Enzymatic DNA modifications for genetic diagnostics

by

Ashleigh E. Rushton

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

School of Chemistry

College of Engineering and Physical Sciences

University of Birmingham

September 2019

Abstract

Techniques for genetic diagnostics are advancing at a rapid pace, with new technologies constantly emerging as we understand an increasing amount about the human genome. Methods of visualising DNA – such as fluorescence *in situ* hybridisation (FISH) – have been used for decades, but novel approaches are now opening up new applications for the technique, such as rapid hybridisation times for faster results, or investigation of areas of the genome that have previously been inaccessible due limitations in the current technology. This thesis shows how methyltransferase (MTase) enzymes can be used as a means to explore different regions of the human genome for various clinical applications.

Chapter three sees the optimisation of the expression of the MTase, M.*Taq*I. This protein is used throughout this thesis, alongside natural cofactor AdoMet and cofactor analogue AdoHcy-6-N_3, to label DNA site specifically. This technology is used for various experiments in the following chapters. Chapter three also attempts to produce mutated versions of other MTases for similar labelling experiments.

Chapter four uses the M.*Taq*I labelling technology to label oligoprobes, short sequences of DNA, for potential use in FISH diagnostics; specifically looking at aneuploidy, which can be indicative of certain cancers. Different conditions are tested to obtain the highest signal to noise ratio, to ensure confident detection of centromeres of the chromosome 17 in patients. These results are used to design a probe set that can simultaneously detect the loss of chromosomes 1, 7 or 17 – which is associated with poor prognosis in acute lymphocytic leukaemia – by labelling each probe with a different colour dye. As oligoprobes can detect

highly homologous sequences, this chapter also explores the use of this technique in potentially detecting single nucleotide polymorphisms in the human genome, which are associated with many diseases such as spinal muscular atrophy.

In Chapter five, this MTase labelling technology is used to produce probes for single genes as opposed to centromeric regions. Focus is on the BCR gene, as it is associated with the BCR/ABL translocation, prevalent in most cases of chronic myeloid leukaemia.

Finally, Chapter six explores DNA mapping as an approach to detect small DNA mutations, by investigating the pattern in fluorescence intensity of two highly similar sequences labelled with the MTase technology. This could enable certain carriers of spinal muscular atrophy to be identified.

Acknowledgements

Firstly, I would like to thank everyone in the Neely group for their ongoing support, especially my supervisor, Rob, who took a gamble letting a biologist into his lab! Special thanks to Andy and Nat for all of their hard work that directly supported various aspects of this thesis. Daz – getting back into academia was difficult, but your patience and guidance made it that little bit easier, and I am incredibly grateful. A huge thank you also to Krystian, Jack, Mo, Su and Emma – all of you helped me to enjoy my downtime at Birmingham, whether it was grabbing a gin at staff house, a pumpkin spiced latte at Starbucks or a body pump class at the gym, it was very much appreciated.

Thank you also to the BeyondSeq project and Horizon 2020 for my funding, which gave me the opportunity to undertake this work, as well as OGT for their help in probe design, and Jamie Webster in the Protein Expression Facility for assisting the manufacture of MTases. Thanks to Jeremy Pike (COMPARE) for all your help with image analysis, in particular the protocol development for Icy.

I'm very grateful to WMRGL for allowing me to come back to the lab and supporting my work, as well as offering suggestions and ideas, in particular the FISH and haemato-oncology hubs (especially Emma, Simon and Mariela). Thank you to all of the fantastic friends I made during my time in Birmingham, especially Bowns, Gav and Dave – no matter how busy you were you'd always find time to answer my random questions about genetics, or for a trip to the pub.

I have always had huge support from all of my family and friends; shout-out to Sean for always being there for me, being a fantastic distraction, and making me laugh through the difficult patches. Thank you to my amazing parents, grandparents and brother, who have always been proud despite not fully understanding what I do, and have offered unwavering support throughout. My in-laws Bren and Chris have continued to have faith in me throughout all stages of this PhD, and this work would not have been possible without their constant encouragement.

I am incredibly lucky to have such strong and inspiring women in my life that have kept me focused when times have been tough. Notably, Luisa, Emily, Natalie, Suzi and Amy – your drive and dedication to your own work has given me strength in mine, and I am eternally thankful for your support.

I would have struggled to write my thesis while working full time if it wasn't for the encouragement from my colleagues at kdm communications, particularly the writing team and Liv and Hannah. I am extremely grateful for all the kind words and cake, and for everything that you've all done to try and make this year as stress-free as possible.

Finally, the biggest thank you to my partner, Chaz, who has endured a lot of tears over the past four years but has encouraged me to see it through. I could not have done this without you, and Jake, counteracting the bad days with so many happy ones (and a lot of wine!). Thank you for sharing the good and the bad, and for your ongoing support.

Table of contents

Table of figures

Table of tables

CHAPTER ONE
Introduction

Introduction

1.1 DNA structure and function

Deoxyribonucleic acid (DNA) is the hereditary material contained within the nucleus of cells. These molecules contain critical genetic instructions for development, function, reproduction and growth, and are of great importance in the understanding of inheritance, as well as the genetic basis behind disease.

1.1.1 The structure of DNA

DNA is made up of nucleotides that contain a phosphate group, a sugar group and a nitrogen base. There are four nitrogen bases; adenine (A), thymine (T), cytosine (C) and guanine (G), and it is the order of these bases within a DNA sequence that determines the DNA's instructions, or genetic code. A and G bases are referred to as purines, while C and T are pyrimidines, their structures are shown in **Figure 1.1**.

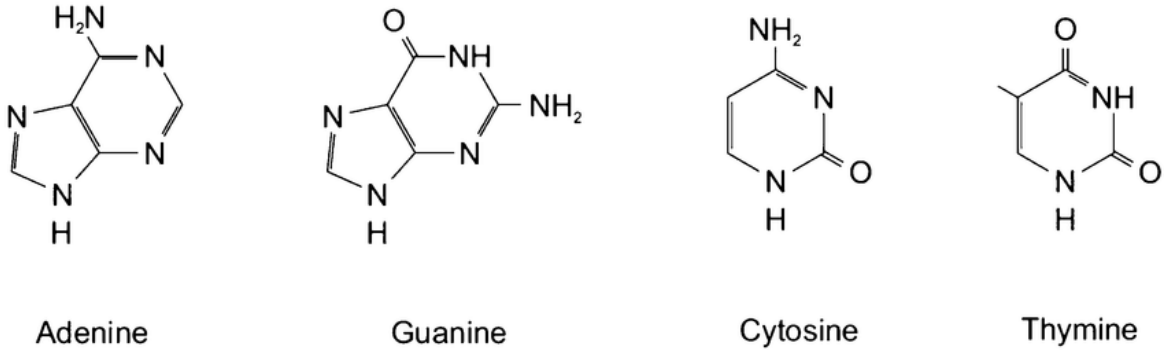


Figure 1.1: Chemical structure of purines adenine and guanine, and pyrimidines cytosine and thymine.

The DNA sequence forms genes, which contain the genetic information that gets transcribed from DNA into ribonucleic acid (RNA), and then translated into proteins, discussed in more detail in **1.1.2**.

DNA was first observed in 1869 by biochemist Frederich Miescher, but the importance of this molecule was not fully appreciated until many years later[1]. Rosalind Franklin and Maurice Wilkins were the first to suggest that DNA formed a helical molecule based on their work using X-ray diffraction[2]. Erwin Chargaff was also investigating the structure of DNA, with notable observations that A, T, C and G were not found in equal quantities (and that this varied among different species) and that the amount of A was always equal to T, and C equal to G[3]. Thanks to this research – and data from a number of other researchers – James Watson and Francis Crick determined that these nucleotide building blocks were arranged in the famous DNA double helix. They published this data in 1953[4], and were awarded the Nobel Prize alongside Maurice Wilkins in 1962.

Each helix within the double helix structure is formed by a chain of nucleotides linked by phosphodiester bonds. The helices are held together by hydrogen bonds between the base pairs; each pair consists of a purine and a pyrimidine as mentioned above, where adenine pairs with thymine and cytosine with guanine. Watson and Crick's original model suggested that there were two hydrogen bonds between bases and, while this is true for A and T, we have since discovered that there are three bonds between C and G.

There are three known conformations of DNA – A-DNA, B-DNA and Z-DNA. Watson and Crick's model describes B-DNA, where the double helix contains one complete turn every 10

base pairs and is 34 Å (3.4 nm) long, **Figure 1.2**. Adjacent base pairs are appropriately 3.4 Å apart, and are stacked via Van der Waals forces. The energy associated with these forces is relatively weak, but as the helical structure contains many bases, there is a large amount of force to stabilise the overall structure of the helix.

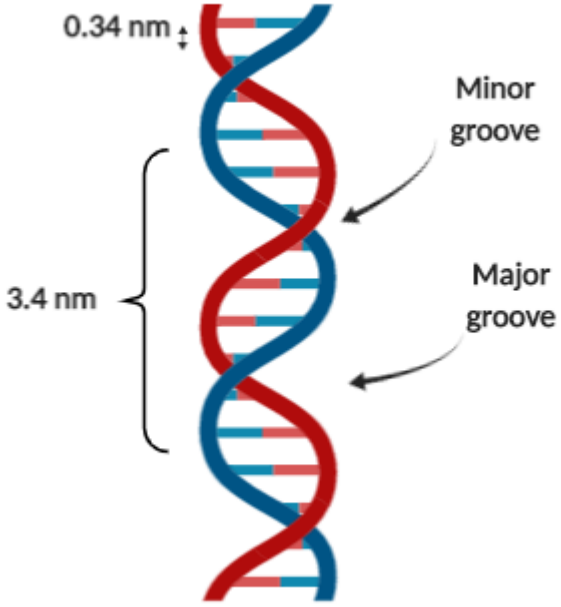


Figure 1.2: The double stranded DNA helix as described by Watson and Crick (1953). Adjacent base pairs are 0.34 nm (3.4 Å) apart, and each complete turn is 3.4 nm in length.

The stacking of bases within the double helix structure of DNA results in the molecule having two asymmetric grooves; the minor and major groove. This is a result of the configuration between the bonds and forces of the base groups. The grooves expose the base edges and are important sites for binding, allowing various proteins to interact with and maintain the DNA to regulate gene activity.

1.1.2 Transcription and translation

Double stranded DNA runs in an antiparallel manner, with the two strands running alongside each other but in opposite directions (each strand is said to run from 5' to 3'). 5' is the phosphate-bearing end, whereas 3' has a hydroxyl group, both strands align with each other in complement, i.e. the DNA sequences pair to their partner as mentioned above (A to T, C to G)[5].

These DNA sequences form genes, which encode for proteins, key molecules responsible for all functions necessary for life, such as cell division. When mutations occur and interrupt the functions of these proteins, this can lead to pathogenesis, which will be discussed later in this thesis. Genes manufacture – or express – proteins in a two-step process; transcription, followed by translation[6], **Figure 1.3**.

Transcription sees the DNA sequence as a template for complementary base-pairing. RNA polymerase II catalyses the formation of pre-mRNA (pre-messenger RNA), which is processed into mature mRNA; a single-stranded copy of the transcribed gene.

During translation, the mRNA is read in triplicate, i.e. three bases as a time; each three bases of mRNA is determined a codon. The order of codons denotes the specific amino acid sequence that is being translated, and the mRNA serves as a template to assemble the chain of amino acids to form the protein.

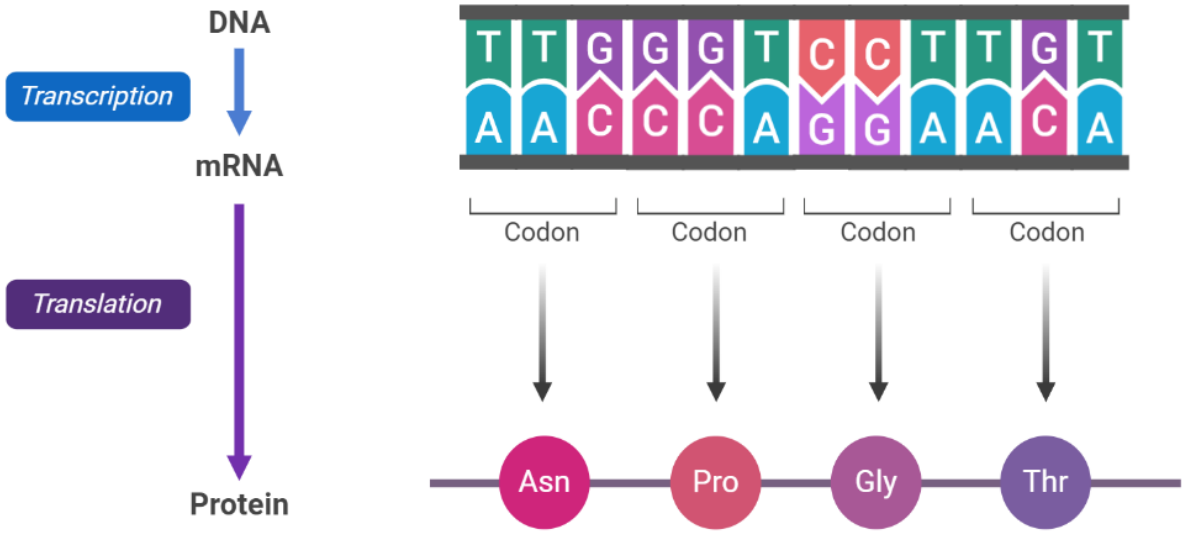


Figure 1.3: Schematic showing the process of transcription and translation. DNA is copied into mRNA by RNA polymerase, before the sequence is read and translated into protein. Each codon codes for an amino acid and builds parts of the protein molecule.

1.1.3 The cell cycle

Cell division is an important process for eukaryotic cells, functioning in tissue growth, repair, and maintenance, and is a critical component of the cell cycle[7]. The cell cycle is an ordered sequence of crucial events that occurs prior to cell division. This process is divided into four stages; first the cell increases in size (gap 1, G1), before copying its DNA (synthesis, S) preparing to divide (gap 2, G2) and then undergoing cell division (mitosis, M), **Figure 1.4**. Eukaryotic cells spend the majority of their life (around 90 %) in interphase, the period of preparation before mitosis. There are a number of proteins – growth factors, growth factor receptors, signal transducers and transcription factors – involved in each of these critical

stages[8]. Many of these act as checkpoint signalling systems to make sure that the cell cycle progresses correctly, with the end of G1 and G2 being vital in detecting DNA damage before continuing into S phase, preventing these errors from being replicated.

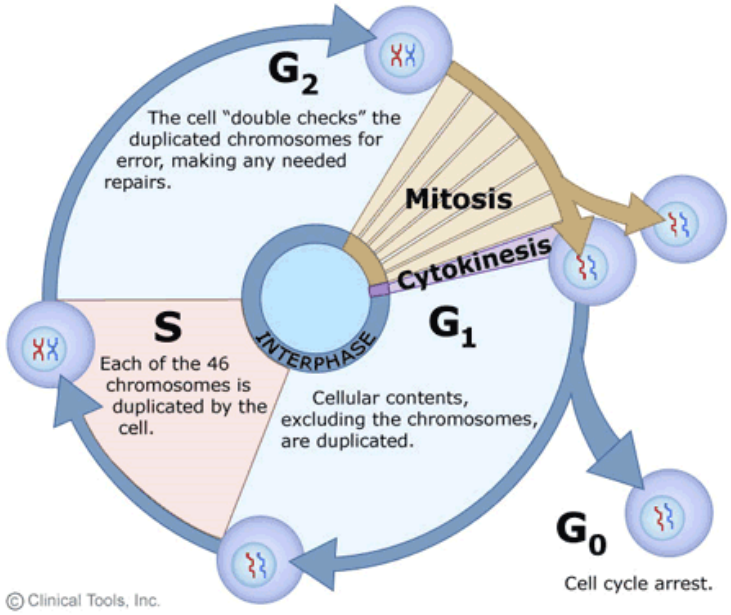


Figure 1.4: The cell cycle is a controlled process of the replication of chromosomal material and, ultimately, division of parent cell to daughter cells. Different cell cycle stages check for error before dividing in mitosis. Image taken from: https://www2.le.ac.uk/projects/vgec/highereducation/topics/cellcycle-mitosis-meiosis.

Proteins, such as p53[9], play a crucial role in the DNA damage response pathway, and mutations within these proteins can cause cells to grow irregularly – instead of being instructed to undergo apoptosis, programmed cell death – resulting in diseases such as cancer[7,10]. These proteins are therefore often key targets for therapeutics, as well as diagnostic markers, which will be explored later in this thesis.

Mitosis is the phase of cell division where two daughter cells are produced, containing the same genetic information as their parent cell[11]. The chromosomes that were replicated during

S phase are divided in a highly controlled process to make sure that each daughter cell receives a copy of each chromosome. Mitosis is divided into five distinct stages: prophase, prometaphase, metaphase, anaphase and telophase, before finally, undergoing cytokinesis, **Figure 1.5**. During these stages, the nuclear envelope surrounding the chromosomes breaks down, and the duplicated chromosomes condense and attach themselves to spindle fibres (composed of microtubules). These spindle fibres help to align the chromosomes before pulling one copy of each to the opposite side of the cell. Once this has completed, the nuclear envelope begins to reform and the chromosomes decondense, before the spindle fibres disassemble and the cell is pinched into two new cells (cytokinesis).

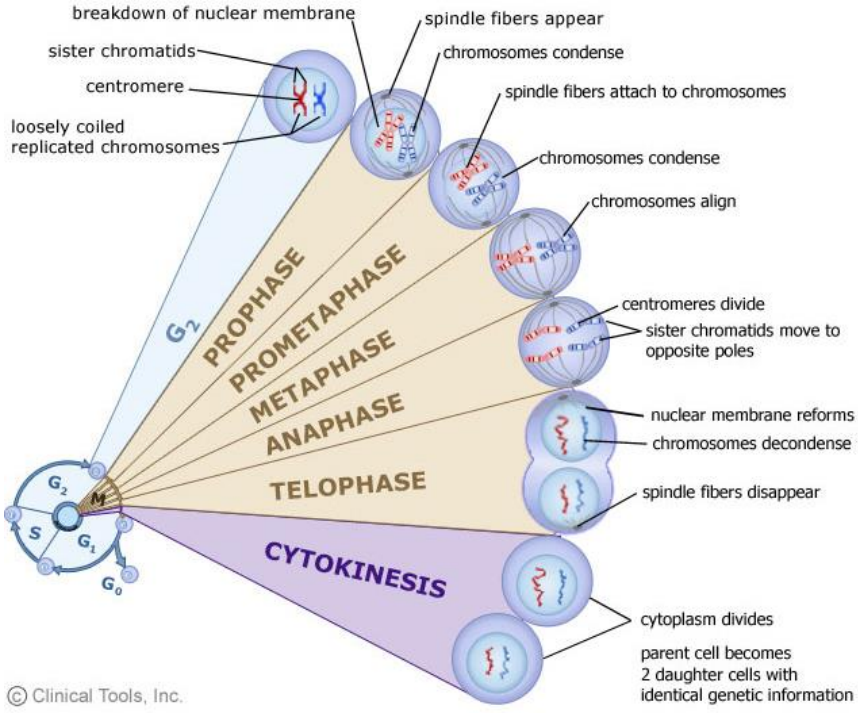


Figure 1.5: Mitosis is the process of cell division whereby daughter cells are produced as exact genetic copies of their parent cells. This process is split into various stages each with distinct characteristics. Image taken from: https://www2.le.ac.uk/projects/vgec/highereducation/topics/cellcycle-mitosis-meiosis.

The aim of mitosis is to produce daughter cells with exact copies of the genetic information as their parent cell – a full set of chromosomes. However, errors can occur during mitosis that result either in the cell being directed to apoptosis or, if the errors go undetected, cause mutations that can give rise to diseases such as cancer[10]. Nondisjunction, for example, is the failure of sister chromatids to separate during cell division – and can also occur during meiosis[12,13], when haploid sex cells are formed from diploid parent cells – and results in a daughter cell with abnormal chromosome numbers (aneuploidy). Aneuploidy is associated with many cancers and genetic diseases such as Down's syndrome (trisomy 21)[14,15].

1.1.4 DNA mutations

DNA mutations are permanent changes to a DNA sequence that have implications that range in severity, potentially affecting the cell's physiology and ability to undergo normal cellular processes, resulting in disease. Mutations also range in how much of the DNA is affected; in some cases it can be a single nucleotide, and in others large segments of a chromosome are altered[16]. These mutations can occur by different mechanisms, for example they could be inherited from a parent, or from DNA failing to replicate correctly during cell division, which could be as a result of external influences – such as radiation or specific chemicals that cause strand breaks or DNA adducts – preventing efficient replication and repair[17–19].

There are a few different types of mutations that can occur, with some having significant clinical implications, and others having little to no effect. Mutations can be either structural or numerical (aneuploidy)[20]. Severity is determined by the location of the mutation within a gene (or genes), and the function of the gene(s) that is affected. As discussed briefly above, aneuploidy – of which monosomy and trisomy are both examples – is a common cause

of many genetic disorders, as well as cancers. Different examples of structural mutations are displayed in **Figure 1.6**.

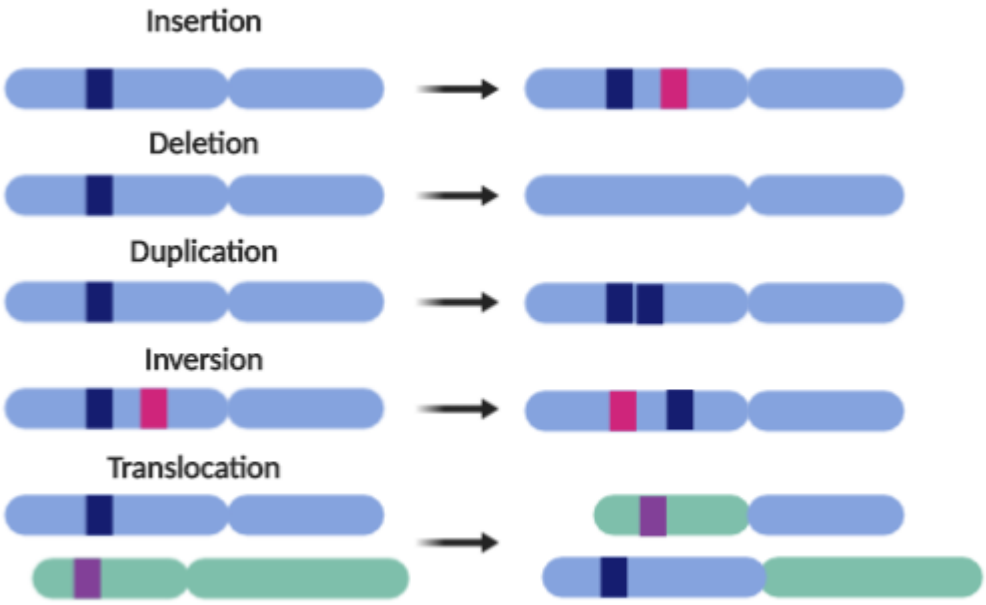


Figure 1.6: Various genetic mutations can occur, some of which give rise to disease. Structural mutations can involve insertion, deletion, duplication, inversion and translocation, where genetic information is lost, gained or transferred to a different part of a single, or multiple, chromosomes.

Insertions or deletions see the addition or loss of genetic material – from a single base to a large section (potentially hundreds of thousands of kilobases) of DNA. The size of the mutation will determine which diagnostic technique is used for detection, as will be discussed in greater detail in **1.2**. These mutations can cause a shift in the codon that is translated, termed a frameshift mutation. Duplication also sees the addition of genetic material where a specific sequence is erroneously repeated. Substitution, or point mutation, sees the change of one nucleotide to another, e.g. an A becomes a G. Changing a single base could also potentially change the codon for translation, resulting in errors in the protein being produced. This is explored later in this thesis, where a single nucleotide change causes a truncated protein to be produced in spinal muscular atrophy (SMA). Translocation is a larger mutation, that sees part of a chromosome swap with part of another, which occurs in many cases of

cancer, and is discussed in greater detail later in this thesis. Inversion is another large mutation, where a DNA segment is flipped 180 ° so that it runs in reverse to the original structure.

1.2 Genetic diagnostic techniques

Cytogenetic techniques such as karyotyping G-banded chromosomes and FISH (fluorescent *in situ* hybridisation) allow us to obtain information on a whole chromosomal level, and hence to detect large genomic rearrangements[21]. These approaches, while effective for detection of certain mutations – such as the formation of the Philadelphia Chromosome[22] through chromosomal translocation in CML (chronic myeloid leukaemia) – are not suitable for diseases that involve smaller mutations e.g. SNPs (single nucleotide polymorphisms) that are present in CF (cystic fibrosis)[23]. Recently, research has increasingly moved from cytogenetics onto molecular genetics, with NGS at the forefront of diagnostic techniques.[24] While NGS is a high throughput technique that can provide results at single base resolution, it typically does so using short reads of ~40-400 bp (Illumina), therefore making it ineffective in analysing large chromosomal rearrangements[25]. Due to the mechanism of NGS and the necessary amplification of target sequences, this also makes the technique difficult to use in diseases where copy number variation (CNV) – a type of structural variation where sections of the DNA sequence are repeated or deleted – may play a role. Ensemble averaging of amplified sections may also cause problems for diagnosing residual diseases characterised by a small subset of abnormal cells, such as in leukaemia. By analysing samples on a single molecule level, this could allow for quantitative information on (ab)normal sequences to be gathered, rather than lost through ensemble averaging. There is a clear gap in potential to diagnose certain diseases effectively and in a less time-consuming manner, which could possibly be filled by the integration of both cytogenetic and molecular techniques[26]. **Figure**

1.7 shows the range in size of genetic mutations, and the most suitable genetic testing strategy.

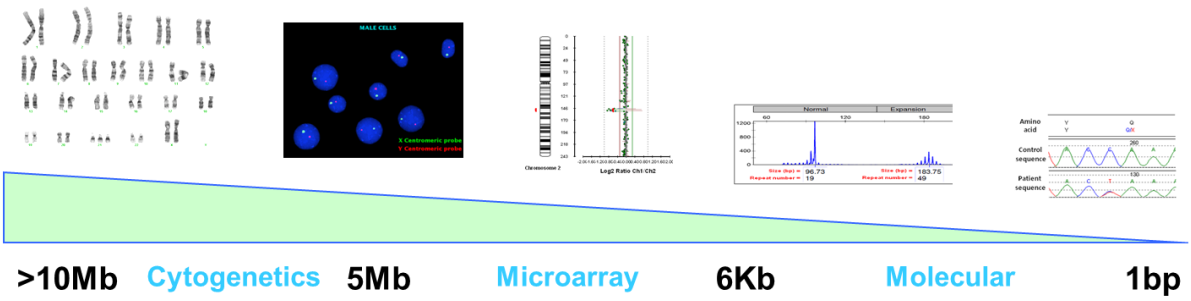


Figure 1.7: Cytogenetic techniques (e.g. karyotyping and FISH) can be used to visualise large genetic mutations, while molecular techniques such as sequencing are better suited to smaller changes, down to the single base pair level.

1.2.1 Methylation and disease

Another disadvantage of most current genetic diagnostic techniques, is that they do not take into account epigenetic information that could be the cause of many different diseases. Epigenetics is starting to gain more attention from researchers who have acknowledged the link between DNA methylation and disease[27]. In mammals, methylation occurs mainly at CpG dinucleotides, converting the DNA base cytosine to 5-methylcytosine (m5C). As DNA methylation is involved in basic gene expression and regulation, as well as cellular differentiation and development, aberrations in methylation can lead to the progression of many genetic diseases such as Prader-Wili, Angelman[28] and Beckwith-Wiedemann syndrome[29]. Hypermethylation of promoter regions on CpG islands of tumour suppressor genes results in silencing of those genes, and has been associated with almost all tumour types[30]. Hypomethylation has also been linked to cancer, as this can lead to chromosomal instability resulting in tumour growth[31,32].

1.2.2 Karyotyping

Karyotyping is a cytogenetic technique that involves the pairing and ordering of a patient's chromosomes to check for large mutations that involve megabases or more of DNA. It is often used in cases of aneuploidy, where it is suspected that there are extra chromosomes, such as in the case of Down's syndrome where the patient has trisomy 21[33], or loss of entire chromosomes such as in Turner syndrome (associated with loss of chromosome X)[34]. Karyotyping can also show structural changes including translocations, deletions and duplications, and can be used to diagnose conditions such as genetic birth defects or cancers. An example of a "normal" 46 XX karyotype is shown in **Figure 1.8**.

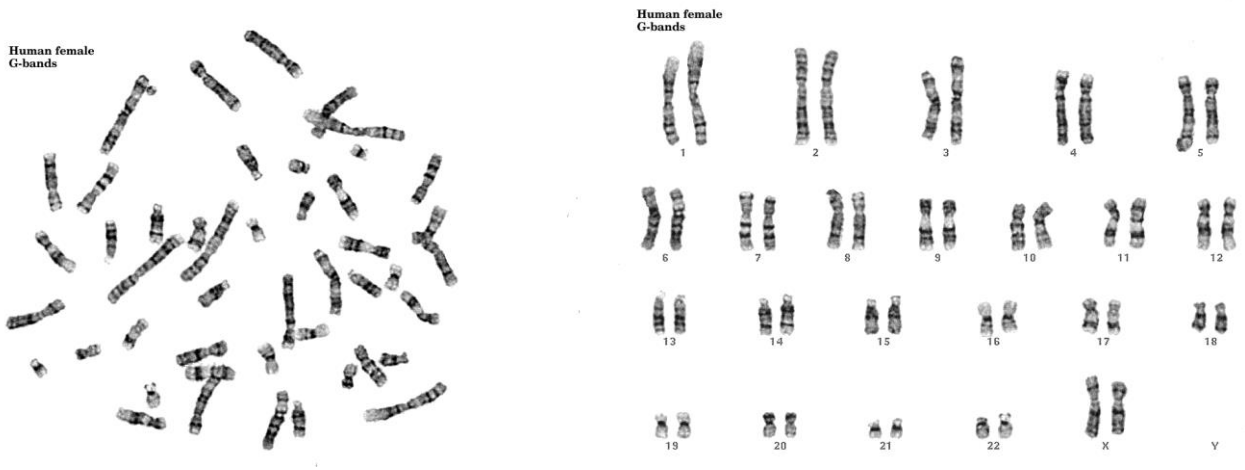


Figure 1.8: A "normal" 46 XX karyotype, before and after pairing and ordering. Taken from: http://www.pathology.washington.edu/galleries/Cytogallery/main.php?file=human%20karyotypes.

1.2.3 Fluorescence *in situ* hybridisation (FISH)

Fluorescence *in situ* hybridisation (FISH) is a molecular cytogenetic technique used to detect and localise specific DNA sequences both in metaphase and interphase cells[35]. It is considered the gold standard cytogenetic method for detecting large chromosomal mutations such as translocations and aneuploidy; FISH is particularly suited to chromosome enumeration as it can be performed on both cells both in interphase and metaphase, saving

time in harvest as synchronisation is not required as it is in karyotyping. In metaphase, resolution is typically between 1 and 3 Mb, and in interphase, mutations of around 50 kb to 1-2 Mb can be detected; the increase of resolution in metaphase is due to chromosome being more condensed and is another advantage of FISH. Due to the high specificity, sensitivity and speed in which this technique can be used, FISH is routinely used both for diagnostics and research for a range of disorders from haematological malignancies to solid tumour samples. The process works by using fluorescently-labelled probes that are designed to be complementary to the target of interest along chromosomes[35]. Once the probe has been deposited onto a slide containing fixed patient cells, it is heated to a temperature capable of denaturing the DNA of both probe – which is typically double-stranded in traditional FISH probes – and sample so that they are single stranded. The temperature is then reduced back down to around 37 ºC to allow the hybridisation of single-stranded probe to the target DNA sequence, though this process can take up to 16 hours. The slides can then be washed and the nuclei visualised using fluorescence microscopy, this process is shown in **Figure 1.9**.

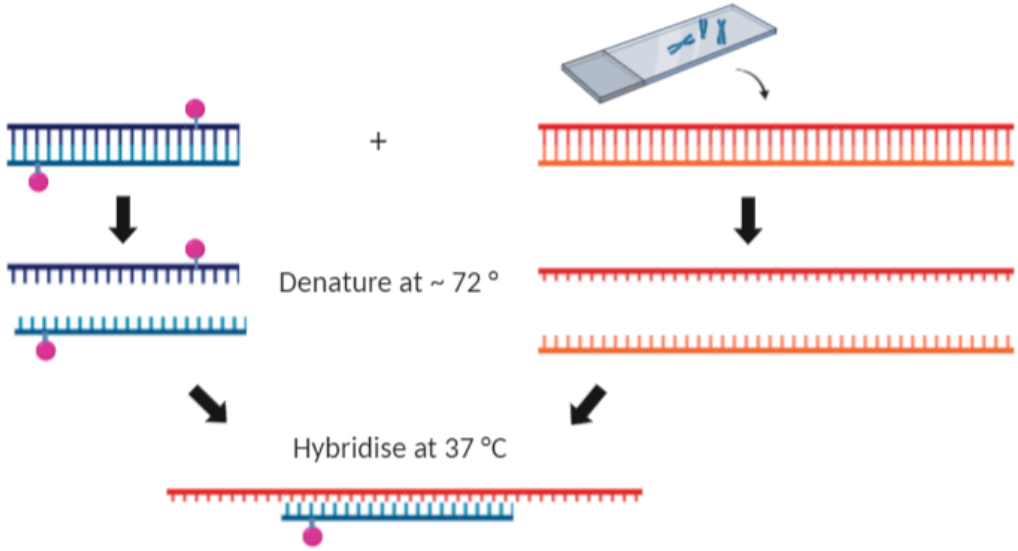


Figure 1.9: Schematic showing the workflow of FISH. Fluorescently-labelled probes are used that are complementary to the target sequence. DNA is denatured by heating to around 72 °C before cooling to hybridise at 37 °C. The sample can be visualised using fluorescence microscopy.

FISH was originally developed in the 1960s[36], but as new methods have progressed in terms of probe labelling and design – increasing the technique's sensitivity – it is being used for a wider range of applications[35,37–40]. From the 1990s, there was a sharp increase in the amount of publications using this technique[41], which has since steadied since the development of sequencing, but there are still new applications and technologies emerging that show promise for future diagnostics. FISH is still widely used for the diagnosis of cancers and other genetic disorders due to its precise and rapid nature, and these new technologies are likely to further improve the way we can treat patients, this is discussed further in Chapters 4 and 5.

Various genetic abnormalities can be highlighted using FISH, such as aneuploidy as seen in Down's syndrome where an extra chromosome 21 is present, gene fusions as seen in certain cancers such as the formation of the Philadelphia Chromosome (BCR/ABL) in CML[42], or

loss/gain of chromosomal material such as a deletion of chromosome 5q[43], commonly associated with myelodysplastic syndrome (MDS). FISH is commonly used to confirm abnormalities that have been identified via other diagnostic techniques such as karyotyping or microarray analysis, or to identify balanced rearrangements or microdeletions that alternative methods were unable to detect.

Despite advances and ongoing research into various aspects of FISH technology, it is clear that there is a call for further improvement in availability of different probes to enhance its potential in both diagnostic and research applications. Currently there are a range of probes (~200) from commercial sources, that are derived from the human genome and used in diagnostics for common genetic diseases[44]. This number however, is relatively small and restrictive, and does not provide an option for diagnosis of less common genetic abnormalities, especially microdeletions and balanced re-arrangements which may not be detectable using other techniques. These probes are generally developed from DNA fragments, collected during the Human Genome Project, that are cloned in bacterial artificial chromosomes (BACs).[45] This BAC library can then be called upon to retrieve probes that are designed for specific loci of interest.

1.2.4 OligoFISH

Oligoprobes are short sequences of DNA (around 50 bp) designed to be complementary to the region of interest (ROI)[46]. Unlike most commonly used FISH probes, they are not derived from BACs, but are designed synthetically[44]. Due to the short length and low complexity of the probe, this leads to faster hybridisation kinetics compared to traditional probes (which can be hundreds of kilobases in length), as well as greater specificity to the target[37]. If these were

to be used clinically, this could result in faster results for patients, making them highly favourable over standard probes. Another benefit of these synthetic probes is the ability to design and tailor them with high specificity to target uncommon abnormalities and variations[38]. This flexibility sets them apart from other FISH probe manufacturers who are only able to create probes for common abnormalities, or those that are easily available within a BAC library. Research has also found that oligoprobes are able to discriminate between cytogenetically indistinguishable homologous samples[44]. Structural variations that differ only at a few bases are able to be detected by these oligoprobes when designed to target these areas[47].

OligoFISH is increasing in popularity due to its extensive capabilities; there is an emerging application of using oligos for FISH in single-molecule and super resolution imaging, with Beliveau *et al.* using Oligopaint probes – single-stranded libraries of fluorescently-labelled oligos – to visualise genomic regions ranging in size from tens of kilobases to many megabases. This will be discussed in more detail in Chapter 5[45].

1.2.5 Fibre-FISH

FISH can also be performed on DNA that has been stretched and immobilised across a microscope slide, to allow visualisation of smaller mutations down to around 1,000 bp; this technique is called fibre-FISH[48,49]. In this way, the physical order of DNA fragments can also be determined and could be used to investigate translocations or duplications/deletions of certain genes. Fibre-FISH can be used in conjunction with restriction mapping to assist in ordering genomes and identifying gaps using coloured "barcodes" for visualisation, **Figure 1.10**[50].

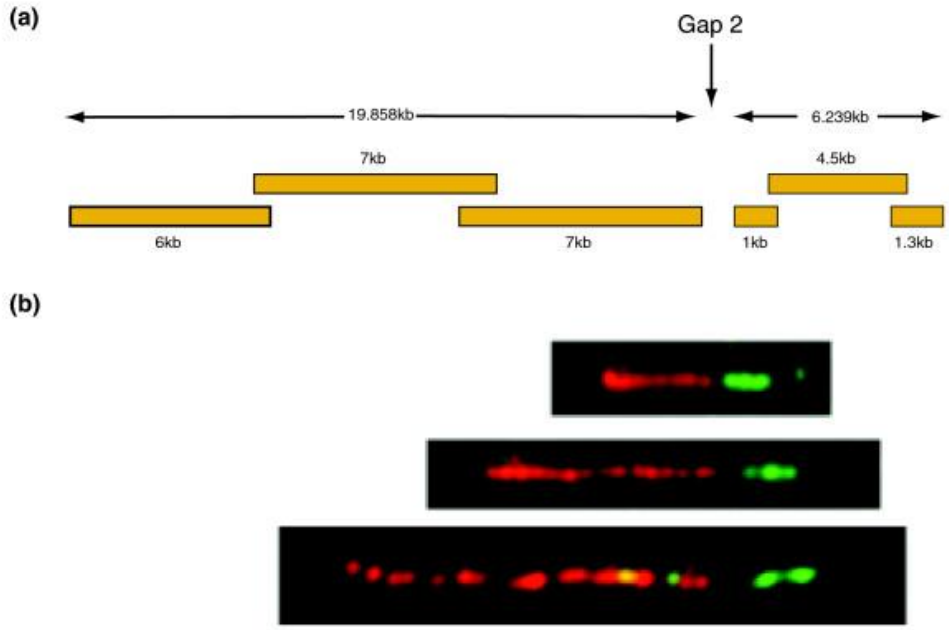


Figure 1.10: Fibre-FISH creates a physical map of DNA fragments (A) allowing visualisation of gaps within a sequence (B). Taken from: Cole *et al.* (2008).

1.2.6 Microarray

DNA microarrays are microscope slides capable of detecting thousands of genes at a time[51]. Each slide, or chip, has probes attached that are complementary to their target of interest, and can detect gene expression (mRNA). The process of running a microarray involves using mRNA samples from both the patient and a reference, which are then converted to cDNA and labelled with fluorescent probes of different colours. These samples are combined and then

hybridised to the probes on the chip. The chip is then scanned to measure the level of gene expression and flags up positions where the patient differs from the reference sample, uncovering potential changes to gene expression which could be indicative of disease. This technique is highly effective in highlighting losses and gains of genetic information that a patient may have, from copy number variation down to SNPs detection, and at a higher resolution than cytogenetic techniques such as FISH.

SNP arrays are a type of array/chip that can be used to investigate slight variations between whole genomes, and are frequently used for large genome-wide association studies to determine disease susceptibility[52]. Everyone has multiple SNPs within their genome[53], and genetic linkage analysis can be performed using SNP arrays to map a person's SNP variants against difference disease loci, providing insight into markers for diseases such as rheumatoid arthritis[54] and prostate cancer[55]. SNP arrays can also be used to detect loss or mutation of a specific allele (loss of heterozygosity (LOH)), which can be associated with oncogenesis. This technique has advantages over similar technologies as it can detect gene conversion events, highlighting the inheritance patterns of alleles from the parents, but they are not able to detect balanced translocations.

1.2.7 qPCR

Quantitative PCR (qPCR) is another technique that uses hybridisation to detect DNA mutations. PCR amplifies a specific region of interest by performing a series of heating and cooling stages to denature the DNA[56]. Primers are designed to be specific to the end of the target DNA and, using DNA polymerase and added deoxynucleotides (dNTPs), the primers extend to synthesise new strands in the cooling stage. As the cycles repeat this amplifies the

amount of DNA in the sample. qPCR uses either dyes that intercalate into double-stranded DNA or fluorescently-labelled primers, so that after amplification the DNA can be detected. This allows direct quantification of a specific DNA target such as in cases of gene amplification or translocations[57,58].

1.2.8 Sanger sequencing

DNA sequencing provides genetic information down to single base resolution, determining the exact position that each base is in. It still remains a challenge to sequence entire genomes due to their complexity, which is why these methods require the DNA to be broken into smaller fragments and reassembled into a consensus sequence. This has become a quicker and less expensive process since the completion of the Human Genome Project in 2003[59].

Sequencing was first investigated in the 1960s, where Robert Holley and colleagues sequenced the first whole nucleic acid sequence – alanine tRNA from *Saccharomyces cerevisiae*[60]. It wasn't until 1977 that a major breakthrough progressed this technology further, with Fred Sanger developing dideoxy chain-termination sequencing[61], **Figure 1.11**, now referred to as Sanger sequencing. This method uses dideoxynucleotides (ddNTPs), molecules similar to dNTPs but lacking a hydroxl group on the 3' carbon. During PCR, dNTPs are amplified by joining at the 3' hydroxyl group; by incorporating ddNTPs into the mix, this prevents the chain from growing further. Each base (A, T, C or G) of ddNTPs is fluorescently labelled, so that when the chain is terminated, the colour of the dye acts as a marker for that base. For Sanger sequencing, fractions of dNTPs and fluorescently-labelled ddNTPs are mixed and amplified by PCR, with each strand randomly terminated during replication by the presence of the ddNTP. These molecules can be applied to capillary

electrophoresis which separates fragments by length and, as they run through the capillary, the fluorescent signal is recorded by a detector. This reports which ddNTP was incorporated into the strand at each point based on peaks in the fluorescence intensity, and the chromatogram acquired can be used to determine the sequence.

Sanger sequencing can be used for fragments of up to around 900 bases in length, beyond this it becomes inefficient and expensive. Shotgun sequencing can improve this technique further

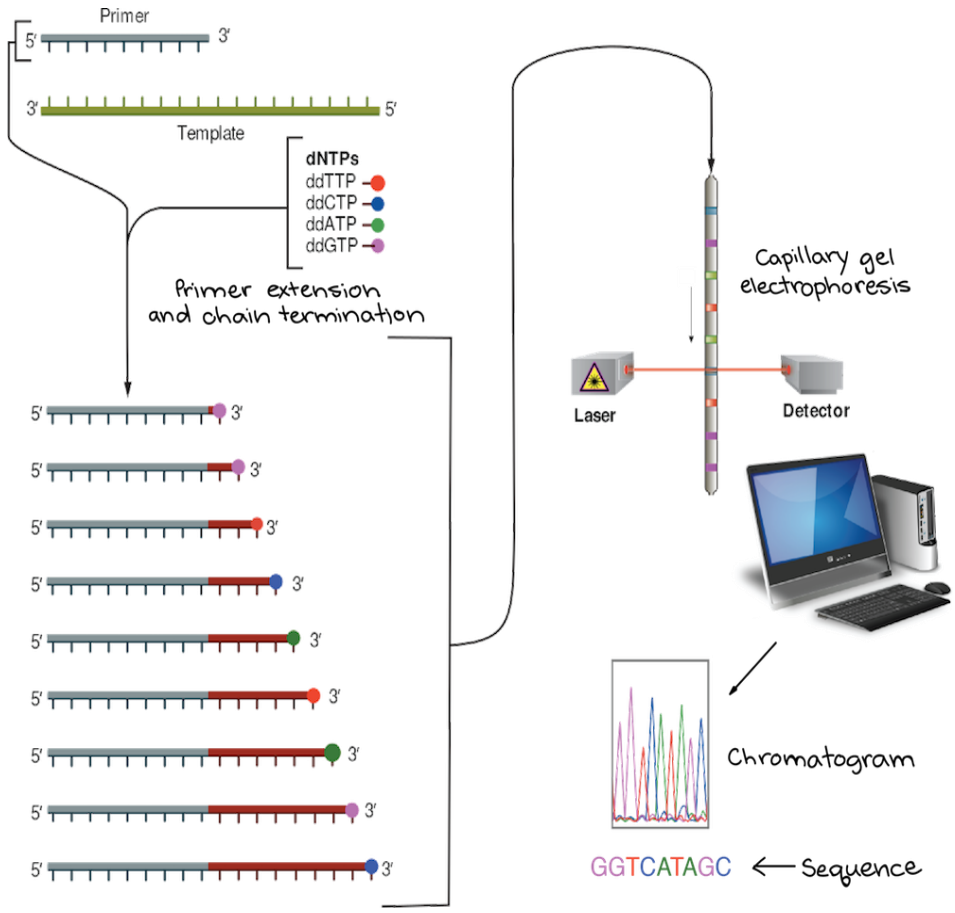


Figure 1.11: Sanger sequencing uses a chain termination technique to detect base position within a DNA sequence. Fluorescently-labelled ddNTPs incorporate into single stranded DNA during extension, terminating the chain, and these fragments can be read and ordered to determine the underlying sequence. Taken from https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/biotechnology/a/dna-sequencing.

by incorporating several rounds of fragmentation of DNA into smaller segments than chain-termination sequencing, which can then be assembled *in silico* to produce a longer, overlapping contiguous sequence[62].

1.2.9 Next generation sequencing

Since the development of Sanger sequencing, new large-scale sequencing techniques that are faster and less expensive have emerged[24,25]. The first human genome took ten years to sequence, at a cost of around $3 billion, but now, the same can be achieved with next generation sequencing (NGS) in a single day, for around $1,000. This progression has allowed NGS to become feasible for clinical applications, and is now a widely used technology.

There are a variety of different NGS techniques that vary slightly, but they all have the same features that distinguish them from Sanger sequencing. NGS processes samples at large scale in parallel, i.e. many sequencing reactions happening at the same time, which means that multiple results can be processed at once. This high-throughput process translates into sequencing potentially thousands of genes at one time, as well as providing deep sequencing and therefore accuracy to detect novel or rare variants. This not only decreases time to result, but always dramatically lowers the cost of sequencing. One of the limitations of NGS, however, is that there can be a significantly higher error rate than traditional Sanger sequencing, and that the reads are much shorter (between 35 and 700 base pairs). Shorter reads can make it more challenging bioinformatically to piece together the genome. Large rearrangements – such as duplications, deletions, insertions and translocations – can be troublesome or impossible to detect, and complex regions in the genome containing repeats

and gaps may be difficult to map. As discussed in **1.1.4**, these large-scale mutations can be associated with a number of diseases, and it is critical that these can be detected accurately.

1.2.10 Single molecule real time sequencing

New long-read technologies have been developed in an attempt to overcome some of the limitations of NGS, one of which is single-molecule real-time (SMRT) sequencing, commercialised by Pacific Biosciences[63]. This technique again uses DNA replication to sequence long fragments of DNA. Thousands of individual wells – named the zero-mode waveguides (ZMWs) – each contain a single DNA polymerase fixed to the wells' transparent bottom, alongside a single DNA template. As a labelled nucleotide is incorporated into the DNA sample in each well, a camera records the emitted light that allows the sequence to be read. As the pore is too small for light to easily pass through, the emitted light is that only of a single nucleotide. On average, SMRT sequencing can typically reach reads of around 20 kb, a huge improvement to other sequencing platforms, making it seem a good choice for deciphering difficult regions of the genome. However, there are some drawbacks to this system, as the flow cells used do not have as high throughput as Illumina NGS platforms. The ZMWs do not always carry out successful sequencing reactions either due to failure of the polymerase to anchor to the ZMW, or inaccurate loading of more than one DNA molecule into the ZMW[64]. Error rate is also high for this technique, with the single-pass error (the rate of error per read) approaching 15 %, which naturally impacts both time and cost, making it not ideal in a clinical context. In 2019, Illumina acquired Pacific Biosciences, with discussion to merge the two techniques to produce a high-throughput long-read hybrid technology with a low error rate, which would significantly improve the current techniques.

1.2.11 Nanopore sequencing

Another type of long-read sequencing is Oxford Nanopore Technologies' nanopore system[65,66]. This is one of the only platforms that does not use the incorporation of nucleotides to detect the sequence, but instead directly reads the bases themselves of single-stranded DNA (or RNA). This means that the technique does not rely on PCR amplification, avoiding the bias that this can produce. For this technique, an electric current is applied across a protein pore. As strands pass through these pores, the current is disturbed and this shift in voltage is noted. The characteristic shift of each nucleotide is recorded and, by training the data according to this, the sequence of an unknown fragment can be determined based on these shifts. This technology currently produces reads of similar size to Pacific Biosciences on average (between 10 and 20 kb[67]) but there are reports of maximum lengths reaching ~ 2 Mbp[68]. It's important to note that fragment length is limited by sample preparation, and not the technique itself, as shearing of the DNA can occur prior to analysis.

One of the main disadvantages of this technique is that the pore only remains operable for a certain numbers of runs before it breaks down. This means that in terms of cost, it may not be feasible for clinical applications, although the company has recently released higher throughput systems such as the GridION to join their expanding range of products[69]. Another challenge is that further work needs to be done to improve the technology's accuracy. The speed that the strand moves through the pore (1 to 5 μs per base) can make deciphering the recording difficult as, if there is noise present, this affects the ability to accurately detect a single nucleotide, and increases error rate[70]. Various groups are working on new algorithms which have greatly improved the technology in their MinION since it first became available[71].

1.3 Fluorescence

As discussed previously, some genetic diagnostic approaches use fluorescence microscopy to visualise molecules such as DNA. These techniques use fluorescently-labelled probes to bind to the target of interest, which allows it to be detected for further analysis.

1.3.1 Fluorescence and emission of light

Fluorescence was described by George Stokes in 1852[20], when he observed that the mineral fluorspa was capable of emitting red light when excited by UV light. Fluorescence occurs as a result of a molecule in the singlet ground state (S_0) absorbing photons of energy, which promotes electrons into a higher-energy orbital, **Figure 1.12**. This excited state (S_1') only lasts a short period of time (nanoseconds) before the electrons begin to relax, releasing energy as photons. The emitted light typically has a longer wavelength and lower energy than that absorbed. This difference in absorption – the Stokes shift – allows sensitive (single molecule) detection of emitted photons in fluorescence-based experiments, **Figure 1.13**.

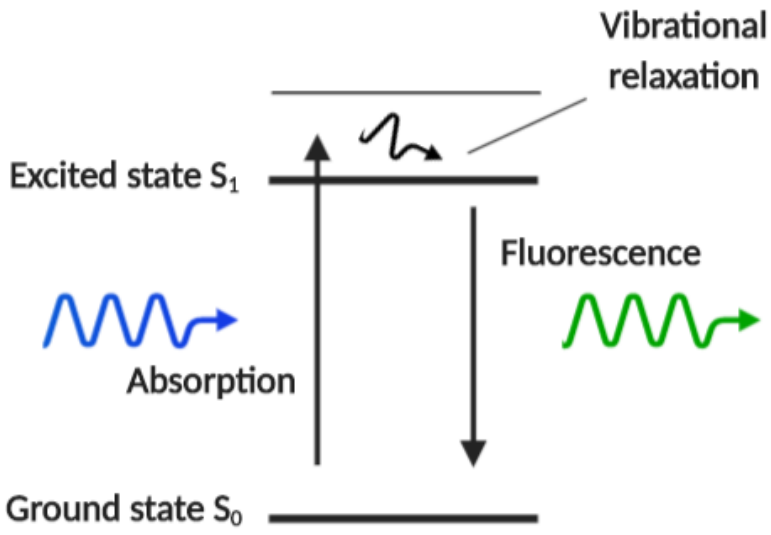


Figure 1.12: Fluorescence occurs when a molecule in the singlet ground state (S_0) absorbs photons of energy, which promotes electrons into a higher-energy orbital. This excited state lasts nanoseconds before the electrons relax and release energy as photons, producing fluorescence.

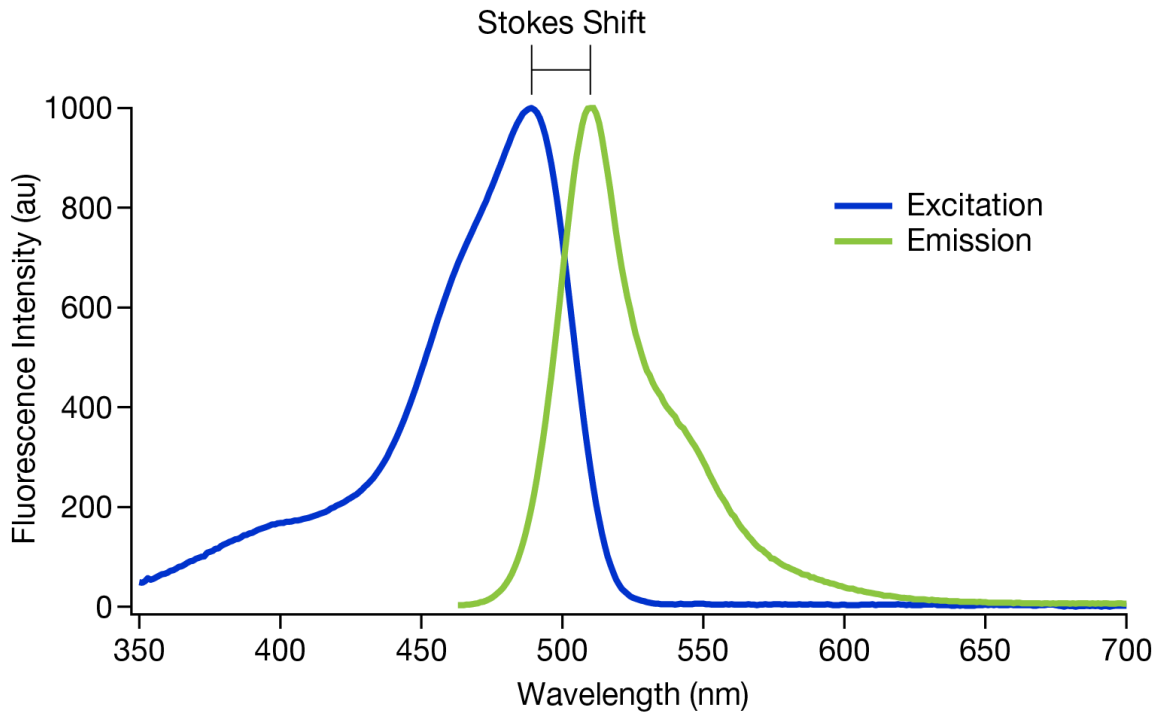


Figure 1.13: Stokes shift is the difference in absorption (excitation) and emission, that allows detection of emitted photons in fluorescence-based experiments. Taken from: https://www.scientifica.uk.com/learning-zone/widefield-fluorescence-microscopy.

1.3.2 Fluorescence microscopy

Fluorescent probes are invaluable tools, as they provide sensitivity and specificity in complex samples such as cells. In this way, they can be used to detect genetic abnormalities that may play a role in the pathogenesis of certain diseases.

Fluorescence microscopy is a popular technique used to visualise biological samples that have been labelled with fluorophores (fluorescence probes)[72]. The most basic fluorescence microscopy technique is widefield, with the set up typically consisting of a light source, a dichroic mirror, excitation and emission filters, an objective lens and a detector/camera, **Figure 1.14**. Excitation is achieved by light – typically from lasers or mercury bulbs –

passing through an excitation filter that only allows specific wavelengths through[72]. The light reflects off of the dichroic mirror, focuses through the objective lens and then hits the fluorophores within the sample. The fluorescent molecules are then excited and emit light as described in **1.3.1**. As the emitted light is a different wavelength to the excitation, this allows the emission filter and dichroic mirror to distinguish between the two, preventing the excitation light from reaching the detector. Emitted light is collected by an eyepiece or a camera for image acquisition and analysis.

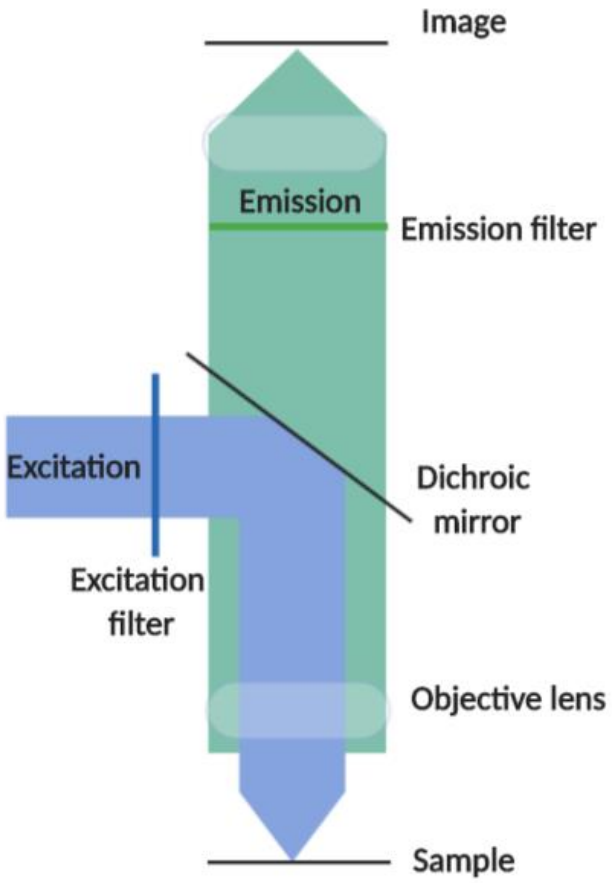


Figure 1.14: Typical widefield microscope set up. Excitation light is passed through a filter, reflected off of the dichroic mirror and focused through the objective lens to excite the fluorophores within the sample. The subsequent emitted light (of a longer wavelength that the excitation light) passes through the dichroic mirror and emission filter, to acquire the image.

While widefield microscopy can produce high-resolution images, there are some limitations that affect it. The whole sample is flooded with illumination during excitation; this causes emitters that are above and below the focal plane to be excited, which can contribute to high background noise in images compared to techniques such as confocal microscopy[73]. Computational techniques such as deconvolution can be used however to improve the resolution of images acquired from widefield microscopy, by removing light that is out of focus and re-assigning blurred noise to source points.

1.4 Methyltransferase enzymes

Methyltransferases (MTases) are enzymes that each recognise, with high specificity, sequences of DNA between 4 and 8 bps in length[74]. Bacteria have developed a unique defence mechanism against viral invasion which uses these DNA MTase enzymes. When viral DNA enters the host, it is unmethylated; to distinguish between self and non-self DNA, the bacteria methylates its own genetic information using MTases. As corresponding endonucleases (restriction enzymes) recognise and act on the same DNA sequences, this allows them to cleave and remove the unmethylated viral DNA, while protecting self-DNA from restriction (as it is blocked by a methyl group)[75]. The MTases' high specificity for DNA sequences has since been used by many research groups to deliver methyl groups to target DNA sequences of interest[76–79].

