruplexes outside the telomeres of any organism.

Our first exploration into providing more explicit evidence of G-quadruplex formation involved applying chemical biology on a genome scale. This started with a small heterocyclic molecule from our lab named pyridostatin (PDS)[43] (owing to its pyridine core and cytostatic properties with cancer cell lines). This molecule was a potent (~200 nM) stabilizer of G-quadruplex DNA (and RNA) that exhibited a strong preference for G-quadruplexes as a broad structural class without measurable interaction with double helical DNA. The PDS scaffold ultimately proved to be sufficiently versatile to be utilized in a variety of G-quadruplex experiments in due course. As part of a collaboration with Jean-François Riou, we showed that PDS could displace telomere binding protein human protection of telomeres 1 (h-pot1) from the telomere ends in human cancer cells which we hypothesized was via a mechanism that involved sequestering the G-rich telomere strand of DNA as a G-quadruplex during replication.[43] This observation was accompanied by a DNA damage event at the telomeres, visualized by an antibody recognizing phosphorylation of histone γH2AX. There was also a hint of some damage events outside the telomeres in other parts of chromosomes. Some years later we followed up these observations by collaborating with a local DNA damage expert, Steve Jackson, in Cambridge, and in these studies we demonstrated that PDS was indeed causing DNA double strand breaks in human cancer cells, in a manner that was functionally-dependent on replication and active transcription.[62] This was consistent with the view that DNA is predisposed towards G-quadruplex formation in functionally active states when the DNA strands are separated. We then set out to carry out an ambitious experiment in which we physically mapped the location of double strand breaks formed within the genomic DNA of cancer cells, as a direct result of treatment with our G-quadruplex ligand, PDS. This was carried out by enriching DNA fragments that we had chemically cross-linked to a DNA damage response marker phosphorylated histone γH2AX, using an antibody (known as a Chromatin Immuniprecipitation Sequencing, or ChIP-seq). The isolated fragments were then sequenced at great depth using Solexa/Illumina sequencing and the sequenced regions aligned against the human genome (Fig. 10). Indeed, this proved to be our first application of the

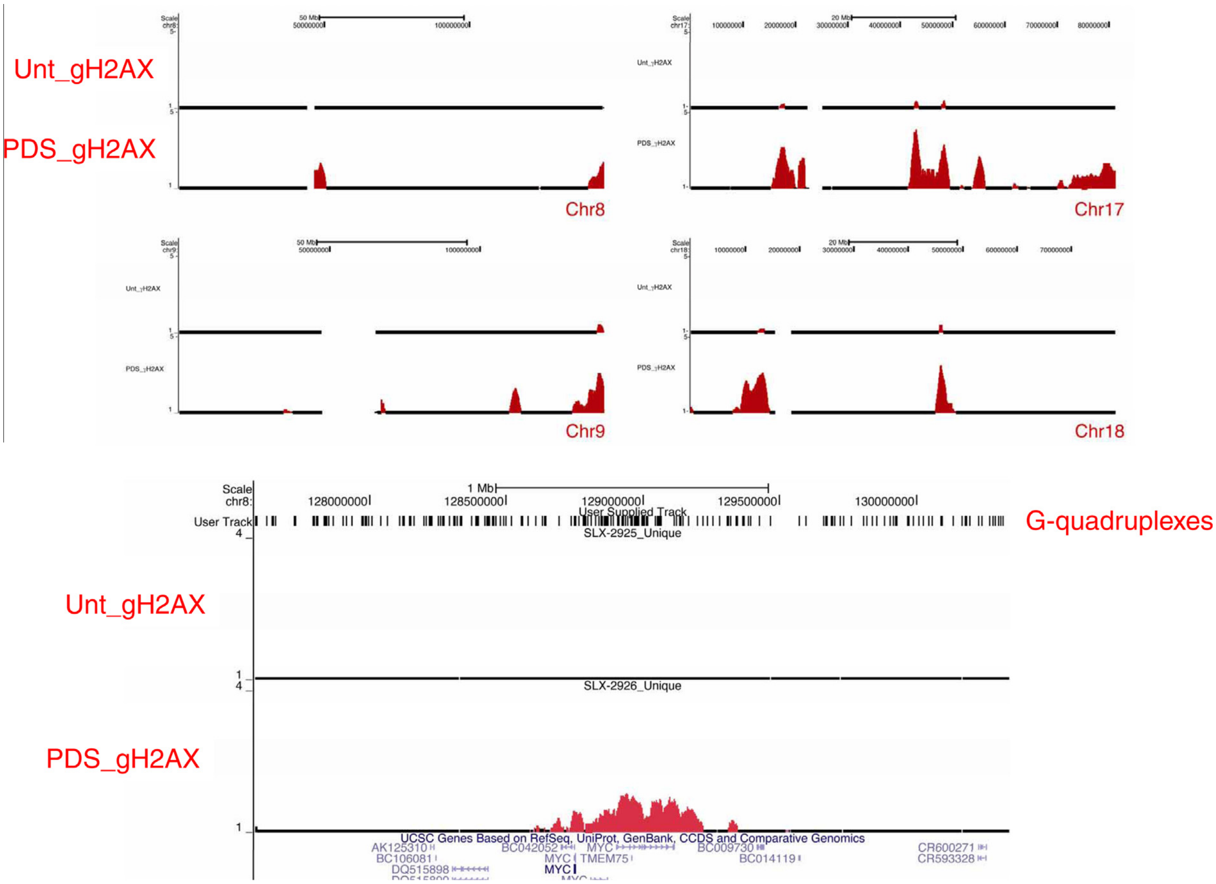


Figure 10. Examples of peaks in chromosomes (upper) and an expansion of a peak in the MYC gene (lower) showing alignment with predicted G-quadruplexes (vertical black bars).

technology that we'd played a role in creating some ten years earlier (!). The outcome of this experiment was the discovery of a surprisingly small number (~fifty) of 'hot spots' in the genome where PDS elicited a functional response. These sites were then traced to specific DNA sequences and upon further analysis using the G-quadruplex predictor *Quadparser*, these peaks were found to map onto G-quadruplex forming regions of the genomes providing gratifying evidence that our ligand was indeed acting at quadruplex forming sites in genomic DNA. We identified a number of cancer-related genes that had been physically targeted by pyridostatin, and noted that one in particular *SRC*, was one of the most prominent genes targeted by pyridostatin. Within the body of the *SRC* gene we identified more than 20 independent quadruplex motifs of which 23 were confirmed to fold into stable G-quadruplexes as judged by CD, UV and ^{1}H NMR Spectroscopy. Furthermore, the transcription of *SRC* was silenced by the action of PDS. Using a metastatic cancer cell model, in which cell motility was driven by over expression of *SRC*, we demonstrated reversal of the metastatic phenotype, using a wound-healing assay.[62] At that time these experiments collectively provided the most compelling data we had seen in our lab in support of G-quadruplex formation and targeting in genomic DNA of human cells. This was shorting followed up by two orthogonal studies each addressing G-quadruplex formation in genomic DNA. We collaborated with one of the pioneers of therapeutic antibody engineering, John McCafferty, to employ phage display and antibody selection to generate high affinity and highly selective single chain antibodies that broadly recognized DNA G-quadruplexes. One such antibody, BG4 was found to bind a number of G-quadruplexes, formed from oligonucleotides based on genomic sequences, with low nanomolar K_d.[63] This antibody allowed us to visualize G-quadruplex formation in the DNA within the nucleus of human cells (Fig. 11). The fluorescence signal generated by immunostaining with BG4 allowed us to quantitate, in a relative sense, the influence of factors that increased or decreased the density of G-quadruplexes in human cells. We made two important subsequent observations. First, that quadruplex density increased naturally during the S phase of the cell

