METHYLTRANSFERASE-DIRECTED LABELLING FOR VISUALISATION AND IDENTIFICATION OF DNA

BY

NATHANIEL OVELIN WAND

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

DOCTOR OF PHILOSOPHY

Physical Sciences of Imaging for the Biomedical Sciences
School of Chemistry
College of Engineering and Physical Sciences
University of Birmingham

March 2018

ABSTRACT

New techniques for rapid identification of complex mixtures of viral and bacterial DNA and for visualising multi-copy plasmids in single bacterial cells have been developed using a combination of methyltransferase-directed labelling, molecular combing, and widefield microscopy.

In Chapter 2 the protocol for methyltransferase-directed labelling was optimised. A maximum labelling efficiency of 50% (as measured by single molecule counting results) was obtained for Atto 647N-labelled pUC19.

In Chapters 3 and 4 the optimised labelling protocol was used to label genomic DNA for optical mapping and identification. The *in silico* results in Chapter 3 show that the combination of techniques used in this thesis represents the 'sweet-spot' for optical mapping and identification of microorganisms. In Chapter 4 this combination of techniques was used along with a new algorithm, for rapid identification of bacteriophages, Adenovirus A, resistance plasmids and bacterial strains in complex mixtures of genomic DNA.

Finally, Chapter 5 uses methyltransferase-directed labelling to investigate the mechanisms bacteria use to maintain and transfer resistance plasmids. Atto647N-labelled pUC19 and pRSET B plasmids were visualised in *E. coli* bacteria and diffusion over several seconds or plasmid segregation over many hours was studied.

ACKNOWLEDGMENTS

I would like to thank all those who have guided and supported me throughout my research project. As everyone knows, science is a team game and without the support of others I would not have got where I have.

First, I would like to thank my main supervisor, Robert Neely for giving me the chance to pursue such a great project and for the help, ideas and guidance he has given me throughout. I would also like to thank my other supervisors, Steve Busby and Iain Styles for their help, ideas and guidance throughout.

Next, I would like to thank all those from the PSIBS doctoral training centre, and the EPSRC for funding. I have had a great time with everyone and I believe that being able to interact with a cohort of students throughout my doctoral project has been a great advantage. Particular thanks go to those from my year: Sophie Glinton, Prema Chola Torres, Suzie Mason, Dan Pencross, Iggy Partarrieu, Siobhan King, Mo Rassul, Emma Meeus, Jack Pearce and Richard Young. We've had some great experiences together and I am grateful for your support and friendship. Special thanks go to Sophie Glinton for putting up with me as a flatmate (and now boyfriend) for the past three years.

I would like to thank those in the Neely group, past and present, for sharing my research project highs and lows with me, for listening to my presentations and for putting up with my healthy scepticism. Special thanks go to those who have helped directly with this project: Mo Rassul, Darren Smith, Ashleigh Rushton, Andrew Wilkinson and Su Wang. Also, special thanks to Lara Horne for being a great project student and helping with the bacterial transformations.

I would also like to thank those in the Busby lab who helped me with the bacterial transformations. In particular, thanks to Rita Godfrey, Doug Browning, Laura Sellers and Dave Lee. Also, thanks go to Michelle Buckner and Roger Grant for supplying precious DNA samples. I would also like to express my gratitude for anyone who has helped me at any other point, but who I may have forgotten.

Finally, I would like to thank my family and friends for listening to me go on about my project and for encouraging me. Special thanks to my parents, Kathryn and John, and my brother Charlie.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 DNA Structure and Function

1.1.1 DNA as the molecule of inheritance

"The apple doesn't fall far from the tree". Humans have always been aware that traits are inherited from one generation to the next and have used this knowledge for selective breeding of domestic plants and animals. However, it wasn't until 1865 that Gregor Mendel, the father of genetics, discovered the fundamental laws of inheritance, which were rediscovered at the start of the 20th century and began the field of classical genetics[1].

Despite this, the molecular basis of inheritance was unknown. In 1928 Frederick Griffith showed that genetic material could be transferred between bacteria[2]. Mice injected with a mixture of heat-killed virulent bacteria and nonvirulent bacteria died, as the cell debris of the virulent bacteria transformed the nonvirulent bacteria. The nature of this 'transforming principle' was determined by Avery and co-workers in 1944[3] when they separated the classes of molecules found in the cell debris and found that only one molecule, deoxyribonucleic acid, DNA, induced the transformation. DNA was confirmed to be the molecule of genetic inheritance by the 1952 Hershey-Chase experiment, in which it was shown, using radioactive isotopes, that primarily DNA, rather than proteins, entered the cell upon phage infection[4].

1.1.2 Structure of DNA

The apparent simplicity of the structure of DNA belied its importance. It was first isolated by Friedrich Miescher in 1869[5] and due to its presence in the cell nuclei, he

termed the substance 'nuclein'. It was known to be a long polymer composed of four types of nucleotide subunits, chemically identical except for the nitrogen base (Figure 1.1)[6]. Each nucleotide consists of a phosphate group, a deoxyribose sugar and one of four bases: adenine (A), guanine (G), cytosine (C) and thymine (T). Adenine and guanine are known as purines, whilst cytosine and thymine are pyrimidines.

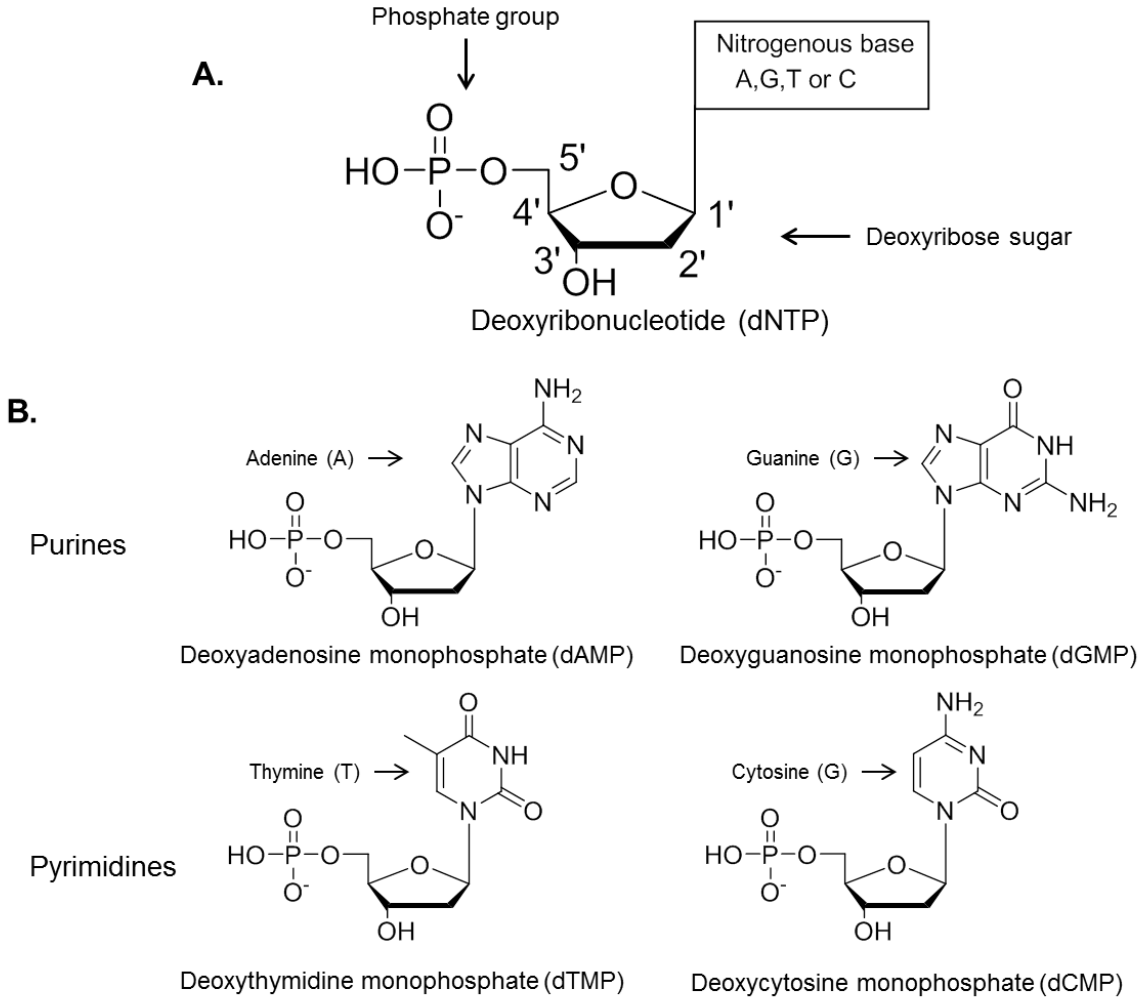


Figure 1.1 **The chemical structure of nucleotides. A) Chemical structure of a nucleotide, showing the basic building blocks: a nitrogenous base, a deoxyribose sugar and a phosphate group. B) The structure of each of the naturally-occurring DNA bases. These are purines (adenine and guanine) and pyrimidines (thymine and cytosine).**

Watson and Crick showed how these simple building blocks were put together when they published the structure of DNA in 1953[7]. X-ray diffraction data by Rosalind Franklin and Maurice Wilkins had suggested DNA was a helical molecule, whilst Erwin Chargaff had established that the amount of pyrimidines and purines was always the same, as was the amount of adenine and thymine and conversely the amount of guanine and cytosine[8]. From these results, Watson and Crick derived the famous double helix (Figure 1.2), in which each helix is a chain of nucleotides linked by phosphodiester bonds. The two helices are held together by hydrogen bonds between base pairs, with each pair consisting of a purine and a pyrimidine, in which adenine pairs with thymine and guanine with cytosine. Two hydrogen bonds are formed between adenine and thymine and three hydrogen bonds between guanine and cytosine. The directionality of a DNA chain is denoted 5' to 3' (named from the orientation of the ribose carbons) and the two helices run in opposite directions, or antiparallel.

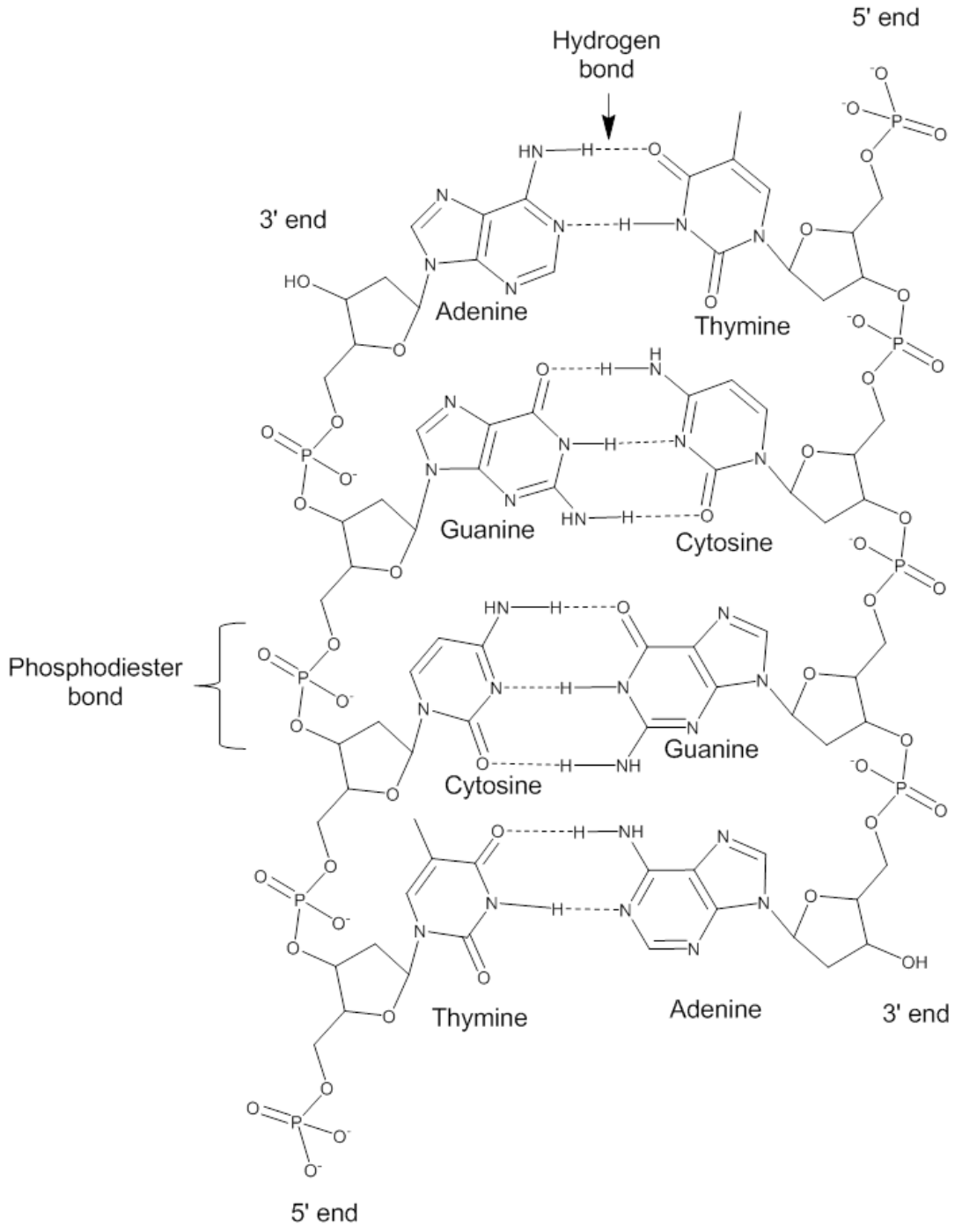


Figure 1.2 **The chemical structure of DNA. The structure is a double helix, in which each helix is a chain of nucleotides held together by phosphodiester bonds. The direction of the chain is denoted 5' to 3' and the two helices run antiparallel. The chains are held together by hydrogen bonding between base pairs, base-stacking interactions and hydrophobic effects. In Watson-Crick base-pairing adenine forms two hydrogen bonds with thymine and guanine forms three hydrogen bonds with cytosine.**

In the 3-dimensional structure of DNA the bases partly stack on top of one another in the double helix structure (Figure 1.3). This allows for favourable electrostatic interactions between bases and the exclusion of water allows for stabilisation by hydrophobic effects. The stacking results in a double helix with two grooves known as the major and minor grooves. The primary physiological form of DNA is the B-form, in which there are 10 bases per helical turn and 0.34 nm between base pairs, although at least two other biologically relevant forms, the A- and Z-form, also exist.

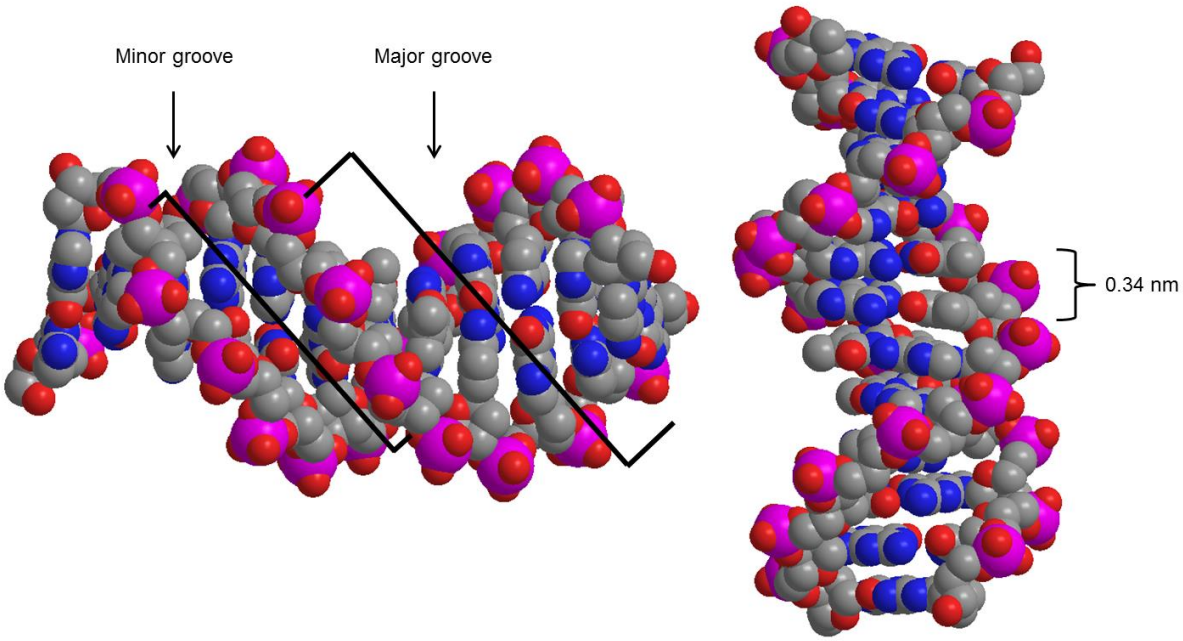


Figure 1.3 **The structure of the B-form of DNA. Crystal structure from Drew *et al*[9]. The primary physiological form of DNA is the B-form, in which there are 10 bases per helical turn and 0.34 nm between base pairs. In DNA the bases partly stack on top of one another in the double helix structure resulting in two grooves known as the major and minor grooves. (Grey=carbon, red=oxygen, blue=nitrogen, pink=phosphorus).**

1.1.3 The Central Dogma

The structure of DNA makes it clear how DNA can be used to store and replicate genetic information. The bases can be considered a four-letter code that spells out biological messages, whilst the complementary base pairing means each helix can be used as a template for replication of DNA.

The flow of genetic information from DNA is explained by the central dogma of molecular biology (Figure 1.4), first stated by Francis Crick in 1958[10], that "information cannot be transferred back from protein to either protein or nucleic acid". This describes the flow of genetic information, in which DNA is transcribed to produce mRNA, which is then translated to synthesise proteins, which are in turn responsible for biological functions in cells.

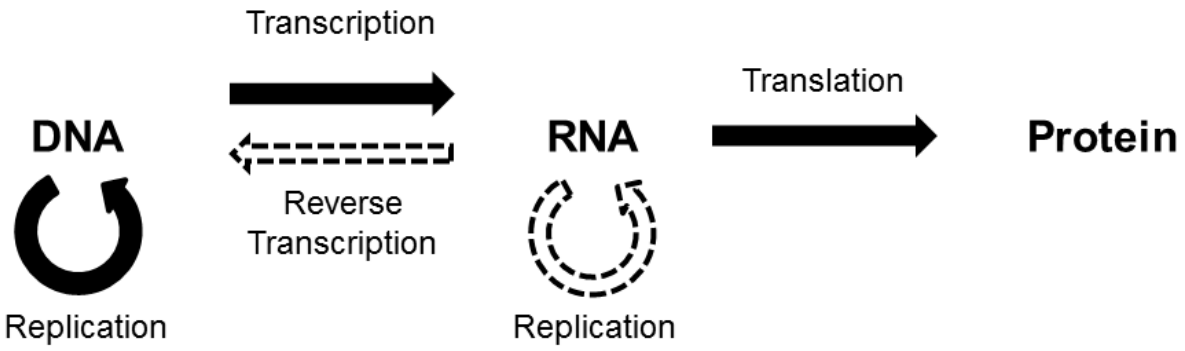


Figure 1.4 **The central dogma of molecular biology. This describes the flow of genetic information in a biological system. There are three general transfers (solid black arrows): replication of DNA; transcription of DNA to mRNA; translation of mRNA to proteins. Two other special transfers are known to naturally occur (dashed arrows): reverse transcription of RNA to DNA and replication of RNA. There are no known transfers of information that occur from proteins to either proteins or nucleic acids.**

Ribonucleic acid (RNA) is also a nucleic acid that shares a similar structure to DNA and also has important biological roles. It is also composed of a chain of nucleotides, but unlike DNA is usually single-stranded, forming secondary structures rather than a double helix. Chemically it is similar to DNA, but thymine is replaced with uracil (lacking a methyl group) and deoxyribose is replaced by ribose (adding a hydroxyl group at the 2' position). These differences give DNA greater stability than RNA, which has more transient roles. The main forms of RNA are: messenger RNA (mRNA), which is used to code for synthesis of proteins; transfer RNA (tRNA), which delivers amino acids for protein synthesis; and ribosomal RNA (rRNA), the RNA component of ribosomes, which are responsible for protein synthesis. Other roles include gene regulation and post-transcriptional modification of RNA.

There are 20 naturally occurring amino acids, which are coded for by the four bases in DNA (Table 1.1). Three letter codes, known as codons, give 64 possible combinations. In 1961 Nirenberg and Matthaei deciphered the first of these 64 combinations[11]. They prepared an extract from bacterial cells, from which protein could be synthesised, before adding artificial RNA, composed of a single repeated codon. Each RNA yielded a specific polypeptide chain composed of a single amino acid which could be identified and used to decode all 64 combinations. These codons also include 'Stop' codons to terminate protein synthesis as well as degeneracy in the code.

1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	Phenylalanine	TCT	Serine	TAT	Tyrosine	TGT	Cysteine	T
	TTC	Phenylalanine	TCC	Serine	TAC	Tyrosine	TGC	Cysteine	C
	TTA	Leucine	TCA	Serine	TAA	STOP	TGA	STOP	A
	TTG	Leucine	TCG	Serine	TAG	STOP	TGG	Tryptophan	G
C	CTT	Leucine	CCT	Proline	CAT	Histidine	CGT	Arginine	T
	CTC	Leucine	CCC	Proline	CAC	Histidine	CGC	Arginine	C
	CTA	Leucine	CCA	Proline	CAA	Glutamine	CGA	Arginine	A
	CTG	Leucine	CCG	Proline	CAG	Glutamine	CGG	Arginine	G
A	ATT	Isoleucine	ACT	Threonine	AAT	Asparagine	AGT	Serine	T
	ATC	Isoleucine	ACC	Threonine	AAC	Asparagine	AGC	Serine	C
	ATA	Isoleucine	ACA	Threonine	AAA	Lysine	AGA	Arginine	A
	ATG	Methionine	ACG	Threonine	AAG	Lysine	AGG	Arginine	G
G	GTT	Valine	GCT	Alanine	GAT	Aspartate	GGT	Glycine	T
	GTC	Valine	GCC	Alanine	GAC	Aspartate	GGC	Glycine	C
	GTA	Valine	GCA	Alanine	GAA	Glutamate	GGA	Glycine	A
	GTG	Valine	GCG	Alanine	GAG	Glutamate	GGG	Glycine	G

Table 1.1 **DNA codons. There are 20 naturally occurring amino acids, which are coded for by the four bases in DNA. Three letter codes, known as codons, give 64 possible combinations and all are shown here. These codons also include, 'Stop' codons to terminate protein synthesis as well as significant degeneracy in the code. This degeneracy is usually in the third base of the codon known as 'third-base wobble'.**

1.1.4 Importance of the DNA sequence

Each non-cloned individual and each species will have a unique genetic code and therefore a unique DNA sequence, producing a unique set of proteins. This can be used to identify species or individuals when identification by phenotype alone is difficult, for instance microorganisms. In addition, the DNA sequence can also provide information about the evolutionary history of an organism. As species evolve changes are accumulated in the DNA sequence, such as mutations, insertions, deletions and translocations. Closely related species will therefore share more sequence identity than species which are less related and this information can be used to produce a phylogenetic tree, detailing the evolutionary history of a group[12].

The DNA code is also related to genetic diseases[13]. A single error in a single gene can cause a protein to be incorrectly coded and therefore affect its folding and function. An example of this is sickle-cell disease, a blood disorder in which a single nucleotide is mutated, (a single nucleotide polymorphism or SNP), resulting in a single amino acid substitution from glutamate to valine. At low oxygen concentrations this change causes protein aggregation and contributes to several health problems. Larger changes in the genome can include: duplications, which are common in many types of cancer and can cause protein to be overexpressed; deletions, for instance having a single copy of the SHOX gene is associated with short stature; and rearrangements[14].

To diagnose disease, both transmitted (e.g. infections) and genetic, it is therefore useful to be able to read the DNA code and identify DNA sequences. This allows for accurate diagnosis of diseases, which can then be targeted with the appropriate treatment. This is an important motivation behind the development of several techniques that are capable of identifying the DNA code.

1.2 Identifying the DNA code

1.2.1 Restriction mapping

One of the earliest techniques used to identify DNA was restriction mapping. Restriction enzymes are naturally occurring enzymes that are used in the restriction-modification (RM) system in bacteria, a defence mechanism used by the bacterium against foreign DNA. This system was first observed in 1952-53 by Luria and Human[15] and Bertani and Weigle[16], who reported that bacteriophages varied in their ability to grow on different host strains.

In 1970 the first restriction enzyme that cleaved DNA into specific fragments, endonuclease R, was described by Smith and Wilcox[17]. This was used a year later by Kathleen Danna and Daniel Nathans to produce specific fragments of SV50[18]. Smith and Wilcox had used sucrose gradients to analyse fragments, but they lacked the resolution for proper separation. Nathans instead used polyacrylamide gel electrophoresis, described by Ulrich Loening[19]. An electric field is applied across the gel, causing DNA molecules to move towards the cathode. Smaller molecules will move through the gel faster than larger molecules, effectively separating molecules by size.

Differences in DNA sequences can be identified by differences in restriction patterns, known as restriction fragment length polymorphisms. For instance those found in human mini-satellites (regions of the genome that contain repeat sequences) were first used by Alec Jeffreys in 1985[20] to identify individuals in the process of DNA fingerprinting. DNA from an individual is extracted, restricted and separated by gel electrophoresis. Mini-satellites are identified by Southern blotting (see later) and since

the length of these repeat regions tends to vary in length between individuals, the resultant restriction patterns will vary and allow for identification.

It is also possible to use multiple restriction enzymes to map DNA (Figure 1.5)[21]. If two restriction enzymes are used and both single and double digests carried out, then the location of the restriction sites can be inferred. This is still commonly used to determine the orientation of a cloned insert but was particularly useful when automated sequencing (see later) was still prohibitively expensive.

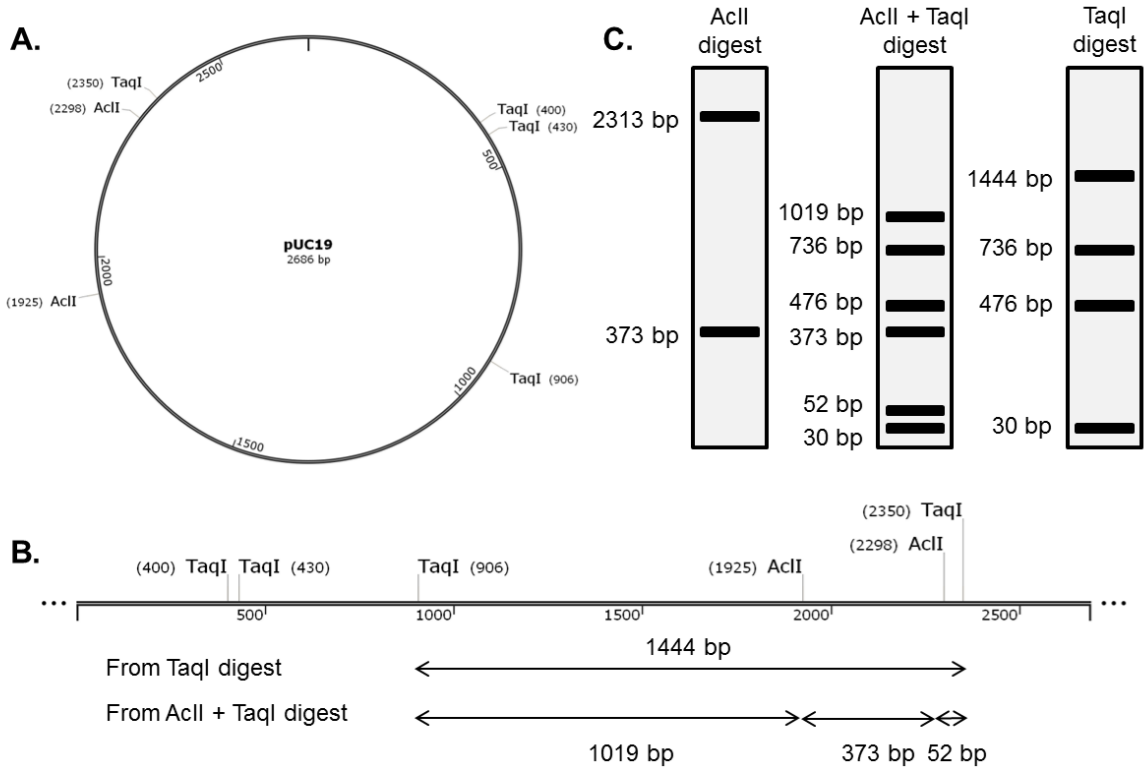


Figure 1.5 **Restriction mapping of pUC19. A) Circular representation of pUC19, showing position of restriction sites for two enzymes: TaqI and AclI. B) Linear representation, showing the distance between restriction sites. C) Example of restriction digests. pUC19 is digested with each restriction enzyme and with a combination of both, then fragments are separated by gel electrophoresis. By comparing differences in restriction pattern, it is possible to map the relative position of the restriction sites. For example, the 1444 bp fragment in the TaqI digest is restricted into three fragments (1019, 373 and 52 bp) in the combined digest, implying the position of the two AclI sites.**

1.2.2 DNA Hybridisation

Although restriction mapping can help identify DNA fragments and map restriction sites it doesn't give any information on the sequence of the DNA. One of the simplest ways to probe the sequence is to take advantage of complementary Watson-Crick base pairing. In these methods single-stranded DNA or RNA will bind, or hybridise, to the complementary base sequence.

This was exploited in 1975 by Edwin Southern, in a technique now known as 'Southern blotting' (Figure 1.6)[22]. Southern used restriction enzymes and gel electrophoresis to separate DNA molecules but wanted to know which restriction fragments contained a specific sequence, complementary to a given RNA. This could be done by a time-consuming process: cutting out the fragments, eluting the DNA and hybridising to RNA. However, to speed up the process Southern transferred the DNA from the gel to a nitrocellulose membrane and then hybridised with radiolabelled RNA, that could be detected by autoradiography. This gives a relatively high throughput method of identifying the position and copy number of specific DNA sequences.

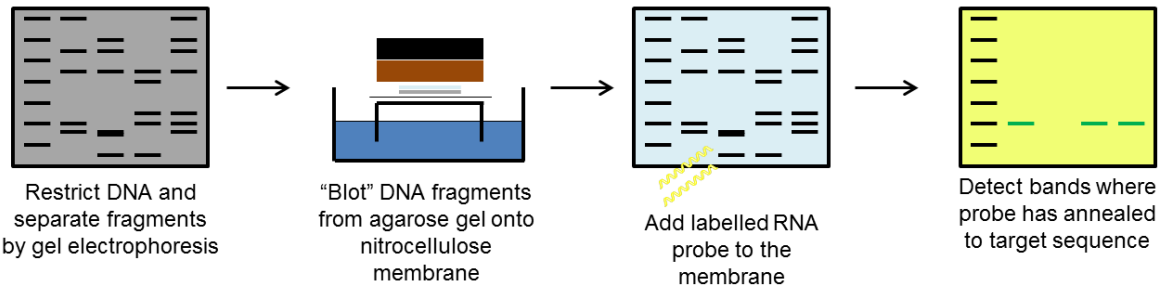


Figure 1.6 **Southern blotting procedure. A DNA sample is restricted, and fragments separated by gel electrophoresis. Fragments are then transferred from the gel onto a nitrocellulose membrane before exposure with a labelled nucleic acid probe. The probe will hybridise to specific DNA sequences which can be detected.**

A related technique is fluorescence in situ hybridization (FISH) (Figure 1.7). This built on earlier methods that used radiolabelled probes[23], in which radiolabelled RNA was incubated with chromosomes and imaged by autoradiography. The first use of fluorescent in situ detection was in 1980 and allowed for significantly better resolution, speed and safety[24]. Single stranded nucleic acid probes are covalently labelled with fluorescent dyes and can be used to detect single genes when bound to chromosomal DNA, which has been attached to a substrate, typically a glass slide.

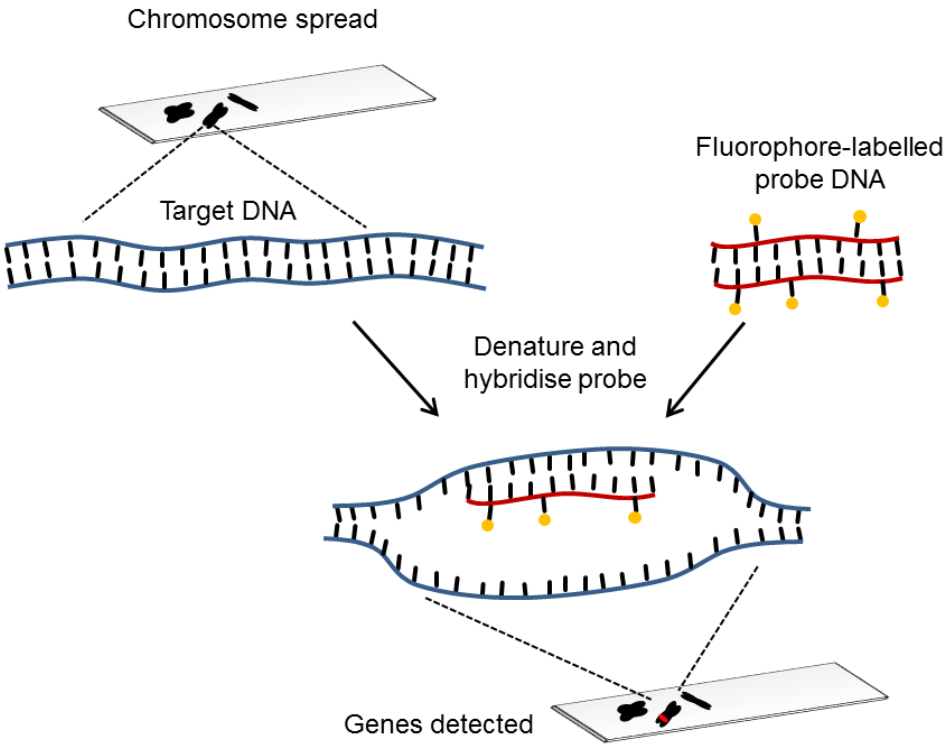


Figure 1.7 **Fluorescent in situ hybridisation (FISH) procedure. An interphase or metaphase chromosome spread is prepared, in which chromosomes are usually fixed to a glass slide. In addition, probe DNA is prepared, ranging in size from around 50 bp to over 100 kbp. This is specific for a region of the target DNA and is fluorescently labelled by, for example, nick translation. The probe and target DNA are denatured, followed by hybridisation on the glass slide, allowing visualisation of the specific region of interest, for example the position of genes.**

The inherent resolution of FISH is limited by the chromatin condensation state to around 1-3 megabases. However, by using de-condensed chromatin and by combing out DNA fibers, in a technique known as fiber-FISH this resolution can be dramatically increased to the order of kilobases (Figure 1.8)[25]. The order and number of genes can be identified, in a way complementary to restriction mapping and, for instance, can show whether gene rearrangements have occurred.