1.4.1 DNA alkylation using MTases

Recent research has shown that it is possible to use the MTases' specificity, and modify their natural co-factor, in order to transfer extended functional groups to specific targets or features of interest on DNA. This enzymatic method allows labelling of the DNA without

modification of the sequence itself. As there are thousands of known DNA MTases targeting more than 250 different recognition sites[80], using these for DNA modification seems a promising approach for many applications, including gaining more knowledge and understanding of gene expression[81], cellular differentiation and the link between methylation pattern and disease[32,82,83], as well as aiding detection of genetic variation within a population[27,81,84].

MTases can be split into two major groups, and interact with either adenine or cytosine, shown in **Figure 1.15A-C**. The m5C class of enzymes mentioned previously (e.g. M.*Hha*I, M.*Bsa*WI) methylate the ring carbon at position 5 of cytosine, converting it to 5-methylcytosine (m5C)[85]. The other group consists of amino MTases which target the exocyclic nitrogen of either adenine (e.g. M.*Taq*I) or cytosine (e.g. M.*Bam*HI) resulting in N6-methyladenine and N4-methylcytosine respectively[80].

Typically, the structure of a bacterial DNA MTase consists of a large and small domain, of which the large contains the cofactor binding site and catalytic domain, and the small accommodates the target recognition domain (TRD) responsible for sequence specific DNA recognition[85,86]. The structure of the catalytic domain remains similar for all DNA MTase enzymes and comparative sequence analysis has shown that within this domain there were 10 conserved motifs in m5C MTases. The conserved motifs (of which I, IV and VI were most conserved)[72,81] were examined using crystal structure analysis and their importance for function confirmed by performing mutagenesis of these conserved residues. Their mutation has a dramatic impact on catalysis, cofactor binding and DNA binding[82]. The ubiquitous co-

factor S-adenosyl-L-methionine (AdoMet), acts as the donor for the transfer of a methyl group to the DNA target, leaving the product S-adenosyl-L-homocysteine (AdoHcy), **Figure 1.15**.

A novel concept for labelling DNA using synthetic cofactors was put forward in 2004[77], by replacing AdoMet's amino acid side chain with a highly reactive aziridine group with fluorophore attached, it was found that the DNA MTase M.*Taq*I was able to catalyse DNA modification to the specific sequence of interest. This method, as well as the use of N-

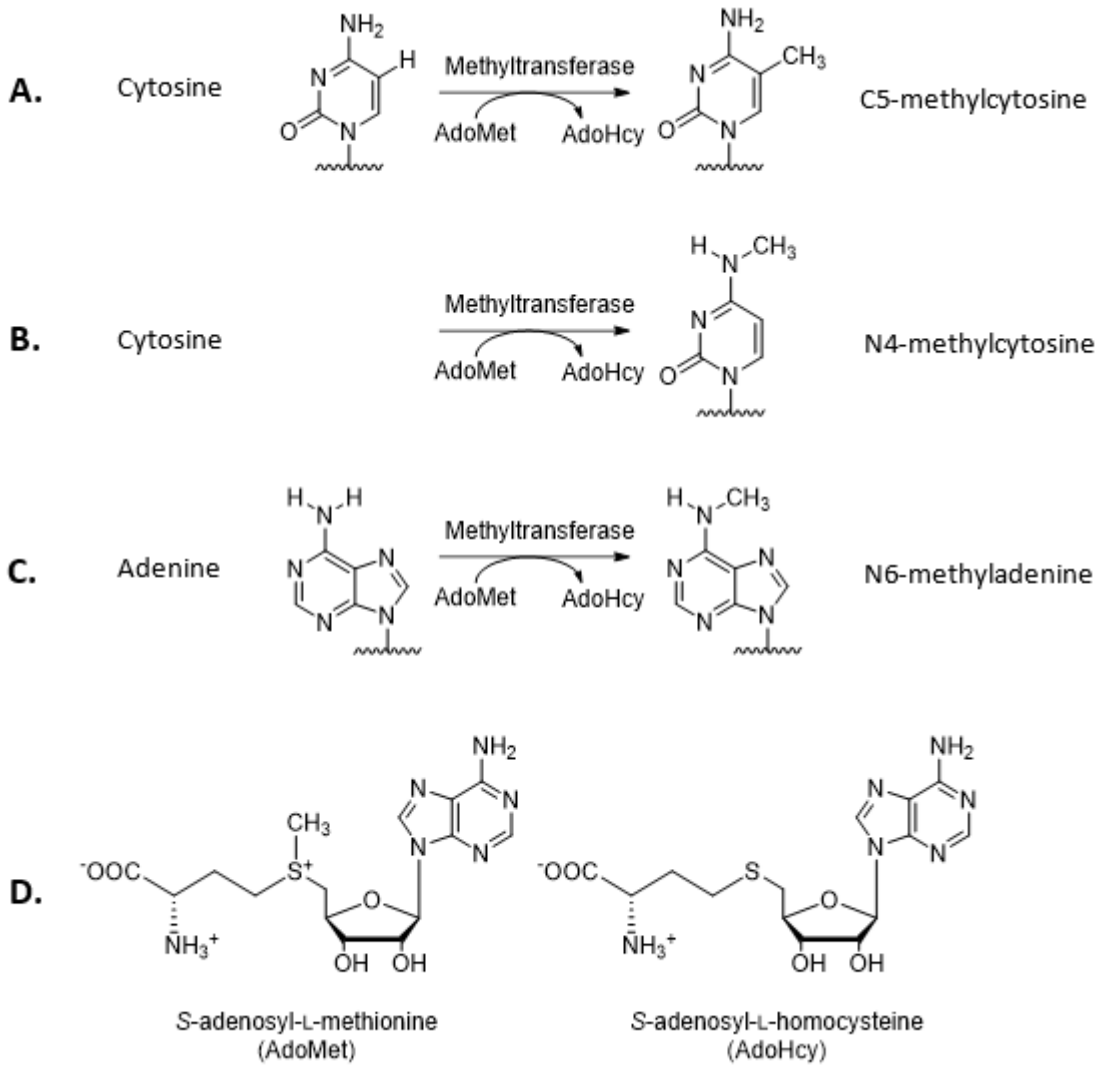


Figure 1.15: MTases transfer methyl groups site-specifically to DNA. The target base differs depending on the type of MTase, resulting in A) C5-methylcytosine B) N4-methylcytosine or C) N6-methyladenine. AdoMet is the cofactor used for methylation, with the leaving product AdoHcy (D).

mustard chemistries as seen in 2005[87], involved the coupling of the entire cofactor itself to the target DNA sequence and was dubbed sequence-specific MTase-induced labelling (SMILing). This chemistry was again used by Schmidt *et al.* in 2008 and successfully used to study cell transfection[88]. In 2006, Dalhoff *et al.* reported the synthesis of the first AdoMet analogue with carbon chains replacing methyl groups[79,89,90], allowing transfer of the extended linear groups alone to specific target DNA sequences, and evaluated the cofactors efficiency in enzymatic reactions with all three MTase classes. The Weinhold and Klimašauskas group continued research into synthetic cofactors with these extended propargylic side chains, with this transfer of activated functional groups being referred to as MTase-directed transfer of activated groups (mTAG)[90]. Using mTAG, many chemical entities can be transferred to a DNA target sequence, for example a fluorophore, that could then be used for optical mapping experiments.

These synthetic cofactors have not provided efficient transalkylation with wild type m5C MTases, however, currently making them somewhat cumbersome to work with. Steric engineering of MTase enzymes has proven successful in more efficiently allowing the transfer of these unnatural groups from cofactor analogues to the target DNA, and will be discussed in greater depth in **1.4.2**. Using two step mTAG labelling for single molecule mapping experiments has proven partly successful for Vranken *et al.* in 2014, while attempting to couple fluorophores to specific sequences of DNA[78]. However, while the MTases were able to functionalise the DNA, the step of coupling the fluorophore (azide–alkyne cycloaddition) resulted in only 60 % labelling efficiency. Single-step labelling, using fluorescent arizidine-based cofactors as discussed above, allowing MTases to deliver a fluorophore directly from the analogue seemed to provide more successful results. Weinhold *et al.* have very recently published work demonstrating that single step labelling can be

efficient in labelling bacterial genomes[29]. This technique could be useful in the ability to rapidly screen organisms and pathogens and for bacterial strain typing.

As described above, AdoMet analogues can be produced to contain extended chemical moieties, such as amine or azide groups, **Figure 1.16C**. When using amine cofactors, such as AdoHcy-6-NH_2, **Figure 1.16A**, the primary amine is transferred, which can react with N-hydroxysuccinimide (NHS) ester dyes. This coupling reaction occurs at slightly alkaline conditions of pH 7.2 – 9, creating a stable amide bond, **Figure 1.16B**.

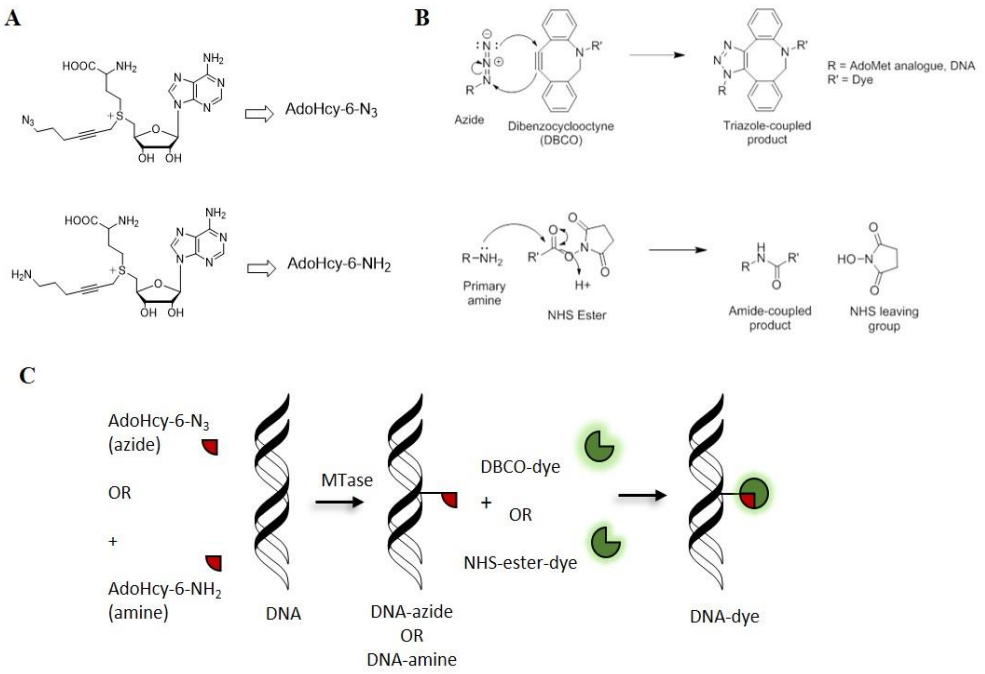


Figure 1.16: A) Diagram of AdoMet analogues: AdoHcy-6-N_3 (top) and AdoHcy-6-NH_2 (bottom) B) SPAAC coupling reaction (top) and amine-NHS coupling reaction (bottom) C) MTase-directed fluorophore coupling: labelling with AdoHcy-6-N_3 or AdoHcy-6-NH_2 using SPAAC and NHS coupling, respectively.

Labelling with DBCO-functionalised dyes enables coupling to azide chains via strain-promoted azide-alkyne cycloaddition (SPAAC) chemistry, and can be used with azide

AdoMet analogues such as AdoHcy-6-N_3, **Figure 1.16A**. The conformational strain on the eight-membered ring in the DBCO dye structure, allows it to react with the azide group of the cofactor without heat or added catalysts, via click chemistry, **Figure 1.16B**. **Figure 1.16C** shows how these AdoMet analogues can be used with the MTases to transfer azide or amine groups to the DNA site specifically, before undergoing SPAAC (azide) or NHS (amine) coupling reactions to label the site with a fluorophore.

In order to test labelling efficiency, DNA protection assays are often used. This involves incubating DNA (usually pUC19 or lambda) with the MTase and cofactor of choice, followed by challenging the DNA with the MTase's corresponding restriction enzyme (i.e. recognising the same DNA sequence). If alkylation is successful, this will protect the DNA against restriction. If alkylation is unsuccessful, the restriction enzyme can cut the DNA at all of its recognition sites. The DNA is then visualised on an agarose gel via electrophoresis. From analysing the presence of bands at certain points in the gel, it is possible to calculate how efficient the MTase has been at labelling the DNA and, quantify the level of protection. This technique will be used extensively in Chapter 3.

1.4.2 Steric engineering of MTases for improved labelling of DNA

As discussed, DNA MTases are enzymes that target short sequences of DNA typically between 4-8 bases long[76]. The novel concept of using synthetic cofactors, with active functional side-chains in place of methyl groups, makes DNA labelling possible[74,76,91]. These side-chains can be transferred to DNA site-specifically using MTases and later attached to a range of biomolecules or dyes. This way of labelling has many advantages, one being that it does not cause damage to the DNA such as that seen with nicking and restriction enzymes[92].

MTase-directed labelling has since been used for several applications such as optical mapping, DNA capture, and visualising DNA *in situ*.

m5C-MTases (such as M.*Bsa*WI), are present in both eukaryotes and prokaryotes, and share similar mechanisms and structure. Comparative sequence analysis has shown that these MTases share conserved sequences (I-X). From examining the crystal structure of these conserved motifs, it was suggested that by making mutations at specific residues this could have a dramatic impact on cofactor binding[76]. By systematically modifying these bases at non-essential positions, this opens-up the cofactor binding pocket, allowing for greater transalkylation with relatively bulky synthetic cofactors, compared to the natural cofactor AdoMet. Research performed by Lukinavičius *et al*[76]., showed that by making double and triple replacements in the amino acid sequence of the MTase M.*Hha*I (which recognises GCGC, methylating the underlined cytosine), efficiency of reactions with AdoMet analogues could be significantly increased. The modifications of M.*Hha*I in this research were performed in the cofactor binding pocket at two or three non-essential positions in the variable region in conserved motifs IV and X, shown in **Figure 1.17**[76].

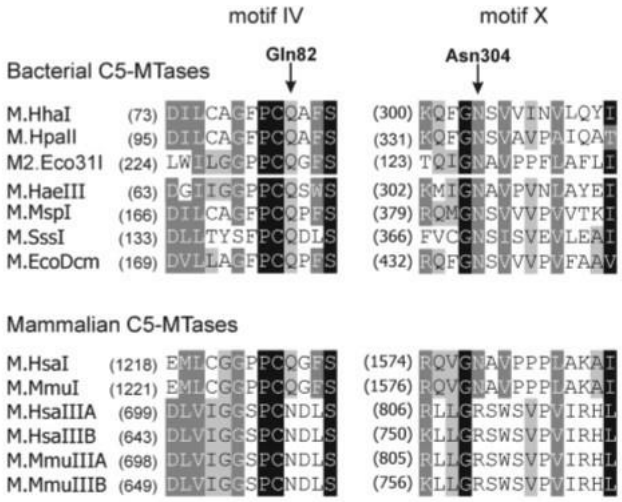


Figure 1.17: Amino acid alignments at conserved motifs IV and X in C5-MTases, a target for directed mutagenesis. Taken from Lukinavičius *et al.*(2012).

In particular, this study found that these mutations led to an increase in synthetic cofactor binding efficiency, and a higher rate of alkyl transfer (**Figure 1.18**[76]), as well as a reduction in the stability of the complex DNA-M.*Hha*I-AdoHcy, meaning reduced affinity towards natural cofactor AdoMet This is important, as it demonstrates that engineered MTases can react with the synthetic cofactors even in the presence of competing AdoMet. Research has also shown, however, that methylation efficiency can vary significantly depending on the cofactor and enzyme combination, so these changes in the MTase structure may not be functional with all AdoMet analogues. This research has shown significant applicability of expanding a range of engineered m5C-MTases to develop a toolbox for covalent sequence-specific labelling of DNA both *ex vivo* and *in vivo*.

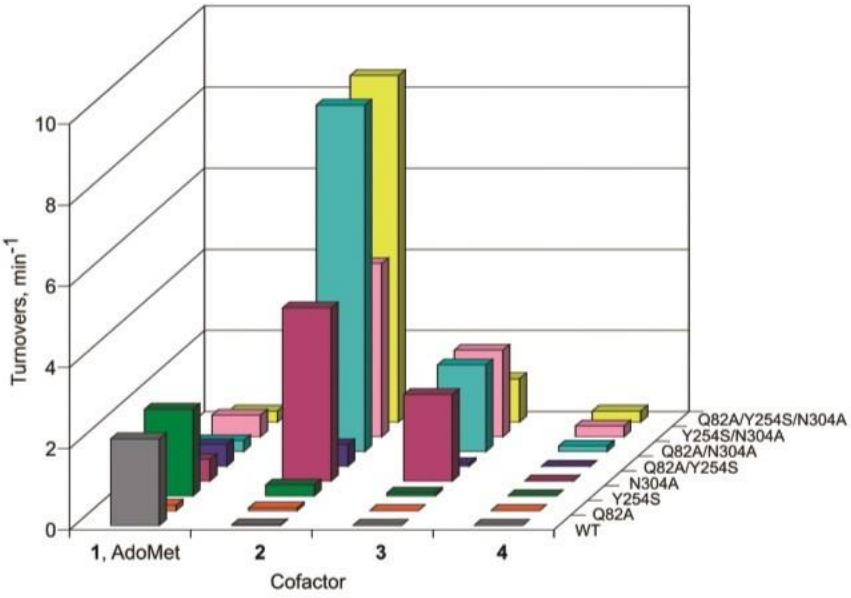


Figure 1.18: Activity of wild type and mutated M.*Hha*I with AdoMet and a range of AdoMet analogues. DNA protection assays were used to acquire turnover estimates. Taken from Lukinavičius *et al.* (2012).

1.5 Optical mapping

Used in conjunction with sequencing, optical mapping has been shown to be a highly useful for a number of applications, mainly as a tool to aid the assembly of genomes[93]. The aim of optical mapping is to provide short pieces of genomic information from a large single molecule of DNA every few thousand bases, thanks to enzymes which target and modify sequences of around 6-8 bps in length. Optical mapping techniques involve the linearisation and extension of single DNA molecules, which are then visualised through fluorescence microscopy[26,78,93–95]. Single molecule techniques such as this can allow a simple route to studying large DNA molecules (up to megabases in length), without the need for DNA amplification or building a complex library. This can lead to a more straightforward approach to assembling and studying whole genomes, even of complex samples. Novel techniques for optical mapping of DNA have been emerging over the past few years and are being

continuously improved[96,97]. These techniques look promising, as they address some of the challenges that current diagnostic techniques face (as discussed in **1.2**) such as issues with CNV, ensemble averaging and reconstruction of the genome after amplification of sequences. Optical mapping is of great interest as it can provide a scaffold to aid genome assembly when used in conjunction with sequencing, as discussed below. Another highly effective application of optical mapping is for strain typing and sequence validation. Research published by Grunwald *et al.* in 2015[98], demonstrated how optical mapping was used to rapidly identify bacterial genomes for both T7 and lambda bacteriophages, which provides both an exciting and important practical application. This technique can therefore directly analyse DNA molecules without any *a priori* knowledge of the sample composition.

1.5.1 Restriction mapping

Initial optical mapping studies in the mid-1990s were developed using restriction enzymes[99,100], and remain the most established. These studies, carried out by the Schwartz lab, have formed the basis of all subsequent optical mapping techniques and have been used to sequence full genomes *de novo*, including the recent publication of the goat genome[101]. Other great achievements using the optical restriction map technique include the assembly of the highly repetitive maize genome[102] in 2009 and the mapping of four human genomes[103] in 2010. More recently it was used to complete the genomic sequence of a new species of bacteria[104]. The technique works by depositing the sample onto a functionalised surface before cutting segments of DNA at specific sites. The DNA is stained, commonly with intercalating dye YOYO-1, imaged and analysed. This provides a scaffold for which

sequencing information can be assembled and sized, as well at the acknowledgement of where gaps lie within the sequence, illustrated in **Figure 1.19**[93].

Despite this being a highly recognised and published method for optical mapping of genomes, there are also some limitations. Restriction enzymes are usually chosen that cut on estimate every 7-10 kbs, this is to prevent small fragments of ~2 kbs or less from dissociating

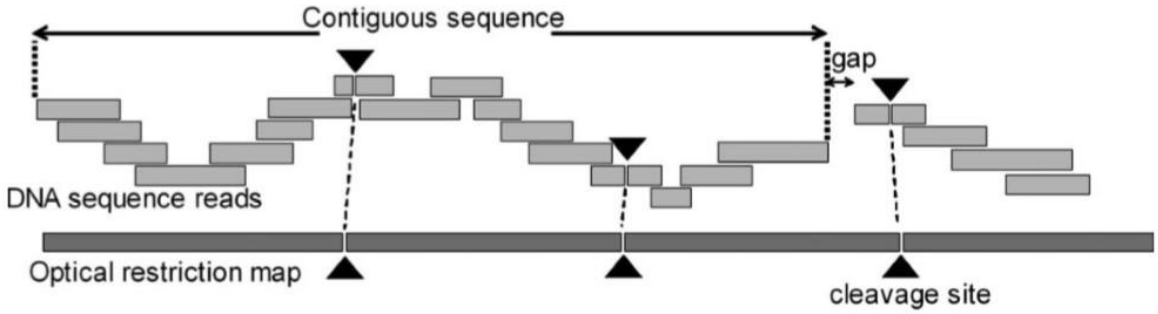


Figure 1.19: Optical mapping can be used in conjunction with sequencing to provide information on location gaps, and the position of contiguous sequences. Taken from Neely *et al* (2011).

from the surface. This means that when the region of interest on the genome is fairly small, as may be the case seeing as the average human gene is only 3 kb, this method is impractical.

1.5.2 Nicking enzymes for optical mapping

Optical mapping with nicking enzymes, uses these enzymes for labelling of the DNA, rather than cutting as in restriction mapping. This is a technique that has only been used for optical mapping fairly recently by Xiao *et al.*[105], despite the labelling by DNA nicking technique being first described in 1970s by Rigby *et al.*[106] Using "nickases" or nicking endonuclease enzymes, the backbone of the target sequence of the DNA is nicked to produce a single strand break. Subsequently, DNA polymerase is added to the reaction to begin DNA synthesis from the site of the nick. The polymerase can be designed to integrate a fluorescently-labelled nucleotide at the nicking site and therefore into the new (short) DNA

strand. An example of this was performed by Xiao *et al.* using the nicking enzyme Nb.BbvCI[107] to introduce a break into the DNA strand of interest, followed by the DNA polymerase integrating Tamra-ddUTP at the nicking site for labelling. The sample was then combed onto a surface for visualisation and localisation of fluorophores via fluorescence microscopy. Using nicking enzymes for optical mapping does have many advantages, mainly the highly specific labelling of target DNA sequences with fluorescent dye. However, as nicks can occur in DNA naturally, this can lead to non-specific labelling by the polymerase, which is a limitation of the technique. This covalent modification approach allows the sample to be extended and analysed after labelling through nanofluidic devices[108].

1.5.3 MTase-directed optical mapping

In 2010, Neely *et al.* proposed a novel idea for mapping using DNA MTases[109]. This optical mapping concept involves direct observation of single molecules of DNA stretched via combing and using MTase enzymes to fluorescently label the DNA sequence specifically. This novel technology allows analysis of the DNA sequence without compromising the sequence's integrity, providing an ordered optical map. The research performed by Neely *et al.* highlights the potential in using MTases as a way to label DNA with both a high level of specificity and at a high density, providing a "DNA fluorocode". This fluorocode would be a simple representation of the DNA sequence, which after imaging can be read as a barcode, as seen in **Figure 1.20**[109].

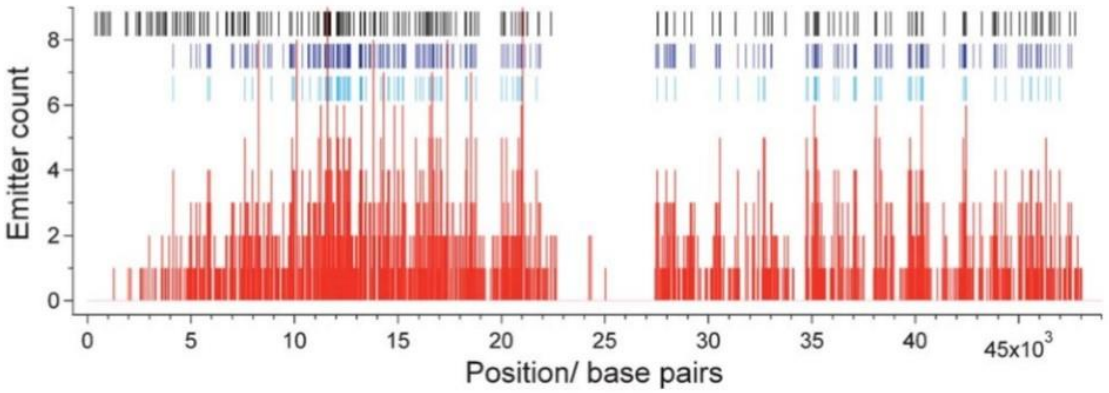


Figure 1.20: Histogram displaying localised fluorophores along lambda DNA. Black lines represent sites of M.*Hha*I, and can be used to produce a DNA fluorocode. Taken from Neely *et al.* (2010).

Due to the range of MTases with different sequence specificities, this study suggests an established toolbox of MTase enzymes each with an appropriate synthetic cofactor would be highly beneficial for labelling and highlighting different mutations or methylation states in various disease states, as well as aiding genome assembly. By using sub-diffraction limit imaging techniques, combined with the highly specific MTase M.*Hha*I (which recognises 5'-GCGC-3') providing high labelling density, Neely *et al.* were successfully able to localise fluorophores with great precision at the single gene level. In order to localise tags, to a precision of ~76 bps, that were in close vicinity and therefore overlapping each other, photobleaching was used to "turn off" emitters in turn. One of the main strengths of this method over other optical mapping techniques is the ability to map genomes in a single experiment, without the need for initial sequencing studies. Neely *et al.* successfully showed that the fluorocode is able to achieve a barcode-like representation of a DNA sequence in the absence of a reference genome, which is therefore able to detect variations that other techniques cannot, such as CNV. Another advantage of MTase directed optical mapping over using restriction enzymes is the ability to provide a much higher density of DNA labelling, with approximately one site every 650 bp, as well as higher precision in localisation of labelled sites. Methylation *in vitro* can also be detected, as the transfer of functional group to

the target sequence will be impossible if the region is already methylated. Comparing results from this sample with a generated reference map *in silico,* means missing fluorescent labels can be detected as methylated sites. This could be of great use when detecting biomarkers for disease. Multicolour labelling of DNA, using both high and low density MTases, could be the next step for MTase-directed optical mapping using two or more MTases to ensure even more confidence in the accuracy of the fluorocode interpretation.

1.6 Concluding remarks

Genetic techniques have evolved rapidly over the past few decades, but there are still limitations to these techniques that prevent detection and diagnosis of various mutations. Developing these technologies further could make it possible to discover new areas of the human genome that have previously been inaccessible, or could provide novel diagnostic and carrier detection tests for different mutations and diseases.

Cytogenetic techniques are effective in the detection of large rearrangements such as translocations and chromosomal aneuploidy, but lack the resolution to determine SNPs and other small mutations. In contrast, molecular diagnostics (e.g. NGS) provides single base pair resolution, but loses the contextual information necessary to retain the sequence position within the genome. This makes larger rearrangements difficult to visualise, as well as diseases where CNV plays a role. Ensemble averaging of amplified sections (a step crucial for the NGS protocol) may also cause problems for diagnosing residual diseases characterised by a small subset of abnormal cells, such as in leukaemia. With the emergence of SMRT and nanopore sequencing, this also demonstrates that there is a trade-off between throughput (and cost), and the long-range information that is necessary for CNV detection.

Optical mapping can provide a solution to the limitations of these traditional techniques by providing long-range contextual information. Combining this technology with highly specific MTase-labelling could allow the visualisation of single DNA molecules, while maintaining the sequence position within the genome. MTase-directed labelling has many benefits over other labelling techniques:

- High specificity ensures that non-specific labelling does not occur;

- Multiple MTases with different recognition sequences can label at variable densities, providing unique intensity profiles for different DNA sequences of interest;
- No damage to the DNA itself, as with nicking enzymes.

The emergence of new probe designs for FISH have also allowed this technique to continue to evolve, and it is gaining increased recognition in its ability to provide a physical map of both small and large genetic mutations. Using oligoprobes for FISH is becoming a popular way to amplify signals, highlighting regions of the human genome that are normally only accessible via sequencing techniques, due to their high specificity. With further development, oligoprobes could be used to visualise SNPs while maintaining sequence context within the genome. There is also potential in these oligoprobes to highlight genetic abnormalities much quicker than current diagnostic techniques, making it a favourable approach for diagnosing diseases that need prompt diagnosis, such as leukaemias.

1.7 Research aims

New diagnostic techniques emerge and evolve as we understand more about the human genome, and there has been substantial progress towards novel techniques to discover and detect links between genetic instability and disease – but it is apparent that there are limitations stopping these technologies from reaching their full potential. This thesis shows how MTases can be used to explore different regions of the human genome – at a high resolution without the need for sample amplification – for various clinical applications.

1.7.1 DNA labelling technology

Using MTase enzymes to deliver functional groups to target DNA sequences with high specificity and at a high density can have numerous practical applications, such as detecting structural genomic rearrangements that can also be biological markers for certain diseases, or for reliably labelling DNA for use in diagnostic tests such as FISH. These techniques are also suited for rapid pathogen identification, which could aid hospitals in strain typing bacteria in samples.

- Optimise the MTase-directed labelling of DNA, primarily with M.*Taq*I; currently synthetic cofactors do not lead to complete labelling with fluorophores.
- Development of MTase toolbox to be used with modified cofactors for DNA mapping.
    - Directed sequence engineering of a range of MTases at specific conserved residues for improved transalkylation with synthetic cofactors.
    - Express and screen a range of mutated proteins with a variety of cofactors to see which demonstrate efficient transalkylation activity.

- Achieve both high- and low-density labelling by producing enzymes that recognise different length recognition sequences, to produce a dual colour map.

1.7.2 Oligoprobes for FISH

By labelling probe sequences using MTase based labelling techniques, it is possible to design probes with any number of fluorophores attached simply by including the specific MTases' recognition sequence within the probe sequence. This is an important feature of this probe design, as it is not limited in the amount of labels that can be added, therefore improving sensitivity when viewing images of samples.

- Optimise the labelling of FISH oligoprobes using MTases technology.
- Optimise the following parameters to ensure strong signals and no cross-hybridisation:
  - Probe design;
  - hybridisation conditions (buffer and temperature);
  - washing conditions (stringency);
  - hybridisation time.
- Investigate the sensitivity of MTase-labelled FISH probes.
- Explore the effects that oligoprobes have on hybridisation times.
- Design and produce centromeric probes for chromosome 1, 7 and 17 – associated with acute lymphoblastic leukaemia (ALL) – as suggested by collaborators at the West Midland Regional Genetics Laboratory (WMRGL).

    - Create a probe cocktail with all three probes labelled in three spectrally-distinguishable colours that will provide a quick diagnosis within one screening.
- Design, produce and optimise probes to target single gene loci (p53 and BCR).

1.7.3 Detection of point mutations

MTase labelling technology could also provide potential in diagnosing disorders and abnormalities that contain SNPs or other small structural changes that current genetic techniques struggle to detect. Spinal Muscular Atrophy (SMA), a neuromuscular disorder, is the most common genetic cause of death in infancy[110]. The disease is characterised by mutations and therefore loss of functionality in the gene SMN1[111]. Nearly identical gene SMN2, which only has one critical nucleotide difference, can be present in variable numbers in patients and therefore restore some of the functionality lost from the SMN1 mutation[112]. This can result in varying levels of severity of the disease. Due to the similarity in sequence, FISH cannot currently be used to distinguish between SMN1 and SMN2, which is critical for carrier detection. Due to variations in the arrangement of the SMN1 gene, with some carriers having two copies of the gene on a single chromosome, and none on the other – discussed in greater detail in **3.1.2 –** this also causes problems in diagnosis using molecular techniques.

- Explore the use of oligoprobes to distinguish between highly similar sequences, such as the SNP in SMA
- Investigate the potential of DNA mapping with MTases to determine slightly different DNA sequences

- M.*Taq*I and M.*Hinc*II have overlapping recognition sites, TCGA and GTYRAC respectively, investigate whether methylating a DNA sequence with M.*Hinc*II can subsequently block the labelling of M.*Taq*I, thus allowing the detection of a slightly altered sequence.
- M.*Hpy*188I – the recognition sequence of which is affected by a SNP in gene SMN1 – could be expressed, purified and screened for potential in SMN1/SMN2 detection after proof-of-concept mapping experiments using M.*Taq*I and M.*Hinc*II.

CHAPTER TWO
Methods and materials

Methods and materials

2.1 Molecular biology

2.1.1 Alignment and sequence engineering

Sequences for wild type M.*Hha*I, M.*Sfo*I and M.*Bsa*WI were obtained from REBASE[113], and Jalview was used to align and identify the conserved regions (**Figure 1.17**). Double point mutations were made at the same non-essential site in each sequence based on the promising transalkylation activity of M.*Hha*I Q82A N304A[76]. The genes were synthesised by IDT DNA, and a Gibson Assembly (**2.1.2**) was performed to insert the gene for the mutated MTases into new vector pRSET-B. These MTases will be expressed and their efficiency with AdoMet analogues (synthesised by Andrew Wilkinson and Krystian Ubych) determined.

2.1.2 PCR and Gibson assembly®

Gibson assembly is a molecular cloning technique that joins multiple DNA fragments – such as genes into new plasmids – in a single isothermal reaction[114]. Primers (**Table 2.1**) were designed for PCR of the fragments needed for Gibson assembly to sub-clone mutated MTase sequences into pRSET-B, **Figure 2.1**, ahead of protein expression.

Plasmid/Gene	**pRSET-B**	**M.H*ha*I Q82A N304A**	**M.S*fo*I T77A D360A**	**M.*Bsa*WI E83A D384A**
Reverse Primer 5' - 3'	CGGATCCTTATCGTCATC	cggatcaagcttcgaattctTTCCAGTTAATACGGCTTG	cggatcaagcttcgaattctTTCCAATTAGCTCGCCTTG	cggatcaagcttcgaattctTTCCAGTTAGACGCCCTC
Forward Primer 5' – 3'	AGAATTCGAAGCTTGATCC	acgatgacgataaggatccgGTGGATCGTATCGAGATC	acgatgacgataaggatccgGTGGATATGCGCTTTGCTG	acgatgacgataaggatccgGTAGACATGACCCGTCGTC

Table 2.1: Primers ordered for Gibson Assembly to subclone mutated MTases into new expression vectors.

A standard NEB Q5 High Fidelity protocol was used in order to amplify the desired fragments using PCR. After amplification, fragments were confirmed by gel electrophoresis, **2.1.5**. Manufacturer's instructions were then followed to perform Gibson assembly using NEB Gibson assembly kit.

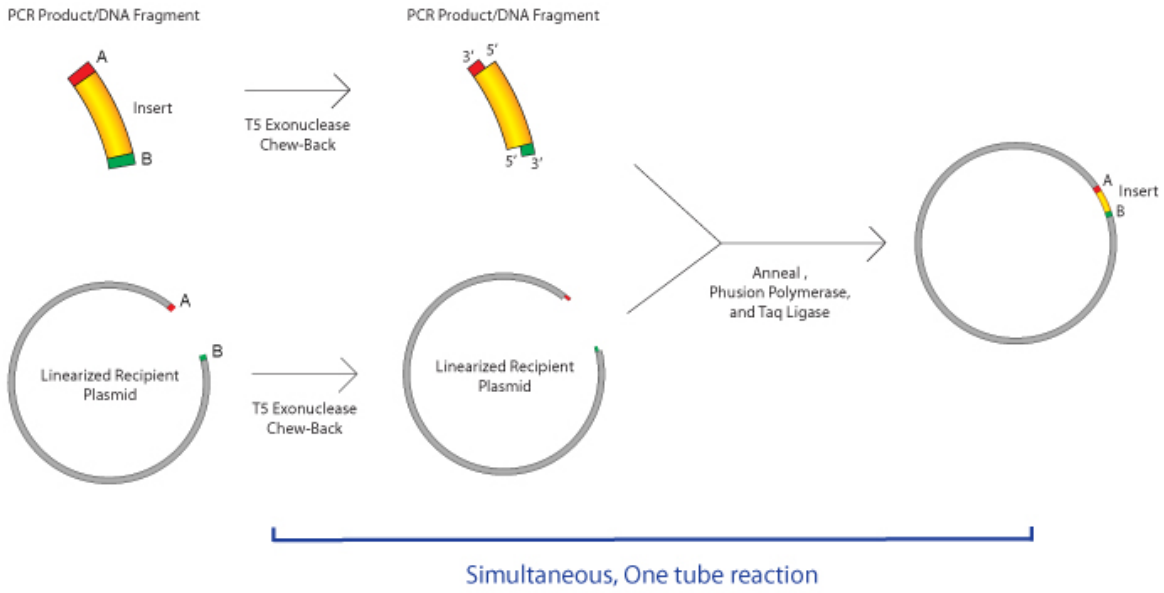


Figure 2.1: Gibson assembly process where "insert" is the gene to be subcloned, into the recipient plasmid. A and B are the forward and reverse primers used to amplify this fragment during PCR before annealing exonuclease chew-back and joining via a ligase reaction. Taken from https://www.addgene.org/protocols/gibson-assembly.

2.1.3 Restriction digests

To ensure that the sub-cloning of mutated sequences was successful, samples were restricted and analysed using gel electrophoresis, before being sequenced. After purification of DNA, a restriction digest was carried out using enzymes BamHI (G^GATCC) and EcoRI (G^AATTC), both ordered from NEB. For each sample two tubes were set up with either BamHI or EcoRI. On ice, 1 µl restriction enzyme (BamHI or EcoRI), 4 µl ThermoFisher 10X FastDigest Green Buffer, 800 ng DNA and water up to 50 µl total volume were mixed. Samples were incubated at 37 ^{o}C for 20 minutes before 40 µl from each tube per sample was mixed into a single fresh microcentrifuge tube to create a double digest. The samples were incubated at 37 ^{o}C for a further 40 minutes. After incubation, 10 µl per sample were analysed using gel electrophoresis, as described in **2.1.5**. Samples were sent to sequencing for confirmation and then expressed.

2.1.4 Sequencing

MTase DNA sequences were confirmed by sequencing using services provided by the School of Biosciences at the University of Birmingham. Typical reactions were made up as shown in **Table 2.2**, and submitted using universal primers designed for the T7 promotor and terminator region.

	Amount
DNA	~500 ng
Forward primer 5’ TAATACGACTCACTATAGGG 3’	10 µM
Reverse primer 5’ GCTAGTTATTGCTCAGCGG 3’	10 µM
Water	up to 10 µl final volume

Table 2.2: Sample prep requirements for sequencing DNA samples in a plasmid with a T7 promotor.

2.1.5 Gel electrophoresis

DNA was analysed using gel electrophoresis for validation of the sequence. Gel electrophoresis is a technique used to separate DNA fragments according to size by running an electric current across the agarose gel, **Figure 2.2**. As DNA fragments are negatively charged, they are pulled towards the positive electrode, with smaller fragments travelling faster, and therefore further, down the gel. This means that the fragments can be identified according to how far down the gel they have travelled. Gel electrophoresis can also be used to check MTase activity.

Unless otherwise stated, a 1 % agarose gel was prepared by mixing 80 ml TAE buffer and 0.8 g agarose. This was heated until the agarose had dissolved, and then poured into the electrophoresis equipment to set. 2 µl 6x NEB loading buffer was added to 10 µl sample and loaded into the lanes, alongside 5 µl NEB 2-log ladder. The gel was run at 120 V for around 45 minutes and then left to stain in Gel Red for 30 minutes. A UV visualiser was used to acquire an image using EtBr filter.

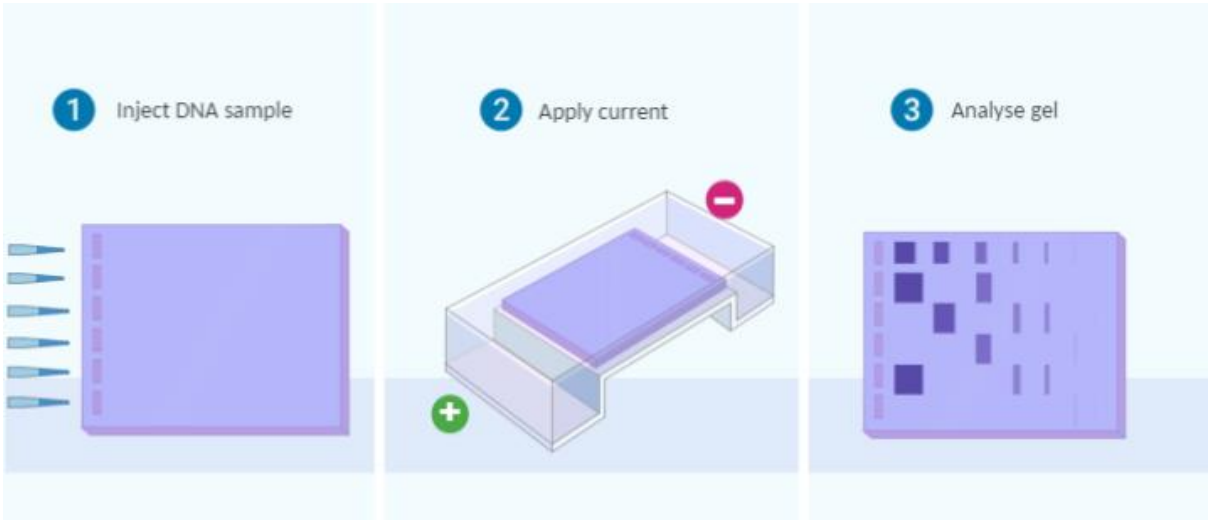


Figure 2.2: Gel electrophoresis separates DNA samples by applying an electric current across an agarose gel. Small, negatively charged DNA fragments run faster through the gel, leaving a distinct size-based pattern that can be used for DNA identification.

2.1.6 Preparation of LB broth and LB plates

LB broth was prepared by mixing 20 g Sigma LB Broth (Lennox) with 1 L water and autoclaving. Once cooled, the appropriate antibiotic was added to the correct working concentration, **Table 2.3**.

LB plates were prepared by autoclaving 500 ml water and 20 g Sigma LB Broth with agar (Miller). Once cooled, the appropriate antibiotic was added to the correct working concentration. Plates were poured in sterile conditions and left to set.

Plasmid	Size (bp)	Antibiotic resistance	Working concentration (of antibiotic)
pET-28c	5367	Kanamycin	100 µg/ml
pRSET-B	2900	Ampicillin	50 µg/ml

Table 2.3: Expression vectors and their appropriate antibiotic (and concentration) for growth.

2.1.7 Bacterial transformation

Competent *E. coli* strains were selected for either DNA amplification (NEB Turbo Competent *E. coli* (High efficiency) – C2984) or protein expression (NEB T7 Express Competent *E. coli* (High efficiency) – C2566) and thawed on ice. In a microcentrifuge tube, 1 µl of DNA stock (up to 1 ng DNA) and 25 µl competent cells were mixed and lightly triturated before being left to incubate on ice for 30 minutes. The sample was heat shocked at 42 ^{o}C for 30 seconds, then returned to ice for 5 minutes. 495 µl room temperature SOC media was added and the sample was left in a shaking incubator, at ~250 rpm and 37 ^{o}C, for 30 minutes to 1 hour. Bacteria was spread on 3 LB plates (containing the appropriate antibiotic), at different volumes; 25 µl, 50 µl and 300 µl before being left at 37 ^{o}C overnight.

The following day, 3 colonies per plate were picked using a pipette tip, and each placed into 5 ml LB media (with appropriate antibiotic) into 50 ml falcon tubes. Samples were again left to incubate overnight in a shaking incubator at ~250 rpm and 37 ^{o}C. For DNA amplification;

samples were purified using a standard Qiagen QIAprep Spin Miniprep Kit the following day, and the DNA was analysed using gel electrophoresis to check that the DNA present was of the correct size, (**2.1.5**) before being sent for sequencing to confirm the sequence was also correct (**2.1.4**). For protein expression the protocol was continued as in **2.1.8**.

2.1.8 Protein expression

The following day, two colonies were picked in sterile conditions using a pipette and placed in 20 ml LB media plus antibiotics (see **Table 2.3**). These starter cultures were left shaking overnight at 37 ºC, 250 rpm. The next day, a 1/80th dilution of the starter culture was made i.e. 40 ml in a 2 L conical flask with 400 ml LB media (with appropriate antibiotic). This was left shaking at 37 ºC, 180 rpm, until reaching an OD A600 of 0.4-0.6. Once at OD 0.4-0.6, IPTG was added to final concentration of 0.5 mM and left shaking at 20 ºC, 180 rpm for 16 hours (unless otherwise stated).

2.1.9 Cell lysis

Cells were spun down at 4 ºC, 4,000 rpm for 20 minutes and the supernatant discarded. Pelleted bacteria cells were resuspended in 20 ml PBS with protease inhibitor and spun again at 4 ºC, 4,000 rpm for 12 minutes. Lysozyme was added to fresh PBS with inhibitor to final concentration 4 mg/ml, the supernatant was again discarded, and resuspended in 25 ml of this solution. The sample was sonicated four times in 30 second on/off pulses before being spun back down at 4 ºC, 4,000 rpm for 14 minutes. The supernatant was collected and pellet discarded. Attempts were made to remove residual AdoMet from the protein complex for some experiments at this point, described in **2.1.11**.

2.1.10 Protein purification

1 ml Ni-NTA agarose beads were spun at 500 x g for 5 minutes, the storage buffer removed and replaced with 1 ml PBS. The beads were washed by inverting the tubes a few times and then being spun in the centrifuge at 500 x g for 5 minutes. This was repeated five times. 1 ml PBS was added to the beads to create a 50 % slurry, 50 µl of which was mixed per 1 ml lysate and left on an end-over-end spin for 1 hour at 4 ºC. PBS was poured through Biorad EconoColumns to wet, before the sample was dropped through the column with a pastette and collected. The sample was run through the column followed by 25 ml (100x the column volume) of wash buffer (PBS, 300 mM NaCl, 20 mM imidazole). 5 ml elution buffer (PBS, 250 mM imidazole) was finally poured into the column, left for 10 minutes and then eluted and collected. An Amicon Ultra 0.5 ml 10 kDa kit was used to exchange the buffer following the manufacturer's instructions. Storage buffer was made up of PBS, 5 mM EDTA and 5 mM β-mercaptoethanol. Protein presence was checked using SDS-PAGE as described in **2.1.12**, and stored in 50-60 % glycerol at -20 ºC.

2.1.11 Removal of bound AdoMet from M.*Taq*I

In an attempt to remove residual AdoMet from M.*Taq*I protein, a palindromic oligo containing the M.*Taq*I recognition sequence (5′CCGCC*TCGA*GGCGG3′) was annealed by incubation at 95 °C for two minutes, before cooling at room temperature overnight. Two equivalents of M.*Taq*I were added to oligos sample before purification, and the mixture was incubated at 50 °C for 30 minutes.

2.1.12 SDS-PAGE

The presence of protein from each eluted fraction was checked using SDS-PAGE. 20 µl of sample was mixed with 5 µl 4x Laemmli Sample Buffer, heated at 100 ºC for 5 minutes, and

spun in a microcentrifuge for 30 seconds. 5 µl ladder (NEB Blue Prestained Protein Standard, Broad Range (11-190 kDa)) was loaded into a precast Protean 4-20 % gel, along with 20 µl of sample and run at 200V for around 45 minutes. The gel was stained in Instant Blue on a rocker for 2 hours, and then washed in water for 1 hour. Gels were visualised on a UV visualiser. Fractions that showed high concentrations of protein based on analysis after SDS-PAGE were pooled together and concentrated using the Amicon Ultra 0.5 ml 10 kDa kit and stored in 50-60 % glycerol at –20 °C.

2.1.13 Western blot

Western blots were performed using anti-His antibodies to check that the bands from the SDS-PAGE gel were the correct protein (purified using their His-tag). 40 µl of each sample was taken pre- and post-purification and boiled at 100 ºC for 5 minutes. Samples were then analysed using SDS-PAGE as in **2.1.12**. A Biorad Trans-Blot turbo transfer cassette was used for the Western blot. A 50 ml solution was made up of 5 % milk in 0.1 % PBST (phosphate buffered saline with Tween-20) and 2.5 g of Marvel original dried skimmed milk, which blocks background proteins. After SDS-PAGE, the gel was placed between the transfer pack filters which were set up as described by the manufacturer. Transferred layer was covered in the 5 % milk 0.1 PBST solution and gently rocked at room temperature for 1 hour. The filter membrane was placed in a pouch with 10 ml of the milk solution and 3 µl anti-His (H1029) antibody from Sigma (a 1/3,000 dilution as per instruction). The pouch was sealed and left rocking overnight at 4 ºC.

The following day, the sample was retrieved and washed in PBST, by gentle rocking for 5 minutes at room temperature. This was repeated with fresh PBST 5 times. The sample was

then covered in 5 % milk 0.1 % PBST with 9 µl (a 1/3,000 dilution) of secondary antibody (CST antimouse IGG (70765, Cell Signaling Technology)). This antibody is conjugated to horseradish peroxidase (HRP) which can be used for chemiluminescent detection, this offers a means to identifying a His-tagged protein. The sample was left gently rocking for 1 hour at room temperature, before the solution (and antibody) was poured off and washed 5 times with PBST (for 5 minutes rocking each time). Excess liquid was gently shaken off the sample before it was placed on cling film. 1 ml of detection reagent (1 and 2) from the ECL kit was added, and left for 5 minutes to develop a signal. The sample was shaken gentle to remove excess liquid, wrapped in sling film and placed in hypercassette. High performance radiography film was placed over the sample and visualised in X-ray developer, with an exposure time of 2 minutes. If bands are present in the image, then His-tagged protein is present.

2.1.14 Protection assay

Activity of protein was checked using a protection assay. Active MTases, in optimal conditions, should methylate DNA, blocking restriction by corresponding restriction enzymes that have the same recognition sequence (e.g. MTase M.*Taq*I and restriction enzyme R.*TaqI* both recognise the sequence, TCGA). This protocol was adapted depending on the MTase and cofactor tested. The following is a general protocol for testing M.*Taq*I methylation with AdoMet, but can also be used for other cofactor analogues such as to test the alkylation activity with AdoHcy-6-N_3.

On ice, a master mix was created by mixing 67 µl molecular grade water, 8 µl 10x NEB CutSmart buffer, 4 µl pUC19 (1,000 ng/µl), 1 µl AdoMet (32 mM) (or 2 µl AdoHcy-6-N_3

(15mM) for final cofactor concentration of ~ 375 mM, and reduce volume of water to total reaction volume 80 µl). The master mix was split into 1x 20 µl and 5x 10 µl and labelled 1-6. A 2x serial dilution was made by adding 1 µl MTase (1 mg/ml) to tube 1 and mixing before adding 10 µl from tube 1 to 2 and continuing until tube 6 (discarding the final 10µl leaving 10 µl in each tube).

The following controls were prepared as shown in **Table 2.4**.

	AdoMet control (Tubes 7&8)	**No cofactor (9&10)**	**No MTase (11&12)**
10x NEB CutSmart	2 µl	2 µl	2 µl
pUC19 (1,000 ng/µl)	1 µl	1 µl	1 µl
SAM (3.5 mM)	0.5 µl	-	-
M.*Taq*I	0.5 µl	0.5 µl	-
Water	16.5 µl	17 µl	17 µl

Table 2.4: Controls set up in protection assay.

All samples were incubated at 50 ^{o}C for 1 hour before adding 0.5µl restriction enzyme (R.*Taq*I) to all tubes except 8, 10 and 12. Samples were again incubated for 1 hour at 65 ^{o}C. 0.5 µl proteinase K was added to all tubes and incubated at 50 ^{o}C for 1 hour before being analysed using gel electrophoresis (**2.1.5**).

Other MTases and cofactor analogues were tested throughout this thesis, using the following buffers. Different MTases have different optimum active temperatures, so this should be determined beforehand, **Table 2.5**. Note that restriction enzymes also require different incubation temperatures as stated by the manufacturer, and this should be considered when performing the assay.

MTase	Recognition sequence	Number of sites on pUC19	Incubation temperature (º C)
M.*Taq*I	TCGA	4	50
M.*Hinc*II	GTYRAC	1	37
M.*Hha*I	GCGC	17	50
M.*Bsa*WI	WCCGGW	3	50
M.*Sfo*I	GGCGCC	1	37

Table 2.5: Table showing different MTases used in protection assays in this thesis, their recognition sequence, and their optimal incubation temperature.

As some buffers contain the chelating agent EDTA, **Table 2.6**, $MgCl_2$ was added (to final concentration 20 mM) in an additional step before restriction to counteract its effects. The presence of $MgCl_2$ could prevent the activity of the restriction enzymes, leading to false positive results (i.e. the DNA looks to be protected, but has actually not fully restricted).

Buffer	Recipe
Cutsmart (NEB)	20 mM tris-acetate, 10 mM magnesium acetate, 50 mM potassium acetate, 100 µl/ml BSA
NEB2 (NEB)	50 mM Tris-HCl, 10 mM $MgCl_2$, 100 mM NaCl, 1 mM DTT
NEB2.1 (NEB)	50 mM Tris-HCl, 10 mM $MgCl_2$, 100 mM NaCl, 100 µl/ml BSA
Low salt buffer	50 mM Tris-HCl, 15 mM NaCl, 0.5 mM EDTA
HincII buffer	50 mM Tris-HCl, 5 mM β-mercaptoethanol, 10 mM EDTA

Table 2.6: Table showing different buffers that were used and their ingredients.

2.2 Fluorescence *in situ* hybridisation

2.2.1 Annealing centromeric hairpin probes

For centromeric hairpin probes, oligos were ordered from IDT and resuspended at a concentration of 100 µM in Tris-EDTA (10 mM Tris and 1 mM EDTA at pH 8). 50 µl of oligo were placed in PCR tubes and heated to 85 ºC for 5 minutes. Hairpin oligos were immediately placed on ice for rapid cooling to form the hairpin structure.

Sequences for oligos were ordered from IDT DNA (**Table 2.7**) following the standard hairpin sequence: CCCTCGATCGATCGATCGACCCTTTTGGGTCGATCGATCGATCGAGGGTTTT

Chromosome	Sequence (5'–3')
1	TTTCAACCTGAACTCACAAG
1	CTCATCAAAGCTACATGGAA
7	AGCGATTTGAGGACAATTGC
7	CCACCTGAAAATGCCACAGC
17	ATCATTGCACTCTTTGAGGAGTACCG
17	ATAATTGCACTTCTTTGAGGCCTACCG

Table 2.7: Table showing different recognition sites for centromeric probes 1, 7 and 17. These sequences were proceeded by the same hairpin sequence.

17CEN sequences were taken from O'Keefe *et al.* (1996)[44].

1CEN sequences were taken from Pironon *et al.* (2010)[115].

7CEN sequences were taken from Waye *et al.* (1987)[116].

2.2.2 Annealing docking and imaging strand probes

For probe designs involving a docking and imaging probe, new sequences were ordered for 17CEN. The docking strand was a single stranded piece of DNA which hybridised directly to the DNA. This sequence was:

5'

ATCATTGCACTCTTTGAGGAGTACCG TTTTTT GGGT GGTT GTTT GTGT TTTG TG TG TTGG 3' for 17CEN 1 and:

5’

ATAATTGCACTTCTTTGAGGCCTACCG TTTTTT GGGT GGTT GTTT GTGT TTTG T GTG TTGG 3’ for 17CEN 2.

The imaging strand was the same, regardless of the target:

5’

GGTCGAGGTCGAGGTCGAGGTTTTCCTCGACCTCGACCTCGACCTTTTCCAACAC ACAAAACACAAACAACCACCC 3’

All oligos were ordered from IDT and resuspended to 100 µM in 1x Tris-EDTA. 50 µl of oligo were placed in PCR tubes and heated to 85 ºC for 5 minutes. The single stranded docking oligos were left to cool slowly overnight, at room temperature, to prevent them forming secondary structures. The hairpin imaging strand was placed directly onto ice for 5 minutes to form the hairpin structure.

2.2.3 Annealing single gene hairpin probes

Oxford Gene Technology (OGT) kindly aided probe design for oligoprobes for the BCR gene, with a hairpin on the end for MTase labelling. 89 potential ROIs were sent from OGT, selecting target regions approximately 350 bp apart, which targeted the BCR gene specifically and fit the parameters needed for the oligoprobe conditions. From the 89 sequences, 83 met the specification of being < 60 bases in length (once M.*Taq*I labelling sites (CCC TCG ACC CTT TTG GGT CGA GGG TT) had been added), ~55 % GC content, and T_m of ~70 °C. These specifications were required in order to keep cost of the oligos low, as well as ensuring that they had similar properties and would hybridise under the same

conditions. Probes were ordered in a 96 well plate from IDT in 1x TE to a concentration of 100 µM, all of which can be seen in **8.2**. 1 µl of each oligo was mixed into a PCR tube and heated to 85 ºC for 5 minutes. The oligos were immediately placed on ice for rapid cooling to form the hairpin structure.

2.2.4 Fluorescently labelling oligoprobes with DBCO dyes

Oligoprobes were labelled site-specifically using M.*Taq*I (produced by myself in the Protein Expression Facility) and AdoHcy-6-NH_2 (synthesised by Andrew Wilkinson), a scheme for this reaction is shown in **Figure 1.16**. 4 µl 10x cutsmart MES pH 5.7 was mixed with 2 µl oligos (diluted 10x in Tris-EDTA), 1 µl AdoHcy-6-NH_2 (15mM), 0.5 µl M.*Taq*I (1 mg/ml) and water (up to 40 µl total volume). The sample was incubated at 50 ºC for 1 – 1.5 hours. 0.5 µl proteinase K was added and incubated for 1 hour at 50 ºC before sample was purified using mini Quick Spin Oligo (Sigma-Aldrich) sephadex columns (column buffer 10 mM Tris-HCl pH 7.5, 1 mM EDTA, 100 mM NaCl). 10 µl DMSO was added to samples, followed by the addition of 1 mM DBCO dye (in DMSO), **Table 2.8**, in 2x excess to M.*Taq*I sites. Sample was left at room temperature overnight for coupling. The following day samples were purified again using Quick Spin Oligo columns to remove any excess dye.

Dye	Source	Reactive group for conjugation	Excitation λ max (nm)	Emission λ max (nm)
TAMRA	Jena Biosciences	DBCO	560	565
Rhodamine Green	Jena Biosciences	DBCO	501	526
Alexa 647	Jena Biosciences	DBCO	651	672
Atto 647N	Sigma Aldrich	NHS-ester	647	661

Table 2.8: Various dyes used for MTase labelling of DNA.

2.2.5 Fluorescently labelling oligoprobes with NHS-ester dyes

For probes to be labelled with NHS-ester dyes and the azide cofactor AdoHcy-6-N_3, an additional step is required. This reaction yields enough dye mix for labelling three probes (three lots of 10 µM oligo). 1.15 µl PBS, 0.8 µl DMSO, 0.175 µl DBCO-amine (20 mM) and 0.21 µl Atto647N (or other NHS-ester dye) (50mM) were mixed and left at 4 °C for 1 to 3 hours. Note, a final concentration of ~60 µM amine-linker is needed per 10 µM labelling sites. Samples were then prepared as in **2.2.4** up until after the first purification with Sephadex columns. 3.5 µl of DMSO was then added to the dye mix and left at room temperature for 15 minutes. 1.68 µl dye mix was taken per sample and added to 8.32 µl DMSO for a 20 % DMSO mixture (10 µl total). The full 10 µl dye/DMSO mix was added to each probe and left covered overnight at room temperature. Samples were purified using Qiagen QIAquick Nucleotide Removal Kit and eluted into 50 µl water.

2.2.6 Preparation of patient sample slides

Anonymised 46 XX/XY white blood cell suspensions were kindly provided by West Midlands Regional Genetics Laboratory (WMRGL).