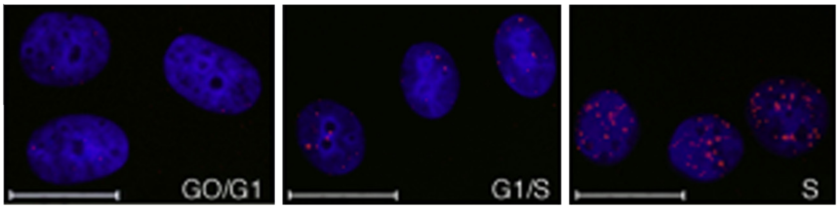


Figure 11. Visualising DNA G-quadruplex structures in the nuclei of human cells.

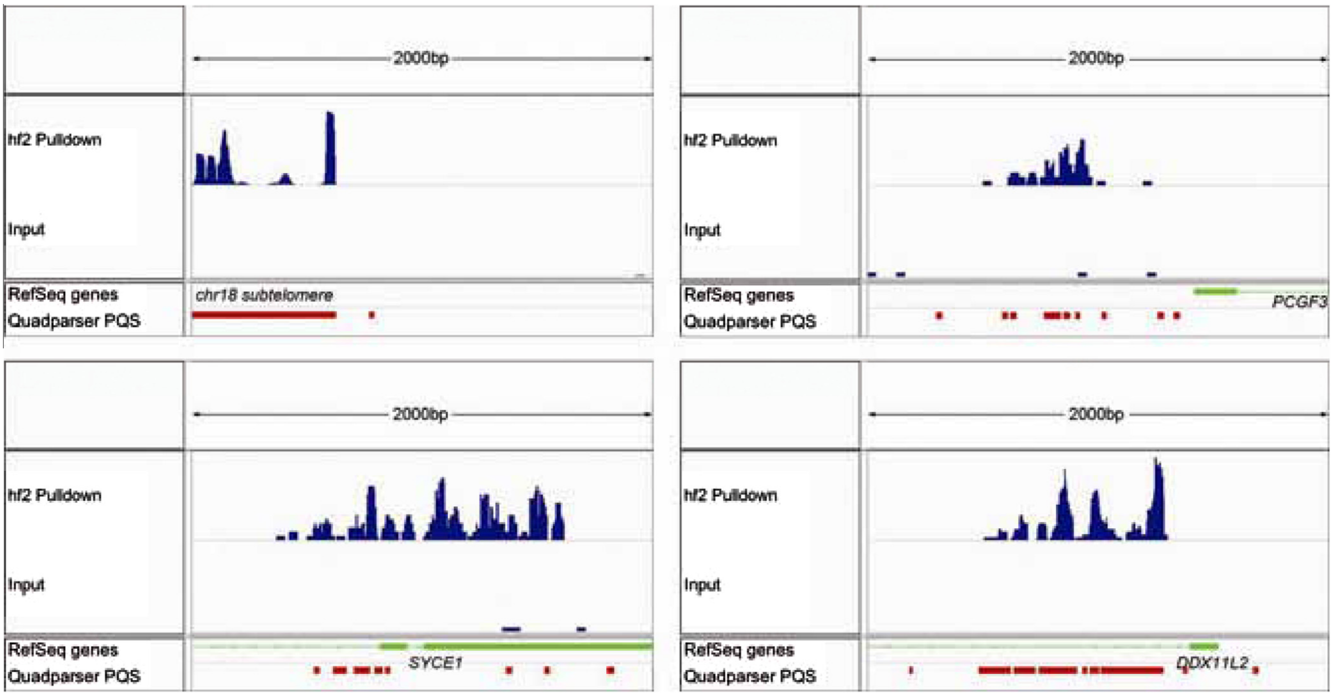


Figure 12. Peaks identified by antibody pulldown correspond to predicted G-quadruplex sequences.

cycle consistent with G-quadruplex formation being naturally high when DNA is being replicated. Second, that addition of a G-quadruplex ligand (PDS) to the tissue culture medium led to an elevation in the observed quadruplex density consistent with the trapping of G-quadruplex structures by the small molecule ligand.[63] At about the same time, in parallel experiments we extracted, then mechanically fragmented genomic DNA, derived from MCF7 breast cancer cells, and used a different G-quadruplex-recognising antibody, hf2,[64] to bind and precipitate DNA fragments in which a G-quadruplex structure was present. The hf2-enriched DNA fragments were subjected to next generation sequencing the enriched peaks were aligned against predicted G-quadruplexes[56] (Fig. 12), to show a high correlation. The data provided experimental evidence of stable G-quadruplexes at defined sites in genomic DNA. For a selection of genes in which we had mapped G-quadruplexes we showed that administration of a G-quadruplex-targeted ligand also altered the transcript levels in those genes, compared to control genes that showed the absence of quadruplexes. So, in a relatively short space of time, we became convinced by this new data that quadruplexes do exist in human cells and that quadruplexes were indeed being physically targeted by representative members of our small molecule quadruplex ligands. Equipped with compelling new data to accompany numerous, credible hypotheses for the mechanistic effects of G-quadruplexes on genome function, the aspiration to generate efficacious small molecules for therapeutics would now appear to be a reasonable goal to pursue over the next decade.

