CHEMICAL BIOLOGY OF DNA AND THE GENOME

SHANKAR BALASUBRAMANIAN[*, †]

[*]*Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, United Kingdom*

[†]*Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, CB2 0RE, United Kingdom*

Status of research on the chemical biology of DNA and the genome

The functions of DNA in living systems are governed by its supramolecular chemistry, molecular recognition and covalent chemistry, all of which occur in the context of a genome and the associated proteins and RNA. Here, I focus on three distinct aspects of DNA, for which chemistry is fundamental.

The canonical DNA double helix conformation, elucidated by Crick, Watson, Franklin, Wilkins and others revealed the importance of cognate hydrogen-bonded base pairing – the molecular basis for information transfer, replication and the genetic code. This in turn led to the importance of understanding ordering of DNA bases – the DNA sequence. Since the late 1970s, Fred Sanger's termination-based DNA sequencing approach provided the means to read the sequence of genomes starting from small genomes of viruses (~thousands of bases) and bacteria (~millions of bases) to the reference human genome (~billions of bases) completed as part of the human genome project. Approaches to high-throughput DNA sequencing (aka next generation sequencing) have, in recent years, transformed this field and its application to science and medicine. I will first share a perspective on next generation sequencing.

Natural, enzyme-mediated covalent chemistry operates on DNA in living systems to create reversibly modified DNA bases. Theses chemical base modifications tend to occur in the structural context of the major groove, with the most prevalent mammalian modification being a methylation at C-5 of cytosine to generate 5-methylcytosine. More recently, it was confirmed that 5-methylcytosine can be oxidized by the ten-eleven translocase (TET) enzymes to form 5-hydroxymethylcytosine, in addition to higher oxidized derivatives. 5-methylcytosine, 5-hydroxymethylcytosine and unmodified cytosine, each represent chemically distinct states of the respective part of the major groove of the DNA double helix. Such chemistry can reprogram the recognition of DNA by proteins that mediate actions, such as transcription of a gene into RNA, and DNA replication. This chemistry constitutes a vital part of the adaptative and responsive molecular mechanisms collectively termed epigenetics. It is imperative to be able to

detect and sequence these epigenetic changes to build an understanding of their functions in normal biology and in disease states. I will then share a perspective on epigenetic DNA sequencing where chemistry is deployed to sense modifications, as well as the four canonical DNA bases.

While the DNA double helix is the most prevalent conformational state of DNA in genomes, there are alternative secondary structures that form through different arrangements of non-covalent interactions. G-quadruplex alternative DNA structures form through an arrangement involving self-recognition of four guanines to form a G-quartet, which can be further stabilized by interaction (or coordination) of a cation species by the lone pair of O6 of each guanine. A single strand of DNA (or RNA) comprising multiple G-rich runs of sequence can fold back to form a G-quadruplex with multiple, stacked G-quartets. Such structures are thermally stable in the lab. I will end by discussing recent progress towards elucidating the importance of these non-canonical DNA structures in biology and the approaches developed to explore them.

Our recent contributions

Our contributions to the sequencing of genetic bases in DNA arose as unintended consequences of a collaborative project with my colleague David Klenerman and our co-workers Colin Barnes and Mark Osbourne, in Cambridge. We were exploring the process of DNA synthesis by a polymerase enzyme by single molecule fluorescence detection. The challenges of experimental realities naturally forced us to consider and innovate formats that were practical. One approach, which was relatively new at that time, was to immobilize the DNA template strand and observe the templated incorporation of fluorescently tagged nucleotides onto a DNA primer using single molecule fluorescence detection. The experiment was carried out by creating a massively parallel DNA array to enable collection of large datasets, whereby each molecule provided information. Several leaps in thinking led us from this basic research to addressing an important methodological problem: How can we get from the large-scale effort deployed the International Human Genome Project to sequence one human reference genome in a decade, to sequencing a genome a day and enable population scale human genome sequencing? Our experiments suggested a way forward. Details of the method and the challenges involved in its development are described elsewhere [1-3] and I will provide a brief overview here. Labelling each of the four activated monomer substrates, the 2'-deoxynucleotidetriphosphates, each with a distinct fluorophore, allows each base to be uniquely detected. A 3'-protecting group on each monomer allows incorporation by the polymerase and restricts the process to a single nucleotide incorporation. Chemistry was needed to modify the natural monomers to achieve this and a suitable DNA polymerase, natural or engineered, was required to incorporate the modified deoxynucleotide triphosphates. Sequential cycles of templated DNA extension, reading the fluorescence mark by imaging the surface and subsequent chemical removal of fluorophore and 3'-protecting group allows one base at a time to be decoded on each DNA fragment. The parallel array allows the process to simultaneously read many (millions-billions) of different DNA sequences. The overall readout being strings of nucleobase sequence from many sites on the array, which could then be computationally assembled, either by overlapping fragments or by alignment to the reference human genome (which had not yet been completed at the start of our work). Klenerman and I