Patient slides were prepared at WMRGL using their standard slide making SOP. Samples were spun at full speed for 5 minutes and the supernatant poured off (performed in one quick pour, note that some liquid will remain in the bottom of the tube). Meanwhile, slides were cleaned in a hood with methanol and tissue. Using a fine tip pipette, a drop of sample was added to the slide and left to dry, before viewing on an optical microscope to detect the presence of cells. If the sample was too thin i.e. not many cells present, another drop may be added. If the sample is overcrowded, a drop of fixative (3:1 methanol:acetic acid) is added to to the tube and resuspended, before being dropping onto a fresh slide as above.

2.2.7 Probe hybridisation for hairpin probes

50 ml denaturation solution (0.07M NaOH, 70 % ethanol) was heated to 72 ºC. 46 XX/XY fixed patient slides containing interphases and metaphases were placed in denaturation solution for 2 minutes followed by a dehydration series (70 %, 85 %, 100 % methanol) for 2 minutes each. Slides were left briefly to airdry. For the original hairpin probes, 5 µl of each variant (or single gene) probe were mixed (~75 ng DNA each) with 5 µl hybridisation buffer (6 mM NaOH, 40-70 % formamide, 20 % dextran sulphate) and 1µl COT-1. For the docking/imaging strand probes, 2.5 µl of each docking strand (17CEN1 and 17CEN2) were mixed with 5 µl imaging strand and 5 µl of hybridisation buffer (40 % formamide). The whole volume was applied to slide. A coverslip was placed over sample

and bubbles removed with pipette tip. Samples were then hybridised at 37 ºC for 15 minutes (1 hour for single gene probes).

2.2.8 Washing slides

The coverslip was removed from slides. Slides were then washed for 5 minutes at high stringency 0.4x SSC/0.3 % NP-40 at room temperature, followed by 5 minutes at low stringency RT 2x SSC/0.1 % NP-40. 10 µl DAPI (nuclear stain Ex/Em 358/461 nm) was applied to each slide to counterstain, and coverslip placed on top. Bubbles were removed using a clean pipette tip.

2.2.9 Imaging slides

Samples were imaged using an inverted, epifluoresence microscope equipped with a 100× objective lens (Nikon, 1.49/oil TIRF) and cooled EMCCD camera (Photometrics, Evolve® 512 Delta). Excitation at 405 nm, 488 nm, 561 nm and 640 nm was achieved using solid state lasers (Coherent, OBIS) to visualise DAPI (nuclear stain), Rhodamine Green and TAMRA/Atto 647N/Alexa 647 respectively. A quad-band filter set for 405 nm, 488 nm, 561 nm and 640 nm lasers was used. Images were analysed using FIJI (Image J).

2.3 DNA mapping

2.3.1 Methylation of lambda DNA

Lambda DNA was methylated with M.*Bse*CI before being labelled with M.TaqI to see if blocked sites could be detected using mapping. The reagents in **Table 2.9** were mixed and incubated at 37 °C for one hour.

Reagent	Source	Concentration	Volume
Water	-	-	80 µl
10x Cutsmart NEB2 buffer	NEB	-	10 µl
M.*Bse*CI	Weinhold lab	4000 U/ml	3 µl
AdoMet	NEB	32 mM	1 µl
Lambda DNA	NEB	500 µg/ml	6 µl

Table 2.9: Table showing concentrations and volumes of reagents for lambda methylation.

2.3.2 Ethanol precipitation

10 µl (0.1x volume) 3M NaCl was added to DNA sample followed by addition of 200 µl EtOH. The tube was gently inverted and spun in a centrifuge at 14,000 x g for 15 minutes. The supernatant was removed and the pellet washed by addition of 200 µl 70 % EtOH. The sample was spun again at 14000 x g for 15 minutes. The supernatant was carefully discarded and pellet air-dried for 2 minutes. The DNA pellet was resuspended in 50 µl water.

2.3.3 MTase-directed labelling of lambda DNA

B*se*CI-methylated and unmethylated lambda were labelled with M.*Taq*I and AdoHcy-6-N_3 for DNA mapping. Solutions were made up for each sample, as shown in **Table 2.10**. Samples were incubated at 50 ºC for 1 to 2 hours.

	Methylated lambda		Unmethylated lambda	
Reagent	Concentration	Volume	Concentration	Volume
M.*Bse*CI methylated lambda	127 ng/µl	42.5 µl	-	-
Unmethylated lambda	-	-	500 µl/ml	4 µl
10x NEB Cutsmart buffer	-	10 µl	-	10 µl
M.*Taq*I	1 mg/ml	2 µl	1 mg/ml	2 µl
AdoHcy-6-NH_3	15 mM	2 µl	15 mM	2 µl
Water	-	43.5 µl	-	82 µl

Table 2.10: Table showing necessary reagents for labelling methylated and unmethylated lambda DNA with M.*Taq*I and AdoHcy-6-NH_3.

Meanwhile, the dye was prepared for labelling (a scheme for this is reaction is shown in **Figure 1.16**). For 2 x 100 µl reactions, the following were mixed: 8 µl 1x PBS, 2.72 µl DMSO, 2.4 µl DBCO-amine 20 mM, 2.88 µl Atto 647N NHS-ester. The dye was left at 4 ºC for between 1 to 3 hours.

After the DNA had incubated for 1 hour, 2 µl proteinase K/10 % triton was added to each sample and left at 50 ºC for 1 hour. The DNA was then purified using ethanol precipitation as

in **2.3.2** (sample can be eluted into 55 µl water if wanting to carry out an additional protection assay (**2.1.14**) at this point.

Before adding the dye to the DNA, 4 µl DMSO was added to the dye mix and it was left at room temperature for 15 minutes. 10 µl of this dye mix was then added to each sample as well as 5 µl 10x PBS. The solutions were left at room temperature overnight to efficiently label the DNA with dye.

2.3.4 Preparation of Zeonex-coated slides

20 mm x 20 mm coverslips were placed in an oven overnight at 450 °C to sterilise. 1.5 % Zeonex solution (Zeon Chemicals) was made by dissolving Zeonex in chlorobenzene (typically 1 bead in 850 µl chlorobenzene). The mix was sonicated to ensure it had fully dissolved. The following day, 30 µl Zeonex mix was added to ovened coverslips and immediately spun at 3000 rpm on a spin coater for 90 seconds. Slips were left to dry in a dessicator overnight before use.

2.3.5 Deposition of DNA on Zeonex

1.5 µl of sample was deposited on the surface of a Zeonex-coated coverslip. A pipette tip was used to contact the droplet and drag across the surface at a speed of 20 mm/min. The samples were imaged using widefield 100x at 640 nm excitation.

2.3.6 Extraction and alignment of barcodes

Custom code was written using MATLAB 2016b for automated extraction, *in silico* generation of barcodes and alignment procedures by Nathaniel Wand and Darren Smith. An annotated copy of the code is available on the University of Birmingham edata website at: https://edata.bham.ac.uk/255/.

CHAPTER THREE

Optimisation of MTase-directed labelling of DNA

3.1 Introduction

3.1.1 Enzymatic modification of DNA and diagnostics

Enzymes are powerful substrate-specific proteins which have been used widely over the past few decades. As discussed in the introduction, MTases are enzymes that catalyse the transfer of methyl groups to specific targets (DNA, RNA or protein); this thesis focuses on DNA MTases. By using the MTases' natural specificity, DNA can be modified site-specifically and analysed, determining the underlying DNA sequence. Since the 1970s[117], enzymes have aided DNA identification and genomic mapping.

There are three enzymatic techniques used for site-specific DNA modification, involving restriction enzymes[94], nicking enzymes[108], or MTases[109]. The use of restriction enzymes in restriction mapping[118] was the first of these techniques to be established. This provided a framework for the development and application of these enzymes in further mapping experiments.

Advances in DNA hybridisation techniques (e.g. FISH) and sequencing technologies (e.g. next generation sequencing (NGS)) have surpassed the use of restriction mapping in rapid DNA identification. Though they are more commonly used, both hybridisation and sequencing techniques have their own set of problems. NGS is currently at the forefront of sequencing technologies – although long-read sequencing is rapidly developing – however it still faces issues with copy number variations (CNV), ensemble averaging, and reconstruction of the genome after amplification of sequences. As this technique focuses on small base differences, it loses any larger structural information. On the-other-hand, cytogenetic techniques such as FISH, focus on much larger regions of interest (ROI). FISH is currently

the gold standard in detecting large rearrangements, amplifications, or deletions of genetic material but it is not possible to detect changes at the single-base level using traditional probes. Optical mapping attempts to overcome some of these challenges.

Since the emergence of optical mapping via restriction enzymes, nicking enzymes, and more recently MTases, have been explored[26,93,94,96]. MTases show great potential as a way to label DNA with both a high level of specificity and at a high density. This provides potential for MTases in a number of diagnostic applications which will be discussed later in this thesis. Labelling DNA sequence-specifically using MTases and stretching single DNA molecules onto a surface via combing, can provide an ordered optical map. This novel technology allows analysis of the DNA sequence without compromising the sequence's integrity and can provide a scaffold to aid genome assembly in conjunction with sequencing.

3.1.2 MTases and SNP detection

Some MTases display highly specific recognition of their target motifs. Therefore, it has been hypothesised that they could be used in detection of single nucleotide polymorphisms (SNPs). SNPs are variations of single nucleotides that occur at a specific position in a DNA sequence. This genetic variation can be the underlying cause for susceptibility to certain diseases e.g. cystic fibrosis, and also impact the severity of those illnesses[23].

Spinal muscular atrophy (SMA) is a recessive neurodegenerative disease characterised by the loss of the SMN1 gene[119]. A nearly identical gene, SMN2, has only one critical nucleotide difference. SMN2 can be present in variable numbers in patients and therefore restores some of the functionality lost from the SMN1 mutation, resulting in varying levels of severity of

the disease[112]. It is possible to be a carrier of SMA if you only have one copy of SMN1, or if you have two copies of SMN1 on one chromosome; a 2:0 "silent" carrier[120], **Figure 3.1.**

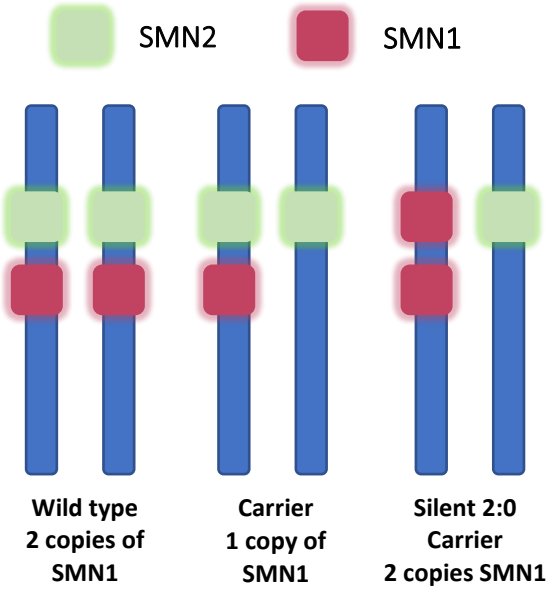


Figure 3.1: Schematic showing copy number and position of SMN1 and SMN2 in different patients. It can be difficult to determine carriers of SMA due to silent carriers with two copies of SMN1 on one chromosome (2:0 carriers). This makes it problematic to distinguish this from the wild type using molecular techniques.

Currently it is not possible to determine with 100 % certainty whether someone is a silent carrier. Molecular techniques, such as multiplex ligation-dependent probe amplification (MLPA) – a variation of multiplex PCR that amplifies multiple targets using with a single primer pair – can detect two copies of SMN1, but cannot determine if they are in the 2:0 formation or 1:1. Cytogenetic techniques, such as FISH, can also not be used for this arrangement, as they do not have the specificity to work at the single-base level.

M.*Hpy*188I is an MTase which targets TCNGA[113]. This sequence is disrupted by a single base change difference in the sequence of SMN1, but not in SMN2. If a patient's DNA could be labelled with M.*Hpy*188I and mapped, it could be possible to determine whether a

patient's SMN1 genes are in the 1:1 or 2:0 formation based on the pattern produced from the fluorophores. In this way it could be possible to detect silent carriers of SMA by locating the exact position of these genes within their genome.

3.2 Aims

The preliminary aim of this chapter was to develop an optimised procedure for expressing and purifying MTases for various labelling and mapping experiments. M.*Taq*I will be tested to ensure it can provide efficient labelling with AdoMet as well as with synthetic cofactor analogues. Expression conditions will be optimised to ensure high protein yield that can reach complete labelling of DNA.

As discussed above, SNPs can be the cause of genetic disorders, but they are troublesome to detect in certain diseases using current diagnostic techniques. As M.*Taq*I and M.*Hinc*II have overlapping recognition sites, TCGA and GTYRAC respectively, it would be interesting to see if methylating a DNA sequence using M.*Hinc*II can subsequently block the labelling of M.*TaqI*, thus allowing the detection of a slightly altered sequence. If M.*Hinc*II is not active after expression, and fails to methylate the DNA with AdoMet, attempts can also be made with M.*Bse*CI, as the recognition site (ATCGAT) also overlaps with that of M.*Taq*I. M.*Hpy*188I is a specific target for this research as its recognition sequence (TCNGA) is affected by a SNP in the SMN1 gene, which is associated with SMA. This chapter will test the activity of both M.*Hinc*II and M.*Bse*CI with AdoMet. If successful, these enzymes will be used to determine if it is possible to detect single-base differences using mapping later in this thesis.

Finally, attempts will also be made to produce various MTases with different recognition sites to be part of a labelling toolbox for mapping. This will involve engineering the MTase sequence to open the cofactor binding pocket. Mutations will be made on the MTases, M.*Hha*I, M.*Sfo*I, and M.*Bsa*WI to produce enzyme for labelling at a range of recognition

sequences. This mutation will hopefully enhance activity with larger synthetic cofactors and make dual-colour high- and low-density mapping possible.

3.3 Results and discussion

This chapter develops the production and application of MTase enzymes in DNA labelling reactions. MTases are highly specific enzymes that – when combined with synthetic AdoMet analogues – can label DNA sequences of interest with fluorophores, which can then be used for downstream diagnostic applications. This technology could be used for a range of applications, from densely labelling probes for FISH, to detecting long- and short-range genomic sequences in DNA mapping.

3.3.1 Optimisation of MTase preparation for directed-labelling of fluorophores

As discussed in the aims and introduction of this chapter, MTases will be used in labelling reactions to attach functional groups site-specifically to DNA. In order to obtain the MTase protein it must be expressed in a bacterial system, *E. coli* in the case of this thesis. M.*Taq*I was chosen due to its versatile nature, and produced using the protocol in **2.1.8**. Once purified, the sample was tested for activity using a protection assay; this determines whether M.*Taq*I could successfully methylate DNA, thus protecting it from restriction by corresponding restriction enzyme R.*Taq*I. M.*Taq*I methylates the underlined adenine within the DNA sequence 5' TCG<u>A</u> 3'. This sequence is present four times within the pUC19 plasmid (**Figure 3.2A**), resulting in a distinct pattern when analysed using gel electrophoresis if fully restricted or protected (**Figure 3.2B**).

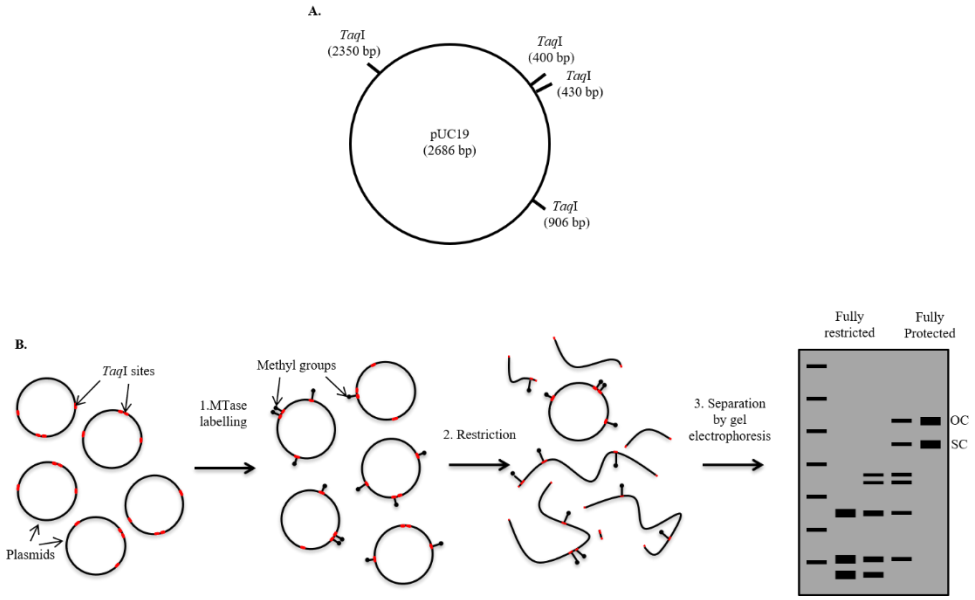


Figure 3.2: A) pUC19 plasmid showing the four *Taq*I (TCGA) sites present in its sequence B) Schematic of a protection assay workflow. pUC19 is labelled using M.*Taq*I at all *Taq*I sites before being restricted and analysed using gel electrophoresis to separate the DNA fragments by size. If the DNA has been successfully methylated by M.*Taq*I, restriction will be blocked and the DNA will remain intact.

It is important to note that when running plasmid DNA on a gel, the fully protected sample can often appear as multiple bands. This is due to the DNA existing in three conformations: supercoiled (SC) (where DNA is wound up tightly in a compact structure), open circular (OC) (where nicks have been introduced to one strand of the plasmid) and linear (nicked at both strands), shown in **Figure 3.3**. Supercoiled DNA runs the fastest through an agarose gel, whereas the larger open circle of OC runs the slowest, and will appear at the top of the gel.

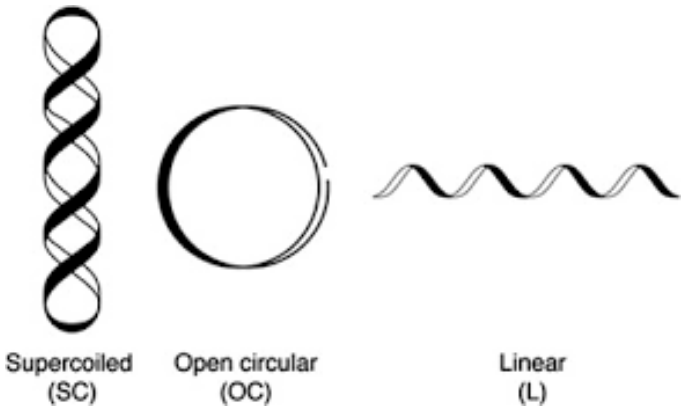


Figure 3.3: Schematic showing supercoiled, open circular (nicked on one strand) and linear DNA (nicked on both strands) conformations, all of which run at different speeds through an agarose gel.

Varying concentrations of M.*Taq*I were incubated with AdoMet (0.4 nM) and pUC19 DNA (~20 nM) as described in **2.1.14**, to check that the DNA could be successfully methylated, and how much of the MTase was needed for full protection. These samples were run on a 1 % agarose gel, with each lane representing a 2 x dilution of M.*Taq*I (from 625 nM in lane 2 down to ~ 80 nM in lane 5). As can be seen in **Figure 3.4**, M.*Taq*I has prevented restriction of pUC19 DNA in lanes 2-5, meaning the protein is functional even at the lowest concentration tested. As pUC19 has four *Taq*I sites, this means that there is a concentration of ~ 200 nM sites to methylate in each well. As lane 5 still shows full protection (with only 80 nM M.*Taq*I methylating 200 nM sites), this demonstrates that the protein has a turnover of at least two times.

In the absence of AdoMet (lanes 7-10), however, there is still partial protection seen, as the DNA is not fully restricted (the control in lane 11 shows full restriction). This suggests that residual AdoMet is being copurified with M.*Taq*I[121] and is preventing full restriction by methylating DNA without added AdoMet.

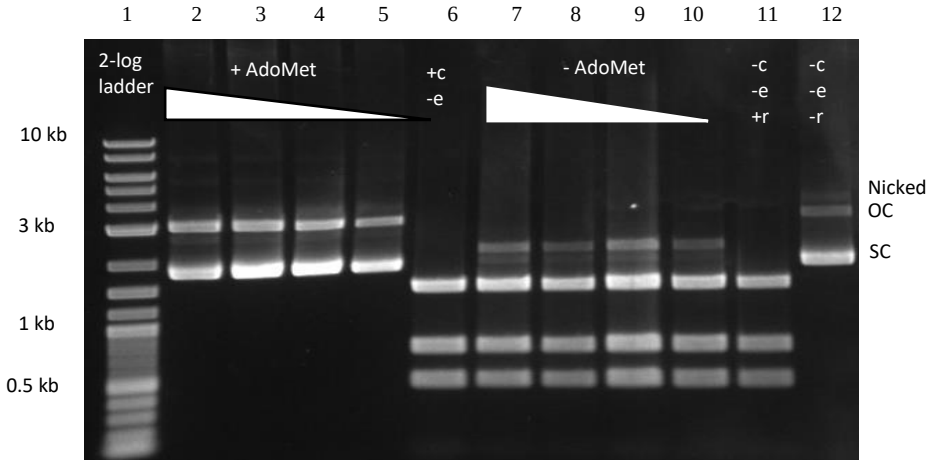


Figure 3.4: Protection of pUC19 DNA using M.*Taq*I and AdoMet
Lane 1 = 2-log ladder, lanes 2-5 = 2x dilution of M.*Taq*I with AdoMet, lane 6 = no M.*TaqI* control, lanes 7-10 = 2x dilution M.*Taq*I no AdoMet, lane 11 = restricted pUC19, lane 12 = unrestricted pUC19.

M.*Taq*I has prevented restriction of pUC19 DNA in lanes 2-5, meaning the protein is functional even at the lowest concentration tested. In the absence of AdoMet (lanes 7-10), partial protection can be seen, as the DNA is not fully restricted. This suggests that residual AdoMet has coeluted with the M.*Taq*I protein during purification.

Having native AdoMet bound to M.*Taq*I protein has been shown to prevent efficient labelling of DNA when using cofactor analogues[120]. As MTase-directed labelling is intended to be used extensively in this thesis, the effects that this residual AdoMet could have on experiments were considered. For DNA mapping, residual AdoMet could cause problems as the sites will not be labelled with fluorophores if there is already a methyl group present, resulting in inaccurate recognition of barcodes. This could lead to false negative/positive results, especially when looking at small mutations, as loss of a fluorophore may suggest a SNP in the DNA sequence, when in fact this was as caused by residual AdoMet methylating

the site and blocking labelling by the AdoMet analogue. Residual AdoMet could also cause problems with labelling of DNA for oligoprobes as again, if methylated, this would block labelling sites for the fluorophores, resulting in reduced sensitivity and brightness. If probes are not bright enough this may lead to inaccurate results as, even if the unlabelled probe binds, it will go undetected.

Attempts were therefore made to remove bound AdoMet from the protein complex to enable efficient labelling of DNA. As cofactor degradation increases at temperatures above 50 ºC[122], it was hypothesised that by heating protein samples to 72 ºC prior to purification, this may cause the AdoMet to dissociate from the M.*Taq*I complex. This was tested by heating half of a lysed cell preparation to 72 ºC for 2 minutes, while processing the other half as normal (described in **2.1.10**). Both samples were purified under the same conditions and their results compared via denaturing SDS-PAGE. Both conditions produced intact M.*Taq*I protein as observed by a strong band at 48 kDa on the SDS-Page gel in **Figure 3.5A and B**. The heat-treated (HT) M.*Taq*I sample, **Figure 3.5B**, also appears to be more pure, as there are a reduced number of non-specific bands in the gel. As the gels in **Figure 3.5A and B** display consistently strong bands at the 48 kDa mark, it suggests that the protein lost in the HT sample is non-specific proteins that have been copurified with the protein of interest (M.*Taq*I). It could be that heating the sample has caused these proteins to degrade, but as M.*Taq*I can withstand higher temperatures it remained intact. The heating process may have also caused the protein to refold differently, reducing the amount of aggregated *M.Taq*I protein, hence the presence of less protein beyond 48 kDa[123].

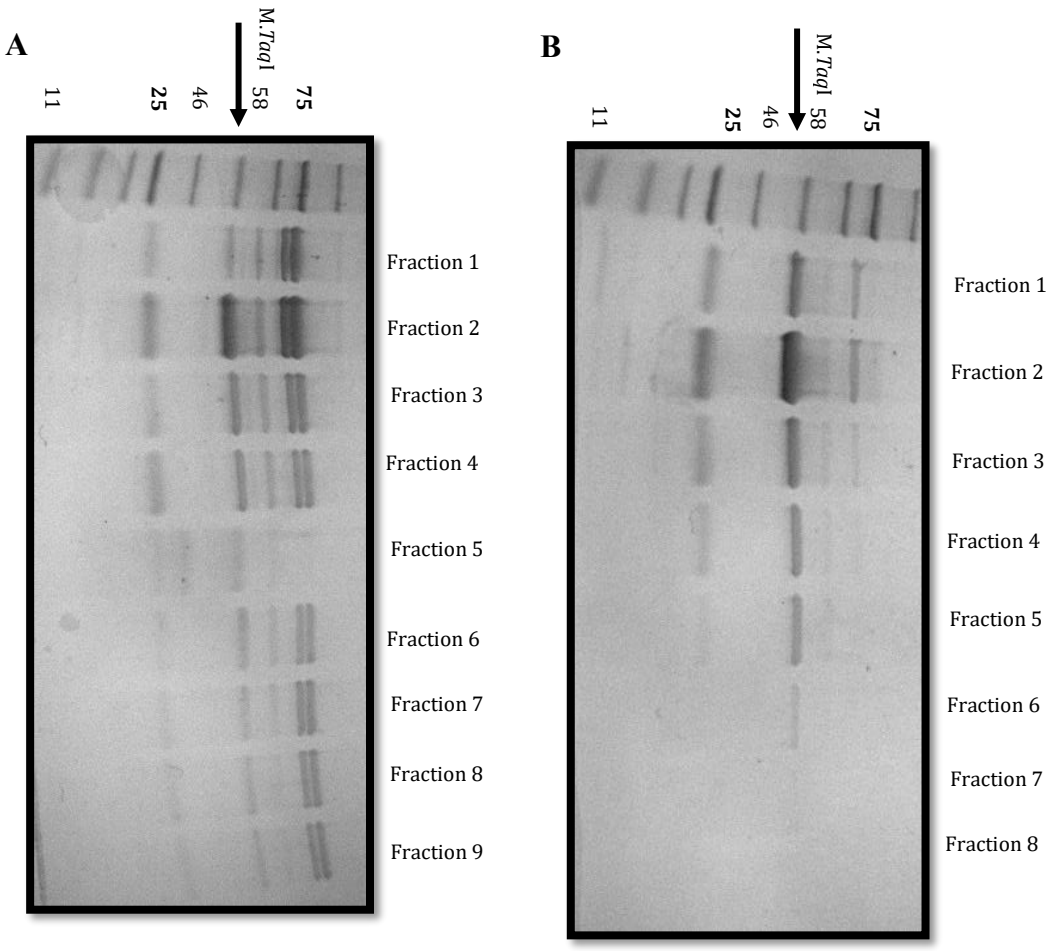


Figure 3.5: SDS-PAGE gel of M.*Taq*I protein (48 kDa) eluted fractions after purification A) without and B) with heat treatment to 72 ºC. Non-specific protein reduction can be seen in B, after heat treatment. M.*Taq*I protein has been produced in both conditions, as highlighted by the strong band at ~ 48 kDa.

Fractions 1 to 4 of the non-heat-treated sample were pooled, as they contained the highest yield of M.*Taq*I protein, as shown by the strong band at 48 kDa in **Figure 3.5**. Fractions 1 to 5 of the heat-treated sample were pooled. Both samples were used to methylate pUC19 DNA to check that they were functional, and to detect residual AdoMet. Protection of pUC19 appears to be equally effective in both as shown in **Figure 3.6**, which indicates that there is no decrease in M.*Taq*I activity with heat treatment. However, in lanes 14 and 15 where no AdoMet was added, there does not seem to be a reduced amount of residual protection. This means that bound AdoMet is still present within the protein complex, and that other attempts for removal could be explored. However, the amount of residual AdoMet, while present, is minimal, suggesting that using this concentration of *M.Taq*I (~300 nM) could still be suitable for further mapping applications.

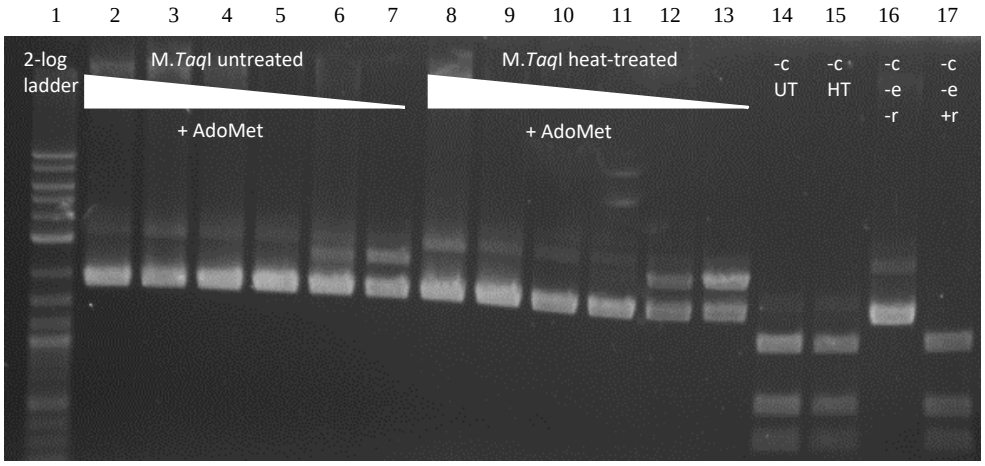


Figure 3.6: Protection assay comparing M.*Taq*I efficiency after heat-treatment Lane 1 = 2-log ladder, lanes 2-7 = 10x dilution untreated (UT) M.*Taq*I with AdoMet, lanes 8-13 = 10x dilution M.*Taq*I protein heat-treated (HT) at 72 ºC, lane 14 = UT M.*Taq*I no AdoMet, 15 = HT M.*Taq*I no AdoMet, 16 = unrestricted pUC19, 17 = restricted pUC19.

Both untreated and heat-treated M.*Taq*I samples were active with AdoMet, as highlighted by full protection in lanes 2-7 and 8-13 respectively. They also both showed residual protection in the absence of AdoMet (lanes 14 and 15), suggesting that heating the sample does not remove residual AdoMet from the M.*TaqI* complex.

In another attempt to remove bound AdoMet, drop dialysis was used. 20 µl of M.*Taq*I was placed on a drop dialysis filter disc (Merck, 0.025 µl pore size) over low salt buffer (10 mM phosphate buffer, 5 mM EDTA) overnight, in an attempt to remove excess salts and inhibitory substances (and bound AdoMet) from the protein. As can be seen in **Figure 3.7**, drop dialysis was ineffective in removing the residual AdoMet as protection can still be seen. This is likely because AdoMet is so tightly bound into the M.*Taq*I complex.

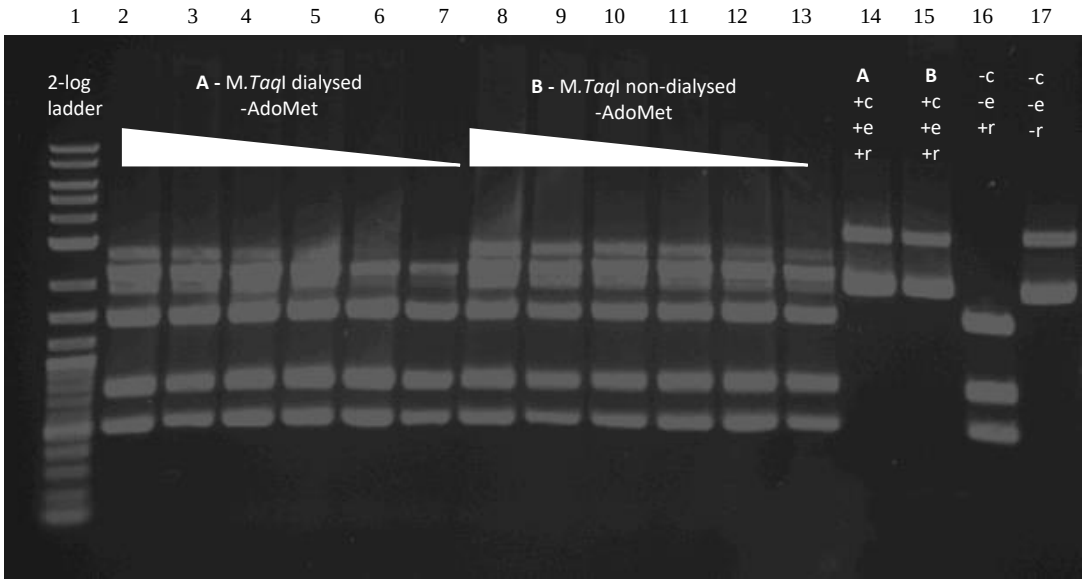


Figure 3.7: Protection assay comparing residual AdoMet protection of pUC19 DNA using A) dialysed and B) non-dialysed M.*Taq*I

Lane 1 = 2-log ladder, lanes 2-7 = 2x dilution of dialysed M.*Taq*I protein (A) no AdoMet, lanes 8-13 = 2x dilution of non-dialysed M.*Taq*I protein (B) no AdoMet, lane 14 = dialysed M.*Taq*I + AdoMet, lane 15 = non-dialysed M.*Taq*I + AdoMet, lane 16 = restricted pUC19, lane 17 = unrestricted pUC19.

Both dialysed and non-dialysed samples showed residual AdoMet protection (lanes 2-7 and 8-13 respectively), suggesting that dialysis does not remove AdoMet from the M.T*aq*I complex.

As M.*Taq*I recognises and methylates the palindromic DNA sequence 5' TCGA 3', incubating the protein sample with oligonucleotides (oligos) containing this sequence (5′CCGCC*TCGA*GGCGG3′) will result in removal of bound AdoMet[121]; bound AdoMet will be used up methylating sites within the oligos.

This was tested in two different experiments, firstly the oligos were added directly after cell lysis during protein production (**2.1.11**). Samples were incubated at 50 ºC for 30 minutes before continuing with the purification as normal. The eluted protein was used in a protection assay as before to directly compare the amount of partial protection of pUC19 DNA. As seen in **Figure 3.8**, both untreated M.*Taq*I (**Figure 3.8A**) and that which had been incubated with oligos (**Figure 3.8B**) are functional, but also appear to have equal amounts of residual bound AdoMet. A longer incubation could have been carried out but this could have caused the protein itself to degrade, resulting in low yield. There is also a large amount of competiting (methylated) genomic DNA within the sample that could be used instead of the bound AdoMet. It could also be that the buffer conditions at this point were not suitable for MTase-labelling, or that the M.*Taq*I protein was not active, and that purifiying the sample and then incubating with the oligos may prove more beneficial. In this way there would be more control over the conditions for the residual AdoMet to methylate the oligos which could be more successful in doing so.

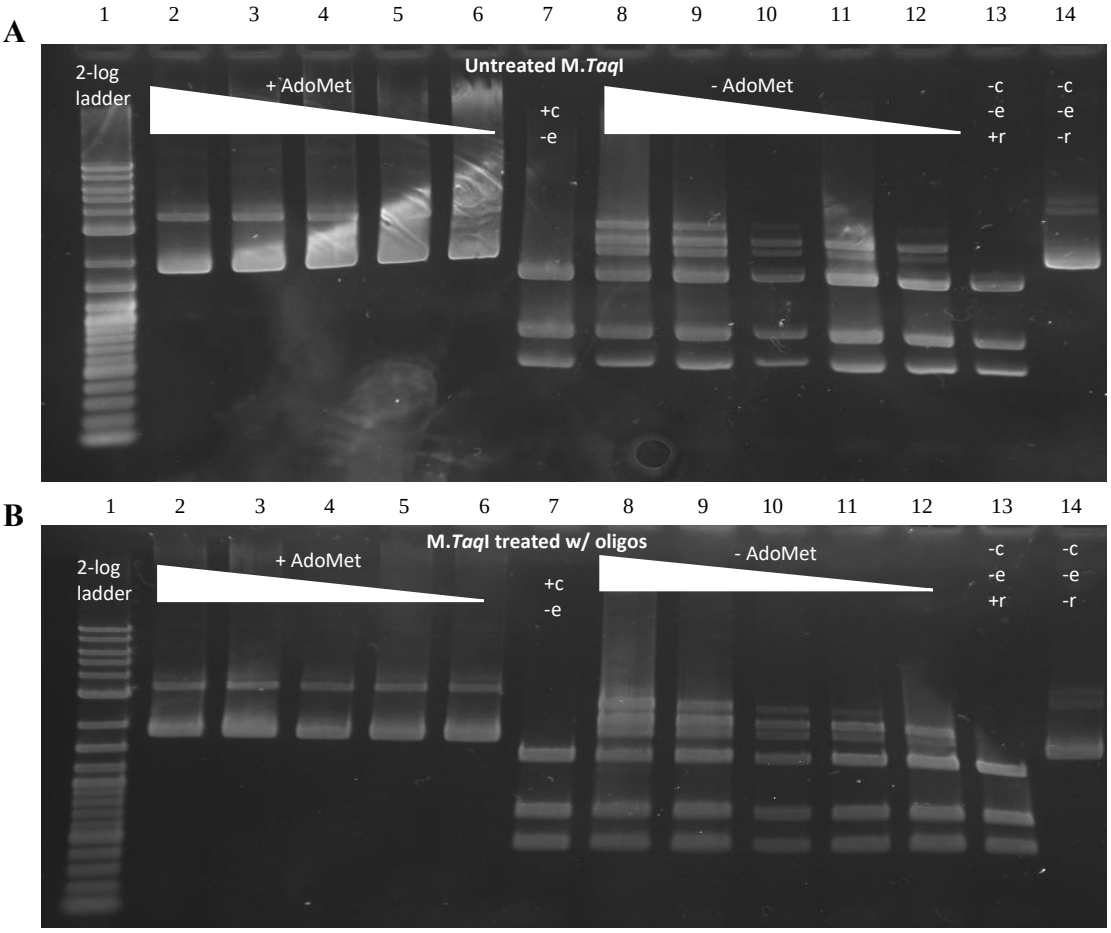


Figure 3.8: Protection assay comparing residual AdoMet protection using M.*Taq*I produced under A) normal conditions and B) lysate incubated with oligos containing M.*Taq*I recognition site for 30 minutes at 37 ºC.

Lane 1 = 2-log ladder, lanes 2-6 = M.*Taq*I with AdoMet, lane 7 = no enzyme control, lanes 8-12 = M.*Taq*I no AdoMet, lane 13 = restricted pUC19, lane 14 = unrestricted pUC19.

Both samples showed residual protection in lanes 8-12, as highlighted by the lack of full restriction. This suggests that incubating M.*Taq*I lysate with oligos containing the MTases recognition site prior to labelling does not remove residual AdoMet.

Attempts were then also made to remove bound AdoMet by adding the oligos after the protein has been purified and eluted. Oligos containing the recognition sequence (TCGA) were incubated with purified M.*Taq*I protein for 30 minutes at 50 °C. The samples were then directly added to a reaction mixture containing pUC19 to detect any residual protection of the DNA. At a glance, the results shown in **Figure 3.9** do not appear to show a reduction of

residual AdoMet. However, when looking at the control in lane 15, which has no enzyme and so should the DNA should not be protected, it shows that restriction has not gone to completion. This suggests that the presence of oligos prevented R.*Taq*I from fully restricting the DNA. The experiment was repeated with an added purification step using anion exchange columns to remove the oligos, to see if this could reduce the residual AdoMet.

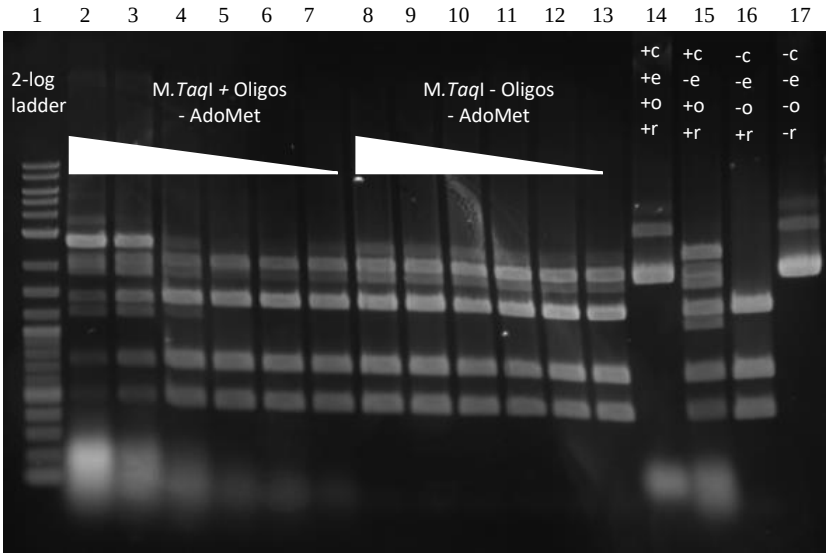


Figure 3.9: Protection assay of pUC19 after incubation of M.*Taq*I protein with and without oligos, to detect residual protection.

Lane 1 = 2-log ladder, lanes 2-7 = M.*Taq*I + 2x dilution of oligo incubated at 65 ºC for 15 minutes beforehand, lanes 8-13 = M.*Taq*I – oligos, lane 14 = unrestricted pUC19 + oligos, lane 15 = pUC19 + oligos no M.*Taq*I, lane 16 = restricted pUC19, lane 17 = unrestricted pUC19.

Lanes 1-4 suggest that the oligos are preventing full restriction of the DNA, and should be removed before testing this condition.

Oligos containing M.*Taq*I labelling sites were incubated with M.*Taq*I at 65 °C for 15 minutes, before being purified using anion exchange columns. Samples were either A) Not purified B) Run through the column once C) Run through the same column twice or D) Run

through a column twice, with fresh columns each time. The samples were then used for a protection assay with a two times dilution of M.*Taq*I and compared for residual protection as seen in **Figure 3.10**.

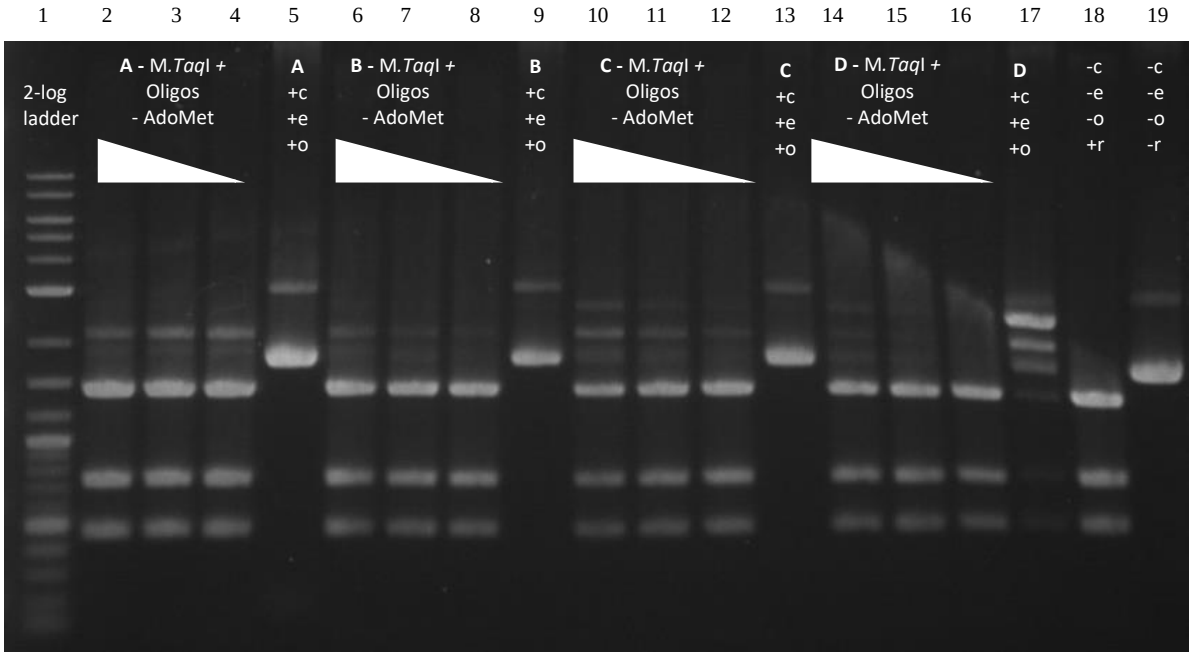


Figure 3.10: Protection assay showing residual protection after incubation of M.*Taq*I protein with oligos. Purification after incubation differed as follows A) No purification B) Flowed through column once C) Flowed through column twice D) Fresh column used

Lane 1 = 2-log ladder, lanes 2-4, 6-8, 10-12, 14-16 = M.*Taq*I + oligos (purified in conditions stated above) no AdoMet, lanes 5, 9, 13 and 17 = M.*TaqI* + oligos (purified in conditions stated above) + AdoMet, lane 18 = fully restricted pUC19, lane 19 = unrestricted pUC19.

Results show that incubating M.*Taq*I with oligos containing its recognition site prior to labelling removes some of the residual AdoMet, as highlighted by the increase in restriction in lanes 6-8, 10-12 and 14-16. This also suggests that oligos need to be removed sufficiently from the preparation to enable restriction to go to completion, as residual protection did not decrease in this instance where the DNA was not purified (A).

Each lane contained 10 µM of oligos, and started with ~ 117 nM of M.*Taq*I (before the 2x dilution). It does seem that the clean up using these columns had been successful; with no purification in lane 2 to 4, there is consistent amounts of unrestricted DNA as in the previous

experiment. However, all of the other samples that had been purified using the anion columns show some level of increased restriction, which is more apparent as the *M.Taq*I concentration decreases. This suggests that incubation with the oligos is capable of removing residual AdoMet but care must be taken to ensure that the appropriate concentration of both M.*Taq*I and oligos is used. If too much M.*Taq*I is used, this could introduce high enough levels of AdoMet to cause the associated problems with residual protection. It is therefore compulsory that the lowest concentration of M.*Taq*I is used – while still being enough to provide adequate alkylation with the cofactor – in order to not unnecessarily add residual AdoMet into the reaction. The third sample in D (lane 16) appears to show no residual protection, using ~39 nM of M.*TaqI*, and samples B and C show limited protection, suggesting that this is the optimal concentration to use if using this approach for AdoMet removal. Each lane with oligos contained 10 µM oligos which corresponds to 80 µM of M.*Taq*I labelling sites (each one has 8 TCGA regions). This highlights the careful consideration needed when labelling using M.*Taq*I; there must be a high enough concentration of the enzyme to ensure that labelling is efficient, but not so unnecessarily high that residual AdoMet is added to the mix. A balance needs to be struck to prevent bound AdoMet potentially blocking fluorophore sites with methyl groups, leading to inefficient labelling and potentially false positives/negatives if the technology is being used for clinical applications.

The results in this chapter suggest that the removal of bound AdoMet is not as straightforward as other research suggests. The presence of this residual AdoMet should be taken into consideration when planning labelling experiments, although it is not certain how much this affects the results. The majority of the DNA does get restricted, implying that there is only a small subset of M.*Taq*I which actually has the AdoMet bound; it is unclear how much this will affect labelling at this point. Further dialysis and extensive washing at various

points in the M.*Taq*I preparation could perhaps be considered in future to remove as much AdoMet as possible. For future experiments, it is worth noting that the minimum amount of M.*Taq*I needed for each reaction should be used. This is to ensure that excess AdoMet is not unnecessarily added to reactions and to reduce unwanted methylation during transalkylation. The proportion of M.*Taq*I carrying residual AdoMet appears to be small, as the majority of the DNA is unaffected at low M.*Taq*I concentrations, suggesting that this may not cause as many labelling problems as initially suspected. If problems are encountered, purification using the anion exchange columns appears to be successful in removing the oligos, and reduces residual protection.

3.3.2 Use of MTases in DNA alkylation

As discussed in **1.4.2**, AdoMet analogues can be used for various transalkylation reactions. These analogues contain functional groups such as an amine or azide which can be transferred to the MTase recognition site instead of a methyl group. These amine/azide cofactors enable the transfer of fluorophores to DNA site-specifically using NHS-ester or strain-promoted azide-alykne cycloaddition (SPAAC) chemistries respectively (**Figure 1.16**). Due to the NHS-ester dyes hydrolysing at low pH, dibenzocyclo-octyne (DBCO) dyes and SPAAC chemistry will be used for the majority of this thesis, as this allows the synthetic cofactors (which degrade at high pH) to be used more readily.

AdoMet analogue AdoHcy-6-N_3, shown in **Figure 3.11**, was produced by Andrew Wilkinson. This synthetic cofactor contains an extended, aliphatic linker terminating in an azide group in place of the methyl group. This means that when incubated with M.*Taq*I, the azide should be

transferred site-specifically to DNA. This was tested using a protection assay as described in **2.1.14**.

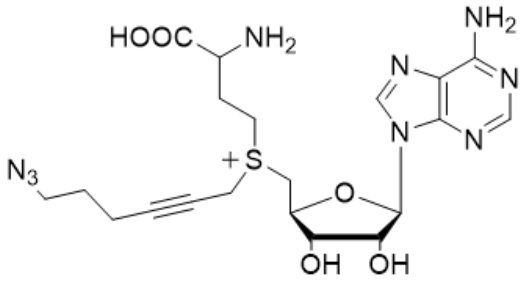


Figure 3.11: Structure of AdoHcy-6-N_3. The functional azide group is transferred by MTases to specific DNA sequences.

A concentration of 375 nM AdoHcy-6-N_3 was used in each lane, with varying concentrations of M.*Taq*I (from ~312 nM in lanes 2, 6 and 10 down to ~78 nM in lanes 4, 8 and 12) and 20 nM pUC19 (200 nM M.*Taq*I sites). As can be seen in **Figure 3.12**, the azide group successfully transferred to pUC19 DNA as indicated by the lack of restriction by R.*Taq*I suggesting that there is a turnover of at least two times. This shows promise for using this combination of enzyme and cofactor for future labelling experiments.

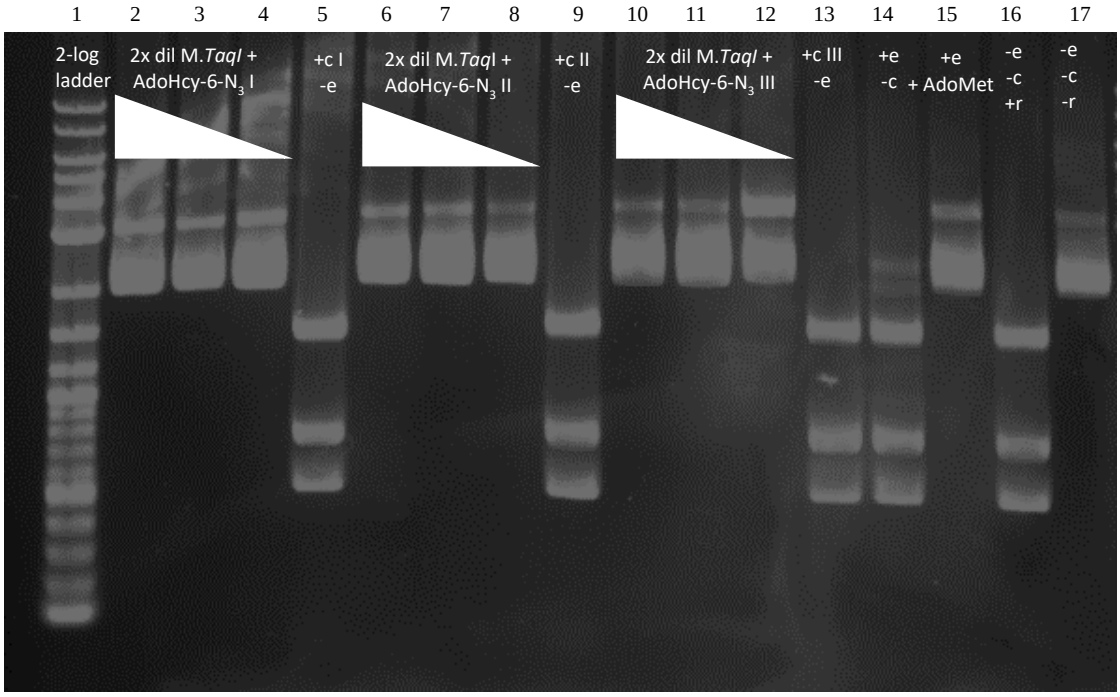


Figure 3.12: Protection assay showing protection of pUC19 DNA by different isomers of azide cofactor AdoHcy-6-N_3 and M.*TaqI*.

Lane 1 = 2-log ladder, lanes 2-4 = 2x dilution M.*Taq*I with AdoHcy-6-N_3 isomer I, lane 5 = control with AdoHcy-6-N_3 isomer I no enzyme, lanes 6–8 = 2x dilution M.*Taq*I with AdoHcy-6-N_3 isomer II, lane 9 = control with AdoHcy-6-N_3 isomer II no enzyme, lanes 10-12 = 2x dilution M.*Taq*I with AdoHcy-6-N_3 isomer III, lane 13 = control with AdoHcy-6-N_3 isomer III no enzyme, lane 14 = control with M.*Taq*I no cofactor, lane 15 = control with AdoMet and M.*Taq*I, lane 16 = restricted pUC19, lane 16 = unrestricted pUC19.

M.*Taq*I protein is active with all isomers of AdoHcy-6-N_3, as highlighted by the lack of restriction of the DNA in each case.

To test the efficiency of MTase labelling, a short hairpin sequence of DNA was ordered that contained one M.*Taq*I recognition site (and therefore two labelling sites), **Figure 3.13**. A hairpin was used to provide a single piece of DNA which contained a double-stranded region with palindromic 5' TCGA 3' site. Previous research has shown that in order for M.*Taq*I to dock and label DNA it must be double stranded[77], and a hairpin design allows for this.

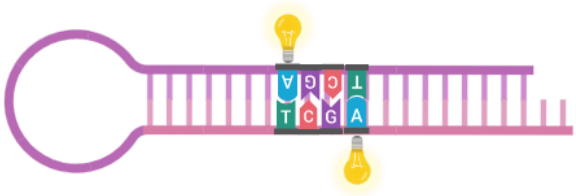


Figure 3.13: Schematic of hairpin oligoprobe design containing one TaqI recognition site within the sequence for labelling with M.*Taq*I.

The hairpin was alkylated with an azide functional group as described in Chapter 2 using AdoHcy-6-N_3 (**Figure 3.11**) and M.*Taq*I. Mass spectrometry was then used to detect the presence of the azide linker (mW 121) after alkylating the oligo (mW 18,378). In this way, it can be observed how efficient labelling is using this approach – as well as if the hairpin structure has formed correctly to enable alkylation – and the proportion of unlabelled DNA within a sample estimated. This is important as it gives some indication as to how residual AdoMet may affect future labelling reactions and therefore how reliable the results are. It may also show that the labelling protocol needs further optimisation to improve the labelling efficiency.

The sample was purified by HPLC by Andrew Wilkinson. **Figure 3.14A** shows the HPLC trace of both unlabelled (black) and alkylated hairpin (blue). Three main peaks of varying sizes can be seen within the alkylated hairpin sample, with the smallest of those corresponding to the large peak of the unlabelled (black) sample. This suggests that the smallest peak is indicative of the labelling reaction not going to completion, as some was left unlabelled. After collecting the peaks and submitting to mass spectrometry, it confirmed that the peaks were unlabelled DNA (**Figure 3.14Bi**) , hemi-alkylated DNA (**Figure 3.14Bii**), and fully-alkylated DNA (**Figure 3.14Biii**). This means the second peak shows that some of the DNA had only the azide linker on one side of the hairpin (the oligo has a mW of 18,378 and

the aizde linker has a mW of 121) potentially due to slow turnover of enzyme leading to partial alkylation. The majority of the sample did end up with labels on both sides, as the spectra shows a peak at 18,620 (18,378 + 121*2). Only a small amount of the sample remained unlabelled (peak at ~18,379) after the reaction. Minimal amounts of M.*Taq* were used in this sample (78 nM) – in order to reduce the amount of residual AdoMet added into the reaction – compared to 2 µM of labelling sites. 100 % labelling of 2,000 nM sites after 1 hour with 78 nM M.*Taq*I would be ~25 turnovers, so it is understandable that complete labelling would not be reached with this concentration. The spectra indicate that approximately 75 % labelling has been achieved (around 1500 nM of labelled sites), which corresponds to ~ 19 turnovers. This supports similar results published by Weinhold *et al.* in 2005[89]. M.*Taq*I is a thermophile, with an optimum working temperature of ~65 ºC[113], however, temperatures of this heat would cause the cofactor to rapidly degrade. A compromise in temperature was used (50 ºC) meaning that a slight sacrifice in enzyme performance was made, this could also affect labelling going to completion, but is not critical.

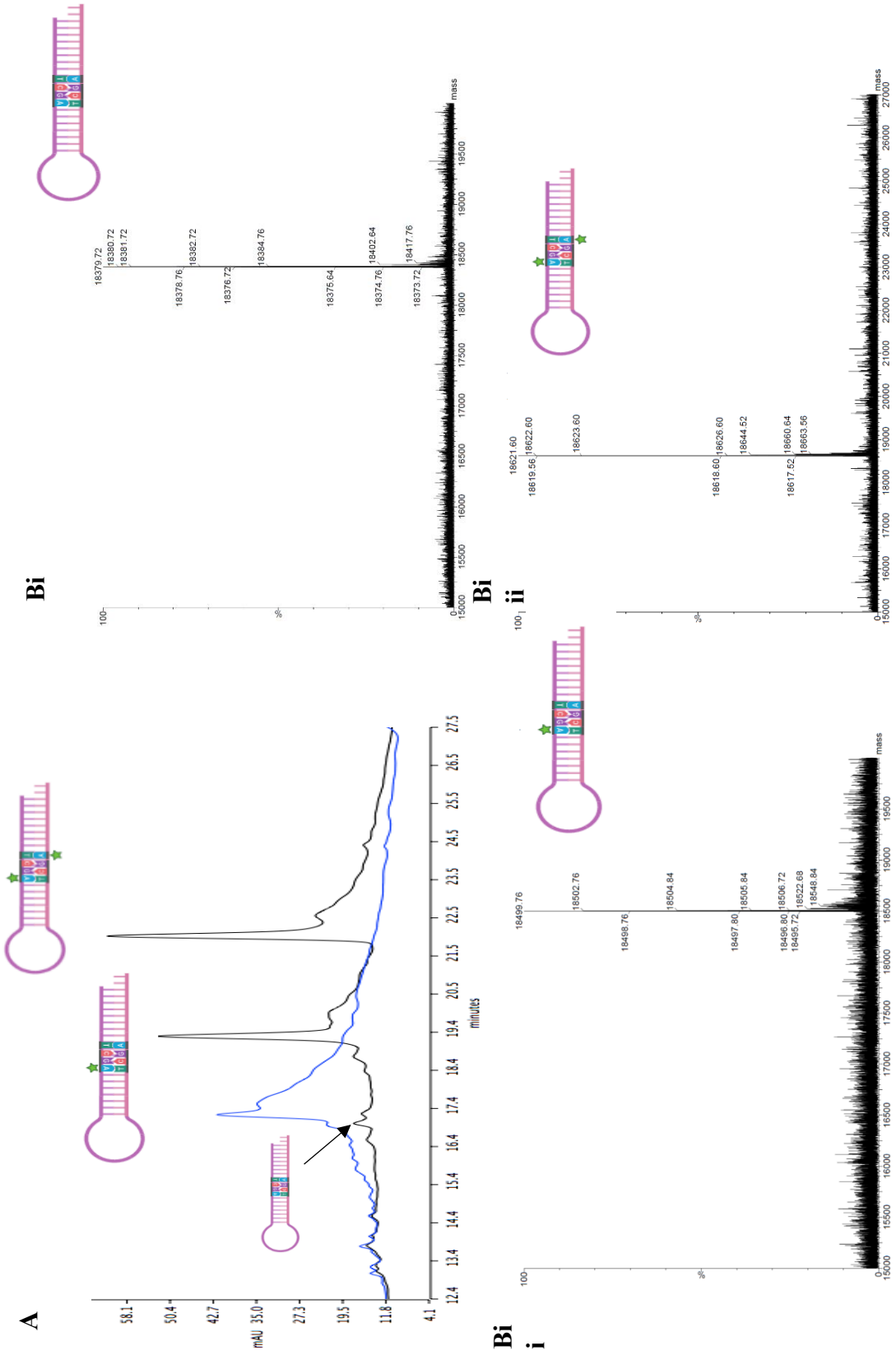


Figure 3.14: Hairpin oligos were alkylated with azide cofactor AdoHcy-6-N_3 and M.*Taq*I A) HPLC trace of unlabelled (blue) and labelled (black) oligo B) Spectra of the three labelled peaks showing i) unlabelled ii) hemi-alkylated and iii) fully-alkylated oligos. This shows that ~ 75 % labelling was achieved with this concetration of M.*Taq*I, undergoing ~19 turnovers.