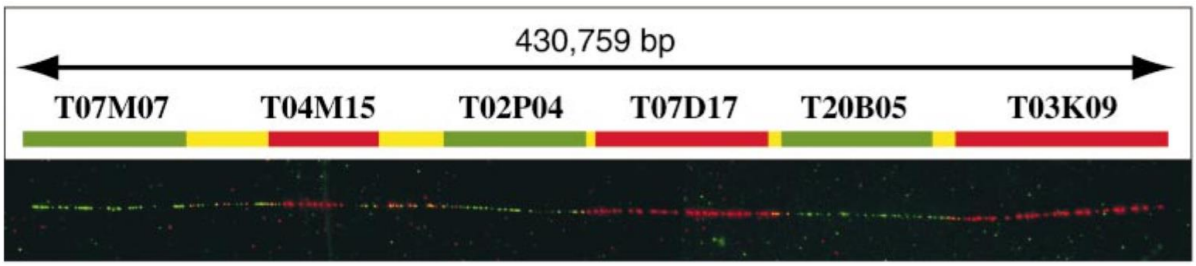


Figure 1.8 **Example of Fiber-FISH. Taken from Jackson *et al*[26]. Fiber-FISH uses the same probe hybridisation as FISH, but target DNA is now stretched when immobilised on glass slides. Molecular combing of DNA follows protein digestion in high salt and detergent, to remove histones and de-condense chromatin. Using this the order and number of genes can be identified. A) A 431 kbp DNA construct, consisting of six bacterial artificial chromosome (BAC) clones (T07M07, T04M15, T02P04, T07D17, T20B05 and T03K09) is used for physical mapping. The clones were detected by alternating green-red probes, with overlapping regions shown in yellow. B) Fiber-FISH signal from a single DNA fiber from the 431 kbp DNA construct.**

DNA hybridisation is also exploited in DNA microarray technology, where the aim is to detect and quantify the expression of thousands of genes at a time[27]. Traditionally, single stranded probes are attached to a solid support in an orderly collection of spots. Each spot contains many copies of a single known sequence and a microarray will contain thousands of ordered spots, therefore can detect thousands of different genes. The target DNA is fluorescently labelled, added to the microarray and allowed to bind to the probes. Non-specific DNA is washed away, and a fluorescent signal is detected from each spot when hybridisation has occurred, allowing for a read-out of the number of copies of

a gene that are present (based on the fluorescence intensity). Microarrays are particularly well-suited to detecting single nucleotide polymorphisms and the variation in expression levels of genes.

The polymerase chain reaction (PCR) is another common technique that exploits hybridisation to DNA (Figure 1.9). PCR is able to amplify a few copies of a specific region of DNA by many orders of magnitude and though not usually a technique that is used for DNA identification directly, it has enabled many of these other technologies, by allowing very large amounts of pure DNA to be produced. PCR was developed by Kary Mullis in 1983 and involves several steps[28]. First the target DNA is melted or denatured, then the reaction is cooled, and specifically designed primers are allowed to anneal at either end of the target DNA. A DNA polymerase is then allowed to extend the primers to synthesise new strands of DNA. The cycle is repeated, with an effective doubling of the number of target DNA molecules during each cycle.

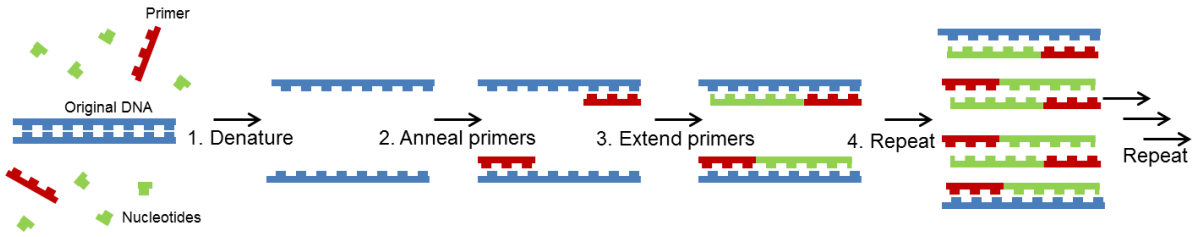


Figure 1.9 **Polymerase chain reaction (PCR) procedure. First the two strands of the target DNA are denatured by heating the mixture to 95°C. The mixture is cooled to 55-70°C to allow the annealing of specific primers to either end of the target DNA, before the primers are extended by heat-resistant DNA polymerase at 72°C. This is cycle is run many times with an effective doubling of DNA in each cycle. For instance, after 30 cycles the target DNA can be amplified 2^{30} or over 1 billion times.**

An extension of this method is quantitative PCR (qPCR), which allows for direct quantification of the amount of sample DNA[29]. The amplification of DNA is monitored in real-time by using dyes that intercalate into double-stranded DNA or by using fluorescently-labelled primers which will fluoresce when annealed to the target DNA. This allows for the presence and quantity of the target DNA to be tested.

These methods share the common feature that a known sequence of interest is required, however when the DNA of a completely novel organism needs to be identified then the DNA code itself must be read.

1.2.3 DNA Sequencing

DNA sequencing is used to provide the base sequence of a piece of DNA[30]. Initial efforts in the 1960s focused on sequencing RNA, for which pure samples were more readily available and for which the molecule is uncomplicated by the complementary strand. By combining analytical chemistry techniques and partial digestion of RNA, Robert Holey and colleagues were able to sequence the first whole nucleic acid in 1965, that of alanine tRNA from *S. cerevisiae*[31]. The first complete protein-coding gene sequence, the coat protein of bacteriophage MS2, was produced in 1972 by Walter Fiers' laboratory[32]. The method used 2-dimensional gel electrophoresis (i.e. a gel run twice in perpendicular directions in two different buffer conditions) to separate and detect radiolabelled, partially digested, RNA fragments, the sequences of which could be deciphered. The MS2 genome was the first complete genome to be published in 1976[33] using the same method.

These methods were adapted to sequence DNA, with two complex methods being developed in the mid-1970s: Alan Coulson and Sanger's 'plus and minus' system[34] and

Allan Maxam and Walter Gilbert's chemical cleavage technique[35]. Using the plus and minus system Sanger and colleagues reported the first DNA sequence, of bacteriophage phiX174 in 1977[36].

However, the most commonly used first generation sequencing technology was the dideoxy chain-termination method, developed by Sanger and colleagues in the same year (Figure 1.10)[37]. Dideoxynucleotides lack the 3' hydroxyl group required for extension of the DNA and therefore, by mixing a fraction of these with standard deoxynucleotides, chain termination is caused randomly at every position within the DNA molecule during replication. If four parallel reactions are carried out for each base and the fragments run on a gel, the order of bases can effectively be read out.

Various improvements were made to this first generation of technology, for instance using fluorescently labelled dideoxynucleotides allows the DNA to be sequenced and read out in a high throughput manner in a single reaction rather than four separate reactions. Rather than using gels, capillary electrophoresis can be used to separate DNA fragments by length. As the DNA is run through the capillary the fluorescent signal can be detected, which will report the dideoxynucleotide that was incorporated at that point and which can therefore be used to record the sequence.

Sanger sequencing can be used to sequence fragments accurately but only those of around 700-900 bases in length, so for sequencing longer fragments 'shotgun-sequencing' can be used[38]. In shotgun sequencing DNA is randomly fragmented to produce a large number of overlapping fragments, which can be sequenced separately, before being assembled *in silico*. This process can be used on small fragments, but also to reconstruct whole genomes. A complementary approach to whole genome sequencing is

to use bacterial artificial chromosomes (BACs). These are DNA constructs based on bacterial plasmids, into which a small fragment (~150 kbp) from a larger genome is inserted. This essentially divides the genomes into sections, which are amplified by replication in bacterial cells, sequenced (for instance by shotgun sequencing) and used to reconstruct the whole genome.

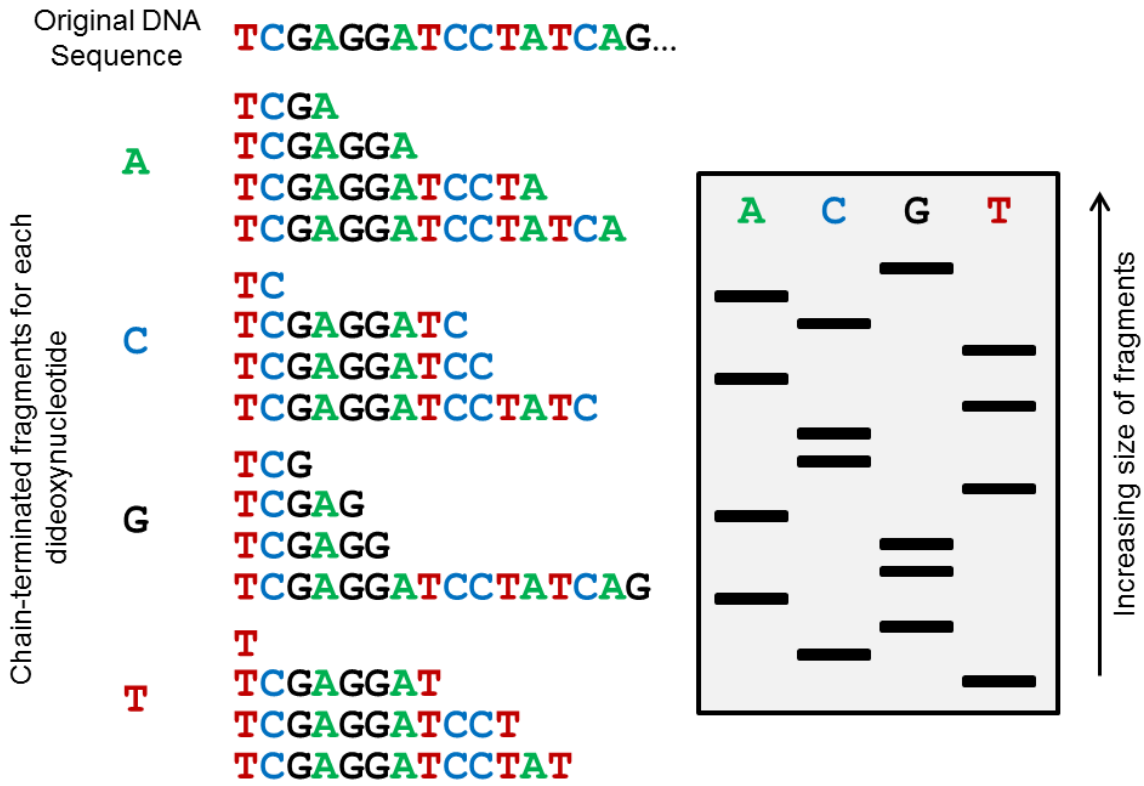


Figure 1.10 **Dideoxy chain-termination (Sanger) method of DNA sequencing. Dideoxynucleotides lack the 3' hydroxyl group required for extension of the DNA and therefore, by mixing a fraction of these with standard deoxynucleotides, chain termination is caused randomly at every position. If four separate reactions are carried out with each of the dideoxynucleotides in turn then the following fragments are generated for the example sequence. When separated by size the sequence can effectively be read out. Automated sequencing uses fluorescently labelled dideoxynucleotides that allow the DNA to be sequenced from a single reaction, rather than four separate reactions.**

A second generation of sequencing technologies has taken advantage of shotgun sequencing to vastly speed up and lower the cost of DNA sequencing. In the late 1990s

Pål Nyrén and colleagues developed a method that inferred nucleotide identity, by measuring pyrophosphate production, as each nucleotide is washed through the system, over template DNA bound to a solid support[39]. This method allowed for real-time observations rather than lengthy electrophoreses, was licensed by 454 Life Sciences and evolved into the first major next generation sequencing (NGS) technology, allowing the mass parallelisation and sequencing of a much greater amount of DNA.

The development of low-cost, rapid NGS technologies has continued in subsequent decades allowing for production of huge amounts of sequencing data. The first human genome took years to sequence and is estimated to have cost roughly $2.7 billion, whilst the same results can now be achieved in days for under $1000, enabling the application of DNA sequencing in a clinical context[40].

However, these technologies generally have greater error rates (~0.1-15%) and shorter reads (35-700bp) than traditional Sanger sequencing methods, which can be problematic. DNA hybridisation techniques have shown that large genomes often have repetitive and complex regions, which are difficult to resolve using short-read sequencing approaches[41]. Large scale structural variations, such as duplications, deletions, insertions, inversions and translocations and ranging typically from one kilobase to many megabases have been linked to several genetic traits, as discussed previously, underlining their importance. Long-read sequencing approaches can overcome some of these issues as they deliver reads in excess of several kilobases. If these long reads can span regions of structural variations, then the sequence can be unambiguously constructed.

The most widely used approach currently for long read sequencing is single-molecule real-time (SMRT) sequencing, commercialised by Pacific Biosciences[42]. This has thousands of individual wells, with transparent bottoms, known as zero-mode waveguides. A DNA polymerase is fixed to the bottom of the well and guides the DNA through the zero-mode waveguide. As a single labelled nucleotide is incorporated a camera records the light emitted, which allows for the sequence to be read.

Another type of long read sequencer is a nanopore sequencer, commercialised by Oxford Nanopore Technologies[43]. This doesn't use the incorporation of nucleotides to detect the sequence, unlike most other platforms, but can be used to directly read the sequence for single-stranded DNA. Strands are passed through a protein pore and as they are the DNA modulates this current that is set across the pore. Shifts in the voltage are characteristic of the DNA sequence and by training the data against known sequences, the sequence of an unknown fragment can be inferred. Large repeats of a single base or base modifications remain challenging to detect however[30].

Currently these long-read sequencing approaches have considerably lower throughput and higher cost than short-read NGS approaches, which has limited their adoption[30]. One major issue is the single-pass error (i.e. error for a single read), which approaches 15% for SMRT. However, since these are randomly distributed a consensus sequence is still reliable, and these platforms seem ideal for *de novo* genome assembly. Single molecule real time sequencing has also been used successfully to fill gaps within the human genome reference sequence[44].

Another method to help understand structural and copy number variation is to return to restriction mapping and more recent applications of similar approaches.

1.3 Optical Mapping

1.3.1 Origins of optical mapping

Optical restriction mapping was first reported in the mid-1990s by the Schwartz laboratory[45]. In the first mapping experiment a restriction digest (as used in restriction mapping) of elongated individual molecules was visualised after fixation in agarose gel. The DNA fragments could be sized and mapped back to the genome based on the known restriction pattern (Figure 1.11).

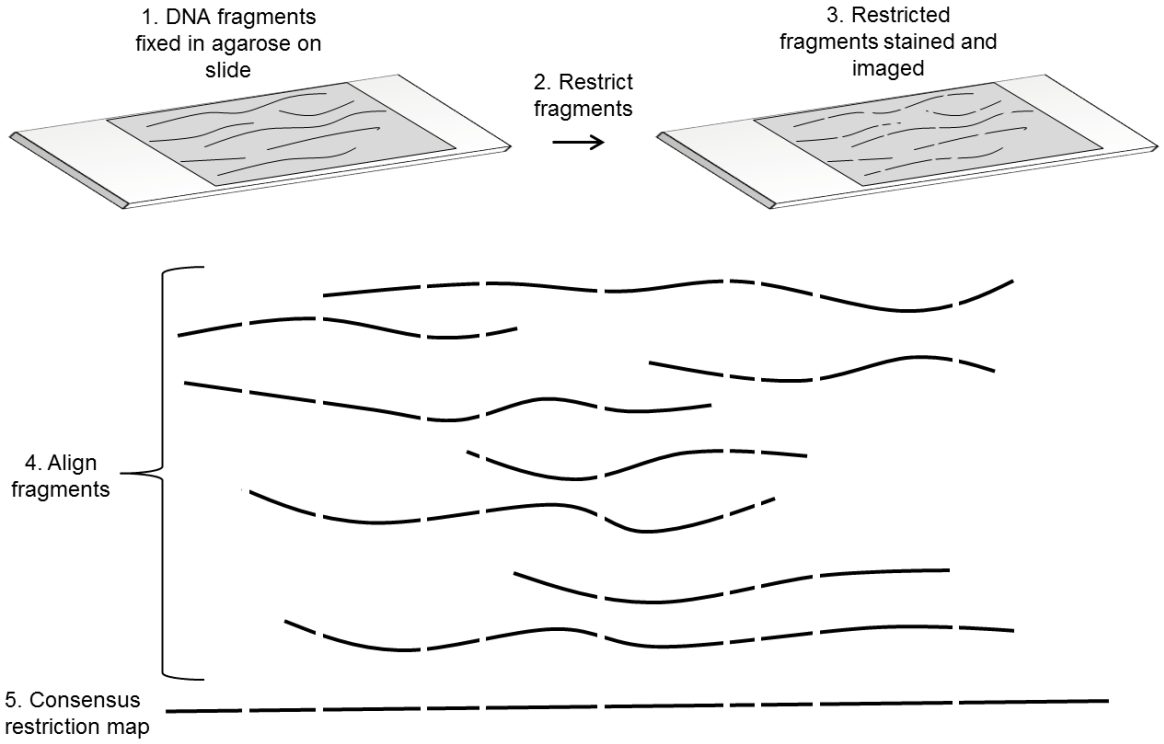


Figure 1.11 Optical restriction mapping procedure. DNA fragments are elongated and fixed in agarose gel, typically on a glass slide. Restriction enzymes are used to digest fragments at specific sites, so when fragments are imaged they can be sized and mapped back to a reference restriction pattern.

The optical restriction map contains long-range information that is complementary to the base-level resolution obtained by NGS methods and can effectively be used as a scaffold for *de novo* genome assembly and gap filling[46,47]. However its use is limited since it is still relatively low throughput, for instance it can take approximately 1 month to scan a human genome image by image[48]. There are also issues with the inherent limit of the minimum size of molecules and significant sizing errors for small fragments.

There have been two broad approaches used to improve the throughput and application of optical mapping: nanofluidic devices[49] and molecular combing of DNA.

1.3.2 Optical Mapping in nanofluidic devices

DNA is a large polymer with a persistence length of around 60 nm[50]. In solution, a free DNA molecule will coil to minimise the free energy of the system, with a size given by the radius of gyration (Figure 1.12A). If the DNA is introduced into a channel it will act in three ways depending on the size of the channel, the persistence length and the radius of gyration. If the channel is larger than the radius of gyration the DNA will act as it does in solution. If the channel is smaller than radius of gyration but still much larger than the persistence length, then the DNA will act like a string of non-interacting blobs. Finally, if the channel is smaller than the persistence length the DNA will be completely uncoiled (Figure 1.12A). If the channels are of the order of the persistence length then the DNA molecule will be confined and extended to a length proportional to the length of the molecule (Figure 1.12B)[50].

DNA can be visualised in nanochannels by fluorescent molecules that bind to DNA. These include dyes that will bind to the major or minor groove in DNA (e.g. DAPI) and dyes which will insert between the stacked DNA base pairs, known as intercalators. One

commonly used intercalating dye is YOYO-1 which has a high binding constant, is highly fluorescent when bound to DNA, but practically non-fluorescent when free in solution. DNA can also be labelled to produce a fluorescence pattern dependent on the underlying DNA sequence (see later).

By visualising dyes that bind to DNA the size of large, individual, DNA molecules can be measured and the location of labels on DNA can be correlated to the DNA sequence. The main advantage of these devices is their flexibility and ease of use to image individual stretched DNA molecules. However, the thermal motion of DNA means the fluorescence pattern along the intensity profile is blurred during imaging. This must be corrected by recording movies of the intensity profile, a stack of which is known as a kymograph (Figure 1.12B). Individual intensity profiles from each frame can be aligned to produce a consensus of the fluorescence pattern[51], although this results in longer image acquisition and a loss in resolution.

DNA nanofluidics were used by Riehn *et al*. in 2005 for visualisation of DNA in restriction mapping[52]. The DNA was confined in nanochannels and the restriction carefully controlled by electrophoresis and the diffusion of magnesium ions and EDTA, to prevent restriction before the DNA entered the channels. The restriction sites of SmaI, SacI and PacI were mapped with a precision of around 1.5 kbp.

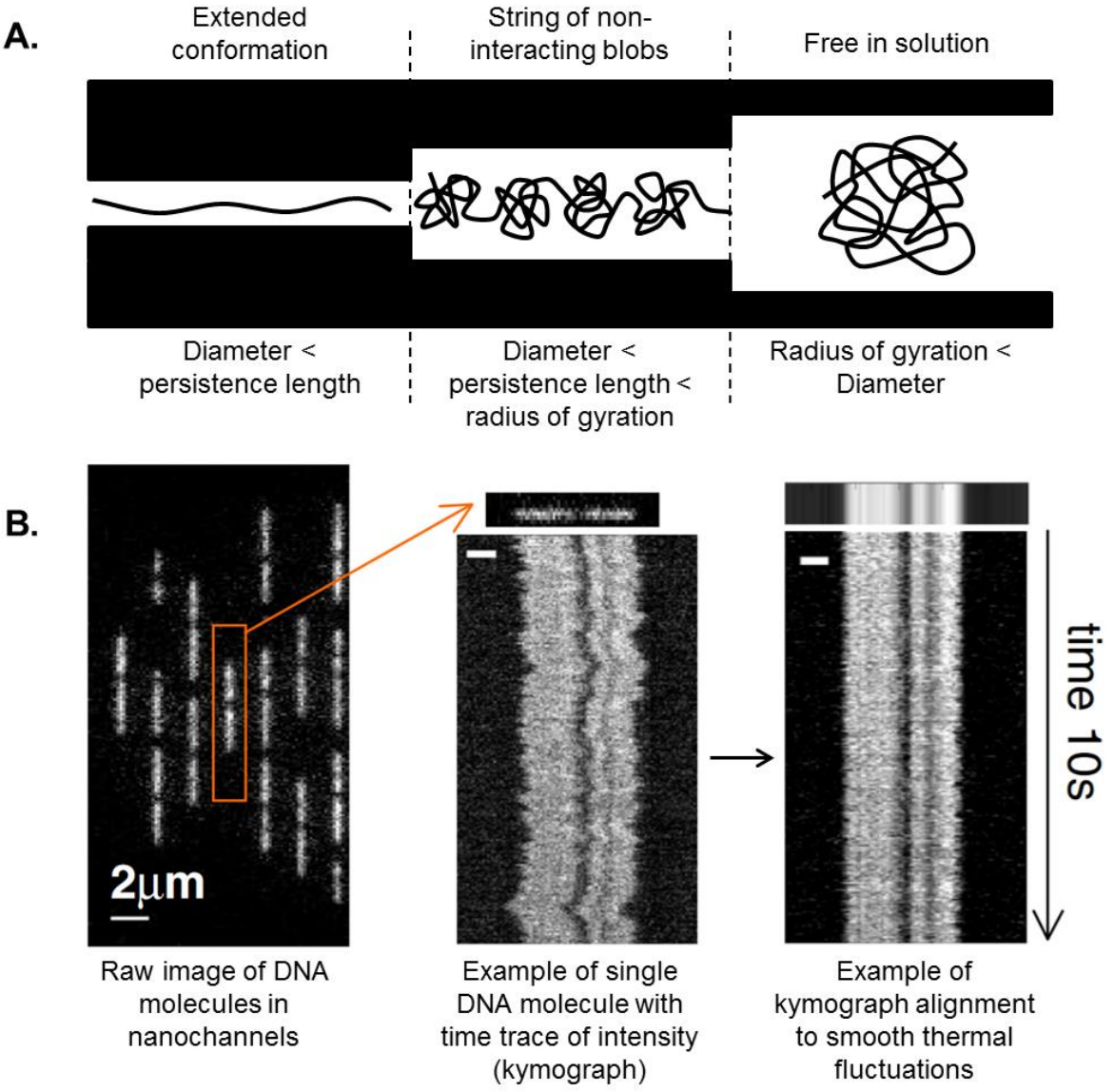


Figure 1.12 **Nanofluidic devices for optical mapping of DNA molecules. A) The behaviour of DNA in channels is dependent on the size of the channel, the persistence length and the radius of gyration of the DNA. If the channel is smaller than the persistence length the DNA will be completely uncoiled. If the channel is smaller than radius of gyration but still much larger than the persistence length, then the DNA will act like a string of non-interacting blobs. If the channel is larger than the radius of gyration the DNA will act as it does in solution. B) Example of optical mapping in nanochannels, adapted from Reisner *et al*[53]. DNA molecules are stained for affinity mapping and confined in nanochannels. A single DNA molecule is highlighted and a time trace of intensity (kymograph) is shown, showing thermal fluctuations of the molecule. The kymograph is aligned to produce a single average intensity profile that can be used for optical mapping.**

1.3.3 Optical Mapping by molecular combing

The deposition and linearization of DNA on a surface is used for fiber-FISH and can also be used for optical mapping of DNA molecules. A detailed experimental and theoretical description was given by Bensimon in the mid 1990s[54]. Typically, DNA is deposited on a surface carrying a net charge, for instance poly-L-lysine, or a surface coated with a hydrophobic compound, for instance PDMS. At low pH DNA molecules will adsorb strongly and non-specifically to the surface, whilst at high pH they will adsorb weakly. In between these extremes, at around pH 6, DNA will bind strongly and specifically to the surface at its extremities[55]. In molecular combing a receding meniscus is used to uniformly stretch the DNA across the surface using this phenomenon (Figure 1.13).

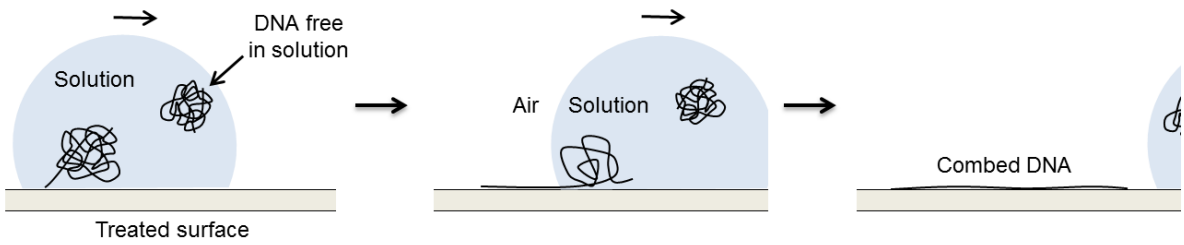


Figure 1.13 Principle of molecular combing of DNA. At around pH 6, DNA in solution will bind strongly and specifically at its extremities to a hydrophobic or positively-charged surface. As the air-solution interface is moved DNA is stretched uniformly on the surface, perpendicular to the receding meniscus.

Ideally molecules are deposited in an absolutely linear and uniformly stretched manner. This enables the length of DNA fragments to be easily extracted (for instance for optical restriction mapping) or a fluorescent pattern along the DNA to be extracted (see later). Molecular combing has been studied extensively to achieve this aim, with the surface coating, pH, ionic strength and manner of deposition all being carefully controlled (Figure 1.14)[56,57]. An early example of its utility was improved restriction mapping using molecular combing techniques to stretch the DNA fragments, allowing for far more

accurate sizing[58]. Good molecular combing from droplets containing picograms of DNA is now possible, but it remains challenging to get ideal deposition of samples[57]. If there are any sheared DNA molecules, or other impurities, in the sample then they will also be deposited on the surface.

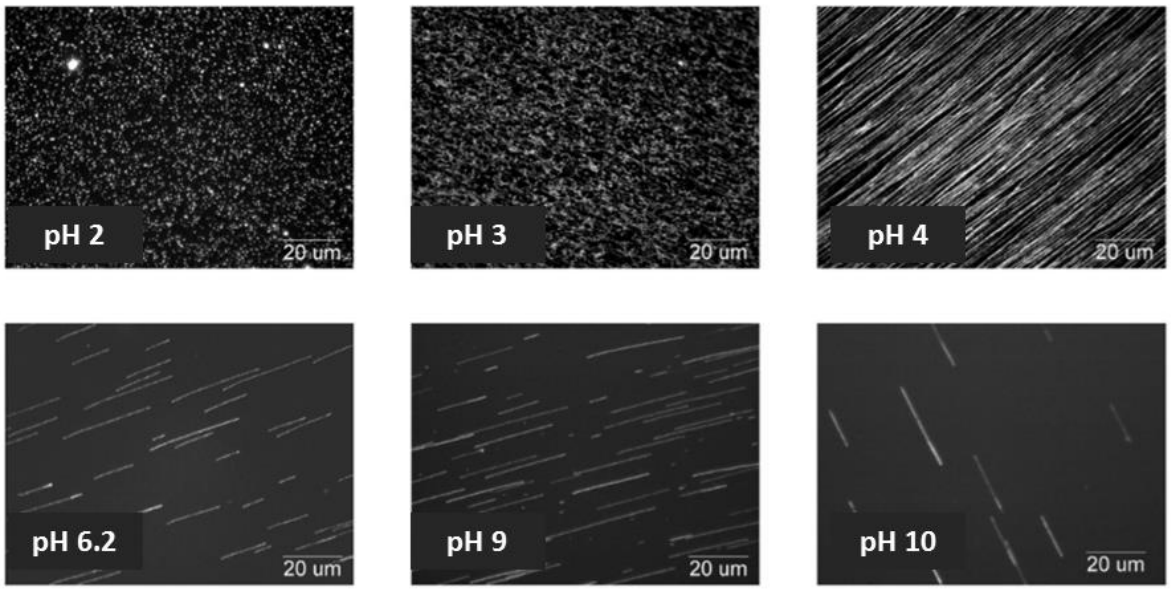


Figure 1.14 pH-dependent molecular combing of DNA on hydrophobic PDMS surfaces. Adapted from Benke *et al*[56]. At pH 2-3 adhesion is too strong to allow stretching of DNA molecules. At pH 4-10 a large, but reducing, number of molecules are adsorbed and stretched.

1.3.4 Labelling DNA for optical mapping

Alternatives to optical restriction maps label rather than cut the DNA, since if labels are added with high specificity, their location on the DNA can be used to map the DNA. This makes handling the DNA more straightforward, since the order of fragments does not need to be preserved and can also be used to overcome the inherent size limitations with restriction mapping. Combing, imaging and size estimation of small fragments (<1 kbp) is virtually impossible, since the fragments will only be a few pixels long. Hence restriction enzymes with a low density of sites are used, for instance an 8-site

recognition sequence occurs on average every 4^8=65,536 bp. However, this makes optical restriction mapping of small genomes (e.g. viral, bacterial, 50-5000 kbp) difficult since there will be few, or no, restriction sites for mapping. Higher density labelling can be used to overcome these issues and can be affinity-based, or enzymatic.

Affinity-mapping (Figure 1.15) generally uses the difference in hydrogen bonding between AT and GC base pairs. AT base pairs can form two hydrogen bonds, whilst GC pairs will form three hydrogen bonds (Figure 1.2), meaning GC-rich regions are more stable. In denaturation mapping the DNA is stained with an intercalating dye, which dissociates from AT-rich regions as they melt before GC-rich regions (Figure 1.15A). This will give a fluorescent pattern along DNA fragments, dependent on the GC-content and degree of melting. This was first demonstrated by Reisner *et al* in 2010 and an example is shown for lambda DNA in Figure 1.12B[53].

An alternative approach is to use a competitive inhibitor to bind to specific regions of the DNA (Figure 1.15B). This produces a fluorescent pattern if used together with an intercalating dye that is excluded from these regions by the inhibitor. For example affinity mapping has been demonstrated with netropsin which competitively binds to AT-rich regions[59] and actinomycin D which binds to GC-rich regions[60]. This mapping approach has been used to develop diagnostic techniques and has been applied in nanofluidic devices to identify bacteriophages[59], resistance plasmids[61,62], bacterial strains[63] and resistance outbreaks[64]. The main limitation of these approaches is that the information content is relatively low. For unique identification of a DNA fragment the fluorescent pattern must contain a number of contrasting peaks and troughs. However, for high density labelling based on GC-content there are relatively small intensity

modulations, which do not give well defined peaks and troughs (Figure 1.15E) and make assignment of the fragment, to a reference DNA sequence, difficult.

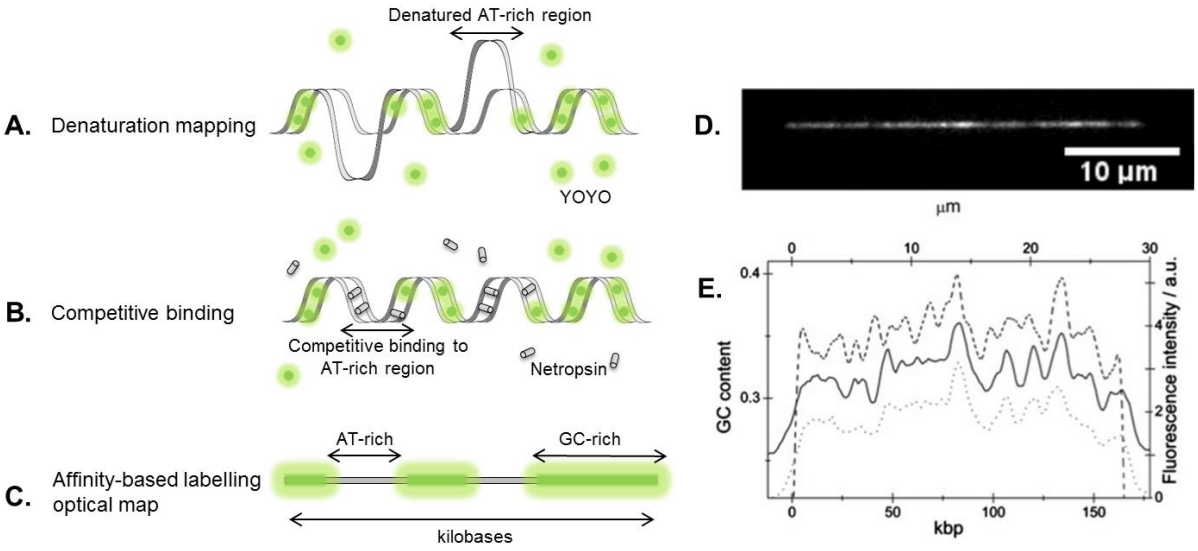


Figure 1.15 Affinity-based optical mapping of DNA. A) Denaturation of AT-rich regions causes dissociation of DNA intercalating dyes (e.g. YOYO). B) Competitive binding to specific regions of DNA (e.g. netropsin to AT-rich regions) prevents binding of intercalating dyes. C) Representation of the expected intensity profile from the approaches in A) and B). D-E) Example of affinity-based approach, taken from Nyberg *et al*[59]. D) A single T4 DNA molecule (166.5 kbp), stained using the competitive binding approach and mapped in a nanochannel. E) Traces of the intensity along the molecule: theoretical (dashed line), raw kymograph (dotted line) and aligned kymograph (solid line).

Enzymatic approaches can be used to increase the information content for reliable optical mapping. Enzymes can have high sequence specificity, therefore can be used to label with lower density which will improve the contrast between peaks and troughs in the intensity signal. These approaches include the use of nicking enzymes (Figure 1.16)[65]. In 2006 Xiao *et al*. used Nb.BbvCl to produce nicks (i.e. breaks in a single strand of the double-stranded DNA) at 5'-GCTGAGG-3' sites, followed by the incorporation of fluorescently-labelled nucleotides by DNA polymerase at the same sites. After combing on a surface, the labelled sites can be readily visualised and used for optical mapping.

The main limitation with this approach is the labelling of non-specific and naturally occurring nicks since the DNA polymerase will be unable to distinguish between these and nicks produced at specific sites. Fragmentation of DNA is also a problem when two nicks are close together.