initially envisaged a system capable of sequencing ~billions of bases per experimental run – 3-4 orders of magnitude improvement on automated Sanger sequencing, and on the scale of a human genome. After early proof of concept work, David Klenerman and I then took a step that was unusual (at least for Cambridge at that time) by creating a company, which we called Solexa. We raised investment and an interdisciplinary team was assembled to develop these concepts and early observations into an integrated system. Some of the key, chemical inventive steps included the development of a compatible chemistry for attaching and controlled detaching of a distinct fluorophore for each of the four bases and protecting the 3'-oxygen of each deoxynucleotide triphosphate. First generation Solexa nucleotides used a hemiaminal precursor at the 3'-oxygen in the form of an azidomethyl group that could be cleanly transformed, by Staudinger reduction with water soluble phosphines, to the hemiaminal which is then hydrolysed in water to release the 3'-OH. The corresponding chemistry was deployed to remove the fluorophore at the same time. The early imaging systems that we built were two-colour excitation, total internal reflectance fluorescence microscopes with emission filtering prior to signal collection by charged coupled detectors, capable of single molecule detection. The single molecule array of DNA fragments was generated by controlled fragmentation and deposition of genomic DNA onto a chemically optimised surface of a fused silica flowcell. We adopted a method developed by Pascal Mayer and co-workers, from Manteia, to enable amplification of single DNA fragments on the surface of the chip to hundreds of copies of identical sequences at the same sites. Sequencing these "clusters" of arrayed DNA molecules gave stronger fluorescence signal, avoided stochastic issues that relate to single molecule detection, helped the accuracy of sequencing and enabled the assembly of a system with lower cost cameras. The first integrated Solexa sequencing systems were manufactured and shipped to genome centres, such as the Broad Institute and the Sanger Institute, in 2006 and could sequence ~a billion bases in a single experiment. Further improvements were made, after acquisition of Solexa by Illumina, to enable high-end systems that sequence ~trillions of bases in ~1 day (effectively a 30-50x human genome an hour) at a cost to the user of ~$100. Thus, DNA sequencing has undergone an improvement in 6-7 orders of magnitude in speed and cost in 1-2 decades. This has enabled human (and other) genome sequencing at population scale. Consequently, we are in a transformative phase in building our understanding of the genetic component of human diseases (including cancer, rare diseases and infectious agents), more optimal treatments and disease detection. Biological research has also been transformed by empowering an individual researcher to have a sequencing capacity which far exceeds that of the world, just two-three decades ago.

The major groove of the DNA double helix exhibits a pattern of hydrogen bond donors and acceptors, external to the base stack, that represents the ordering of the base-pairs. Proteins can read this pattern, providing a mechanism for sequence-specific recognition of critical proteins, such as transcription factors, that selectively bind to and operate at genes or genomic regions in cells. The narrower minor groove also has sequence-dependent hydrogen bonding capability and while being too small for many protein features, such as an alpha-helix, sequence-specific polyamide small molecules pioneered by Peter Dervan, can bind and recognize these features.