3.3.3 Potential for use of MTases in small-scale mutation detection

The MTase M.*Hpy*188i has a recognition site that is interrupted by the absence of SMN1 (**Figure 3.15A**), the loss of which is associated with SMA. As discussed in the introduction of this chapter, carriers of a 2:0 mutation of SMA are impossible to detect using current methods as SMN1 cannot be distinguished between highly homologous SMN2, which differs at one critical nucleotide. Mapping this region could allow detection of this nucleotide difference (determined by a loss of fluorophore in SMN2) while maintaining sequence context (**Figure 3.15B**). This means that DNA mapping could highlight whether someone has the SMN1 gene in a 1:1 ratio, or is a 2:0 silent carrier.

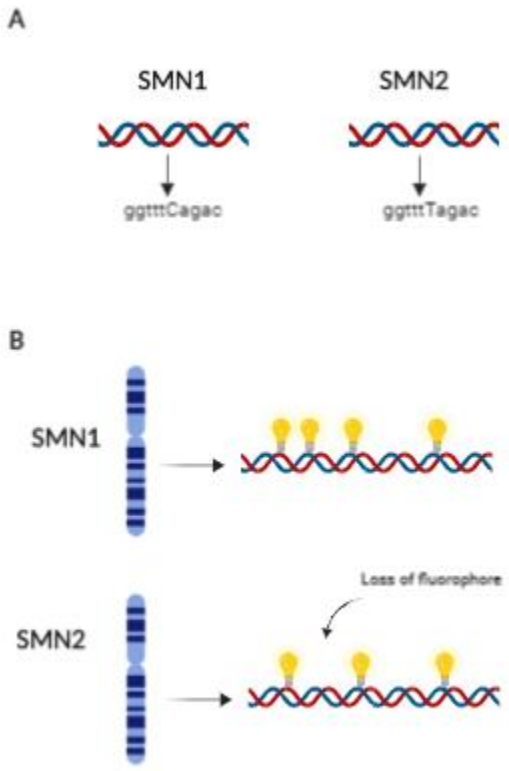


Figure 3.15: Schematic displaying A) The critical nucleotide difference between SMN1 and SMN2 that disrupts M.*Hpy*188i labelling sites B) A representation of loss of fluorophore in DNA mapping of SMN1/SMN2 with M.*Hpy*188i. This could be used to detect 2:0 carriers of SMA.

We sought to simulate this point mutation using a model system; the use of MTase M.*Hinc*II was considered as its recognition site, 5' GTYRAC 3', overlaps with that of M.*Taq*I. This will prevent a subset of M.*Taq*I sites from being modified (and labelled) if first methylated with M.*Hinc*II. By using DNA mapping, the lack of fluorophore could potentially be observed, therefore allowing detection of small deletions down to base pair level. M.*Hinc*II was expressed, as described in **2.1.8,** from clones produced by Dr Robert Neely. The enzyme was tested for activity with AdoMet using a protection assay. As there is only one *Hinc*II site on pUC19, cutting with just R.*Hinc*II would result in linear DNA and be difficult to distinguish from uncut plasmid. Two restriction enzymes that cut at different sites were therefore used to ensure that DNA fragments of dissimilar sizes were produced, so that there was a noticeable difference (i.e. multiple bands) when analysed using gel electrophoresis.

As can be seen in lanes 2-4 of **Figure 3.16**, the enzyme is achieving partial protection of the DNA, as it is not being fully restricted. It could be that the enzyme needs to be more concentrated or that turnover is too slow to achieve full protection. Full protection will need to be achieved if M.*Hinc*II is to be used for labelling reactions.

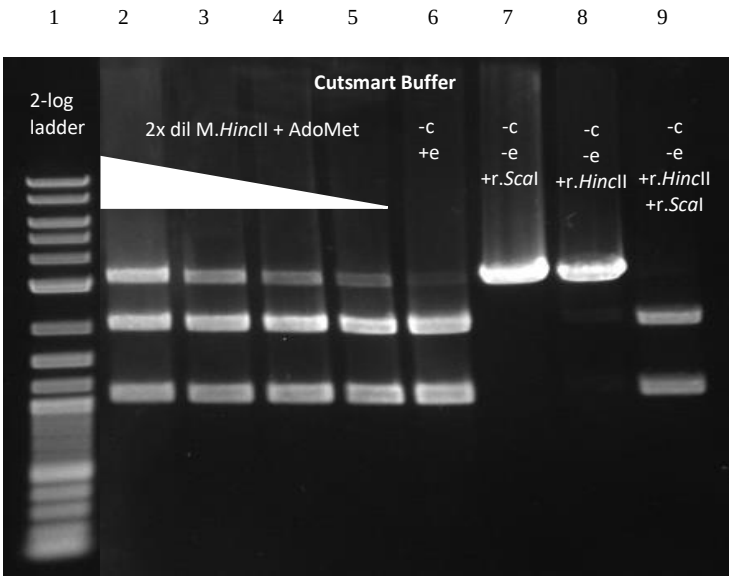


Figure 3.16: Protection assay showing protection of pUC19 DNA by AdoMet and M.*Hinc*II.

Lane 1 = 2-log ladder, lanes 2-5 = 2x dilution M.*Hinc*II with AdoMet, lane 6 = control with enzyme, no AdoMet, lane 7 = restricted with r.*Sca*I, lane 8 = restricted with r.*Hinc*II, lane 9 = fully restricted.

M.*Hinc*II is partially active with AdoMet as the DNA is not fully restricted (as in control lane 9) in lanes 2-5. It is not fully protecting the DNA, however, and will need to be optimised before further use.

Previously, the commercially available NEB Cutsmart Buffer (50 mM potassium acetate, 20 mM tris-acetate, 10 mM magnesium acetate, 100 μg/ml BSA) had been used for protection assays. Buffers are usually composed of a weak acid, salts, and potentially additives. These components provide a stable environment for the reaction by keeping the pH constant. This occurs by taking up protons released during the reaction or releasing them as they are consumed. The slightest change in pH or salt concentration can impair many biochemical processes including enzymatic reactions. As the buffer can play a role in the efficiency of the enzymes, alternates were considered for use with M.*Hinc*II.

Two alternative buffers were tested based on their successful use in similar enzymatic reactions: buffer A[76] (50 mM tris-hydrochloride, 15 mM sodium chloride 0.5 mM EDTA, 2

mM β-mercaptoethanol, 0.2 mg/ml BSA), and buffer B^{124} (50 mM tris-hydrochloride, 5 mM β-mercaptoethanol, 10 mM EDTA). EDTA is commonly added into buffers to chelate multivalent cations and stop DNA degradation. The restriction assay protocol for buffers containing EDTA therefore had an additional step before restriction whereby cations were added back into the mixture. β-mercaptoethanol reduces disulphide bonds between cysteine residues, preventing protein aggregation to hopefully achieve higher MTase activity. Additives such as BSA can also be added in an attempt to stabilise the enzyme.

Both buffer A and B appear to make no significant difference to the activity of M.*Hinc*II, **Figure 3.18** and **Figure 3.17** respectively, as both just show partial protection in lanes 2-5. It does appear that there is slightly more protected DNA with buffer B at lower enzyme concentrations – for instance when comparing lane 5 of both gels – suggesting that M.*Hinc*II is more active with buffer B. This suggests that the kinetics of M.*Hinc*II are too slow to reach complete protection; a higher concentration of M.*Hinc*II will improve this, as can be seen from the steady decrease in protection in gels. However, in order to achieve this, the construct and expression conditions of this enzyme should be re-evaluated.

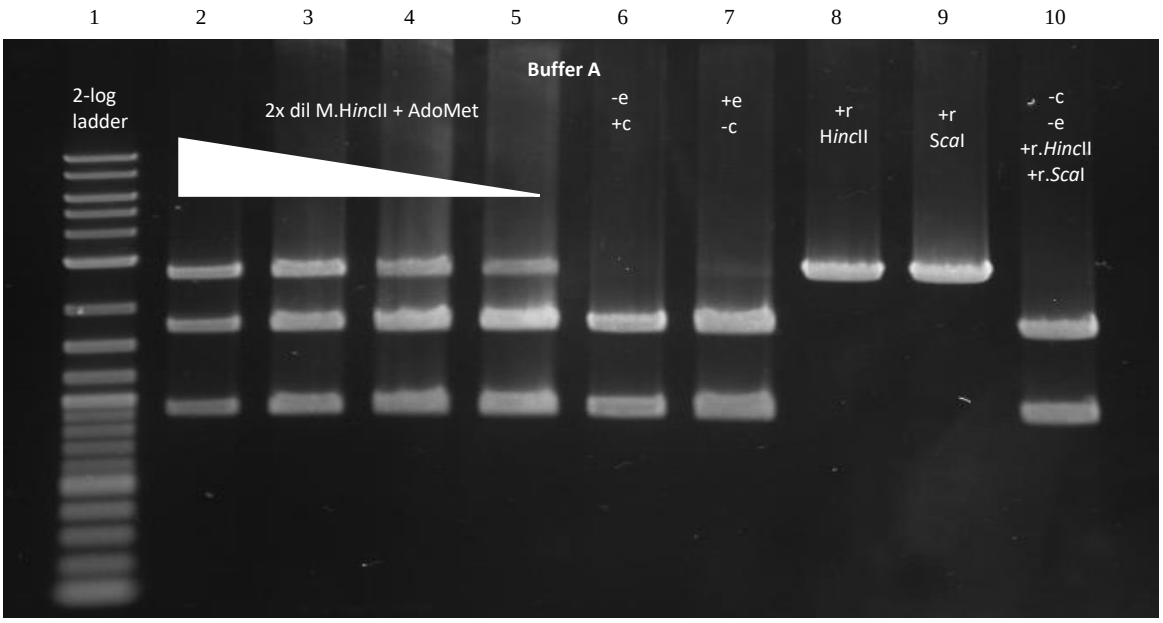


Figure 3.18: Protection assay showing protection of pUC19 DNA by AdoMet and M.*Hinc*II using "Buffer A"

Lane 1 = 2-log ladder, lanes 2-5 = 2x dilution M.*Hinc*II with AdoMet, lane 6 = control with AdoMet, no enzyme, lane 7 = control with enzyme, no AdoMet, lane 8 = restricted with R.*Hinc*II, lane 9 = restricted with R.*Sca*I, lane 10 = fully restricted.

M.*Hinc*II is partially active with AdoMet as the DNA is not fully restricted (as in control lane 10) in lanes 2-5. It is not fully protecting the DNA, however, and will need to be optimised before further use.

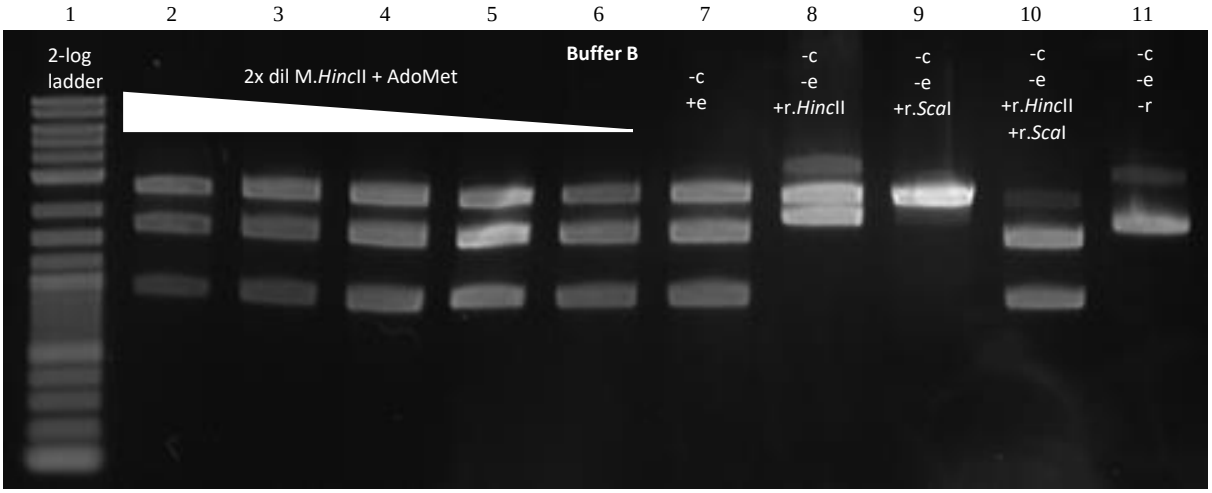


Figure 3.17: Protection assay showing protection of pUC19 DNA by AdoMet and M.*Hinc*II using "Buffer B"

Lane 1 = 2-log ladder, lanes 2-6 = 2x dilution M.*Hinc*II with AdoMet, lane 7 = control with enzyme, no AdoMet, lane 8 = restricted with R.*Hinc*II, lane 9 = restricted with R.*Sca*I, lane 10 = fully restricted, lane 11 = fully protected.

M.*Hinc*II is partially active with AdoMet as the DNA is not fully restricted (as in control lane 10) in lanes 2-5. It is not fully protecting the DNA, however, and will need to be optimised before further use.

Fortunately, an alternative enzyme to M.*Hinc*II is available; M.*Bse*CI. M.*Bse*CI was considered as its recognition site (ATCGAT) also overlaps and blocks a subset of M.*Taq*I sites. The literature has also shown that this enzyme works well with synthetic cofactors in alkylation reactions[125]. M.*Bse*CI expression was attempted using clones from Dr Robert Neely, however purification of active protein was unsuccessful. Upon troubleshooting it was noted that the His-tag of the M.*Bse*CI construct was on N-terminus, as opposed to the C-terminus as it is in the literature[126]. This may cause problems with the protein correctly folding and therefore affect its stability and functionality.

Active M.*Bse*CI protein was kindly provided by the Weinhold lab[127] and tested in a protection assay to determine its activity. A comparison was made between three different buffers, *Bse*CI buffer (100 mM Tris-HCl, 500 mM NaCl, 200 mM MgCl, 500 µM EDTA, 20 mM β-mercaptoethanol)[128], NEB Cutsmart and NEB2 (buffer A, B, and C respectively), and enzyme activity, as described in **2.1.14**. As can be seen from **Figure 3.19,** the enzyme was active with all three buffers, with slightly decreased activity in *Bse*CI buffer A, suggesting that these are not the optimum conditions for methylation with this enzyme. NEB2 buffer was selected as the most efficient buffer for methylation. After successfully methylating lambda DNA, M.*Bse*CI was later used in Chapter 5 for further labelling experiments.

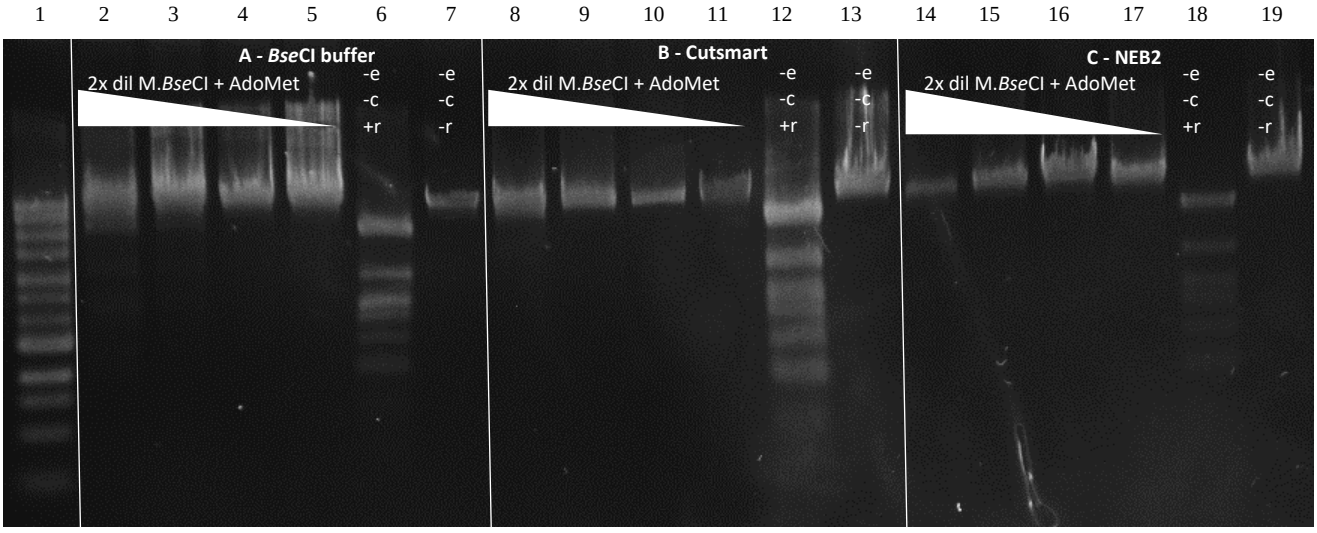


Figure 3.19: Protection assay showing protection of lambda DNA by AdoMet and M.*Bse*CI using A) BseCI buffer, B) Cutsmart buffer, and C) NEB2 buffer.

Lane 1 = 2-log ladder, lanes 2-5 = 2x dilution M.*Bse*CI with AdoMet in Buffer A, lane 6 = fully restricted, lane 7 = fully protected, lane 8-11 = 2x dilution M.*Bse*CI with AdoMet in Buffer B, lane 12 = full restricted, lane 13 = fully protected, lane 14-17 = 2x dilution M.*Bse*CI with AdoMet in Buffer C, lane 18 = fully restricted, lane 19 = fully protected.

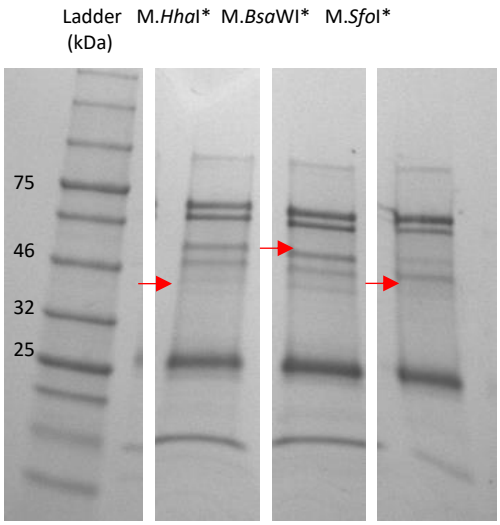
M.*Bse*CI is active with AdoMet in all buffers as the DNA is protected in all test lanes.

3.3.4 Producing mutated enzymes for human genome mapping

As discussed in **1.1.4**, enzymes containing mutations at certain points in their DNA sequence affect which amino acid is produced. This change in protein structure results in the opening-up of the cofactor binding pocket. By changing the structure of this pocket, bulkier cofactor analogues can interact with the protein. It has therefore been suggested that these mutated MTase enzymes have a greater affinity to AdoMet analogues and are more efficient when performing transalkylation reactions. The promising results from Lukinavičius *et al.*[76], discussed in **1.4.2**, compared how enzyme M.*Hha*I Q82A N304A worked with a range of AdoMet analogues. They then extended this to other m5c MTases, which suggests that we could replicate this. By testing different mutated enzymes with AdoMet analogues, it was

hypothesised that a toolbox of MTases could be developed for various mapping experiments. M.*Hha*I was chosen after the success shown in the literature, and M.*Bsa*WI and M.*Sfo*I were selected due to them also being m5C MTases (i.e. those that produce C5-methylcytosine). This means that all three enzymes share a similar amino acid sequence, and that the conserved motifs within them can be used to help select amino acids – corresponding to those identified as "activating" in the work by Lukinavičius *et al.*[76] – for mutation.

DNA sequences for M.*BsaW*I, M.*Sfo*I, and M.*Hha*I were aligned using Jalview to locate these mutations. The mutated sequences, M.*BsaW*I E83A D384A, *M.Sfo*I T77A D360A, and M.*Hha*I Q82A N304A (which will be referred to as M.*Bsa*WI*, M.*Sfo*I*, and M.*Hha*I* henceforth) were ordered from IDT DNA. The genes for each of the MTase enzymes were sub-cloned into the vector pRSET-B using Gibson assembly (described in **2.1.2**). Successful cloning was confirmed via sequencing. The enzymes were expressed in *E. coli* using the protocol in Chapter 2 and analysed using SDS-PAGE (**2.1.12**) to verify the protein expression levels. As can be seen from **Figure 3.20**, expression-levels of all three proteins were low. The gel shows no obvious bands where the proteins should be, indicated by red arrows. This could be due to the various expression conditions not being optimised beforehand, and a general method being used.



Protein	Size (kDa)
M.*Hha*I*	~37
M.*Bsa*WI*	~47
M.*Sfo*I*	~42

Figure 3.20: SDS-PAGE gel showing low expression levels of mutated M.*Hha*I*, M.*Bsa*WI* and M.*Sfo*I*. Red arrows indicate expected bands from the proteins' size in (kDa) shown in table (right).

Focus was shifted to M.*BsaW*I* and M.*Sfo*I*, as they had not been tested before and could prove useful as enzymes for dual-colour DNA mapping. Protein expression conditions were altered for M.*BsaW*I* and M.*Sfo*I* in an attempt to increase yield. The proteins were grown up at a larger volume and lysed more carefully using an Emulsiflex (a homogeniser as opposed to a sonicator). The Emulsiflex would hopefully offer less shearing whilst keeping the protein at a constant temperature, resulting in higher yield. Protease inhibitors were also used in all buffers to prevent protein degradation. The yield for M.*Sfo*I* significantly improved, with a strong band appearing at the 42kDa mark, **Figure 3.21**. Conditions for M.*Bsa*WI* still needed further optimisation for improved yield.

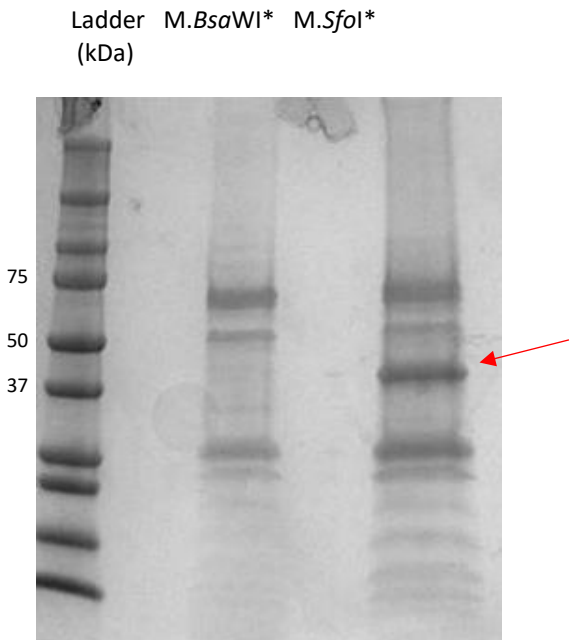


Figure 3.21: SDS-PAGE gel showing M.*Bsa*WI* and M.*Sfo*I* protein after optimised expression. Presence of M.*Sfo*I* protein is indicated by red arrow.

As M.*Bsa*WI is derived from a thermophile, *Bacillus stearothermophilus*[113], expression was attempted again without lowering the overnight incubation temperature after induction. This meant that the culture was left to grow for 16 hours at 37 ºC as opposed to 20 ºC. This change in temperature had a significant effect on yield, as can be seen by a much stronger band at 47kDa in **Figure 3.22B**, compared to **Figure 3.22A.** A Western blot was also used to confirm that the protein was present throughout the purification, as shown by a strong band in **Figure 3.22C.**

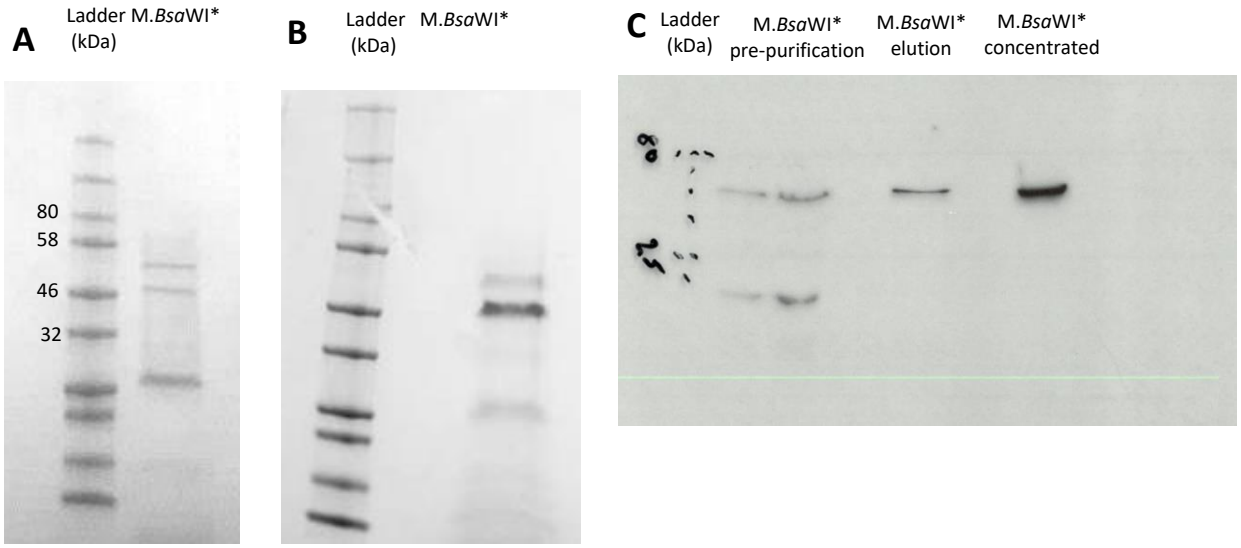


Figure 3.22: SDS-PAGE gel showing M.*Bsa*WI* grown at A) 20 ºC and B) 37 ºC after induction. Increase in yield is visible at the higher temperature. C) Western Blot confirming M.*Bsa*WI* protein with His-tag using anti-His antibody pre- and post-purification.

These proteins were tested for their ability to alkylate DNA using both AdoMet and AdoHcy-6-N_3 via a protection assay. A low salt buffer was initially used (50 mM tris-hydrochloride, 15 mM sodium chloride 0.5 mM EDTA, 2 mM β-mercaptoethanol, 0.2 mg/ml BSA). M.*Sfo*I* appeared to show slight protection with AdoMet as seen in lanes 2-4, **Figure 3.23A**. As the enzyme was intended to be used for dual-colour mapping, it would ideally have increased activity with AdoHcy-6-N_3. This was not that case, and no alkylation was seen, as indicated by full-restriction of DNA in lanes 2-4, **Figure 3.23B**.

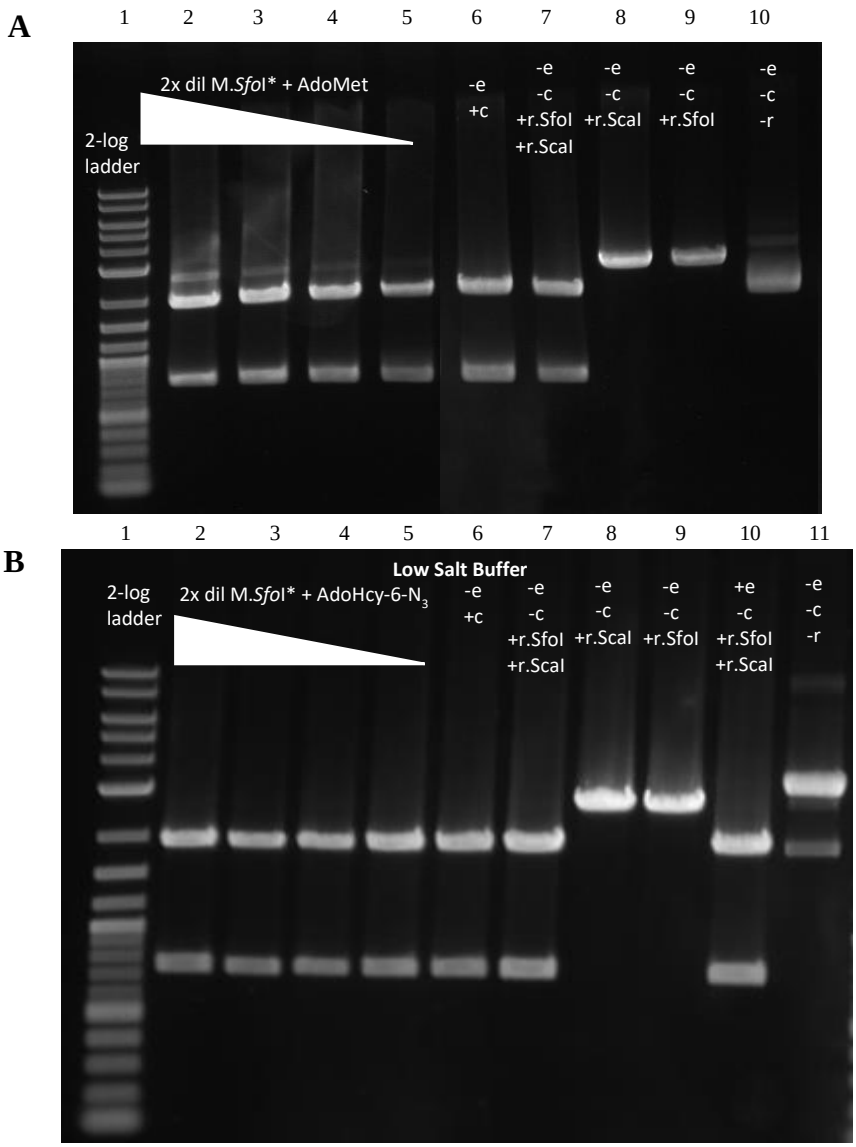


Figure 3.23: Protection assay showing protection of pUC19 DNA with M.*Sfo*I* in low salt buffer using A) AdoMet

Lane 1 = 2-log ladder, lanes 2-5 = 2x dilution M.*Sfo*I* with cofactor, lane 6 = no enzyme + cofactor, lane 7 = fully restricted, lane 8 = restricted with R.*Sca*I, lane 9 = restricted with R.*Sfo*I, lane 10 = fully protected.

B) AdoHcy-6-N_3

Lane 1 = 2-log ladder, lanes 2-5 = 2x dilution M.*Sfo*I* with cofactor, lane 6 = no enzyme + cofactor, lane 7 = fully restricted, lane 8 = restricted with R.*Sca*I, lane 9 = restricted with R.*Sfo*I, lane 10 = enzyme no cofactor plus restriction enzymes, lane 11= fully protected.

M.*Sfo*I was slightly active with AdoMet, as highlighted by the slight protection in A lanes 2-5, but will need to be optimised if intended for further use. M.*Sfo*I was not active with AdoHcy-6-N_3 as highlighted by complete restriction in the test lanes of B.

M.*Bsa*WI* was tested with AdoHcy-6-N_3 in low salt buffer and appeared to show promise. As **Figure 3.24** shows, there is slight protection of DNA with the cofactor analogue.

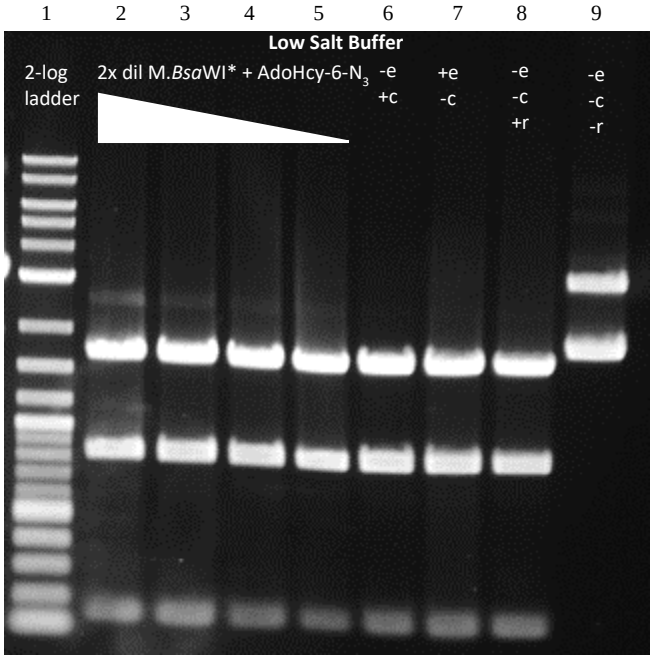


Figure 3.24: Protection assay showing protection of pUC19 DNA with M.*Bsa*WI* in low salt buffer using AdoMet and B) AdoHcy-6-N_3
Lane 1 = 2-log ladder, lanes 2-5 = 2x dilution M.*Bsa*WI* with cofactor, lane 6 = no enzyme + cofactor, lane 7 = no cofactor + enzyme, lane 8 = fully restricted, lane 9 = fully protected.

There is slight protection of DNA with M.*Bsa*WI and AdoHcy-6-N3 as highlighted by the lack of full restriction in lanes 2-4. Increasing the concentration of the enzyme may improve protection.

As there are three *Bsa*WI sites on pUC19 DNA, this means there is a site concentration of around 120 nM. Tube 1 contains approximately 875 nM of M.*Bsa*WI, which should be plenty to alkylate the DNA. It may be the case that turnover is slow for this enzyme and so full alkylation cannot be achieved under these conditions. By altering the buffer conditions, it was thought that the rate of alkylation may be able to be improved. As M.*BsaW*I* had shown some signs of activity, it was then tested with both NEB2 and Cutsmart buffer and both AdoMet and AdoHcy-6-N_3, **Figure 3.25**.

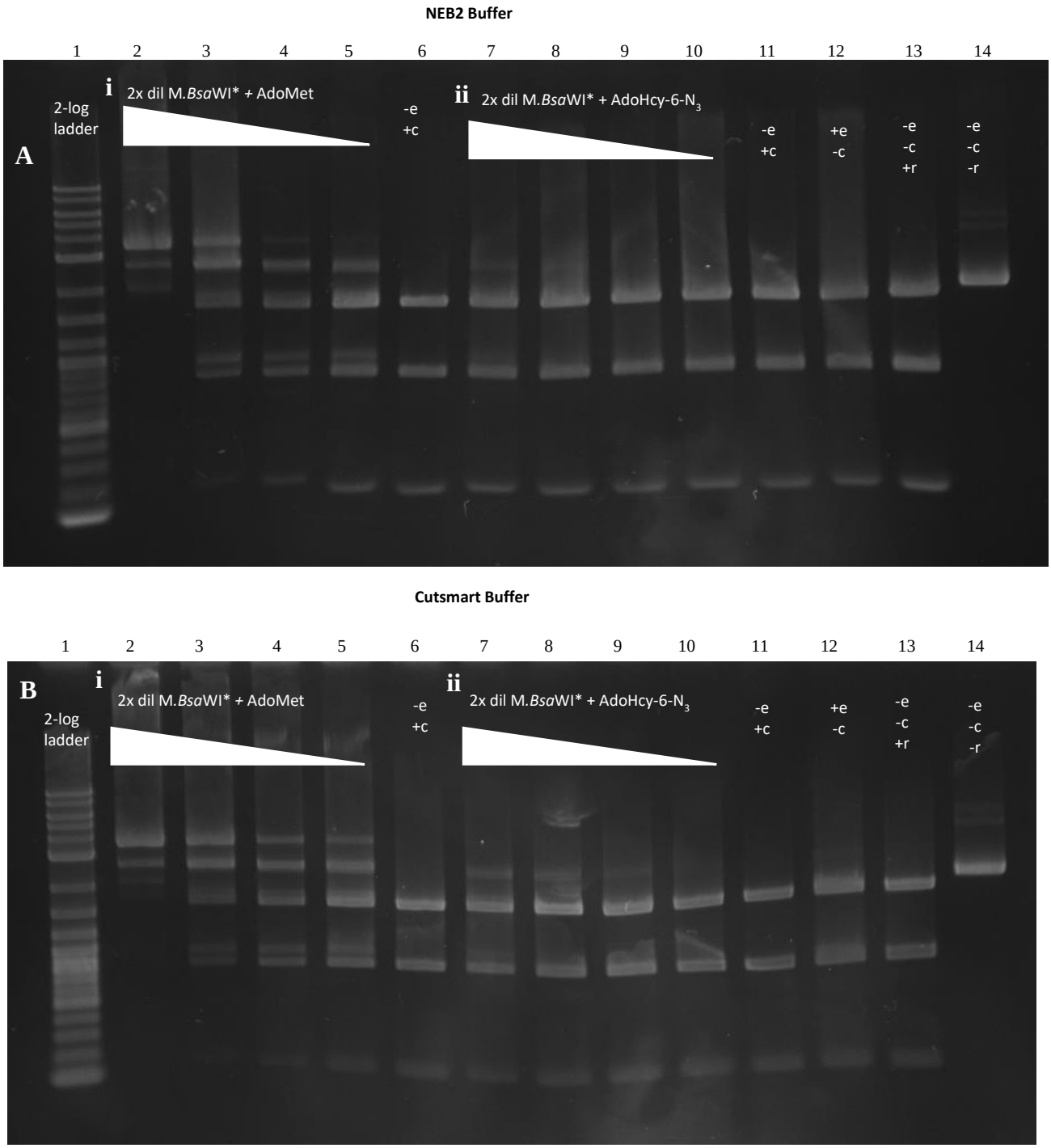


Figure 3.25: Protection assay showing protection of pUC19 DNA by i) AdoMet ii) AdoHcy-6-N_3 and a 2x dilution of M.*Bsa*WI in A) *Bsa*WI buffer, B) Cutsmart buffer and C) NEB2 buffer:

Lane 1 = 2-log ladder, lanes 2-5 = 2x dilution M.*BsaW*I with AdoMet, lane 6 = no enzyme + AdoMet, lane 7-10 = 2x dilution M.*BsaW*I with AdoHcy-6-NH_3, lane 11 = no enzyme + AdoHcy-6-N_3, lane 12 = no cofactor + enzyme, lane 13 = fully restricted, lane 14 = fully protected.

A higher concentration of M.*Bsa*WI has shown that it is more active with AdoMet, but not with AdoHcy-6-N_3. Full protection is still not seen in either samples, as there is slight restriction even at the highest concentration (lane 2).

A higher concentration of M.*Bsa*WI was used in an attempt to reach full protection with AdoMet, however, this still could not be achieved, **Figure 3.26.**

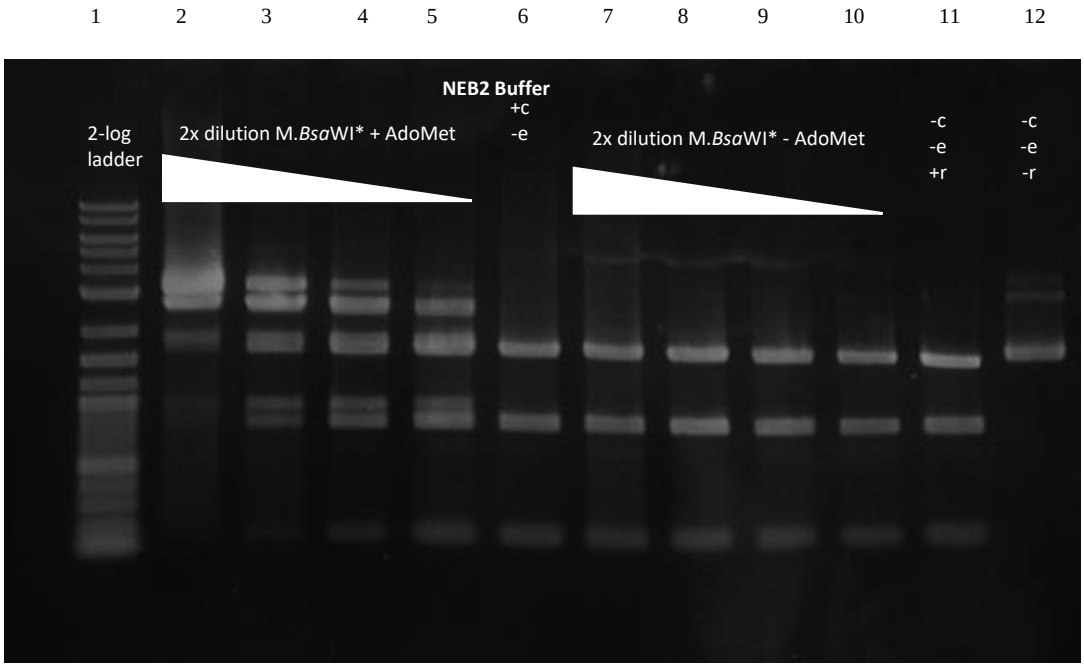


Figure 3.26: Protection assay showing protection of pUC19 DNA by AdoMet and a 2x dilution of M.*Bsa*WI in NEB2 buffer:

Lane 1 = 2-log ladder, lanes 2-5 = 2x dilution M.*BsaW*I with AdoMet, lane 6 = no enzyme + AdoMet, lane 7-10 = 2x dilution M.*BsaW*I - AdoMet, lane 11 = fully restricted, lane 12 = fully protected.

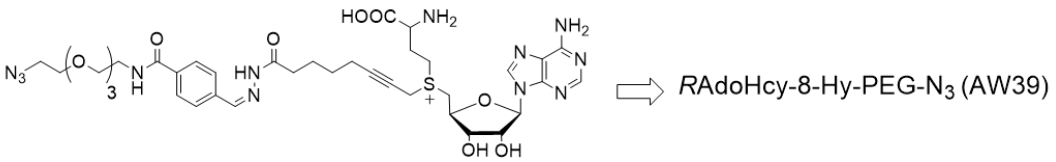
Increasing the concentration of M.*Bsa*WI led to near full protection of DNA with AdoMet as shown by the lack of resitrction in lane 1. Complete protection is still not seen, suggesting that enzyme turnover is too slow for mapping experiments.

Lane 1 shows that using 1.75 µM concentration of M.*Bsa*WI does increase the level of protection, demonstrating that the turnover of enzyme must be very slow, and therefore a high concentration is needed. The DNA is still not fully protected, however, and so the conditions will not be suitable for labelling with M.*Bsa*WI.

After M.*Bsa*WI* failed to protect DNA using AdoHcy-6-N_3 an alternative cofactor analogue was used. AdoHcy-8-Hy-PEG-N_3, shown in **Figure 3.27A**, is a much larger azide-analogue produced by Andrew Wilkinson, which has shown higher turnover rates than AdoHcy-6-N_3 in

assays with M.*Taq*I (unpublished data). It may be also that this cofactor analogue is more suited to the mutation made to the cofactor binding pocket of M.*Bsa*WI, and therefore has greater affinity to it. A protection assay was attempted using this cofactor analogue, but no activity was seen, resulting in full restriction of the DNA, **Figure 3.27B**.

A



B

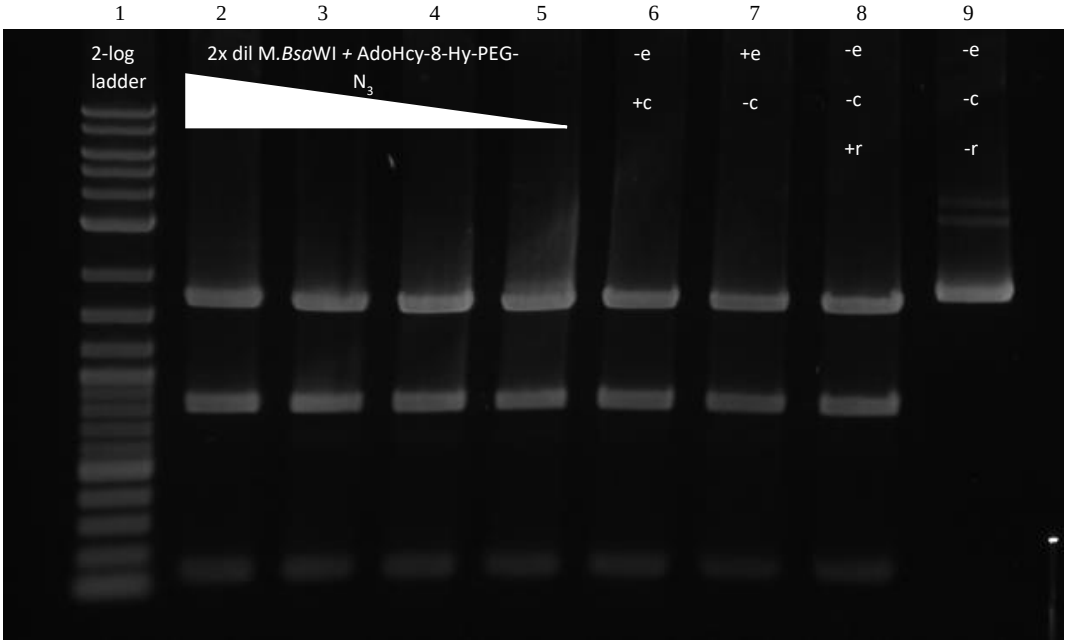


Figure 3.27: A) Structure of AdoMet analogue AdoHcy-8-Hy-PEG-N_3 B) Protection assay showing protection of pUC19 DNA by AdoHcy-8-Hy-PEG-N_3 and a 2x dilution of M.*Bsa*WI in NEB2 buffer:

Lane 1 = 2-log ladder, lanes 2-5 = 2x dilution M.*BsaW*I with AdoHcy-8-Hy-PEG-N_3, lane 6 = no enzyme + AdoHcy-8-Hy-PEG-N_3, lane 7-10 = 2x dilution M.*BsaW*I - AdoHcy-8-Hy-PEG-N_3 lane 11 = fully restricted, lane 12 = fully protected.

M.*Bsa*WI is not active with AdoHcy-8-Hy-PEG-N_3, as demonstrated by full restriction in all test lanes.

It was decided that M.*Bsa*WI will not be suitable for use in this thesis for labelling experiments. None of the mutated MTases proved to be efficient with the synthetic cofactors tested. This could be due to the mutation to the binding pocket not being favourable for AdoHcy-6-N_3 or AdoHcy-8-Hy-PEG-N_3.

It may be that the structure should be investigated further and protein expression optimised to ensure that the best mutations have been made to the cofactor pocket, and that the protein is correctly folded and active. Other synthetic cofactor analogues could also be tested to see if this mutated enzyme has a greater affinity to them, and therefore higher levels of activity. It may also be interested to see how well wild type M.*Bsa*WI interacts with the synthetic cofactors, as it may be that actually the WT has greater activity and is more suitable for labelling than that mutated versions. This should be considered if looking to continue with dual colour, long- and short-range DNA mapping.

3.4 Conclusion

3.4.1 Protection assay summary

This chapter saw a number of protection assays test the activity of different enzymes and cofactors. There following conclusions were made:

- M.*Taq*I protein is active with both AdoMet and AdoHcy-6-N_3, and therefore suitable for labelling experiments.
- Residual AdoMet coeluting with M.*Taq*I protein may inhibit labelling, but can be removed via incubation with TCGA-containing oligos if required.
- The lowest concentration of M.*Taq*I protein should be used in order to not unintentionally add residual AdoMet into the reaction. Low concentrations of M.*Taq*I can still result in full labelling, as demonstrated by mass spec.
- Mutant enzymes (M.*Bsa*WI and M.*Sfo*I) are not fully active with any of the cofactors tested, and are not suitable for further experiments at this point.
- Enzyme M.*Hinc*II is not active with the cofactors tested and is not suitable for further experiments at this point.
- M.*Bse*CI enzyme can be used instead of M.*Hinc*II, and is active with AdoMet. M.*Bse*CI can therefore be used for methylation and mapping experiments with M.*Taq*I.

3.4.2 General conclusion

This chapter shows expression of five enzymes, three of which have not been yet been reported. DNA alkylation was successful with M.*Taq*I, but others were not active and able to achieve full labelling. MTase activity could be improved with optimised protein expression and increased yield to increase turnover rate, as well as further structural studies.

M.*Taq*I has been successfully produced and is efficient in alkylating DNA with AdoHcy-6-N_3, as confirmed by mass spectrometry. It has since been documented that using a higher concentration of M.*Taq*I leads to complete labelling of DNA, leaving none of the sample hemi-labelled, in work carried out by Andrew Wilkinson (not documented). This suggests that labelling efficiency is higher than that stated, and often reaches completion. This technology will be used extensively throughout the rest of this thesis for various imaging applications. Results from gel electrophoresis suggested that around 0.4 nM M.*Taq*I per nM of sites is a suitable amount to ensure full labelling without unnecessarily adding too much excess AdoMet into the reaction. Results from mass spectrometry showed that M.*Taq*I has a turnover of 19 in an hour of labelling, and so an even lower MTase concentration than that tested in gel electrophoresis could be used. Care should be taken to use as little M.*Taq*I as possible in each reaction to ensure that excess residual AdoMet is not unnecessarily added into the mix, which could result in methylation rather than labelling of the DNA of interest. If single-molecule work is to be carried out, further investigation into removal of residual AdoMet may be necessary to ensure efficient labelling. This may involve extensive washing of the protein during the purification process[121].

If wanting to continue the production of a toolbox for dual-colour labelling, further rational design and structural work into mutant enzymes should be carried out. This will ensure that the mutations made are optimal for the specific synthetic cofactors produced. Random mutagenesis could also be explored as a means to test different mutations with a range of cofactor analogues. More work into optimisation of expression conditions should also be performed for the enzymes to ensure that an adequate yield is reached, and that the protein is correctly folded.

CHAPTER FOUR

Optimisation of oligoprobes for chromosome enumeration

4.1 Introduction

This chapter looks to explore the cytogenetic mutations that can occur and are indicative to diseases such as cancer, in particular investigating chromosome aneuploidy. FISH is a technique that is frequently used to diagnose large mutations such as the loss or gain of entire chromosomes, due to its rapid nature and ability to be used on both interphase and metaphase cells[35]. This means that when harvesting samples they do not necessarily have to be synchronised (i.e. applying mitotic blocks to achieve as many cells in metaphase as possible – as is the case for karyotyping), which again saves time and speeds up time to result, which is crucial for many patients that need treatment as quickly as possible. The use of oligoprobes for FISH is investigated in this chapter, in an attempt to further improve turnaround times when looking at chromosome enumeration. The MTase-labelling technology tested in Chapter 3 is used to label oligoprobes to see if it can be effectively applied to this technique, and whether this results in bright probes that are easily detected, with short preparation times.

4.1.1 Detecting genetic instability in cancer

As discussed in the introduction, fluorescence *in situ* hybridisation (FISH) is a cytogenetic technique used to detect and localise specific DNA sequences both in metaphase and interphase cells. Due to the high specificity, sensitivity and speed in which this technique can be carried out, FISH is routinely used both for diagnostics and research for a range of disorders from haematological malignancies to solid tumour samples[35,129].

As discussed in **1.1.4**, genetic instability – which includes both numerical and structural chromosomal abnormalities – is a key hallmark of many cancers[10,20,130]. Aneuploidy, for example, is a numerical abnormality that involves the loss of gain of an entire chromosome, or in some cases multiple chromosomes. This is likely caused by errors that occur in cell

division, described in **1.1.3**, which can result in improper spindle assembly and separation of sister chromatids during mitosis (or meiosis). Uncontrolled cell division is a key characteristic of cancer, due to mutations in genes that encode cell cycle regulator proteins, such as tumour suppressor genes. This means that DNA damage or mutations that give rise to cancer can often go undetected, allowing the mutated cells to rapidly proliferate.

Acute lymphoblastic leukaemia (ALL) is an aggressive form of cancer that affects white blood cells in both adults and children, requiring immediate treatment[131]. Although ALL is rare – with around 650 new diagnoses every year in the UK – it is the most common type of childhood cancer, with approximately 85 % of cases affecting those under the age of 15[132]. Survival rate for children with ALL has recently been reported to be around 90 %[133], with babies and adults significantly lower at 50 %[134] and 35 %[135] respectively. Prompt diagnosis of patients is key in order to administer treatments to patient with ALL as quickly as possible, in a hope to improve prognosis.

ALL is a complex disorder that typically emerges when a lymphoblast gains multiple mutations in the genes that affect blood cell development; research has suggested that these mutations can be inherited – e.g. mutations to the genes p53, CDKN2A/2B or IKZ1 – or caused by environmental risk factors such as radiation[131,133,135]. Individually, one mutation puts a person at low risk of developing ALL, but this increases significantly if there are multiple mutations present[131]. There are many structural and numerical mutations that have been linked to ALL, such as mutations in C-MYC, a transcription factor involved in increased cell division, and translocations of two genes BCR and ABL to form a BCR/ABL fusion gene. The BCR/ABL mutation encodes a tyrosine kinase that promotes cell division,

which is a mechanism that cancer cells then use to rapidly divide and grow. This is a common mutation also found in CML, and is investigated in Chapter 5. As discussed above, aneuploidy is also a hallmark in many cancers, including ALL, where multiple chromosomes can be missing, or duplicated, in each cell; the use of FISH for chromosome enumeration will be explored in this chapter.

4.1.2 Probe design for enumeration

FISH probes are designed to be complementary to the DNA sequence within the region of interest, and labelled with fluorophores. As discussed previously, these probes are annealed to a slide containing fixed patient cells, after heating to a temperature capable of denaturing the DNA – or chemical denaturation – of both probe and patient so that it is single stranded, before cooling to 37 ºC, for hybridisation. Probes can be used on their own or in combination in a probe cocktail. A mixture of fluorescent dyes (usually red, green and blue) can be used to label the different probes and view different abnormalities within one screening.

There are three main types of probes that are routinely used in FISH – locus-specific, centromeric/telomeric and whole chromosome paint, **Figure 4.1** – the use of which depends on which chromosomal abnormality is being investigated[136].

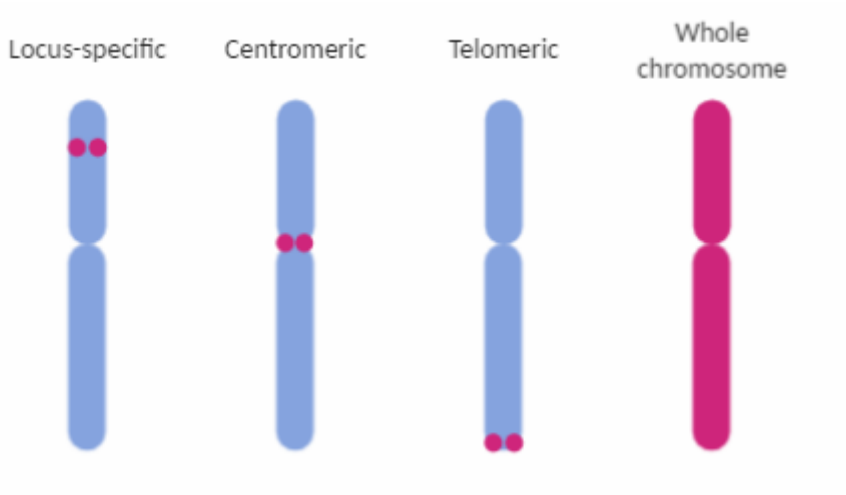


Figure 4.1: There are four main probe types. Locus-specific probes target particular regions or genes, and are useful for loss/gain of a region or translocations. Centromeric and telomeric probes bind to repetitive regions at the middle or ends of a chromosome, respectively. Whole chromosome paints are useful for origin of marker chromosome analysis.

Locus-specific probes are designed to by complementary to a region or a gene of interest, and can therefore detect gene translocations, deletions and amplification; this is explored in Chapter 5. Whole chromosome paints can be used as enumeration probes, but are mostly used to determine whether unbalanced or balanced chromosomal rearrangements have occurred, or to identify the origin of additional material found within the cell[137].

Centromeric and telomeric probes target α-satellite repetitive regions located at the centromere, where two sister chromatids meet, or on the telomeres at the end of chromosomes[138–140]. Centromeric probes are generally used as enumeration probes, and are highly useful in many genetic disorders including those associated with trisomies, as well as cases of cancer. ALL in particular, is a complex disorder that characterises itself in many ways, one of which being the loss or gain or various chromosomes; this can be detected using

centromeric FISH probes. There are complexities in designing centromeric probes for FISH, however, due to the repetitive nature of the centromere.

As stated above, centromeres form part of a specialised DNA sequence that join each pair of sister chromatids, **Figure 4.2**.

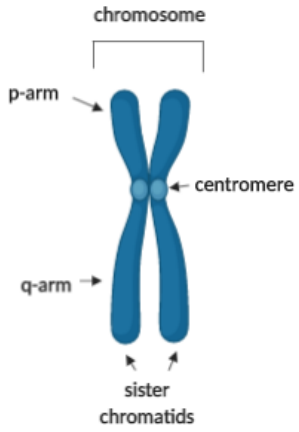


Figure 4.2: Chromosome with centromere linking two sister chromatids in the middle.

In most eukaryotic cells, the centromere's DNA sequence is made up of large arrays of non-coding repetitive DNA (satellite DNA) and, in humans, the primary centromeric unit is called α-satellite (or alphoid DNA)[141]. This satellite DNA is the main component of all centromeres, and a structural constituent of heterochromatin (condensed DNA)[142]. Each unit is based on a 171 bp A-T rich monomer, which makes up a higher order repeat (HOR) that is tandemly repeated potentially hundreds to thousands of times, spanning megabases in total[143].

Monomers from each specific chromosome are reported to be 50 to 70 % identical[144,145] so there are regions that are highly repetitive across all chromosomes, and FISH probes need to be designed for this region carefully to avoid cross-hybridisation[143,146,147]. For chromosome

specific units, each repeat within a HOR is approximately 97 to 100 % identical, and so this conserved region can be used for probe design. O'Keefe *et al.*, investigate the 2.7 kb alpha-satellite HOR repeat unit for 17CEN in a paper in 1996,[148] which is comprised of 16 monomers each of 171 bp. They report the use of oligoprobes to study this region, noting that the probes can distinguish between two highly homologous sequences within this region, differing at only 4 bases to each other. These probes have since been used in many studies to investigate methods of detecting both of these highly similar sequences[148–150]. These sequences will be used in this thesis to explore the use of oligoprobes for 17CEN enumeration.

4.1.3 Oligoprobes for FISH

Oligoprobes are short sequences of DNA (around 50 bp) designed to be complementary to the region of interest[18]. Unlike most commonly used FISH probes, they are not derived from BACs, but are designed synthetically[16]. Due to the short length and low complexity of the probe, this leads to faster hybridisation kinetics compared to traditional probes (which can be hundreds of kilobases in length), as well as greater specificity to the target[19]. If these were to be implemented clinically, this could result in faster results for patients, making them highly favourable over standard probes. These faster hybridisation times also allow potential for same-day diagnosis of diseases that typically take multiple rounds to detect. As stated above, for presentation of ALL at WMRGL, an initial round testing for BCR/ABL, MLL and TEL/AML1 is performed, which takes 16 hours to hybridise. If the initial round shows no abnormalities, a second round looking highlighting IGH and E2A is tested, again with an overnight hybridisation. With shorter hybridisation times, both rounds could be performed and analysed on the same day, increasing diagnosis time significantly.

Another benefit of these synthetic probes is the ability to design and tailor them with high specificity to target uncommon abnormalities and variations[20]. This flexibility sets them apart from other FISH probe companies who are only able to create probes for common abnormalities, or those that are easily available within a BAC library. There is an increasing amount of sequence information available now, thanks to the advancement of sequencing technologies, making it easier to design oligoprobes to unique ROIs[45,151–153].