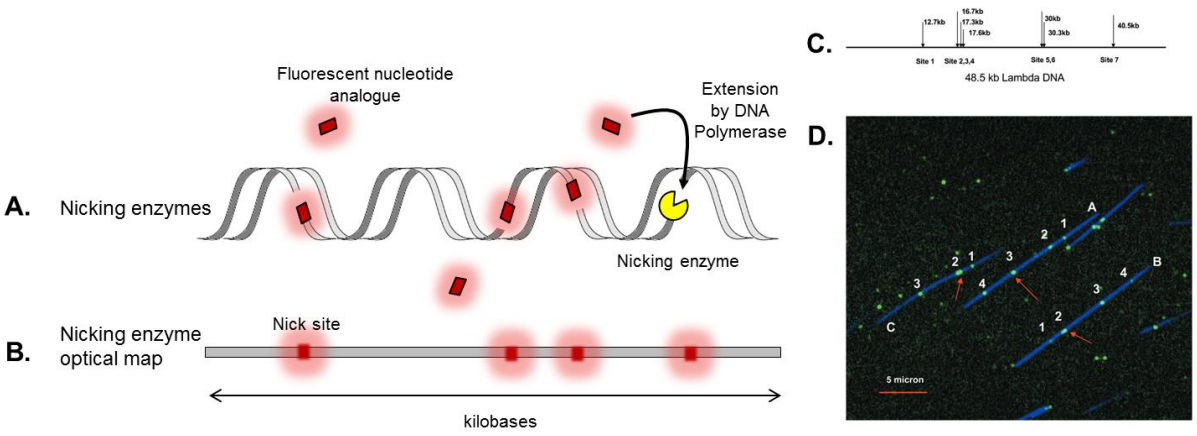


Figure 1.16 **Nicking enzyme approach for optical mapping of DNA. A) Nicking enzymes produce nicks at specific sites, followed by the incorporation of fluorescently-labelled nucleotides by DNA polymerase. B) Representation of the expected intensity profile. C-D) Example of nicking enzyme approach, taken from Xiao *et al*[65]. C) Expected nick sites for Nb.BbvCI (5'-GCTGAGG-3') and lambda DNA (48.5 kbp). D) Composite image of linearised lambda DNA (blue) and labelled nick sites (green). Four spots only, corresponding to clusters of the seven nick sites, are visible due to the diffraction limit. Fragments A and B are fully labelled, whilst fragment C has three spots visible.**

The first use of optical mapping using nicking enzyme was by Jo *et al.* in 2007 to map three BACs, labelled using the same method[66]. Das *et al*. used a different nick labelling scheme[67], where DNA polymerase lacking 5′-3′ exonuclease activity was used to incorporate fluorescently-labelled nucleotides as normal, but also leaving a short single-stranded flap of bases that could subsequently be targeted with fluorescently-labelled nucleotide probes. Therefore, as well as mapping nick sites, specific target sequences can be mapped, which has clear applications when only specific areas of large genomes are of interest. These approaches have been extended further to dual colour[47] (using two

nicking enzymes) and by using CRSIPR/Cas9 approaches to increase specificity for identification of individual genes[68].

An alternative enzymatic approach is to use methyltransferase-directed labelling, which can label DNA highly specifically, without damage and with high density. Neely *et al*. first reported optical mapping using methyltransferase-directed labelling in 2010[69]. For comparison, in optical restriction maps, the density of modified sites is typically only one site per 10-100kb, whilst the average density for M.TaqI, which labels DNA at 5'-TCGA-3' sites, is one site every 256 bases. This is the labelling method that will be used for optical mapping methods in this research.

1.4 Methyltransferase-directed labelling of DNA

1.4.1 DNA methyltransferases

DNA methylation is found in many organisms ranging from bacteria and viruses to mammals[70]. The enzymes that undertake DNA methylation are known as DNA methyltransferases and fall into three groups, depending on the target of methylation and therefore the final product: C5-methylcytosine, N4-methylcytosine and N6-methyladenine (Figure 1.17A-C). All known classes of DNA methyltransferases use the cofactor S-adenosyl-L-methionine (AdoMet, Figure 1.17D) as the methyl donor and most are active as monomeric enzymes. Methyltransferases that transfer methyl groups to other substrates, including proteins, RNA or small molecules (e.g. histamine), are also common, but will not be discussed here.

DNA methyltransferases are the other part of the restriction-modification system in bacteria, mentioned previously in section 1.2.1 (restriction mapping). In bacteria, methylation and the respective restriction by the sister endonuclease occurs within the same DNA target, typically consisting of a palindromic sequence, up to 8 base pairs in length. When foreign DNA is introduced into a bacterium it will be cleaved by the restriction enzymes, whilst at the same site the bacteria will methylate its own DNA, protecting it from destruction. In eukaryotes, modification occurs predominantly at CpG sites and has more diverse roles, including gene regulation. There are now thousands of known methyltransferases with specific recognition sites[71].

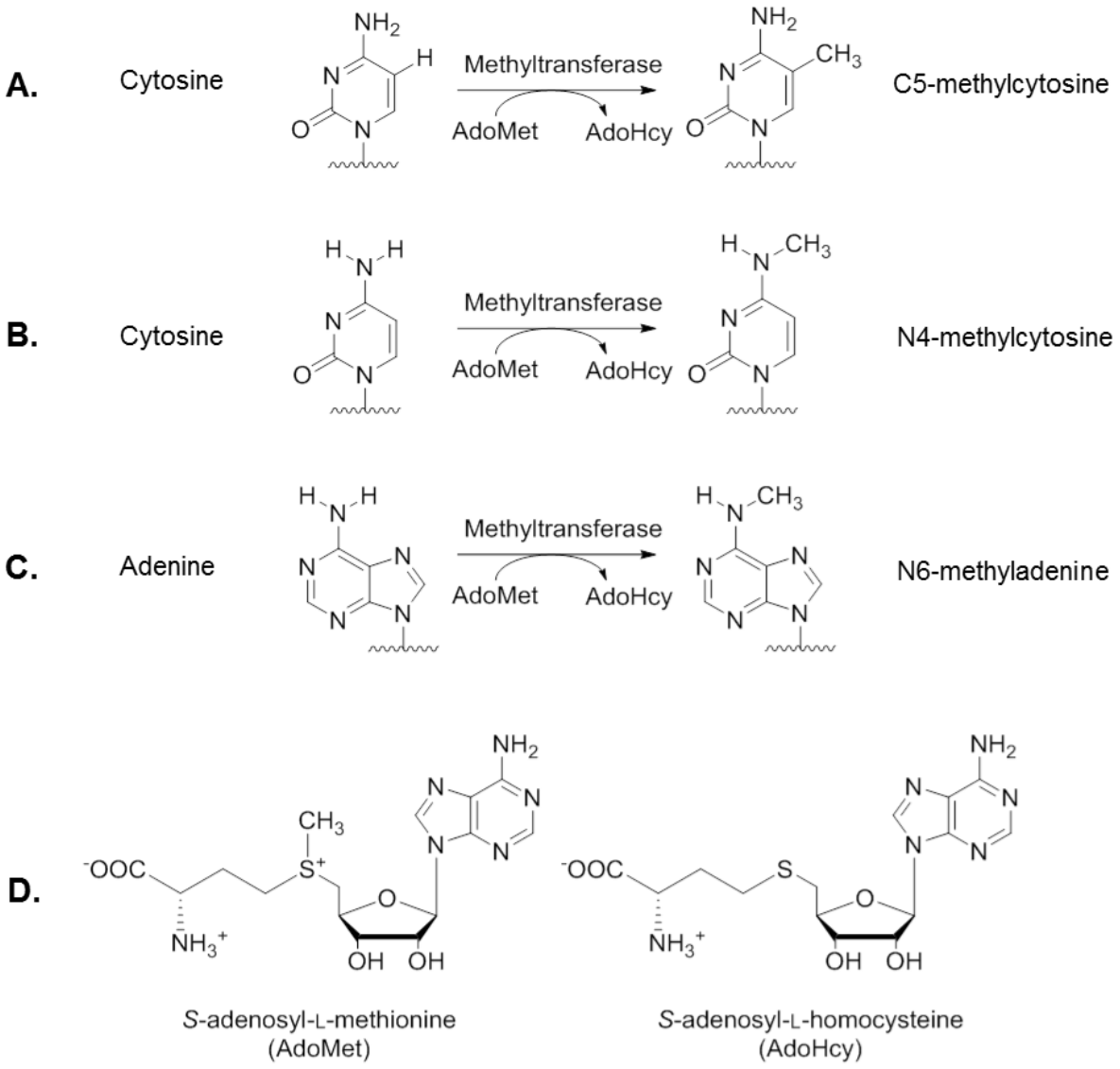


Figure 1.17 **The transfer of methyl groups to DNA by methyltransferases. A-C) Schematic representation of the types of DNA methylation catalysed by different groups of methyltransferases: A) C5-methylcytosine, B) N4-methylcytosine and C) N6-methyladenine. D) The structure of the substrate used for methylation, S-adenosyl-L-methionine (AdoMet) and the product, S-adenosyl-L-homocysteine (AdoHcy).**

1.4.2 Synthetic AdoMet analogues

Since methyl-transfer from AdoMet is highly specific and efficient, DNA methyltransferases are an attractive target for the transfer of functionalised groups to DNA. This can be achieved using synthetic AdoMet analogues, which transfer groups other than methyl groups and which can be divided into two main classes: aziridinoadenosines and doubly-activated AdoMet analogues.

Methyltransferase-directed modification of DNA with synthetic AdoMet analogues was first reported in 1998 by Elmar Weinhold's group, using aziridinoadenosines (Figure 1.18)[72]. A reactive aziridine group replaces the homocysteine moiety in AdoMet, allowing the whole cofactor to be covalently transferred to DNA. Subsequently in 2004 a fluorescent reporter group was attached to the adenine base in order to sequence-specifically label the DNA[73,74]. Propargyl substitutions on the 5'-N of the base have also been used to enable click chemistry[75]. However, aziridinoadenosines require a stoichiometric amount of enzyme for labelling DNA since the reaction product is a potent inhibitor of methyltransferases and therefore prevents rapid catalytic turnover. Also, the cofactors are highly reactive and may be unstable or cause non-specific labelling. For instance, aromatic nitrogen mustards are known DNA-alkylating agents.

In 2005 Dalhoff *et al*. first reported the synthesis of doubly-activated AdoMet analogues (Figure 1.18), with extended carbon chains which could be efficiently and specifically transferred by DNA methyltransferases[76,77]. It was known that ethyl and propyl groups were inefficiently transferred by DNA methyltransferases, however the addition of an unsaturated bond at the beta position to the sulfonium centre improved the efficiency of transalkylation.

This has been extended to allow the addition of various functional groups, including: amines[78], allowing conjugation with NHS-esters; alkynes[79–81] and azides[81], both capable of click chemistry; ketones[82], which can react with hydroxylamines or hydrazides; or directly coupled to fluorophores[83] . Therefore, there exists a wide variety of AdoMet analogues which can be used by methyltransferases to covalently attach a range of reactive groups to DNA, site-specifically and with high efficiency[84–86].

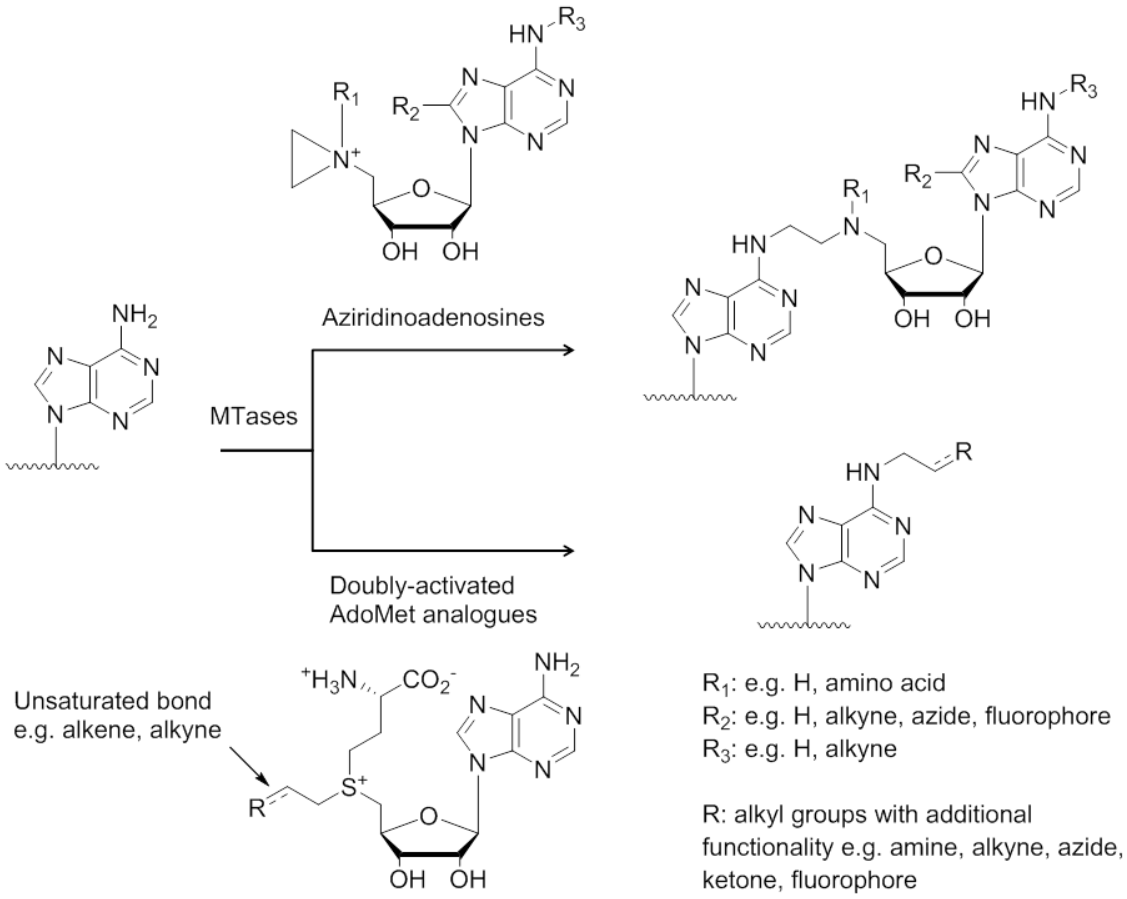


Figure 1.18 Overview of methyltransferase-directed labelling of DNA by synthetic AdoMet analogues. Labelling can occur using two main classes of molecule: aziridinoadenosines (top) and doubly-activated AdoMet analogues (bottom). Aziridinoadenosines will transfer the whole cofactor and require stoichiometric amounts of methyltransferase. In contrast doubly-activated AdoMet analogues transfer only the extended carbon chain in a catalytic manner. Various functional or reporter groups can be transferred to DNA bases by both classes of molecule.

1.4.3 Methyltransferases for labelling DNA

A number of DNA methyltransferases have been reported which can be used for the transfer of functional groups from AdoMet analogues (Table 1.2). This highlights the flexibility of this labelling approach, since these methyltransferases have different target sequences, two to six base pairs in length, which allows labelling density to be varied. For example, in the first report of doubly-activated AdoMet analogues a DNA methyltransferase of each class was used[76]: the DNA N6-adenine methyltransferase M.TaqI (5'-TCG<u>A</u>-3')[87]; DNA C5-cytosime methyltransferase M.HhaI (5'-G<u>C</u>GC-3')[88] and DNA N4-cytosine methyltransferase M.BcnIB (5'-C<u>C</u>SGG-3').

DNA methyl-transferase	**Target Sequence** (Modified base underlined)
M.TaqI	TCG<u>A</u>
M.HhaI*	G<u>C</u>GC
M.SssI*	<u>C</u>G
M.BseCI	ATCG<u>A</u>T
M.BcnIB*	C<u>C</u>SGG
M2.Eco31I*	GGT<u>C</u>TC
M.EcoRI	GA<u>A</u>TTC
M.HpaII	C<u>C</u>GG
M.XbaI	TCTAG<u>A</u>
M.FokI	GG<u>A</u>TG and C<u>A</u>TCC

* Mutant enzyme

Table 1.2 Alternative DNA methyltransferases for labelling of DNA. A number of different methyltransferases have been reported, varying in their efficiency and target sequence.

It was noted that a mutation within the AdoMet binding pocket of M.HhaI increased the rate of transalkylation[76], which suggested that engineering the cofactor binding sites can improve methyltransferase-directed labelling efficiency. If this could be applied more widely, to the thousands of known DNA methyltransferases, since many of them share structurally similar AdoMet binding pockets, this would allow for a whole library of

modifications of DNA at specific sequences. This also raises the prospect of *in vivo* modification of DNA, since it is possible a methyltransferase could be engineered to preferentially transalkylate using an AdoMet analogue, rather than naturally-occurring AdoMet.

This type of engineering has been demonstrated by the Klimašauskas group, who modified the cofactor binding pocket of M.HhaI by directed mutagenesis and systematic replacement of three non-essential positions with smaller residues[89,90], leading to substantial increases in activity. Analogous replacements of residues in DNA C5-cytosine methyltransferases M.HpaII (5'-C<u>C</u>GG-3') and M2.Eco31I (5'-GGT<u>C</u>TC-3') also showed catalytic activity could be obtained.

1.4.4 Applications of methyltransferase-directed labelling of DNA

One application of methyltransferase-directed labelling is the selective capture of DNA molecules (Figure 1.19). The enrichment of DNA samples is important for NGS methods, as there is such a large amount of sequencing data produced (see section 1.2.3). Also, DNA methylation in humans is associated with some cancers and so selective capture based on methylation state may have diagnostic uses. If DNA is selectively modified by methyltransferases it can be captured by the reactive group. For instance alkyne-modified lambda bacteriophage DNA has been captured using click chemistry onto silica-based beads[91] and amine- and azide-modified genomic DNA has been captured by conjugation with NHS- or DBCO-biotin respectively onto streptavidin-coated beads[92].

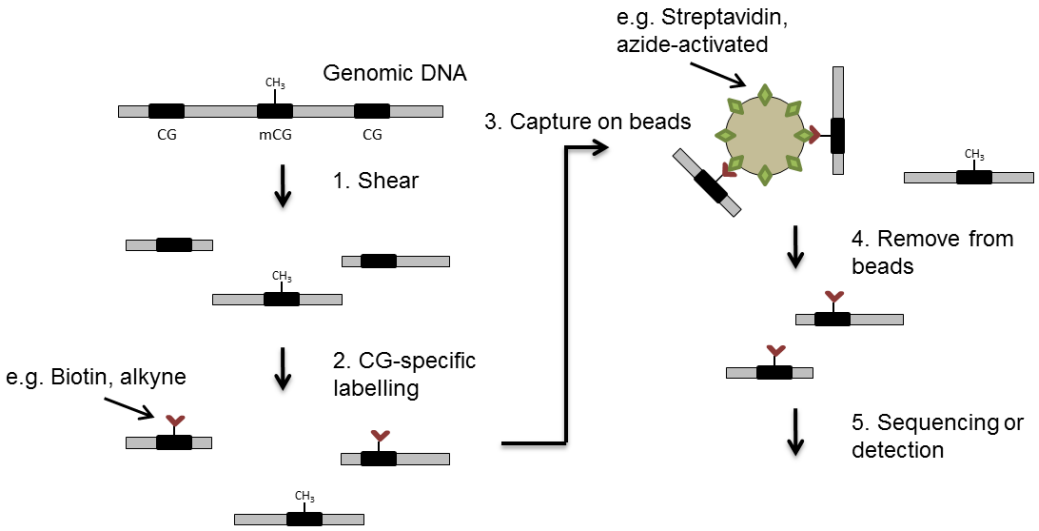


Figure 1.19 Procedure for DNA capture of unmethylated genomic DNA. Genomic DNA is sheared and labelled will functional groups (e.g. biotin, alkyne) at specific sites (e.g. CpG sites). If a site is already methylated, then no labelling occurs. Labelled fragments are captured onto activated beads (e.g. streptavidin or azide-coated) to separate from methylated fragments. The fragments are removed from beads prior to detection or sequencing.

Another application has been the use of methyltransferase-directed labelling for optical mapping, as previously discussed in section 1.3.4 (Figure 1.20). The main advantages over nicking enzymes are that the DNA is labelled without fragmentation, with greater efficiency and with greater density. Neely *et al*. first reported optical mapping using methyltransferase-directed labelling in 2010 with M.HhaI (5'-G<u>C</u>GC-3' sites)[69] and similar results have subsequently been reported with M.TaqI (5'-TCG<u>A</u>-3' sites)[93]. These applications used molecular combing to stretch DNA molecules, which enabled super-resolution of fluorophores to localise sites with 80 bp resolution (see 1.6.2 for a discussion of super-resolution microscopy). Following deposition the fluorophores are photobleached and based on the stochastic nature of this bleaching individual emitters can be localised[94]. Kim *et al.* have used mapping as a reference to locate protein-binding sites[95]. M.BseCI (5'-ATCGAT-3') was used to fluorescently label T7 DNA molecules with

an aziridinoadenosine cofactor, to which labelled T7 RNA Polymerase was bound. After combing on a poly-L-lysine surface, DNA molecules could be extracted and although the labelling was rather sparse, the methyltransferase sites could be used as a reference to localise promoters with five-fold greater accuracy than using distance measurements alone.

M.TaqI-labelled bacteriophage DNA has also been mapped using nanofluidic devices by the Ebenstein group[83]. Here bacteriophage DNA was identified by the intensity profile, however because of the relatively high density of labels and thermal motion, high resolution maps could not be obtained. More recently however the same group has demonstrated super-resolution mapping in silicon nanochannels, using a low density of labels and single molecule tracking, to overcome the effect of thermal motion[96].

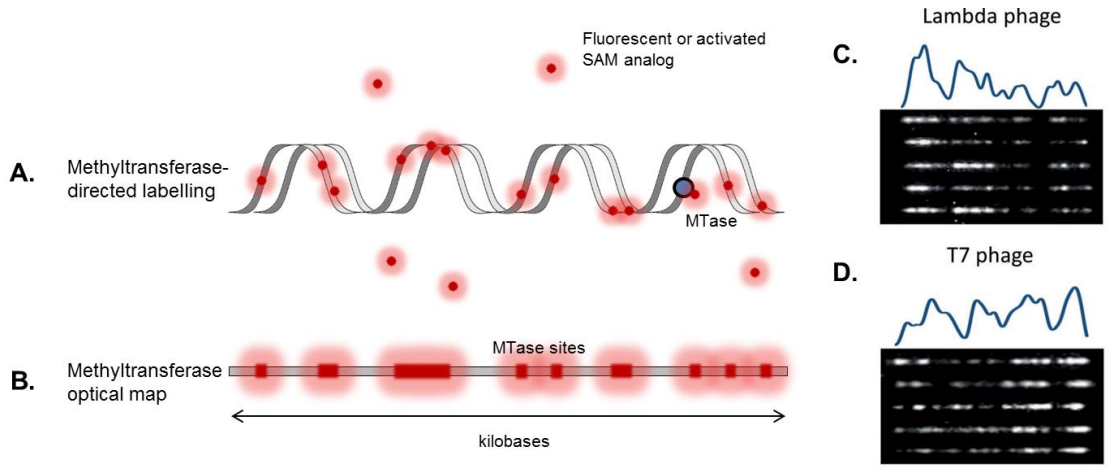


Figure 1.20 **Methyltransferase-directed labelling approach for optical mapping of DNA. A-B) Schematic of methyltransferase-directed approach for enzymatic-based labelling of DNA. A) Methyltransferase label DNA with fluorophores at specific sites, either directly or via coupling (e.g. amines and NHS esters, azides and alkynes via Click chemistry). B) Representation of the expected intensity profile. C-D) Example of methyltransferase-directed labelling approach, taken from Grunwald *et al*[83]. M.TaqI (5'-TCGA-3' sites) was used to directly transfer fluorophores to bacteriophage DNA, which were subsequently mapped in nano-channels. Expected intensity profiles and representative molecules are shown for: C) lambda (48.5 kbp) and D) T7 (40.0 kbp) bacteriophage DNA.**

The attachment of fluorophores to DNA can also allow its visualisation *in situ*, for instance in FISH, or in live-cell imaging (Figure 1.21). There are a number of approaches for fluorescent-labelling of plasmids[97], but methyltransferase-directed labelling gives an advantage over other approaches since it allows for non-destructive, covalent attachment of fluorophores at specific positions and in densities that can be readily controlled. Schmidt *et al.* used an aziridinoadenosine cofactor and M.TaqI to covalently label plasmids pUC19 and pRB322 with Cy3 dyes, before transfection of mammalian cells and visualisation of plasmids[98]. The transfection of similar plasmids in bacteria, rather than eukaryotic cells, has not been reported, but is a topic that could be explored further using this labelling technique.

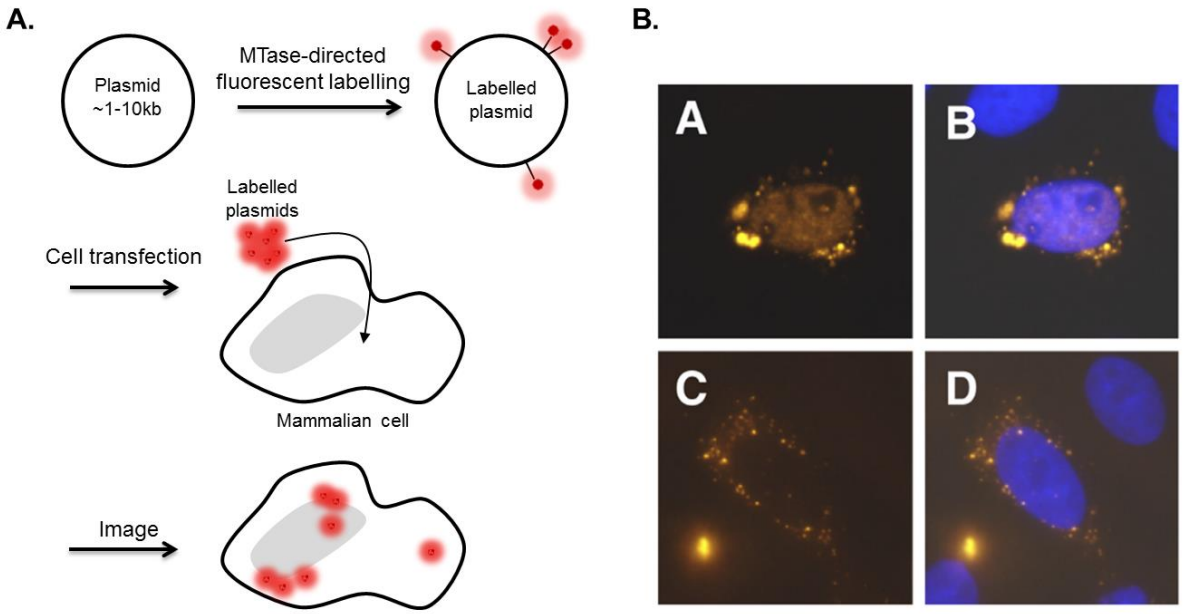


Figure 1.21 **Localisation of plasmids by methyltransferase-directed fluorescent labelling. A) General procedure for localisation of plasmids. Plasmids are fluorescently labelled by methyltransferase-directed labelling before cell transfection and imaging. B) Example of cells transfected with labelled plasmids, taken from Schmidt *et al*[98]. pUC19 (2686 bp) was labelled with Cy3 using an aziridinoadenosine cofactor before transfection into CHO-K1 cells. The plasmid (yellow) was found in the cytoplasm (A and C) and the nucleus (A only). Overlays of A and C with DAPI staining of the nucleus (blue) are shown in B and D respectively.**

1.5 Bacteria, plasmids and antibiotic resistance

1.5.1 Bacteria structure and overview

Bacteria are prokaryotic single cell organisms which do not contain any membrane-bound organelles, in contrast to eukaryotic cells. They were among the first lifeforms on Earth, are typically a few microns in length and come in a variety of forms [99]. The small size of cells and lack of membrane-bound organelles belies the complexity of the underlying subcellular architecture of bacteria, which are highly ordered and dynamic cells[100].

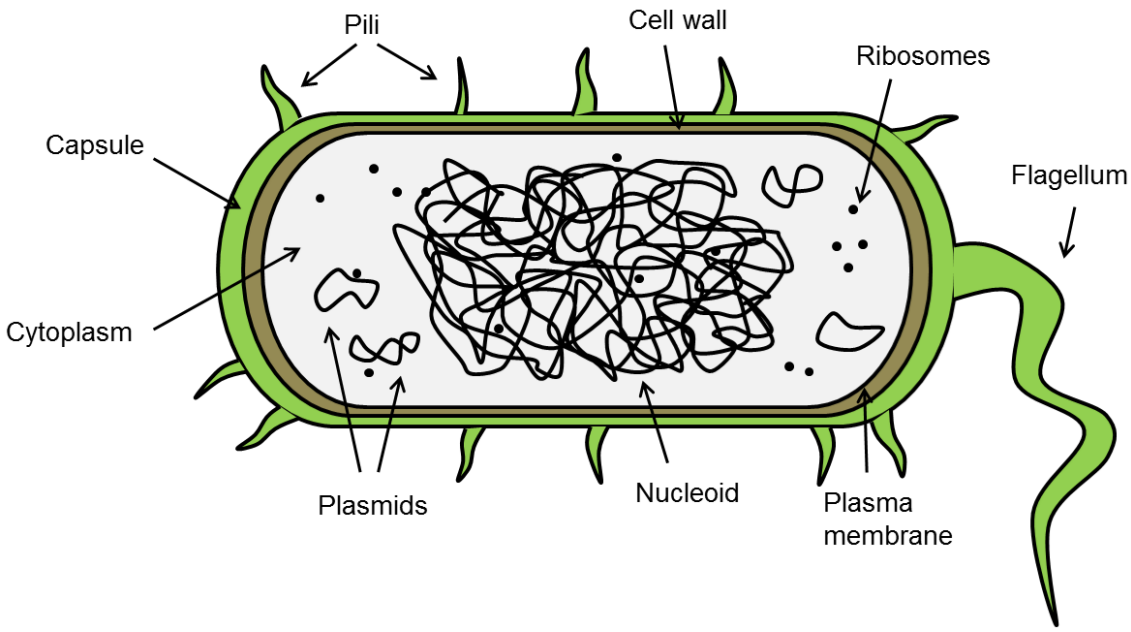


Figure 1.22 **Typical structure of a bacteria cell. The cytoplasm, where most essential functions are carried out, is enclosed by a cell envelop composed of several layers: a plasma membrane; a peptidoglycan cell wall; and in some species a capsule composed of polysaccharides. In addition, many species have hair-like projections, known as pili, that aid attachment and some have flagella that aid movement. The cytoplasm contains the genetic material and ribosomes for protein synthesis. Many bacteria have a single large, circular genome which is localised in the nucleoid as well as small extrachromosomal pieces of DNA called plasmids.**

The basic features of bacteria are shown in Figure 1.22. However, like eukaryotic cells, they also have a range of cytoskeletal proteins and intercellular signalling systems to coordinate growth, as well as to localise proteins and DNA to specific subcellular localisations at specific times. Apart from chromosomal DNA, which is responsible for most cellular functions, bacteria can contain DNA in the form of plasmids.

1.5.2 Plasmids and antibiotic resistance

Plasmids are small pieces of DNA that are found naturally in bacteria and have essential roles in metabolism, pathogenesis and resistance. The term 'plasmid' was first used by Lederberg in 1952 to describe any extrachromosomal genetic particle[101,102]. They are separate from the chromosome and capable of replicating independently. Plasmids are usually circular, the copy number (i.e. the number of copies in a cell), can range from a few copies to hundreds, whilst the size can range from around 1 kbp to 100 kbp[103].

Plasmids carry genes that include those which promote replication, maintenance and proliferation of the plasmid, but also genes that help the host to adapt to the environment. Plasmids are capable of being transferred between bacteria, (horizontal gene transfer), as well as between generations of bacteria (vertical gene transfer) and are dynamic enough to control the expression of genes that may only be useful transiently. One of the most important classes of gene spread by plasmids is genes for antibiotic resistance.

Antibiotics are antimicrobial drugs used to treat bacterial infections. Paul Ehrlich first developed a drug to specifically target disease-causing microbes in 1910, to treat syphilis[104]. In 1929 Fleming discovered penicillin[105], which was successfully synthesised in 1940[106] and went into mass production and distribution in 1945, revolutionising

medicine in the 20th century. However even before the extensive use of penicillin, it was observed that some bacteria could use enzymes to degrade it, therefore developing antibiotic resistance[107].

Generally antibiotic resistance is conferred by the conjugation of mobile genetic elements, the most important of which are transposons and plasmids[108]. Genes are expressed from these genetic elements and the proteins that are synthesised are used in a variety of strategies to confer resistance. These strategies include modification, destruction, removal and reduced uptake of the antibiotic molecule and modification and protection of the target site of the antibiotic.

Antibiotic resistance is now one of the greatest public health threats and by 2050 it has been estimated that the societal and financial cost, if not tackled, will be US$100 trillion[109]. As well as developing new treatments and novel antibiotics, preventive measures are recommended, such as public awareness, good sanitation and reducing and controlling the use of antibiotics in healthcare, agriculture and the environment. Also recommended is the development of new rapid molecular diagnostics to quickly and accurately diagnose infections and inform early appropriate antibiotic therapy[110]. This would improve clinical outcomes, and reduce antibiotic use, limiting the selection of antibiotic-resistant bacteria.

Current molecular diagnostics take around two days, as bacteria are cultivated and then identified[111]. Meanwhile the patient is treated based on empirical observations and the likely pathogens. This obviously leads to ineffective treatment and inappropriate use of antibiotics. Rapid diagnosis by biomarkers or identification without the need for

cultivation would accelerate diagnosis. Optical mapping therefore has the potential to be applied as it needs only a small amount of pure DNA.

The mechanisms of horizontal and vertical gene transfer are also being investigated in this context, to help fully understand mechanisms for development of antibiotic resistance[112].

1.5.3 Plasmid organisation and dynamics

To ensure plasmid transmission to daughter cells bacteria have active partitioning systems. There are a number of well-characterised partition system which share similarities[113]. The Type I and II are the most common partition systems, both involving three components, an adaptor protein, a motor protein, and a centromere (a specific sequence on the plasmid). The adaptor protein binds to the centromere and recruits the motor protein, leading to filament formation and either the plasmids are pulled apart (type I) or pushed toward opposite ends of the cell (type II).

It is thought that no such active partition mechanisms are involved for high copy number plasmids. Instead it is thought they are randomly segregated during division, meaning by chance each daughter cell should retain at least a single copy[114]. However there is some debate about this mechanism, as microscopy of fluorescently-labelled plasmids has shown clustering at the poles of cells[115,116]. This has been attributed to displacement by the nucleoid[117] although there is also evidence of plasmids randomly distributed throughout the cytoplasm, observable by localisation of individual plasmids (Figure 1.23)[118]. It has also been shown that localization can be driven by other processes, for instance transcription[119].

The investigation of plasmid localisation, clustering and dynamics showcases how fluorescence microscopy, and its recent advances beyond the diffraction limit, has become an invaluable tool for studying biological systems. Figure 1.23 shows examples of how both conventional widefield fluorescence microscopy (Figure 1.23B and E) and super-resolution microscopy (Figure 1.23C and F) can be used to investigate a biological system. These types of fluorescence microscopy are also vital for optical mapping of DNA and so the principles will be discussed in more detail.