Decoding modified bases in DNA

During the quest to elucidate the chemical structure of DNA, it became apparent that there exists a remarkably broad chemical repertoire of naturally occurring derivatives of DNA nucleobases. Many non-canonical nucleobases were initially discovered in the genomic DNA of phage and microorganisms that showed a rich chemical diversity of functionality primarily on the non-Watson–Crick edges of the DNA bases (Fig. 13). For all such bases the enzyme-mediated modifications have the potential to dynamically alter the fundamental properties of DNA including the structure, duplex stability and molecular recognition, particularly of the major groove. There is much to be elucidated about in which genomes such modifications occur, where (in the genome) and when they occur, and why? Perhaps one of the most chemically intriguing classes of modification being the glucosylated hydroxymethyl derivatives of cytosine and uracil found in some T-even bacteriophage[65,66] and in *Trypanosoma brucei*[67] (Fig. 13). While the advent of Sanger sequencing chemistry in the 1970s prompted an emphasis on research involving decoding the sequence of G, C, A and T in genomes, there has been rather limited exploration of non-canonical DNA bases. An exception was the modification 5-methylcytosine (5mC), which has become established as an important,

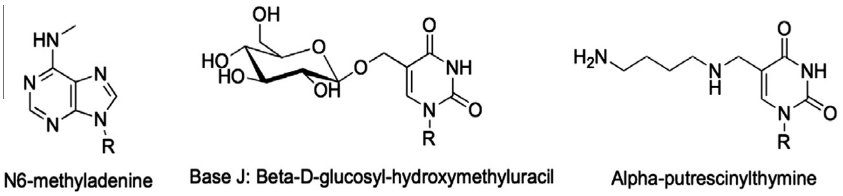


Figure 13. Examples of non-canonical bases found in nature.

heritable chemical DNA modification that can influence the expression of genes in higher organisms.[68] The methyl group in 5mC is donated from S-adenosylmethionine to C by a DNA methyltransferase enzyme, and is known to occur predominately at: cytosine-phosphate-guanine (CpG) dinucleotide sites in the genomic DNA of mammals; CpG, CpHpG and CpHpH sites (where H = A, C or T) in plants; and at restriction sites in bacterial to protect 'self' DNA from cleavage by restriction enzymes. The methylation of C alters the physical properties of DNA, conferring greater duplex stability[69] and importantly it alters the major groove recognition properties of DNA, which has a consequence for the proteins, for example, transcription machinery, that recognize DNA particularly at promoters of genes. Thus, the C-methylation patterns constitute a mechanism to dynamically reprogram gene expression in the genome. These methylation patterns can be preserved during DNA replication, and in some circumstances methylation is effectively erased by failure to preserve these patterns during replication. There has been considerable interest in the search for a mechanism to explain the apparent replication-independent de-methylation of 5mC that occurs under certain biological circumstances such as in non-dividing neural cells[70] and in the pre-replication paternal zygote.[71]

There was an important inflection in this field during 2009, when two papers from the Heinz[72] and Rao[73] laboratories were simultaneously published on the robust detection and discovery of 5-hydroxymethylcytosine (5hmC) in mammalian DNA. The alpha-ketoglutarate dependent dioxygenase ten-eleven translocase (TET) family of enzymes were found to be responsible for the oxidation of 5mC to 5hmC in the genomic DNA of mammals. Subsequently it was also discovered by Carell[74] and others[75] that the next cytosine oxidation level, 5-formylcytosine (5fC), was also present in the genomic DNA of mammals. Albeit at trace levels, further oxidation to 5-carboxylcytosine (5caC) was also evident somewhat abruptly expanding our awareness of the repertoire of non-canonical bases in mammals to 5mC plus 5hmC, 5fC and 5caC. At the time of publication of the 2009 papers on 5hmC, a colleague from hematology, Tony Green, phoned me to discuss the potential implications of the discovery for human genome function. He also challenged me to develop a chemistry that could detect 5hmC and enable one to decode the modification by DNA sequencing. It was invigorating to think about sequencing chemistry, once again, and I felt reasonably assured that the selective chemical manipulation of the heterobenzylic hydroxyl group was tractable. The key was to achieve this in a way that would lead

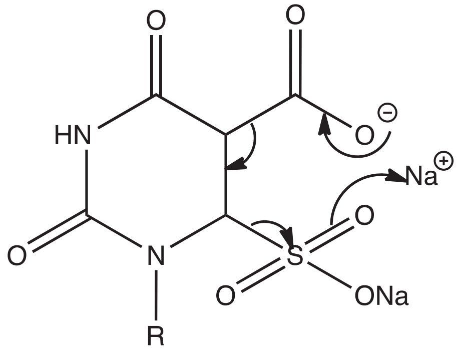


Figure 15. Proposed decarboxylation on reacting bisulfite with 5-carboxythymidine.

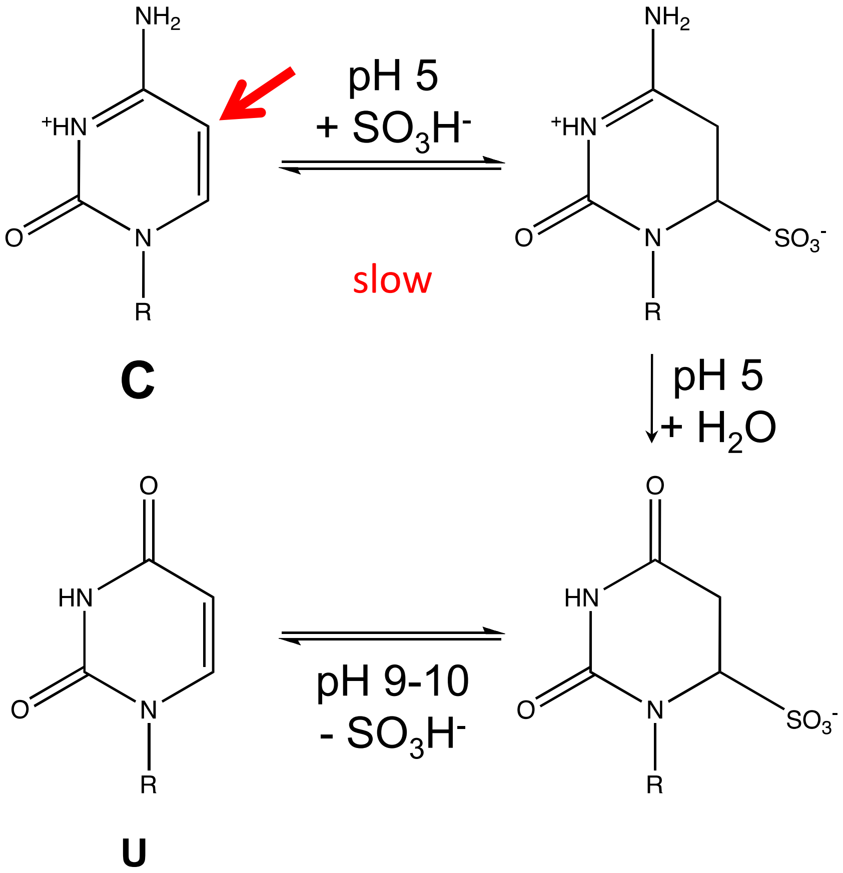


Figure 14. The transformation of C to U by bisulfite, which is resisted when C is methylated at C-5 (red arrow).