4.1.4 Labelling of probes

As discussed briefly in Chapter 1, there are various ways to fluorescently label DNA, many of which vary in the amount of specialist training and reagents required. As MTase-directed labelling could be performed easily with a kit simply containing a buffer, MTase and cofactor analogue, it seems like a suitable method for labelling of oligoprobes in both a diagnostic and research lab. It is also incredibly easy to incorporate MTase recognition sites into oligoprobe sequence designs, allowing control over the position, and quantity, of fluorophores, giving the user a high level of control over the sensitivity of the probe. Currently, BACs are mostly labelled using nick translation, a method developed by Rigby and Paul Berg in 1977[106]. This is a fairly time-consuming process, where DNA is treated with deoxyribonuclease I (DNase I) which creates nicks in the phosphate backbone of the sequence. This is followed by the addition of DNA Polymerase I which, by 5'-3'exonuclease activity, replaces nucleotides at nicked sites with fluorescently tagged dNTPs. It also usually requires the addition of ligases to fill in any non-specific nicks. In contrast, using MTases is a much simpler protocol, which requires less steps, less reagents and takes less time, while offering increased flexibility over probe design.

4.1.5 Fluorophore choice and properties

As discussed in the introduction of this thesis, fluorescent probes are invaluable tools, as they allow the investigation of the structure and function of biomolecules, as well as the sequence of DNA within an organism's genome. In this way, they can be used to detect genetic abnormalities that may play a role in the pathogenesis of certain diseases.

When using FISH in the clinic, it is common for multiple probes to be used simultaneously in order to detect a range of mutations. Quite often this is to check for abnormalities that routinely occur together, or to rule multiple mutations out at once. It is often also the case that two probes are used as one can be a positive control for a single gene probe (i.e. using a control probe for the centromere of chromosome 17, while using a single gene probe for the same chromosome, such as p53). For this reason, it is necessary that multiple fluorophores are used that are stable, spectrally distinguishable and bright. For MTase labelling, it is also crucial that these fluorophores are compatible with our labelling technology, which is discussed further in the results section of this chapter.

In this chapter, different fluorophores are used in an attempt to perform multiple hybridisations of various loci at once. This means that fluorophores had to be considered that would be efficient with both the chemistry being used, and the microscope set up (i.e. the excitation and emission must be suitable for the lasers and filters in the lab). Other factors that can be considered are the quantum yield (i.e. the number of photons emitted per photons absorbed, which determines the fluorophores' efficiency) and photostability of the fluorophores[154–156]. Quantum yield is directly proportional to how bright the probe will be, and so the higher this is, the higher the potential SNR. Photostability is important as the

fluorophores are irreversibly destroyed by photobleaching when exposed to high laser powers and long excitation times, which will result in decreased SNR.

The main deciding factor for choosing fluorophores for oligoprobes in this thesis was that they were spectrally distinguishable and compatible with the MTase labelling technology. There are now many dyes that have been developed across the UV/vis spectrum that are commercially available and capable of easy coupling chemistry. Dyes can easily be chosen based on their brightness and photostability, and their emission and excitation checked to ensure that they do not overlap with each other spectrally if wanting to use more than one dye simultaneously.

4.2 Aims

The aim of this chapter is to explore the use of MTase-labelled oligoprobes in FISH. Due to the short size of the oligoprobes in comparison to the more commonly used BAC-derived probes, it is hypothesised that hybridisation to patient slides should occur in a significantly decreased time. As the probes are labelled with MTases rather than other labelling methods, this also allows full flexibility in the location of the fluorophore, as well as being able to easily add additional fluorophores to the probe design, increasing sensitivity.

After discussions with the scientists at West Midlands Regional Genetics lab (WMRGL), who use FISH regularly to diagnose ALL, it emerged that they often perform several "rounds" of FISH tests in order to determine the exact mutations that a patient has. This can also determine the prognosis and direction for treatment, as some cytogenetic subtypes have worse prognosis than others. Mutations that are tested for are initially BCR/ABL, MLL/ TEL/AML1, followed by IGH and E2A, these look for various translocations of amplifications that could be involved in the development of ALL. Enumeration of various chromosomes (1, 7, 17 followed by X, 6 and 10) is then tested, with loss of 1, 7 or 17 being associated with poor prognosis for ALL patients. As each hybridisation typically takes 16 hours, this process can often take multiple days to reach a diagnosis using the current FISH protocol. It was hypothesised that by using oligoprobes, it may be possible to speed up the time to result for these tests, potentially providing the option to perform several tests in a single day. This could be a more efficient way of diagnosing patients, as multiple mutations can be detected in a significantly shorter timeframe, reaching the conclusion of ALL much quicker. WMRGL suggested that focus was initially placed on designing probes to detect loss of chromosome 17, to see if using the MTase labelling technology combined with oligoprobes could provide a quicker and more efficient means to obtaining results for patients

potentially with ALL in critical need of timely treatment. This chapter will aim to label oligoprobes designed specifically for the 17CEN region with M.*Taq*I and AdoHcy-6-N_3 to see if hybridisation times can be improved, and that bright probes can be detected. Conditions – including wash buffer stringency, probe concentration, hybridisation buffer components and probe design – will be analysed and optimised in order to be confident that the oligoprobes can hybridise consistently, and correctly identify the region of interest (ROI). If successful, probes for the centromere of chromosome 1 and 7 will also be explored.

Research has found that oligoprobes are able to discriminate between cytogenetically indistinguishable homologous samples[44]. Structural variations that differ only at a few bases are able to be detected by these oligoprobes when designed to target these areas[148]. This provides huge potential in diagnosing disorders and abnormalities that contain SNPs or other small structural changes that current genetic techniques struggle to detect. Spinal Muscular Atrophy (SMA), a neuromuscular disorder, is the most common genetic cause of death in infancy[157]. The disease is characterised by mutations and therefore loss of functionality in the gene SMN1[111]. Nearly identical gene SMN2, which only has one critical nucleotide difference, can be present in variable numbers in patients and therefore restore some of the functionality lost from the SMN1 mutation[112]. This can result in varying levels of severity of the disease. Due to the similarity in sequence, traditional FISH cannot currently be used to distinguish between SMN1 and SMN2, which is critical for successful diagnosis and treatment. Due to variations in copy number on SMN2 gene, as well as having different conformations of SMN1 (1:1 or 2:0 carriers) this also causes problems in diagnosis using molecular techniques. Using oligoprobes that can distinguish between highly similar sequences may prove to be an invaluable technique, and will be explored further in this chapter, by attempting to detect highly homologous centromere sequences.

4.3 Results and discussion

The overall aim of this chapter was to test and optimise the use of MTase-labelled oligoprobes in FISH. Using M.*Taq*I and AdoHcy-6-N_3 that had been tested in the previous chapter, this labelling technology was used to attach fluorophores to a DNA hairpin targeting the centromeric region of chromosome 17 – followed by the centromere of chromosomes 1 and 7. This chapter hopes to produce an oligoprobe that can successfully bind to the target, and be detectable using fluorescence microscopy. Automating the process of assessing signal to noise ratio were also investigated in order to determine the best conditions for the probes. Using oligoprobes may make it possible to achieve faster hybridisation times than current FISH workflows, which could result in faster turnaround times, quicker results for patients and prompt treatment for a range of diseases. They can also be designed for any region of the genome and engineered to avoid specific repetitive sequences.

4.3.1 Hairpin oligoprobes for rapid chromosome enumeration

As discussed in the introduction of this chapter, oligoprobes have been proven to hybridise more rapidly to ROIs in FISH. This is due to their short size (typically < 100 bases) resulting in faster hybridisation kinetics[38,46]. The potential for the use of oligoprobes in diagnostic laboratories is vast. These short hybridisation times could mean faster patient diagnosis, ultimately leading to quicker administration of treatment. After discussion with West Midlands Regional Genetics Laboratory (WMRGL), it was decided that as proof of concept a probe for the centromere of chromosome 17 (17CEN) would be designed. Loss of 17CEN can be indicative of acute lymphocytic leukaemia (ALL), and is currently tested (along with a probe for chromosome 1 and 7) using FISH in a 16-hour hybridisation. As ALL is a complex disease, it can take many rounds of FISH to come to a conclusive diagnosis. This means that with each round taking 16 hours, it can be days before a diagnosis is met, and a patient's

condition could rapidly deteriorate while waiting for treatment. With the ability of oligoprobes to potentially hybridise to ROIs much faster than traditional probes, this technology could be used to provide rapid diagnosis in laboratories. Multiple tests could also be performed in a single day, rather than waiting for an overnight hybridisation with each round of testing, increasing the chance of a timely delivery of treatment to the patient.

Oligoprobes were designed using short sequences targeting a region within the centromere of chromosome 17. The sequence was obtained from work by O'Keefe *et al.*[44] and confirmed by performing a BLAT search against the human genome. A BLAT search is a tool that allows DNA sequences to be compared about the human genome, highlighting matches of 95 % or greater for 25 bases or more. This region was selected as it was identified as being unique to chromosome 17 and should not hybridise elsewhere under stringent wash conditions. According to the study, this region of DNA can exist in two highly homologous forms. The two variants, which differ at 4 base positions, will be referred to as 17CEN1 and 17CEN2. Humans are thought to either have 17CEN1, 17CEN2, or a mixture of both within their chromosome 17 centromeres, which will be explored later in this thesis. A hairpin design included a double-stranded portion of the probe, shown in **Figure 4.3A**. This double-stranded region is necessary for future MTase labelling – which will be tested if the initial design is successful after trialling with prelabelled probes – so that the MTase can bind to and label the DNA strand[77]. To test the efficiency of the hairpin probe design to hybridise uniquely to chromosome 17, and to see if such as small piece of DNA can be detected, 17CEN1 and 2 were ordered from IDT DNA with a Texas Red NHS-ester (Abs 596 Em 613) conjugated to the 5' end. The two variants were mixed to ensure the area could be detected (in case the patient only had one or the other). The prelabelled 17CEN probe was hybridised to a patient sample as described in **2.2.7**, and visualised using an inverted, epifluorescence microscope

equipped with a 100x objective lens (Nikon, 1.49/oil TIRF) and cooled EMCCD camera (Photometrics, Evolve® 512 Delta). Excitation at 405 nm and 561 nm was achieved using solid state lasers (Coherent, OBIS) to visualise DAPI (nuclear stain) and Texas Red (probe) respectively. As can be seen in **Figure 4.3B**, the probe successfully bound to chromosome 17 and could be clearly detected under these conditions. This confirms not only that the hairpin design can efficiency bind to the human genome and is detectable using FISH, but also that the recognition sequence is specific to that loci, and does not hybridise elsewhere under these conditions. The result also came with just a 15-minute hybridisation time – significantly lower than the 16-hour hybridisation time used currently by clinical laboratory protocols. This could have a huge impact on the current turnaround times of FISH results.

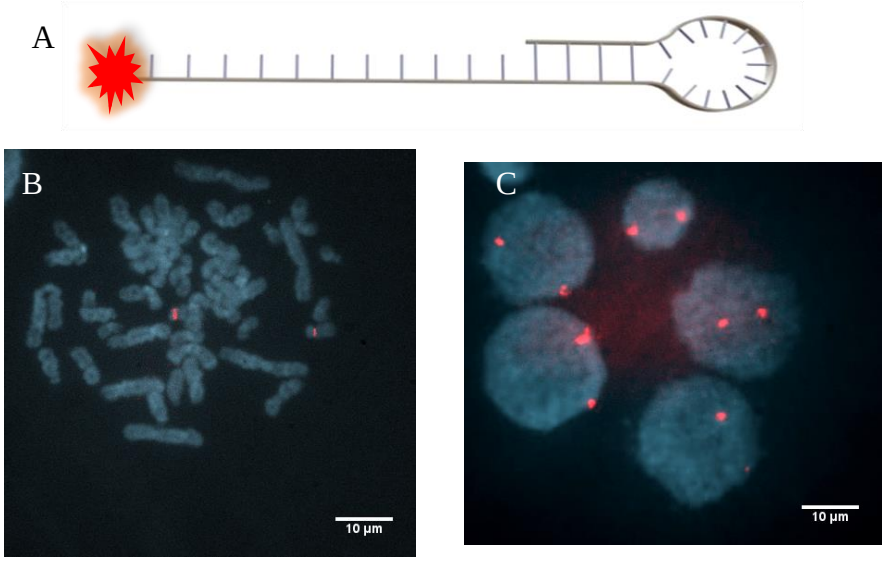


Figure 4.3: A) Schematic of hairpin probe design with fluorophore at 5' end B) A hairpin 17CEN probe visualised by FISH on human methaphase and C) interphase nuclei cells after 15-minute hybridisation.

While the prelabelled 17CEN probes were successful in highlighting the ROI, they were expensive to order; 100 nM of Texas Red-labelled oligo (the lowest concentration that could be ordered from IDT for this size oligo) was ~£70, and both 17CEN1 and 2 needed to be

ordered to account for both homologues of that region. This means that if different ROIs were needed – which would be the case for anything other than repetitive centromeric/telomeric regions – the cost of labelling oligoprobes in this way would be too high for many diagnostic laboratories. If wanting to look at single genes for example, where potentially hundreds of oligos are across the ROI, the price would increase substantially. By labelling the hairpins using MTase-labelling technology the price can be lowered significantly, as the enzyme and cofactor are manufactured inhouse and at a high yield, with only microlitres of each being used per labelling reaction. As discovered in Chapter 3, MTase labelling is best when using the lowest possible concentration of M.*Taq*I, and the amounts of cofactor and enzyme needed to label the oligoprobes would be incredibly small for a large batch of probes. This technology could therefore lower the price of probes significantly. Labelling with MTases also allows control over how many fluorophores can be attached to the probe by incorporating extra MTase recognition sites within the probe design, as well as offering the flexibility of where to place the sites within the probe sequence, or even to have different MTase recognition sites within one probe for dual colour. The hairpin structure in the sequence of the oligoprobe, shown in **Figure 4.4**, is critical to the labelling procedure as MTases recognise, and bind to, double stranded DNA sequences[77].

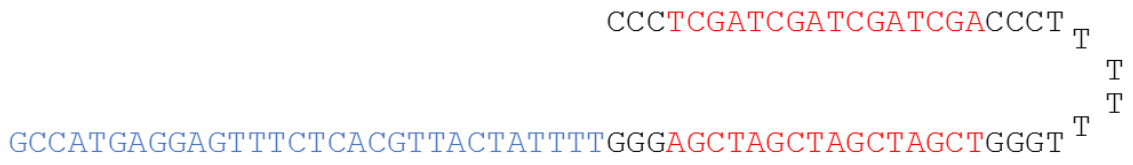


Figure 4.4: Schematic of oligoprobe targeting 17CEN1. By altering the amount of red "TCGA" sites within the sequence, extra fluorophores can be attached to the probe by M.*Taq*I, which should increase sensitivity.

By incorporating a hairpin into the design, this allows the recognition sequence to occur as a double stranded piece of DNA, leaving the binding site open for hybridisation to the ROI. By labelling using MTases, the number of labelling sites can be altered by the addition of extra recognition sequences within the design. Addition of extra sites could increase sensitivity, or increase the chance of a probe having a label attached, which will be explored later in this chapter. This means that if alkylation does not reach 100 % completion, i.e. not all sites receive azide functionality, that even a hemi-methylated strand could potentially be detected if the fluorophore is bright enough. M.*Taq*I (recognition sequence TCG*A*) will be used to label the oligoprobes, due to its efficiency with synthetic cofactors – namely AdoHcy-6-N_3 – as shown in Chapter 3.

The 17CEN hairpin sequences were ordered from IDT without the attached fluorophore, to be labelled using the MTase technology at a significantly reduced cost (100 nM unlabelled oligo was ~£10 whereas 100 nM Texas Red-labelled oligo was ~£70). Based on these prices, if wanting to order multiple colours and multiples ROIs – which would be necessary if wanting to use these probes in a diagnostic lab – the prelabelled probes would be far too expensive for many laboratories to consider. As M.*Taq*I and AdoHcy-6-N_3 were successfully used for alkylation experiments in Chapter 3, they were now used to attempt labelling of the oligoprobes. 17CEN1 and 17CEN2 were labelled with DBCO dyes using SPAAC click-chemistry as discussed in Chapter 3. There are limited DBCO dyes commercially available, but TAMRA DBCO (Abs 553/Em 575) was initially tested as its properties were like that of previously successful Texas Red, and excitation was achieved using a 561 nm laser.

The oligoprobe was labelled and hybridised to the sample for 15 minutes as described in **2.2.4** and **2.2.7**. As can be seen in **Figure 4.5**, the oligoprobes can clearly be detected under the conditions used. This means that the oligoprobe design is suitable for MTase-labelling experiments, opening up the potential for this technology to be used in wider FISH applications for various mutations and diseases. This could have a significant impact on turnaround-times for patient results in clinical laboratories, and shows huge potential for prompt diagnosis of diseases such as ALL where rapid treatment is required. The flexibility and affordability of this technology could be revolutionary to dianogstic labs that require quick and efficient testing in often complicated situations, such as in cases of ALL where the karyotype is can be complex and requires multiple rounds of testing.

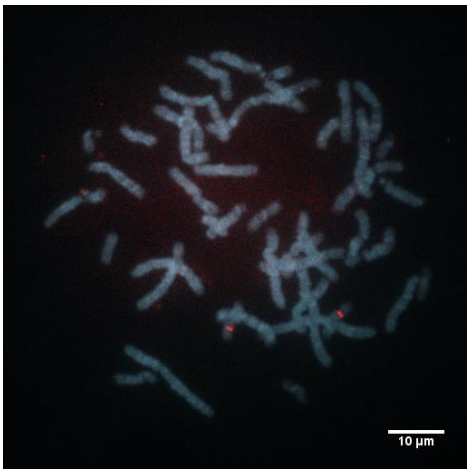


Figure 4.5: Human metaphase nuclei on 46XX/XY sample showing a successful 15-minute hybridisation of MTase-labelled 17CEN1/2 with TAMRA DBCO.

4.4 Optimisation of oligoprobe design

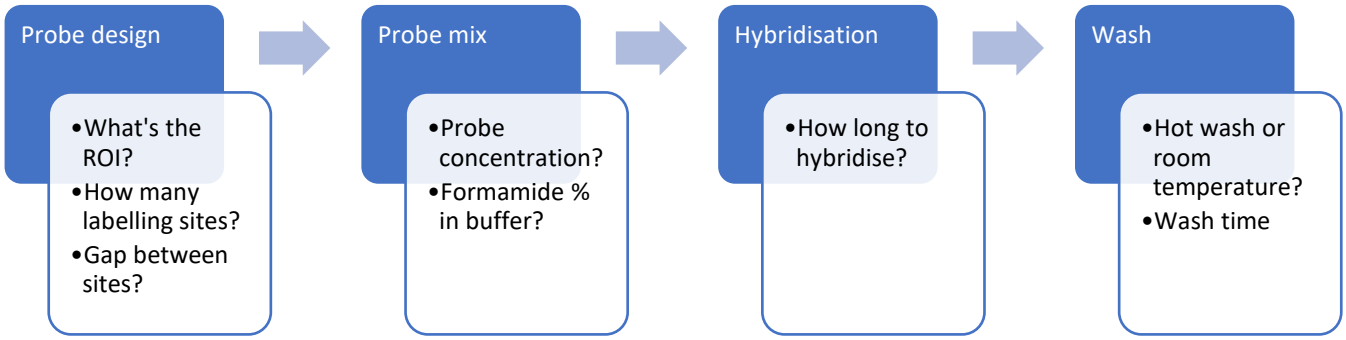


Figure 4.6: Scheme of FISH process and points for optimisation. Different parameters were tested in terms of probe design, probe mixture, hybridisation conditions and wash stringencies in order to produce the highest SNR.

To obtain efficient and reproducible results, conditions for MTase-FISH need to be optimised. There were many factors to consider at each stage of the protocol, such as the oligoprobe design itself, number of fluorophores, concentration of probe, hybridisation conditions, and wash stringencies (**Figure 4.6**), all of which are discussed in this chapter. It's important to note that each condition was tested with samples from the same patient – prepared under the exact same conditions and at the same time – to keep variables as consistent as possible, as some patients may have more of less copies of the repetitive centromeric region. This could result in some probes appearing brighter than others if a patient had more copies of that region, due to the increased number of sites rather than the change in condition. Conditions were assessed qualitatively by manual visualisation and, where possible, quantitatively using the image analysis software, Icy[159]. Quantitative analysis was not always possible if the samples had particularly high background, or if samples were of poor quality. Protocols were designed for Icy to detect interphase nuclei (**Figure 4.7A**), as well as the spots within them, and obtain the fluorescence intensity of those areas. A Gaussian filter and Otsu threshold (**Figure 4.7B**) were set to detect cell outlines (the ROI)

from the 405 channel, and tools for spot detection (using the 561 channel) within those ROIs detected probes (**Figure 4.7C**). Background and artefacts visible outside of the ROIs (such as the spot seen outside of the nuclei in **Figure 4.7C**) were ignored by the software. Intensity of both the nuclei and the spots (probes) were measured in the 561-channel so that signal to noise ratio (SNR) could be calculated.

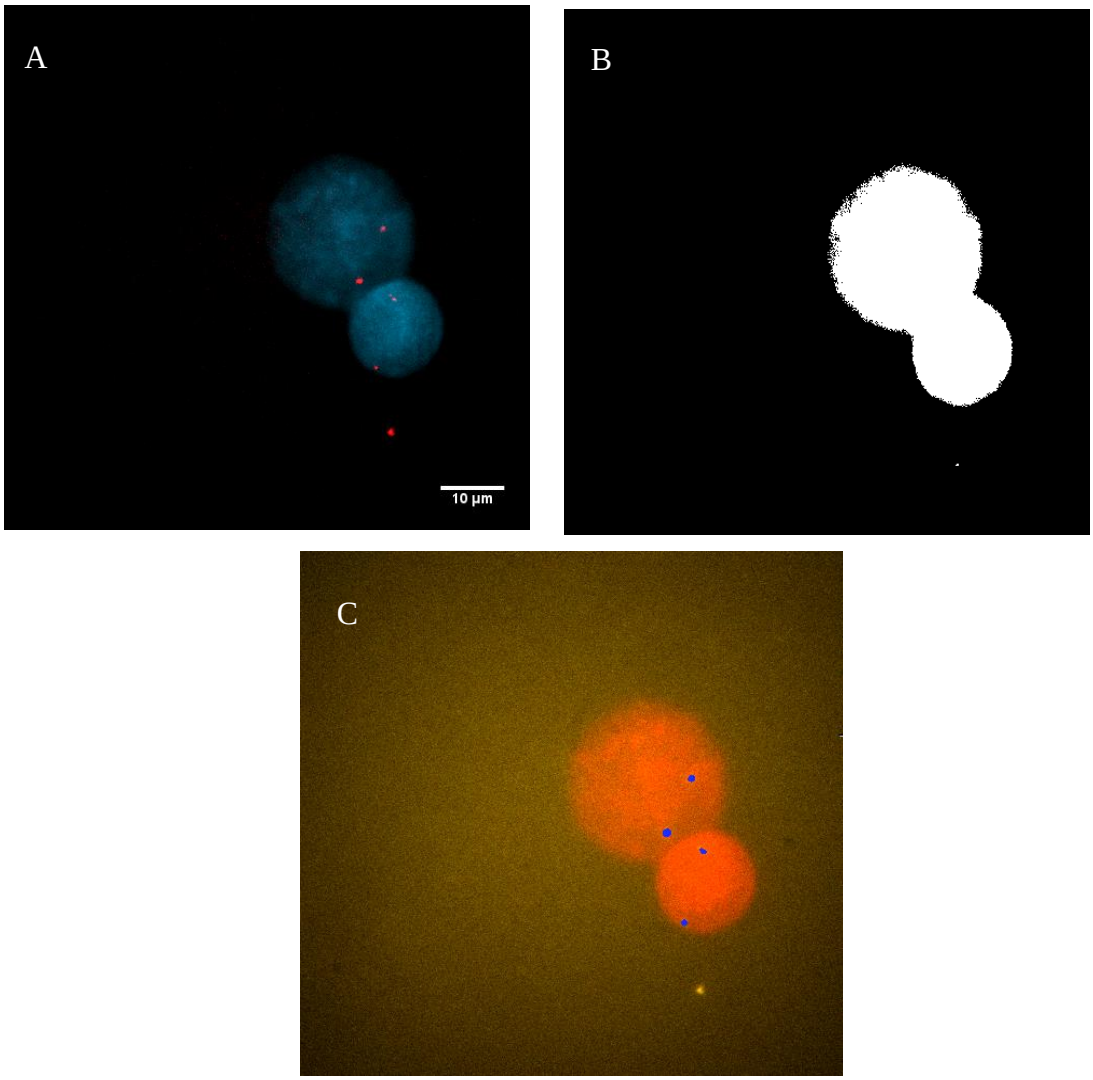


Figure 4.7: A) FISH image of human interphase nuclei with two signals in each B) Icy thresholds to locate interphase nuclei (the ROI) C) Spot detector locates probes (shown in blue) within the interphase nuclei, and ignores those outside. The values for fluorescence intensity for these regions can be collected and used to calculate SNR.

SNR was calculated by dividing the average signal fluorescence by the average background intensity of the nuclei. This shows how bright the probe is in comparison to the background fluorescence, and so the higher the SNR, the better the probe for confident and accurate detection.

First, the probe concentration was examined. After purification, 17CEN (1 and 2 combined) measured at ~ 14 ng/µl. A range of concentrations of TAMRA-labelled oligoprobes were tested and visualised (0 ng/µl, 1 ng/µl, 2 ng/µl, 4 ng/µl, 6 ng/µl, 8 ng/µl and 10 ng/µl). For each of them a standard amount of hybridisation buffer (5 µl) was used, as this is a typical volume used at WMRGL and would be easy to incorporate into the current protocol. Probes were visible for each of the examples, with no obvious improvement in SNR as the concentration increased, as seen in **Figure 4.8**.

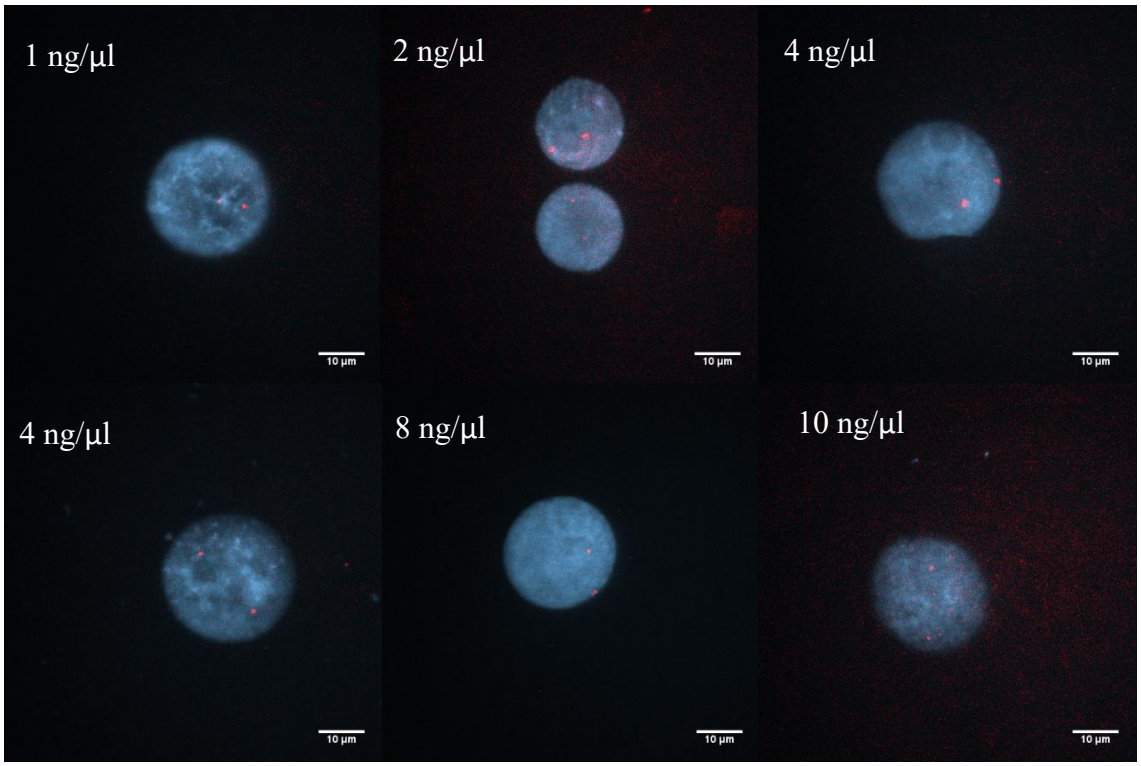


Figure 4.8: Interphase nuclei showing increasing concentration of 17CEN probes labelled with TAMRA DBCO. Probes were visible in all conditions and changing the concentration did not appear to have a significant impact on signal strength.

It is interesting to note that with a low concentration (1 ng/µl) of probe, signals are still detectable using FISH, although not very bright. From 2 ng/µl, background fluorescence appears to have slightly decreased – but not significantly – this may mean that an increased number of probes have hybridised to the target, resulting in higher signal to noise. All samples do have some background fluorescence, potentially caused by excess free-probe on the slide being trapped under the coverslip, suggesting that the wash conditions should be optimised.

Going forward, it appears that adding extra probe does not have a significant effect on SNR based on these qualitative results, and so as little as possible should be used in order to save cost. Adding too much unnecessary probe may also have an effect on the amount of background if not sufficiently washed away from the sample before imaging, or if the stringency is not high enough to remove all of the non-specifically bound probe. It may also be that the volume of hybridisation buffer could have an effect on the amount of probe that can penetrate and bind to the sample. Dextran sulphate – one of the key components in the hybridisation buffer – acts as a volume exclusion agent, enhancing hybridisation by creating a crowded environment, increasing the effective local probe concentration[160]. In future, it could be useful to test different amounts of hybridisation buffer, and therefore dextran sulphate concentrations, to see if this has an effect on hybridisation.

Quantitative analysis allows us to gain further insight into the optimum conditions for probe concentration. Between 50 and 100 interphase nuclei images of each sample were taken and analysed using Icy. From these images, the mean fluorescence intensity (MFI) of the probe were calculated using spot detection with the intensity of the 561 laser for that region. SNR

was calculated by dividing the MFI of the spots by the MFI of the whole nucleus in the 561 channel; a higher SNR means that the signal is brighter than the background and is the best choice for FISH conditions. The results in **Figure 4.9**, appear to confirm the qualitative result, that increasing the probe concentration does not have a significant effect on SNR. There is no noteworthy change in the median SNR (as shown by the horizontal line across each box plot) or mean (shown as an X) across all concentrations. All of the boxes overlap, demonstrating the that the data is similar across each condition. The "whiskers" of each box plot are also fairly consistent, suggesting again that there is not much variation between the conditions, and that most of the data for each plot is close to the mean value.

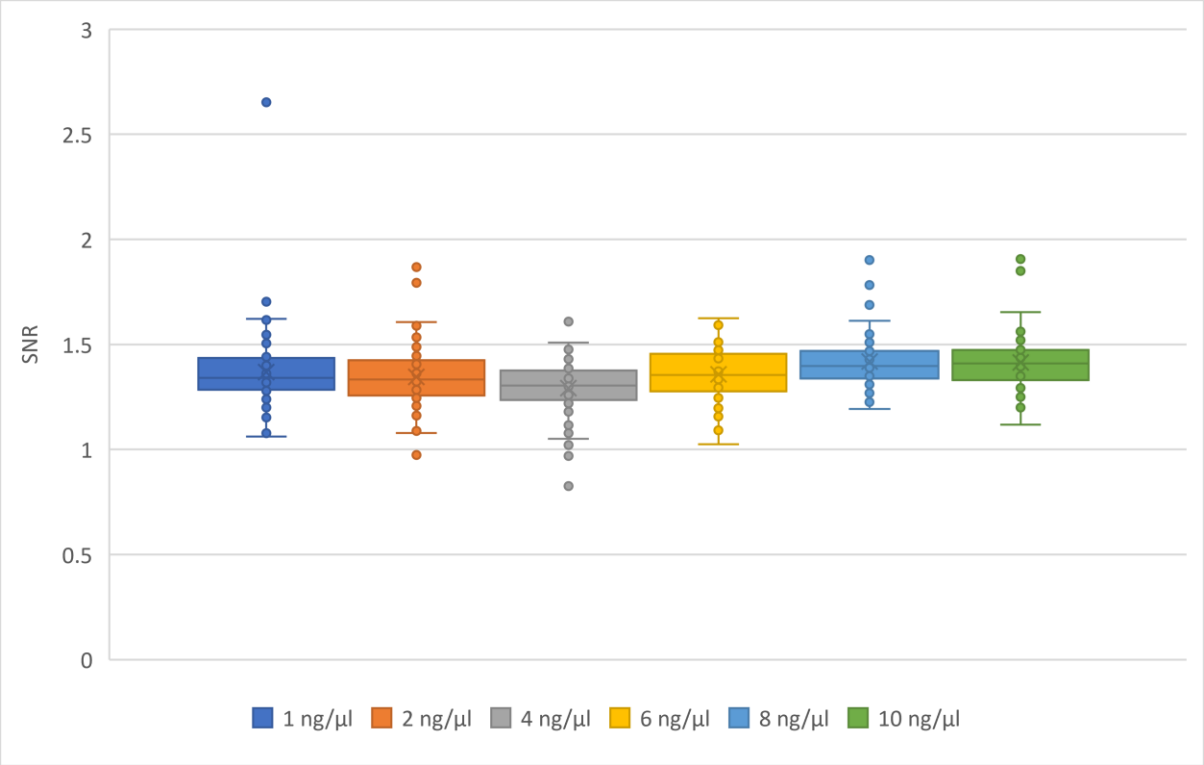


Figure 4.9: Box plot showing SNR of different concentrations of 17CEN TAMRA-labelled probe. There was no significant variation between conditions.

It is important to note that, as stated above, to keep costs down, it would be therefore be optimal to use the least amount of probe necessary for detection. Testing wash stringencies

may be a way to further reduce background noise in the cells and increase sensitivity of the probes, as it could prevent non-specific binding; this will be explored later in this chapter.

FISH probes are already expensive to order and, if wanting to use prelabelled-oligoprobes, single gene detection - i.e. needing a new probe for each loci – could be impractical due to cost. It is not certain how many fluorophores are needed to detect a signal, but the fewer, the better, in terms of cost per test. By producing an alternative that is just as bright, or brighter, than the probes currently on the market, this could reduce the number of probes needed (i.e. the concentration needed) for a visible signal, and therefore reduce cost. By labelling oligoprobes with MTases, this allows flexibility in how many labels are attached to the probe. Incorporating TCGA sites into the probe designs means any number of M.*Taq*I sites can be added, potentially making the probe brighter. A bright probe, combined with faster hybridisation rates, could have a significant impact in terms of quality of – and confidence in – results for diagnostic labs.

Five different probe designs were tested, each with either a different number of TCGA sites, or different "spacers" between sites. Probes with one, two or three sites placed directly next to each other were used, as well as one with two sites separated by 2 bp and one with two sites separated by 15 bp (**Figure 4.10**). 2 bp spacing was chosen to see if this could affect MTase labelling efficiency, as the extra bases were thought to be potentially needed for the MTase to dock onto the strand. The 15 bp spacing was chosen in an attempt to prevent quenching of fluorophores as typical donor-acceptor systems see quenching of dyes at a distance of <50 Å, which equates to <15 bases[161].

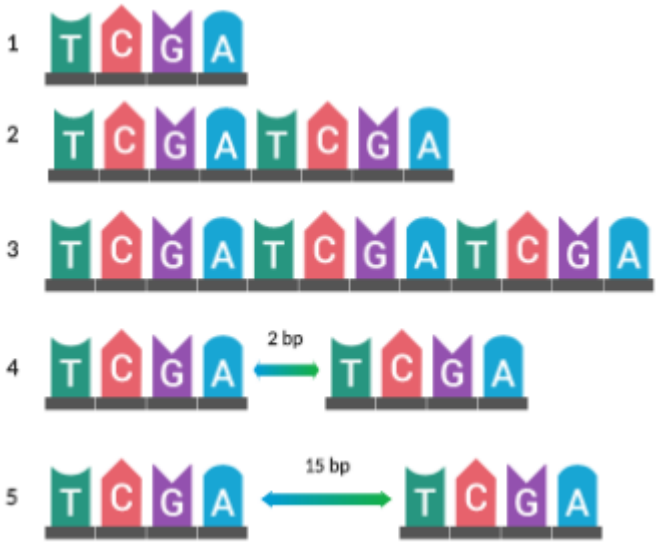


Figure 4.10: Five different probe designs were ordered with varying amounts of M.*Taq*I sites or spacers to see if this had an impact on SNR.

Figure 4.11 shows that increasing the number of labelling sites did not appear to significantly improve the SNR. The median and mean SNR do seem to slightly increase with the number of sites, but not significantly so. This could be further tested with the addition of more sites into the sequence, but this would also increase the cost per probe, which should be a consideration during design. The 2 bp spacing between sites also appeared to have no effect on labelling efficiency, as the 2 bp sample has very similar SNR median and interquartile values as the 2-site sample with no spacing. It could be that in both samples the fluorophores are quenched from being too close to one another, which is why they have similar SNR to the 1-site and 3-site samples. Having 15 bp spacing between sites does seem to have had a positive effect on SNR, with the median of this sample being slightly higher than all other samples, including the 3-site sample. An increased spacing of 30 bp could be considered to see if this has an even bigger impact on the intensity of the fluorophores, although this increase in probe size could also increase hybridisation time and cost. A 3-site probe with 15 bp could also be considered for testing to see if this again increases SNR, but this would be

more expensive to order and could be seen as unnecessary if the 1-site probes are detectable. This may need to be considered for single gene probes where there are significantly less probes bound to the region than with repetitive centromeric regions.

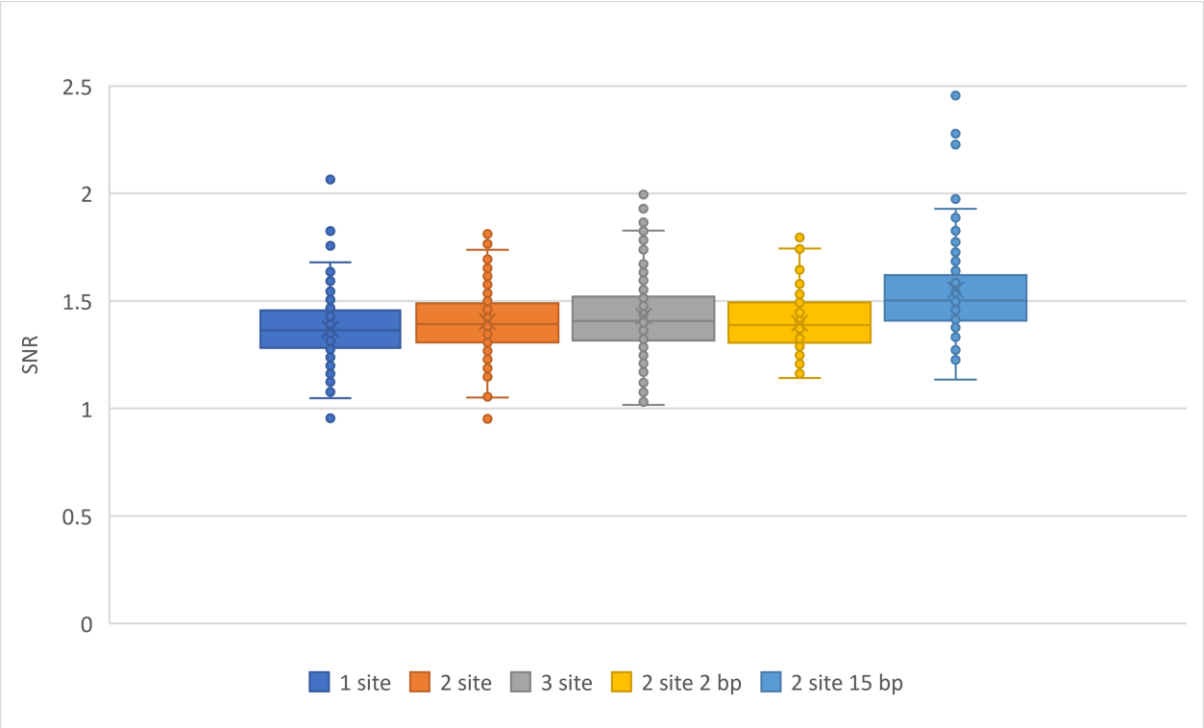


Figure 4.11: Graph showing SNR of different 17CEN probe designs with varied number of labelling sites, or different linker spacing between sites. There appears to be no significant difference, although the 15 bp spacer has a slight increase in SNR.

Recent work by Schröder et al has explored the effect of closely positioned dyes on fluorescence, stating that a stronger signal of fluorescence was obtained with distances of around five base pairs[162]. A distance of around 7 bps indicated permanent separation of dyes, but the oligoprobe used in **Figure 4.6**, with 2 bp linker (and therefore six base pairs between dyes) does not necessarily support this. It is important to note, however, that as M.*Taq*I is palindromic, there is often a dye on both strands that will inevitably interact with one another. A different enzyme (with a non-palindromic recognition site) could be considered if this

appears to be a problem, but the results here do not indicate that this affects the ability to detect the probes.

4.5 Optimisation of oligoFISH conditions

As discussed in the introduction of this chapter, formamide is a commonly used solvent in FISH hybridisation buffers. The addition of formamide results in the destabilisation of DNA complexes as it competes with hydrogen bond formation between Watson-Crick base pairs. Addition of 1 % formamide to a probe mix lowers the melting temperature (T_m) of probe:target by 0.72 °C[163]. T_m is defined by the temperature at which 50 % of double stranded DNA (dsDNA) denatures to become single stranded DNA (ssDNA). This is determined by the probe:target length, as well as the C:G composition. The oligoanalyzer tool on IDT's website was used to determine the T_m of probe:target for 17CEN1 and 2, which were 58.6 °C and 59 °C respectively. Careful consideration of formamide concentration in the hybridisation buffer was needed due to the low probe:target T_m to prevent accidental denaturation during the hybridisation step at 37 °C. Standard hybridisation buffer typically contains ~70 % formamide, which may prove too stringent for small oligoprobes compared to traditional probes (the longer the probe the higher the melting temperature). If too low a stringency is used, however, this may lead to a higher level of non-specific binding resulting in cross-hybridisation and high background.

Hybridisation buffers were made up as listed in **Table 4.1** to include 30 %, 40 %, 50 %, 60 % or 70 % formamide. These buffers were tested along with Cytocell Hybridisation Buffer B (~70 % formamide) with 17CEN1/2 TAMRA DBCO following the protocol in Chapter 2. Each sample was visualised on the microscope and between 50 and 100 images of interphase

nuclei were taken. These images (a representative selection shown in **Figure 4.12**) were analysed using Icy to determine the signal intensity for each visible probe.

Buffer	T_m **17CEN probe:target**
30% formamide	48.2 °C
40% formamide	44.6 °C
50% formamide	41 °C
60% formamide	37.4 °C
70% formamide	33.8 °C
Cytocell 70% formamide	33.8 °C

Table 4.1: Table showing melting temperatures (T_m) of 17CEN probe to target DNA. As formamide concentration increases, the melting temperature of probe to target decreases, which will need to be considered when performing heated steps during the FISH protocol.

From manually assessing the samples by eye, the optimum formamide concentration appeared to be between 40 and 50 per cent. Any higher than this and the probe began to get lost in the background, as contrast in the image needed to be increased to be able to visualise the spots. This suggests that at higher concentrations of formamide, there are not as many probes bound to the region, resulting in a reduction of signal to noise. This is important to consider if planning on moving to other probe designs that are targeting single genes, as each oligo would have a unique sequence to bind to (as opposed to a repetitive sequence in centromeres), and if hybridisation is not efficient enough the site may not be detected at all.

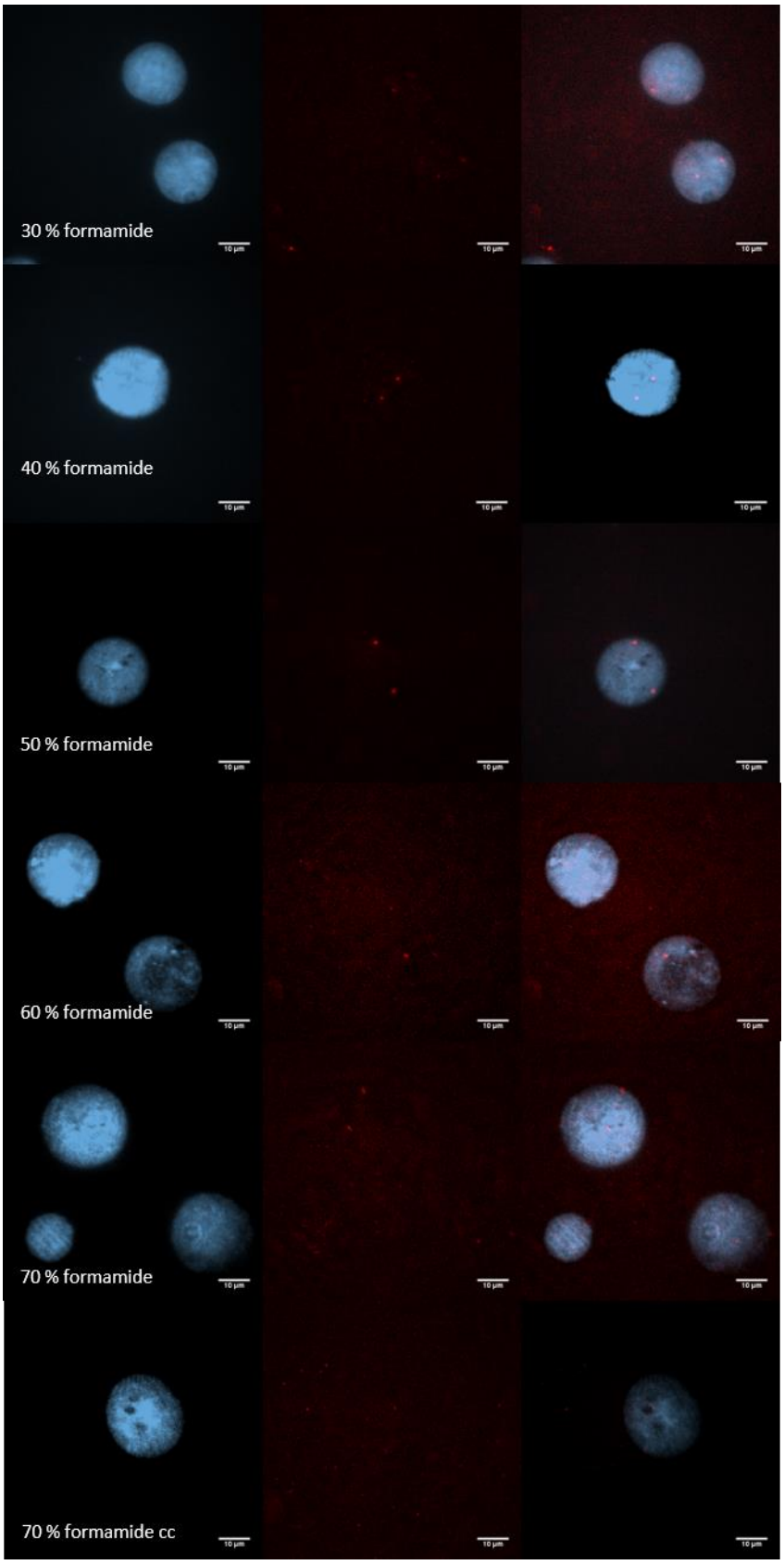


Figure 4.12: Interphase nuclei showing hybridisation of 17CEN probes with varying percentages (30-70 %) of formamide in buffer. Probe signal appeared to reduce as formamide increased, due to less probe binding to the ROI.

Quantitative data showed the same as qualitative, as can be seen in **Figure 4.13**; optimum formamide percentage appeared to be ~40-50 % as this resulted in the highest SNR. SNR medians and means for 40 and 50 % formamide samples are significantly higher than 60 % formamide and above, and the range within the interquartile regions are also higher. 30 % formamide appears to not be stringent enough, resulting in lower signal to noise as there will be non-specifically probes bound that contribute to the background. Going forward, 40 % seems to be the best option for this probe design as it has the least variability across all images (as indicated by a reasonably small box and whiskers), and has the highest mean and median SNR.

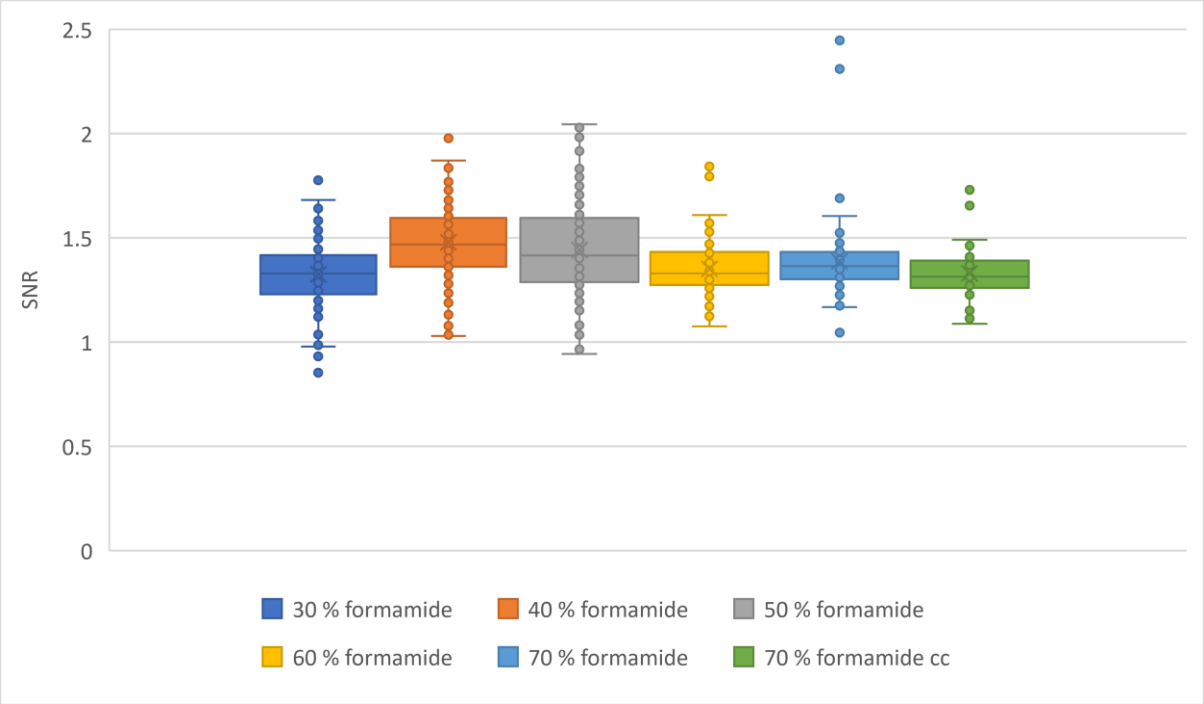


Figure 4.13: Graph showing SNR of 17CEN probe with different percentages of formamide within the hybridisation buffer. Between 40 and 50 % formamide seemed to have a positive effect on SNR.

From looking at the T_ms of the probe:target for different formamide conditions in **Table 4.1**, this supports the results in **Figure 4.12** and **Figure 4.13**. The probes were hybridised to the

patient sample at 37 °C. Increasing the formamide percentage to a point where probe:target is reduced to near the hybridisation temperature means that there is a chance that some of the probes simply will not anneal to the target, and remain as a single-stranded probe. This means that while some probes are annealed and present, the hybridisation is not 100 % efficient. When using 30 % formamide, the condition was not stringent enough, resulting in a higher background which Icy detected (meaning a lower SNR).

Another factor influencing stringency of the reaction is the post-hybridisation wash. This wash removes any non-specifically bound probes from the slide, as well as any free-dye in the solution. The more stringent the wash (i.e. the higher the temperature or lower the salt concentration), the less background will be present in the sample due to removal of weakly bound probes to non-target regions. High salt in the buffer destabilises charge repulsion between the negatively-charged phosphate backbone of the double-stranded DNA, therefore making the complex more stable[160,164]. This means that in using lower salt concentration in the high stringency wash, this reduces the weakly bound probes that are bound non-specifically, reducing background. Heat is also important as when applied to double-stranded DNA, it disrupts the hydrogen bonds between base pairs, again destabilising the complex, and reducing the amount of non-specifically bound probe[160]. Typically, at WMRGL, slides are washed for 2 minutes at 72 ºC in a high-stringency wash buffer (0.4x SSC, 0.3 % IPEGAL), followed by 30 seconds at room temperature in a low-stringency wash buffer (2x SSC, 0.1% IPEGAL). The combination of low ionic concentration and high temperature of the high-stringency buffer destabilises the bond between the probe and any mismatched targets and hence, washes off any non-specifically bound probes.

As oligoprobes are designed to bind specifically to their ROI[44], this makes them more accurate than BAC-derived probes, which bind around the whole region without having 100 % sequence homology. If an oligoprobe was non-specifically bound to a region without 100 % sequence homology, then the T_m of the bound double-stranded region would be low due to the short size of the oligo, meaning that the wash (or the formamide in the hybridisation buffer) would denature the short piece of double-stranded DNA and dissociate the probe. This means that oligoprobes may need a less stringent wash than larger BAC-derived probes. A 72 ºC wash (as carried out in the standard WMRGL protocol) is higher than the T_m of the probe:target, and so room temperature washes were also tested to see the effects on hybridisation.

Five wash conditions were considered as shown in **Table 4.2**. These explored various wash times, as well as a heated wash, to find the optimum condition for bright signals with low background.

Sample	High stringency (0.4x SSC, 0.3% IPEGAL)	Low stringency (2x SSC, 0.1% IPEGAL)
1	1 minute RT	1 minute RT
2	2 minutes 72 ºC	30 seconds RT
3	2 minutes RT	30 seconds RT
4	5 minutes RT	5 minutes RT
5	10 minutes RT	10 minutes RT

Table 4.2: Table showing different wash conditions tested with 17CEN probes, to see if the different stringencies would affect the SNR of the oligoprobes.

The samples used for this experiment consistently had higher levels of background than usual. This was possibly due to poor quality sample from harvesting, or that the sample had deteriorated. In particular, the background on the lower stringency washes was high, leaving artefacts and high levels of non-specifically bound probes within the nucleus. Using Icy to analyse these samples proved difficult due to the high level of background, and so manual, qualitative analysis was performed. Images from each condition were observed and visually analysed to detect signals. The images, found in **Figure 4.14**, were selected to be representative of the interphase nuclei acquired for each condition.

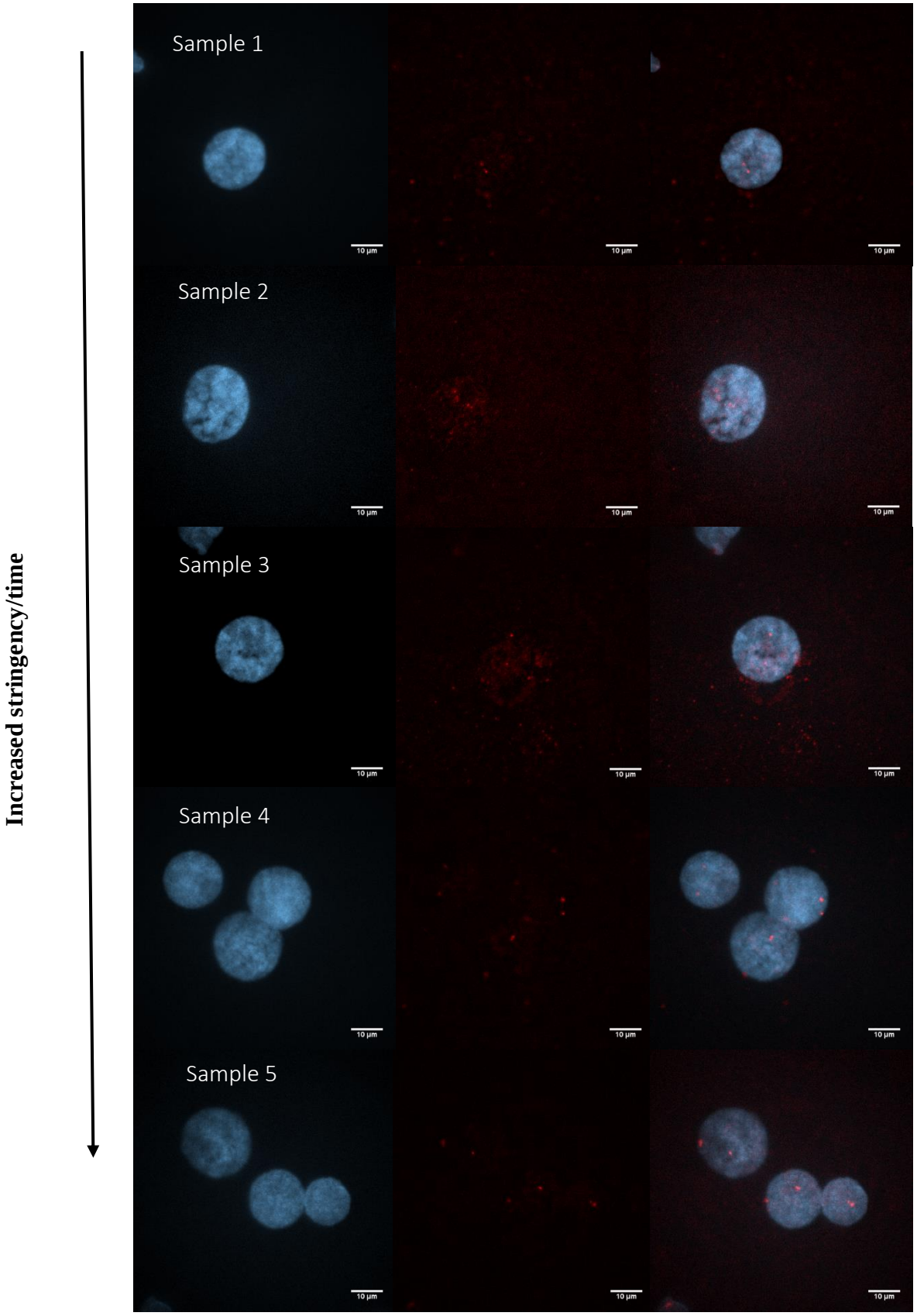


Figure 4.14: Interphase nuclei cells showing the effects of increased stringency in wash buffer conditions. Each condition shows the image in the 405 (DAPI), 561 (TAMRA) and merged 405/561 channel. Increased stringency appeared to improve SNR significantly.

As can be seen from the different samples, having a 2-minute high-stringency wash or less did not efficiently remove much background from the slides, as indicated by the high amount of non-specifically bound probe in the nuclei of samples 1-3, **Figure 4.14**. While it is possible to detect signals for these repetitive centromeric probes under these conditions, if attempting recognition of single genes, where there is potentially a lot less probes binding to the ROI, the background may be too high if the signals are not bright enough. If probe is bound non-specifically, it may take more time in the high-stringency wash buffer (where the salt concentration is lower to destabilise the double-stranded DNA complex) for the strands to denature. With high levels of background, this reduces the confidence a clinician can have when making a diagnosis, and so to increase the reliability, a longer, higher-stringency wash should be used. Performing a 5-minute (or 10-minute) high- and low-stringency wash, sample 4 (and 5) in **Figure 4.14**, appeared to reduce background in this sample significantly, and should be considered in future. While this is slightly longer than the wash times used in the clinic, the reduction in hybridisation time means the whole process is still considerably quicker using oligoprobes, and would still significantly reduce time to result.

As all hybridisations so far were successful at only 15 minutes, various times were tested to see if this allowed more probes to bind to the ROI and therefore, a stronger signal. Although enough probes have bound to 17CEN to be detectable after 15 minutes in previous experiments, it may be that not all of the sites had actually been labelled within the ROI – as there are potentially thousands of repeats within the centromere – and so there is potential for an even brighter signal.