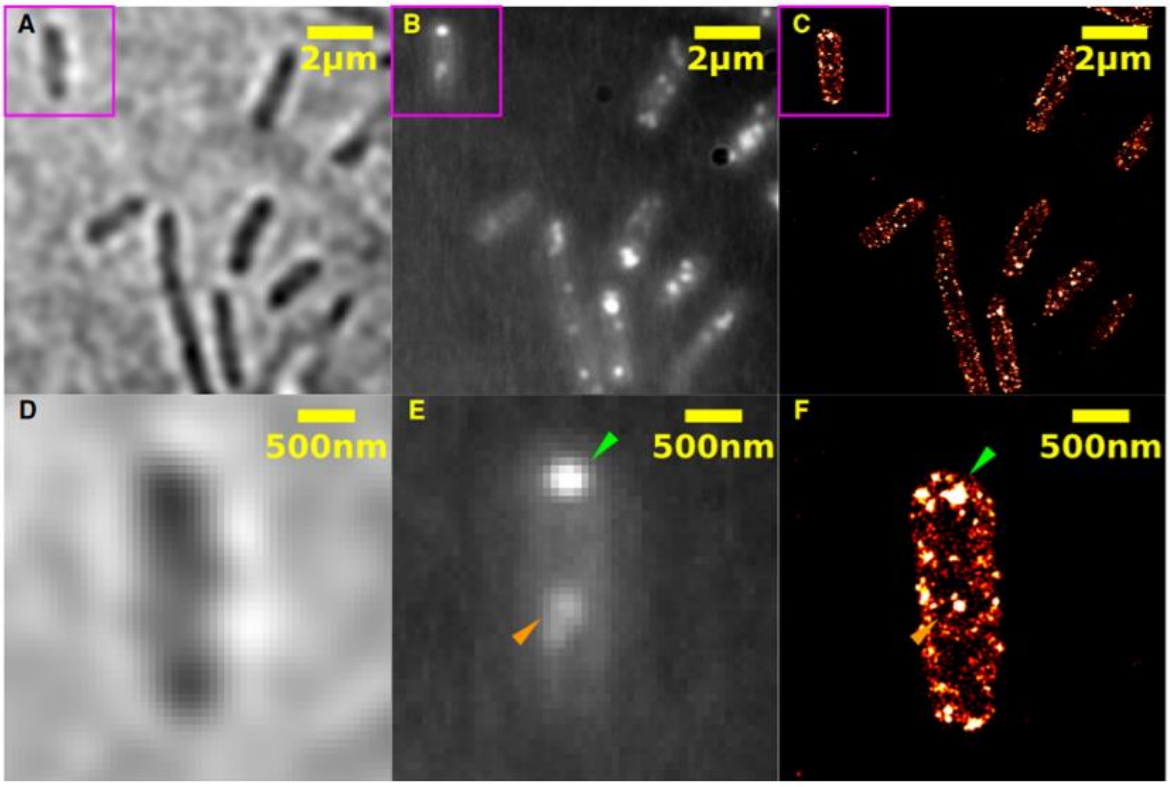


Figure 1.23 **Plasmid localisation in bacteria. Taken from Wang *et al*[118]. Single-molecule fluorescence in situ hybridization (smFISH) was used to localise a ColE1-derivative plasmid in fixed and permeabilised Escherichia Coli. A) Immobilised bacteria viewed in brightfield. B) Conventional widefield fluorescence microscopy. Fluorescent foci are clusters of smFISH-labelled plasmids. C) E. Coli imaged by localisation microscopy. D-F) Show an enlargement of the highlighted region in A-C). In the localisation image, as well as clusters of plasmids around the poles of the cell, there are plasmids located within the nucleoid region.**

1.6 Fluorescence and Microscopy

1.6.1 Fluorescence and the emission of light

Fluorescence was first described by George Stokes in 1852[120]. He coined the term when he observed that the mineral fluorspar emitted red light when illuminated by UV light. Fluorescence is a type of luminescence (a termed introduced by Eilhard Wiedemann in 1888[121]), which is the general term for emission of UV, visible or infrared photons from an electronically excited species.

The processes involved in luminescence are illustrated in Figure 1.24[122]. In a molecule, electrons occupy orbitals which exist at discrete energy levels (many vibrational levels are associated with each orbital), the lowest of which is known as the ground state, S_0. A pair of electrons in an occupied orbital will have the opposite spin, meaning the total quantum spin is zero, which is termed a singlet state. When light interacts with matter it is either scattered or absorbed. If absorbed, a photon can promote an electron from S_0 to one of the vibrational levels in an unoccupied orbital, for example the first singlet excited state, S_1. In principle the electron spin is unchanged, so the transition will be singlet-singlet. However, if the election undergoes conversion to another state and changes spin, the overall quantum spin is one and the state is known as a triplet state, e.g. first excited triplet state, T_1.

The excited species is de-excited by a number of possible processes: including internal conversion, fluorescence, intersystem crossing and phosphorescence. Internal conversion (IC) is a non-radiative transition between two electronic states of the same spin multiplicity. If this is from S_1 to S_0 then vibrational relaxation towards the lowest vibrational level of S_0 can occur without emission of light. In fluorescence the S_1 to S_0

relaxation is accompanied by the emission of a photon. Usually this occurs from the lowest vibrational level of S_1.

Another process for de-excitation is intersystem crossing (ISC), a non-radiative transition between electronic states of different multiplicities, for instance from S_1 to T_1. This is in principle forbidden, however coupling between the spin magnetic moment and the orbital magnetic moment can be large enough to make such a transition possible. From T_1 a molecule may be de-excited by further ISC and vibrational relaxation, or a radiative de-excitation in a process called phosphorescence.

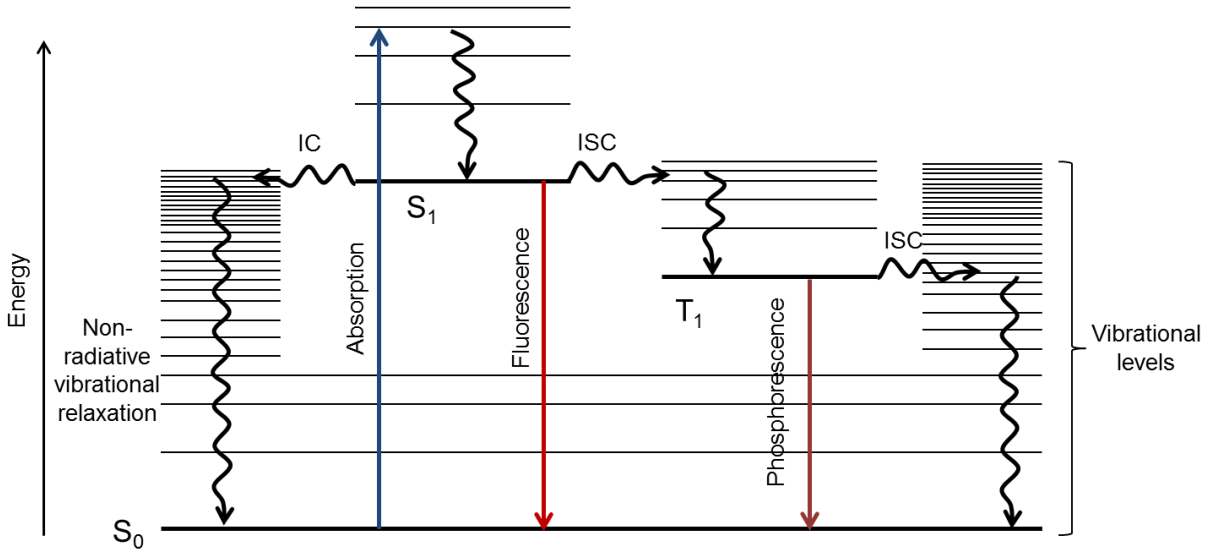


Figure 1.24 **The energy states and transfers involved in photoluminescent processes. In a molecule, electrons occupy orbitals which exist at discrete electronic energy levels, including the ground state (S_0), excited singlet states (S_1, S_2...) and excited triplet states (T_1, T_2...). Each electronic state will have many vibrational levels. When absorbed a photon can promote an electron from S_0, to one of the vibrational levels in an unoccupied orbital, for example the first singlet excited state, S_1, from which vibrational relaxation towards the lowest vibrational level can occur. Internal conversion (IC) is a non-radiative transition between two electronic states of the same spin multiplicity (e.g. S_1 and S_0). Intersystem crossing (ISC) is a non-radiative transition between electronic states of different multiplicities (e.g. S_1 and T_1). Fluorescence is a radiative transition from an excited singlet state to the ground state (e.g. S_1 to S_0). Phosphorescence is a radiative transition from an excited triplet state to the ground state (e.g. T_1 to S_0).**

Some of the main factors that should be considered when choosing a luminescent probe are: the maximum excitation and emission wavelengths; the difference between them (Stokes shift); the molar absorption coefficient; the quantum yield; and the lifetime and photostability of the probe.

The maximum excitation and emission wavelengths are determined by the energy levels of the system. For instance, in linear and cyclically conjugated systems it is usually the overlap between π-orbitals that determines the wavelength of absorption. Greater conjugation can therefore lower the energy of the transition and consequently lengthen the maximum excitation wavelength. Typically, fluorophores are selected that are suitable for the microscope setup (e.g. the wavelength of the excitation laser) and which are excited far from background fluorescent species. For instance, most cells exhibit a natural fluorescence from a range of species (e.g. flavins, NADH, collagen), termed 'autofluorescence', that must be overcome. Practically a large Stokes shift also usually makes it easier to distinguish fluorescent species.

The brightness of a fluorophore depends on its ability to absorb and emit photons. The ability to absorb photons is described by the molar absorption coefficient, which is the amount of light absorbed, at a given wavelength, depending on the concentration of the fluorophore. The efficiency of photon emission is described by the quantum yield, which is the number of fluorescence photons emitted per excitation photon absorbed. Fluorescence brightness is proportional to the product of the molar absorption coefficient and quantum yield, so usually high absorption coefficient and quantum yields are desirable.

Another factor is the lifetime of the excited state. Non-radiative de-excitation processes will compete with fluorescence when they take place on a timescale comparable with the lifetime of the excited state. This means for example fluorescence is typically more intense than phosphorescence, as the lifetime is much shorter. This is because in phosphorescence the emission occurs from a triplet state to the singlet ground, which is a 'forbidden' transition and therefore kinetically slow.

Under a high intensity of photons, the limiting factor is often the photostability of a fluorophore. If fluorophores were infinitely photostable then even dim fluorophores could be used with very high laser powers and long excitation times. However, in reality fluorophores are eventually destroyed, in a process known as photobleaching, which reduces the signal to noise ratio. This process can also be exploited to probe dynamic processes, for instance by using the recovery of fluorescence due the diffusion of fluorescent molecules, or for localisation (see section 1.6.2). During photobleaching the fluorophore is irreversibly destroyed and will no longer fluoresce. This usually involves a permanent structural change from the excited state, for example conversion to a triplet excited state and subsequent destruction of covalent bond, for instance by oxygen free radicals[123,124].

The choice of fluorophore is dependent on the application and careful consideration of these parameters. Generally, for imaging in biochemical applications fluorophores should be: excitable at a suitable wavelength; detectable at a suitable wavelength; bright; photostable; soluble and easily synthesised. The main types of fluorophores that are used to fulfil these requirements are small organic molecules[125], fluorescent proteins[126] and quantum dots[127].

The first fluorophores based on small organic molecules was quinine sulfate, identified by John Herschel in 1845[128]. Since then a whole range of organic fluorophores have been developed, many based on fluorescein[129], rhodamine[130], BODIPY[131] and cyanines[132] (Figure 1.25). Development of derivatives and novel small organic molecules is ongoing, but due to the difficulty of predicting photophysical properties, rational design has proved difficult. However, there are many dyes widely available across the UV/vis spectrum, capable of a range of easy coupling chemistry and which are relatively bright and photostable. Also, their size means they are unlikely to interfere with biological function.

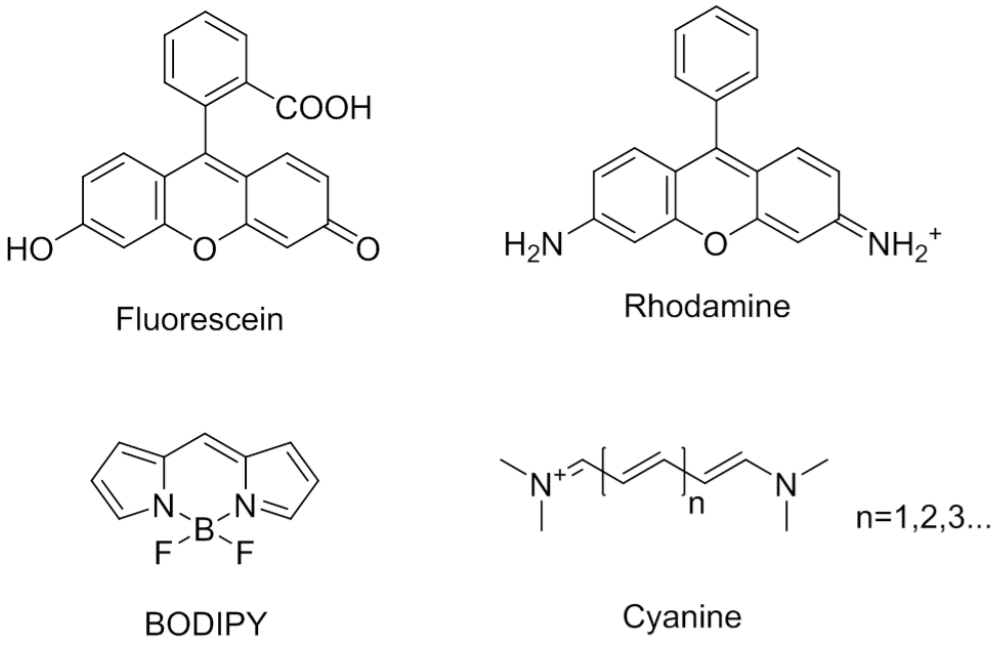


Figure 1.25 Common organic fluorophores. Many commercially available small organic fluorophores are derivatives of these four molecules: fluorescein, rhodamine, BODIPY and cyanine. Note the large linear or cyclically conjugated systems that allow the absorption of light in the visible spectrum.

1.6.2 Fluorescence Microscopy

These probes are used along with microscopy to probe the structure and function of biomolecules. When producing an image there is an inherent trade-off between image resolution, imaging speed, signal-to-noise and photobleaching[133]. Each available technique will provide advantages in some of these aspects but be poor in others, so careful consideration is necessary.

Optical microscopy uses the transmission or reflectance of visible light to produce an image. Various techniques exist to increase contrast in images produced by light, such as phase contrast and dark illumination microscopy, however one of the most powerful tools available is fluorescence microscopy. Here it is the fluorophores, discussed previously, that provide the contrast in the image by emitting light. If the fluorophores are conjugated to specific structures, then they can be used to study localisation and dynamics.

Several fluorescence microscopy techniques exist, with various advantages and disadvantages[133]. The most basic technique is wide-field fluorescence microscopy (Figure 1.26.A), which typically consists of a light source, a dichroic mirror, excitation and emission filters, an objective and a detector. Photons from the light source are focused by the objective onto the sample for excitation of fluorophores. The photons emitted from fluorophores following de-excitation are collected by the objective and used to produce a magnified image of the position of the fluorophores, which may be observed through microscope eyepieces or by a camera. The dichroic mirror is used to separate photons from the light source and the sample, by reflecting photons from the light source and allowing photons from the sample (at a longer wavelength due to the

Stokes shift) to pass through. These systems now typically have lateral resolution of up to 200 nm and millisecond time resolution.

The maximum resolution is fundamentally limited by the diffraction limit, described by the Abbe diffraction limit, first described in 1873[134]. The maximum lateral resolution is approximately $0.61\lambda/NA$, where λ is the wavelength of light and NA is the numerical aperture of the microscope. The numerical aperture is a measure of how much light can be collected by the objective, given by $n\sin(\alpha)$, where α is one half the angular aperture of the objective. For a high NA, (around 1.4) the maximum lateral resolution is around 200 nm.

The main drawback with wide-field fluorescence microscopy is poor axial resolution as a result of light emitted from out-of-focus planes. Structured illumination microscopy is a form of wide-field microscopy that overcomes this limitation by inserting a moveable grid pattern into the optical path of the excitation light. This creates a pattern that can be used to reduce light from out-of-focus planes and in some cases can be used to improve lateral resolution two times beyond the diffraction limit[135]. However, multiple images must be taken which can lead to photobleaching.

Laser scanning confocal microscopy is another common form of fluorescence microscopy that looks to overcome the axial limitations of wide-field microscopy (Figure 1.26.B). Typically, in laser scanning confocal microscopy a pinhole is inserted in front of the light source. The point of light generated by this is focused on the sample and light collected as normal; however, another pinhole is inserted into the optical path prior to the detector. This prevents light which originates from above or below the plane of focus from reaching the detector and eliminates the light from out-of-focus planes. However,

to construct an image the point of light must be scanned across the whole sample which means photobleaching is a major concern and the speed of acquisition is inherently slow. Multipoint or slit confocal microscopes have been used to accelerate image acquisition.

The axial resolution of laser scanning confocal microscopy is still only on the order of 600-1000 nm. This can be improved by using total internal reflection fluorescence microscopy (TIRF) (Figure 1.26.C). Here an oblique angle of excitation is used, which when set beyond the critical angle to the coverslip, means the light will undergo total internal reflection. This sets up an evanescent wave which only propagates around 200 nm above the coverslip surface, exciting only those molecules close to the surface. As well as increasing axial resolution this can reduce background emission from out-of-focus planes, and since the only requirement is angled excitation light, it is easy to implement.

Only fluorophores in the first 200 nm can be excited however. This has led to the development of light sheet microscopes[136]. These illuminate the sample from a plane orthogonal to the imaging plane, therefore have the same advantage as TIRF, of eliminating out-of-focus light, but also allow for 3D images to be produced with fast image acquisition.

Finally the other main type of fluorescence microscopes are those that are aimed at breaking the diffraction limit by an order of magnitude or more[137]. These fall into two main categories: stimulated emission depletion (STED) microscopy and single-molecule localization methods (Figure 1.27).

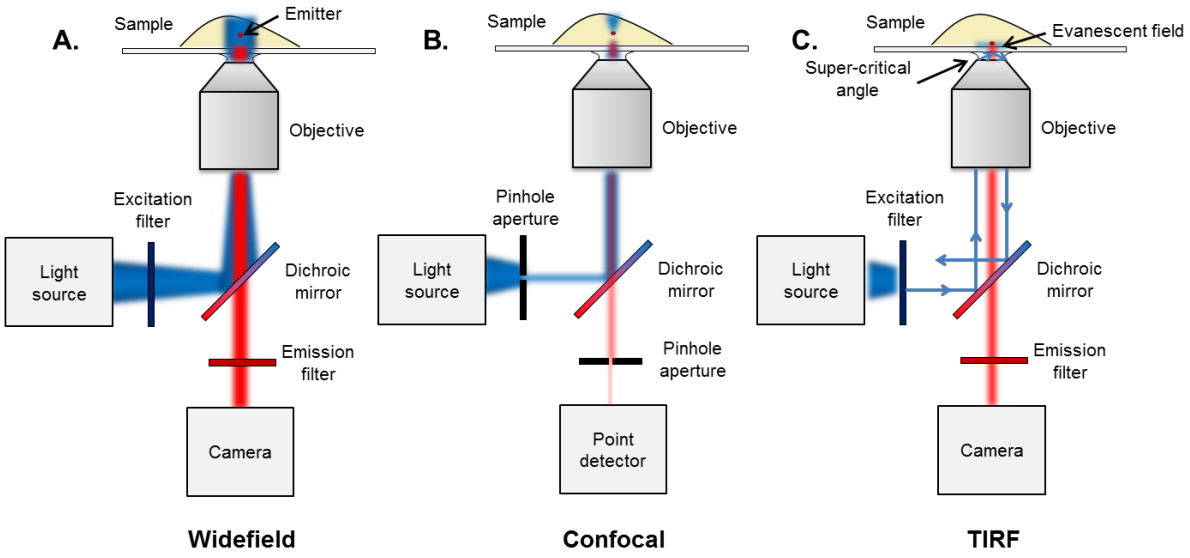


Figure 1.26 **Common fluorescence microscopy techniques. A) A typical widefield microscope. Excitation light is passed through a filter, before being delivered to the sample. Out of focus planes are illuminated, causing poor signal to noise when the emitted light is collected and used to generate an image. B) A typical laser scanning confocal microscope. Many features are shared with a widefield microscope, but there is the addition of pinhole apertures to prevent out of focus planes being illuminated. However, this also means a sample must be scanned to generate a whole image. C) A typical total internal reflection fluorescence (TIRF) microscope. Many features are also shared with widefield microscopy, but the incident laser is tilted beyond the critical angle. Total internal reflection occurs and sets up an evanescent field close to the surface of the sample, eliminating light from out of focus planes and increasing axial resolution. However, imaging is restricted to the surface of the sample.**

In STED an excitation laser is used, as in laser scanning confocal microscopy, but a second doughnut-shaped laser is also used to quench the fluorophores where the excitation spot and doughnut overlap. Depending on the power of the STED laser this typically gives lateral resolutions of around 20-70 nm. The photobleaching caused by the STED laser can be problematic, limiting the maximum resolution that can be obtained.

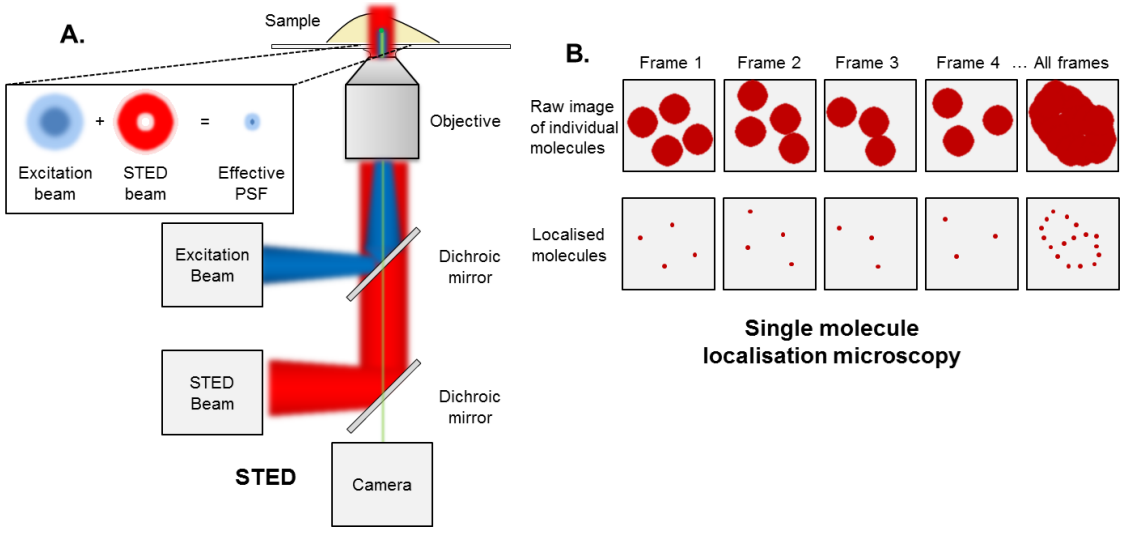


Figure 1.27 **Approaches for super-resolution microscopy. The maximum resolution of optical microscopy is fundamentally limited by the diffraction limit, described by the Abbe diffraction limit ~$0.61\lambda/NA$ ~200 nm. Super-resolution approaches break this diffraction limit by an order of magnitude or more. A) Stimulated emission depletion (STED) microscopy. A second doughnut-shaped laser (STED beam) is used to quench the fluorophores where the excitation beam and doughnut overlap. This generates a narrower effective point spread function (PSF) that can be used to scan the sample. B) Single molecule localisation microscopy (e.g. PALM, STORM). Stochastic switching of fluorophores means that in any single frame there are no overlapping emitters. This means Gaussian fitting can be used to localise individual molecules and if many frames are combined then all molecules can be localised.**

In comparison, single-molecule localization methods approaches, such as PALM and STORM, exploit the localisation of individual fluorophores. If fluorophores can be stochastically switched on and off, such that there are no overlapping fluorophores, the position of individual molecules can be estimated by fitting a 2D Gaussian profile. By taking many images, the positions of individual molecules can be combined to form a single super-resolution image with lateral resolution of the order of 10-30 nm. The advantage of these approaches over STED is that standard fluorescence microscopes, for instance TIRF, can be used, as long as stochastic fluctuations can be generated. However, the number of images that must be required sets a limitation on the speed of acquisition to typically a few seconds per frame at best.

1.7 Conclusion

1.7.1 Overview

Antibiotic-resistance is an increasing problem, which needs to be tackled and better understood. It is usually conferred by the transfer of mobile genetic elements, such as resistance plasmids, which need to be identified and their gene organisation and transmission investigated. A large number of techniques have been developed to identify and image DNA and a broad overview of techniques has been given here, highlighting their development, strengths, weaknesses and applications.

For identification of the DNA code, hybridisation techniques, (e.g. Southern blots, FISH, DNA microarrays), can identify specific genes or SNPs, but lack the single base information required to understand the whole genome. In contrast, DNA sequencing techniques (e.g. NGS, SMRT sequencing) can provide base-pair resolution for unknown genomes, but lack the contextual information required to easily assemble large genomes.

Optical mapping of DNA can help bridge the gap between these traditional techniques, as it provides long-range contextual information that can be used to rapidly identify complex mixtures of DNA as well as large-scale genomic variations (Figure 1.28). In addition, it can be used on small amounts of genetic material and can give single molecule information that is otherwise obscured in a large population. This makes it an ideal method to apply to studying resistance plasmids.

Optical mapping can be carried out on single DNA molecules in nanofluidic devices or using molecular combing. Nanofluidic devices provide clean images of individual DNA

molecules, but thermal fluctuations reduce the apparent resolution and require time lapses to be taken. Imaging DNA molecules that have been combed is inherently faster and higher resolution is possible, but suffers from experimental difficulties in obtaining single, well-separated and stretched molecules.

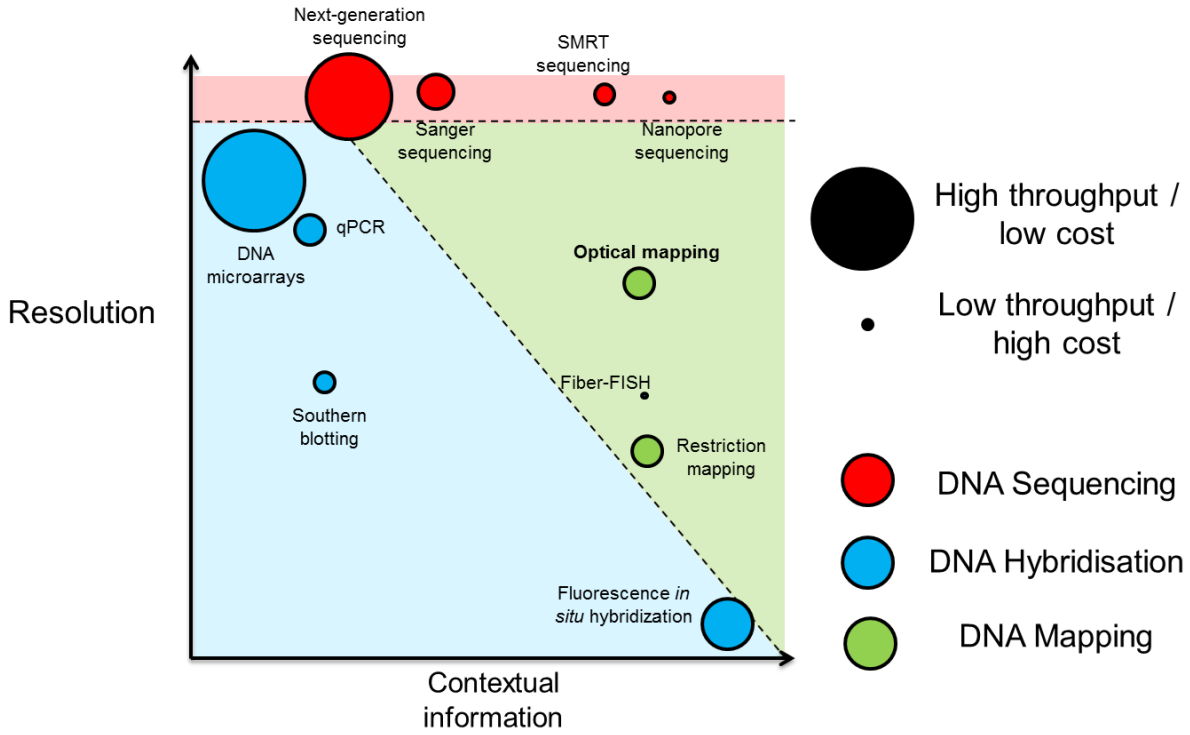


Figure 1.28 **Overview of techniques for DNA identification. There are three main techniques families of techniques that can be used to identify DNA: sequencing, hybridisation and mapping. Sequencing approaches give single base resolution of unknown sequences, but there is generally a trade-off between longer reads (giving contextual information) and throughput/cost. Hybridisation approaches can be used with specific, known sequences to give either good resolution (e.g. microarrays to detect SNPs) or good long-range information (e.g. FISH). However, DNA mapping techniques bridge these two, by providing information on genome structure with ~100-1000 bp resolution, with moderate throughput and cost, for unknown or known sequences.**

Several labelling methods that rely on the underlying sequence can be used to obtain unique patterns for optical mapping. Affinity labelling approaches are generally easy to apply but suffer from relatively low information content (i.e. how different intensity profiles are) compared to enzymatic approaches. Methyltransferase-directed labelling can be used at variable densities to provide unique intensity profiles for DNA identification and doesn't cause DNA damage or suffer from many non-specific labels.

1.7.2 Aim and objectives

The aim of this research is to develop novel applications for methyltransferase-directed fluorescent labelling of DNA. The first objective is to develop a robust technique for identification of viruses, resistance plasmids, bacterial populations and other complex mixtures of DNA. Taking the strengths and weaknesses of current techniques into account, this research will use an optical mapping approach, utilising methyltransferase-directed labelling and molecular combing of DNA fragments. A combination of small organic fluorophores and fluorescence widefield/TIRF microscopy will be used for imaging.

The other objective of this research is to apply methyltransferase-directed labelling to visualise plasmids transfected into bacteria. The development of this technique will allow it to be applied for study of the partition mechanism of high copy number plasmids and of the transfer and maintenance of low copy number plasmids, such as resistance plasmids.

To achieve these aims this thesis will be split into three main sections. The first (CHAPTER 2) will focus on the optimisation of methyltransferase-directed labelling, applied to fluorescent labelling for optical mapping and visualisation of plasmids. The

next section (CHAPTER 3 and CHAPTER 4) will focus on the computational and experimental aspects of optical mapping using methyltransferase-directed labelling. These approaches will be applied on simple genomes and mixtures, through to more complex mixtures. In the final section (CHAPTER 5) methyltransferase-directed labelling will be applied to small plasmids, followed by transformations into bacteria for visualisation of localisation and dynamics.

CHAPTER 2 OPTIMISATION OF METHYLTRANSFERASE-DIRECTED FLUORESCENT LABELLING OF DNA

Robert K. Neely provided supervision and guidance for the research undertaken in this chapter. Nathaniel O. Wand (the author) designed, performed and analysed all labelling experiments, including restriction assays and single molecule counting experiments. Andrew Wilkinson designed, performed and analysed all stability and purity experiments.

2.1 Introduction

2.1.1 Enzymes and DNA identification

Enzymes have been used to identify specific DNA sequences and map genomes since the 1970s. This family of techniques exploits the natural sequence specificity of enzymes, to help produce a unique pattern that is representative of the underlying DNA sequence. For example restriction enzymes were used to produce specific fragments of DNA by Danna and Nathans in 1971[18], which were separated by gel electrophoresis and visualised. This is a powerful technique, since as well as confirming the identity of DNA it allows the relative position of enzyme recognition sites to be found, which can be used to determine the orientation and position of an insert in a cloning vector.

For the identification of microorganisms, restriction mapping has been largely superseded by DNA hybridisation techniques (e.g. DNA Microarrays, FISH), to detect specific, known sequences, and DNA sequencing techniques (e.g. Sanger, NGS), to map and sequence whole genomes. However, the specificity of enzymes has been exploited again more recently, for the identification of DNA by optical mapping. This has the potential to bridge the resolution and contextual information provided hybridisation

and sequencing techniques, in particular to determine the arrangement of genomes in a costly and timely manner. In optical mapping individual DNA molecules are stretched and imaged, before being identified by unique patterns, which are representative of the underlying sequence. Optical mapping, using restriction enzymes to generate the unique pattern, was first demonstrated in the mid-1990s[45]. Since then a number of other enzymes have been used to sequence-specifically label, rather than cut, DNA, for instance nicking enzymes[65] and DNA methyltransferases[69].

2.1.2 Methyltransferase-directed labelling of DNA

DNA methyltransferases and restriction enzymes together form the restriction-modification system in bacteria, of which thousands of examples are now known[71]. In bacteria, methylation and the respective restriction by the sister endonuclease occurs within the same DNA target, consisting of a palindromic sequence, typically 2 to 8 base pairs in length. DNA methyltransferases are an attractive target for labelling DNA, since they can be used to modify DNA efficiently, covalently, without damage, with a broad range of densities, and with high specificity and fidelity[138,139].

In nature all known classes of DNA methyltransferases use the cofactor *S*-adenosyl-L-methionine (AdoMet) as the methyl donor. It has been demonstrated that synthetic AdoMet analogues can also be used to catalyse the transfer of more complex chemical groups to DNA. Fluorophores and other modifications can be targeted to specific sites in a DNA sequence, efficiently and non-destructively[86]. This labelling has been used for a number of applications, such as capturing DNA, visualising it *in situ,* as well as for optical mapping. The efficacy of this approach for these applications will depend on the efficiency of the labelling. For instance, optical mapping of 100% labelled DNA

fragments will be more reliable than optical mapping of fragments with only 10% labelling, since it is the unique pattern that is required for alignment to a reference (see section 3.2.4).

2.1.3 Quantifying DNA methyltransferase labelling efficiency

Quantifying labelling efficiency has not proved a straightforward task and there has been little attempt to systematically assess the efficiency of labelling by synthetic cofactors. Traditionally the protection of DNA by methyltransferases is quantified by restriction assays[140] (Figure 2.1A). In these a restriction enzyme, which recognises the same sequence as the methyltransferase, is incubated with the methylated DNA. Non-methylated sites will be cleaved, and the DNA sample can be analysed by gel electrophoresis.

Figure 2.1C shows the restriction pattern that can be expected, depending on the labelling efficiency, for pUC19, which is 2686 base pairs long, and the DNA methyltransferase M.TaqI (5'-TCGA-3'), for which there are four sites on the plasmid (Figure 2.1B). When pUC19 is unlabelled, and therefore fully digested, three fragments are clearly visible (1444, 736 and 476 bp). The smallest fragment (30 bp) is not visible since the intensity of bands is dependent on the mass of DNA. When pUC19 is partially labelled there is incomplete digestion and a number of longer fragments appear. When labelling reaches 100% one single band is seen, as the DNA is unrestricted. In practice the plasmid will exist in a supercoiled conformation, so a fragment will be present that runs faster than the linear fragment. In acidic solutions and at high temperatures, the DNA can be damaged. Nicking will cause the plasmid to adopt the open circular conformation, whilst a break will lead to linear DNA. This can be a problem in real

experiments, since the expected restriction pattern for 50-60% labelling and the restriction pattern for the fully protected plasmid can be difficult to discriminate.