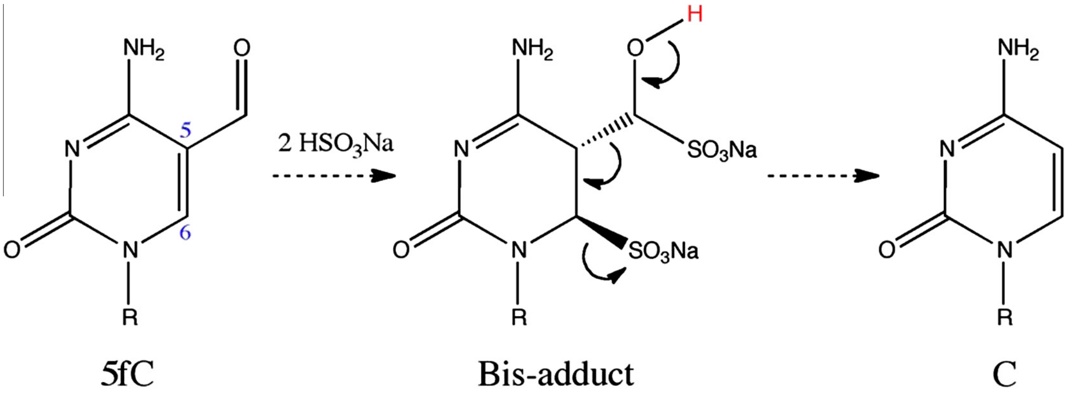


Figure 16. Deformylation of 5fC.

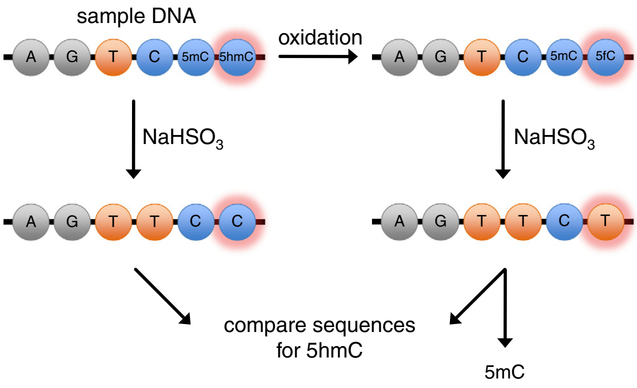


Figure 17. Oxidative Bisulfite-Sequencing (oxBS-seq).

to a read-out that could be determined by some form of DNA sequencing, whilst ensuring the chemistry was aqueous-compatible and orthogonal to the other chemical functionalities of DNA. At that time 5mC was being decoded by exploiting the chemical reactivity of sodium bisulfite with the cytosine base, first reported in 1970[76,77] (Fig. 14). Under acidic conditions bisulfite adds across the C5–C6 double bond of cytosine to form a covalent adduct that, having lost its aromaticity, is then quickly hydrolysed with loss of the exocyclic amine. Then, upon raising the pH the bisulfite is eliminated restoring the aromaticity of the base, which has now been transformed to uracil. Thus the overall transformation is conversion of C to U, which is read as a T by DNA sequencing, owing to the altered Watson Crick hydrogen-bonding pattern. When C is methylated at C5, as in 5mC, the overall conversion is considerably slower. Thus, DNA containing both methylated and unmethylated cytosine can be reacted with bisulfite such that all Cs are converted to U whilst all 5mCs remain unconverted. Thus when bisulfite-treated DNA is subjected to sequencing, only the methylated Cs would be read as a C, since 5mC preserves the Watson–Crick hydrogen-bonding pattern of a C, allowing one to decode all the positions that are methylated. There exists a huge body of work studying the role of DNA methylation in basic biology and disease.[78] However, all such studies were carried out without the knowledge that 5hmC also existed in the DNA. It was quickly established that reaction of 5hmC with bisulfite led to formation of cytosine-5-methylsulfonate, preserving the Watson–Crick hydrogen-bonding pattern of C, but without conversion to a uracil

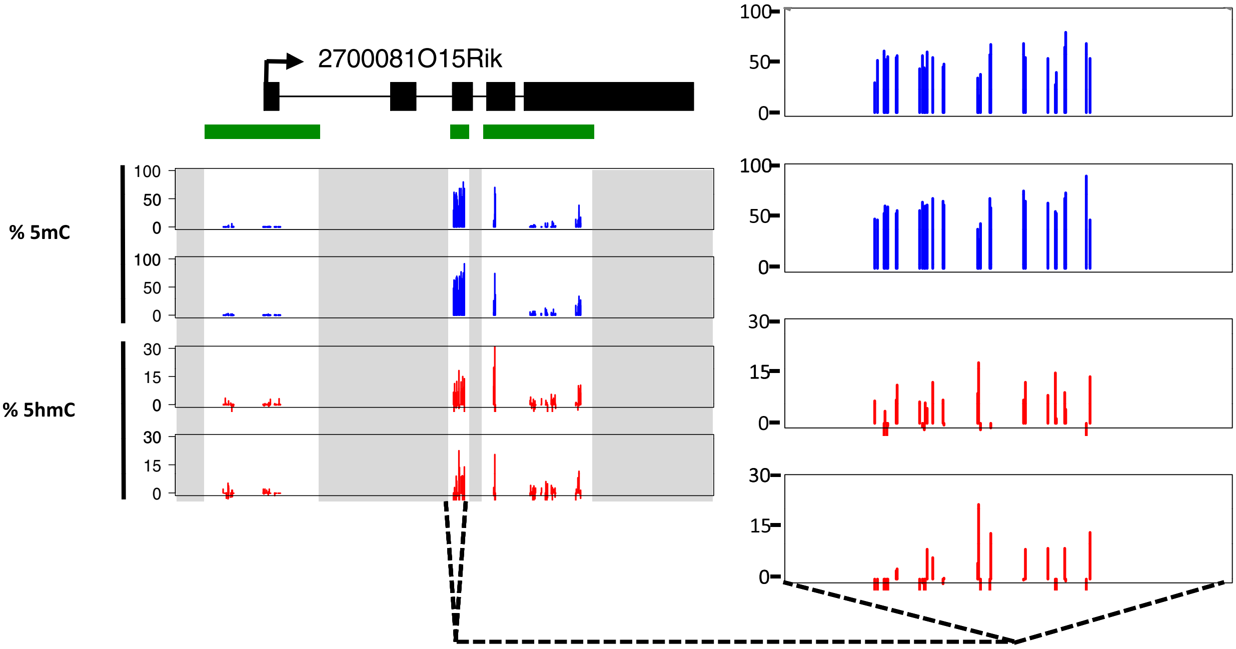


Figure 18. Example data showing single base resolution sequencing of 5mC (blue) and 5hmC (red) from DNA of mouse embryonic stem cells.