Methylation of C-5 of cytosine by natural DNA methyltransferase enzymes, using S-adenosylmethionine as a methyl donor, results in a methyl group protruding at the corresponding location of the major groove, as opposed to a hydrogen atom for unmodified cytosine. 5-hydroxymethylcytosine was first observed in bacteriophage by Wyatt and Cohen in 1952. In 2009, the labs of Anjana Rao and Nathaniel Heinz showed 5-hydroxymethylcytosine to be a naturally occurring modified base in mammalian DNA, generated by enzyme oxidation by the TET family of ketoglutarate-dependent dioxygenases. The dynamic processes of DNA methylation and hydroxymethylation have the potential to modulate DNA-protein interactions with consequent downstream effects including control of gene expression across the genome. Thus, this epigenetic chemistry is of vital importance to biological function, most notably in the differentiation of pluripotent stem cells into specialized cell lineages during the development of organisms. Other key areas of biological importance include the reprogramming of cells for cell therapy and epigenetic shifts during the onset of disease and aging. There is some level of understanding about the biology of DNA methylation, though much more remains to be understood. However, far less is known about the biology of 5-hydroxymethylcytosine. Martin Bachman in my lab used heavy isotope labelling in cells and in mice to show 5-hydroxymethylcytosine was largely a stable modification in somatic tissues [4], suggesting that it might be important. Biological and medical research has comprised a substantial component of genetics and genetic sequencing, more so after improvements in sequencing, described earlier. Conversely, the sequencing of epigenetic modifications was not enabled until recently and this is an active area of interest in my laboratory. Given chemical modifications such as cytosine methylation or hydroxymethylation occur on the major groove edge of the base, they cannot be directly distinguished from each other or unmodified cytosine by an approach that relies solely on canonical base pairing, such as Sanger sequencing or Solexa-Illumina sequencing. Work by Shapiro in the 1970s showed that bisulphite could react rapidly with cytosine, by addition across the C5-C6 double bond of cytosine at acidic pH, to generate a non-aromatic adduct that undergoes hydrolytic deamination which after elimination of bisulfite at mildly basic pH affords uracil. Substitution of a methyl group at C5 slows down the kinetics of bisulfite reaction. Thus, DNA subjected to bisulfite treatment transforms all C bases to U (equivalent base pairing to T), with methylated C not converting, owing to the slower reaction. Comparative sequencing (with and without bisulfite) can identify sites in the genome that are methylated. Hydroxymethylcytosine behaves similarly to 5-methylcytosine in bisulfite-conditional sequencing and so both modifications are indistinguishable. The first sequencing approach to resolve this issue came from my PhD student Michael Booth who discovered that ruthenate or perruthenate could selectively chemically oxidise 5-hydroxymethylcytosine to 5-formylcytosine (and some 5-carboxycytosine), which upon bisulfite treatment transforms to uracil (deformylation followed by hydrolytic deamination and elimination of bisulfite). Thus, oxidation followed by bisulfite (OxBisulfite) conditionally converts

hydroxymethylcytosine to uracil (as well as cytosine to uracil), without altering 5-methylcytosine. This chemistry provided conditional approaches for sequencing and resolving cytosine, 5-methylcytosine and 5-hydroxymethyl cytosine [5]. Chuan He's lab provided an alternative, still bisulfite-dependent, approach for resolving 5-hydroxymethyl cytosine, called TAB-seq [6]. There are drawbacks with any bisulfite-based sequencing procedure: 1. bisulfite causes a side reaction where cytosine de-pyrimidinates generating abasic sites then DNA cleavage, fragmentation and substantial loss of sample DNA; 2. The C-to-T changes in sequencing at all unmodified Cs effectively reduces the genome complexity to three bases making sequence alignment difficult and detection of natural C-to-T mutations (the most common mutation in cancers) challenging; 3. Conditional chemical changes of the base identity require two or three independent sequencing runs and subtractive analysis of complex datasets to distinguish C, 5-methylcytosine and hydroxymethylcytosine, which is expensive, cumbersome and error-prone. We developed a solution to address these caveats in a biotech company that I cofounded, called Biomodal, through the work of Jens Füllgrabe, Walraj Gosal and others [7]. We enabled "six letter" sequencing of all four genetic bases plus 5-methylcytosine and 5-hydroxymethylcytosine. The solution comprised taking fragments of sample DNA and complementing each original strand, with epigenetic base modifications with a "copy" strand devoid of modifications, connected via a DNA hairpin loop. A series of enzymatic steps, then allows information on the sample strand to be transferred to the copy strand, such that standard sequencing of both strands, after opening the hairpin, allows the sequence of six letters to be decoded through a two-base code. The full details have now been published. Most cytosine modifications occur at C-p-G dyads, which have complementary G-p-C dyads on the opposite strands that can form nine different combinations of C, mC and hmC across strands. My co-worker Jack Hardwick very recently adapted six-letter sequencing to a method that resolves epigenetic information across the strand [8]. Early data on stem cell DNA using this method suggests while mC is largely symmetric across strands, hmC is almost entirely asymmetric, providing insights into a previously unknown epigenetic code, to be elucidated in biology. The most recent chemical sequencing method from my lab came from the work of David Schmidl and Sidney Becker who created an unnatural base pair capable of resolving the less abundant, natural DNA modification 5-formylcytosine [9]. A one-step, efficient chemical conversion of 5-formylcytosine with malonitrile gives an adduct that base pairs with protonated 3,7-dideazaadenine, while permitting canonical genetic base pairs. The concept was demonstrated via Sanger sequencing and with more development work and extension could lead to practical approaches to simultaneous genetic and epigenetic sequencing. Chemical approaches have therefore provided ways of accessing epigenetic information while retaining genetics. This paves the way to build a greater understanding of the roles of DNA epigenetics in biology and in disease states alongside genetics.