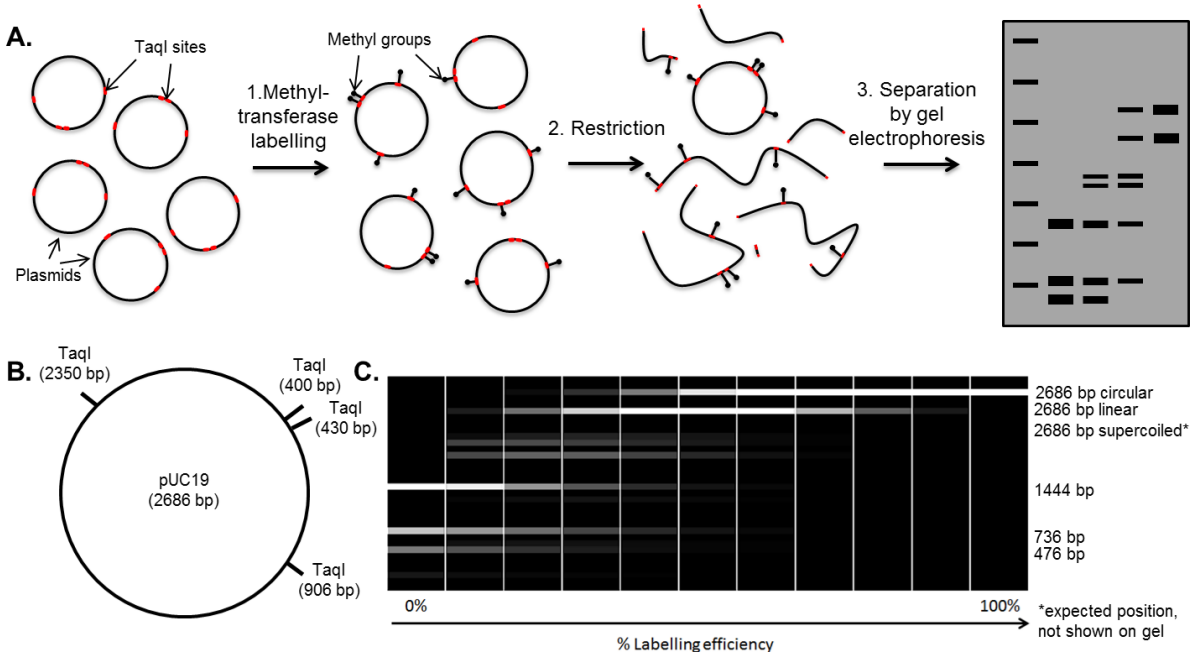


Figure 2.1 **Quantifying labelling efficiency by restriction assays. A) General restriction assay procedure. First DNA is labelled at specific sites (e.g. TaqI, 5'-TCGA-3' sites) by the methyltransferase. Next the sister restriction enzyme, that will cut at the same site, is incubated with the DNA and will fragment un-methylated DNA. The resulting restriction fragments are separated by gel electrophoresis. B) Map of pUC19 showing TaqI sites. C) Expected restriction pattern for pUC19 restricted by R.TaqI from 0% to 100% methylation efficiency.**

As well as this issue, restriction assays do not directly report on the labelling efficiency by synthetic cofactors. For instance the site may be hemi-methylated only (R.TaqI doesn't restrict hemi-methylated sites[71]) or the site may not be labelled with the desired synthetic group, it could for instance be labelled with a methyl group only. This can lead to misleading results and alternative methods should be used as validation.

For optical mapping using methyltransferases, labelling by fluorophores gives the unique pattern for DNA identification. However, simple quantification of labelling efficiency by absorption spectroscopy is not possible since the concentration of DNA is

typically of the order of nanomolar. For instance, pUC19 at a concentration of 100 ng/µl is equivalent to around 50 nM, which for 8 dyes per plasmid equates to 0.4 µM dye. For Atto647N, with a molar extinction coefficient of 1.5×10^5 $M^{-1}cm^{-1}$, and a path length of 1 cm this would equate to an absorbance of just 0.06. To address this we have previously reported a single molecule counting procedure[141] (Figure 2.2), which in tandem with restriction assays, can now be used to give a more comprehensive description of labelling efficiencies.

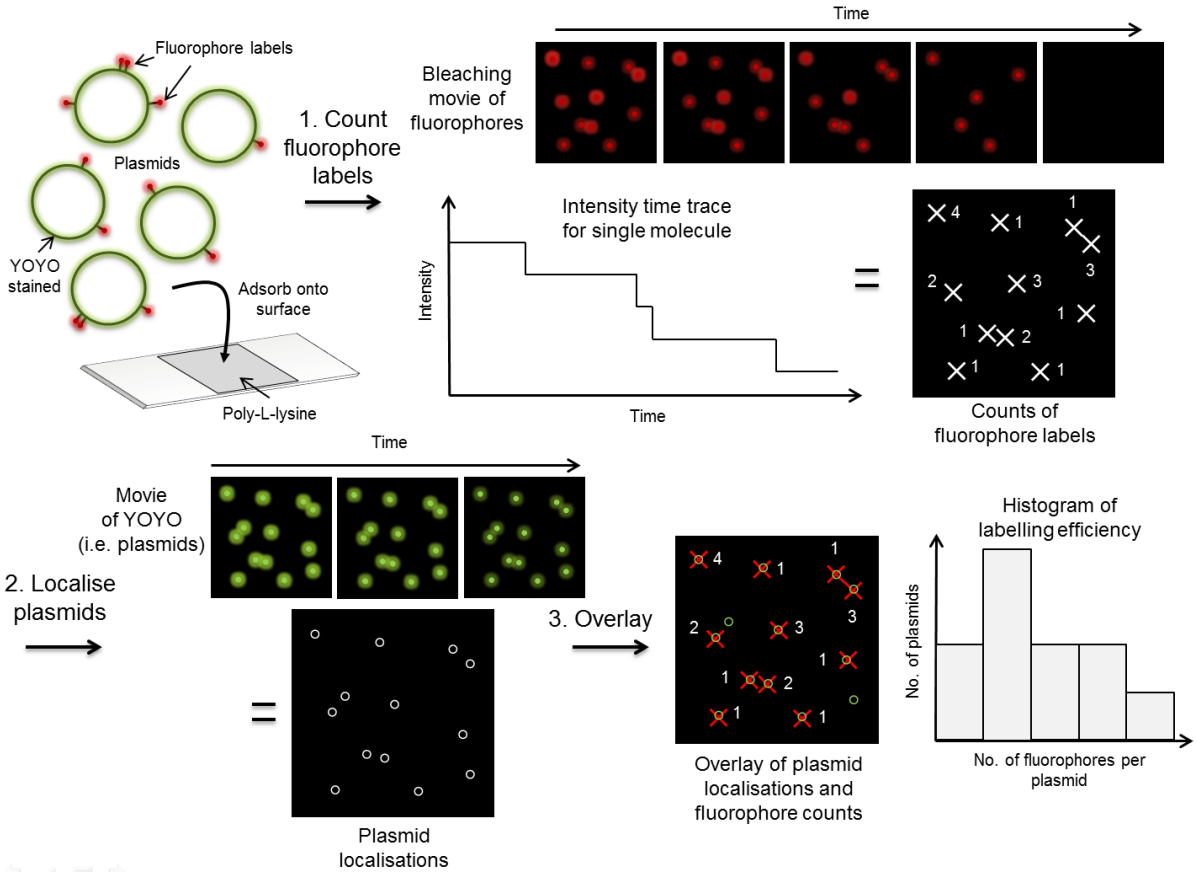


Figure 2.2 **Single molecule counting procedure. Fluorescently-labelled plasmids are stained with YOYO-1 and adsorbed onto a poly-L-lysine surface. A bleaching movie is taken of the fluorophore labels and used to localise them. An intensity time trace for each molecule is effectively used, in which individual bleaching steps correspond to individual fluorophores. A separate movie is taken of the YOYO-1 and used to localise plasmids. The fluorophore and plasmid localisations are overlaid using custom Matlab software to produce a histogram which reports the number of fluorophores counted per plasmid.**

In the single molecule counting procedure fluorescently-labelled plasmids are stained with YOYO-1 (a DNA intercalator) and adsorbed onto a positively-charged poly-L-lysine surface. A bleaching movie is then taken of both the fluorophore and of YOYO-1. The YOYO-1 movie can then be used to localise plasmids, whilst the number of fluorophores can be counted by running backwards through the bleaching movie and localising fluorophores as they appear. This can be done using freely available Localizer software[94]. The positions of fluorophores and positions of plasmids can be used to determine the overlap, and therefore used directly to count the number of fluorophores coupled per plasmid (i.e. a maximum of 8 per pUC19 plasmid).

Figure 2.3 shows the expected distributions for pUC19 (generated by stochastic modelling), for labelling efficiencies ranging from 10 to 100%. At low labelling efficiencies, only a few fluorophores are counted per plasmid, for example at 20% labelling efficiency the peak is at 1 fluorophore per plasmid. At 50% labelling efficiency a distribution of plasmids is expected, centred on 4 fluorophores per plasmid. Finally, at very high labelling efficiencies most plasmids should have all 8 fluorophores.

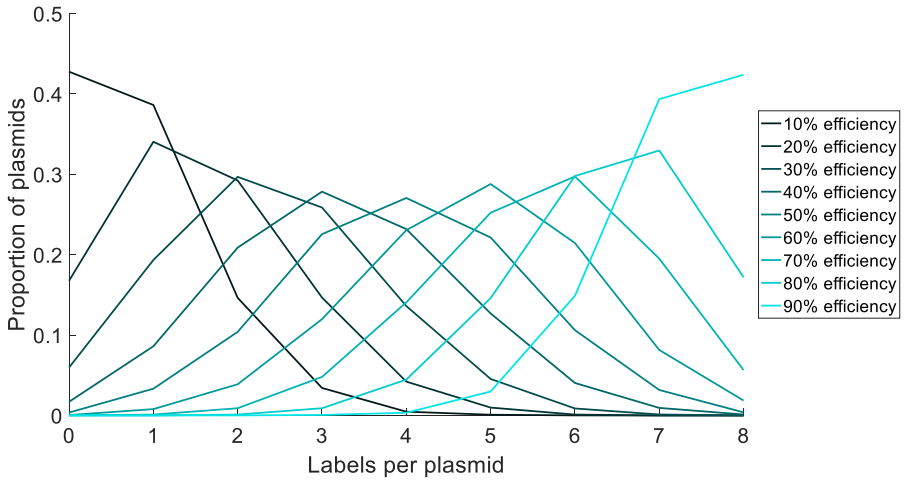


Figure 2.3 Calculated distributions for single molecule counting, for pUC19 labelled by M.TaqI. From 10-90% labelling efficiency.

2.1.4 Overview

Here a range of experimental factors that affect the efficiency of fluorescent labelling of DNA by M.TaqI will be tested. The optimisation of fluorescent labelling demonstrates how the maximum labelling efficiency by methyltransferases can be achieved and should serve as a guide for other methyltransferase labelling strategies.

The factors that will be discussed include: dye and cofactor coupling strategies; the effect and removal of bound AdoMet; the decomposition of the cofactor; the reaction conditions, including the reaction buffer, temperature and time of reaction; dye purity and choice; and the choice of methyltransferase. Finally, the reliability of single molecule counting experiments and the overall labelling efficiency will be discussed.

2.2 Results and Discussion

2.2.1 Dye and cofactor coupling strategies

Here four strategies for the conjugation of fluorophores to DNA are discussed. These involve pre- or post-transalkylation coupling of the fluorophore using amine-NHS or strain-promoted alkyne-azide cycloaddition (SPAAC) coupling (Figure 2.4). These are well known reactions that can be readily used in aqueous conditions. The amine-NHS reaction couples primary amines to NHS esters to form peptide bonds, but suffers from competing hydrolysis of the NHS-ester.[142] For example at pH 7.0 at 0°C the NHS-ester half-life is around 4-5 hours, whilst at pH 8.6 at 4°C this is reduced to 10 minutes. The dibenzocyclooctyne (DBCO) and azide groups used in the SPAAC coupling are by contrast far more stable.[143]

The structures of the AdoMet analogues that have been used for these schemes are shown in Figure 2.5. In all the analogues the methyl group in AdoMet has been replaced by an extended side chain. Side chains are activated by incorporating a double or triple bond in the beta position to the sulfonium centre[76] and a terminal functional group is incorporated for coupling.

The length of the sidechain is also important[81]. But-2-ynyl cofactors which contain an electron-withdrawing group close to the unsaturated bound are susceptible to nucleophilic attack, for instance by water. However, by using hex-2-ynyl groups, extending the carbon chain by two units, the stability of cofactors is markedly increased. Longer chains are also beneficial for post-transalkylation coupling, as they permit better accessibility and enhanced reactivity of the terminal functional group. There is some

increased steric hindrance, however M.TaqI has a cofactor binding pocket which can accommodate longer chains well[144].

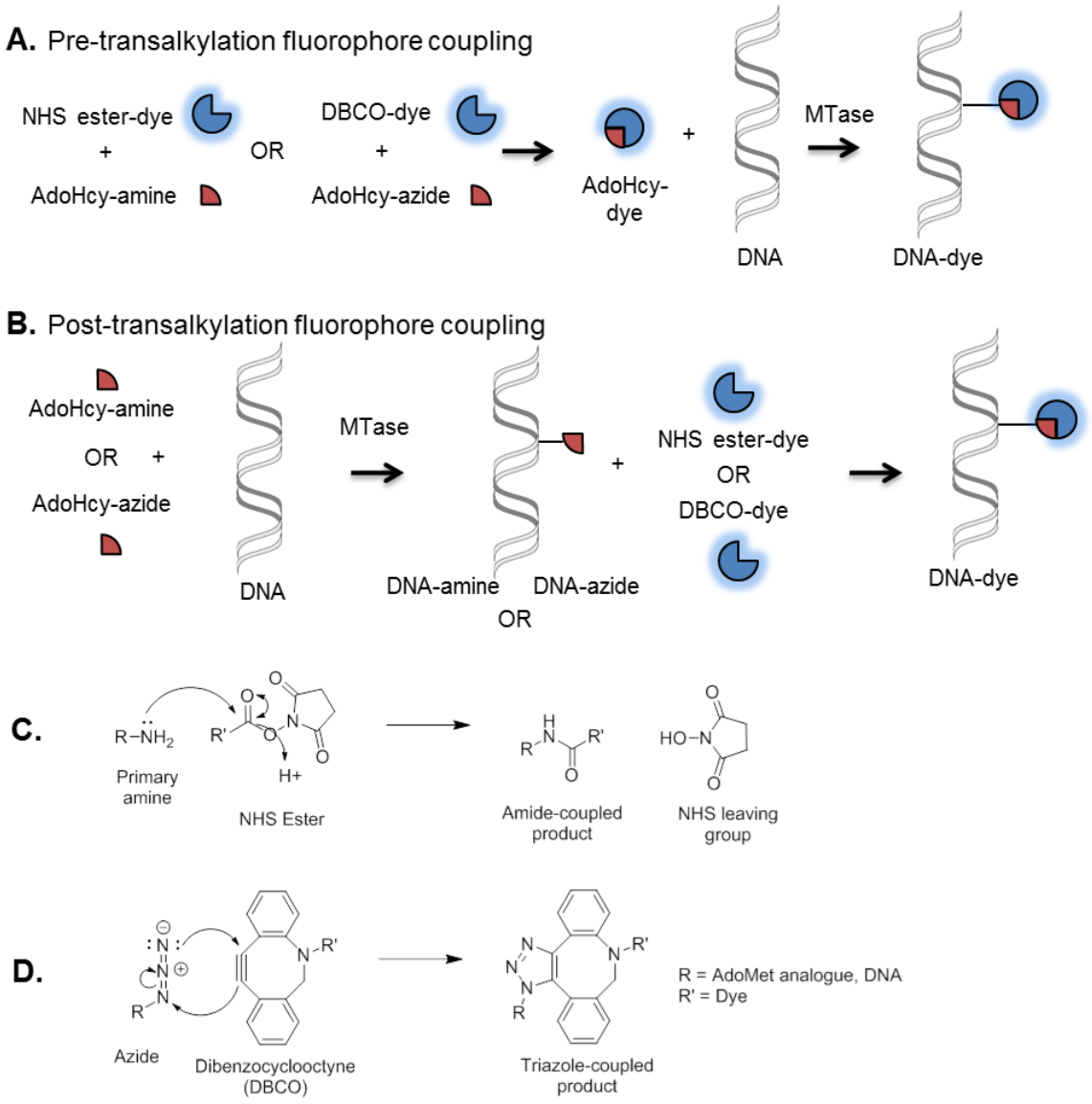


Figure 2.4 **Labelling and reaction schemes. A) Pre-transalkylation fluorophore coupling: labelling with AdoHcy-dye after amine-NHS coupling or SPAAC coupling. B) Post-transalkylation fluorophore coupling: labelling with AdoHcy-amine or -azide followed by amine-NHS coupling or SPAAC coupling to dye. C) Amine-NHS coupling reaction. D) SPAAC coupling reaction.**

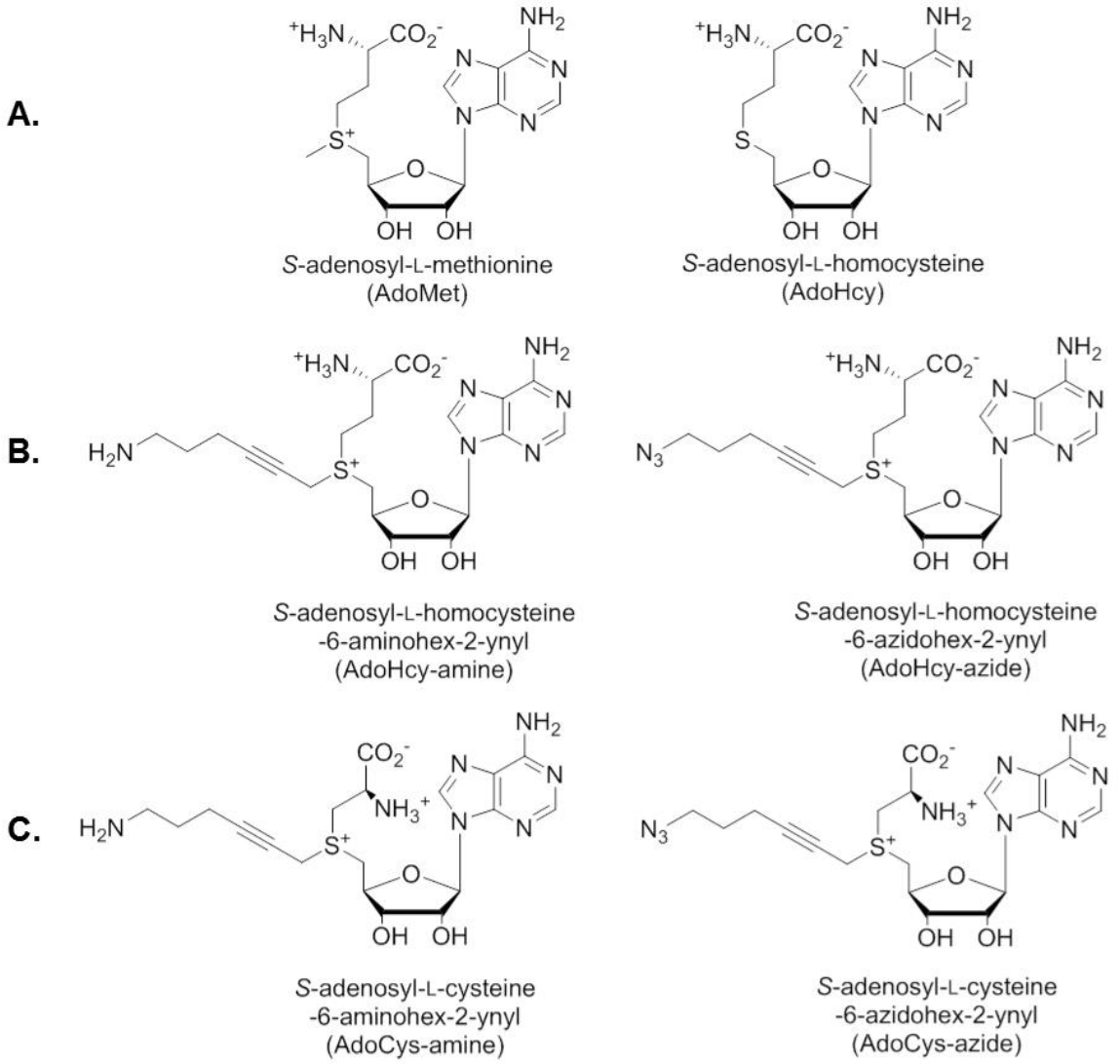


Figure 2.5 **AdoMet analogues used for transalkylation. A) Naturally occurring cofactors: AdoMet and AdoHcy. AdoMet is the natural cofactor used for DNA methylation. AdoHcy is the natural product of DNA methylation and can be used as a precursor for synthesis of AdoMet analogues. B) Hex-2-ynyl AdoMet analogues: AdoHcy-amine and AdoHcy-azide. Both cofactors replace the methyl group in AdoMet with a hex-2-ynyl chain. AdoHcy-amine has a primary amine group at the end of this chain, whilst AdoHcy-azide has an azide functional group. C) Hex-2-ynyl AdoCys analogues: AdoCys-amine and AdoCys-azide. Both cofactors replace the homocysteine in AdoHcy with a cysteine and the methyl group in AdoMet with a hex-2-ynyl chain. AdoCys-amine has a primary amine group at the end of this chain, whilst AdoCys-azide has an azide functional group.**

Novel cysteine cofactors which differed from homocysteine analogues by one fewer carbon in the amino acid chain were also tested. These would have applications if the enzyme binding pocket could be designed to accommodate them and should show enhanced stability, as ring closing of the amino acid is one of the prominent decomposition pathways. However, in this application they showed significantly lower activity with the wild-type M.TaqI enzyme and were therefore not used further. Restriction assays are shown for each of these cofactors (Supplementary Figure 7.3: AdoHcy-amine, AdoCys-amine and AdoCys-azide and Figure 2.13: AdoHcy-azide).

2.2.2 Comparing restriction assays and single molecule counting

It has been clear from previous labelling experiments that although restriction assays appear to show full protection of DNA (to the extent that any restricted fragments are beyond the limits of detection), this does not equate to 100% fluorescent labelling of DNA[141]. This is shown in Figure 2.6 where a two-step labelling scheme is used: first the DNA is transalkylated using M.TaqI and AdoHcy-azide; then the azide functionalised groups are SPAAC-coupled to a fluorophore. Figure 2.6A, lanes 2-7, show full protection of pUC19 by M.TaqI and AdoHcy-azide at high concentrations of M.TaqI. For comparison, the concentration of pUC19 is around 20 nM, therefore the concentration of labelling sites for M.TaqI is around 200 nM, and the concentration of M.TaqI in lane 2 is around 300 nM. This means for lane 4 (which appears to have complete protection) there is turnover of at least two times. The corresponding single molecule counting results after SPAAC coupling however only show a maximum of around 45% labelling efficiency (Figure 2.6B and Figure 2.6C).

It is important to note that M.TaqI is not preventing restriction by binding to DNA (Supplementary Figure 7.1) and also that there is no evidence that the palindromic target sites are only being modified on one site. The expected distributions this would give in single molecule counting experiments can be modelled and compared to experimental results (Supplementary Figure 7.2).

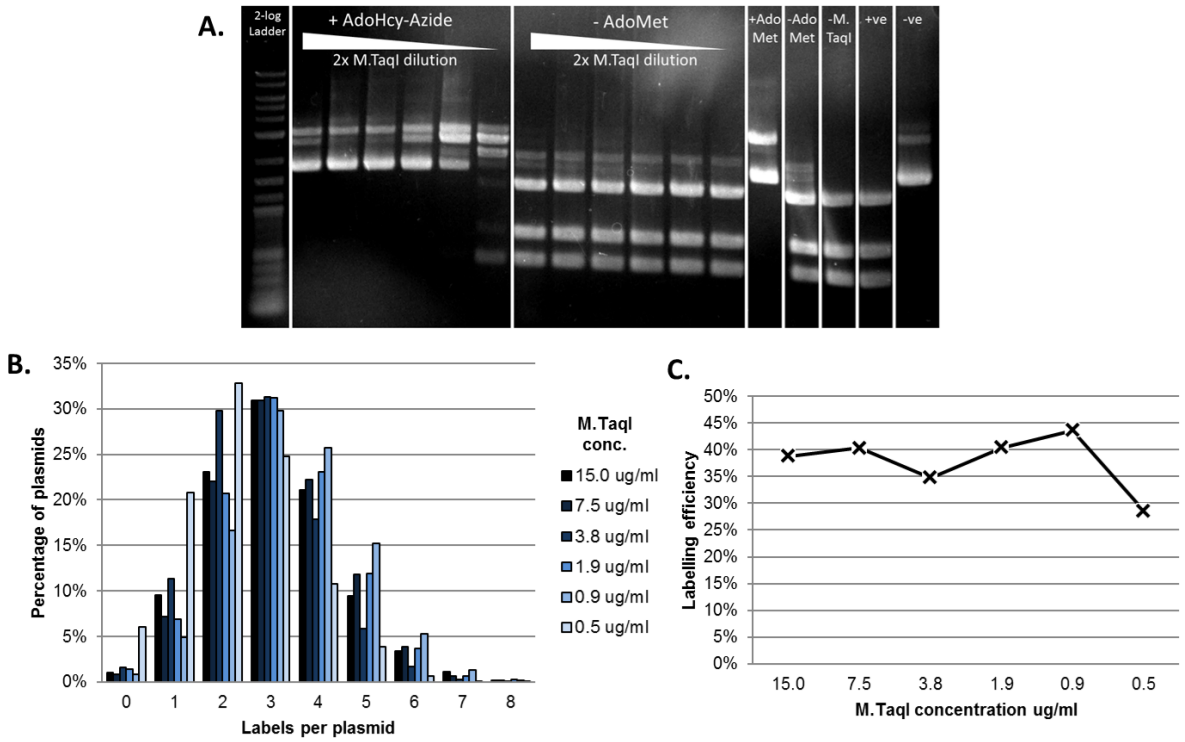


Figure 2.6 **M.TaqI labelling of pUC19 with AdoHcy-azide. A) Restriction assay for M.TaqI labelling of pUC19 with AdoHcy-azide and without added cofactor. Lane 1 = 2 log ladder; lanes 2-7 = AdoHcy-azide, 2x dilution of M.TaqI; lanes 8-13 = no cofactor, 2x dilution of M.TaqI; lane 14 = AdoMet control; lane 15 = no cofactor control; lane 16 = no M.TaqI control; lane 17 = restricted pUC19; lane 18 = unrestricted pUC19. B) Single molecule counting results, for labelling conditions equivalent to lanes 2-7, followed by post-transalkylation coupling with TAMRA-DBCO. C) Labelling efficiencies from counting results**

2.2.3 Effect and removal of bound AdoMet

One of the factors which could cause this apparent discrepancy is the protection of DNA by AdoMet instead of the desired AdoMet analogue, leading to a methyl group being transferred instead of the reactive chemical moiety. It has been previously reported that purified M.TaqI is able to protect DNA even when no AdoMet is added to the reaction. Bound AdoMet is co-purified with M.TaqI and has been observed in, for example, the crystal structure, despite not being deliberately added to the crystallisation buffer[144]. Figure 2.6A, lanes 8-13, are consistent with this and show protection of DNA despite no AdoMet being added to the reaction and there is a concomitant increase in labelling efficiency with a reduction in M.TaqI concentration (Figure 2.6B and Figure 2.6C).

However, it also of note that the amount of protection by bound AdoMet is non-linear, in other words as the enzyme concentration is halved the protection does not also half. Therefore, at even very low concentrations of M.TaqI labelling by bound AdoMet is seen. This may be due to complex characteristics of the enzyme and labelling reaction. For instance, for other methyltransferases it has been shown that an enzyme dimer may be formed during labelling[145] and that AdoMet may bind in different sites and in different conformations[146].

Previously reported methods have used incubations with oligonucleotides to remove this bound AdoMet[144]. However, the results here do not suggest this has a significant effect and indeed this can reduce labelling efficiency if oligonucleotides are not used carefully. Oligonucleotides containing the labelling site (e.g. TCGA for M.TaqI) can be added directly to the reaction mixture before pUC19 is added. In principle this should

use up all the bound AdoMet before the transalkylation reaction is carried out on pUC19, as M.TaqI uses the bound AdoMet to methylate the oligonucleotides.

This does appear to prevent protection in the restriction assays when no AdoMet analogue is added (Figure 2.7A, lanes 1-6), and there is still significant protection when AdoHcy-azide is added (Figure 2.7A, lanes 7-12). However, single molecule counting results suggest there is no improvement in labelling efficiency. The best labelling is achieved when no oligonucleotides are added (Figure 2.7B and Figure 2.7C), likely due to labelling of the oligonucleotides in preference to the pUC19, due to the large concentration of labelling sites on oligonucleotides. For comparison lane 1 contains 10 μM oligonucleotides, equivalent to 80 μM labelling sites, whilst pUC19 is at a concentration of around 20 nM, which is equivalent to 160 nM labelling sites.

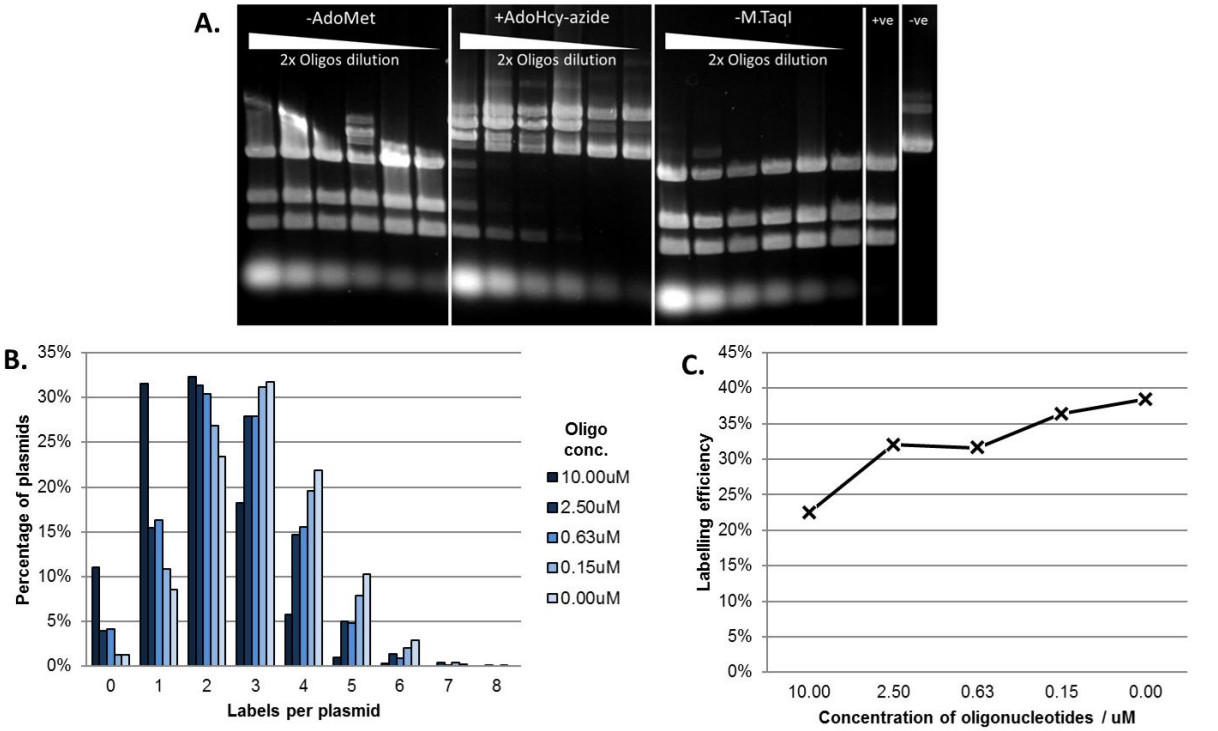


Figure 2.7 **M.TaqI labelling of pUC19 with AdoHcy-azide in the presence of oligonucleotides. A) Restriction assay for M.TaqI labelling of pUC19 with AdoHcy-azide and without added cofactor, in the presence of oligonucleotides. Lanes 1-6 = no cofactor, 2x dilution of oligonucleotides; lanes 7-12 = AdoHcy-azide, 2x dilution of oligonucleotides; lanes 13-18 = no M.TaqI, 2x dilution of oligonucleotides; lane 19 = restricted pUC19; lane 20 = unrestricted pUC19. B) Single molecule counting results for conditions equivalent to lanes 7-12, but a 4x dilution of oligonucleotides starting from lane 8 and finishing with no oligonucleotides, followed by post-transalkylation coupling with TAMRA-DBCO. C) Labelling efficiencies from counting results**

A more careful method is to incubate M.TaqI with oligonucleotides, followed by removal of the oligonucleotides before labelling of pUC19. This has been carried out using anion exchange columns after overnight incubation of M.TaqI with oligonucleotides. However even after overnight incubation there is still evidence of bound AdoMet present (Figure 2.8A, lanes 1-6) and labelling efficiency is not improved (Figure 2.8C). This is in contradiction of previous studies[144].

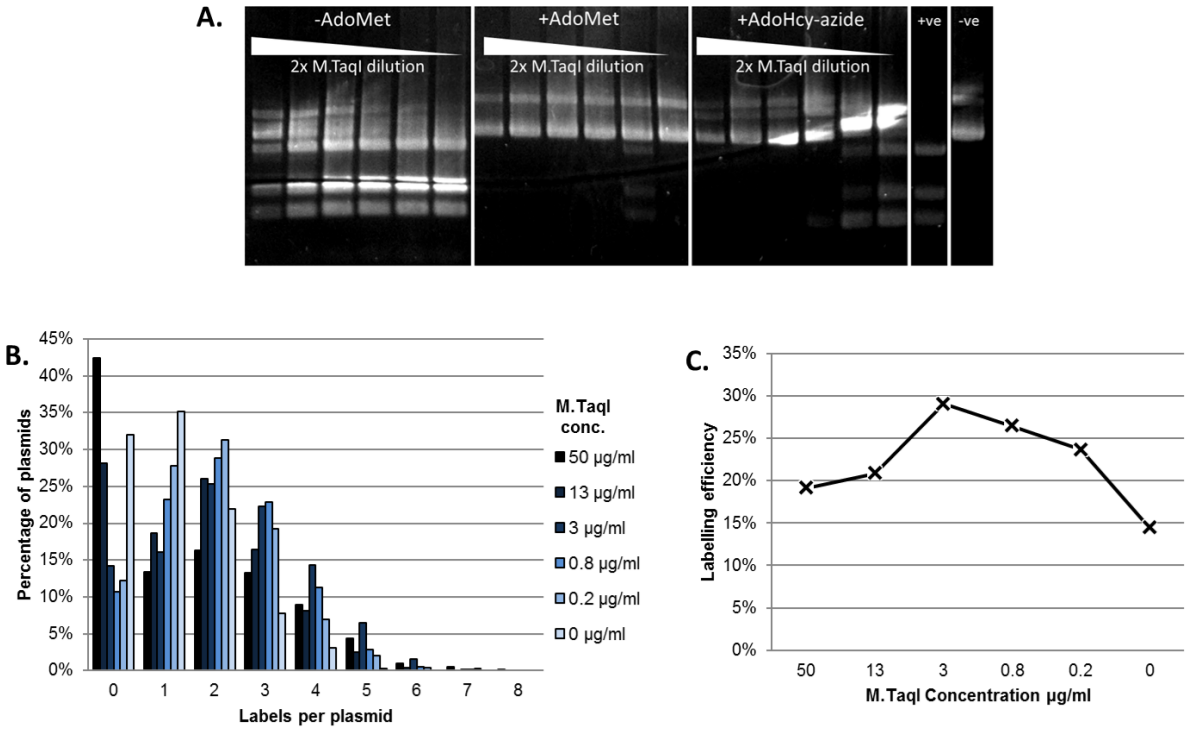


Figure 2.8 **M.TaqI labelling of pUC19 with AdoHcy-azide after incubation with oligonucleotides. A) Restriction assay for M.TaqI labelling of pUC19 with AdoHcy-azide, AdoMet and without added cofactor, after incubation with oligonucleotides. Lanes 1-6 = no cofactor, 2x dilution of M.TaqI; lanes 7-12 = AdoMet, 2x dilution of M.TaqI; lanes 13-18 = AdoHcy-azide, 2x dilution of M.TaqI; lane 19 = restricted pUC19; lane 20 = unrestricted pUC19. B) Single molecule counting results for conditions equivalent to lanes 13-18, but a 4x dilution of oligonucleotides starting from lane 13 and finishing with no oligonucleotides, followed by post-transalkylation coupling with TAMRA-DBCO. C) Labelling efficiencies from counting results.**

Here the single molecule counting results also show that at very high enzyme concentrations, where there is significant protection by bound AdoMet, there are a large amount of plasmids with no labels (Figure 2.8B). This distribution is surprising, since it is not predicted by the modelling carried out previously (Figure 2.3 and Supplementary Figure 7.2). When modelling labelling efficiency the chance of labelling each site on the pUC19 was independent of the labelling of the other three sites. However here it appears that if one site is methylated by bound AdoMet then there is an increased probability of other sites being methylated. At this point it is not clear what causes this behaviour.