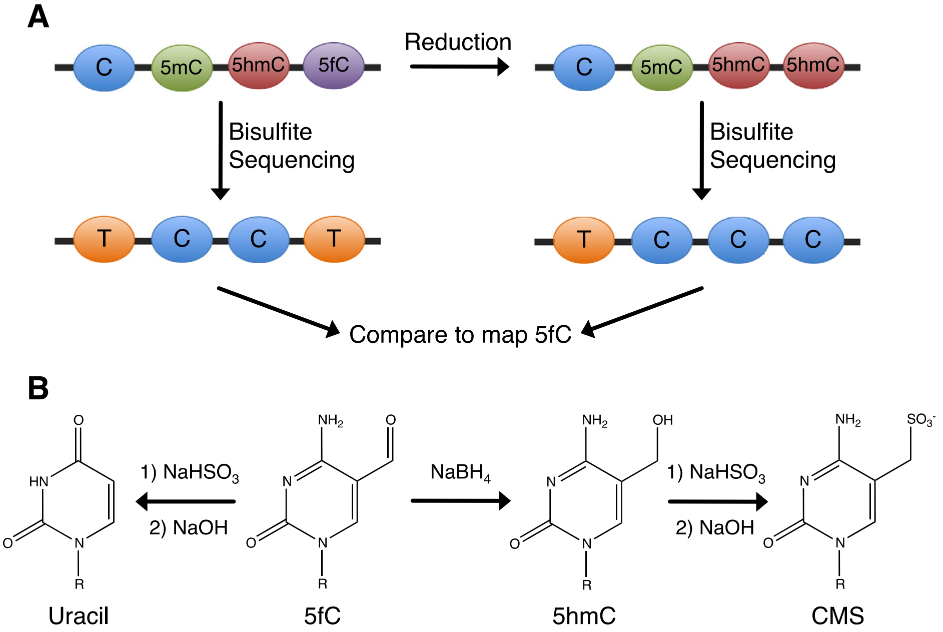


Figure 19. A scheme for reduced bisulfite sequencing (RedBS-seq) of 5fC.

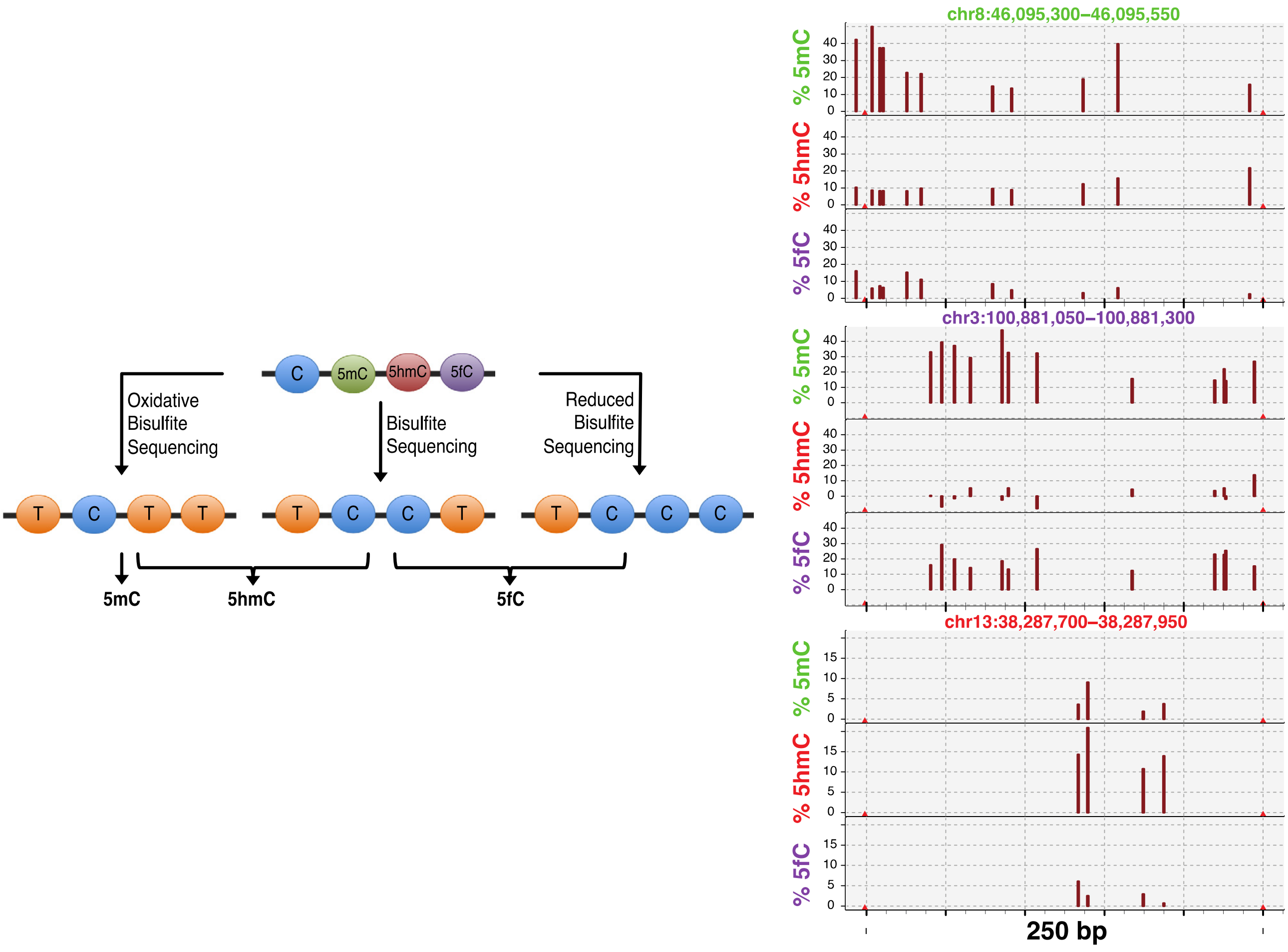


Figure 20. Decoding 5mC, 5hmC and 5fC at single base resolution.

derivative.[79,80] Therefore 5hmC is actually indistinguishable from 5mC by bisulfite-based sequencing and that is the case for the many published studies that used this approach. Our thoughts on how to detect 5hmC and resolve it from 5hmC were influenced by the work of Isono et al. on polyoxin structure elucidation, published in 1969[81] describing the facile decarboxylation of 5-carboxyuracil upon reaction with bisulfite, proposed to proceed via a bisulfite adduct (Fig. 15). We reasoned that, by analogy, reaction of bisulfite with 5-carboxycytosine would result in rapid decarboxylation to form cytosine, then hydrolytic deamination (although not necessarily in that order) to generate uracil. This was confirmed by suitable model reactions with derivatives of the 5-carboxycytosine 5caC) monomer. Our goal then became to devise a selective chemical oxidation of 5hmC to 5caC, to enable a controlled, two-step conversion of 5hmC to U, that differentiate 5hmC from 5mC in DNA. We explored many approaches for the clean chemical oxidation of 5hmC to 5caC. Part of the challenge was to achieve high efficiency, whilst avoiding oxidation of the canonical bases G, C, A, T and also 5mC. Having Steve Ley as a colleague, we naturally thought to include salts of perruthenate derivatives as potential oxidants.[82] While TPAP itself would not work, as we were working in aqueous medium, the potassium salt of perruthenate gave promising results when reacted with partially protected 5hmC nucleoside monomer as a model. It was fairly straightforward to mediate the clean transformation of 5hmC to 5fC, but we struggled to push the oxidation all the way to the carboxylate derivative. At this point we made an unexpected discovery that 5fC converted to U when reacted with bisulfite. The rate of this transformation was sufficiently fast to distinguish 5mC from 5fC (derived from 5hmC by oxidation). We proposed a possible reaction mechanism via the bisulfite adduct with the 5fC followed by fragmentation to cytosine that then converts to U via the standard hydrolysis mechanism, then elimination at elevated pH (Fig. 16). Armed with this new reactivity, we now had a scheme for decoding 5hmC, and differentiating it from 5mC and C, as depicted in Figure 17. DNA containing 5mC and 5hmC could be treated with bisulfite to convert C to U leaving both 5hmC and 5mC to be read as a C by sequencing. In a parallel reaction we would take another aliquot of the same DNA, oxidize with perruthenate, converting all 5hmCs to 5fCs and then treating with bisulfate, that would lead to an overall conversion of C and 5hmC to U, leaving only 5mC as C. By comparing the difference between those two sequencing reactions, we could cleanly discriminate 5hmC from 5mC. We have termed this process oxidative bisulfite sequencing or OxBS-seq for short. In collaboration with my colleague Wolf Reik, we carried out the first example of single base sequencing with 5hmC 5mC and C, on genomic DNA from mouse embryonic stem cells.[83,84] Figure 18 shows an example of single base resolution data. Shortly after this, an alternative and elegant method for decoding 5hmC at single base resolution, termed TAB-seq, was published by Chuan He and co-workers, involving the blocking of 5hmC by enzymatic glucosylation, then enzymatic oxidation of 5mC to 5caC followed by bisulfite.[85]