Five different time-points were used for hybridisation: 2 minutes, 15 minutes, 1 hour, 5 hours, overnight (16 hours). As can be seen in **Figure 4.15**, SNR appears to increase significantly up to an hour of hybridisation and then remain constant. The mean and median marks of the box plots for 1 hour and above are significantly higher than the 2-, and 5-minute hybridisations, as well as the upper-quartile range. This suggests that it takes an hour for all probes within the set to anneal to their target. Interestingly, the probes are visible after hybridisation times as short as 2 minutes, which shows that enough probes do rapidly bind to the target in this time to be detected. This suggests that a 15-minute hybridisation (or even a 2-minute hybridisation) may be sufficient if rapid enumeration is needed – which may be the case if a patient needs urgent treatment – and this could prove to be revolutionary for FISH diagnostics. For non-repetitive probes, however, a longer (1-hour hybridisation) should be considered, as this could improve SNR by ensuring that there is enough time for each individual probe to find its unique target and be detectable.

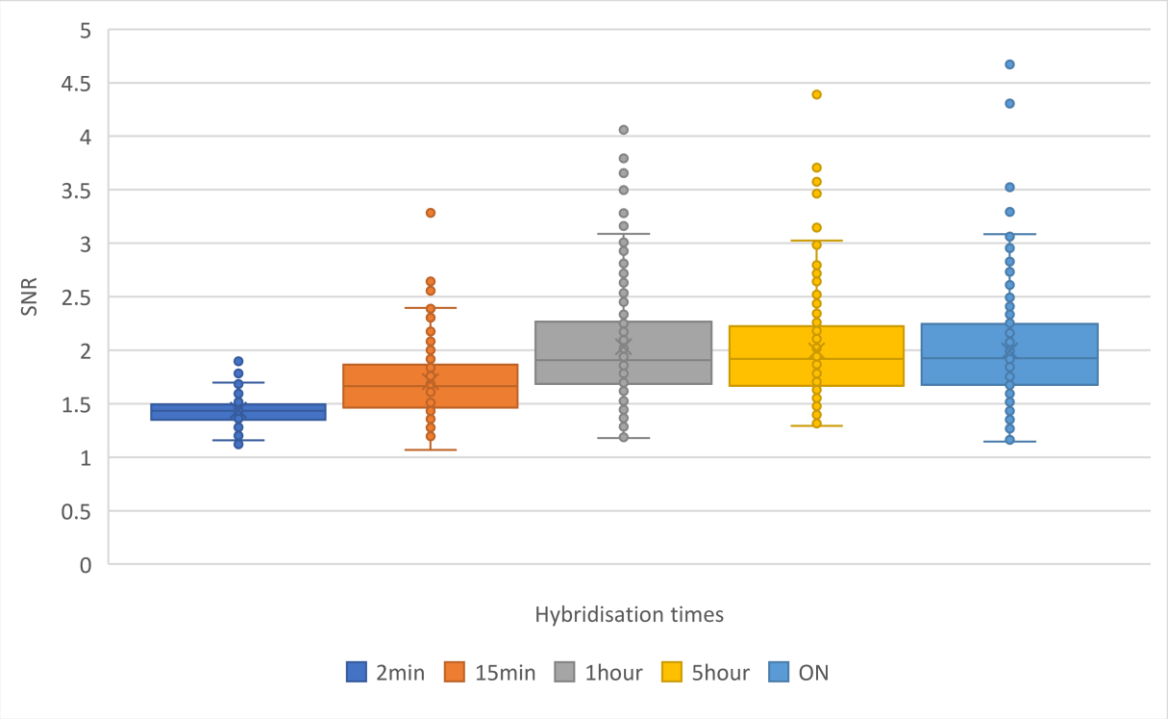


Figure 4.15: Graph showing SNR of 17CEN probes after various hybridisation times. SNR increases significantly if probes are hybridised for an hour or longer, possibly due to more oligos having time to bind to the ROI.

4.5.1 MTase-labelled oligoprobes as a diagnostic tool for ALL

Once 17CEN labelling with TAMRA DBCO was optimised, attempts moved to finding other fluorophores that could be used for MTase labelling of different oligoprobes; by creating a toolbox of optimised conditions for different coloured dyes, this could allow multiple oligoFISH probes to be hybridised simultaneously. This would mean that in a single test, multiple abnormalities could be detected at once, which would be useful in cases where the karyotype shows various mutations (such as in ALL as discussed). Dyes were chosen that were compatible with the microscope filter and lasers available, and had to be distinguishable from other dyes used, i.e. their excitation/emission spectrums did not significantly overlap.

A dye was needed from the far-red end of the spectrum, and so Alexa 647 DBCO (Abs 648 nm/Em 671 nm) was originally selected. The probe was labelled and hybridised to a 46 XX/XY sample as described in **2.2.4** and **2.2.7**, the conditions being the same as with successful detection of 17CEN-TAMRA. When visualising the sample, however, background fluorescence was high, making it impossible to reliably detect a distinguishable ROI. This could be due to the negative charge of the Alexa 647 DBCO dye, see **Figure 4.19**, preventing full hybridisation to the negatively charged DNA, which results in excess free probe, or non-

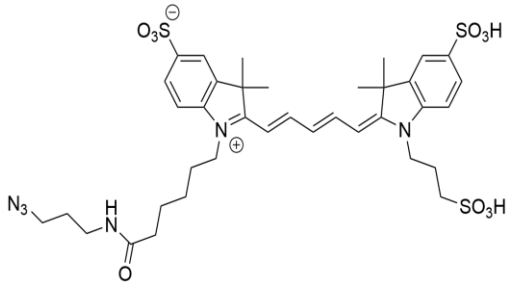


Figure 4.16: Alexa 647 DBCO structure. The fluorophore carries a negative charge which may prevent hybridisation to negatively-charged DNA, making it unsuitable for use with oligoprobes.

specifically bound probe. An optimised, higher stringency wash step could be tested to see if this removes excess dye, but this is not ideal if wanting to use in conjunction with TAMRA-labelled probes. Fluorophores will only be considered suitable if they can be detected using similar conditions to the TAMRA-labelled probes, so that they can be used simultaneously in a test.

Atto 647N NHS-ester (Abs 647 nm/Em 661 nm) had previously been successfully used for MTase labelling technology with amine cofactor AdoHcy-6-NH_2[96], so this dye was then considered. Atto 647N is a cationic dye, carrying a positive charge, and its features also show that it has excellent fluorescence quantum yield and high photostability, making it a good candidate to test. As AdoHcy-6-N_3 had been working well for labelling of oligoprobes, this was still used, with the addition of a DBCO-amine linker added into the reaction with the NHS-ester, as described in **2.2.5**.

Using Atto 647N-labelled probes was not as straightforward as using TAMRA and, while probes could be detected, it appeared to also show free dye binding to DNA non-specifically, which has been a problem reported in previous literature[156]. Even after a high stringency wash (10 mins 0.4x SSC/0.3 % IPEGAL followed by 10 mins 2x SSC/0.1 % IPEGAL), background signal in images was high (**Figure 4.20A**), possibly due to excess free dye in the probe mixture itself. A more intense purification (i.e. one with extra wash steps in the protocol) was tested in an attempt to remove the excess dye from the probe mix. Originally, mini Quick Spin Oligo (Sigma-Aldrich) sephadex columns were used for purification, designed for removing unincorporated nucleotides from a labelled oligo sample. These allow larger molecules to pass through while retaining those that are smaller (such as unlabelled

DNA or free dye). While this is a simple and quick protocol to follow – with only a two-step spin procedure – this does not include a wash step, and relies on the sorting of molecules by size to purify the sample. In contrast, Qiagen's QIAquick Nucleotide Removal Kit, contains a silica membrane which binds the oligos before a subsequent wash step using an ethanol-based buffer. This additional wash step removes excess salts and dyes. Using the QIAquick columns for purification of oligoprobes before hybridisation resulted in significant improvement in visible signal to noise ratio of samples, and probes were more easily –and reliably – detectable (**Figure 4.20B**). Using the QIAquick columns, this shows that Atto 647N may be a suitable fluorophore to use with TAMRA to form a multicolour probe mix for detection of multiple mutations in a single test.

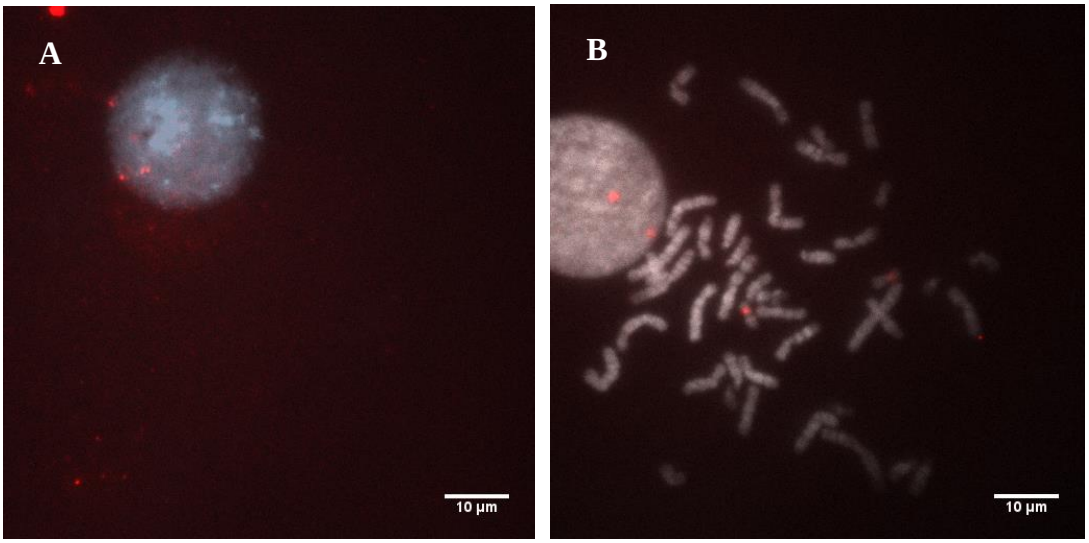


Figure 4.17: Human 46 XX/XY interphase nuclei hybridised with 17CEN and Atto 647N and purified using A) Quick Spin Oligo columns B) QIAquick columns containing and extra wash step in the protocol. Using the QIAquick columns showed improved reduction in background fluorescence, resulting in higher reliability of oligoprobe detection.

In order to produce a mixed probe containing three colours, a "green" dye (with an emission wavelength between 500–565 nm) also needed to be tested. Labelling was performed with Rhodamine Green DBCO dye (Abs 501 nm/Em 526 nm) using the standard labelling

protocol in Chapter 3 (including the use of the sephadex columns for oligo purification). Hybridisation was performed using the same optimised conditions as shown in Chapter 2. The dye was coupled to the probe and excited with a 488 nm laser. As can be seen in **Figure 4.21**, the probe successfully bound to 17CEN and was easily detectable with the standard oligoprobe conditions (as used with TAMRA in the previous chapter). This result suggests that Rhodamine Green DBCO would be a suitable choice for dye to be used in conjunction with TAMRA to provide a mixed multicolour probe. The next step is to test if all three fluorophores that have been successfully used for labelling (TAMRA, Rhodamine Green and Atto 647N) can be efficiently distinguished in a single test.

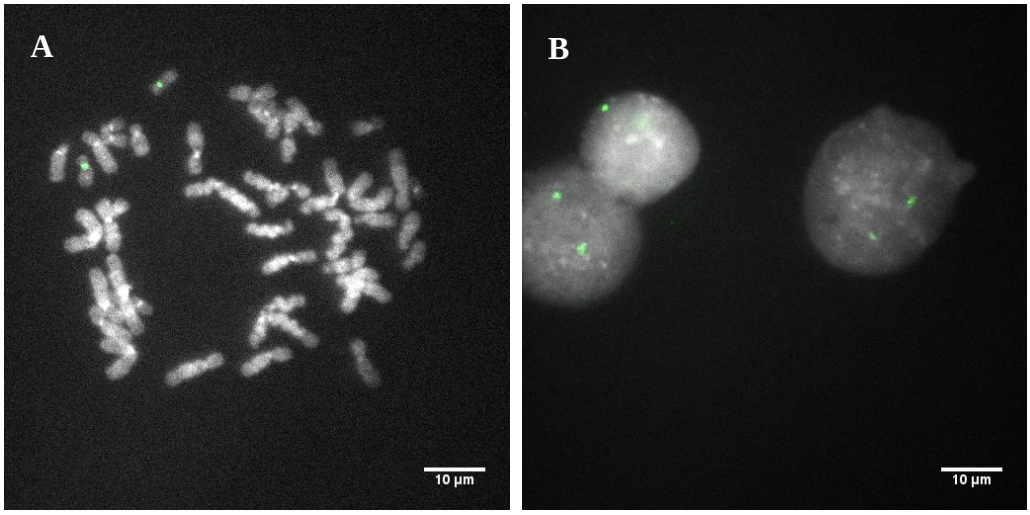


Figure 4.18: Human 46 XX/XY A) metaphase and B) interphase nuclei. Both images show 17CEN oligoprobes labelled with rhodamine green DBCO dye successfully hybridising to the target of interest (17CEN) under standard conditions optimised in Chapter 3.

As mentioned in previously, ALL is a complex disease characterised by a number of chromosomal abnormalities, one of which being loss of chromosome 1, 7 or 17. At WMRGL, they perform several rounds of FISH for a patient with suspected ALL, with one of the rounds being a simultaneous test for loss of chromosome 1, 7 and 17 using probes designed

for the centromeres of these chromosomes. With the oligoprobes hybridising efficiently in just 15 minutes, a screen for hypodiploidy of these three chromosomes could be performed in a significantly shorter space of time (currently 16 hours in the clinic). This also means that if the test came back negative (i.e. the patient does not have that mutation), another round of FISH could be quickly carried out within the same day to identify the correct mutation. Currently, a negative result would result in another overnight hybridisation being set up and could take days to determine the actual mutation that is present. Faster hybridisation times could result in rapid cancer diagnosis for a patient, which means that they could be quickly placed on the correct treatment, and hopefully have a dramatic impact on their prognosis.

The centromere for 7CEN was the next loci to be investigated for MTase-labelling. The sequence for 7CEN was taken from a paper looking at rapid chromosome enumeration[165], and checked on USCG genome browser to ensure that it mapped to the correct region. The BLAT function on the website was used for this, which displays sequences (of 25 bases or more) within the human genome that match with 95 % or greater similarity; in this way it is possible to check whether the sequence being used maps to the correct chromosome, and is unique in doing so, i.e. it will not hybridise elsewhere. Probes were ordered from IDT (the sequence can be found in **Table 2.7**), annealed as in Chapter 2, and labelled with TAMRA DBCO. A 15-minute hybridisation was performed with the 7CEN TAMRA-labelled probes and the sample analysed by excitation at 561 nm. As can be seen in **Figure 4.22A**, 7CEN was efficiently detected. Samples hybridised with both 7CEN (TAMRA) and 17CEN (Rhodamine Green) were then also tested, with **Figure 4.22B** showing that the oligoprobes were successful in highlighting these regions of interest simultaneously.

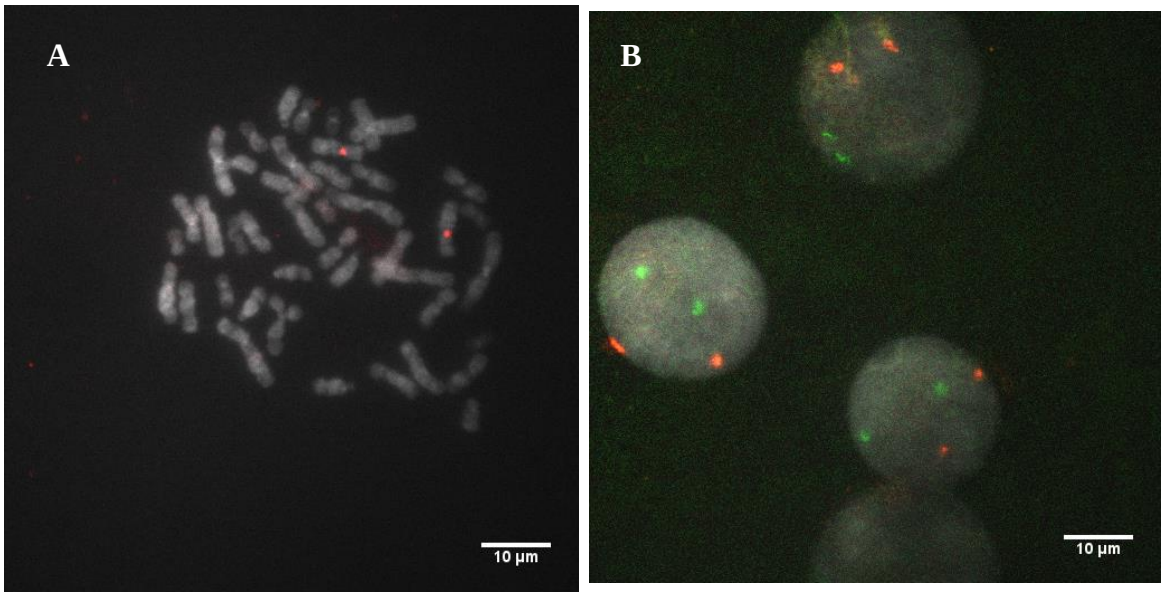


Figure 4.19: A) Human metaphase nuclei showing 7CEN (TAMRA) successfully hybridising to target. B) Human interphase nuclei showing 7CEN (TAMRA) and 17 CEN (Rhodamine Green) simultaneously highlighting the ROIs using standard oligoprobe protocol.

Due to the complexity of the alpha satellite region of chromosome 1 – this chromosome shares much of its tandemly repeated DNA with many other chromosomes[147,166,167], care had to be taken to ensure a unique region was targeted. There is still large gap in sequencing data for assembled centromere regions of the human genome due to the limitations of current sequencing techniques. This is because the tandem repeats of each chromosome are so similar, making these regions difficult to distinguish; this is something that the emergence of technologies such as nanopore sequencing are helping to tackle, and scientists are optimistic that the gaps will be filled in the near future. 1CEN is incredibly repetitive, and initial attempts to find a unique region (taken from current literature[168] and labelled, hybridised and washed as described in Chapter 2) failed, as demonstrated by the cross hybridisation present in **Figure 4.23**. This shows that, while the probe appears to have hybridised to the centromere of chromosome 1, it was not specific enough to that region, and has also hybridised to numerous other centromeres that contain the same tandem repeat. This sequence is therefore unsuitable for detection of 1CEN.

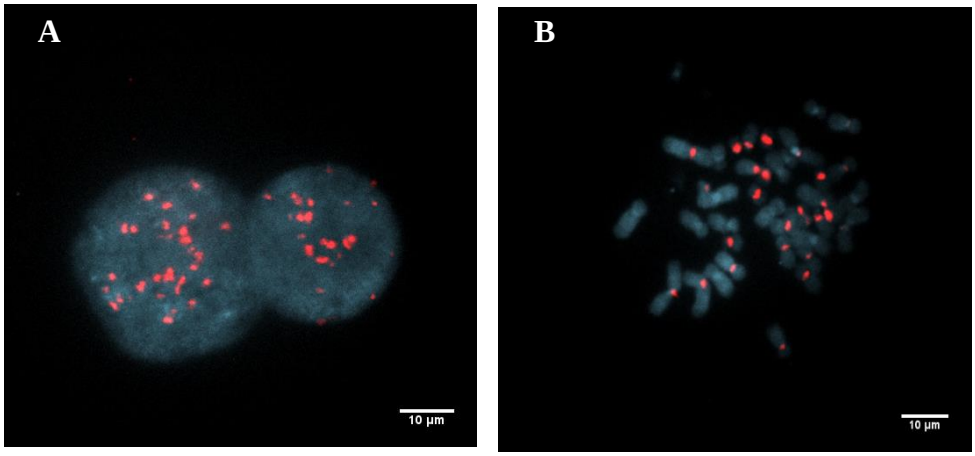


Figure 4.20: A) Interphase and B) metaphase nuclei showing cross hybridisation of 1CEN probe labelled with TAMRA DBCO. The sequence chosen for 1CEN oligoprobe was not unique to this loci, and is present within the centromere of other chromosomes.

After searching the literature, a different region of 1CEN[115] was targeted and the sequence ordered from IDT. Probes were labelled as follows: 1CEN (Atto647N), 7CEN (TAMRA), 17CEN (Rhodamine Green) using the protocol in Chapter 2, and tested with a 15-minute hybridisation. This sequence appeared to be unique to 1CEN and, as can be seen in **Figure 4.24**, the MTase-labelled probes successfully bound to their appropriate targets.

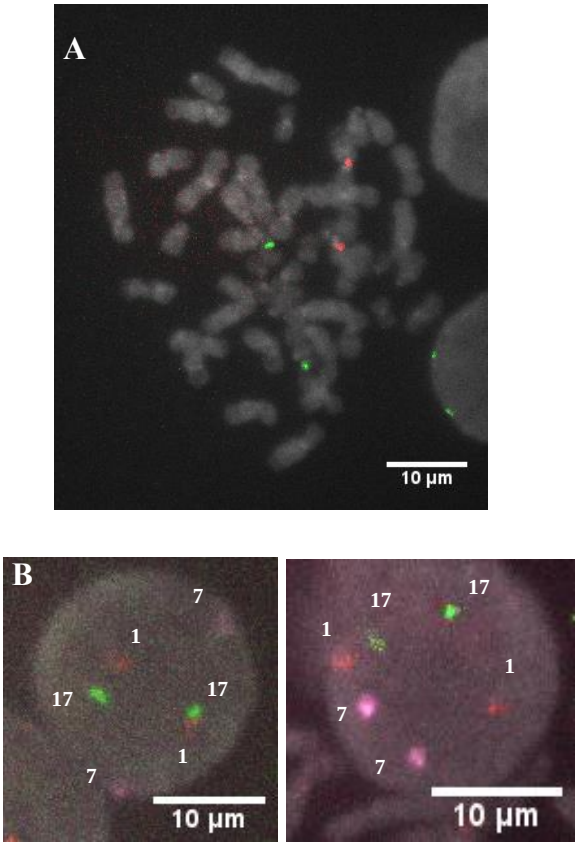


Figure 4.21: A) Metaphase nuclei showing 1CEN (TAMRA) hybridised with 17 CEN (Rhodamine Green). B) Zoomed in interphase nuclei showing 1CEN (TAMRA), 7CEN (Atto 647N) and 17CEN (Rhodamine Green) all hybridised simultaneously in 15 minutes.

Probes did not appear in every nucleus, however, and it may be that further optimisation of the conditions or sequences is needed in order to see a more homogenous result. This may prove difficult as the different oligoprobes could require slightly different hybridisation or wash conditions, and a compromise in these may affect the quality of some of the results. A

balance will need to be found in order to optimise conditions for each sequence so that can all be detected reliably. It would be useful to test the three probes with different wash and hybridisation conditions to determine which is the best compromise in order to see each one clearly, and with certainty, to avoid false negatives or positives.

This oligoFISH protocol has shown a significant reduction in the time taken to prepare the samples and get results (shown in **Figure 4.25A**), which could potentially have huge implications in terms of turnaround times for diagnostics. This is of particular interest for diseases such as certain cancers, where quicker administration of treatment could directly improve prognosis for a patient. Slide preparation beforehand from patient samples was identical to the SOP currently used at WMRGL and so would require no changes to current protocol. Denaturation of the DNA was performed manually by incubation at 72 °C for two minutes in buffer (2M NaOH/100 % EtOH) before being passed through a dehydration series (2 minutes 100 %, 85 % and 75 % MeOH); this mimics the protocol that popular commercial FISH probe manufacturer Cytocell uses. As stated previously, current probes are then hybridised to the slide for typically 16 hours (or overnight). These MTase-labelled oligoprobes have shown a significant decrease in the time taken to hybridise – with results for enumeration in as little as 2 minutes – due to their much smaller size than traditional FISH probes; this could have a huge impact on turnaround times, prompt treatment and, ultimately, a more positive prognosis for the patient. As discussed in the previous chapter, it does seem that increasing the time of the wash (5 minutes at 0.4x SSC/0.3 % IPEGAL and then 5 minutes at 2x SSC/0.1 % IPEGAL) may be useful to remove excess background and gain a clearer signal, but this is a minor time loss compared to the saving of hybridisation times (see **Figure 4.25B**), and could even be improved by further optimisation of probe design.

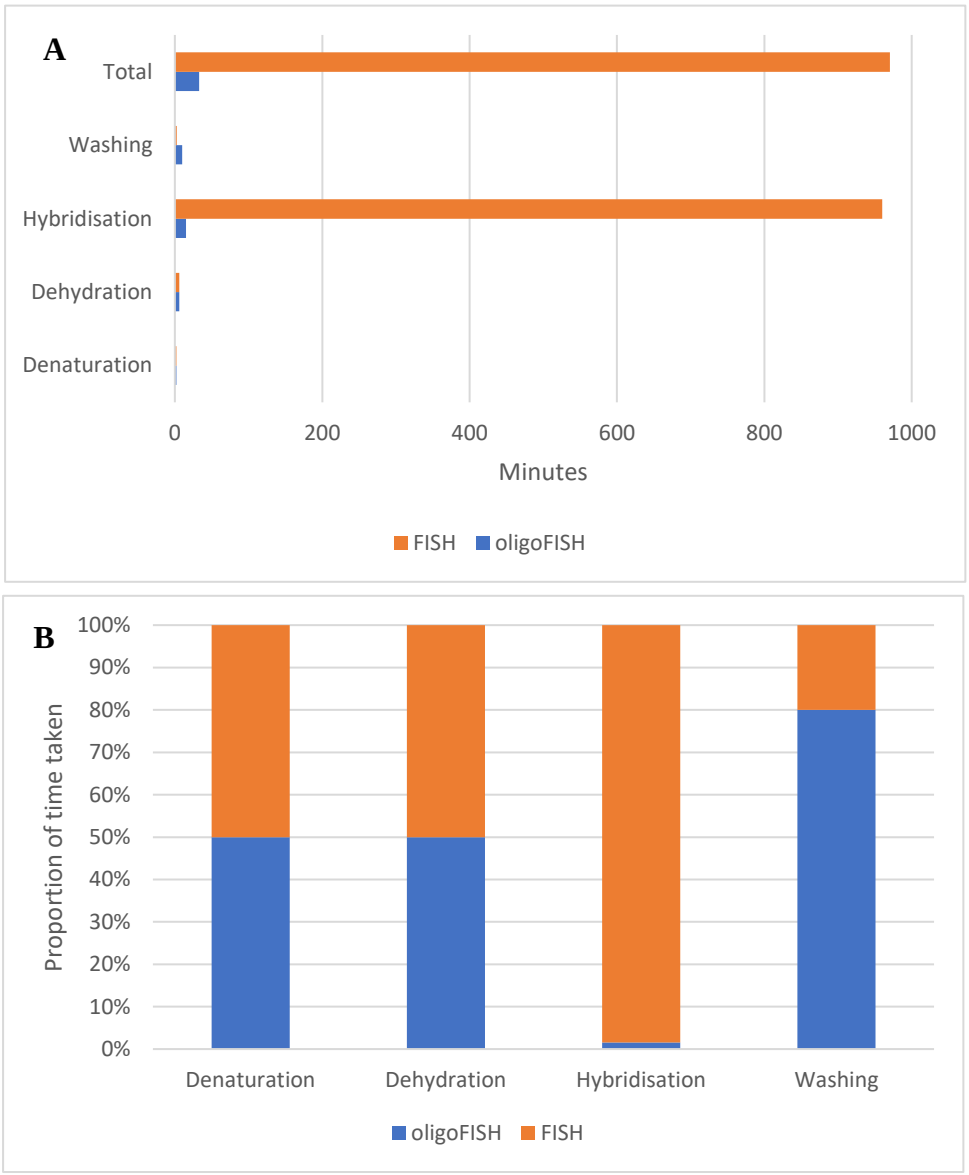


Figure 4.22: Graphs showing A) Total time taken for each step in minutes of oligoFISH and FISH protocols B) Comparison of typical time taken for each step of both oligoFISH and FISH for centromeric probes. Hybridisation is significantly quicker in oligoFISH making the total time to result much faster.

While using oligoprobes does have distinct advantages in hybridisation time and specificity, it is important to note that they appear to much more sensitive to both sequence design and wash conditions. This means that increased thought must be put into the parameters of each

probe when designing them, for instance the length, GC content (which determines the melting temperature) and specificity to the ROI. Going forward, as more sequencing information becomes available through advancements in technology such as nanopore sequencing, it may become easier to design these short oligoprobes bioinformatically and target the exact region required. This may provide valuable information to produce probes that have more favourable characteristics, making them less sensitive to the wash conditions.

Despite further optimisation being needed, this method still shows huge potential for the use of oligoprobes in FISH to detect multiple genetic abnormalities – such as those associated with ALL – in a single, rapid test.

4.5.2 Detection of small base differences

Another valuable attribute of oligoprobes, is their apparent ability to distinguish between highly homologous sequences[44,148]. This is due to their short size in comparison to the more commonly used BAC derived probes. This specificity means that despite FISH typically being a cytogenetic technique, oligoprobes could be used to detect differences in sequences down to the molecular level.

As discussed in the introduction of this chapter, the ability to achieve single base resolution combined with long range sequence information is vital for detection of SNPs in cases such as SMA. Current sequencing efforts are capable of detecting mutations down to the single base but lose sequence context in the process, resulting in loss of information as to where the SNP is on the specific chromosome – i.e. are they a 1:1 or 2:0 carrier of the SMA gene, **Figure 4.16**. Current FISH and other cytogenetic techniques can also not be used, as SMN1 bears a huge resemblance to SMN2, only differing in five positions, and current probes are not sensitive enough to detect this subtle difference in sequence.

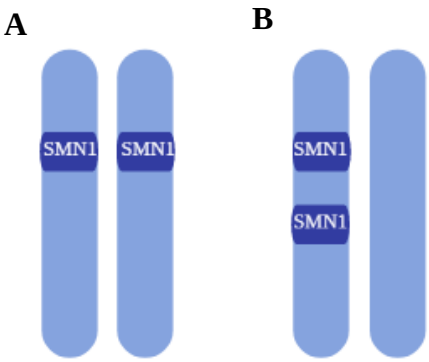


Figure 4.23: Schematic showing A) A 1:1 carrier of the SMN1 gene, with reduced risk of a child with SMA B) A 2:0 carrier of SMN1, with an increased risk of a child with SMA. 2:0 carriers are impossible to detect using current diagnostic techniques.

The region targeted by the 17CEN probes in this chapter contained two highly homologous sequences, hence two slightly different sequences (17CEN1 and 17CEN2) were used simultaneously. These sequences differ only at 4 base positions. To investigate the potential for SNP detection with the oligoprobes, experiments were carried out as a proof of concept to see if the probes could distinguish the differences between patients that had copies of 17CEN1, 17CEN2 or a combination of both. The various combinations of 17CEN1/2 that a patient can have within their chromosome pairs are shown in, **Figure 4.17**.

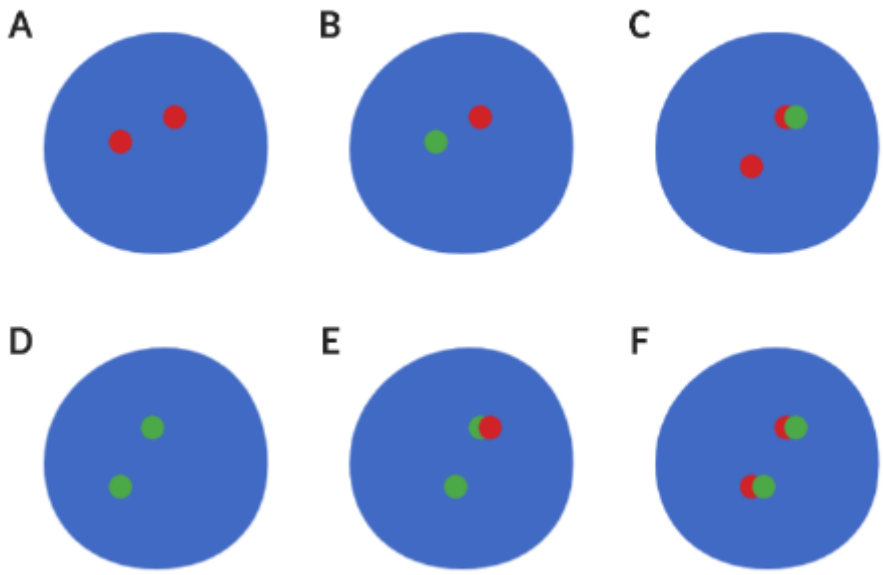


Figure 4.24: Schematic showing the variant combinations of 17CEN a patient could have across chromosome pairs, with red symbolising 17CEN1 and green 17CEN2. A) 17CEN1 only B) 17CEN1 and 17CEN2 C) 17CEN1 and 17CEN1/2 D) 17CEN2 only E) 17CEN2 and 17CEN1/2 F) 17CEN1/2.

A mixture of 17CEN1 labelled with TAMRA and17CEN2 labelled with Rhodamine Green were hybridised to patient samples using the standard oligoprobe protocol. Results seemed to be inconsistent; while there were some patients whose nuclei did appear to show both probes, the quality of the samples was poor and so it was difficult to be confident in the result. Some of the results did seem to show a difference between patients who had both 17CEN1 and 17CEN2 in in equal quantities, such as in **Figure 4.18**, which appears to show a patient with 1 copy of 17CEN1 (one red signal) and one centromere containing both 17CEN1 and 2 (a mixture of red and green signal, circled).

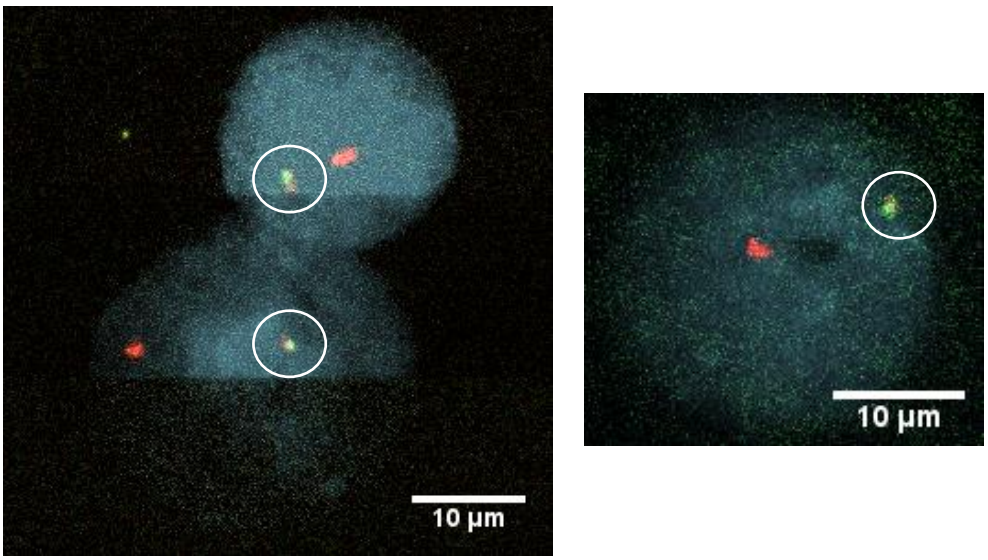


Figure 4.25: Interphase nuclei hybridised with 17CEN1 TAMRA and 17CEN2 rhodamine green. Results seems to indicate that the patient has one copy of 17CEN1 and one of 17CEN1/17CEN2, as shown by the mixed red/green signal, circled.

This result shows promise for the technique in detecting SNPs if conditions and probe design are optimise, particularly to remove excess background to get clearer and brighter signal. If wanting to use oligoprobes to detect SNPs, it is crucial that the differences between SMN1 and SMN2 can be distinguished, therefore the signal must be bright enough to detect (as there will be significantly less probes bound to the individual ROI as opposed to repetitive

centromeric probes). Probe design and hybridisation conditions will need further optimisation to explore the potential of use of oligoprobes for SNP detection.

4.6 Conclusions and future work

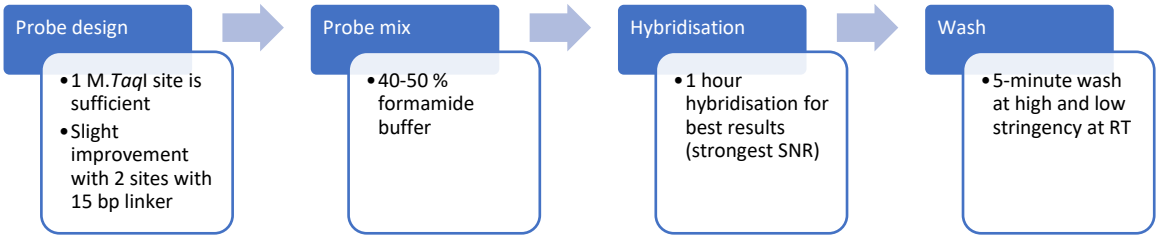


Figure 4.26: Optimised conditions for 17CEN oligoprobes to achieve the best SNR for each parameter tested either quantitatively, qualitatively or both.

In this chapter, oligoprobes have successfully highlighted the centromere of chromosome 1, 7 and 17. The MTase labelling technology tested in Chapter 2 has been used to label these short hairpin sequences using M.*Taq*I, AdoHcy-6-N_3 and three different dyes. The main outcomes of this chapter, and the optimised conditions that should be considered going forward, are summarised in **Figure 4.26**.

The main accomplishment of these oligoprobes is that they can rapidly hybridise in as little as two minutes. For many cancers and genetic diseases, prompt diagnosis is directly correlated to an improved prognosis, and so this quick turnaround of results could have huge implications on patient health. It also means that in complex cases where multiple rounds of testing need to be performed to reach a conclusive diagnosis, numerous tests can be performed in a single day – rather than having to wait for an overnight hybridisation – again speeding up the time to results and supporting timely administration of treatment. The results of this chapter do show that it is likely to take approximately an hour for the complete set of oligos to bind to the target, but in urgent cases, probes can still be detected down to as little as two-minute hybridisation. As centromeric probes have potentially thousands of repeats, they still appear to be detectable with this short hybridisation time, a longer hybridisation time

should be considered when testing non-repetitive probes to ensure that there is enough probes bound to the ROI to be detectable.

Formamide percentage also showed to have an effect on SNR, with 40 to 50 % formamide being the optimum amount to use. The addition of formamide destabilises double stranded DNA by lowering the melting temperature. This result highlights the importance and sensitivity of hybridisation and SNR, and suggests that these conditions should be optimised for each new oligo ROI to ensure that the formamide concentration does not affect the probe binding during the hybridisation step.

This chapter also discovered that washing slides for at least five minutes at both low and high stringency (at room temperature) significantly reduced background, washing away any probe that had bound non-specifically to the nuclei. This significantly increased the brightness of the probes and should be considered in future to achieve the highest SNR. Despite this washing step being slightly longer than the traditional method used by WMRGL, the significant decrease in time for hybridisation means that a result is still produced in a fraction of the time.

The concentration of probe used did not seem to make a difference to the SNR. As little probe as possible should be used going forward to reduce cost and background from non-specifically bound probe. The number of labelling sites incorporated into the design also appeared to make a limited difference to SNR, however this could be because not enough sites were added to make an impact. Adding additional sites to the probe design would

significantly increase the cost of the probes, however, and also increase hybridisation times, so is not considered necessary at this point.

The oligoprobes were reasonably successful in highlighting variant sequences for probes that differed at only 4 base positions, but this will need to be optimised in order to confirm this. Perhaps it would be wise to test a sequence that not as repetitive as centromeric loci, but contains more copies than a single-copy loci, as the next step. Probe design could also be optimised in an attempt to boost the signal of each probe, increasing the sensitivity.

In the future, it may be useful to see if these small oligoprobes have other applications in different tissues types, e.g. paraffins. As they are much smaller than traditionally used probes, it may be that they are more permeable into tissue sample and may bind more efficiently. The potential for oligoprobes to detect highly homologous sequences could also be used to study inheritance and evolution over time, by comparing homologs from parents across generations of families. This could provide valuable insight into the evolution of disease.

CHAPTER FIVE

Optimisation of oligoprobes for single genes

5.1 Introduction

MTase-labelled oligoprobes have been used in Chapter 4 for chromosome enumeration associated with acute lymphoblastic leukaemia (ALL), targeting repetitive centromeric regions of chromosomes 1, 7 and 17. This chapter uses this optimised technology to detect single genes that are linked to specific cancers, including chronic myeloid leukaemia (CML).

5.1.1 Genetic abnormalities and cancer

As discussed in the previous chapter, aneuploidy (the presence of an abnormal number of chromosomes within a cell) can be indicative of various cancers including ALL[131]. There are a number of other clinically significant mutations, many of which can be diagnosed and monitored using FISH.

Cells are constantly exposed to a variety of stresses from the environment that lead to DNA damage, which results in mutations that induce genomic instability and can lead to the development of cancer if the cells are not sent to apoptosis or senescence.[169] **Figure 5.1**.

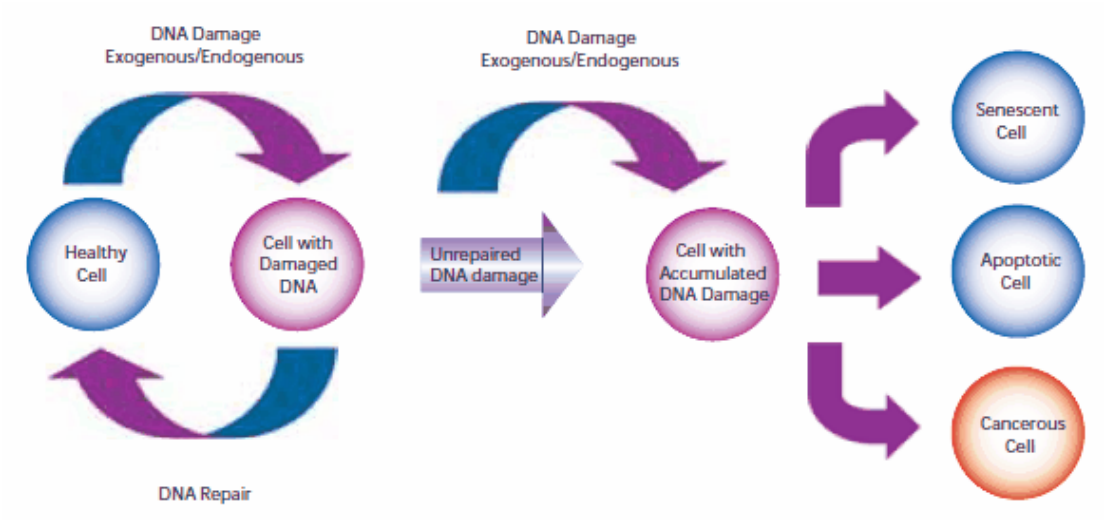


Figure 5.1: The DNA repair process involves several proteins. When the repair process fails, damage accumulates in the cell and it is either directed to apoptosis, or senescence (where the cell no longer divides but remains within the cell). Uncontrolled cell division of damaged DNA can lead to cancer. Taken from: https://www.sigmaaldrich.com/technical-documents/articles/biofiles/dna-damage-and-repair.html.

p53 is a tumour suppressor protein that is inactivated in around half of all human cancers, making it the most common genetic mutations in human cancer[170]. This protein is often referred to as the "guardian of the genome" as it is involved in cell cycle arrest, sending damaged DNA to be repaired before replication. If the DNA cannot be repaired, p53 induces apoptosis of the cell, eliminating the risk of mutation, and therefore cancer, from the cell line. p53 mutations are therefore a key target for FISH probes as mutation within this gene can lead to uncontrollable growth of cancerous cells, and is implicated in many different cancers[9,171].

Chronic myeloid leukaemia (CML), a cancer of the blood and bone marrow, was the first cancer to be associated with a clear genetic abnormality[172]. The translocation of chromosome 9 and 22 is present in 95 % of cases of CML. As a result, part of the BCR gene (chromosome 22) fuses with the ABL gene (chromosome 9), producing the BCR/ABL gene-fusion known as the Philadelphia chromosome, **Figure 5.2**. This makes it an important region to study for both efficient diagnostics and new approaches for cancer therapy, and it is usually the first mutation tested when a patient has suspected CML.

Figure 5.2: Schematic showing the BCR/ABL gene-fusion that creates the Philadelphia chromosome, commonly associated with CML.

As discussed in the introduction of this thesis, there are many different probe types depending on the mutation that is being investigated. Chapter 4 explored the use of repetitive centromeric probes, which are used for whole chromosome enumeration, a mutation that occurs in many cancers such as ALL, as well as in genetic disorders such as Down's syndrome. If wanting to look for amplification or loss of a specific gene – such as loss of p53 – which again is common in many different cancers, gene-specific probes covering the whole

ROI can be used. Another common probe that is used is for translocations, which are important biomarkers for different cancers. As stated above, a translocation commonly occurs in cases of CML where part of chromosome 9 fuses with part of chromosome 22. This mutation is investigated using a break-apart probe for the genes BCR and ABL, on chromosomes 22 and 9 respectively. Break-apart probes are designed to flank either side of the point where the genes will split during the translocation, **Figure 5.3**. In this way, either side of the ROI can be detected and monitored, and the translocation detected. If using a break-apart probe for a single gene, probes can be designed with different colours either side of the break point so that the colours will split if a translocation is present, or if part of the gene is missing.

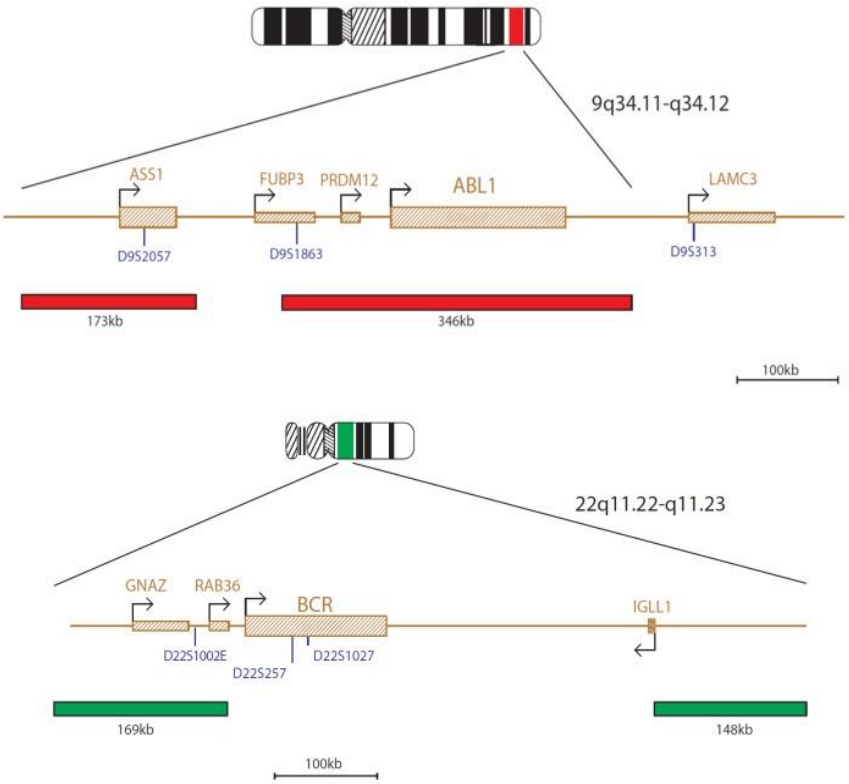


Figure 5.3: Break-apart probes are designed to flank the ROI so that certain colour patterns can be observed if there is a translocation present. Taken from www.cytocell.com/probes/14-bcrablabl1-translocation-dual-fusion.

As stated above, the BCR/ABL gene fusions occurs in ~ 95 % of all cases of CML, but it is also significantly prevalent in cases of ALL[173], and in rare occasions in acute myeloid leukaemia (AML)[174]. In both AML and ALL, this translocation is associated with an extremely poor prognosis, often not responding well to chemotherapy alone, and therefore requiring an urgent bone marrow transplant. FISH is essential in highlighting these translocations rapidly, and with oligoprobes making this process even quicker, this could be revolutionary for treatment of cancer patients.

5.1.2 Branched probes for signal amplification

If oligoprobes are to be used for SNP detection, only a small portion of DNA will be binding to the ROI, unlike potentially thousands that bind for centromeric regions. This means that the probes will need to have exceptionally high SNR, and the probe design will need to amplify the fluorophore signal so that it is bright enough to detect with certainty. Many different methods for probe amplification have been explored in the literature in order to achieve high SNR in these challenging situations, often involving the multiplexing of numerous DNA (or RNA) oligos.

Multiplexing imaging strands is a popular method, as it allows a single binding site with multiple fluorophores attached. One such method, coined clampFISH (click-activated FISH)[175], has a primary probe that binds to the sequence of interest (**Figure 5.4A**), before secondary and tertiary probes bind to the first. This pattern of multiplexing continues, with the signal of the probe effectively doubling with each round of amplification, **Figure 5.4B**[175].

This method claims to achieve both high specificity and up to 400x signal amplification, which could potentially be used to determine SNPs.

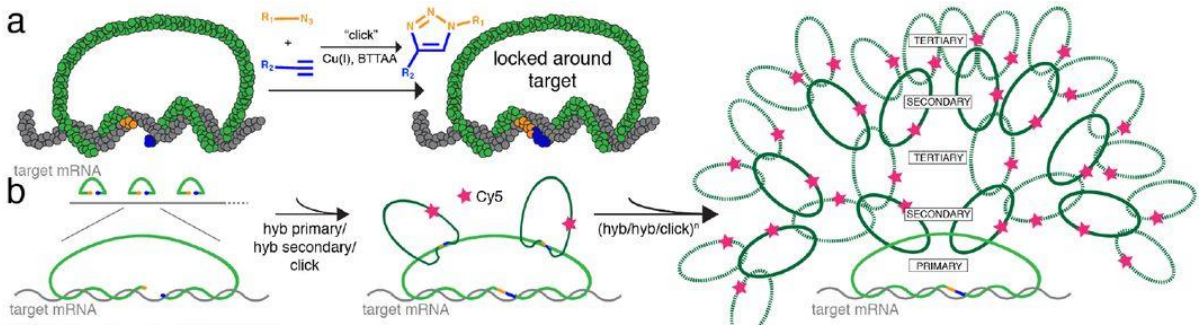


Figure 5.4: clampFISH uses a primary probe that binds to a ROI, which then undergoes multiplexing to amplify the signal. Taken from Rouhanifard *et al.* (2018).

Other methods include the targeted programmed growth of detectable concatemers *in situ* using enzymatic rolling circle amplification (RCA)[176] or hybridisation chain reaction (HCR)[177,178]. Both of these techniques result in the generation of multiple copies of DNA strands that can amplify a signal when bound to the ROI.

This assembly of dendritic "branched" DNA structures to create large DNA scaffolds that fluorescent probes can bind to has proven successful in amplifying signal for many lab groups[179,180]. Beliveau *et al.* have published several papers exploring this in a technique they have coined Oligopaint[45,181,182]. Oligopaint amplifies libraries of single-stranded fluorophore-conjugated oligos that can be used to visualise regions ranging from tens of kilobases to megabases[45], as well as small mutations and SNPs. Each oligo is designed to be complementary to a short stretch of the target genome, as well as a region that binds to a secondary oligo that further enhances fluorescent signals, **Figure 5.5**.

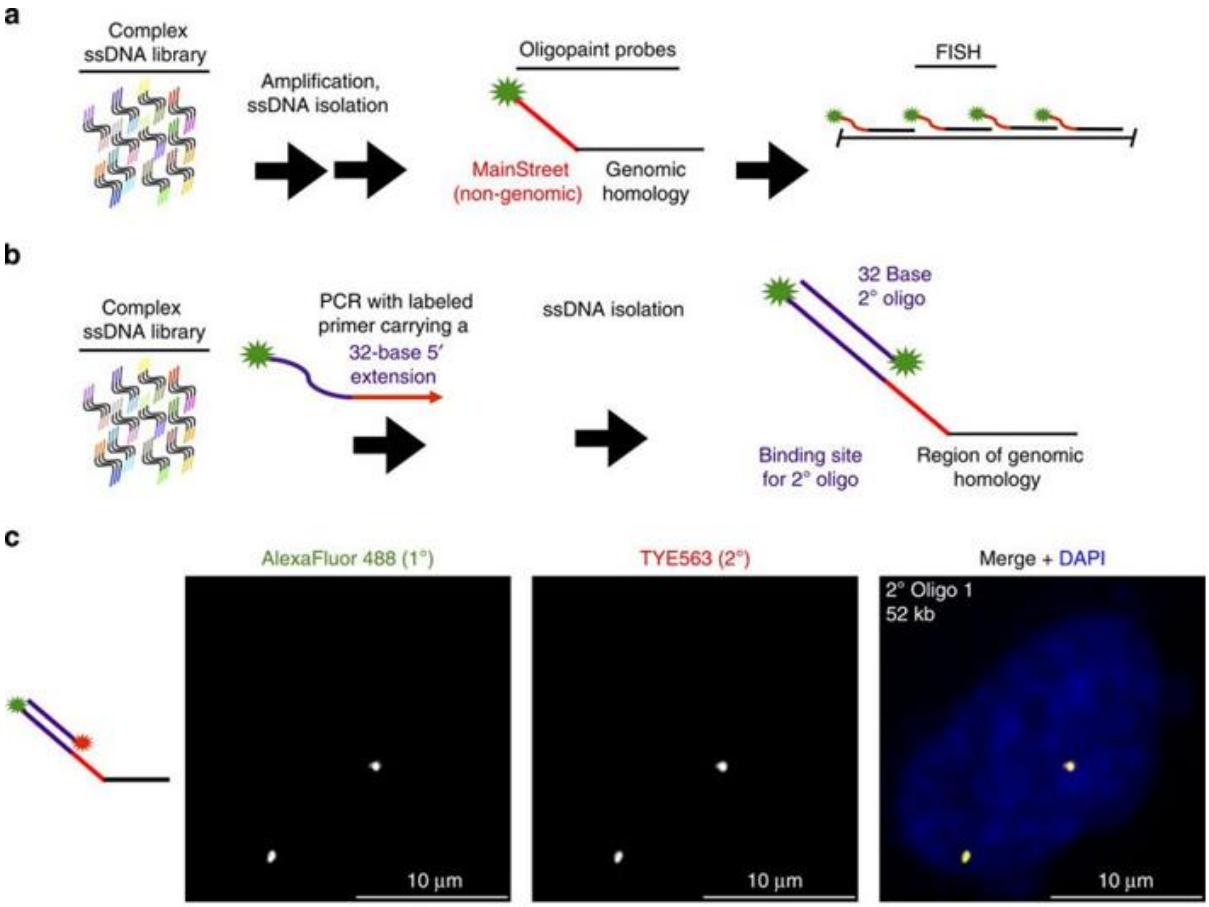


Figure 5.5: Oligopaint amplifies libraries of single-stranded fluorescently-labelled oligos that bind to a ROI in a unique way to amplify probe signal. This technique has been used to visualise regions from tens of kilobases to megabases, as well as SNPs. Taken from Beliveau *et al.* (2012).

This technique was further developed by Beliveau *et al.* into SABER (signal amplification by exchange reaction) multiplexed imaging, **Figure 5.6**[183]. This involves using the same oligo-based FISH probes with long, single-stranded DNA concatemers that acts as a scaffold to bind short complementary fluorescent imaging strands. The authors show that SABER can amplify RNA and DNA FISH signals 5- to 450-fold in fixed cells and tissues.

This technique claims to provide an inexpensive way to amplify the signal of both RNA and DNA FISH probes in fixed cells and tissues, and could again be used to amplify signals to visualise SNPs.

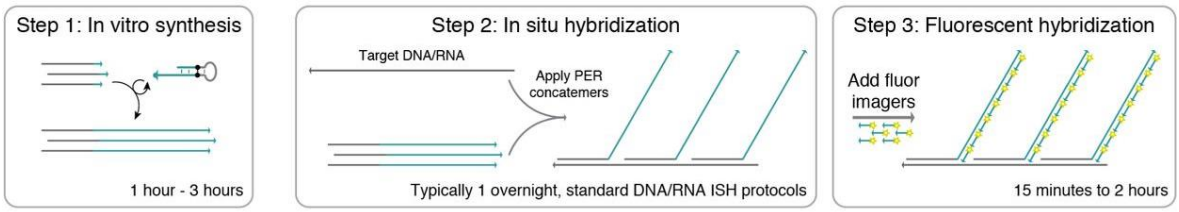


Figure 5.6: SABER amplifies FISH imaging even further using a multiplex probe approach, where multiple fluorescently-labelled oligos bind to a single oligo. Taken from Beliveau *et al.* (2019).

Despite these studies showing the promise of multiplexed FISH probes, the strategies are still fairly complicated to implement. Designing oligoprobes remains troublesome for some areas of the human genome, and a knowledge of bioinformatics is necessary to ensure that unique regions are chosen. Optimisation of the individual techniques for each ROI would also be imperative, as well as testing different sample types and regions of the genome. Probe amplification will be explored in this thesis, but alternative methods for SNP detection will also be investigated.

5.2 Aims

As the genes p53 and BCR/ABL are commonly involved in mutations resulting in cancer development, these will be areas of focus for oligoprobe production. If the same 15-minute hybridisation is achievable as in Chapter 4 with the centromeric probes, this could provide a very powerful tool for rapid diagnosis of these diseases, and therefore timely treatment for patients.

A new probe design will be tested to investigate if multiplexing may be useful for detecting small mutations such as SNPs. As proof of concept this will be tested using 17CEN, as the probes used in Chapter 4 were successful in highlighting this ROI with this design and target sequence.

5.3 Results and discussion

The overall aim of this chapter was to use the optimised MTase-labelling technology from previous chapters to develop probes for single genes.

5.3.1 Single gene detection using oligoprobes

Production of MTase-labelled FISH probes for the repetitive centromeres of chromosome 1, 7 and 17 – where only a single target was used for each – was reasonably successful. Probes were visible after just 15 minutes of hybridisation, although optimisation is needed to improve the consistency of results. The next goal was to try and test the capability of oligoprobes to detect non-repetitive, single genes. As discussed in the introduction of this chapter, gene detection is common in FISH to detect loss or amplification of a specific gene, as well as chromosomal translocations, all of which can be involved in various pathogenic pathways. Like centromeric probes, these gene probes also take 16 hours to hybridise, and so a rapid test – potentially provided by using oligoprobes – could have implications in turnaround times for patients. This is useful, as stated previously, for patients with cancers that need prompt treatment or for pre- or postnatal cases that need urgent diagnosis. Again, with a shorter hybridisation time, multiple tests could be performed consecutively in a single day rather than waiting for an overnight hybridisation if a result comes back negative.

As BAC probes do not need 100 % fidelity to bind, i.e. they do not need to match the target of interest exactly, one long piece of DNA can be used to target the ROI. However, as oligoprobes are short and highly specific, these probes need to be designed to tile across the entire gene region. In order to product oligoprobes for a gene, for example, multiple sequences will have to designed to tile across the entire ROI. This is a challenge compared to designing oligoprobes for centromeres – which contained a single sequence repeated

hundreds to thousands of times – as numerous sequences will need to be used in order to achieve a bright signal, **Figure 5.7**. Several probes will have to be designed to each cover a small and specific portion of the ROI, and they all must have similar characteristics to one another to ensure consistent hybridisation. They must all also be unique to the ROI to ensure that they do not hybridise elsewhere in the genome.

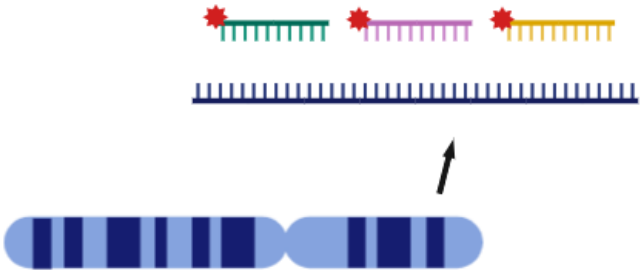


Figure 5.7: Simplified schematic showing labelled oligoprobes tiled across a ROI. Multiple probes are needed in order to produce a signal bright enough to detected.

As proof of concept, the p53 gene was investigated, as this is commonly mutated in cases of cancer[170], as discussed in the introduction of this chapter. p53 is also present on chromosome 17, and so this could allow the use of the 17CEN probes as a control, to check that it is highlighting the correct chromosome.