Another method we propose to remove bound AdoMet is incubation of M.TaqI with a competitive inhibitor. Sinefungin, a naturally occurring competitive inhibitor of AdoMet (Figure 2.9A)[147], was incubated with M.TaqI prior to the transalkylation reaction. Figure 2.9B shows an example restriction assay for this reaction. In lanes 1-6 there is a decrease in protection by bound AdoMet at higher sinefungin concentrations, however importantly there is still some protection even at very high concentrations of sinefungin. This is unexpected, since the bound AdoMet is at a concentration of no greater than the M.TaqI, around 300nM, whilst the dissociation constants for AdoMet and sinefungin are 2.0 μM and 0.34 μM respectively. Therefore, at such high concentrations of sinefungin, for example lane 1 contains 10 mM sinefungin, virtually all the bound AdoMet should be displaced by sinefungin, which is incapable of labelling pUC19.

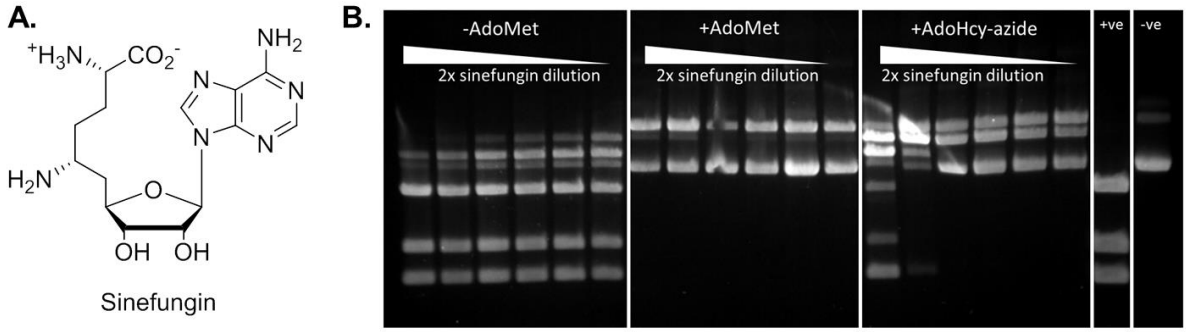


Figure 2.9 **M.TaqI labelling of pUC19 with AdoHcy-azide after incubation with sinefungin. A) Structure of sinefungin. This is structurally similar to AdoMet and AdoHcy (Figure 2.5A) and is a naturally occurring competitive inhibitor of transalkylation by AdoMet. B) Restriction assay for M.TaqI labelling of pUC19 with AdoHcy-azide, AdoMet and without added cofactor, after incubation with sinefungin. Lanes 1-6 = no cofactor, 2x dilution of 10 mM sinefungin; lanes 7-12 = AdoMet, 2x dilution of 10 mM sinefungin; lanes 13-18 = AdoHcy-azide, 2x dilution of 10 mM sinefungin; lane 19 = restricted pUC19; lane 20 = unrestricted pUC19.**

These results suggest removal of bound AdoMet is not straightforward and its presence should be taken into consideration. The non-linear protection of DNA by bound AdoMet may explain some of the difficulty in complete removal of bound AdoMet. Extensive washing and dialysis of methyltransferases as well as oligonucleotide or sinefungin treatments should be considered to remove as much bound AdoMet as possible[148]. It is also important to note that the concentration of M.TaqI should be minimised to reduce labelling by bound AdoMet, which requires optimisation of reaction conditions.

2.2.4 Efficiency of coupling strategies and decomposition of cofactor

In the previous section a two-step labelling scheme was used: first the DNA is transalkylated using M.TaqI and AdoHcy-azide; then the azide functionalised groups are SPAAC-coupled to a fluorophore. However, a number of alternative strategies were outlined in Figure 2.4, which should each be considered. For example, it has been shown previously that the alternative two-step labelling scheme using post-transalkylation amine-NHS coupling is remarkably inefficient[141]. This is due to competing processes such as hydrolysis of the NHS ester and dye decomposition.

For amine-NHS coupling pre-transalkylation coupling of the AdoMet analogue, AdoHcy-amine, to the NHS ester dye is a more successful strategy and showcases the extent to which very large synthetic groups can be transferred to DNA by M.TaqI. However, this approach still has a number of drawbacks, particularly the breakdown of the AdoMet analogue. AdoMet is well known to be unstable at high pH[149], however amine-NHS coupling reactions require a slightly alkaline pH (7.2-9.0) to occur.

Therefore, balancing the pH for maximum coupling but minimal degradation of the cofactor and NHS-ester is difficult. Various conditions were trialled for coupling. 1xPBS

does not buffer the mixture sufficiently, so the buffer is too acidic, (due to the formic acid used for cofactor storage), whilst 0.1M sodium bicarbonate, pH 8.3, is too basic and gave poor labelling, presumably as a result of cofactor degradation. A 10x PBS solution, pH 7.4, at 4^{o}C, gives the best conditions, balancing coupling and degradation of the cofactor and NHS-ester.

Cofactor degradation will also continue during the transalkylation reaction and a balance between conditions must be struck. More acidic conditions will slow cofactor degradation but reduce the activity of M.TaqI. The competition between these is shown in Figure 2.10, where pH is varied between 5.75 and 6.75. With AdoMet as the cofactor the activity of M.TaqI slightly decreases with decreasing pH (lanes 7-11), however with the pre-transalkylation coupled cofactor, AdoHcy-Atto647N (lanes 2-6), there is an increase in protection with decreasing pH, until pH 5.75, where the inactivity of the enzyme becomes the overriding factor.

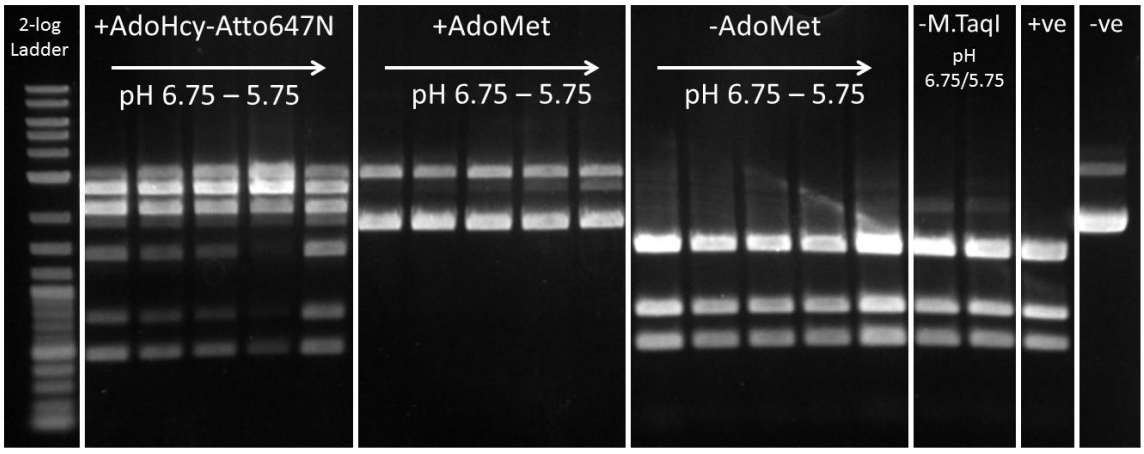


Figure 2.10 Variation in protection with pH. M.TaqI labelling of pUC19 with AdoHcy-amine coupled pre-transalkylation to Atto647N NHS Ester. Lane 1 = 2 log ladder; lanes 2-6 = AdoHcy-Atto647N, pH 6.75-5.75; lanes 7-11 = AdoMet, pH 6.75-5.75; lanes 12-16 = no M.TaqI control, pH 6.75-5.75; lanes 17/18 = no added cofactor, pH 6.75/5.75; lane 19 = restricted pUC19, lane 20 = unrestricted pUC19.

These results can also be clearly demonstrated in single molecule counting results (Figure 2.11). Here just two pHs are tested: pH 5.7 and pH 7.2, corresponding to commercial CutSmart from New England Biolabs with Tris as the buffer system, and an identical buffer with MES as the buffering agent. The restriction pattern in Figure 2.11A indicates that protection is more complete at lower pH (lanes 2-6 compared to lanes 7-11), and this is confirmed in the single molecule counting results (Figure 2.11B), which give labelling efficiencies of 35% and 26% at pH 5.7 and 7.2 respectively.

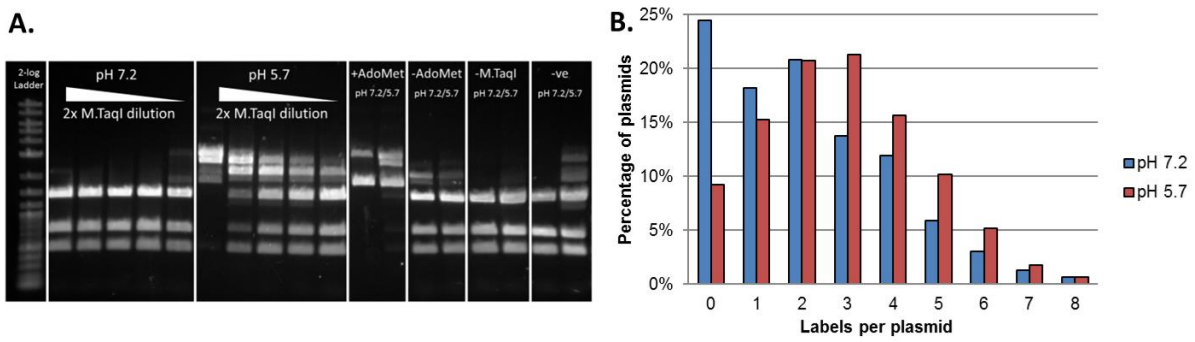


Figure 2.11 Variation in labelling efficiency at pH 7.2 vs pH 5.7. M.TaqI labelling of pUC19 with AdoHcy-amine coupled pre-transalkylation to Atto647N NHS Ester. A) Restriction assay. Lane 1 = 2 log ladder; lanes 2-6 = pH 7.2, 2x dilution of M.TaqI; lanes 7-11 = pH 5.7, 2x dilution of M.TaqI; lanes 12/13 = AdoMet control, pH 7.2/5.7; lanes 14/15 = no cofactor, pH 7.2/5.7; lanes 16/17 = no M.TaqI, pH 7.2/5.7; lanes 18/19 = restricted pUC19 , pH7.2/5.7. B) Single molecule counting results. Labelling at pH 7.2 (blue) and pH 5.7 (red) with the same concentration of M.TaqI.

SPAAC coupling to the cofactor can also be used pre-transalkylation. Here the coupling can be carried out at low pH in 0.05% formic acid to slow degradation of the cofactor, since there are no competing reactions as during amine-NHS coupling[143]. However, the same dependence on pH is seen for the pre-transalkylation reaction (Figure 2.12). Labelling efficiencies of 25% and 17% at pH 5.7 and 7.2 are seen respectively.

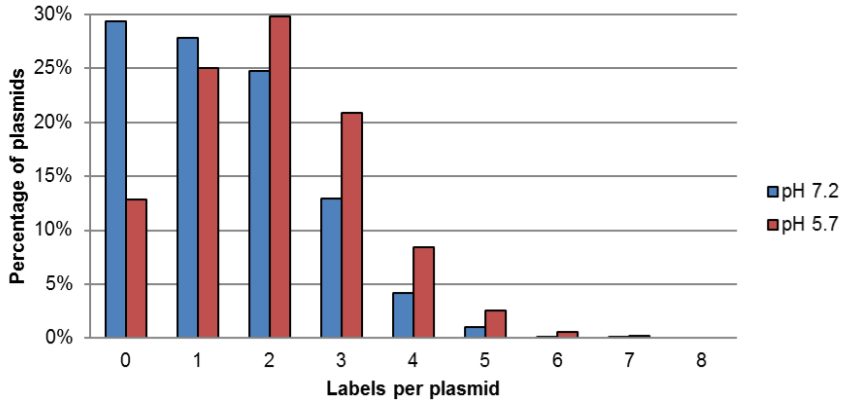


Figure 2.12 Variation in labelling efficiency at pH 7.2 vs pH 5.7. M.TaqI labelling of pUC19 with AdoHcy-azide coupled pre-transalkylation to TAMRA-DBCO. Single molecule counting results. Labelling at pH 7.2 (blue) and pH 5.7 (red), with the same concentration of M.TaqI.

An alternative approach is to return to SPAAC coupling post-transalkylation. It has previously been shown that coupling efficiency is high and the kinetics fast[141]. This also has the advantage that coupling with AdoHcy-azide alone is not so dependent on pH (Figure 2.13), lanes 2-7 at both pHs show similar restriction patterns. This is either because the cofactor is more stable, as the side-chain introduced is not such a capable leaving group, or because the transalkylation reaction is faster and so the competing degradation of the cofactor is not as important. Single molecule counter results for this approach are reported in Figure 2.6-Figure 2.8.

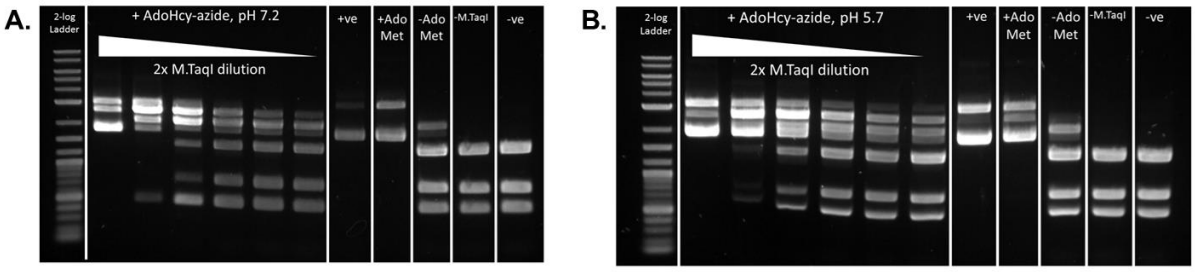


Figure 2.13 Variation in labelling efficiency at pH 7.2 vs pH 5.7. M.TaqI labelling of pUC19 with AdoHcy-azide. A) pH 7.2, B) pH 5.7. Lane 1 = 2 log ladder; lanes 2-7 AdoHcy-azide, 2x dilution of M.TaqI; lane 8, unrestricted pUC19; lane 9, AdoMet and M.TaqI; lane 10, M.TaqI only; lane 11, AdoHcy-azide only; lane 12, restricted pUC19.

The decomposition of the AdoMet analogues can be followed by HPLC[81] (carried out by Andrew Wilkinson). Figure 2.14A shows the HPLC trace of AdoHcy-Azide over the course of 160 minutes, in CutSmart, pH 5.7 and 7.2 at 50°C. The peaks can be assigned to the cofactor (3) and its breakdown products (1, 2, 4, 5) and the area of these peaks used to report on the concentration of the cofactor during a typical labelling reaction. Figure 2.14B shows that after little over an hour (around the length of a typical reaction) the cofactor is completely degraded at pH 7.2, but at pH 5.7 only around half of the initial cofactor has degraded.

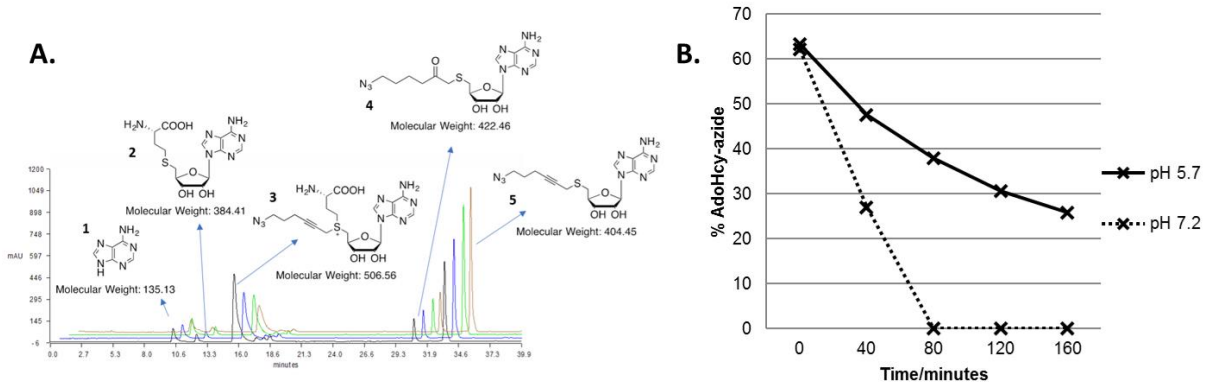


Figure 2.14 Decomposition of AdoHcy-azide. A) HPLC trace for breakdown of AdoHcy-azide. Taken over 160 minutes, every 20 minutes, incubated at 50°C at pH 5.7 or pH 7.2. Peaks can be assigned by mass spectrometry. B) Decomposition of AdoHcy-azide over 160 minutes at 50°C, pH 7.2 (green), pH 5.7 (red). HPLC experiments carried out by Andrew Wilkinson.

These results suggest that the decomposition of the AdoMet analogues is happening rapidly enough to compete with the transalkylation reaction. These reactions can be driven to completion by a high concentration of M.TaqI, but this is undesirable due to labelling by bound AdoMet. However careful consideration of the labelling strategy and reaction conditions can be used to maximise labelling efficiency.

2.2.5 Reaction conditions

It is desirable to use the minimum amount of methyltransferase, not only to reduce the effect of bound AdoMet, but also to reduce the cost of materials, particularly when labelling large quantities of DNA. In the previous section it was shown that varying the pH of the reaction mixture can greatly affect enzyme activity. As well as the pH there are a number of factors that should be considered, including the composition of the reaction buffer (e.g. the buffer system, types of salt, additives) and the reaction conditions (e.g. temperature and time). The general reaction buffer components and reaction conditions are shown in Table 2.1.

Reaction buffer: NEB CutSmart

20mM Tris-acetate ← Buffer
50mM Potassium Acetate } Salts
10mM Magnesium Acetate
100µg/ml BSA ← Additive

pH 7.9@25°C
~pH 7.3@50°C

Reaction conditions: 1 hour @ 50°C

Table 2.1 Reaction conditions for M.TaqI-directed labelling. There are two main parts of the reaction that should be optimised: the reaction buffer, for which the pH, buffer system, salts and additives should be considered; and the reaction conditions, in which the time and temperature of reaction should be considered.

Buffers are used to keep the pH of a solution constant and are usually composed of a weak acid and one of its salts. They will resist changes in pH when a small amount of acid or base are added to a solution, by removing the added protons or hydroxide ions. The range of pH that a buffer is useful for is based on the pKa of the weak acid, HA,

where pH = pKa + log [A-]/[HA]. At the mid-point of the buffering range pH = pKa and [A-] = [HA], and small amounts of protons or hydroxide ions will not affect the pH.

Ideally biological buffers should have a pKa of around 6-8 (the pH range most enzymes work at), a low ionic strength and be inert in the biological system, highly soluble and readily available. In the past inorganic buffers were used for biological systems, e.g. phosphate, borate, bicarbonate, however many of these buffers are not inert, for instance they can inhibit enzymes. So these are usually replaced by buffers developed by Norman Good and colleagues[150–152]. These remain crucial tools in biology and examples of these are shown in Figure 2.15.

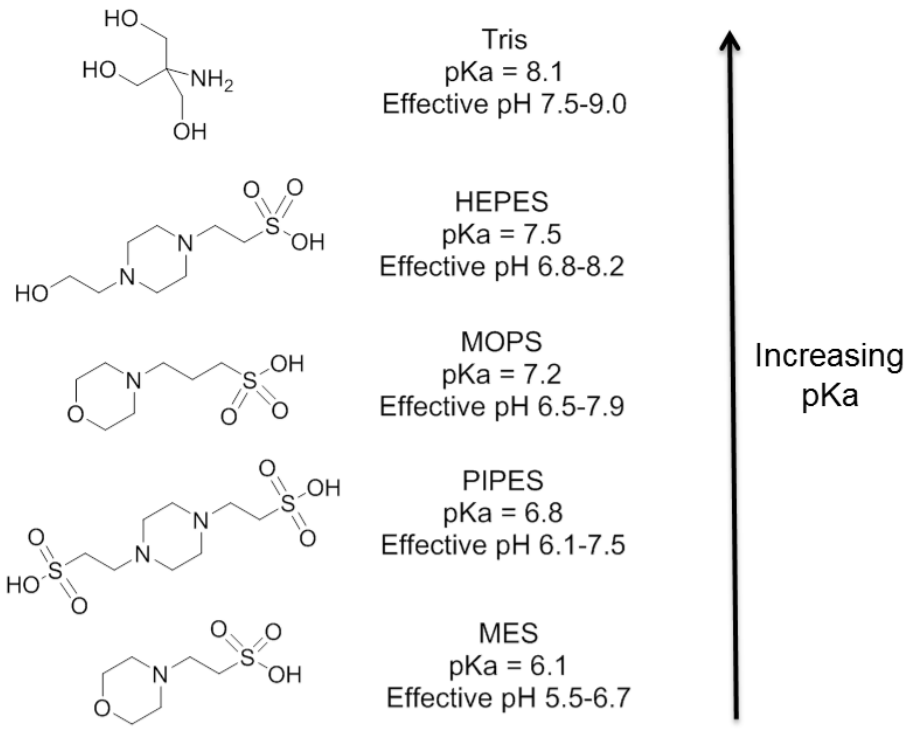


Figure 2.15 Good's buffers. These are readily available, highly soluble buffers, which are commonly used as biological buffers as they have pKa values of 6-8 and are relatively inert. HEPES and PIPES are based on piperazine and MOPS and MES are based on morpholine. Here the buffers are presented in increasing pKa, which gives an indication of the effective pH range.

In the commercially available NEB Cutsmart buffer, Tris is used as the buffer, and it is commonly used elsewhere, since it is highly soluble, inexpensive, has a high buffer capacity and is generally inert. However, it has a number of drawbacks. The pKa of Tris is 8.06, which is at the top end of desired pH for most biological systems, and it has a high temperature and concentration dependence. For instance, at 25^{o}C NEB CutSmart is pH 7.9, but at 50^{o}C (the reaction temperature for M.TaqI) it is around pH 7.3. Tris is also a primary amine, which will couple to NHS esters and is cationic, therefore interacts strongly with DNA. This is known to cause problems for DNA-binding enzymes, for instance the restriction enzyme EcoRV is less active in Tris buffer[153], and the DNA methyltransferase Dam is inhibited by Tris[154].

Based on this knowledge a number of Good's buffers were trialled, to investigate the effect on M.TaqI (Figure 2.15). These range in pKa from 6.1 to 7.5 and therefore at around pH 7 will charged to different extents e.g. MES is largely anionic at pH 7 and will not interact with DNA. The results of these tests for M.TaqI and AdoHcy-azide, coupled to TAMRA pre-transalkylation are shown in Figure 2.16. There is no significant difference between buffers (lanes 2-5), suggesting that buffer identity does not affect M.TaqI, unlike Dam methyltransferase.

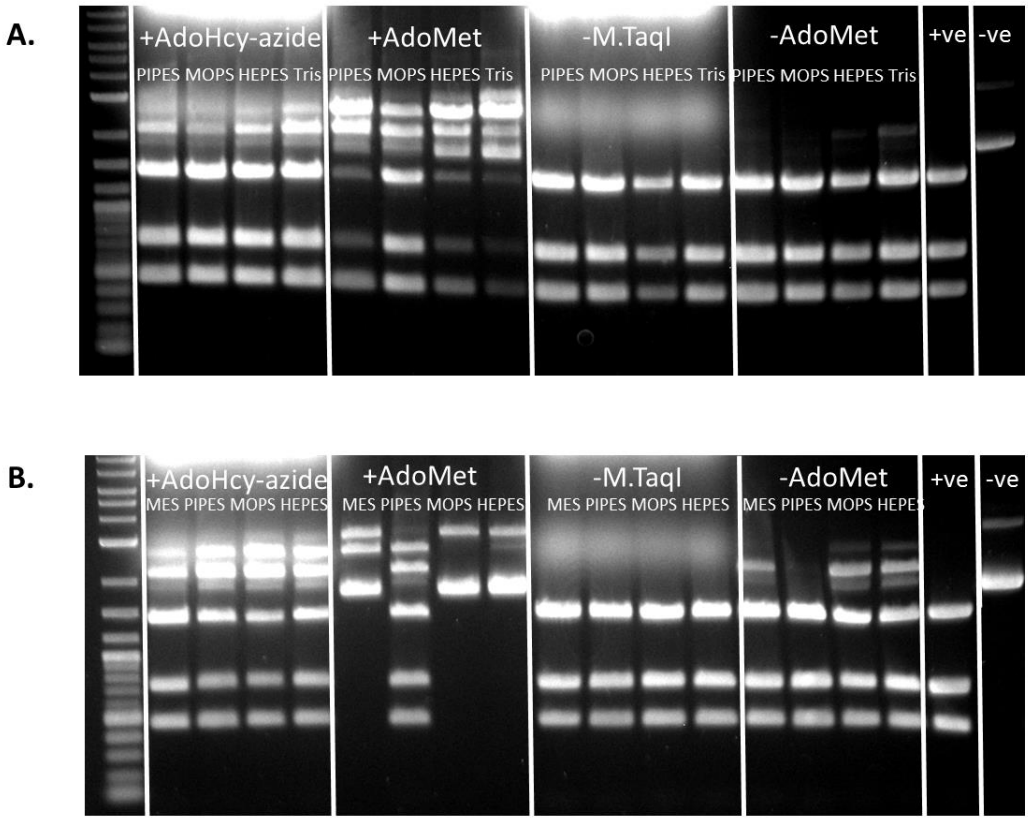


Figure 2.16 Variation in labelling efficiency with buffer system. M.TaqI labelling of pUC19 with AdoHcy-azide, coupled pre-transalkylation to TAMRA. A) pH 7.3: PIPES, MOPS, HEPES and Tris. B) pH 6.7: MES, PIPES, MOPS, HEPES. Lane 1 = 2 log ladder; lanes 2-5 = AdoHcy-TAMRA; lanes 6-9 = AdoMet; lanes 10-13 = AdoHcy-TAMRA only; lanes 14-17 = M.TaqI only; lane 18 = restricted pUC19; lane 19 = unrestricted pUC19.

Another component of the labelling buffer is the identity and concentration of salt. Cations and anions can interact with amino acid residues and therefore affect the structure and function of enzymes. This hasn't been tested in detail, however Figure 2.17 shows an example where pH was adjusted by varying amounts of sodium hydroxide, and alongside this the same amount of sodium in the form of sodium chloride. As well as increasing protection at lower pH (lanes 2-7), as seen previously, this shows increasing protection at decreasing sodium concentrations (lanes 8-13). This suggests that salt concentration may also be important, as well as pH, and perhaps should be investigated further.

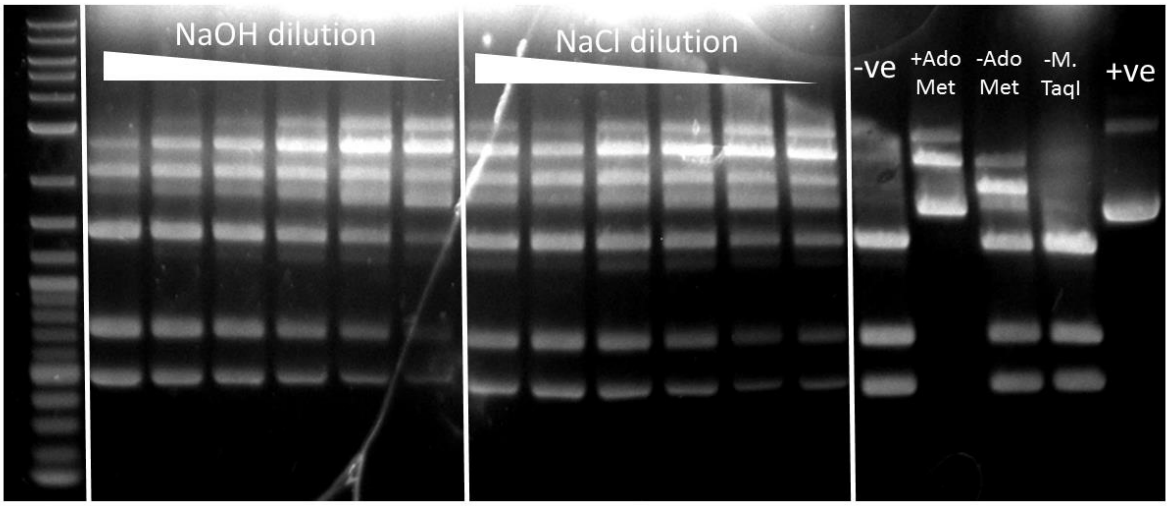


Figure 2.17 Variation in labelling efficiency with salt concentration. M.TaqI labelling of pUC19 with AdoHcy-azide, coupled pre-transalkylation to TAMRA, in MES CutSmart, pH 5.75. Lane 1 = 2 log ladder; lanes 2-7 = AdoHcy-TAMRA, 2x dilution of NaOH, lane 2 = 10mM NaOH, ~pH 6.5; lanes 8-13 = AdoHcy-TAMRA, 2x dilution of NaCl, lane 8 = 10mM NaCl; lane14 = restricted pUC19; lane 15 = AdoMet control; lane 16 = no cofactor control; lane 17 = no M.TaqI control; lane 18 = unrestricted pUC19.

The final part of the reaction buffer are the additives. These are generally added to stabilise and increase the activity of enzymes, for instance BSA is added to CutSmart for these reasons and to prevent dilute protein solutions degrading or binding to the reaction tube. Other additives can be used such as: DTT, to reduce disulphide bonds that form between cysteine residues (e.g. Grunwald *et al*[155]); a surfactant, such as Triton X-100, to prevent protein aggregation; or EDTA, to chelate multivalent cations. In the reactions carried out here DMSO is commonly added to improve solubility of dyes, which are generally poorly soluble in aqueous solution.

As well as the reaction buffer, the reaction conditions should be carefully controlled to optimise the reaction. For example, M.TaqI has a maximum efficiency at 65°C and only around a quarter as much activity at 37°C. However, the cofactor will decompose faster at higher temperatures, so a balance is necessary. Figure 2.18 shows that in an hour the

reaction does not go to completion at 30-40°C (lanes 7 and 8), but is complete at 50°C (lane 9), the temperature that has been used for all other reactions.

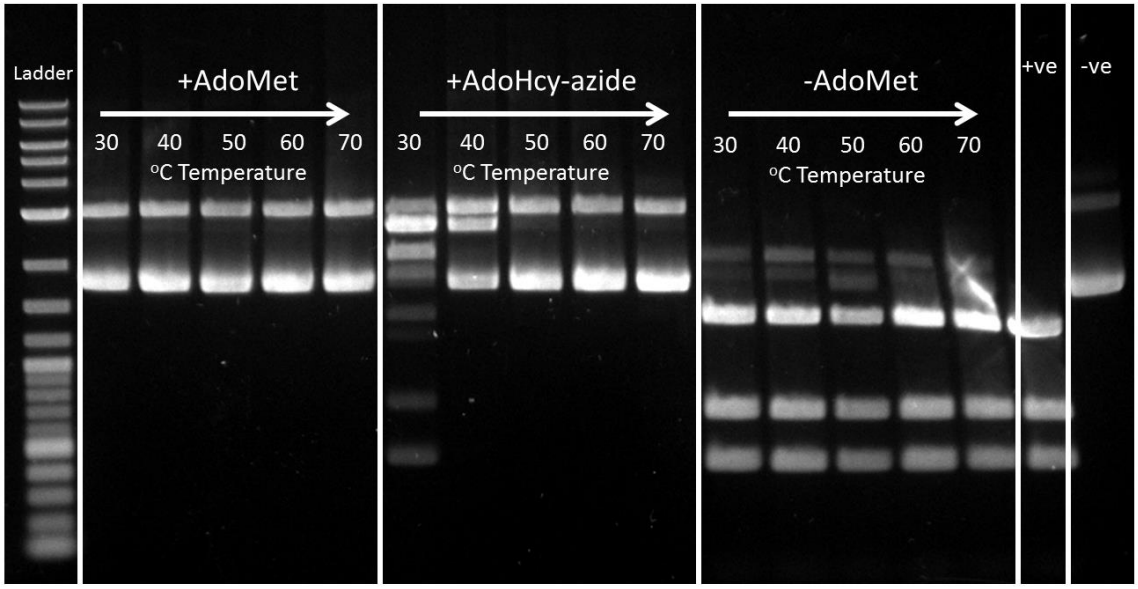


Figure 2.18 Variation in labelling efficiency with temperature. M.TaqI labelling of pUC19 with AdoHcy-azide. Lane 1 = 2 log ladder; lanes 2-6 = AdoMet, 30-70°C; lanes 7-11 = AdoHcy-azide, 30-70°C; lanes 12-16 = no cofactor, 30-70°C; lane 17 = restricted pUC19; lane 18 = unrestricted pUC19.

The time the reaction takes depends on how rapidly the transalkylation occurs. This is shown in Figure 2.19, which shows when AdoMet is used, at high M.TaqI concentrations, the reaction is very rapid and complete protection is seen within 5 minutes (lanes 2-6). In contrast the AdoHcy-azide reaction is not complete until around 40 minutes (lanes 7-11), which is on the same timescale as cofactor decomposition (Figure 2.14) and why decomposition of the cofactor is so important to consider.

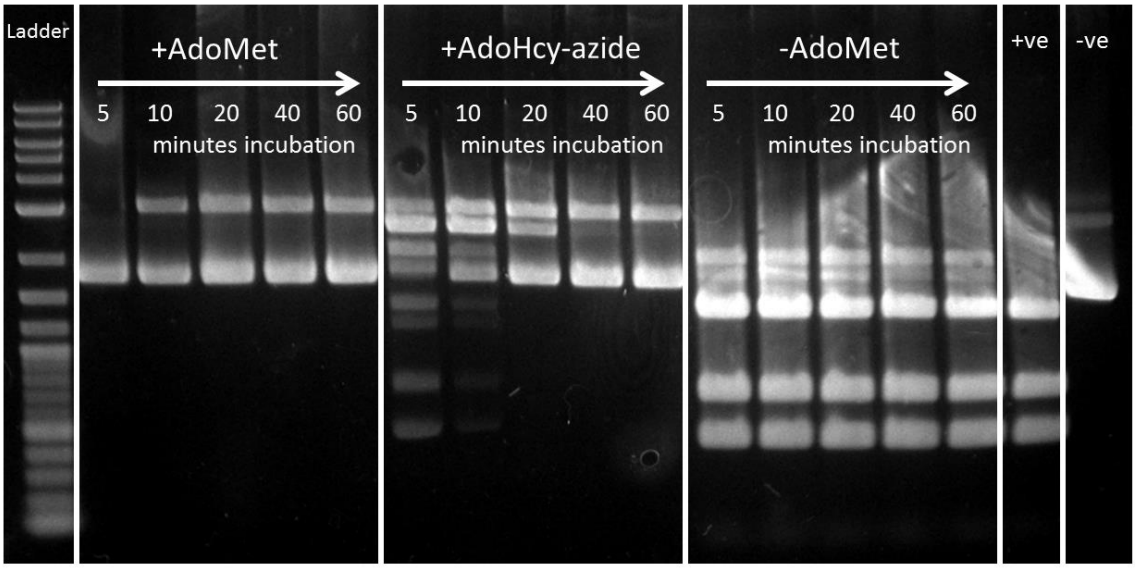


Figure 2.19 Variation in labelling efficiency with time of reaction. M.TaqI labelling of pUC19 with AdoHcy-azide. Lane 1 = 2 log ladder; lanes 2-6 = AdoMet, 5-60 minutes; lanes 7-11 = AdoHcy-azide, 5-60 minutes; lanes 12-16 = no cofactor, 5-60 minutes; lane 17 = restricted pUC19; lane 18 = unrestricted pUC19.