Next we turned to the challenge of decoding naturally occurring 5-formylcytosine (5fC) in DNA. Having established that 5fC transforms to U upon treatment with bisulfite, we adopted the opposite reasoning to the oxidative bisulfite reaction of 5hmC—that is, chemical reduction of naturally occurring 5fC back to 5hmC would preclude the transformation of 5fC to U upon reaction with bisulfite (Fig. 19). Once again we optimized this reaction using a 5fC monomer as a model and demonstrated clean, quantitative conversion by sodium borohydride with no detectable cross-reactivity with other functional groups in DNA. After demonstrating the method first by Sanger sequencing on synthetic oligonucleotides[86] we exemplified its utility on mouse embryonic stem cell DNA. The genomic DNA was first digested with the restriction enzyme Msp1, to select a CpG-rich subset of the genome (comprising ~23 million CpG sites). The DNA was then subjected to three parallel sequencing reactions: BS-seq, oxBS-seq and redBS-seq and the data processed to reveal the levels of 5mC, 5hmC and 5fC present at each 'C-position' integrated across the population of cells at single base resolution. Thus we were able to systematically decode 5mC, 5hmC and 5fC at single base resolution and generate the first high-resolution map of all three modifications (Fig. 20). An important outcome of this experiment was to establish that while overall levels of 5hmC (0.055% of all Cs) and 5fC (0.0014% of all Cs) were low compared to 5mC (1.2% of all Cs) integrated over the whole genome, the average level of each modification was relatively high at the sites where they existed: 17.4% for hmC, 22.8% 5fC and 25.4% 5mC. There is an alternative method for detecting 5fC from the He lab called fCab-seq.[87] Thus, the decoding of 5mC, 5hmC and 5fC in genomic DNA has been well and truly unlocked and we will see in due course how such methods elucidate its importance to life sciences and medicine.

Acknowledgments

I appreciate the vital importance of the many people who have formed my environment throughout my scientific career, many of whom I have already mentioned. It has been a privilege to have a talented and motivated group of coworkers with whom it has always been enjoyable exploring science and nature. I am proud and delighted to see so many former co-workers now succeeding as independent principal investigators. The Chemistry Department at Cambridge is a very special and inspiring place with wonderful colleagues. Steve Ley and Jeremy Sanders have been champions and mentors to me throughout my independent career. I have learned a great deal through productive, long-term collaborations with great scientists: David Klenerman (single molecule biophysics), Stephen Neidle (G-quadruplex structures), Ashok Venkitaraman (cancer biology), Wolf Reik (epigenetics). I have also been fortunate to participate in the commercialization (and improvement) of some of our methods with many outstanding scientists in Solexa and Illumina (sequencing) and Cambridge Epigenetix (decoding epigenetic bases). I continue to be motivated by curiosity-driven, exploratory research and expect that some of it, in due course, may turn out to be useful. I am grateful to the Biotechnology and Biological Sciences Research Council, Cancer Research UK, the European Union, the Wellcome Trust and the philanthropic contributions of Dr. Herchel Smith for funding my research. I thank Leanne Moden for her assistance in preparing this manuscript.

References and notes

1. Furey, W. S.; Joyce, C. M.; Osbourne, M. A.; Klenerman, D.; Peliska, J. A.; Balasubramanian, S. *Biochemistry* **1998**, *37*, 2979.
2. Watson, J. D.; Crick, F. H. *Nature* **1953**, *171*, 737.
3. Brown, D. M.; Todd, A. R. *Chem. Soc.* **1952**, 52.
4. Brown, D. M.; Fried, M.; Todd, A. R. *Chem. Ind.* **1953**, 352.
5. Maxam, A. M.; Gilbert, W. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 560.
6. Sanger, F.; Nicklen, F.; Coulson, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 5463.
7. Osbourne, M. A.; Balasubramanian, S.; Furey, W. S.; Klenerman, D. *J. Phys. Chem. B.* **1998**, *102*, 3160.
8. Osbourne, M. A.; Furey, W. S.; Klenerman, D.; Balasubramanian, S. *Anal. Chem.* **2000**, *72*, 3678.
9. Osbourne, M. A.; Barnes, C. L.; Balasubramanian, S.; Klenerman, D. *J. Phys. Chem. B* **2001**, *105*, 3120.
10. Keller, R. A.; Ambrose, W. P.; Goodwin, P. M.; Jett, J. H.; Martin, J. C.; Wu, M. *Appl. Spectrosc.* **1996**, *50*, 12A.
11. Balasubramanian, S. *Angew. Chem., Int. Ed.* **2011**, *50*, 12406.
12. Balasubramanian, S. *RSC Chem. Commun.* **2011**, *47*, 7281.
13. Bentley, D. R.; Balasubramanian, S.; Swerdlow, H. P., et al. *Nature* **2008**, *456*, 53.
14. Staudinger, H.; Meyer, J. *Helv. Chim. Acta* **1919**, *2*, 635.
15. Gololobov, Y. G. *Tetrahedron* **1981**, *37*, 437.
16. Adessi, G.; Matton, G.; Ayala, G.; Turcatti, J. J.; Mermod, P.; Maye, E.; Kawashima, E. *Nucleic Acids Res.* **2000**, *28*, e87.