The field of G-quadruplex and G-tetrad nucleic acid structural motifs has origins in observations reported by Ivar Bang (in 1910), structural work by Gellert and Davis

(1962) and biophysical experiments by Sen and Gilbert (1988). My lab was drawn to this research in 1998 when my student Sachin Patel was exploring the role of G-quadruplex structure at chromosome ends in telomeric sequences and mechanisms of telomerase-mediated DNA synthesis. I was interested in the key questions: Could G-quadruplexes form in living systems? If so, where in the genome might they form? What is their function? I will give a summary of some key findings from lab. Some of our earlier work involved stimulating collaboration with Stephen Neidle in London. Biophysical experiments provided clues as to which primary DNA sequences were predisposed to forming G-quadruplex folded structures under near-physiological salt and pH. Julian Huppert used the data to generate a search algorithm, *Quadparser*, and interrogated the recently completed reference human genome, to find hundreds of thousands of potential G-quadruplex forming motifs [10]. We observed an enrichment these motifs in sequence elements immediately upstream of the start of protein coding genes (gene promoters) as well as in some repetitive genomic elements. Vicky Chambers, Giovanni Marsico and others in my lab, later developed a biophysical G-quadruplex-sensing sequencing approach, G4-seq, to experimentally corroborate these findings [11]. Plückthun and Lipps had engineered an antibody to show G-quadruplex formation at the telomeres of ciliates, while Giulia Biff in my lab generated a single chain antibody, *BG4*, and used this to demonstrate the presence of G-quadruplexes in the nuclei of a variety of human cell lines [12] and tissues. Robert Hänsel-Hertsch then developed a method to use *BG4* to map G-quadruplex formation and dynamics in the chromatin of human cells [13]. This work demonstrated that G-quadruplex formation was coupled to an open chromatin state, active transcription and specific epigenetic histone posttranslational modifications indicative of a link between G-quadruplexes and transcriptionally active genes. In subsequent studies, my lab has further elucidated the relationship between G-quadruplex formation, transcription of proximal genes and cell function. Genome editing of natural G-quadruplex sequences in cells by Isabel Esain-Garcia [14] has helped us to confirm that the folding of G-quadruplexes can, at least for some genes, promote active transcription and the corresponding assembly of chromatin proteins at the gene promoter. These fundamental studies have also given rise to the opportunity to develop small molecules that stabilise cellular G-quadruplexes that impart site-specific strand breaks that are potentially lethal. Indeed, we established the concept of synthetic lethality of G-quadruplex stabilisation in cells with genetic alterations found in many cancers such as BRCA2. Raphaël Rodriguez provided some of these foundational observations in a study where we collaborated with Steve Jackson [15]. Together with Sam Aparicio, we recently carried out pre-clinical investigations on a G-quadruplex stabilising probe molecule [16] that subsequently entered an investigational clinical trial.