The reaction conditions can also be controlled to reduce the effect of cofactor decomposition. For example, AdoHcy is one of the decomposition products, will also be generated as a by-product during labelling, and is known to inhibit methyltransferases. Therefore, removal of AdoHcy should result in an increased rate of reaction and a reduced concentration of M.TaqI should be required. AdoHcy hydrolase is an enzyme which catalyses the hydrolysis of AdoHcy to adenosine and homocysteine, effectively removing AdoHcy from the reaction mixture (Figure 2.20A). At low methyltransferase concentrations where inhibition by AdoHcy is important, this can be used to increase the rate of reaction (Figure 2.20B, lanes 2-8).

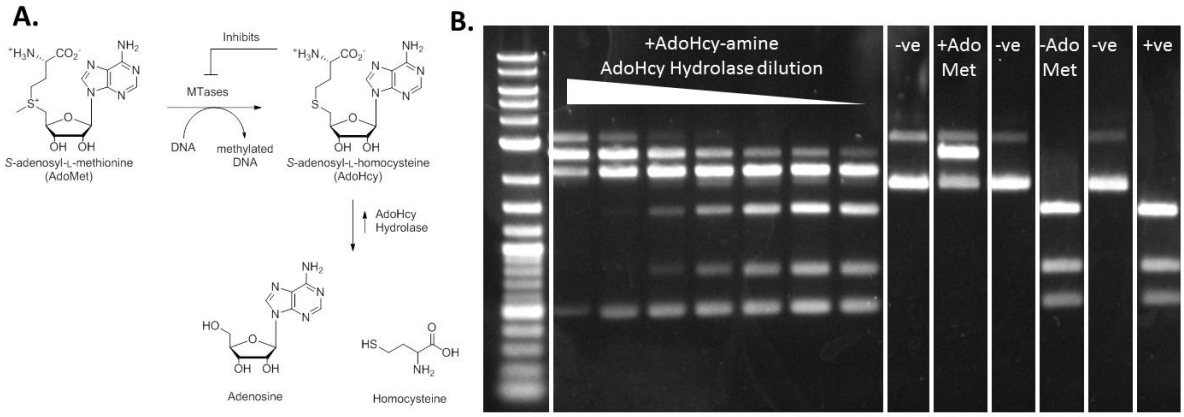


Figure 2.20 Effect of AdoHcy hydrolase. A) General reaction scheme. Methyltransferases use AdoMet for labelling of DNA, but the product, AdoHcy inhibits this reaction. AdoHcy Hydrolase removes AdoHcy by hydrolysis to adenosine and homocysteine. B) Variation in labelling efficiency with concentration of AdoHcy Hydrolase. M.TaqI labelling of pUC19 with AdoHcy-amine. Lane 1 = 2 log ladder; lanes 2-8 = AdoHcy-amine, 2x dilution of AdoHcy Hydrolase; lanes 9, 11, 13 unrestricted pUC19, 5-60 minutes; lane 10 = AdoMet control; lane 12 = no cofactor control; lane 14 = restricted pUC19.

2.2.6 Cofactor and dye purity and choice

A range of factors which affect the rate of reaction have been discussed. By reducing the required methyltransferase concentration, the amount of labelling by bound AdoMet, which results in discrepancies between restriction assays and single molecule counting experiment, can be reduced to a minimum. However, the purity of the dye and cofactor can also lead to differences in restriction assay and single molecule counting results. If the correct reactive chemical moiety is not transferred to the DNA and there is no fluorescent dye, then the DNA may be protected but no fluorophore will be counted.

The AdoHcy-azide cofactor is 61% pure, after storage at -20^{o}C for two months, as determined by Mass spectrometry following HPLC (Figure 2.21A, carried out by Andrew Wilkinson). However, it is important to note that although some of the breakdown products will slow the reaction (for instance AdoHcy, peak 2) by inhibiting M.TaqI none

of the breakdown products are AdoMet analogues and therefore none should lead to protection of DNA by methyltransferases. This is therefore not expected to be a major contribution to the discrepancy between the restriction assays and single molecule counting results.

The DBCO-TAMRA dye used in all SPAAC coupling reactions thus far is purchased commercially and is around 94% pure by HPLC (Figure 2.21B). The two major peaks can be attributed to the two isomers supplied. However, this does not report on how much of the dye is fluorescent, which is an important consideration. If a fluorophore is in a dark state, e.g. photobleached, oxidised or quenched, then it will not be counted in single molecule counting experiments and this will reduce the apparent labelling efficiency.

For example the photophysical properties of TAMRA are known to be affected when conjugated to DNA and other biomolecules[156,157]. TAMRA is hydrophobic and therefore tends to aggregate in aqueous solutions at high concentrations, for instance when local concentration is increased by labelling of the same biomolecule. These dye-dye interactions can lead to fluorescence self-quenching but also a splitting of the normal absorption peak at around 550 nm to obtain an additional peak at 520nm. These kinds of effects may reduce the apparent labelling efficiency, even if DNA has been labelled with a fluorescent dye.

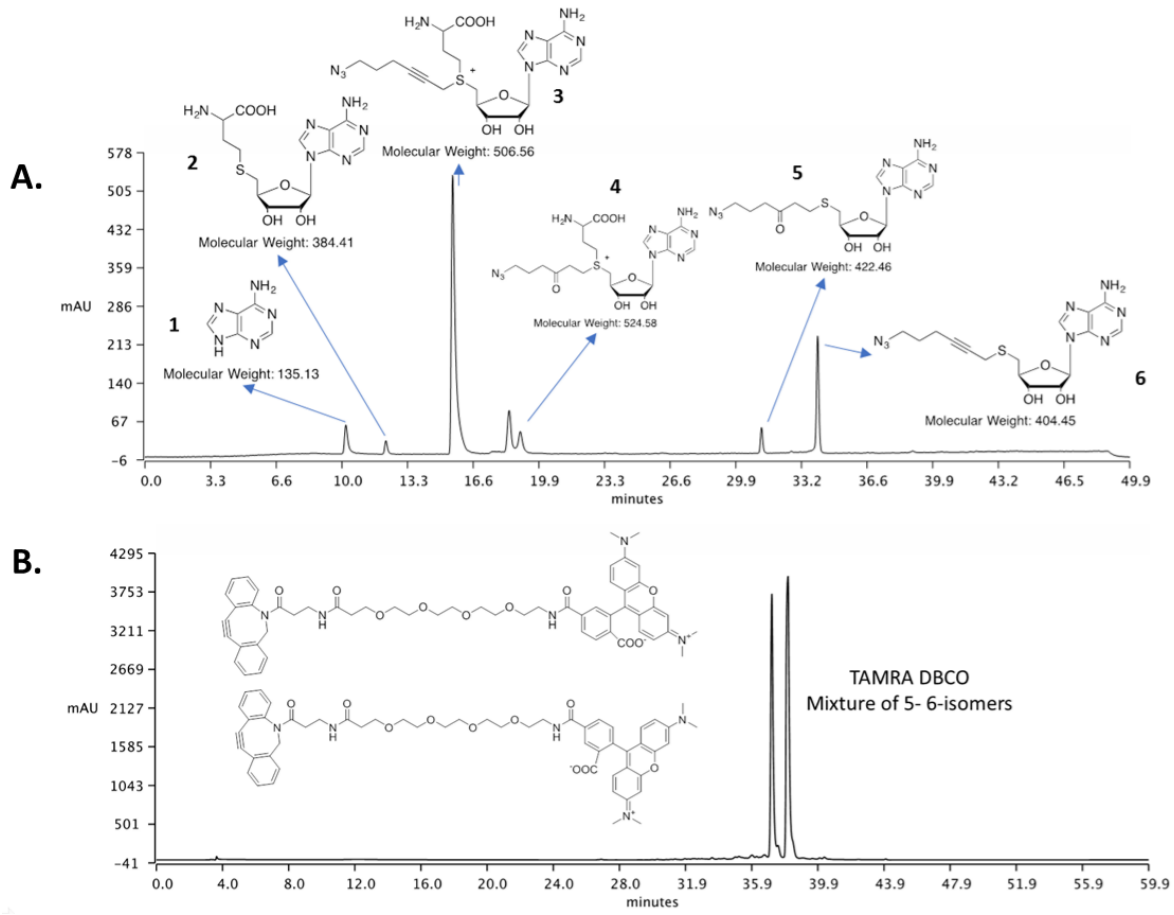


Figure 2.21 **HPLC traces to test purity of cofactor and dye. A) HPLC of AdoHcy-azide after 2 months @-20°C, 61% pure. The major peak is the AdoHcy-azide cofactor (3), however decomposition has occurred, primarily via loss of the amino acid (6). AdoHcy is only a minor product (2). B) HPLC of TAMRA-DBCO, 94% pure. Two major peaks are seen, corresponding to the two isomers.**

To consider these effects several different commercial dyes have been tested, the structures of which are given in Figure 2.22A. Here TAMRA can be directly compared to two alternative DBCO dyes, Texas Red and Cy5.5 and to Atto647N and Atto565 which were NHS esters coupled via an amine-DBCO intermediate. Single molecule counting results are given in Figure 2.22B and Figure 2.22C, for azide-modified and unmodified pUC19, coupled to dyes overnight.

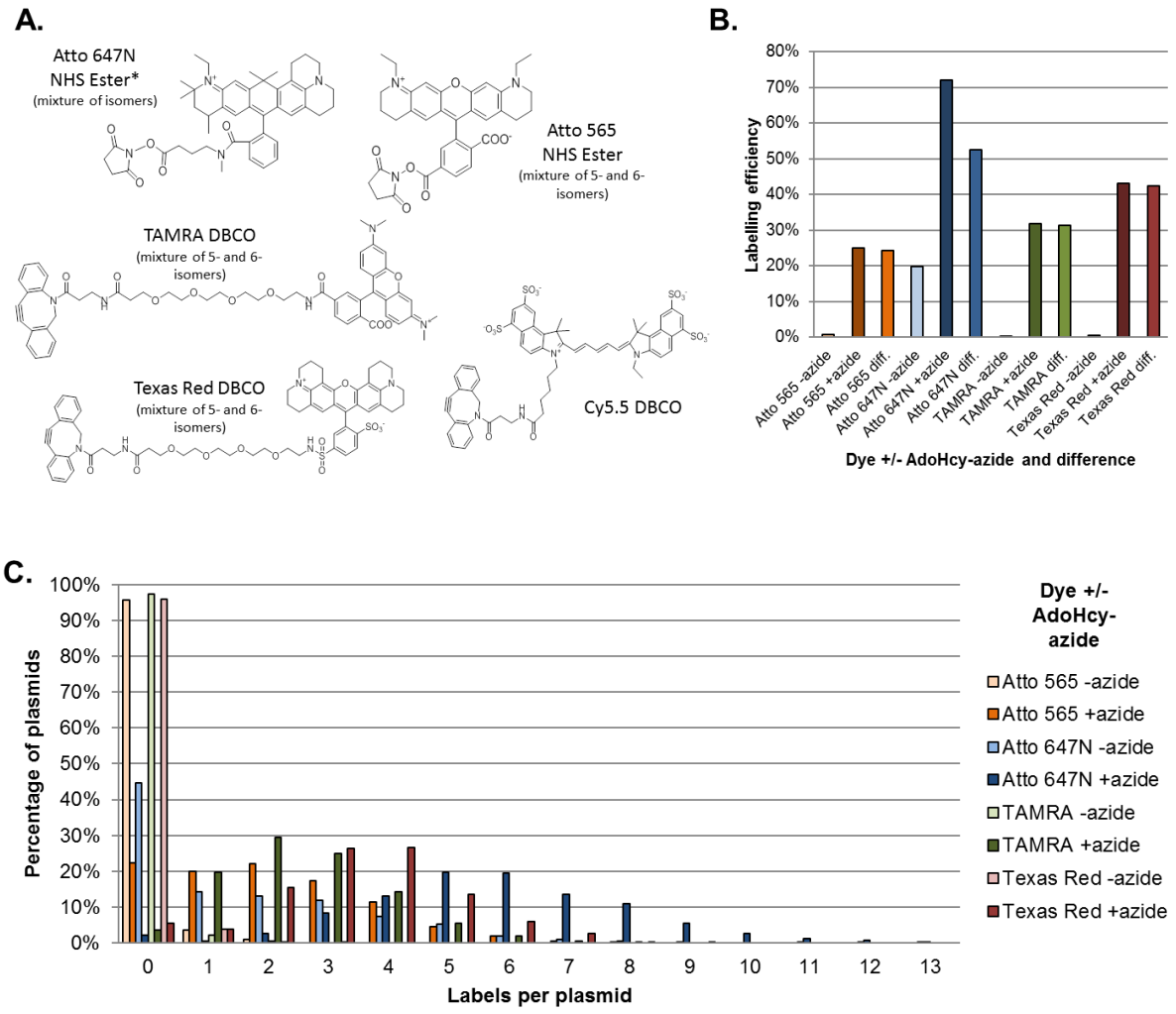


Figure 2.22 Single molecule counting results for different commercial dyes. All were coupled post-transalkylation. pUC19 was unlabelled or labelled with M.TaqI and AdoHcy-azide. A) Structures of commercial dyes, Atto647N NHS Ester*, Atto565 NHS Ester, TAMRA DBCO, Texas Red DBCO and Cy5.5 DBCO. B) Labelling efficiencies for each dye, unlabelled and labelled pUC19 and the difference. C) Single molecule counting results for each dye, unlabelled and labelled pUC19. *Full structure not available from supplier but with perchlorate ion this would fit molecular weight.

Cy5.5 is an anionic dye and gives no labelling[141], whilst Atto647N is cationic and shows an especially large degree of non-specific labelling, i.e. labelling of unmodified DNA. Of the neutral dyes Atto565 gives the poorest labelling, due to either the purity of the dye (only given as ≥70% by the manufacturer) or the extra coupling step. Texas Red gives the best labelling but is within the range of results seen previously for TAMRA. The

degree of photobleaching or quenching is difficult to speculate on, however it is clear from these results that the choice and condition of the dye is important to achieve the maximum apparent labelling efficiency.

2.2.7 Other methyltransferases

In the literature there several different enzymes that have been used for methyltransferase-directed labelling of DNA, showing the flexibility of this labelling approach. Various sequence specificities are available to give different labelling densities, which may be useful for different applications. The results above are all for M.TaqI, which has proved particularly robust for this application, possibly due to the nature of the cofactor binding site and the high pH and temperature tolerance inherent to the enzyme. However, other enzymes will have different optimum reaction conditions and may favour different labelling strategies (e.g. different cofactors).

Here two alternative methyltransferases have been used to fluorescently label DNA: M.HhaI (5'-G<u>C</u>GC-3') and M.MpeI (5'-<u>C</u>G-3'). The main characteristic of these enzymes is that they will label DNA with different densities. For example, for pUC19 there are only four M.TaqI sites, but 17 M.HhaI and 173 M.MpeI sites.

Restriction assays are shown for M.HhaI and M.MpeI in Figure 2.23 and show that labelling is incomplete. For commercially available M.HhaI there is negligible activity. This is consistent with the literature, since it has been reported that mutations of the binding site are required to activate the enzyme for transfer of extended groups[90]. M.MpeI protection by AdoHcy-azide appears to be complete, however there appears to be a large amount of bound AdoMet present. Here the number of sites is very large, which will make full protection difficult if it is not optimised.

Single molecule counting experiments can be used to support these results. There is no non-specific labelling for dye coupling, post-transalkylation, with TAMRA and the number of labels directed by M.HhaI is insignificant. When compared to M.TaqI there should be ~60x more labels for M.MpeI, however this is not seen. Labelling density is only marginally increased for M.MpeI when compared to M.TaqI, suggesting that labelling efficiency is low.

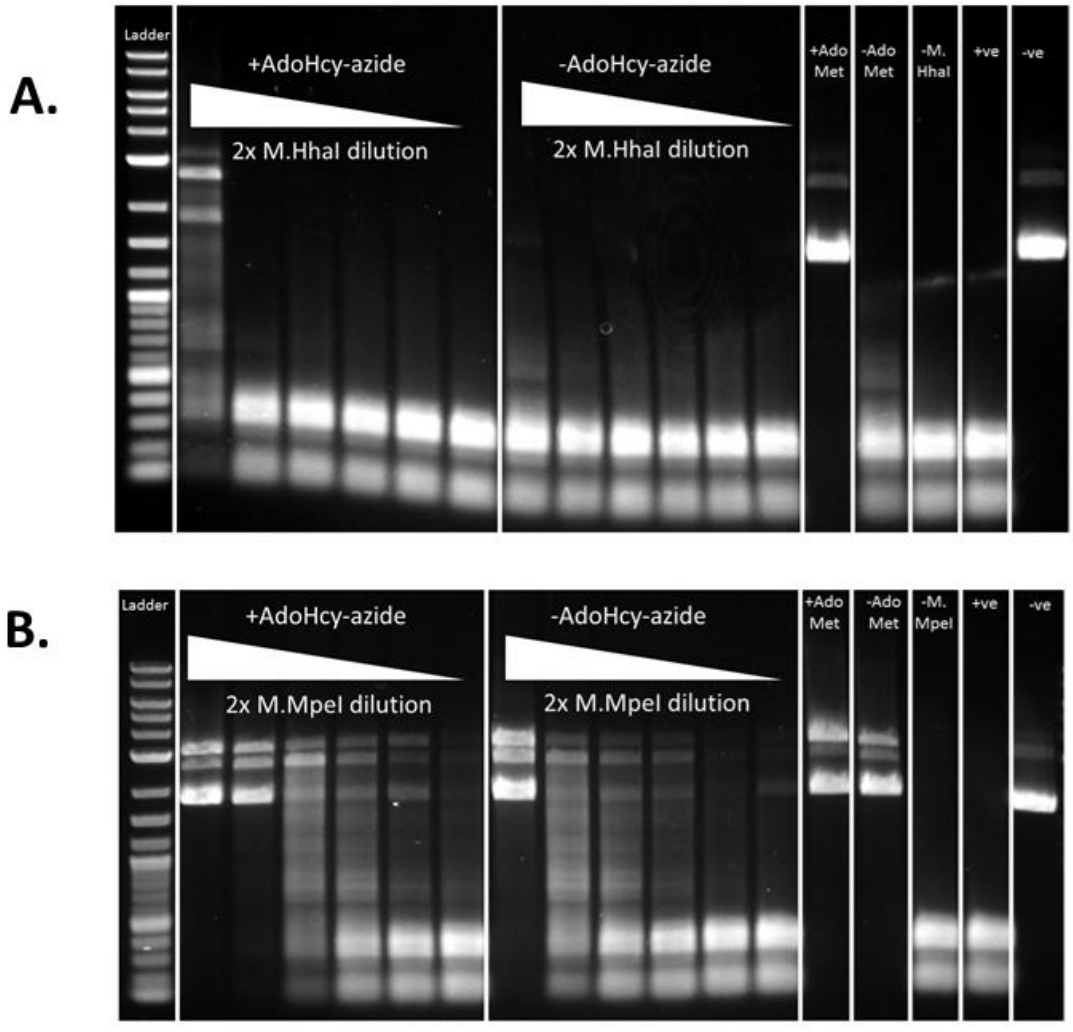


Figure 2.23 Restriction assays for alternative methyltransferases: A) M.HhaI and B) M.MpeI. Lane 1 = 2 log ladder; lanes 2-7 = AdoHcy-azide, 2x dilution of methyltransferases; lanes 8-13 = no AdoHcy-azide, 2x dilution of methyltransferases; lane 14 = AdoMet control; lane 15 = no cofactor control; lane 16 = 9, 11, 13 unrestricted pUC19, 5-60 minutes; lane 10 = AdoMet control; lane 12 = no cofactor control; lane 14 = restricted pUC19.

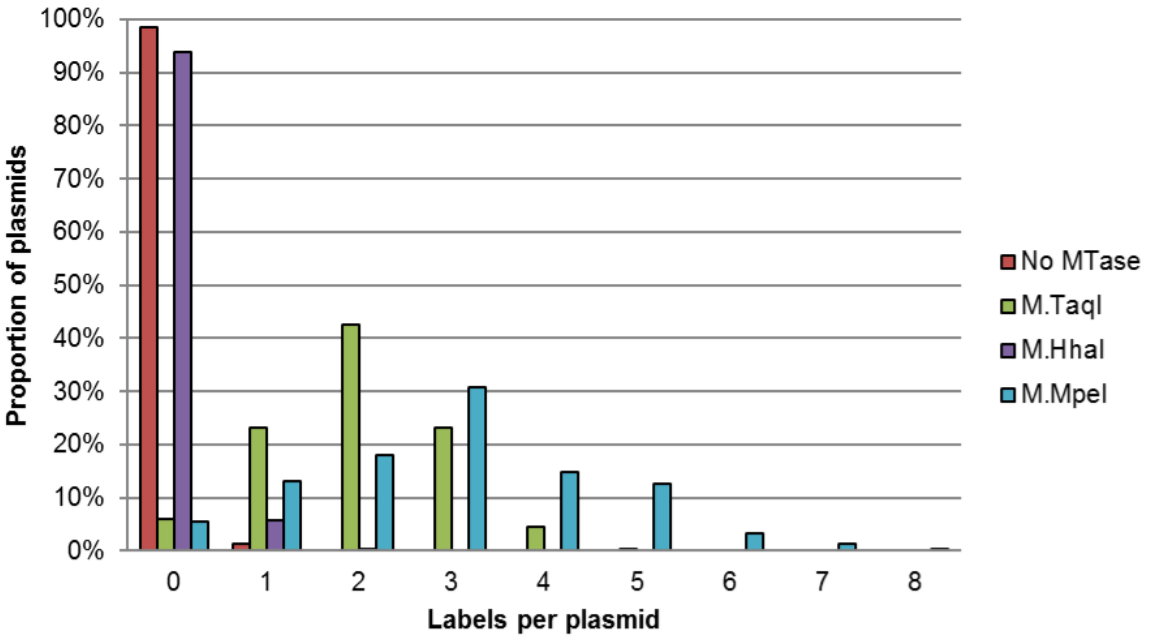


Figure 2.24 **Single molecule counting results for alternative methyltransferases. pUC19 was labelled with AdoHcy-azide and coupled post-transalkylation to TAMRA-DBCO using: No methyltransferase, M.TaqI (four 5'-TCG<u>A</u>-3' sites), M.HhaI (seventeen 5'-G<u>C</u>GC-3' sites) and M.MpeI (173 5'-<u>C</u>G-3' sites).**

2.2.8 Reliability of single molecule counting

Single molecule photobleaching experiments have been used in a variety of systems to calculate labelling efficiency[158–160]. A range of labelling efficiencies has been reported, commonly around 70% for labelling of biomolecules using coupling strategies like those discussed here. However, the experimental limitations are generally not discussed in detail, although it is known that noise, high numbers of fluorophores and simultaneous bleaching events can all lead to underestimation in single molecule counting experiments[161]. When a brightly, but sparsely, labelled molecule bleaches we expect to see several bleaching steps, from which the number of fluorescent labels can be determined (Figure 2.2). However, consider an example intensity trace for a single molecule labelled by M.MpeI (Figure 2.25B). The brightness and number of fluorescent labels makes it difficult to reliably determine individual bleaching steps.

These problems are illustrated in Figure 2.25A. There is a rapid bleaching event at the beginning of illumination, followed by blinking of the molecule as fluorophores begin emitting once more, with relatively high noise. If bleaching is too rapid it is not possible to count the number of individual bleaching events, also a fluorophore that is in a dark state initially, but then emits later in the time trace will not be included in the counting results. Along with incomplete bleaching, these effects will lead to an underestimate in the number of fluorophores. There is also an issue with cumulative noise in the bleaching approach, since any incorrectly identified events will affect the rest of the bleaching movie.

These problems demonstrate how counting results will be unreliable for large numbers of labels. However, a comparison can still be made based on the initial intensity. For example, the intensity of M.MpeI-labelled pUC19 molecules should be ~50x that of M.TaqI-labelled molecules, however this is not the case. See for example the trace in Figure 2.25B for an M.MpeI-labelled pUC19 molecule, which starts at an intensity of around 16,000, and compare this to a trace for M.TaqI labelling in Figure 2.26A, which was recorded using similar laser intensity and gain, and begins at an intensity of around 8,000. These are typical results and reflect the incomplete protection from the restriction assay (Figure 2.23).

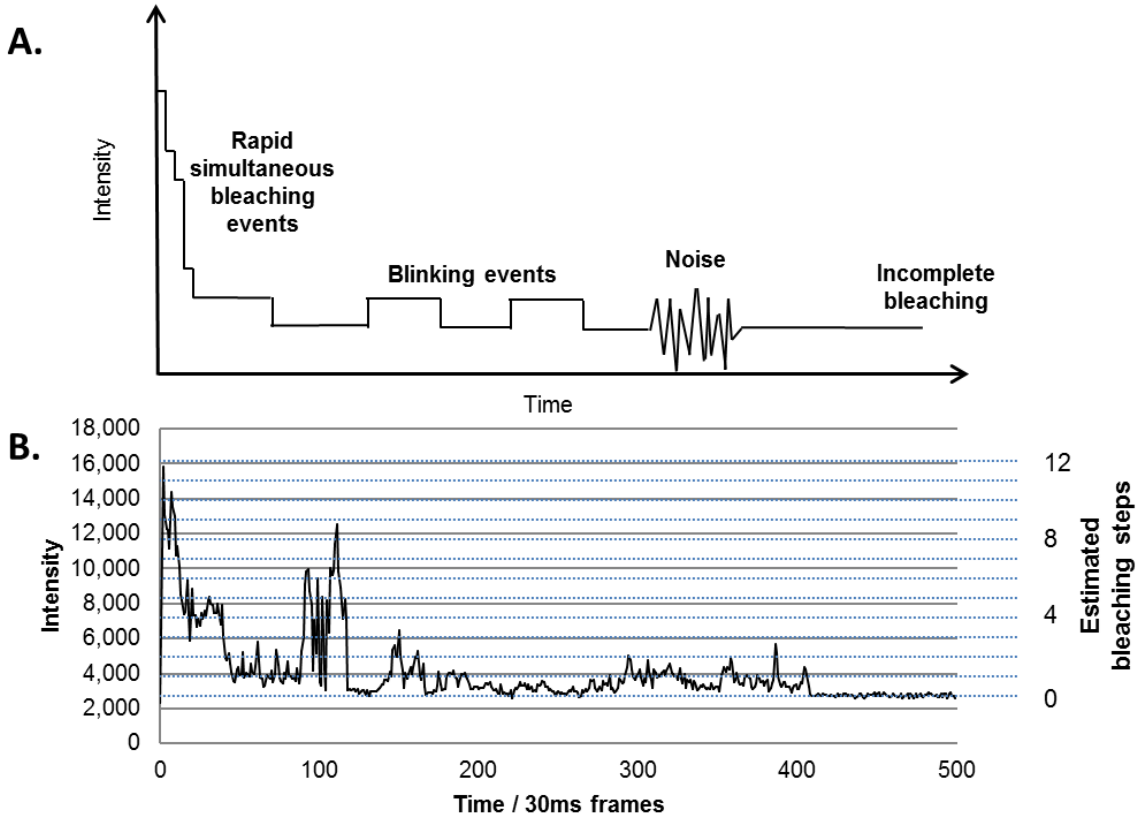


Figure 2.25 Single molecule counting limitations. A) Schematic of issues in single molecule intensity time traces. This shows rapid and simultaneous bleaching events, noise and blinking events, which can all lead to problems when estimating the number of fluorophores. B) An example intensity trace for M.MpeI-directed TAMRA-labelled pUC19. Here the limitations can clearly be seen, and it is difficult to reliably identify bleaching steps (blue dashed lines). However, the initial intensity does give an indication of the number of fluorophores.

Example traces for single molecules, for the results in Figure 2.22, are shown below (Figure 2.26). Each of these traces displays the rapid bleaching and blinking characteristics that will lead to an underestimate of the number of labels. TAMRA and Texas red in particular have lower signal to noise, therefore a relatively high laser intensity is needed, and they are rapidly bleached in these experiments. This could suggest that the difference between the dyes may be due to errors in counting, rather than differences in the inherent labelling efficiency. These problems are the main

limitations with single molecule counting approaches to assess labelling efficiency, however the apparent efficiency is still relevant when assessing the quality of labelling for optical mapping.

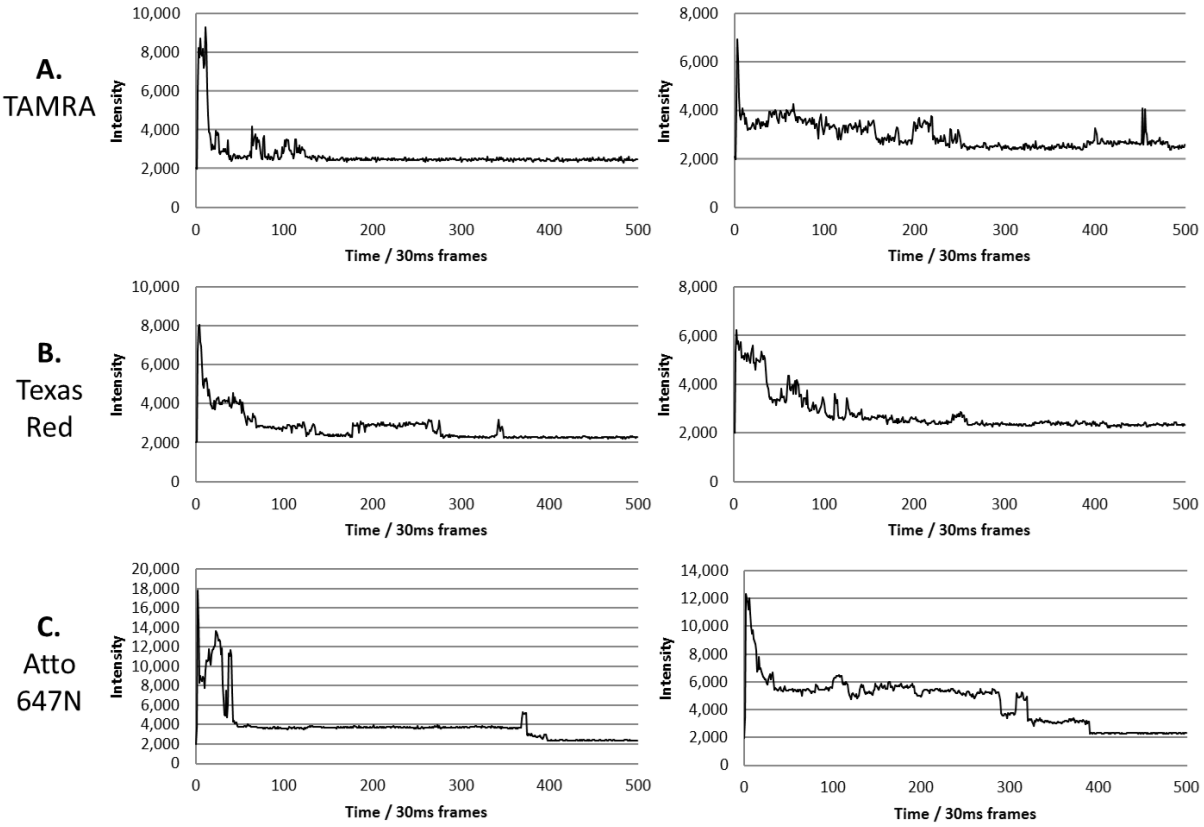


Figure 2.26 Intensity traces for single pUC19 molecules, labelled with AdoHcy by M.TaqI and coupled post-transalkylation with: A) TAMRA, B) Texas Red and C) Atto 647N. Each of these display similar characteristics: rapid simultaneous bleaching events at the beginning of the trace, followed by transient blinking events. TAMRA and Texas Red also have a large amount of noise in the traces. All these make estimation of the number of fluorophores difficult.

Experimentally it is also important to consider incomplete bleaching if photostable dyes are used, although this is not usually a problem for the dyes used here (e.g. Figure 2.26). Incomplete bleaching can be caused by low laser power, or alternatively a time trace that is too short. Supplementary Figure 7.4 shows that if care is taken to completely bleach the sample then single molecule counting results will be repeatable.

2.3 Conclusion

In CHAPTER 2 several factors have been considered to optimise methyltransferase labelling by synthetic AdoMet analogues. These factors have been discussed in detail for fluorescent labelling by M.TaqI, with applications for optical mapping and imaging DNA. However, these results will be broadly applicable to other labelling strategies, other methyltransferases and synthetic groups.

When labelling with methyltransferases it is important to consider the efficiency of each step of the reaction, as one would during any synthesis. The factors that influence labelling efficiency are shown in Figure 2.27A and include: 1) the purity of the starting materials including the cofactor and dye; 2) the yield of the coupling reaction; 3) the yield of the labelling reaction, including labelling by bound AdoMet and factors which can inhibit labelling; and 4) the imaging experiment, including factors which will lead to non-photoactive dye.

Since the research for this thesis was completed, results have been obtained by the Neely group that have shed further light upon the discrepancy between the restriction assays and single molecule counting results and provide more robust evidence for steps 1-3. Andrew Wilkinson labelled hairpin oligonucleotides 60 bp in length, which contained one TaqI site. The oligonucleotides were run on HPLC before labelling; after labelling with the cofactor; and after labelling with TAMRA.

One relatively broad peak in the HPLC trace is obtained for pure, unlabelled DNA. Labelled DNA splits into two peaks of equal size, attributable to each DNA with either one or two labels. There are no other significant peaks, in particular no peak that can be

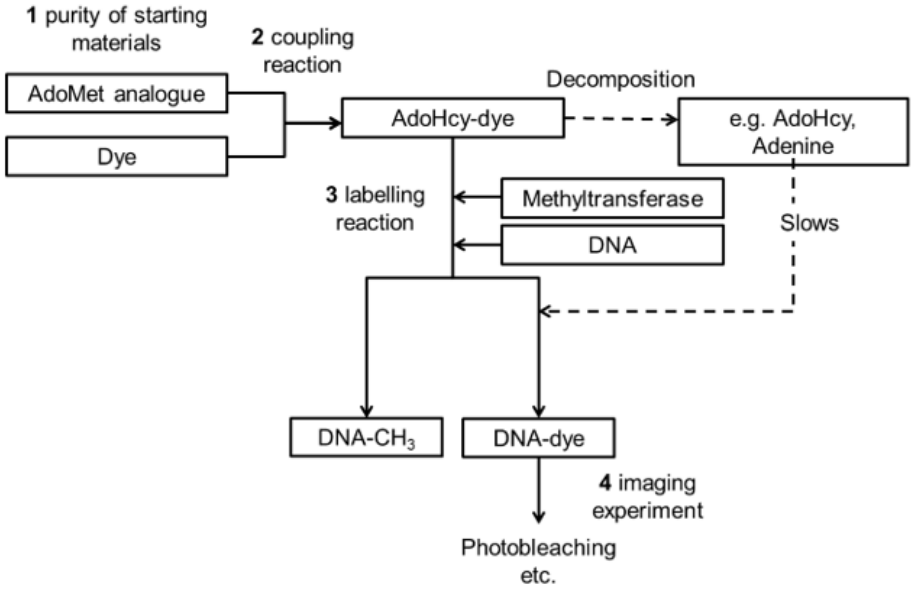
attributed to DNA labelled by bound SAM. This demonstrates a labelling yield (step 3) of around 75%. When the dye is coupled to the DNA there are still only two peaks obtained, meaning the yield of the coupling reaction (step 2) is greater than 90%. This evidence means that if labelling is carefully performed, greater than 75% efficiency for steps 1-3 can be achieved. However, for imaging experiments (step 4) the number of labels that are actually fluorescent will be limiting, as discussed in sections 2.2.6 and 2.2.8.

Regardless, fluorescent labelling efficiency approaching 50% has been demonstrated. The contribution of each step to this final efficiency is illustrated in Figure 2.27B. Steps 1-3 have been carefully optimised to achieve maximum efficiency (80% or greater), although it should be noted that a decrease in efficiency at any step can have a large effect on the overall labelling efficiency.