17. Sanger, F.; Air, G. M.; Barrell, G. G.; Brown, N. L.; Coulson, A. R.; Fiddes, C. A.; Hutchison, C. A.; Slocombe, P. M.; Smith, M. *Nature* **1977**, *265*, 687.
18. http://www.illumina.com/.
19. http://www.illumina.com/technology/solexa_technology.ilmn.
20. Jones, S. J. M.; Laskin, J.; Li, Y. Y.; Griffith, O. L.; An, J.; Bilenky, M.; Butterfield, Y. S.; Cezard, T.; Chuah, E.; Corbett, R.; Fejes, A. P.; Griffith, M.; Yee, J.; Martin, M.; Mayo, M.; Melnyk, N.; Morin, R. D.; Pugh, T. J.; Severson, T.; Shah, S. P.; Sutcliffe, M.; Tam, A.; Terry, J.; Thiessen, N.; Thomson, T.; Varhol, R.; Zeng, T.; Zhao, Y.; Moore, R. A.; Huntsman, D. G.; Birol, I.; Hirst, M.; Holt, R. A.; Marra, M. A. *Genome Biol.* **2010**, *11*, R82.
21. Alexandrov, L. B.; Nik-Zainal, S.; Wedge, D. C.; Aparicio, S. A.; Behjati, S.; Biankin, A. V.; Bignell, G. R.; Bolli, N.; Borg, A.; Børresen-Dale, A. L.; Boyault, S.; Burkhardt, B.; Butler, A. P.; Caldas, C.; Davies, H. R.; Desmedt, C.; Eils, R.; Eyfjörd, J. E.; Foekens, J. A.; Greaves, M.; Hosoda, F.; Hutter, B.; Ilicic, T.; Imbeaud, S.; Imielinski, M.; Jäger, N.; Jones, D. T.; Jones, D.; Knappskog, S.; Kool, M.; Lakhani, S. R.; López-Otín, C.; Martin, S.; Munshi, N. C.; Nakamura, H.; Northcott, P. A.; Pajic, M.; Papaemmanuil, E.; Paradiso, A.; Pearson, J. V.; Puente, X. S.; Raine, K.; Ramakrishna, M.; Richardson, A. L.; Richter, J.; Rosenstie, P.; Schlesner, M.; Schumacher, T. N.; Span, P. N.; Teague, J. W.; Totoki, Y.; Pajic, M.; Papaemmanuil, E.; Paradiso, A.; Pearson, J. V.; Puente, X. S.; Raine, K.; Ramakrishna, M.; Richardson, A. L.; Richter, J.; Rosenstie, P.; Schlesner, M.; Schumacher, T. N.; Span, P. N.; Teague, J. W.; Totoki, Y. *Nature* **2013**, *500*, 415.
22. http://www.raredisease.org.uk/.
23. Saunders, C. J.; Miller, N. A.; Soden, S. E.; Dinwiddie, D. L.; Noll, A.; Alnadi, N. A.; Andraws, N.; Patterson, M. L.; Krivohlavek, L. A.; Fellis, J.; Humphray, S.; Saffrey, P.; Kingsbury, Z.; Weir, J. C.; Betley, J.; Grocock, R. J.; Margulies, E. H.; Farrow, E. G.; Artman, M.; Safina, N. P.; Petrikin, J. E.; Hall, K. P.; Kingsmore, S. F. *Sci. Transl. Med.* **2012**, *4*, 154ra135.
24. Jacob, H. J.; Abrams, K.; Bick, D. P.; Brodiel, K.; Dimmock, D. P.; Farrell, M.; Geurts, J.; Harris, J.; Helbling, D.; Joers, B. J.; Kliegman, R.; Kowalski, G.; Lazar, J.; Margolis, D. A.; North, P.; Northup, J.; Roquemore-Goins, A.; Scharer, G.; Shimoyama, M.; Strong, K.; Taylor, B.; Tsaih, S.; Tschannen, M. R.; Veith, R. L.; Wendt-AndraeL, J.; Wilk, B.; Worthey, E. A. *Sci. Transl. Med.* **2013**, *7*, 194cm5.
25. http://www.genomicsengland.co.uk/.
26. Gellert, M.; Lipsett, M. N.; Davies, D. R. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *1962*, 48.
27. Bang, I. *Biochem. Z.* **1910**, *26*, 293.
28. Sen, D.; Gilbert, W. *Nature* **1988**, *334*, 364.
29. Sundquist, W. I.; Klug, A. *Nature* **1989**, *342*, 825.
30. Blackburn, E. H. *Nature* **1991**, *350*, 569.
31. Salazar, M.; Thompson, B. D.; Kerwin, S. M.; Hurley, L. H. *Biochemistry* **1996**, *35*, 16110.
32. Fletcher, T. M.; Sun, D.; Salazar, M.; Hurley, L. H. *Biochemistry* **1998**, *37*, 5536.
33. Sun, D.; Thompson, B.; Cathers, B. E.; Salazar, M.; Kerwin, S. M.; Trent, J. O.; Jenkins, T. C.; Neidle, S.; Hurley, L. H. *J. Med. Chem.* **1997**, *40*, 2113.
34. Isalan, M.; Patel, S. D.; Balasubramanian, S.; Choo, Y. *Biochemistry* **2001**, *40*, 830.
35. Patel, S. D.; Isalan, M.; Gavory, G.; Ladame, S.; Choo, Y.; Balasubramanian, S. *Biochemistry* **2004**, *43*, 13452.
36. Parkinson, G. N.; Lee, M. P.; Neidle, S. *Nature* **2002**, *417*, 876.
37. Phan, A. T.; Modi, Y. S.; Patel, D. J. *J. Am. Chem. Soc.* **2004**, *126*(28), 8710.
38. Schouten, J. A.; Ladame, S.; Mason, S. J.; Cooper, M. A.; Balasubramanian, S. *J. Am Chem Soc* **2003**, *125*, 5594.
39. Jantos, J. K.; Rodriguez, R.; Ladame, S.; Shirude, P. S.; Balasubramanian, S. *Am. Chem. Soc.* **2006**, *128*, 13662.
40. Bejugam, M.; Sewitz, S.; Shirude, P. S.; Rodriguez, R.; Shahid, R.; Balasubramanian, S. *J. Am. Chem. Soc.* **2007**, *129*, 12926.
41. Shirude, P. S.; Gillies, E. R.; Ladame, S.; Godde, F.; Shin-ya, K.; Huc, I.; Balasubramanian, S. *J. Am. Chem. Soc.* **2007**, *129*, 11890.
42. Dash, J.; Shirude, P. S.; Balasubramanian, S. *Chem. Commun.* **2008**, 3055.
43. Rodriguez, R.; Muller, S.; Yeoman, J. A.; Trentesaux, C.; Riou, J. F.; Balasubramanian, S. *J. Am. Chem. Soc.* **2008**, *130*, 15758.
44. Dash, J.; Shirude, P. S.; Hsu, S.-T. D.; Balasubramanian, S. *J. Am. Chem. Soc.* **2008**, *130*, 15950.