Oligoarray[151] – a free software that generates gene-specific oligonucleotides based on certain input parameters – was used to mine the human genome for short unique sequences from the p53 region. 20 sequences were chosen (**8.1**) that had a GC content of over 50 % and were reasonably evenly spread across the 19,149 base pairs of the gene, in order to span the whole region. These were checked using BLAST to determine sequence identity, and to ensure that

they would not hybridise elsewhere on the genome. Sequences were ordered from IDT in the same hairpin formation as the centromeric probes, labelled with TAMRA DBCO, and prepared for FISH as in **2.2.4**.

Unfortunately, despite attempting both a short (15-minute) and overnight hybridisation, the probes were not detectable, **Figure 5.8**. An overnight hybridisation was performed in case it took longer for these oligos to anneal as it is not repetitive DNA. As the hairpin design, MTase-labelling chemistry and the TAMRA DBCO dye have all been confirmed to be compatible with this protocol, this gives an idea to the areas that may need optimising. It could be that the probe sequences that were chosen were not ideal for this region; if those particular sequences are variable among individuals then it may be that the oligoprobes are too specific to use. Extra analysis would be required to ensure that the targets are not regions that have SNPs or other variants from person to person. As we know that oligoprobes are extremely sensitive to both hybridisation and wash conditions, it could be that these could be optimised for this design, however, as there is no signal detected at all, this suggests that it is more likely to be a problem with the sequence itself. It could also be that the probe density is not enough, i.e. more than 20 probes are needed in order to see a signal. This could be rectified by potentially changing the probe design to incorporate an amplified labelling stand, and will be explored later in this chapter.

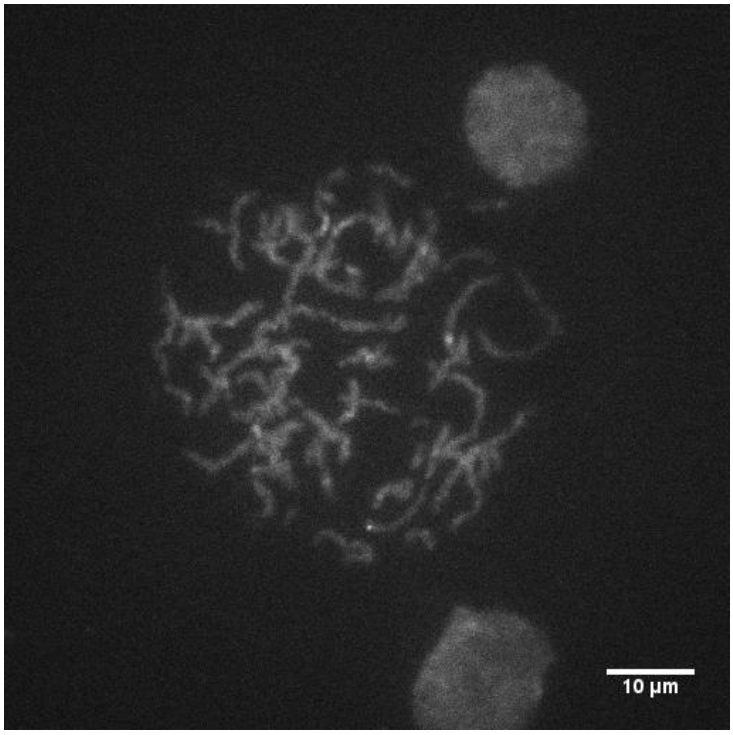


Figure 5.8: Metaphase and interphase human nuclei after p53 overnight hybridisation. No probes were detected suggesting that they need to be redesigned.

Discussions were had with WMRGL and Cytocell (Oxford Gene Technology (OGT)) on the potential of MTase-labelled oligoprobes in other FISH projects. Cytocell was interested in exploring this technology due to the fact that these oligos can be designed bioinformatically, and there is control over fluorophore position and number, which could enhance the probes' sensitivity. There is also potential for these probes to be used for rapid diagnosis in flow cytometry-based applications, coined by some as flowFISH[184,185]. It is also favourable that these oligoprobes are inexpensive to produce and, even with low concentration of probe, they were successful in rapidly detecting ROIs (for centromeres at least). Cytocell expressed interest in using oligoprobes to detect BCR/ABL translocations and offered to assist in the design of probes for the BCR regions as proof of concept.

As discussed in the introduction of this chapter, chronic myeloid leukaemia (CML), a cancer of the blood and bone marrow, was the first cancer to be associated with a clear genetic abnormality[172]. The translocation of chromosome 9 and 22 is present in 95 % of cases of CML. As a result, part of the BCR gene (chromosome 22) fuses with the ABL gene (chromosome 9), producing the BCR/ABL gene-fusion known as the Philadelphia Chromosome, **Figure 5.2**. This makes it an important region to study for both efficient diagnostics and new approaches for cancer therapy.

Currently, FISH is the gold standard for testing for this translocation, making it an interesting target for the oligoprobes and, in the clinic, probes typically hybridise in 16 hours. The BCR gene on chromosome 22 was selected as the initial ROI as proof-of-concept for this technology. As the probes would no longer be detecting repetitive DNA, each oligoprobe would need to be designed to target a unique region of the gene.

OGT kindly aided probe design for oligoprobes for the BCR gene. 89 potential ROIs were sent from OGT, selecting target regions approximately 350 bp apart, which targeted the BCR gene specifically and fit the parameters needed for the oligoprobe conditions. From the 89 sequences, 83 met the specification of being < 60 bases in length (once M.*Taq*I labelling sites had been added), ~55 % GC content, and T_m of ~70 °C. These specifications were required in order to keep cost of the oligos low, as well as ensuring that they had similar properties and would hybridise under the same conditions. This is four times more oligoprobes than were ordered when attempting to target p53 and should be detectable; the sequences were picked using the software OGT's use for their research and diagnostic work, and so were confident that they would be specific to the ROI if the hybridisation conditions can be optimised.

Probes were ordered in a 96 well plate from IDT and labelled with TAMRA DBCO using M.*Taq*I and AdoHcy-6-N_3 as described in Chapter 2, **Figure 5.9**. As results from Chapter 3 suggested that it takes an hour for all probes to anneal to the ROI, hybridisations were set up for the BCR probe for 15 minutes, one hour, and overnight. Initially, 40 % formamide was used in the hybridisation buffer for these samples, as this was seen to be the optimum stringency for the oligoprobes in Chapter 3.

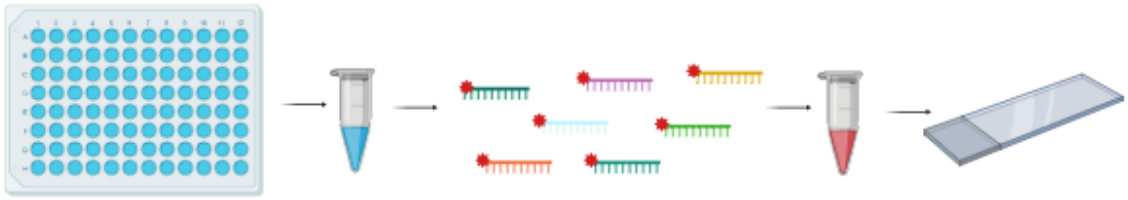


Figure 5.9: Schematic of the BCR oligoprobe workflow. Oligos were ordered in a 96-well plate, pooled and labelled with the MTase M.*Taq*I, before being mixed with hybridisation buffer and hybridised to the patient sample.

The images in **Figure 5.10** show that when using 40 % formamide, there is a high level of cross-hybridisation. A faint signal can be seen on chromosomes which look to be chromosome 22, however with such high background it is difficult to say with certainty. Two faint signals can be seen in every sample at all time-points. By increasing the stringency in these conditions, it may be possible to inhibit probes binding non-specifically and reduce the level of cross-hybridisation. It is important to note that these oligos have been designed to have a higher T_m than those used in the 17CEN experiments, and so a higher formamide percentage may be suitable.

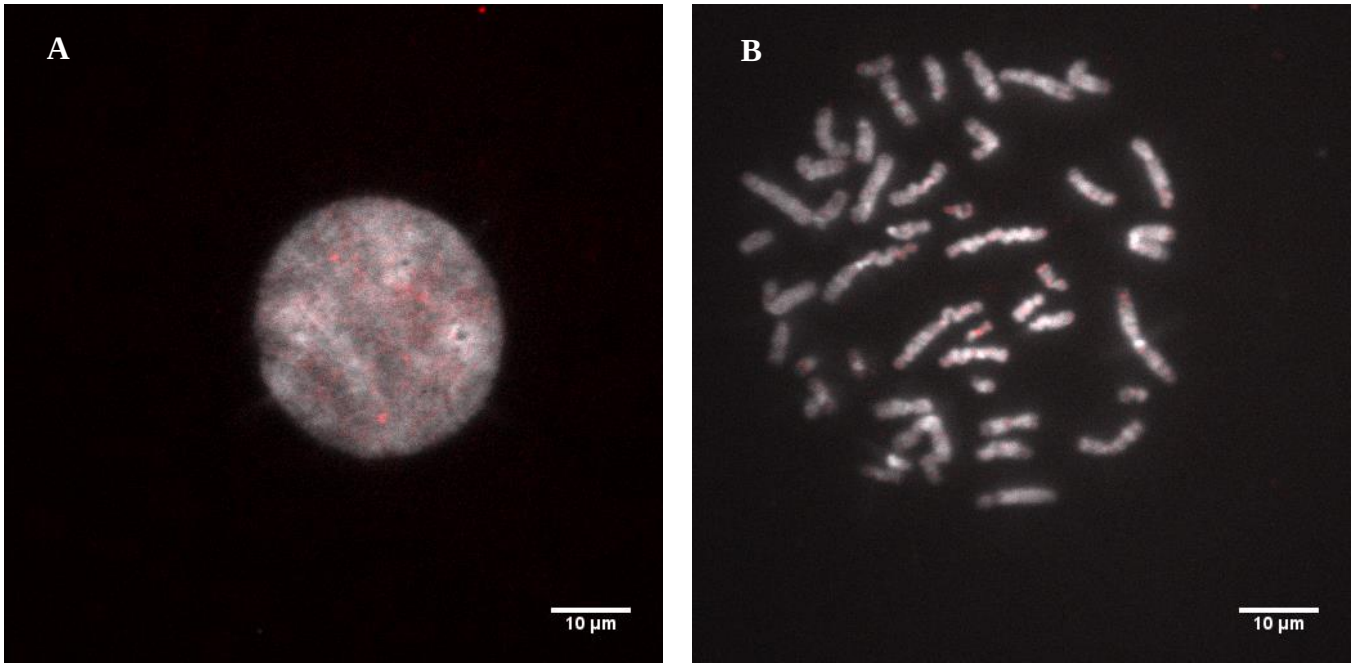


Figure 5.10: Images showing A) Interphase nucleus with BCR probes showing high levels of background B) Metaphase nucleus containing BCR probe with slight cross-hybridisation. An increase in stringency should be tested to attempt to remove non-specifically bound probes.

Samples were set up using 50 %, 60 %, and 70 % formamide hybridisation buffers, and left to hybridise for one hour. At 50 % formamide, high levels of cross-hybridisation were still observed. However, when increasing the formamide concentration to 60 % and above, clear signals can be seen for the ROI with a significant reduction in background (**Figure 5.11**). This again highlights how sensitive these oligoprobes are to the hybridisation conditions; increasing the formamide concentration reduces the amount of non-specifically bound probes to other areas of the human genome, which in turn amplifies the SNR. A careful balance needs to be found where the formamide concentration, and therefore the melting temperature of the DNA, is enough to allow all of the oligoprobes to bind to their target, while reducing the amount that seem to associate to sequences elsewhere. It could also be that as there is only a limited amount of fluorophores that could be bound to each region – because the DNA is not repetitive like centromeric sequences – the probes were not bright enough to produce a good SNR compared to the background. Optimised probe designs could be explored to amplify the signal and improve SNR by incorporating more fluorophores that bind to each ROI.

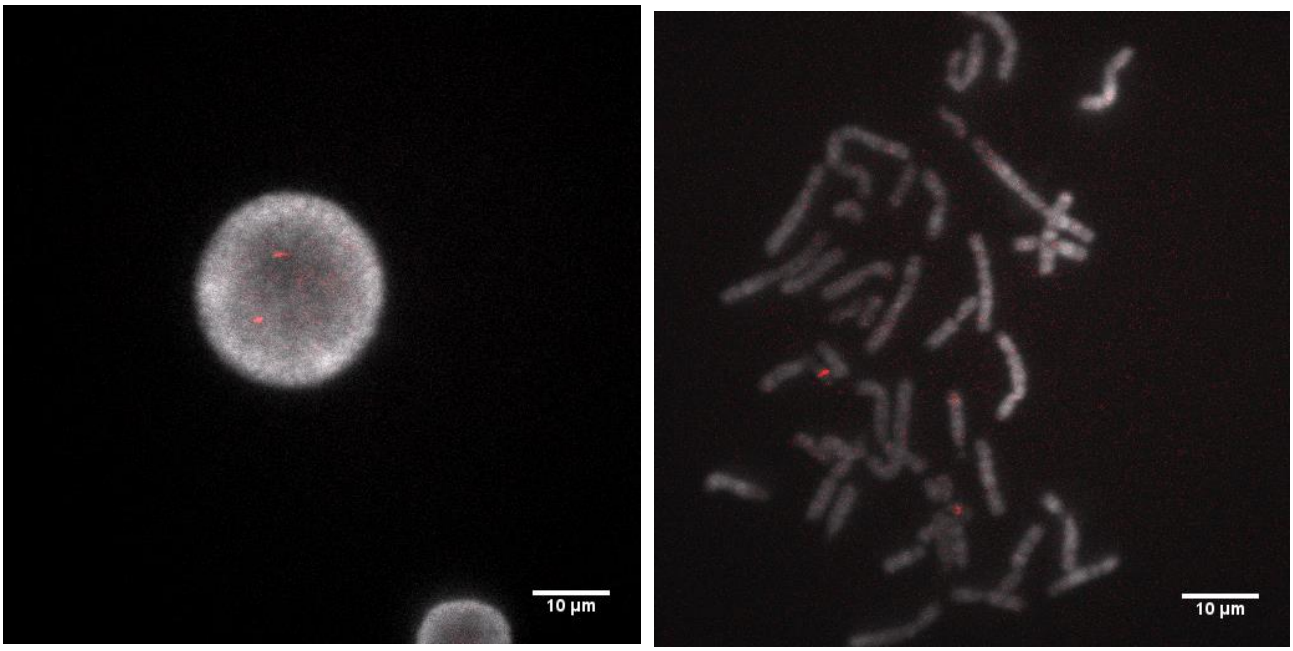


Figure 5.11: Interphase and metaphase cells showing clear signal for BCR probes using 70 % formamide hybridisation buffer and a one-hour hybridisation.

5.3.2 Exploring the potential of branched oligoprobes

The current probe design may need to be engineered in a way to efficiently detect SNPs by improving the brightness of the probe. The idea of a new probe design able to extend and add-on extra fluorescent blocks would mean that a signal could be detected for just one ROI. Oligos were synthesised using a different design built on the success of the hairpins; the idea being that a primary "docking strand" would hybridise to the ROI, and a secondary "imaging strand" to the docking strand as shown in **Figure 5.12A**. This would allow for a batch of imaging strands to be labelled, ready to be used for any new ROIs – saving valuable time – as well as the potential for various ROIs to be labelled with different MTases all in one reaction. As proof of concept, the design was based on the previously successful 17CEN sequence, and the probe was labelled and hybridised using an adapted protocol in Chapter 3. Results, **Figure 5.12B**, showed that this design could successfully hybridise to the target, showing potential for this as a way to amplify the target area for visualisation of SNPs.

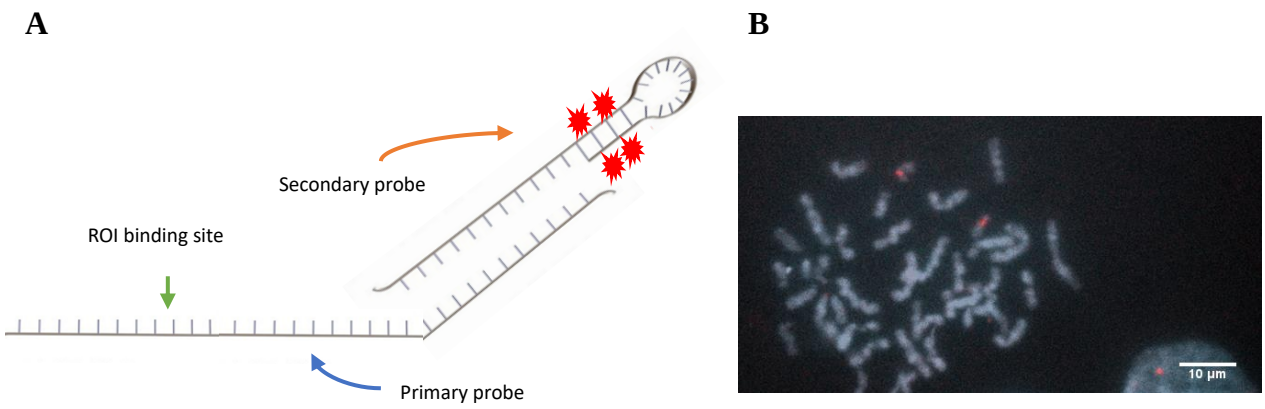


Figure 5.12: A) New probe design involving a primary 'docking strand' that binds to the ROI, and a secondary fluorescently-labelled 'imaging strand' that binds to the primary strand for detection using fluorescence microscopy B) Interphase and metaphase nuclei using new oligoprobe design labelled with TAMRA to detect the 17CEN loci.

Further work should be done to test this design – and others – to calculate the best SNR for the best chances of this being successful in SNP detection. While this design did work with

the centromeric repetitive probes – and may prove useful for future rapidly-hybridising single gene probes – the signal may still not be bright enough to detect SNPs. The probes did not appear to be significantly brighter than when using the original oligoprobes, and background in some samples was still present, which could cause problems such as recording false negative/positives when moving to SNP detection. Improved SNR from optimised hybridisation and washing conditions will need to be investigated before moving onto SNP detection for this approach. In the meantime, other options for long range sequence context combined with single molecule resolution were explored.

5.4 Conclusions and future work

Following on from the successful production of BCR oligoprobes, the next logical step would be to produce oligoprobes for the ABL gene to be able to test for the BCR/ABL translocation. Ideally these would be tested on patients with a normal karyotype initially, to test for probe efficiency, followed by testing on patients with the confirmed translocation. If these probes were successful in highlighting the translocation, they could be suitable in the clinic for diagnosing CML – and in a quicker timeframe. Care will need to be taken to ensure that the hybridisation and wash conditions are similar for both the BCR and ABL probe so that they can both be tested simultaneously.

Alternative probe designs could be explored, including a dendritic probe design, where fluorescently-labelled oligos are multiplexed to provide a bright signal. Once an optimised probe design has been achieved, further investigation into using oligoprobes to detect small variants and SNPs could be performed. In order for this to be successful, i.e. for the signal to be bright enough, a design which produces the highest SNR should be investigated.

CHAPTER SIX

Attempting SNP detection with DNA mapping

6.1 Introduction

The introduction of this thesis described some of the genetic tests that can be used to determine carriers of certain diseases, or as a diagnostic, and now, this chapter attempts to overcome some of the limitations of these techniques.

6.1.1 DNA mapping

DNA mapping is an alternative method that can be used to detect potential SNPs in cases such as SMA[93,186]. While sequencing allows detection down the single-base resolution, the sample preparation fragments the genome into smaller fractions beforehand, and sequence context is lost. This means that, while you can detect the SNPs, it is not always easy to see where this lies within the whole genome, and if there are gaps and repeat regions then this information is inaccessible. This is where DNA mapping could prove invaluable, as it allows both long range sequence information to be detected, and could potentially also allow single molecules to be uncovered. In this way, DNA mapping could bridge the gap between cytogenetic and molecular DNA technologies, enabling SNP detection while visualising a larger region of the genome.

In 2010, Neely *et al.* proposed a novel idea for mapping using DNA MTases[109]. This involved direct observation of single molecules of DNA stretched via molecular combing (discussed in 6.1.3) and using MTase enzymes to fluorescently label the DNA sequence specifically. This novel technology allows analysis of the DNA sequence without compromising the sequence's integrity, providing an ordered optical map. The resulting "fluorocode" provides a visual representation of the DNA sequence.

6.1.2 DNA extension

In order to be able to localise fluorescent tags along a DNA strand, extension and linearisation of the molecule is essential as, in solution, DNA is in a random coil conformation. This can be approached in various ways, either across a solid surface or linearised in solution. In 1998, as discussed in **1.2.5**, Fibre FISH[39] – a form of FISH that involves the stretching of chromosomes – was one of the earliest techniques to use the DNA extension approach. One of the main limitations of this technique is the ability to uniformly stretch the DNA and therefore accurately measure DNA length.

6.1.3 Molecular combing

One means to provide extension of DNA is to stretch the molecule and deposit it along a solid surface. This could be by fixing DNA to a surface via positively charged amines e.g. using polylysine[100] or (3-aminopropyl)tiethoxysilane (APTES)[99] and applying extension force. Although, as mentioned earlier when referring to Meng *et al.* and Cai *et al.* respectively, this has been shown to result in non-uniform stretching (around 85 % partial extension) rendering accurate distance measurements a challenge. A more reproducible technique is DNA molecular combing. Molecular combing was first developed in 1994 by Bensimon *et al.*[187] and later reviewed by Bensimon and Herrick in 2009[130]. This method involves the preparation of a hydrophobic surface, which the DNA sample is then deposited onto. Through hydrophobic interactions of the exposed bases at the end of the DNA with the surface, ends of the DNA bind to the surface, and the rest of the molecule is stretched out of the solution in a linear fashion, **Figure 6.2**. Tethering of the DNA ends has been found to be most successful at around a pH of 6 as, at a lower pH, the DNA molecules will adsorb

strongly – and non-specifically – to the surface, and at higher pH they will adsorb too weakly[188]. Once the droplet has been placed, the air-water contact line (meniscus) provides stretching forces from surface tension to unravel the DNA in the droplet's direction of travel.

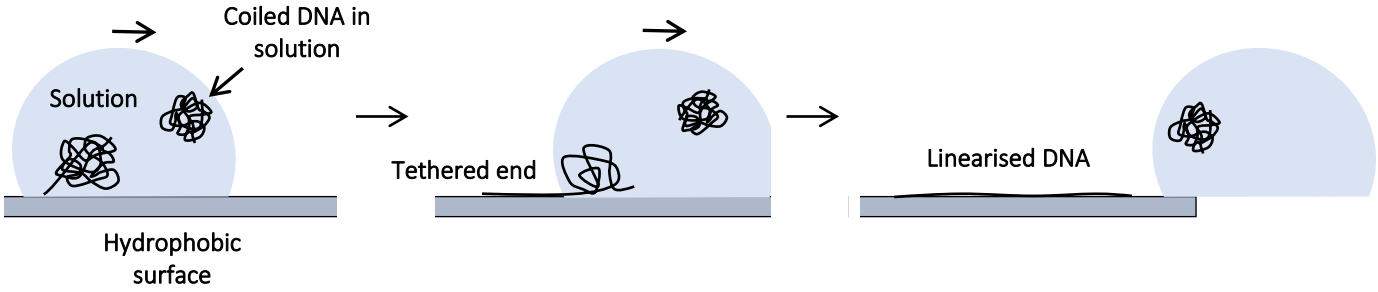


Figure 6.1: Molecular combing to produce linearised DNA. At ~ pH 6, exposed bases of DNA in solution will bind strongly and specifically to a hydrophobic surface. As the meniscus is moved, the DNA is stretched uniformly onto the surface.

Silane chemistry is often used to prepare the hydrophobic surface for combing[189], although more recently many groups have gained more success using polymer spin coating with examples such as poly(methylmethacrylate) (PMMA)[188]. In 2014, Vranken *et al.*[78] used super resolution microscopy and MTase-directed click chemistry, to provide an optical map for bacteriophage genomes. Using molecular combing onto hydrophobic slides made using alkyl-silane, the group obtained around 70 % labelling at target sites, with approximately one site every 500 bp. This approach offers potential for this technique in DNA mapping, as it bridges the gap between typical sequencing outputs and traditional long-range mapping experiments. Deen *et al.*[97] have very recently, in 2015, had great success in using the polymer Zeonex for coating the slides for deposition. When compared to coating with other polymers, this group found that Zeonex was significantly more efficient in DNA capture and uniform surface coverage, and achieved very promising results from only picograms of material, as shown in **Figure 6.2**[97]. This study has demonstrated the potential for molecular combing, without the

need for amplification of samples, and may prove highly beneficial if concentrations of DNA available are very low (i.e. picograms per microlitre).

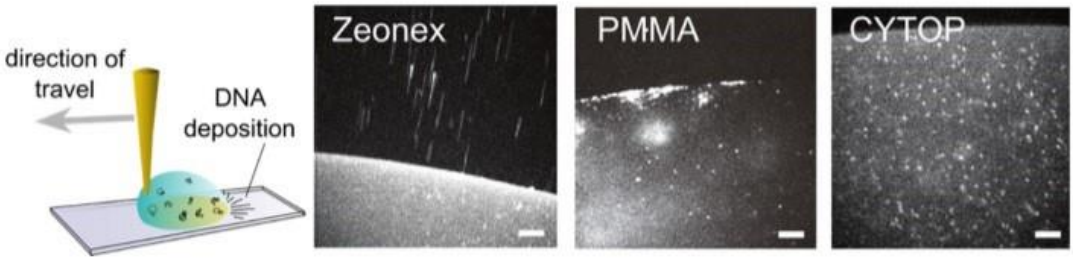


Figure 6.2: A representation of DNA combing and the receding air-water interface created as the droplet is moved in the direction of travel (left). Microscope images of deposition of DNA on three different polymer surface, of which Zeonex had the most efficient deposition (right). Taken from Deen *et al.* (2015).

6.1.4 Nanofluidic devices

Using nanofluidic devices is a popular method of extended DNA molecules without immobilisation, allowing linearisation in the solution phase[190]. Stretching of DNA is driven by the confinement in small channels with dimensions less than DNA persistence length (~50 nm)[191], which can lead to extension of the DNA to lengths of around 60-70 % of its theoretical (solution phase) length. This method is of great interest as it is a high throughput method, allowing hundreds of DNA molecules to be passed through the channels and mapped in parallel, and therefore rapidly. This is a necessity, if optical mapping is to keep up to speed with other genetic techniques such as NGS. Restriction mapping has been shown to be successful whilst using nanofluidic devices by Riehn *et al.*[190], which allows the DNA to pass through the channels while maintaining the order of the sequences. A drawback of this technique however, is that the DNA is always in motion, which negatively affects the resolution that can be acquired. An interesting direction for nanofluidics may be to combine this high throughput technique with a fluorescent labelling technique such as with MTases or nicking enzymes[108]. Alternatively, another group have also shown that mapping via

nanofluidics is also possible from denaturing DNA using heat and formamide[192], and visualising the sequence-specific melting using YOYO-1.

6.1.5 MTases and DNA mapping

Since the emergence of optical mapping via restriction enzymes, nicking enzymes, and more recently MTases, have been explored[26,93,94,96]. MTases show great potential as a way to label DNA with both a high level of specificity and at a high density. This provides potential for MTases in a number of diagnostic applications which will be discussed later in this thesis. Labelling DNA sequence-specifically using MTases and stretching single DNA molecules onto a surface via combing, can provide an ordered optical map. This novel technology allows analysis of the DNA sequence without compromising the sequence's integrity and can provide a scaffold to aid genome assembly in conjunction with sequencing.

By combining both high- and low-density labelling by producing enzymes that recognise different length recognition sequences, it will be possible to produce a dual colour map, which will be highly useful in diagnosing genetic disorders. Enzymes with recognition sequences between 4-8 bps will be produced and screened with cofactors to produce an MTase-labelling toolbox.

Advances in DNA hybridisation techniques (e.g. FISH) and sequencing technologies (e.g. next generation sequencing (NGS)) have surpassed the use of restriction mapping in rapid DNA identification. Though they are more commonly used, both hybridisation and sequencing techniques have their own set of problems. NGS is currently at the forefront of sequencing technologies – although long-read sequencing is rapidly developing – however it

still faces issues with copy number variations (CNV), ensemble averaging, and reconstruction of the genome after amplification of sequences. As this technique focuses on small base differences, it loses any larger structural information. On the-other-hand, cytogenetic techniques such as FISH, focus on much larger regions of interest (ROI). FISH is currently the gold standard in detecting large rearrangements, amplifications, or deletions of genetic material but it is not possible to detect changes at the single-base level. Optical mapping attempts to overcome some of these challenges.

6.1.6 MTases and SNP detection

Some MTases display highly specific recognition of their target motifs. Therefore, it has been hypothesised that they could be used in detection of single nucleotide polymorphisms (SNPs). SNPs are variations of single nucleotides that occur at a specific position in a DNA sequence. This genetic variation can be the underlying cause for susceptibility to certain diseases e.g. cystic fibrosis, and also impact the severity of those illnesses[23].

Spinal muscular atrophy (SMA) is a recessive neurodegenerative disease characterised by the loss of the SMN1 gene[119]. A nearly identical gene, SMN2, has only one critical nucleotide difference. SMN2 can be present in variable numbers in patients and therefore restores some of the functionality lost from the SMN1 mutation, resulting in varying levels of severity of the disease[112]. It is possible to be a carrier of SMA if you only have one copy of SMN1, or if you have two copies of SMN1 on one chromosome; a 2:0 “silent” carrier[120], **Figure 6.3.**

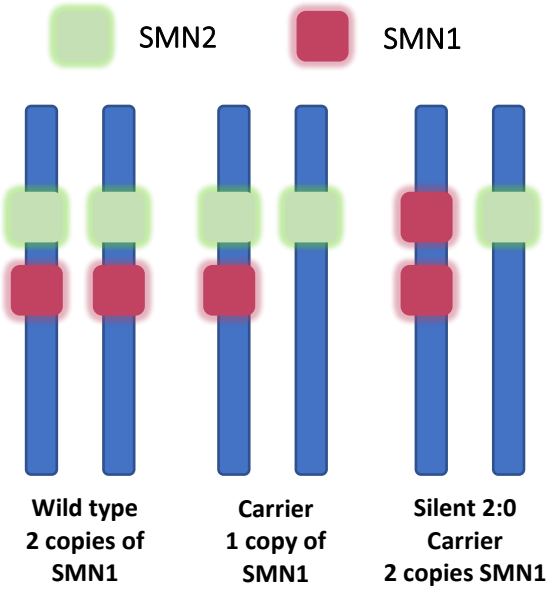


Figure 6.3: Schematic showing copy number and position of SMN1 and SMN2 in different patients. It can be difficult to determine carriers of SMA due to silent carriers with two copies of SMN1 on one chromosome (2:0 carriers). This makes it problematic to distinguish this from the wild type using molecular techniques.

Currently it is not possible to determine with 100 % certainty whether someone is a silent carrier. Molecular techniques, such as multiplex ligation-dependent probe amplification (MLPA) – a variation of multiplex PCR that amplifies multiple targets using with a single primer pair – can detect two copies of SMN1, but cannot determine if they are in the 2:0 formation or 1:1. Cytogenetic techniques, such as FISH, can also not be used for this arrangement, as they do not have the specificity to work at the single-base level.

M.*Hpy*188I is an MTase which targets TCNGA[113]. This sequence is disrupted by a single base change difference in the sequence of SMN1, but not in SMN2. If a patient's DNA could be labelled with M.*Hpy*188I and mapped, it could be possible to determine whether a patient's SMN1 genes are in the 1:1 or 2:0 formation based on the pattern produced from the

fluorophores. In this way it could be possible to detect silent carriers of SMA by locating the exact position of these genes within their genome.

6.2 Aims

As discussed in previous chapter, M.*Hpy*188I, is an MTase that targets TCNGA. This sequence is disrupted in the sequence of SMN1, but not in SMN2. If a patient's DNA could be labelled with M.*Hpy*188I and mapped, it could be possible to determine whether a patient's SMN1 genes are in the 1:1 or 2:0 formation based on the pattern produced from the fluorophores. In this way it could be possible to detect silent carriers of SMA. As proof of concept, this chapter will explore the use of M.*Bse*CI for DNA methylation, to observe whether this will block M.*Taq*I labelling at overlapping sites.

6.3 Results and discussion

An alternative method to SNP detection that was also considered is DNA mapping. As described in **1.5.3**, this technique could be used in conjunction with sequencing to provide valuable information on sequence context, while also detecting small rearrangements or differences to the reference genome; a severe limitation of current techniques. Physical maps display both long- and short-range sequence information. This could prove invaluable in cases of SMA to detect the location of SMN1 genes on potential parents to determine whether they have a 1:1 or 2:0 genotype, and could potentially distinguish between highly homologous SMN1 and SMN2. In this way, carrier detection will be improved significantly, and appropriate measures can be put in place if a couple is planning to have a child, and can be used to calculate risk of SMA development. SMN1 and SMN2 differ at one critical nucleotide position on exon 7. This difference in sequence disrupts the recognition sequence of MTase M.*Hpy*188I, resulting in a loss of fluorophore if attempting to map the region. This means that DNA mapping using this MTase could potentially identify the presence of the SMN1 gene and therefore if they are a silent 2:0 carrier.

Mapping involves labelling the DNA – in the case of this thesis, with MTases – and then stretching single DNA molecules along a hydrophobic surface. The sample is visualised using fluorescence microscopy and the pattern of fluorophores analysed to determine the DNA's underlying sequence.

6.3.1 Blocking alkylation with methlyation

DNA combing was carried out following the protocol described by Deen *et al*[97]. Optimum combing, which has been investigated in great detail by other researchers[26,78,93,96], resulted in uniformly-stretched individual DNA molecules of around 1.52 times the crystallographic length. As a proof-of-concept, lambda DNA was methylated with M.*BseC*I DNA, followed by labelling with Atto 647N using M.*Taq*I and AdoHcy-6-N_3 as described in Chapter 2. As the M.*BseC*I recognition site (ATCGAT) overlaps with M.*Taq*I's (TCGA), we would expect some of the TCGA sites to be blocked, and therefore a loss of fluorophore. On lambda, 15 out of 121 M.*Taq*I sites should be blocked by M.*BseC*I methylation. If it is possible to detect single loss of a fluorophore from a reference sequence, such as with M.*BseC*I methylation and M.*Taq*I labelling, this provides hope that this method may be suitable for discrimination of SMN1 from SMN2.

A protection assay was carried out on M.*Bse*CI-methylated, M.*Bse*CI- and M.*Taq*I-methylated and unmethylated lambda to confirm whether the 15 'blocked' TaqI sites could be detected at this level. This would be determined by a subtle change in pattern of the gel due to different DNA fragments being produced by restriction enzymes. Gel electrophoresis does not have the necessary resolution to determine exact DNA differences down to single base

resolution, but it may indicate whether M.*Bse*CI methylation will block M.*Taq*I labelling for future mapping experiments.

Lambda was methylated with either M.*Bse*CI or both M.*Bse*CI and M.*Taq*I, and restriction was attempted with R.*Cla*I (which has the same recognition sequence as M.*Bse*CI (AT^CGAT)) and R.*Taq*I (T^CGA). Results will determine whether M.*Bse*CI methylation will block R.*Taq*I restriction, and whether this can be detected in the gel.

As can be seen in lane 2 of **Figure 6.4**, M.*Bse*CI efficiently methylates lambda DNA, preventing it from restriction by its corresponding restriction enzyme R.*Cla*I. As the recognition site of M.*Bse*CI overlaps with M.*Taq*I, methylation with M.*Bse*CI should block 15 out of the 121 M.*Taq*I sites present on lambda. Lanes 3 and 9 allow comparison of M.*Bse*CI methylated and unmethylated DNA cut with R.*Taq*I. These results show the very subtle difference in pattern of restricted lambda as expected by the blocked cutting sites.

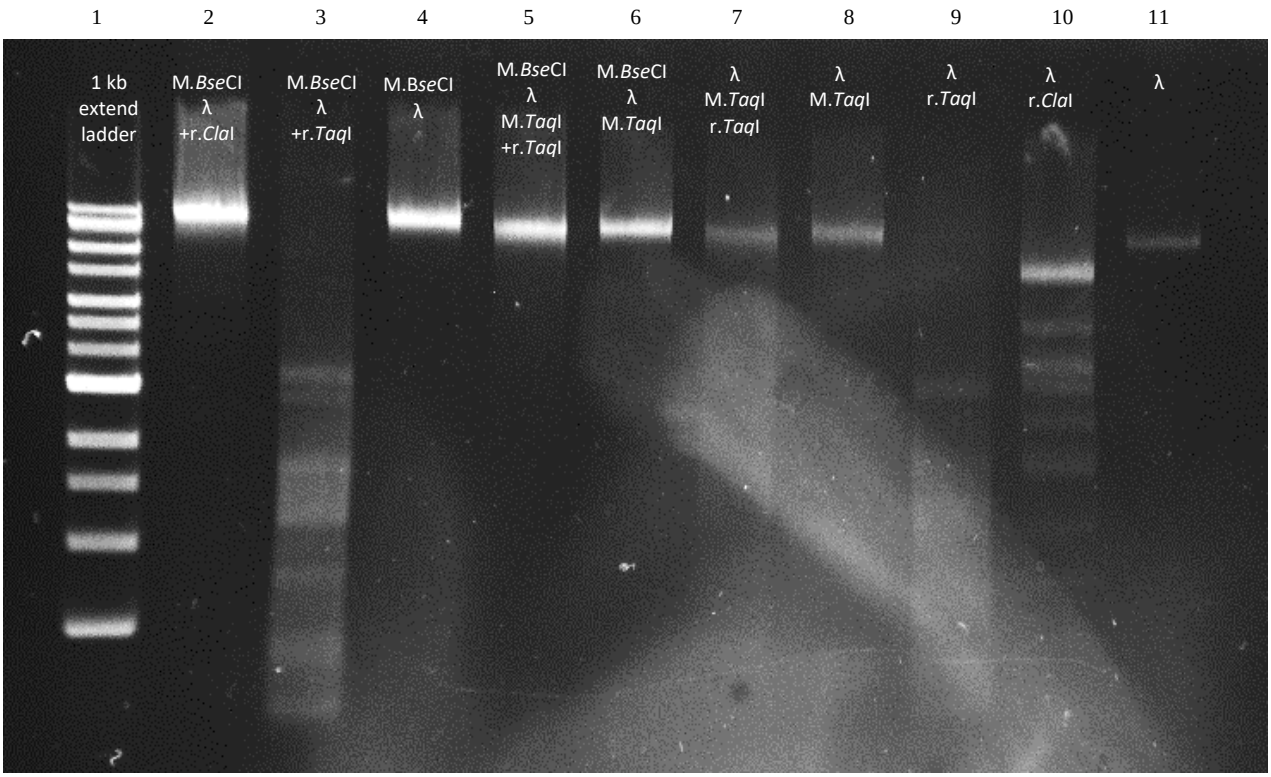


Figure 6.4: Protection assay of M.*Bse*CI methylated lambda.
Lane 1 = 1 kb extend ladder, lane 2 = M.*Bse*CI methylated lambda + R.*Cla*I restriction, lane 3 =M.*Bse*CI methylated lambda + R.*Taq*I restriction, lanes 4 and 5 = M.*Bse*CI + M.*Taq*I methylated lambda + and – R.*Taq*I restriction, lanes 6 and 7 = M.*Taq*I methylated lambda + and – R.*Taq*I restriction, lane 8 = unmethylated lambda + R.*Taq*I restriction, lane 9 = unmethylated lambda + R.*Cla*I restriction, lane 10 = unmethylated lambda.

M.*Bse*CI has blocked restriction of some M.*Taq*I sites, as highlighted by the change in restriction pattern between lane 3 and 9.

The blocked pattern in lane 3 demonstrates the potential of using mapping to uncover further information on the slightly altered sequences, at a higher resolution. While the gel gives an indication that there is a difference in the sequences, it does not provide detail down the single base – and would therefore not be able to detect SNPs – and does also not provide long range information such as sequence context. This does show that mapping could be a feasible option for distinguishing the difference between two slightly different patterns (differing at

15 fluorophore positions). As the protection assay showed efficient blocking of M.*Taq*I alkylation by methylation with M.*Bse*CI, the sample was used for DNA mapping.

6.3.2 DNA mapping with MTases

Both unmethylated and M.*Bse*CI-methylated lambda were labelled with Atto647N using M.*Taq*I. The sample were deposited on a Zeonex surface as described in **2.3.5** – which involved dragging the fluorescently-labelled DNA droplet across a hydrophobic coverslip – and visualised via fluorescence microscopy. The challenge was to discover whether it is possible to determine if the sample was unmethylated or methylated, therefore efficiently detecting the loss of 15 out of the 121 M.*Taq*I labels. The schematic in **Figure 6.5** displays a visual representation of how M.*Bse*CI methylation can block M.*Taq*I labelling at those overlapping sites. This difference in pattern of fluorescence (i.e. the loss of fluorophore at

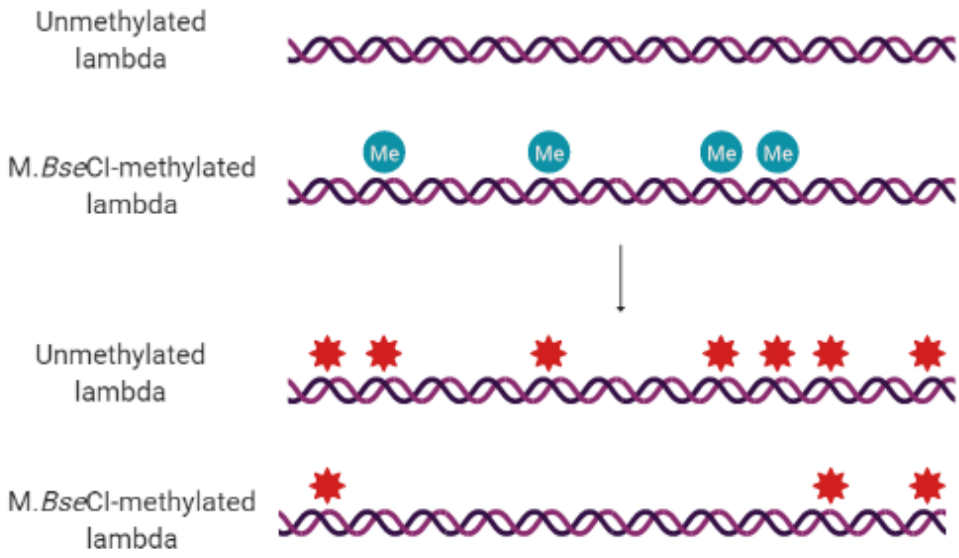


Figure 6.5: Schematic illustrating how M.*Bse*CI methylation of lambda DNA can block M.*Taq*I fluorescent labelling at those sites due to an overlap of recognition sequence. This could serve as a point of concept for detecting SNPs in the human genome.

those sites) should be detected using the DNA mapping approach, showing potential for distinguishing highly homologous sequences.

A grid of images was taken for each sample and analysed using a MATLAB script written by Nathaniel Wand and Darren Smith. The script extracts each linear barcode from the stacked image before cleaning them (e.g. removing any barcodes that are obviously the incorrect size/length or intensity). 'Junk' barcodes are removed at this point, which includes those that have occurred due to poor combing technique – resulting in overlapping barcode artefacts – and contaminated DNA during the labelling preparation. Poor combing often resulted from using either too high a concentration of DNA, and so the sample was too dense, or too little DNA, leaving the sample too sparse. Care also had to be taken to ensure that the equipment was flat when running so that the DNA travelled in a straight line, and that the speed was consistent to stop the pipette tip from 'jolting' out of place when depositing the DNA. For simple alignment, the barcodes are then each aligned to the known reference sequence (i.e. unmethylated lambda or M.*Bse*CI-methylated (blocked) lambda). First, M.*Bse*CI-blocked lambda was analysed and compared to both a blocked and unblocked reference genome to see if the blocked sites could be detected by the subtle change in intensity.

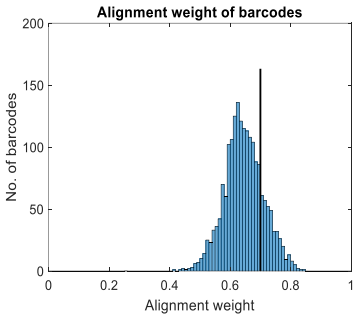
Barcodes were extracted from the sample and an alignment weight calculated for each compared to the reference, **Figure 6.6A/E.** Barcodes with an alignment weight of over 0.7 were considered to be a "good" fit, based on *in silico* data produced by Nathaniel Wand (and documented in his thesis), and from these a consensus barcode was formed, **Figure 6.6B/F**. As can be seen from **Figure 6.6C/G**, for both samples the middle of the genome produced the most contributing barcodes, this is probably due to this being the most well mapped area

during deposition, as the ends of the images tended to be either too sparse, or too dense. **Figure 6.6D/H** show the mean intensity of well-aligned barcodes after background subtraction using a rolling ball average; the mean alignment displays any discrepancies between the experimental consensus barcode and the reference. The experimental barcodes fit both references fairly well – which is to be expected as they differ at only 15 sites – but it seems that at around 15 kbp in particular, the intensity profile is more suited to the blocked reference. When compared to the unblocked reference, the first of the two peaks of the 15 kbp region is shorter than the second peak, suggesting that this is a blocked site.

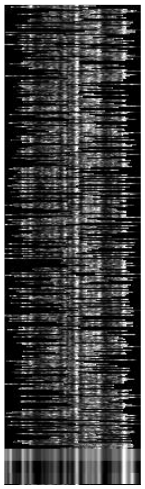
M.*Bse*CI-blocked sample

Blocked lambda reference

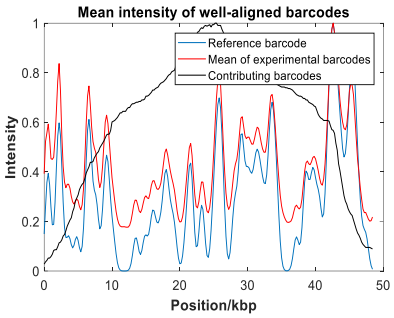
A



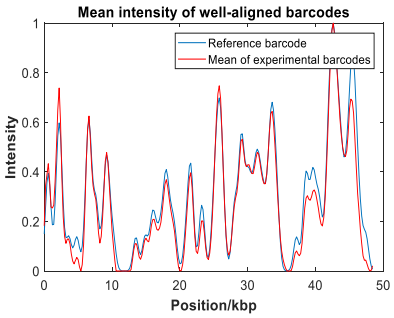
B



C

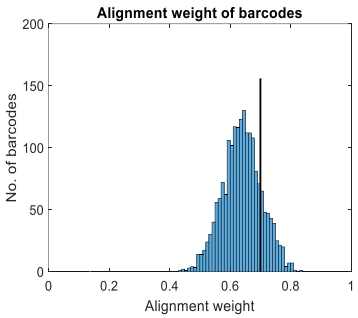


D

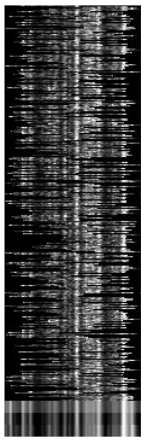


Unmethylated lambda reference

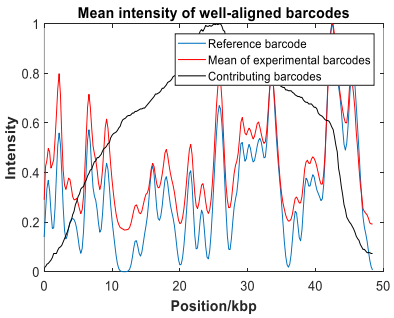
E



F



G



H

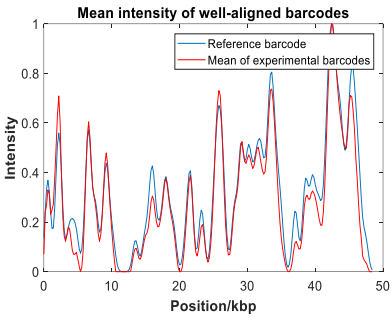


Figure 6.6: M.*Bse*CI "blocked" lambda DNA was labelled with Atto 647N at M.*Taq*I sites after M.*Bse*CI methylation. Barcodes were extracted and aligned to a blocked (A) and unblocked (E) lambda reference. Those that had an alignment weight of 0.7 or higher were combined to form a reference barcode (B/F). The mean intensity of well-aligned barcodes (0.7 or higher) was compared to the reference barcode produced to see how many barcodes fit to each region (C/G) and how well they fit each genome (D/H). The blocked lambda sample appeared to fit slightly better to blocked lambda reference.

To confirm whether this was a position where lambda was blocked, the M.*Bse*CI-methylated sample was compared to the two references (unmethylated and blocked) on a single plot, **Figure 6.7**. The dotted grey line displays the positions of the blocked sites on the genome, which explains the drop in intensity at these points, most noticeably at 15 kbp as mentioned above.

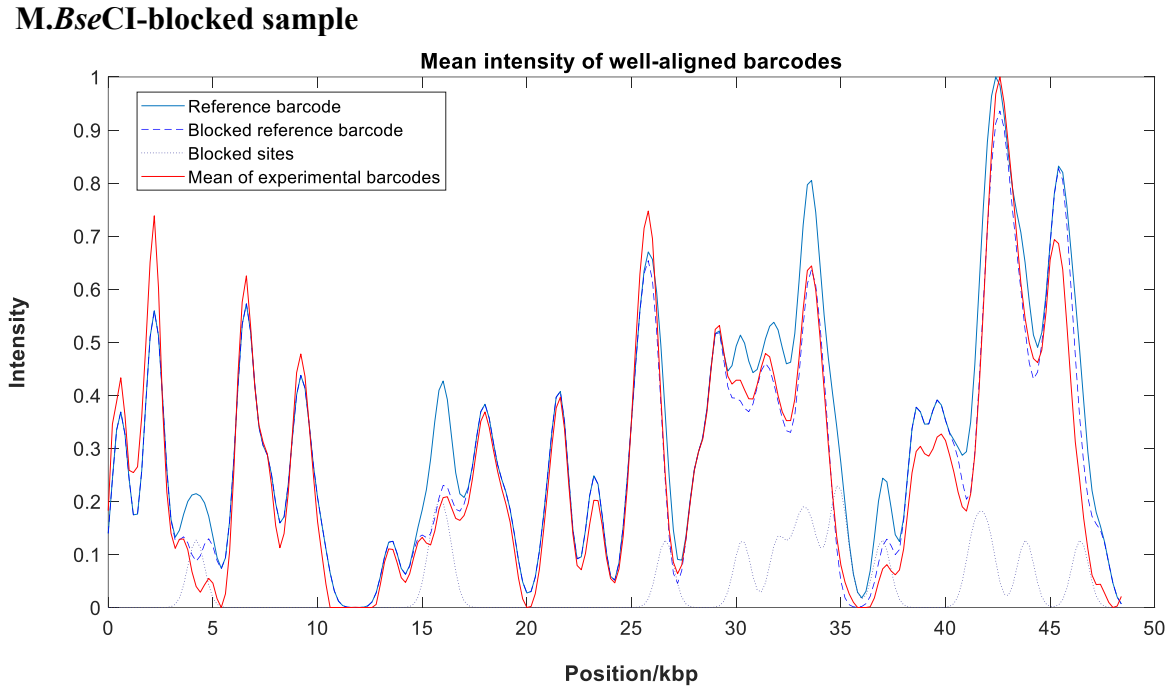
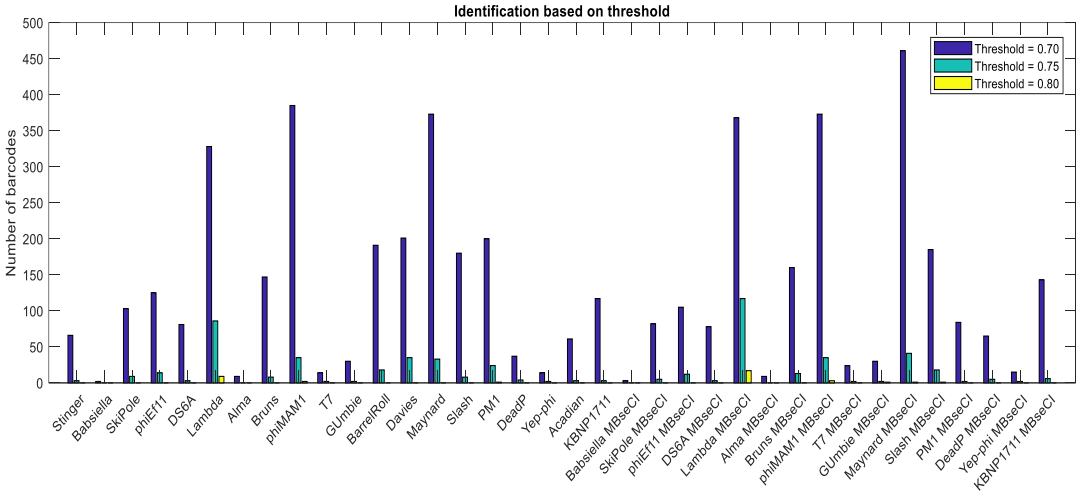


Figure 6.7: "Blocked" lambda sample was labelled with Atto 647N at M.*Taq*I sites after M.*Bse*CI methylation. The mean of experimental barcodes after extraction and alignment was compared to the reference barcodes for blocked and unblocked lambda. The experimental sample barcodes were a better fit to the blocked reference barcode, with notable intensity shifts at the point of blocked sites (dotted grey line).

The sample was then compared to a mixed database of genomes to see if the correct sequence could be identified. For this, a slightly different script was used that, instead of comparing each individual barcode to the "known" reference, ran in a loop comparing every barcode to each other. While this took longer to perform, this meant that a more accurate match could be made, as it would not be attempting to fit them all to a "known" genome.

As can be seen in **Figure 6.8A,** some barcodes mapped to many other references, which is due to the larger genomes sharing a high level of sequence identity with lambda. As the threshold increases, however, it does point more to the correct reference of M.*Bse*CI-methylated lambda, with barcodes only mapping to lambda and blocked-lambda past a threshold of 8. While this has successfully identified the correct genome from a potential pool of others, the fact that it is not 100 per cent certain – mapping to both unblocked and blocked – may cause problems if wanting to be used for diagnostics or screening for clinical samples; further optimisation will be needed if this is going to be used as a reliable test for SMA carrier detection based on this result.

M.*Bse*CI-blocked sample



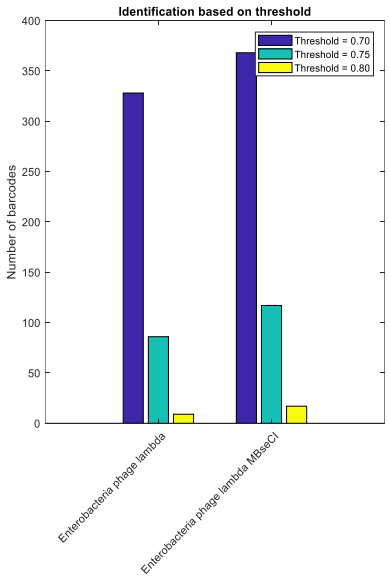
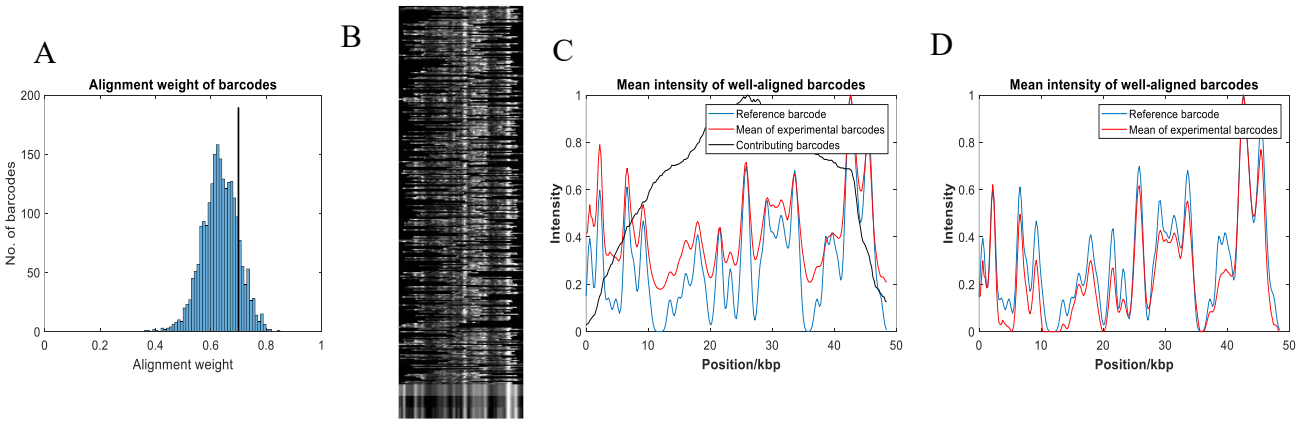


Figure 6.8: Barcodes from the M.*Bse*CI-blocked sample were compared against references from a range of other genomes. The intensity pattern mapped to many of the other genomes as many of them are much larger and contain similar sequences to lambda. Barcodes of threshold 0.8 and higher only mapped to lambda, with the most barcodes correctly mapping to M.*Bse*CI-blocked lambda. This shows that while this technique can align to the correct genome, it may cause erroneous results unless improved further.

The experiment was repeated with unmethylated DNA, to see how well this matched to both the blocked and unblocked references. The data was analysed in the same way as the previous sample, with **Figure 6.9** displaying the results. Again, the middle of the genome was better represented within the sample, shown in **Figure 6.9C/G,** and the barcodes mapped fairly well to both references. One notable difference is again in the 15 kbp region in **Figure 6.9D/H**; D shows that some of the barcodes do follow the same pattern as the blocked reference, possibly due to the DNA not being completely labelled, resulting in coincidental loss of fluorophores at this point. H does map this region much more closely to the unmethylated reference, following the correct pattern of a slightly higher first peak at ~15 kbp compared to the second, where a blocked site would be present. Improvements will need to be made to the mapping process (both deposition and analysis) if wanting to make more confident assumptions/genome identification based on these results.

Unblocked sample

Blocked reference



Unmethylated reference

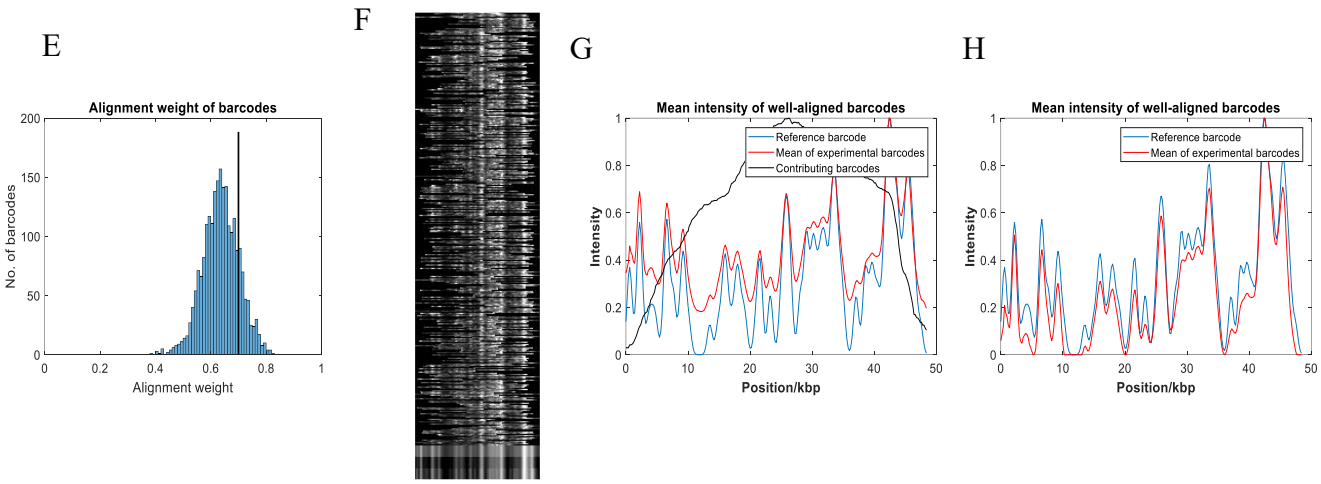


Figure 6.9: Unmethylated lambda sample labelled with Atto 647N using M.*Taq*I. Barcodes were extracted and aligned to a blocked (A) and unblocked (E) lambda reference. Those that had an alignment weight of 0.7 or higher were combined to form a reference barcode (B/F). The mean intensity of well-aligned barcodes (0.7 or higher) was compared to the reference barcode produced to see how many barcodes fit to each region (C/G) and how well they fit each genome (D/H). The unblocked lambda sample does not appear to fit the unmethylated reference significantly better than the blocked reference, improvements will need to be made for more confident correct identification.