An outlook on the future

Approaches for affordable DNA (and RNA) sequencing at scale will continue to improve, though it is currently not clear whether there is a vision or a real requirement for

a further six orders of magnitude improvement. There are technologically impressive single molecule sequencing approaches using nanopores (e.g. Oxford Nanopore Technologies) or real-time fluorescence (Pacific Biosciences) that can provide information from very long, continuous stretches of DNA and resolve DNA (and RNA modifications). I envisage the most important future progress in genome sequencing will stem from its application to science and medicine. Epigenetic sequencing chemistries have largely been restricted to the laboratories that created them, and some have been further developed into robust, exportable technologies that are being commercialised, scaled and will ultimately be democratised to enable discovery science in biology and in medicine. I envision that the G-quadruplex structural motif will become more broadly accepted as a natural regulatory feature of DNA and chromatin and associated biology and if the small molecules developed to target these structures have real potential as a differentiated class of drugs, we will likely see the first such medicines in the next decade or so. Chemistry has played a fundamental role in unravelling characteristics of DNA that are vital for understanding life. Advances in methodologies will continue to drive transformation in the life sciences.

Acknowledgements

I have had many truly outstanding students, postdocs and collaborators to work with over the past three decades. There have been brilliant scientists and staff at Solexa and subsequently Illumina, that I have had the pleasure of working with who are responsible for the development and continuous improvement of next generation sequencing. It has also been a pleasure working with colleagues at biomodal on the development of epigenetic sequencing. My senior biologist colleague David Tannahill has been a vital contributor to all our work over the past fifteen years, and I also thank him for proofreading this manuscript. I thank the BBSRC, Cancer Research UK, the Wellcome Trust and the benefaction of Dr Herchel Smith for supporting my research.

References

1. D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton *et al*., *Nature* **456***,* 53 (2008).
2. R. Rodriguez, Y. Krishnan, *Nature Biotechnology* **41**, 1709 (2023).
3. S. Balasubramanian, *Chem. Commun.***47**, 7281 (2011).
4. M. Bachman, S. Uribe-Lewis, X. Yang, M. Williams, A. Murrell et al., *Nature Chemistry* **6**, 1049 (2014).
5. M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger et al., *Science* **336**, 934 (2012).
6. M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, L. Zhang *et al.*, *Cell*, **149**, 1368 (2012).
7. J. Füllgrabe, W. S. Gosal, P. Creed, S. Liu, C. K. Lumby *et al., Nature Biotechnology* **41**, 1457 (2023).

8. J. S. Hardwick, S. Dhir, A. Kirchner, A. Simeone, S. M. Flynn, *et al., Proc. Natl. Acad. Sci. USA* **122**, e2512204122 (2025).
9. D. Schmidl, S. M. Becker, J. M. Edgerton, S. Balasubramanian *Nature Chemistry* doi.org/10.1038/s41557-025-01925-6 (2025).
10. J. L. Huppert and S. Balasubramanian, *Nucl. Acids Res.*, **33**, 2908 (2005)
11. V. S. Chambers, G. Marsico, J. M. Boutell, M. Di Antonio, G. P. Smith *et al., Nature Biotechnology* **33**, 877 (2015).
12. G. Biffi, J. McCafferty, D. Tannahill and S. Balasubramanian, *Nature Chemistry* **5**, 182 (2013).
13. R. Hänsel-Hertsch, D. Beraldi, S. V. Lensing, G. Marsico, K. Zyner *et al., Nature Genetics* **48**, 1267 (2016).
14. I. Esain-Garcia, A. Kirchner, L. Melidis, R. de Cesaris Araujo Tavares, S. Dhir, *Proc. Natl. Acad. Sci. USA* **121**, e2320240121 (2024).
15. R. Rodriguez, K. M. Miller, J. V. Forment, C. R. Bradshaw, M. Nikan *et al., Nature Chemical Biology* **8**, 301 (2012).
16. H. Xu, M. Di Antonio, S. McKinney, V. Mathew, B. Ho *et al., Nature Communications* **8**, 14432 (2017)