45. Waller, Z. A. E.; Shirude, P. S.; Rodriguez, R.; Balasubramanian, S. *Chem. Commun.* **2008**, 1467.
46. Woodford, K. J.; Howell, R. M.; Usdin, K. *J. Biol. Chem.* **1994**, *269*, 27029.
47. Simonsson, T.; Kubista, M. *Nucleic Acids Res.* **1998**, *26*, 1167.
48. Siddiqui-Jain, A.; Grand, C. L.; Bearss, D. J.; Hurley, L. H. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11593.
49. Phan, A. T.; Kuryavyi, V.; Burge, S.; Neidle, S.; Patel, D. J. *J. Am. Chem. Soc.* **2007**, *129*, 4386.
50. Wei, D.; Parkinson, G. N.; Reszka, A. P.; Neidle, S. *Nucleic Acids Res.* **2012**, *40*, 4691.
51. Hsu, S. T.; Varnai, P.; Bugaut, A.; Reszka, A. P.; Neidle, S.; Balasubramanian, S. *J. Am. Chem. Soc.* **2009**, *131*, 13399.
52. Raiber, E.; Kranaster, R.; Lam, E.; Nikan, M.; Balasubramanian, S. *Nucleic Acids Res.* **2012**, *40*, 1499.
53. Balasubramanian, S.; Hurley, L. H.; Neidle, S. *Nat. Rev. Drug Disc.* **2011**, *10*, 261.
54. Murat, P.; Balasubramanian, S. *Curr. Opin. Genet. Dev.* **2014**, *25*, 22.
55. Huppert, J. L.; Balasubramanian, S. *Nucleic Acids Res.* **2005**, *33*, 2908.
56. Todd, A. K.; Johnston, M.; Neidle, S. *Nucleic Acids Res.* **2005**, *33*, 2901.
57. Huppert, J. L.; Balasubramanian, S. *Nucleic Acids Res.* **2007**, *35*, 406.
58. Huppert, J. L.; Bugaut, A.; Kumari, S.; Balasubramanian, S. *Nucleic Acids Res.* **2008**, *36*, 6260.
59. Schaffitzel, C.; Berger, I.; Postberg, J.; Hanes, J.; Lipps, H. J.; Pluckthun, A. *PNAS* **2001**, *98*, 8572.
60. Paeschke, K.; Simonsson, T.; Postberg, J.; Rhodes, D.; Lipps, H. J. *Nat. Struct. Mol. Biol.* **2005**, *12*, 847.
61. Paeschke, K.; Juranek, S.; Simonsson, T.; Hempel, A.; Rhodes, D.; Lipps, H. J. *Nat. Struct. Mol. Biol.* **2008**, *15*, 598.
62. Rodriguez, R.; Miller, K. M.; Forment, J. V.; Bradshaw, C. R.; Nikan, M.; Britton, S.; Oelschlaegel, T.; Xhemalce, B.; Balasubramanian, S.; Jackson, S. P. *Nat. Chem. Biol.* **2012**, *8*, 301.
63. Biffi, G.; Tannahill, D.; McCafferty, J.; Balasubramanian, S. *Nat. Chem.* **2013**, *5*, 182.
64. Fernando, H.; Rodriguez, R.; Balasubramanian, S. *Biochemistry* **2008**, *47*, 9365.
65. Warren, R. A. *Annu. Rev. Microbiol.* **1980**, *34*, 137.
66. Kornberg, S. R.; Zimmerman, S. B.; Kornberg, A. *J. Biol. Chem.* **1961**, *236*, 1487.
67. Van Leeuwen, F.; Taylor, M. C.; Mondragon, A.; Moreau, H.; Gibson, W.; Kieft, R.; Borst, P. *Proc. Natl. Acad. Sci.* **1998**, *95*, 2366.
68. Klose, R. J.; Bird, A. P. *Trends Biochem. Sci.* **2006**, *31*, 89.
69. Thalhammer, A.; Hansen, A. S.; El-Sagheer, A. H.; Brown, T.; Schofield, C. J. *Chem. Commun.* **2011**, *47*, 5325.
70. Guo, J. U.; Su, Y.; Zhong, C.; Ming, G. L.; Song, H. *Cell* **2001**, *145*, 423.
71. Mayer, W.; Niveleau, A.; Walter, J.; Fundele, R.; Haaf, T. *Nature* **2000**, *403*, 501.
72. Kriaucionis, S.; Heintz, N. *Science* **2009**, *324*, 929.
73. Tahiliani, M.; Koh, K. P.; Shen, Y.; Pastor, W. A.; Bandukwala, H.; Brundno, Y.; Agarwal, S.; Iyer, L. M.; Liu, D. R.; Aravind, L.; Rao, A. *Science* **2009**, *324*, 930.
74. Pfaffeneder, T.; Hackner, B.; Truss, M.; Münzel, M.; Müller, M.; Deiml, C. A.; Hagemeier, C.; Carell, T. *Angew. Chem., Int. Ed.* **2011**, *50*, 7008.
75. Ito, S.; Shen, L.; Dai, Q.; Wu, S. C.; Collins, L. B.; Swenberg, J. A.; He, C.; Zhang, Y. *Science* **2011**, 333.
76. Shapiro, R.; Servis, R. E.; Welcher, M. *J. Am. Chem. Soc.* **1970**, *92*, 422.
77. Hayatsu, H.; Wataya, Y.; Kai, K.; Iida, S. *Biochemistry* **1970**, *9*, 2858.
78. Feinberg, A. P. *Nature* **2007**, *447*, 433.
79. Huang, Y.; Pastor, W. A.; Shen, Y.; Tahiliani, M.; Liu, D. R.; Rao, A. *PLoS ONE* **2010**, *5*, e8888.
80. Hayatsu, H.; Shiragami, M. *Biochemistry* **1979**, *18*, 632.
81. Isono, K. et al *J. Am. Chem. Soc.* **1969**, *91*, 7490.
82. Ley, S. V.; Norman, J.; Griffith, W. P.; Marsden, S. P. *Synth.-Stuttgart* **1994**, *7*, 639.
83. Booth, M. J.; Branco, M. R.; Ficz, G.; Oxley, D.; Krueger, F.; Reik, W.; Balasubramanian, S. *Science* **2012**, *336*, 934.
84. Booth, M. J.; Ost, T. W. B.; Beraldi, D.; Bell, N.; Branco, M.; Reik, W.; Balasubramanian, S. *Nat. Protoc.* **2013**, *8*, 1841.
85. Yu, M.; Hon, G. C.; Szulwach, K. E.; Song, C. X.; Zhang, L.; Kim, A.; Li, X.; Dai, Q.; Shen, Y.; Park, B.; Min, J. H.; Jin, P.; Ren, B.; He, C. *Cell* **2012**, *149*, 1368.
86. Booth, M. J.; Marsico, G.; Bachman, M.; Beraldi, D.; Balasubramanian, S. *Nat. Chem.* **2014**, *6*, 435.
87. Song, C. X.; Szulwach, K. E.; Dai, Q.; Fu, Y.; Mao, S. Q.; Lin, L.; Street, C.; Li, Y.; Poidevin, M.; Wu, H.; Gao, J.; Liu, P.; Li, L.; Xu, G. L.; Jin, P.; He, C. *Cell* **2013**, *153*, 678.