Figure 6.10 shows unlabelled barcodes that fit to the blocked reference with a threshold of over 7. The fact that there are barcodes that do seem to incorrectly fit to this reference highlights the problems that may be encountered if trying to use this as a technique for SNP detection. It may be that the labelling and deposition protocols need to be optimised to ensure that fluorophores are present at all unblocked *Taq*I sites before mapping, as inefficient labelling/deposition would lead to a lower intensity, and therefore give false results.

Unblocked sample

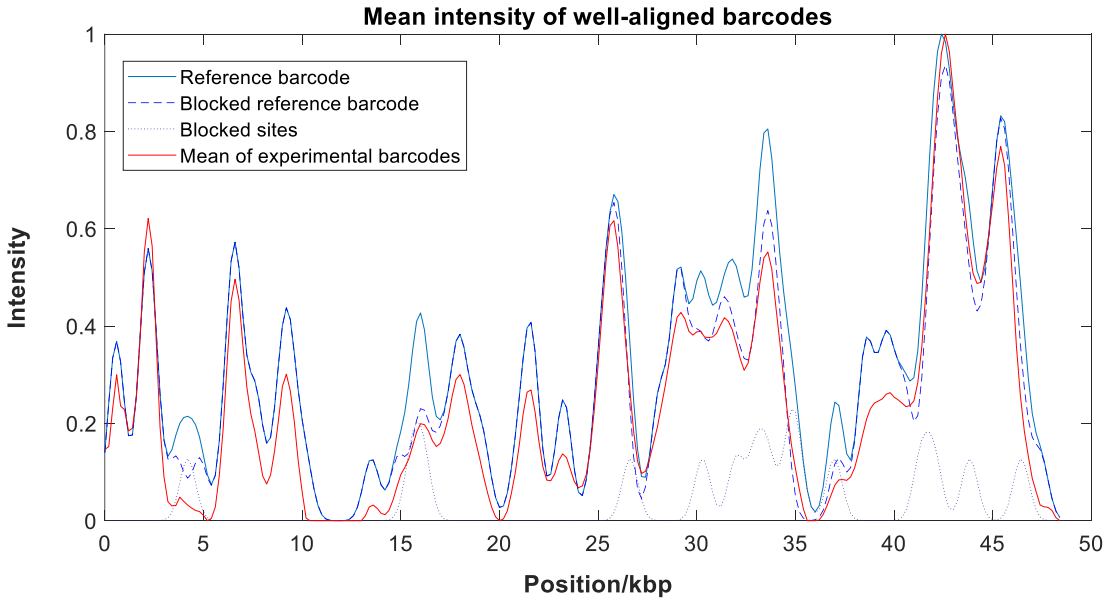
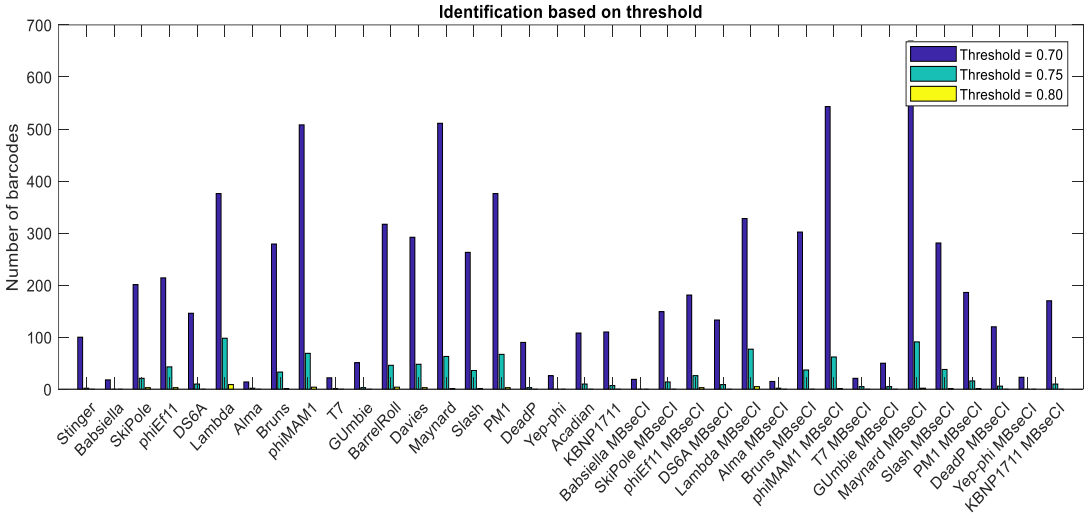


Figure 6.10: Unmethylated lambda was labelled with Atto 647N at M.*Taq*I sites. The mean of experimental barcodes after extraction and alignment was compared to the reference barcodes for blocked and unblocked lambda. The experimental sample barcodes did not seem to fit more closely to the unblocked reference, possibly due to ineffective labelling of the DNA during sample preparation resulting in loss of fluorophores at labelling sites (dotted grey line).

Again, when attempting to map the samples without a “known” genome, it was possible to correctly identify the sample; in this case as unblocked lambda, **Figure 6.11**. The majority of the barcodes did map to the correct genome, which does show potential for this technique. However, until the protocol has been optimised to provide certainty – i.e. only identifying the correct genome at a high threshold – it would not be suitable for SNP detection for carrier testing in the clinic.

Unblocked sample



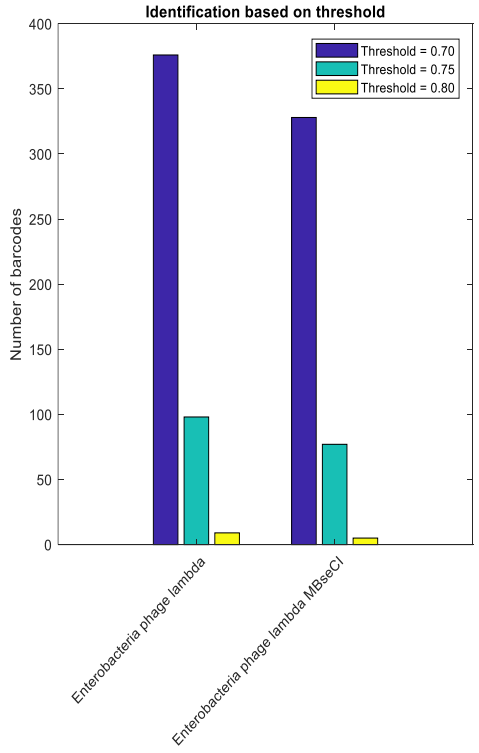


Figure 6.11: Barcodes from the unblocked sample were compared against references from a range of other genomes. The intensity pattern mapped to many of the other genomes as many of them are much larger and contain similar sequences to lambda. Barcodes of threshold 0.8 and higher only mapped to lambda, with the most barcodes correctly mapping to unmethlyated lambda. This shows that while this technique can align to the correct genome, it may cause erroneous results unless improved further.

6.4 Conclusions and future work

As an alternative route to SNP detection to oligoprobes, the *Hpy*188i MTase could be produced in an attempt to detect the SNP in SMN1. The work in this chapter suggests that the current mapping protocol is not currently sensitive enough to identify these SNPs with enough certainty for the clinic, but it shows promise for further work in the future. Labelling and deposition protocols should be optimised in an attempt to improve the reliability of this technique.

If the protein expression of mutated MTases can be optimised to ensure higher yield and activity with synthetic cofactors, further mapping experiments could be carried out. It would be interesting to explore dual colour mapping by combining labelling with different MTases that target both high- and low-density DNA sequences, to map regions of the human genome. This could support DNA sequencing by providing a visual scaffold to help determine the order of specific DNA motifs – a common problem with current sequence techniques.

DNA mapping provides the opportunity to overcome the limitations of sequencing highly repetitive regions of the human genome, or those that contain gaps, but the conditions for preparation and analysis need to be improved to ensure that it is a reliable technique.

CHAPTER SIX
General discussion and future work

General discussion and future work

7.1 General discussion

This thesis has explored the use of MTase-directed labelling of DNA for various applications, including diagnostics. MTases offer a range of advantages over other labelling techniques due to their high specificity and precision, which is ideal for microscopy-based technologies – such as FISH – that require a reliable, efficient, and cost-effective method of labelling DNA to uncover the underlying mutations attributing to many diseases.

The following points have been achieved in this thesis:

1. Active M.*Taq*I protein has been successfully expressed, and an optimised protocol for this process has been developed. M.*Taq*I has been used in a number of labelling reactions and is active with both AdoMet and AdoHcy-6-N_3.
2. Engineered MTases were expressed, with one (M.*Bsa*WI) showing partial activity with AdoMet. This could be useful in future mapping experiments and should be pursued if wanting to continue this work.
3. Oligoprobes were successfully produced for 17CEN using a hairpin design. These probes were then labelled with M.*Taq*I and SPAAC chemistry and were able to detect the ROI using FISH, in as little as a two minute hybridisation.
4. Following the success of 17CEN, probes were also produced for 1CEN and 7CEN and labelled with different dyes. All three colours could be detected simultaneously using FISH, which could significantly improve time to result for patients with ALL.

5. Oligoprobes were successfully produced for detection of the BCR gene. This shows potential for the use of these probes for single gene detection, as well as for translocations such as BCR/ABL.
6. Oligoprobes designed for the highly homologous 17CEN1 and 17CEN2 were somewhat successful in highlighting these slightly different sequences, but design and conditions will need to be optimised in order to be reliable. This shows potential for oligoprobes to be used for SNP detection.
7. Using M.*Bsa*WI to methylate lambda DNA before labelling with M.*Taq*I showed that when mapping this sample, it is possible to detect the small difference in fluorophore pattern caused by the blocked methylation sites. This again shows promise for SNP detection but will need to be optimised if used in clinical applications.

7.1.1 Optimisation of MTases in labelling reactions

Chapter three explored the optimal protocol for producing high yields of active M.*Taq*I protein, as well as different methods to attempt to remove residual AdoMet from the M.*Taq*I protein complex. Complete removal of AdoMet was not successful, but mass spectrometry results of alkylated DNA showed that using an appropriate concentration of M.*Taq*I for the number of labelling sites resulted in the majority of sites being alkylated (as opposed to methylated from residual AdoMet). This suggests that AdoMet may not cause as much of a problem to labelling as originally suggested, and that optimum concentration of M.*Taq*I is more important in order to have enough to label the DNA but without using it in excess. Results suggested that around 0.4 nM M.*Taq*I per nM of sites is a suitable amount to ensure full labelling without unnecessarily adding excess AdoMet into the reaction. Results from mass spectrometry showed that M.*Taq*I has a turnover of 19 in an hour of alkylation, and so an even lower MTase concentration than that tested in gel electrophoresis could potentially

be used. In order to fully remove AdoMet from the complex, extensive washing of the M.*Taq*I protein during purification could be tested.

Chapter three also saw the investigation of different MTase mutations – that were introduced to the DNA sequence based on previous literature – and their expression, and the effect that these engineered sites had on alkylation with synthetic cofactor analogues. Engineering MTases for DNA mapping applications was mostly unsuccessful. More research into the structural changes to the cofactor pocket should be carried out for each protein and specific cofactor analogues if wanting to pursue this as an approach for dual colour mapping. M.*Bsa*WI showed the most promise, offering near-full protection in the presence of AdoMet – but none with AdoHcy-6-N_3 – but further work would need to be performed to enable this protein to be active with other cofactors for labelling; testing wild type M.*Bsa*WI may be interesting for future work to see if there is an improvement of its activity with the cofactors. and this approach should not be dismissed as a powerful tool for mapping. Optimisation of the conditions for expression of these proteins should also be performed to achieve a high enough yield of active protein. Wild type M.*Bse*CI was successfully used to methylate lambda DNA, and was used DNA mapping applications in Chapter 5.

Optimising oligoprobe design and conditions for FISH

Chapter four showed the exciting potential of MTase-labelled DNA as oligoprobes in FISH, and demonstrated that these small hairpins could anneal to patient samples significantly quicker than the traditional FISH protocol. Probes designed for the centromere of chromosome 17 (17CEN) hybridised in as little as two minutes, which could have a huge impact on the turnaround times for patient results. Chapter five showed that probes for 1CEN

and 7CEN could also be annealed rapidly – alongside 17CEN – to test for enumeration of all three probes simultaneously. These probes were labelled with the dyes TAMRA, Rhodamine Green and Atto 647N in order to be distinguished from one another. This rapid hybridisation could potentially improve prognosis for patients with complex forms of cancer such as ALL, which rely on prompt diagnosis.

Conditions should be further optimised to ensure that all oligoprobes are as efficient, and bright, as current BAC probes. Different parameters were explored in Chapter three in order to achieve the best SNR; these included probe concentration, number of M.*Taq*I sites in probe design, formamide percentage in hybridisation buffer, hybridisation time and wash stringency.

Formamide percentage was one of the conditions that had the most effect on SNR, with 40 to 50 % formamide being the optimum amount. The addition of formamide destabilises double stranded DNA by lowering the melting temperature. After calculating the melting temperature of the 17CEN sequence, it showed that a concentration of 60 % and over lowered the melting temperature to 37 ºC or less, which meant that a hybridisation temperature of 37 ºC was too high for all of the oligos to bind, and resulted in decreased SNR. Formamide percentages of less than 40 had the opposite problem; there was an increased amount of non-specifically bound probe contributing to noise and decreasing the SNR. This result highlighted the importance and sensitivity of hybridisation and SNR and suggested that these conditions should be optimised for each new oligo ROI.

Hybridisation time also showed to have an affect on SNR and, although probes could be detected after only two minutes, SNR increased with hybridisation time up to one hour. This

suggests that it takes approximately one hour for all oligos to anneal but, in case of centromeric probes where there are many probes binding and contributing to the signal, results could be detected after as little as two minutes.

The number of M.*Taq*I sites within the probe sequence did not appear to have a significant impact on the SNR, but this may have been different if even more sites were added. Adding extra sites would increase the total oligo length however, which could have an effect on hybridisation times, and would significantly impact cost, therefore it seemed unnecessary at present as the probes could be detected with only one M.*Taq*I site.

The final condition to have a significant impact on the brightness of the probes' signal is the wash conditions. Washing slides at 72 ºC as per standard FISH protocols was not appropriate for the oligos, as this denatured and removed many of the probes that had bound to their target. Washing the slides for at least five minutes at both low and high stringency (at room temperature) significantly reduced background, washing away any probe that had bound non-specifically to the nuclei, resulting in bright signals.

In chapter five, probes were also successfully designed for the gene BCR, which shows potential for this technology to be used for other mutations such as amplifications, deletions or translocations. Results showed that higher formamide percentages were needed for the BCR oligoprobes (70 % formamide as opposed to 40 % for the centromeric probes), as they have purposely been designed to have higher T_ms. This demonstrates that careful design of oligos is crucial to ensure efficient hybridisation, and prevent the oligos from becoming denatured from the ROI during the heating steps.

Detection of SNPs

Chapter five also explored the potential of oligoprobes to detect much smaller mutations, due to their apparent ability to distinguish between highly homologous sequences. As both 17CEN1 and 17CEN2 – two sequences that only differ at 5 base positions – can be present in the centromere of chromosome 17, these probes were labelled with different fluorophores to see if this difference could be detected. The results were inconsistent as there was a high amount of background in the samples, however it did appear that some nuclei contained different combinations of 17CEN1 and 17CEN2, offering promise for this technique.

To improve SNR, probe amplification should be explored to achieve a much brighter probe, as if oligoprobes were to be used to detect a SNP within a single gene, a single fluorophore would be difficult to detect. A new probe design was tested which included a primary probe binding to the ROI, and a labelled secondary probe binding to that. This probe was successful in detecting the ROI and shows potential for this branched oligo approach to potentially add numerous "labelled" sites onto a ROI, amplifying the signal further. MTase-labelling lends itself nicely to this technique, as a pool of labelled probes could be ready for use for different ROIs, providing a quick and cheap option for probe design.

DNA mapping was also tested to see if this approach could determine SNPs, notably for detection of carriers of the SMN1/SMN2 genes which are implicated in SMA. Methylation with M.*Bse*CI blocks 15 M.*Taq*I sites on lambda DNA due to an overlap in the two enzymes' recognition sites. By methylating the DNA with M.*Bse*CI before labelling with M.*Taq*I, results showed that, after deposition and mapping of the DNA, "blocked" lambda could be distinguished from "unblocked" the majority of the time, despite the sequences only differing

at 15 sites. The pattern of fluorescence (barcode) was extracted and assigned to genomes by code written in MATLAB to determine the sequence of the DNA. The majority of barcodes were matched correctly, but there were some barcodes in both samples that were assigned to the incorrect genome. This discrepancy could be attributed to inefficient deposition or sample preparation, resulting in potential shearing of DNA, or overlapping strands, which would need to be optimised in order to achieve more reliable results, especially if being used for clinical applications.

7.2 Future work

The MTase M.*Taq*I has been successfully used for many applications in this thesis – and offers many advantages over other labelling techniques – which opens up the scope for other MTases to be used. The development of MTases with different recognition sites would be useful to produce a "toolbox" of different enzymes that are active with the cofactor analogues. This could prove useful for a variety of projects, in particular where dual colour labelling would be advantageous such as when wanting to map large and complex regions. If using mutated MTases, research should be made into the exact structural changes to the cofactor pocket to ensure that they are suitable, and their gene construct and expression should be optimised to ensure an active protein is produced.

MTases are ideal for DNA mapping work, as they provide a method of labelling DNA without damaging the bases. If deposition and analysis can be optimised, this provides huge potential for mapping to be used in conjunction with sequencing to detect specific ROI, such as when monitoring SMA carrier detection. If this can be achieved, the enzyme M.*Hpy*188i should be investigated, as its recognition is disrupted in the protein SMN1, potentially allowing the detection of SMN1 and SMN2, while retaining sequence context; which is currently not possible using other techniques.

MTase-labelled oligoprobes have shown huge potential for being used in clinical applications to rapidly diagnose a range of diseases. Due to their short size, and therefore increased specificity, oligos need to be designed extremely carefully to ensure the correct ROI is targeted, there are no SNPs in that region, and that it does not cross-hybridise. Oligoprobes for FISH is an area of growing interest, and there is increasingly becoming more software to

help with probe design, including Oligoarray and iFISH[153]. As sequencing is providing an increasing amount of information on the human genome, this will open up the potential for these highly specific oligoprobes to target even more regions, and be applied to many more diseases and biomarkers.

Further work should be carried out using MTase-labelled oligoprobe including investigating the use of the hairpin design in translocations (e.g. BCR/ABL), and trialling new amplified/branched probe designs to improve SNR. It may also prove useful to test how well oligoprobes can highlight regions of the human genome that contain repetitive DNA, but with less repeats than large centromeric copies, to serve as a proof-of-concept before attempting the more difficult feat of SNP detection. An interesting target could be looking at the CAG trinucleotide repeat associated with Huntington disease (HD)[193]. In healthy patients, the CAG region is repeated between 10 and 35 times, but in patients with HD, the same sequence can be repeated more than 120 times. By targeting this region, this could prove the versality of MTase-directed oligoprobes to detect smaller repeats than those associated with centromeres, acting as an interim step between centromeric regions and SNPs.

If oligoprobes can be successfully optimised for SNP detection, these probes have potential not only to detect SNPs associated with genetic diseases, but to differentiate homolog chromosomes within family groups[150,194]. This could play a huge part in future genetic studies to monitor inheritance, fertility, and evolution, offering further insight into both humans and animals, and how we are changing with the world around us.

Supplementary information

8.1 p53 oligoprobe sequences

The following DNA sequences (for the gene p53) were ordered from IDT DNA before being fluorescently-labelled using M.*Taq*I d the protocol in **2.2.4**. Each sequence was ordered following the standard hairpin sequence:

CCCTCGATCGATCGATCGACCCTTTTGGGTCGATCGATCGATCGAGGGTTTT

1. AACTTTGCTGCCACCTGTGT
2. GTAGGACATACCAGCTTAGATTT
3. TTCAGGTCATATACTCAGCCCTG
4. TGCCTTCCTAGGTTGGAAAG
5. AGTTGCTTCAACTACAGGCCT
6. TACGATGGTGTTACTTCCTGATA
7. TGTGTAACAGTTCCTGCATGGG
8. ACTGATTGCTCTTAGGTCTGGC
9. TTATCCATCCCATCACACCCT
10. TGTGAGTGGATCCATTGGAAG
11. AAAGAAGTGCATGGCTGGTGA
12. ACATTTATTGAGCCCAAGCAGG
13. TAAAGGAGCTGTTTGGTAGGG
14. ATTTGTATCCTGGCCCACTGATG
15. TTGATAACAGGGCGTCCACA
16. AAACAGAGGAACAGACTGGGC
17. CTATTGACTAAGGATGTTCAGCA
18. TTTGTGCCGTACTTACGTCATC
19. TTCCTCTTACTTGGCAGAGG
20. TGGATTGGGTAAGCTCCTGACT

8.2 BCR oligoprobe sequences

The following DNA sequences were ordered from IDT DNA before being fluorescently-labelled using M.*Taq*I and the protocol in **2.2.3** and **2.2.4**. Each sequence was ordered following the hairpin sequence: CCCTCGACCCTTTTGGGTCGAGGGTTTT

1. ACCTCAGGCTGGCTGTTGAGAGATT
2. TGACTTCCCTGCTCTGGGTTGTGGTTCT
3. TTTGTACCAAGGCTGGGAGGCACTCAGTGACTT
4. CCTGGTTTATCCAGCATCTGGGATTGTC
5. AGTGCATCTCCTGGGTCTGCCCTTATA

6. TGTCCCTGGAGTTTCTGCAGAGCTGT
7. TTCCAGATTCTGTTGGGTTCGTTGCGTCAGC
8. CCTTGAGAGCATTGAGGAAGCATTGAGGGGCTA
9. TCTGCACTCCAGACTGGGGTTCTTTCT
10. TTCGCTCTGATGTCTCCAGTGGTGACAGTACCT
11. CACCAACCATGCACCAGTGGATTCTGA
12. ACCATCAGGGCGACATGCACTTTGGTTCTCTGT
13. GATAACTCCCAAGCATCACACTGTCC
14. TCAGCTCCTTCCCAGAGGATTTTAGGCACACAG
15. AGGTATAATCCAGTGTCAGTCTGCAGTGGTGGG
16. GTGAGGGAAAGCTGAAATTGTTGCCAAAGGGGG
17. GCCACCAACATTAGCAACAAGGTGCTGCT
18. TGTCTCAGAGTCAGGTGTCTGAAATGTCCTGGG
19. TCTCTCTCCACAGCTCTGCTCTACAAG
20. ATGGTGCTCACACAAGCTCTGTCCACAAAGCTG
21. ATGACGGTGAAGAAGGGAGAGGTGAGTGT
22. CCTACTTCCCCCTGAGTGCTTTCAT
23. AGTGTCCAGGGGGAACAGCTTTTGTCA
24. GTGAACCTGACTGTAGTTGCCTCAGAACCACCT
25. AACACATGGGGCTTGCTTTCCTCCT
26. ATTCCTTGTCTTTGCAGCAGGGTGGGAACATGG
27. CTGAGATGCCTGCTCTTTCTCTTCTACCGAC
28. TTGCTTCACAAAGGCAGGGGCCTGGATCT
29. AGATGCTGCTGATCAGTTGGGCACTCCAA
30. ATTGCTACCTGCTGAGCCTGGGCAAGTCT
31. GGCAGAGGAGAACCAAGGTCTTTCA
32. CAAAGTTTGCAAGGGGTGCTACGGAGA
33. TTTCTTGGGGACCAGAGAGTCTGCA
34. AAACCTCTGGAGTCTGCCACATCCCTGCATAGA
35. TTCTTCCAGACTGGCCTTCCTGGGAA
36. TGTGACTGTCACATTCCCACCTGCAGAGGACAT
37. AAAGGGAGTGTTGTCCTGCCAACTG
38. TTTTCATACACAGACTCCCATGGCCCC
39. TTATGCCGGCTTTGGGATGCAGTCAGGATTGTG
40. TTCTCTCAGATGAGAGTTGCACAGGTGGGTG
41. TGACACTCAGTAGCCTTGCTGAAGG
42. TACTGCAGTCCTTCCCGAAGGACCTCAGTGT
43. TTCCTACCCTTCCACTTATGGGCACCA
44. GGGGAGACACTGGGTTTTCACACTCTCTGTT
45. TTAGGGGTGACTTACCTAGACATGCCCATTCAGCA
46. ATGGCTGGCTTCTTGCCAATTCTGGATCTCCAG
47. AGCCCTGATGTGTTAGCAGGACAGTGAGATG
48. TGAAGGACAGCTTCATGGTGGAGCT
49. GCACTATTGCAGAAAGGTCACCTCAGGACCCAT
50. AGTTATCCTCGGCATAGGCGTGCACACACT
51. CGGTCACATGTTCAGAGTGTCTGTTCCCAGGAA

52. TTCCAGATGGTGGATGAACTGGAGGCAGT
53. ACCCCTGTAAGCTCTCAGCTCTTGGA
54. TGTCCCCCATACAAGCTACCCTGAT
55. TCAATCAAAGGTTAGCCAGGCCAGAGGAG
56. GGCAGGGATGTTGGTAAAAGTTTCTTCTCTCCGC
57. AATGGGAAGGTGAGGCTGTGGCATCT
58. AGATCCCACCTGGTTACCTCCATGTCCCTAA
59. ATTTGCGTAGCCAGGGCGGAGATAACT
60. ACTCATTTCCCCACTGCCCTGTGAT
61. TCATCATTCTCACCTATGCAGAGCCACCTCTCG
62. ACCAGCACTGCACTTGAGAGCCAAGT
63. GGGACTAGTGGACTTTGGTTCAGAAGGAAGAGC
64. GACCCCCTCTGCTGTCCTTGGAACCTTATTA
65. TGTGGGGAAACAGGGAGGTTGTTCAGATGACCA
66. ACCTTCACCCCACAGCAGAGCAGATTT
67. GTCCTGTCTGTGAGCAATACAGCGTGACA
68. ACGACTTCTCCAGCACTGAGCTGCTT
69. GAACGAATGTTGTGGGAAGTCCCGTTTCCCA
70. CAGGTGGGGCACAGGATATTTTCCACT
71. GGCCAGTAGGTGACGTGTCCAAGAGATTT
72. ATCCATGAGAGGTGCCATTTCCCAGCTTCTGCA
73. GAAGATCTGGACTTGGGGACACTCACATGTTCC
74. AGACAACTGGAGAGCTCGGGGAGCAGTTTTT
75. ACGGTCTCATGCCAGGGGTGCTTACAAGGAATA
76. ATGCATGGCGTCCTTTTTCATGCAGCC
77. CCCCTATCTGTGGTCTAGACCCAATTTCTAGGG
78. ACCAGGGTTTCCTGGAGGATCATAGCT
79. CAAATCTTTACCAAGTGCTGGCCTCACCCCCTT
80. GGAGTACTTAGTGCTGGTCTCCTTTGAGATCCG
81. GGTCCTTGCAGCAGATCTTTGAGAGAGCTCA
82. CCAGGTAAAGGGAGGTTCAGATTCTGCCAACCA
83. ATGCACGTGACCTGTGCTCTTCTGTCAGTCTAG

Bibliography

1. Dahm, R. Friedrich Miescher and the discovery of DNA. *Developmental Biology* **278**, 274–288 (2005).
2. Rapoport, S. Rosalind Franklin: Unsung Hero of the DNA Revolution. *The History Teacher* **36**, 116–127 (2002).
3. Chargaff, E., Zamenhof, S. & Green, C. Human Desoxypentose Nucleic Acid: Composition of Human Desoxypentose Nucleic Acid. *Nature* **165**, 756–757 (1950).
4. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
5. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
6. McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *PNAS* **94**, 814–819 (1997).
7. Hartwell, L. H. & Weinert, T. A. Checkpoints: controls that ensure the order of cell cycle events. *Science* **246**, 629–634 (1989).
8. Kamb, A. *et al.* A cell cycle regulator potentially involved in genesis of many tumor types. *Science* **264**, 436–440 (1994).
9. Morgan, S. E. & Kastan, M. B. p53 and ATM: cell cycle, cell death, and cancer. *Adv. Cancer Res.* **71**, 1–25 (1997).
10. Evan, G. I. & Vousden, K. H. Proliferation, cell cycle and apoptosis in cancer. *Nature* **411**, 342–348 (2001).
11. King, R. W., Jackson, P. K. & Kirschner, M. W. Mitosis in transition. *Cell* **79**, 563–571 (1994).
12. LeMaire-Adkins, R., Radke, K. & Hunt, P. A. Lack of Checkpoint Control at the Metaphase/Anaphase Transition: A Mechanism of Meiotic Nondisjunction in Mammalian Females. *The Journal of Cell Biology* **139**, 1611–1619 (1997).

13. Angell, R. First-Meiotic-Division Nondisjunction in Human Oocytes. *The American Journal of Human Genetics* **61**, 23–32 (1997).

14. Hassold, T. & Hunt, P. To err (meiotically) is human: the genesis of human aneuploidy. *Nat Rev Genet* **2**, 280–291 (2001).

15. Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. *Nature Reviews Genetics* **13**, 189–203 (2012).

16. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

17. Cooke, M. S., Evans, M. D., Dizdaroglu, M. & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. *The FASEB Journal* **17**, 1195–1214 (2003).

18. Ye, X. *et al.* Defective S Phase Chromatin Assembly Causes DNA Damage, Activation of the S Phase Checkpoint, and S Phase Arrest. *Molecular Cell* **11**, 341–351 (2003).

19. Hang, B. *et al.* Thirdhand smoke causes DNA damage in human cells. *Mutagenesis* **28**, 381–391 (2013).

20. Bayani, J. *et al.* Genomic mechanisms and measurement of structural and numerical instability in cancer cells. *Seminars in Cancer Biology* **17**, 5–18 (2007).

21. Bačovský, V., Hobza, R. & Vyskot, B. Technical Review: Cytogenetic Tools for Studying Mitotic Chromosomes. in *Plant Chromatin Dynamics* 509–535 (Humana Press, New York, NY, 2018). doi:10.1007/978-1-4939-7318-7_30.

22. Fielding, A. K. Current treatment of Philadelphia chromosome-positive acute lymphoblastic leukemia. *Haematologica* **95**, 8–12 (2010).

23. Gisler, F. M., von Kanel, T., Kraemer, R., Schaller, A. & Gallati, S. Identification of SNPs in the cystic fibrosis interactome influencing pulmonary progression in cystic fibrosis. *Eur. J. Hum. Genet.* **21**, 397–403 (2013).

24. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).

25. Behjati, S. & Tarpey, P. S. What is next generation sequencing? *Arch Dis Child Educ Pract Ed* **98**, 236–238 (2013).

26. Levy-Sakin, M. & Ebenstein, Y. Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy. *Current Opinion in Biotechnology* **24**, 690–698 (2013).

27. Hatzimichael, E., Lagos, K., Sim, V. R., Briasoulis, E. & Crook, T. Epigenetics in diagnosis, prognostic assessment and treatment of cancer: an update. *EXCLI J* **13**, 954–976 (2014).

28. Mann, M. R. W. & Bartolomei, M. S. Towards a Molecular Understanding of Prader-Willi and Angelman Syndromes. *Hum. Mol. Genet.* **8**, 1867–1873 (1999).

29. Weksberg, R., Shuman, C. & Beckwith, J. B. Beckwith–Wiedemann syndrome. *Eur J Hum Genet* **18**, 8–14 (2010).

30. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11995–11999 (1993).

31. Pfeifer, G. P., Kadam, S. & Jin, S.-G. 5-hydroxymethylcytosine and its potential roles in development and cancer. *Epigenetics & Chromatin* **6**, 10 (2013).

32. Gaudet, F. *et al.* Induction of tumors in mice by genomic hypomethylation. *Science* **300**, 489–492 (2003).

33. Korenberg, J. R. *et al.* Down syndrome phenotypes: the consequences of chromosomal imbalance. *PNAS* **91**, 4997–5001 (1994).

34. Stochholm, K., Juul, S., Juel, K., Naeraa, R. W. & Højbjerg Gravholt, C. Prevalence, Incidence, Diagnostic Delay, and Mortality in Turner Syndrome. *None* **91**, 3897–3902 (2006).

35. Bishop, R. Applications of fluorescence in situ hybridization (FISH) in detecting genetic aberrations of medical significance. *Bioscience Horizons* **3**, 85–95 (2010).

36. Levsky, J. M. & Singer, R. H. Fluorescence in situ hybridization: past, present and future. *Journal of Cell Science* **116**, 2833–2838 (2003).

37. Matera, A. G. & Ward, D. C. Oligonucleotide probes for the analysis of specific repetitive DNA sequences by fluorescence in situ hybridization. *Hum. Mol. Genet.* **1**, 535–539 (1992).

38. Aurich-Costa, J., Zamechek, L., Keenan, P. & Bradley, S. Oligo fluorescence in situ hybridization (oligo-fish), a new strategy for enumerating chromosomes in interphase nuclei. *Fertility and Sterility* **88**, S86 (2007).

39. Raap, A. K. Advances in fluorescence in situ hybridization. *Mutat. Res.* **400**, 287–298 (1998).

40. Kwon, S. Single-molecule fluorescence in situ hybridization: Quantitative imaging of single RNA molecules. *BMB Rep* **46**, 65–72 (2013).

41. Huber, D., Voith von Voithenberg, L. & Kaigala, G. V. Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering* **1**, 15–24 (2018).

42. Oncogene - BCR/ABL: from molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia. http://www.nature.com/onc/journal/v21/n56/full/1206082a.html.

43. List, A. *et al.* Lenalidomide in the Myelodysplastic Syndrome with Chromosome 5q Deletion. *New England Journal of Medicine* **355**, 1456–1465 (2006).

44. O'Keefe, C. L., Warburton, P. E. & Matera, A. G. Oligonucleotide Probes for Alpha Satellite DNA Variants Can Distinguish Homologous Chromosomes by FISH. *Hum. Mol. Genet.* **5**, 1793–1799 (1996).

45. Beliveau, B. J. *et al.* Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *PNAS* **109**, 21301–21306 (2012).

46. Tinawi-Aljundi, R. *et al.* One-year monitoring of an oligonucleotide fluorescence in situ hybridization probe panel laboratory-developed test for bladder cancer detection. *Res Rep Urol* **7**, 49–55 (2015).

47. Conner, B. J. *et al.* Detection of sickle cell beta S-globin allele by hybridization with synthetic oligonucleotides. *PNAS* **80**, 278–282 (1983).

48. Ersfeld, K. Fiber-FISH: fluorescence in situ hybridization on stretched DNA. *Methods Mol. Biol.* **270**, 395–402 (2004).

49. Raap, null *et al.* Fiber FISH as a DNA Mapping Tool. *Methods* **9**, 67–73 (1996).

50. Cole, C. G. *et al.* Finishing the finished human chromosome 22 sequence. *Genome Biol.* **9**, R78 (2008).

51. Lemoine, S., Combes, F. & Le Crom, S. An evaluation of custom microarray applications: the oligonucleotide design challenge. *Nucleic Acids Res* **37**, 1726–1739 (2009).

52. Lin, M. *et al.* dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics* **20**, 1233–1240 (2004).

53. Common SNPs explain a large proportion of the heritability for human height | Nature Genetics. https://www.nature.com/articles/ng.608.

54. Thomson, W. *et al.* Rheumatoid arthritis association at 6q23. *Nat Genet* **39**, 1431–1433 (2007).

55. Olama, A. A. A. *et al.* Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet* **41**, 1058–1060 (2009).

56. Bustin, S. A. Developments in real-time PCR research and molecular diagnostics. *Expert Review of Molecular Diagnostics* **10**, 713–715 (2010).

57. VanGuilder, H. D., Vrana, K. E. & Freeman, W. M. Twenty-five years of quantitative PCR for gene expression analysis. *BioTechniques* **44**, 619–626 (2008).

58. Bartley, P. A. *et al.* Sensitive detection and quantification of minimal residual disease in chronic myeloid leukaemia using nested quantitative PCR for BCR-ABL DNA. *International Journal of Laboratory Hematology* **32**, e222–e228 (2010).

59. Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: Lessons from Large-Scale Biology. *Science* **300**, 286–290 (2003).

60. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

61. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467 (1977).

62. Green, E. D. Strategies for the systematic sequencing of complex genomes. *Nat Rev Genet* **2**, 573–583 (2001).

63. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res* **46**, 2159–2168 (2018).

64. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* **13**, 278–289 (2015).

65. Eisenstein, M. Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology* **30**, 295–296 (2012).

66. Schneider, G. F. & Dekker, C. DNA sequencing with nanopores. *Nature Biotechnology* **30**, 326–328 (2012).

67. Urban, J. M., Bliss, J., Lawrence, C. E. & Gerbi, S. A. Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. *bioRxiv* 019281 (2015) doi:10.1101/019281.

68. Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* **8**, (2019).

69. Feng, Y., Zhang, Y., Ying, C., Wang, D. & Du, C. Nanopore-based Fourth-generation DNA Sequencing Technology. *Genomics, Proteomics & Bioinformatics* **13**, 4–16 (2015).

70. Bayley, H. Sequencing single molecules of DNA. *Current Opinion in Chemical Biology* **10**, 628–637 (2006).

71. Loman, N. J. & Watson, M. Successful test launch for nanopore sequencing. *Nature Methods* **12**, 303–304 (2015).

72. Lichtman, J. W. & Conchello, J.-A. Fluorescence microscopy. *Nat Methods* **2**, 910–919 (2005).

73. Gustafsson, M. G. L. Nonlinear structured-illumination microscopy: Wide-field fluorescence imaging with theoretically unlimited resolution. *PNAS* **102**, 13081–13086 (2005).

74. Cheng, X. & Roberts, R. J. AdoMet-dependent methylation, DNA methyltransferases and base flipping. *Nucleic Acids Res* **29**, 3784–3795 (2001).

75. Naderer, M., Brust, J. R., Knowle, D. & Blumenthal, R. M. Mobility of a Restriction-Modification System Revealed by Its Genetic Contexts in Three Hosts. *J Bacteriol* **184**, 2411–2419 (2002).

76. Lukinavičius, G., Lapinaitė, A., Urbanavičiūtė, G., Gerasimaitė, R. & Klimašauskas, S. Engineering the DNA cytosine-5 methyltransferase reaction for sequence-specific labeling of DNA. *Nucleic Acids Res* **40**, 11594–11602 (2012).

77. Pljevaljčić, G., Schmidt, F. & Weinhold, E. Sequence-specific Methyltransferase-Induced Labeling of DNA (SMILing DNA). *ChemBioChem* **5**, 265–269 (2004).

78. Vranken, C. *et al.* Super-resolution optical DNA Mapping via DNA methyltransferase-directed click chemistry. *Nucleic Acids Research* **42**, e50–e50 (2014).

79. Dalhoff, C., Lukinavicius, G., Klimasauskas, S. & Weinhold, E. Synthesis of S-adenosyl-L-methionine analogs and their use for sequence-specific transalkylation of DNA by methyltransferases. *Nature protocols* **1**, 1879–86 (2006).

80. Roberts, R. J. *et al.* A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* **31**, 1805–1812 (2003).

81. Ananiev, G. E. *et al.* Optical mapping discerns genome wide DNA methylation profiles. *BMC Molecular Biology* **9**, 68 (2008).

82. Boultwood, J. & Wainscoat, J. S. Gene silencing by DNA methylation in haematological malignancies. *British Journal of Haematology* **138**, 3–11 (2007).

83. Shivapurkar, N. & Gazdar, A. F. DNA Methylation Based Biomarkers in Non-Invasive Cancer Screening. *Curr Mol Med* **10**, 123–132 (2010).

84. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* **39**, 457–466 (2007).

85. Lauster, R., Trautner, T. A. & Noyer-Weidner, M. Cytosine-specific type II DNA methyltransferases. A conserved enzyme core with variable target-recognizing domains. *J. Mol. Biol.* **206**, 305–312 (1989).

86. Pósfai, J., Bhagwat, A. S., Pósfai, G. & Roberts, R. J. Predictive motifs derived from cytosine methyltransferases. *Nucleic Acids Res* **17**, 2421–2435 (1989).

87. Comstock, L. R. & Rajski, S. R. Conversion of DNA methyltransferases into azidonucleosidyl transferases via synthetic cofactors. *Nucleic Acids Res* **33**, 1644–1652 (2005).

88. Schmidt, F. H.-G. *et al.* Sequence-specific Methyltransferase-Induced Labelling (SMILing) of plasmid DNA for studying cell transfection. *Bioorganic & Medicinal Chemistry* **16**, 40–48 (2008).

89. Grazvydas Lukinavicius Christian Dalhoff. Direct transfer of extended groups from synthetic cofactors by DNA methyltransferases. *Nature chemical biology* **2**, 31–2 (2006).

90. Lukinavičius, G. *et al.* Targeted Labeling of DNA by Methyltransferase-Directed Transfer of Activated Groups (mTAG). *J. Am. Chem. Soc.* **129**, 2758–2759 (2007).

91. Lukinavičius, G., Tomkuvienė, M., Masevičius, V. & Klimašauskas, S. Enhanced Chemical Stability of AdoMet Analogues for Improved Methyltransferase-Directed Labeling of DNA. *ACS Chem. Biol.* **8**, 1134–1139 (2013).

92. Poh, W. J., Wee, C. P. P. & Gao, Z. DNA Methyltransferase Activity Assays: Advances and Challenges. *Theranostics* **6**, 369–391 (2016).

93. Neely, R. K., Deen, J. & Hofkens, J. Optical mapping of DNA: Single-molecule-based methods for mapping genomes. *Biopolymers* **95**, 298–311 (2011).

94. Samad, A., Huff, E. F., Cai, W. & Schwartz, D. C. Optical mapping: a novel, single-molecule approach to genomic analysis. *Genome Res.* **5**, 1–4 (1995).

95. Levy-Sakin, M. *et al.* Towards Single-Molecule Optical Mapping of the Epigenome. *ACS nano* **8**, 14–26 (2014).

96. Wand, N. O. *et al.* DNA barcodes for rapid, whole genome, single-molecule analyses. *Nucleic Acids Res* **47**, e68–e68 (2019).

97. Deen, J. *et al.* Combing of Genomic DNA from Droplets Containing Picograms of Material. *ACS Nano* **9**, 809–816 (2015).

98. Grunwald, A. *et al.* Bacteriophage strain typing by rapid single molecule analysis. *Nucleic Acids Research* (2015) doi:10.1093/nar/gkv563.

99. Cai, W. *et al.* Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proc Natl Acad Sci U S A* **92**, 5164–5168 (1995).

100. Meng, X., Benson, K., Chada, K., Huff, E. J. & Schwartz, D. C. Optical mapping of lambda bacteriophage clones using restriction endonucleases. *Nat Genet* **9**, 432–438 (1995).

101. Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat Biotech* **31**, 135–141 (2013).

102. Zhou, S. *et al.* A Single Molecule Scaffold for the Maize Genome. *PLoS Genet* **5**, e1000711 (2009).

103. Teague, B. *et al.* High-resolution human genome structure by single-molecule analysis. *PNAS* **107**, 10848–10853 (2010).

104. Teng, C. *et al.* Whole-Genome Optical Mapping and Finished Genome Sequence of Sphingobacterium deserti sp. nov., a New Species Isolated from the Western Desert of China. *PLoS One* **10**, (2015).

105. Xiao, M. *et al.* Rapid DNA mapping by fluorescent single molecule detection. *Nucleic Acids Res* **35**, e16 (2007).

106. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. *Journal of Molecular Biology* **113**, 237–251 (1977).

107. Heiter, D. F., Lunnen, K. D. & Wilson, G. G. Site-specific DNA-nicking mutants of the heterodimeric restriction endonuclease R.BbvCI. *J. Mol. Biol.* **348**, 631–640 (2005).

108. Das, S. K. *et al.* Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucl. Acids Res.* 673 (2010) doi:10.1093/nar/gkq673.

109. Neely, R. K. *et al.* DNA fluorocode: A single molecule, optical map of DNA with nanometre resolution. *Chem. Sci.* **1**, 453–460 (2010).

110. Sproule, D. M. & Kaufmann, P. Therapeutic developments in spinal muscular atrophy. *Ther Adv Neurol Disord* **3**, 173–185 (2010).

111. Lefebvre, S. *et al.* Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* **80**, 155–165 (1995).

112. Monani, U. R. *et al.* A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2. *Hum. Mol. Genet.* **8**, 1177–1183 (1999).

113. Official REBASE Homepage | The Restriction Enzyme Database | NEB. http://rebase.neb.com/rebase/rebase.html.

114. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

115. Pironon, N., Puechberty, J. & Roizès, G. Molecular and evolutionary characteristics of the fraction of human alpha satellite DNA associated with CENP-A at the centromeres of chromosomes 1, 5, 19, and 21. *BMC Genomics* **11**, 195 (2010).

116. Waye, J. S., England, S. B. & Willard, H. F. Genomic organization of alpha satellite DNA on human chromosome 7: evidence for two distinct alphoid domains on a single chromosome. *Mol Cell Biol* **7**, 349–356 (1987).

117. Danna, K. & Nathans, D. Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae. *Proc. Natl. Acad. Sci. U.S.A.* **68**, 2913–2917 (1971).

118. Schwartz, D. C. *et al.* Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science* **262**, 110–114 (1993).

119. Pearn, J. Classification of spinal muscular atrophies. *Lancet* **1**, 919–922 (1980).

120. Feldkötter, M., Schwarzer, V., Wirth, R., Wienker, T. F. & Wirth, B. Quantitative Analyses of SMN1 and SMN2 Based on Real-Time LightCycler PCR: Fast and Highly Reliable Carrier Testing and Prediction of Severity of Spinal Muscular Atrophy. *Am J Hum Genet* **70**, 358–368 (2002).

121. Schluckebier, G., Kozak, M., Bleimling, N., Weinhold, E. & Saenger, W. Differential binding of S-adenosylmethionine S-adenosylhomocysteine and Sinefungin to the adenine-specific DNA methyltransferase M.TaqI11Edited by T. Richmond. *Journal of Molecular Biology* **265**, 56–67 (1997).

122. Huber, T. D. *et al.* Functional AdoMet Isosteres Resistant to Classical AdoMet Degradation Pathways. *ACS Chem Biol* **11**, 2484–2491 (2016).

123. Vabulas, R. M., Raychaudhuri, S., Hayer-Hartl, M. & Hartl, F. U. Protein Folding in the Cytoplasm and the Heat Shock Response. *Cold Spring Harb Perspect Biol* **2**, (2010).

124. Ito, H., Sadaoka, A., Kotani, H., Hiraoka, N. & Nakamura, T. Cloning, nucleotide sequence, and expression of the HincII restriction-modification system. *Nucleic Acids Res* **18**, 3903–3911 (1990).

125. Braun, G. *et al.* Enzyme-Directed Positioning of Nanoparticles on Large DNA Templates. *Bioconjugate Chemistry* **19**, 476–479 (2008).

126. Kapetaniou, E. G. *et al.* Purification, crystallization and preliminary X-ray analysis of the BseCI DNA methyltransferase from Bacillus stearothermophilus in complex with its cognate DNA. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **63**, 12–14 (2006).

127. Hanz, G. M., Jung, B., Giesbertz, A., Juhasz, M. & Weinhold, E. Sequence-specific Labeling of Nucleic Acids and Proteins with Methyltransferases and Cofactor Analogues. *J Vis Exp* (2014) doi:10.3791/52014.

128. Rina, M. & Bouriotis, V. Cloning, purification and characterization of the BseCI DNA methyltransferase from Bacillus stearothermophilus. *Gene* **133**, 91–94 (1993).

129. Huber, D., Voith von Voithenberg, L. & Kaigala, G. V. Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering* **1**, 15–24 (2018).

130. Herrick, J. & Bensimon, A. Introduction to Molecular Combing: Genomics, DNA Replication, and Cancer. in *DNA Replication* (eds. Vengrova, S. & Dalgaard, J. Z.) 71–101 (Humana Press, 2009). doi:10.1007/978-1-60327-815-7_5.

131. Inaba, H., Greaves, M. & Mullighan, C. G. Acute lymphoblastic leukaemia. *Lancet* **381**, (2013).

132. Acute Lymphocytic Leukemia - Cancer Stat Facts. https://seer.cancer.gov/statfacts/html/alyl.html.

133. Hunger, S. P. & Mullighan, C. G. Acute Lymphoblastic Leukemia in Children. *http://dx.doi.org/10.1056/NEJMra1400972* https://www.nejm.org/doi/10.1056/NEJMra1400972 (2015) doi:10.1056/NEJMra1400972.

134. Brown, P. Treatment of infant leukemias: challenge and promise. *Hematology Am Soc Hematol Educ Program* **2013**, 596–600 (2013).

135. Paul, S., Kantarjian, H. & Jabbour, E. J. Adult Acute Lymphoblastic Leukemia. *Mayo Clinic Proceedings* **91**, 1645–1666 (2016).

136. Liehr, T., Starke, H., Weise, A., Lehrer, H. & Claussen, U. Multicolor FISH probe sets and their applications. *Histology and histopathology* (2004).

137. Speicher, M. R., Ballard, S. G. & Ward, D. C. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nat Genet* **12**, 368–375 (1996).

138. Cartwright, I. M. Modified PNA Telomere and Centromere FISH Protocols. in *Radiation Cytogenetics: Methods and Protocols* (eds. Kato, T. A. & Wilson, P. F.) 101–105 (Springer New York, 2019). doi:10.1007/978-1-4939-9432-8_12.

139. Grady, D. L. *et al.* Highly conserved repetitive DNA sequences are present at human centromeres. *Proc Natl Acad Sci U S A* **89**, 1695–1699 (1992).

140. Idziak, D., Robaszkiewicz, E. & Hasterok, R. Spatial distribution of centromeres and telomeres at interphase varies among Brachypodium species. *J. Exp. Bot.* erv369 (2015) doi:10.1093/jxb/erv369.

141. Mehta, G. D., Agarwal, M. P. & Ghosh, S. K. Centromere identity: a challenge to be faced. *Mol Genet Genomics* **284**, 75–94 (2010).

142. Knight, J. C. *Human Genetic Diversity: Functional Consequences for Health and Disease*. (OUP Oxford, 2009).

143. Waye, J. S. & Willard, H. F. Structure, organization, and sequence of alpha satellite DNA from human chromosome 17: evidence for evolution by unequal crossing-over and an ancestral pentamer repeat shared with the human X chromosome. *Mol Cell Biol* **6**, 3156–3165 (1986).

144. Sullivan, L. L., Chew, K. & Sullivan, B. A. α satellite DNA variation and function of the human centromere. *Nucleus* **8**, 331–339 (2017).

145. Choo, K. H., Vissel, B., Nagy, A., Earle, E. & Kalitsis, P. A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucleic Acids Res* **19**, 1179–1182 (1991).

146. Baumgartner, A., Weier, J. F. & Weier, H.-U. G. Chromosome-specific DNA Repeat Probes. *J Histochem Cytochem* **54**, 1363–1370 (2006).

147. Warburton, P. E. *et al.* Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* **9**, 533 (2008).

148. O'Keefe, C. L. & Matera, A. G. Alpha Satellite DNA Variant-Specific Oligoprobes Differing by a Single Base Can Distinguish Chromosome 15 Homologs. *Genome Res.* **10**, 1342–1350 (2000).

149. Antson, D.-O., Mendel-Hartvig, M., Landegren, U. & Nilsson, M. PCR-generated padlock probes distinguish homologous chromosomes through quantitative fluorescence analysis. *Eur J Hum Genet* **11**, 357–363 (2003).

150. O'Keefe, C. L., Griffin, D. K., Bean, C. J., Matera, A. G. & Hassold, T. J. Alphoid variant-specific FISH probes can distinguish autosomal meiosis I from meiosis II non-disjunction in human sperm. *Hum Genet* **101**, 61–66 (1997).

151. Rouillard, J.-M., Zuker, M. & Gulari, E. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucl. Acids Res.* **31**, 3057–3062 (2003).

152. Navin, N. *et al.* PROBER: oligonucleotide FISH probe design software. *Bioinformatics* **22**, 2437–2438 (2006).

153. Gelali, E. *et al.* iFISH is a publically available resource enabling versatile DNA FISH to study genome architecture. *Nature Communications* **10**, 1636 (2019).

154. Nyberg, L., Persson, F., Åkerman, B. & Westerlund, F. Heterogeneous staining: a tool for studies of how fluorescent dyes affect the physical properties of DNA. *Nucleic Acids Res* **41**, e184 (2013).

155. Hughes, L. D., Rawle, R. J. & Boxer, S. G. Choose Your Label Wisely: Water-Soluble Fluorophores Often Interact with Lipid Bilayers. *PLoS One* **9**, (2014).

156. Zanetti-Domingues, L. C., Tynan, C. J., Rolfe, D. J., Clarke, D. T. & Martin-Fernandez, M. Hydrophobic Fluorescent Probes Introduce Artifacts into Single Molecule Tracking Experiments Due to Non-Specific Binding. *PLOS ONE* **8**, e74200 (2013).

157. Lunn, M. R. & Wang, C. H. Spinal muscular atrophy. *The Lancet* **371**, 2120–2133 (2008).

158. Zohar, H. & Muller, S. J. Labeling DNA for Single-Molecule Experiments: Methods of Labeling Internal Specific Sequences on Double-Stranded DNA. *Nanoscale* **3**, 3027–3039 (2011).

159. Icy. http://icy.bioimageanalysis.org/.

160. Ostromohov, N., Huber, D., Bercovici, M. & Kaigala, G. V. Real-Time Monitoring of Fluorescence in Situ Hybridization Kinetics. *Anal. Chem.* **90**, 11470–11477 (2018).

161. Jares-Erijman, E. A. & Jovin, T. M. FRET imaging. *Nat Biotechnol* **21**, 1387–1395 (2003).

162. Interchromophoric Interactions Determine the Maximum Brightness Density in DNA Origami Structures | Nano Letters. https://pubs.acs.org/doi/pdf/10.1021/acs.nanolett.8b04845.

163. Blake, R. D. & Delcourt, S. G. Thermodynamic Effects of Formamide on DNA Stability. *Nucleic Acids Res* **24**, 2095–2103 (1996).

164. Haar, F. M., Durm, M., Hausmann, M., Ludwig, H. & Cremer, C. Optimization of Fast-FISH for alpha-satellite DNA probes. *J. Biochem. Biophys. Methods* **33**, 43–54 (1996).

165. Gosden, J. & Lawson, D. Rapid chromosome identification by oligonucleotide-primed in situ DNA synthesis (PRINS). *Hum. Mol. Genet.* **3**, 931–936 (1994).

166. Baumgartner, A., Weier, J. F. & Weier, H.-U. G. Chromosome-specific DNA Repeat Probes. *J Histochem Cytochem* **54**, 1363–1370 (2006).

167. Waye, J. S. & Willard, H. F. Structure, organization, and sequence of alpha satellite DNA from human chromosome 17: evidence for evolution by unequal crossing-over and an ancestral pentamer repeat shared with the human X chromosome. *Mol Cell Biol* **6**, 3156–3165 (1986).

168. Genomic Characterization of Large Heterochromatic Gaps in the Human Genome Assembly. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003628.

169. Mutation, DNA Repair, and DNA Integrity | Learn Science at Scitable. https://www.nature.com/scitable/topicpage/dna-damage-repair-mechanisms-for-maintaining-dna-344/.

170. Joerger, A. C. & Fersht, A. R. Structure–function–rescue: the diverse nature of common p53 cancer mutants. *Oncogene* **26**, 2226–2242 (2007).

171. Joerger, A. C. & Fersht, A. R. Structure–function–rescue: the diverse nature of common p53 cancer mutants. *Oncogene* **26**, 2226–2242 (2007).

172. Nowell, P. C. Discovery of the Philadelphia chromosome: a personal perspective. *J Clin Invest* **117**, 2033–2035 (2007).

173. Groffen, J. *et al.* Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell* **36**, 93–99 (1984).

174. Soupir, C. P. *et al.* Philadelphia chromosome-positive acute myeloid leukemia: a rare aggressive leukemia with clinicopathologic features distinct from chronic myeloid leukemia in myeloid blast crisis. *Am. J. Clin. Pathol.* **127**, 642–650 (2007).

175. Rouhanifard, S. H. *et al.* Exponential fluorescent amplification of individual RNAs using clampFISH probes. *bioRxiv* 222794 (2018) doi:10.1101/222794.

176. Lizardi, P. M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat Genet* **19**, 225–232 (1998).

177. Choi, H. M. T. *et al.* Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* **145**, dev165753 (2018).

178. Dirks, R. M. & Pierce, N. A. Triggered amplification by hybridization chain reaction. *PNAS* **101**, 15275–15278 (2004).

179. Battich, N., Stoeger, T. & Pelkmans, L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nature Methods* **10**, 1127–1133 (2013).

180. Player, A. N., Shen, L.-P., Kenny, D., Antao, V. P. & Kolberg, J. A. Single-copy Gene Detection Using Branched DNA (bDNA) In Situ Hybridization. *J Histochem Cytochem.* **49**, 603–611 (2001).

181. Boettiger, A. N. *et al.* Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* **529**, 418–422 (2016).

182. Beliveau, B. J. *et al.* Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nat Commun* **6**, (2015).

183. SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues | Nature Methods. https://www.nature.com/articles/s41592-019-0404-0.

184. Baerlocher, G. M. & Lansdorp, P. M. Telomere length measurements in leukocyte subsets by automated multicolor flow-FISH. *Cytometry Part A* **55A**, 1–6 (2003).

185. Rufer, N., Dragowska, W., Thornbury, G., Roosnek, E. & Lansdorp, P. M. Telomere length dynamics in human lymphocyte subpopulations measured by flow cytometry. *Nat Biotechnol* **16**, 743–747 (1998).

186. Braz, G. T. *et al.* Comparative Oligo-FISH Mapping: An Efficient and Powerful Methodology To Reveal Karyotypic and Chromosomal Evolution. *Genetics* **208**, 513–523 (2018).

187. Bensimon, A. *et al.* Alignment and sensitive detection of DNA by a moving interface. *Science* **265**, 2096–2098 (1994).

188. Allemand, J. F., Bensimon, D., Jullien, L., Bensimon, A. & Croquette, V. pH-dependent specific binding and combing of DNA. *Biophys J* **73**, 2064–2070 (1997).

189. Michalet, X. *et al.* Dynamic Molecular Combing: Stretching the Whole Human Genome for High-Resolution Studies. *Science* **277**, 1518–1523 (1997).

190. Riehn, R. *et al.* Restriction mapping in nanofluidic devices. *PNAS* **102**, 10012–10016 (2005).

191. Tegenfeldt, J. O. *et al.* The dynamics of genomic-length DNA molecules in 100-nm channels. *PNAS* **101**, 10979–10983 (2004).

192. Reisner, W. *et al.* Single-molecule denaturation mapping of DNA in nanofluidic channels. *PNAS* **107**, 13294–13299 (2010).

193. Penney, J. B., Vonsattel, J.-P., Macdonald, M. E., Gusella, J. F. & Myers, R. H. CAG repeat number governs the development rate of pathology in Huntington's disease. *Annals of Neurology* **41**, 689–692 (1997).

194. Farré, M. *et al.* Novel Insights into Chromosome Evolution in Birds, Archosaurs, and Reptiles. *Genome Biol Evol* **8**, 2442–2451 (2016).