Step 4 appears limiting and is more relevant to the use of methyltransferase labelling for optical mapping. It appears that there is a large amount of dye molecules that are not fluorescent, although it is hard to speculate on the reasons for this. Therefore, alternative dyes should be tested to obtain the highest fluorescent labelling efficiency possible.

These results also demonstrate that simple restriction assays or HPLC experiments do not give a complete picture of fluorescent labelling and alternative methods, such as single molecule counting, should be used to give a more accurate measure of labelling efficiency when the labelling is to be used for visualisation of the DNA.

A



B

Step	Factor	Estimated effect on yield	Reference
1	Purity of cofactor	90-100%	HPLC (Section 2.2.6)
	Purity of dye	70-95%	HPLC (Section 2.2.6), manufacturer
2	Yield of coupling reaction	90-100%	HPLC (A. Wilkinson)
3	Yield of labelling reaction	50-100%	Restriction assays (Section 2.2.4, 2.2.5, 2.2.7), HPLC (A. Wilkinson)
	Labelling with SAM	80-100%	Restriction assays (Section 2.2.3), HPLC (A. Wilkinson)
4	Non-photoactive dye	20-70%	Counting experiments v restriction assays (Section 2.2.2), HPLC (A. Wilkinson)
	Counting experiment	50-80%	Counting experiments (Section 2.2.8)
Overall	**Steps 1-4**	**<50%**	**Counting experiments (e.g. Fig 2.6)**

Figure 2.27 **Overview of factors influencing labelling efficiency during fluorescent labelling of DNA by methyltransferases. A) Schematic illustrating the steps that must be considered to maximise labelling efficiency. B) Efficiencies of each step are estimated based on results and discussion as referenced. Overall, high labelling efficiency can be achieved for steps 1-3, but during step 4 there is a large decrease in efficiency due to non-photoactive dye (e.g. quenched or bleached)**

As well as providing a fluorescent labelling efficiency that is sufficient for optical mapping (see CHAPTER 3 and CHAPTER 4) and localisation of plasmids (see CHAPTER 5), these results should represent an important step towards broadening the application of methyltransferase-directed labelling for future experiments.

2.4 Materials and Methods

2.4.1 AdoMet analogues and Enzymes

AdoMet-azide (50 mM, in 0.1% formic acid) was provided by Andrew Wilkinson. AdoHcy-amine (15 mM, 0.1% formic acid), AdoCys-amine (8.1 mM, 0.1% formic acid) and AdoCys-azide (55 mM, 0.1% formic acid) were provided by Volker Leen's group, KU Leuven. They were prepared by methods as described by Lukinavičius *et al.*[81]

M.TaqI (0.3 mg/ml) was provided by Ashleigh Rushton and M.MpeI (10 mg/ml) provided by Su Wang. Both were expressed, purified and stored based on standard techniques. M.HhaI was purchased from New England Biolabs (NEB).

2.4.2 Restriction assays

Full details of the conditions for each restriction assay are given in the Appendix, section 7.2. A typical 10 µl lane would be prepared as follows: A 10 µl solution containing 1x CutSmart (NEB), 500 ng pUC19 (NEB), 0.15 µg M.TaqI and 320 µM AdoMet (NEB) is incubated at 50°C for 1 hour. 0.5 µl R.TaqI (20,000 units/ml, NEB) is added and the sample is incubated at 65°C for 1 hour followed by addition of 0.5 µl 18 mg/ml Proteinase K (NEB)/0.1% Triton X-100 (Sigma-Aldrich) and incubation at 50°C for 1 hour. 2 µl 6x purple loading dye (NEB) is added and the sample is run by gel electrophoresis followed by post-staining by gel red (Cambridge BioScience) and visualisation.

2.4.3 Single Molecule Counting – Labelling conditions

Full details of the conditions for each counting experiment are given in the Appendix, section 7.2. A typical 40 µl reaction would be carried out as follows. A 40 µl solution

containing 1x CutSmart, 2 µg pUC19, 0.6 µg M.TaqI and 750 µM AdoMet-azide is incubated at 50°C for 1 hour. 2 µl 20 mg/ml proteinase K is added and incubated at 50°C for 1 hour, before purification by GenElute PCR Clean-up kit (Sigma-Aldrich) and elution into 50 µl 1xTE (Sigma-Aldrich). 10 µl DMSO (Sigma-Aldrich) and 0.3 µl 50 mM TAMRA-DBCO (Jena Bioscience) are added and incubated at room temperature for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 50 µl 1xTE.

2.4.4 Single Molecule Counting – Imaging conditions

For imaging, a 50 µl mixture of 50% DMSO, 0.5xTE, 0.2 µM YOYO-1 and ~5 ng DNA was incubated at 37°C for 30 minutes. 450 µl 1xTE was added and 100 µl was placed on poly-l-lysine coated coverslips for ~30 seconds. Subsequently, the sample was washed with 3 ml molecular grade water and dried. Samples were imaged using an Olympus IX81 inverted widefield/epifluorescent/TIRF microscope equipped with 491, 560, 640 lasers, and two Hamamatsu CCD cameras (Orca R2 and high-speed ImageEm).

CHAPTER 3 OPTICAL MAPPING FOR IDENTIFICATION OF COMPLEX MIXTURES OF VIRAL AND BACTERIAL DNA – *IN SILICO* GENERATION AND ALIGNMENT OF DNA FRAGMENTS

Robert K. Neely and Iain B. Styles provided supervision and guidance for the research undertaken in this chapter. Nathaniel O. Wand (the author) designed, performed and analysed all labelling experiments including optical mapping experiments. Nathaniel O. Wand (the author) also developed and performed all extraction and alignment procedures unless otherwise stated.

3.1 Introduction

3.1.1 Identification of microorganisms

The rapid identification of microorganisms is critical for samples ranging from water[162] or soil[163], to clinical samples such as blood or urine[164]. One application is the rapid diagnosis of infections, which can help inform appropriate antibiotic therapy[111]. Current molecular diagnostics can take around two days, as bacteria are cultivated and then identified. Meanwhile, the patient is treated based on empirical observations and the likely pathogens, which leads to ineffective treatment and inappropriate use of antibiotics. Antibiotic resistance is now one of the greatest public health threats and by 2050 it has been estimated that the societal and financial cost, if not tackled, will be US$100 trillion[109].

Various methods have been employed to accelerate diagnosis without cultivation. DNA hybridisation techniques, (e.g. PCR, DNA microarrays), can identify specific pathogens or resistance genes. This is useful for following outbreaks or for common pathogens but will not allow for comprehensive screening since hybridisation techniques rely on

known targets. In contrast, DNA sequencing techniques (e.g. NGS, SMRT sequencing) can be used to recognise all pathogens. However, most high-throughput, low-cost technologies use short sequence reads, which lack the contextual information required to easily assemble large genomes, making the task of identifying pathogens non-trivial.

The base-pair resolution offered by sequencing is not necessary for the unique identification of pathogens. For example, restriction enzymes have been used since the 1970s to identify and map viral genomes[21], exploiting the natural sequence-specificity of enzymes. More recently, optical mapping of DNA has been demonstrated, using enzymatic and affinity-based approaches to generate unique fluorescent patterns for DNA identification. Optical mapping is a complementary technique to sequencing, as it provides long-range contextual information that can be used to rapidly identify both complex mixtures of DNA and large-scale genomic variations.

3.1.2 Optical mapping of DNA

Optical mapping of DNA fits into several broad categories: mapping in nanofluidic devices or by molecular combing; and sequence labelling by affinity or enzymatic approaches. A summary of each of these approaches, including examples, is given in Figure 3.1. There are also a number of excellent reviews[165,166].

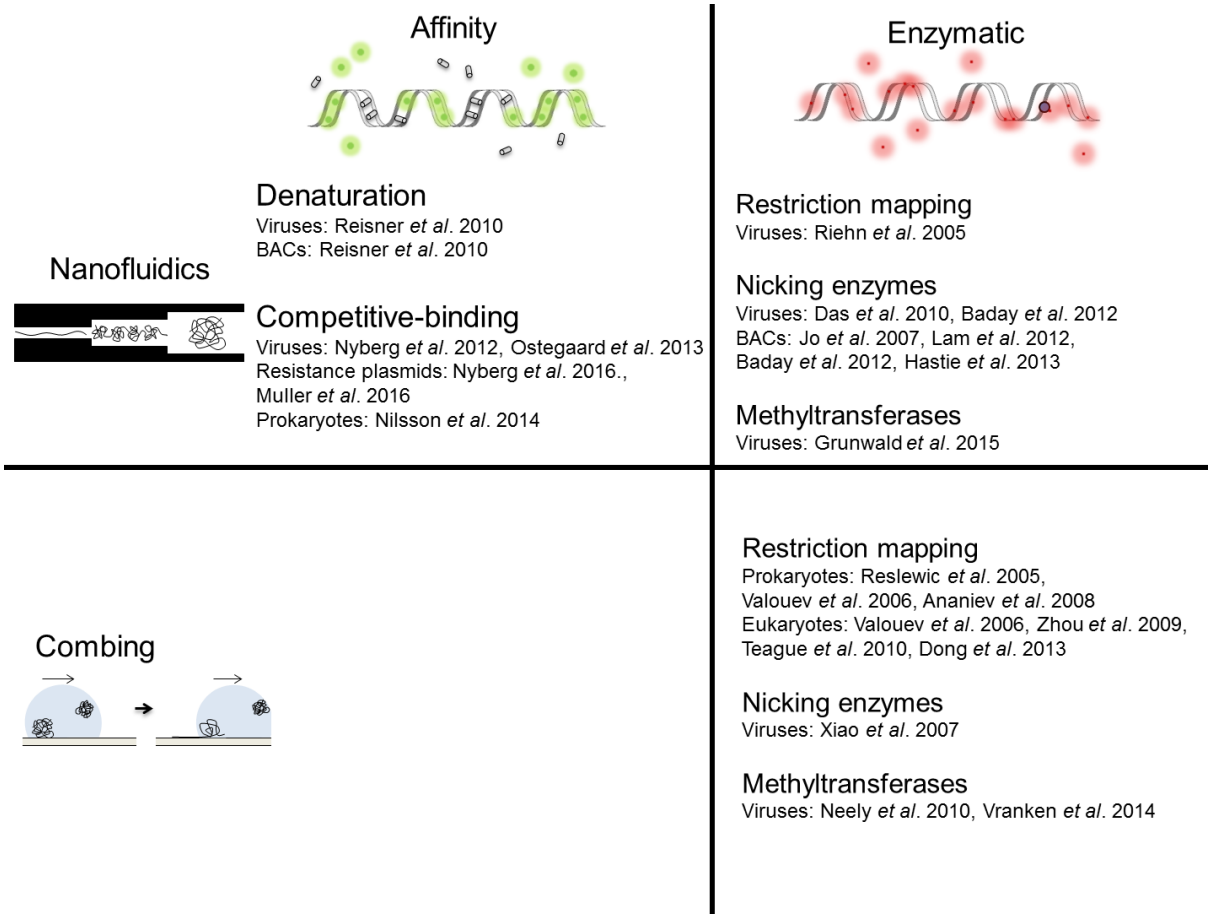


Figure 3.1 Overview of optical mapping for DNA identification. Optical mapping of DNA is split into several broad categories: mapping in nanofluidic devices or by molecular combing; and sequence labelling by affinity or enzymatic approaches. There are many examples available for these techniques, of which various publications are highlighted (and cited in the main text).

In nanofluidic devices DNA molecules are introduced to channels which have a diameter of the order of the persistence length of DNA (~50nm). DNA will be confined in the channels and extended to a length proportional to the length of the molecule. The main advantage of these devices is their flexibility and ease of use to image individual stretched DNA molecules. However, the thermal motion of DNA means the fluorescence pattern is blurred during imaging. This must be corrected by recording movies and kymograph alignment, although this results in longer image acquisition and a loss in resolution.

An alternative approach is to the deposition, linearization and fixation of DNA on a surface. This is generally achieved by molecular combing. At around pH 6, DNA will bind strongly and specifically to hydrophobic or cationic surfaces at its extremities[55]. Therefore, a receding meniscus can be used to uniformly stretch DNA across the surface. The main advantage over nanofluidic devices is that the DNA is fixed, enabling more straight-forward high-throughput or super-resolution imaging. However, if there are any sheared DNA molecules, or other impurities, in the sample then they will also be deposited on the surface which can make extracting the length and fluorescent pattern along individual DNA molecules impossible.

To generate a fluorescent pattern, enzymatic or affinity-based approaches have been used. The earliest optical mapping approaches were based on restriction enzymes, and these are still commonly used. Typically, these methods use molecular combing to deposit DNA on a surface, before digestion at specific sequences, producing an ordered restriction map up to several megabases in length. These have been used extensively on large genomes, particularly as a scaffold for *de novo* genome assembly and gap filling[48,167,168]. However, the low density of restriction sites used for mapping means large DNA molecules are required for reliable assignment to a reference. It is slow and difficult to extract large DNA fragments[169] which has made the identification of microorganisms challenging, although maps for several prokaryotes have been reported[170–172]. Restriction mapping has also been applied in microfluidic devices on lambda DNA[52], but the information content (i.e. the number and size of fragments) is too low for reliable identification of short genomes.

Affinity-based approaches have been used as an alternative to map shorter genomes. Typically, in these approaches the difference in hydrogen bonding between AT and GC base pairs is used to generate a unique pattern, either by exploiting the difference in melting temperature or by binding to a molecule that competes with a DNA stain. These methods have far denser labelling than restriction mapping and have been exploited to map viral genomes[53,59,60], resistance plasmids[61,62] and bacterial genomes[63]. However, the information content (i.e. the number of peaks and valleys) in these approaches is relatively low, which can make unambiguous assignment difficult. It is also difficult to use such approaches for molecular combing since the dyes are not covalently bound.

An alternative approach which covalently labels DNA and has an intermediate labelling density is to use enzymatic labelling. These approaches include labelling with nicking enzymes, which produce breaks in a single strand of the double-stranded DNA, at specific sites. DNA at these sites can be extended with fluorescently-labelled nucleotides for visualisation. This approach has been used to map viral genomes[65,67,173] and BACs, to aid genome assembly[46,47,66]. The main limitations with this approach are the labelling of non-specific and naturally occurring nicks and the fragmentation of DNA when two nicks are close together.

These limitations are avoided by using methyltransferase-directed labelling. DNA methyltransferases and restriction enzymes together form the restriction-modification system in bacteria, of which thousands of examples are now known[71]. Methyltransferases can be used to covalently transfer complex chemical groups, such as fluorophores and other modifications, to specific sites in a DNA sequence[86]. The labelling density is variable, depending on the target site, typically 2 to 8 base pairs in length,

which spans the labelling densities available for other approaches. This labelling has been used in both nanofluidic devices and with molecular combing, to identify viral genomes[69,83,93]. No commercial kit is currently available for methyltransferase-directed labelling of DNA. However, this approach was used and developed in CHAPTER 2 and can therefore be used in this research.

3.1.3 Procedures for matching DNA

Alongside experimental techniques for optical mapping, automated procedures for aligning and identifying DNA molecules have been developed. The procedures that have been reported vary depending on the labelling density and length of DNA fragments. An excellent review of computational methods used for low density labelling methods (e.g. restriction mapping, nicking enzymes) is given by Mendelowitz and Pop[174].

The type of data given by low density labelling methods is an ordered set of fragment lengths, estimated by imaging DNA fragments and automated processing (Figure 3.2). This data commonly contains a number of errors, such as sizing errors and missing or extra restriction sites or fragments. Also, this data will only span individual DNA molecules, so if they originate from a larger genome they must be assembled and combined to construct a genomic map. The alignment procedure used must consider these characteristics.

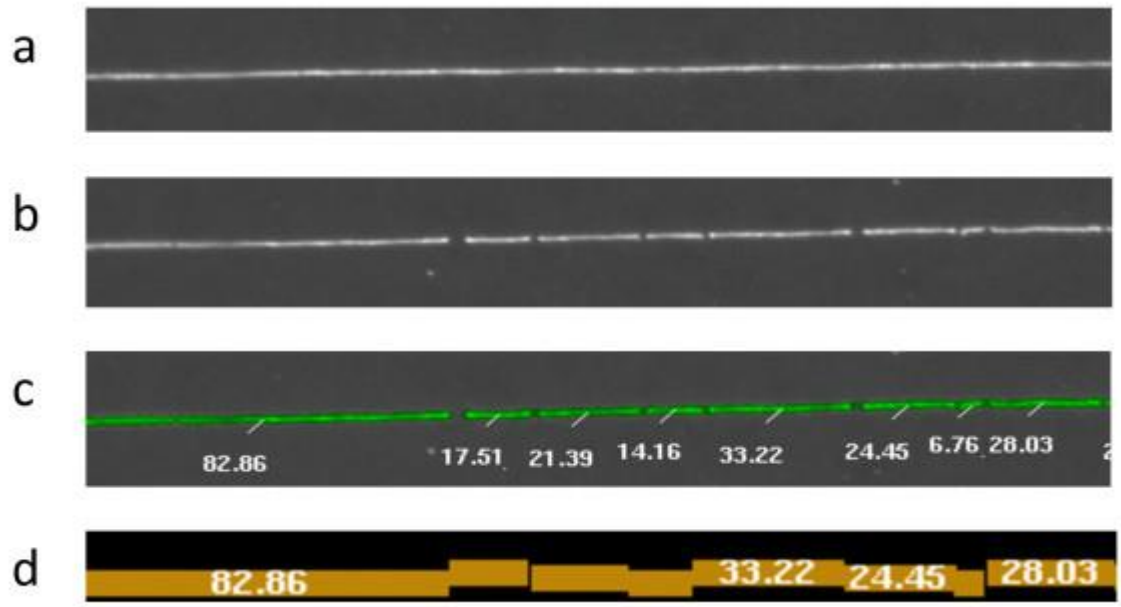


Figure 3.2 **Data generated by optical restriction mapping. Taken from Mendelowitz and Pop[174]. a) DNA samples are deposited onto coated glass surfaces and stained for visualisation. b) Restriction enzymes are used to digest the DNA at specific sites. c) Fragments are automatically extracted, and sizes estimated. d) An ordered set of fragment lengths is generated for each molecule.**

The exact alignment procedure will also depend on the problem. If there is already a consensus map (e.g. for known, sequenced genomes) then individual maps can be aligned to the consensus map. This can be used for identification of pathogens or investigations of large scale structural variations, for instance to aid assembly of genome sequences. Alternatively, if the sequence is unknown then individual maps must be aligned to each other, to produce a *de novo* consensus map. If a consensus map is generated this can be used to identify genomes more rapidly, since only the consensus map must be identified, rather than many inaccurate individual maps.

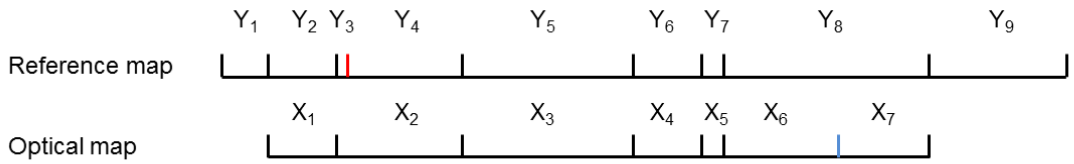
Dynamic programming algorithms are usually used for alignment since they can accommodate missing and false restriction sites and missing fragments. The alignment scoring functions used can also allow for small sizing errors. Dynamic programming

algorithms are usually based on the Smith-Waterman algorithm[175,176]. In this algorithm two strings are compared, in this case a list of segments (e.g. distance between restriction sites), by matching each segment from the optical map to each segment of the reference.

An example is shown in Figure 3.3. The first segment of the optical map (X_i where i=1) is compared to all segments of the reference map (Y_j for j=1 to 9) and a score is given depending on the quality of the match (e.g. the similarity in length) and stored in a matrix. Next the second fragment (X_2) is compared to all segments of the reference map (Y_j for j=1 to 9) and scored based on the quality of its match, but also the quality of the match between X_{i-1} and Y_{j-1}, i.e. the previous segments. This is continued for all segments and the alignment of the DNA fragment becomes clear along the diagonal of the scoring matrix. The score of the last segment of the optical map is the final alignment score of the map and the highest score can be used to find the diagonal path through the matrix. The Smith-Waterman algorithm has been modified to take in to account false restriction or missing restriction by the possibility of combining two or more segments at each[175,176].

Approaches for higher density labelling methods (e.g. affinity-based, methyltransferases) can use Dynamic programming algorithms if the position of labels can be localised (Figure 3.4B). For instance, by photobleaching the fluorophores the position of individual emitters can be estimated, based on the stochastic nature of this bleaching. However this type of localisation usually requires DNA fragments to be fixed to a surface, for instance via molecular combing, since time-lapses are required for the localisation procedure[93]. However, if labelling is more sparse and conditions are

carefully controlled then localisation can be achieved in nanofluidic devices[96]. The main drawback with localisation is the relatively long acquisition time and the processing required for localisation, which both reduce throughput.



Scoring matrix

Reference map segments

Optical map segments	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9
X_1	0	1	0	0	0	1	0	0	0
X_2	0	0	2		0	0	1	0	0
X_3	0	0	0	0	3	0	0	1	0
X_4	0	1	0	0	0	4	0	0	1
X_5	0	0	1	0	0	0	5	0	0
X_6	0	0	0	2	0	0	0	6	0
X_7	0	0	0	0	2	0	0		0

Score = 1 If segment lengths match well

Figure 3.3 **The Smith-Waterman algorithm for alignment of restriction maps. Every segment of an optical map (X_1-X_7) is compared to every segment of a reference map (Y_1-Y_9). The first segment (X_1) is compared to all segments of the reference map (Y_1-Y_9) and a score is given depending on the quality of the match (e.g. the similarity in length) and stored in a matrix. Next the second fragment (X_2) is compared to all segments of the reference map (Y_1-Y_9) and scored based on the quality of its match, but also the quality of the match between X_{i-1} and Y_{j-1}, i.e. the previous segments. Two or more segments can be combined for comparisons, for example if a restriction site is missing (red) or a false restriction site is present (blue). The score of the last segment of the optical map is the final alignment score of the map and the highest score can be used to find the diagonal path through the matrix (green).**

To increase throughput, alignment procedures that do not require localisation can be used, and these typically use intensity profiles automatically extracted along DNA molecules[51,83] (Figure 3.4A). These are then matched by assessing the overlap between a reference signal and the intensity profile, typically using cross-correlation. The reliability of alignment by cross correlation will depend on the characteristics of the profiles. For instance bright regions in a signal are a known problem for cross correlation, as a high score will be recorded when two bright signals are aligned, despite other parts of the profile being misaligned[177].

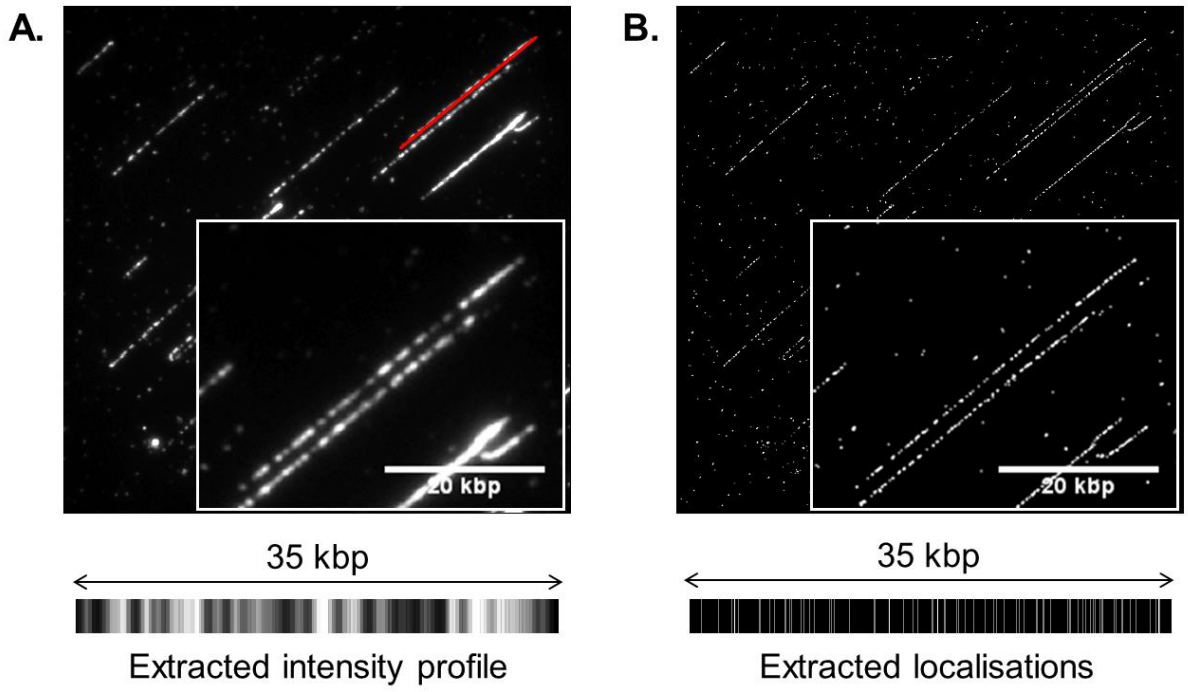


Figure 3.4 **Data for alignment of densely-labelled DNA fragments. Shown here is an image of M.TaqI-directed, Atto647N-labelled T7 phage DNA, combed onto a modified surface (image from Robert Neely). A) The intensity profile can be extracted along individual DNA molecules. For example, the intensity of the zoomed molecule (highlighted in red) is shown. This can be aligned by convolution or cross-correlation with a reference intensity profile. B) Alternatively, individual labels can be localised, for instance by photobleaching. The localisation of fluorophores along the same molecule is shown. Localisations can be used in conjunction with Smith-Waterman algorithms as shown in Figure 3.3.**

3.1.4 Overview

In this thesis, optical mapping using methyltransferase-directed labelling and molecular combing will be used to identify microorganisms. This combination is well suited to rapid identification of pathogens, since it is high-throughput and can be used for intermediate labelling density, which is suitable for reliable identification of microorganisms. Intensity profiles will be used rather than localisations for alignment, as throughput will be higher.

In CHAPTER 3 the methyltransferase-directed fluorescent labelling (based on optimisation carried out in CHAPTER 2), and combing of DNA fragments will be discussed. This will be followed by detailed *in silico* generation and alignment of intensity profiles to test the effect of various experimental parameters and for comparison with other labelling and intensity extraction techniques.

3.2 Results and discussion

3.2.1 Methyltransferase-directed labelling and deposition of genomic DNA

A two-step labelling scheme is used for methyltransferase-directed labelling of DNA: first the DNA is transalkylated using M.TaqI and an AdoMet analogue (AdoHcy-azide); then the azide functionalised groups are coupled to a fluorophore using strain-promoted azide-alkyne cycloaddition (SPAAC). This method was optimised for pUC19 in CHAPTER 2.

A restriction assay can be used to show the concentration of M.TaqI required for complete labelling. Figure 3.5 shows a restriction assay for T7 bacteriophage genomic DNA (henceforth referred to as 'T7'). pUC19 has one TaqI site every 672 bp, whilst T7 has one TaqI site every 360 bp, therefore a higher concentration of M.TaqI is required to fully protect T7. This is also seen for bacterial genomic DNA which has on average one TaqI site every 325 bp (Supplementary Figure 7.5). A completely random genome will have one TaqI site every 4^4=256 bp.

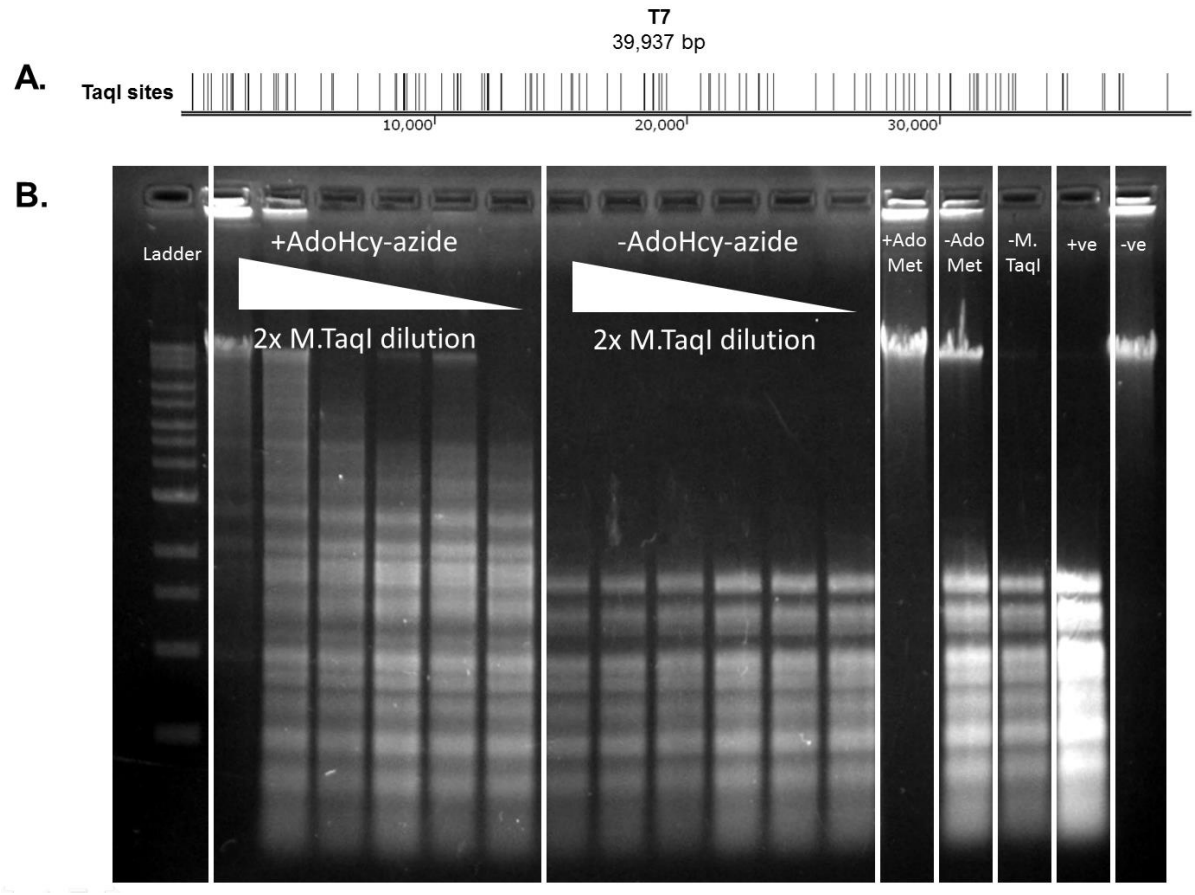


Figure 3.5 **Restriction assay for T7 bacteriophage DNA. A) Map of T7 DNA. It is approximately 40kbp in length and contains 111 TaqI sites (1 every 360bp). B) Restriction assay for M.TaqI labelling of T7 DNA with AdoHcy-azide and without added cofactor. Lane 1, 2 log ladder; lanes 2-7, AdoHcy-azide, 2x dilution of M.TaqI; lanes 8-13, no cofactor, 2x dilution of M.TaqI; lane 14, AdoMet control; lane 15, no cofactor control; lane 16, no M.TaqI control; lane 17, restricted pUC19; lane 18, unrestricted pUC19.**

DNA molecular combing was carried out as described by Deen et al[57], with the exception that 100 mM sodium phosphate pH 5.7/5% DMSO was used as the buffer. The combing buffer was modified to include sodium cations, which have been shown to help combing efficiency[56], whilst DMSO was added to improve dye solubility and prevent DNA secondary structure formation.

Ideal deposition of DNA would leave uniformly-stretched, well-spaced, straight, individual molecules. A good example of experimental deposition is shown in Figure 3.6.

Molecules are uniformly-stretched to around 1.52 times the crystallographic length (Supplementary Figure 7.6) and only have a slight curve, which is only problematic when extracting intensity profiles of larger molecules. However, deposition of single DNA molecules at reasonable densities is not straightforward. This has not been tested systematically here, although the factors that affect deposition have been described in great detail elsewhere[54–56].

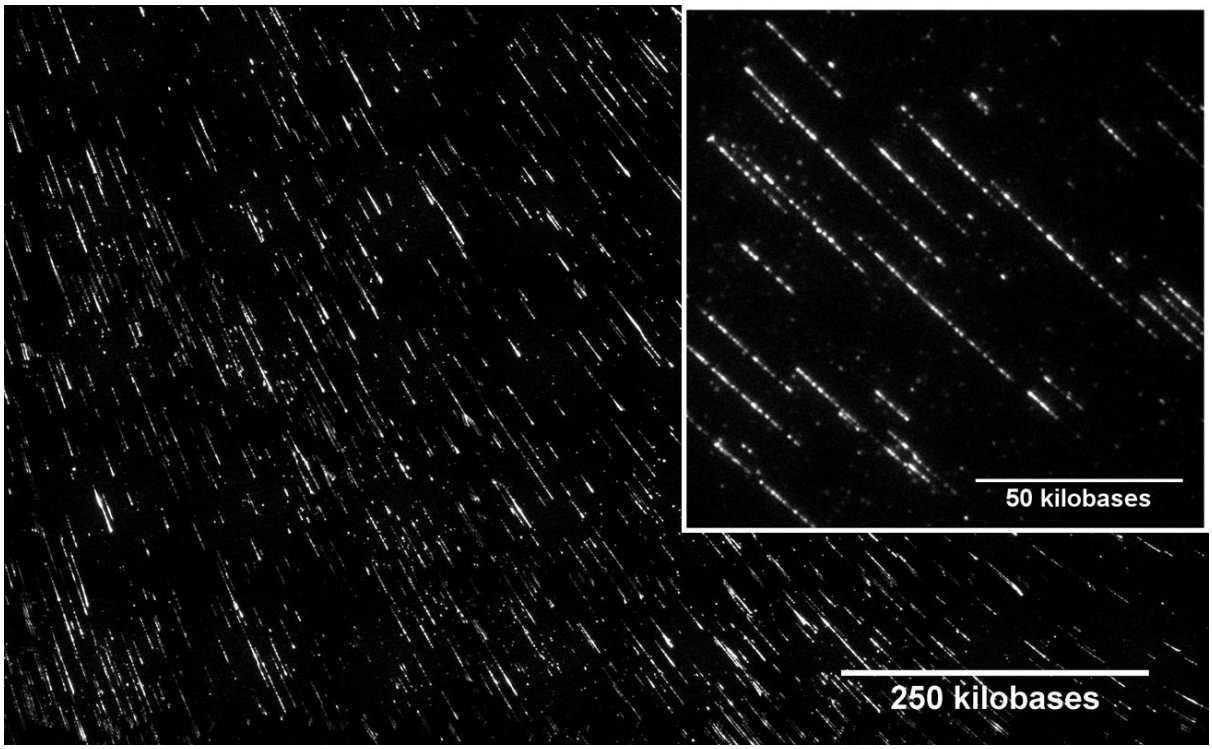


Figure 3.6 **Typical DNA deposition. 100pg/µl M.TaqI-directed Atto647N-labelled T7 DNA is combed at 20mm/min, in 100mM sodium phosphate pH 5.7/5% DMSO, onto zeonex-covered glass cover slips. A stitched image of several frames is shown along with a zoom that shows individual DNA molecules. In ideal combing experiments individual DNA molecules are well spaced, uniformly stretched, and straight.**

Most crucial is the concentration and deposition speed (the speed at which the meniscus is moved across the surface). High concentrations or slow deposition speeds can cause DNA fibre formation (Supplementary Figure 7.7 and Figure 3.7), which means individual molecules are difficult to extract. It can be impossible to distinguish overlapping DNA

molecules, which will prevent proper extraction of intensity profiles and affect subsequent alignment (Figure 3.7A). On the other hand, low deposition density will lower the throughput of the experiment (Figure 3.7C), so a balance must be struck (Figure 3.7B).

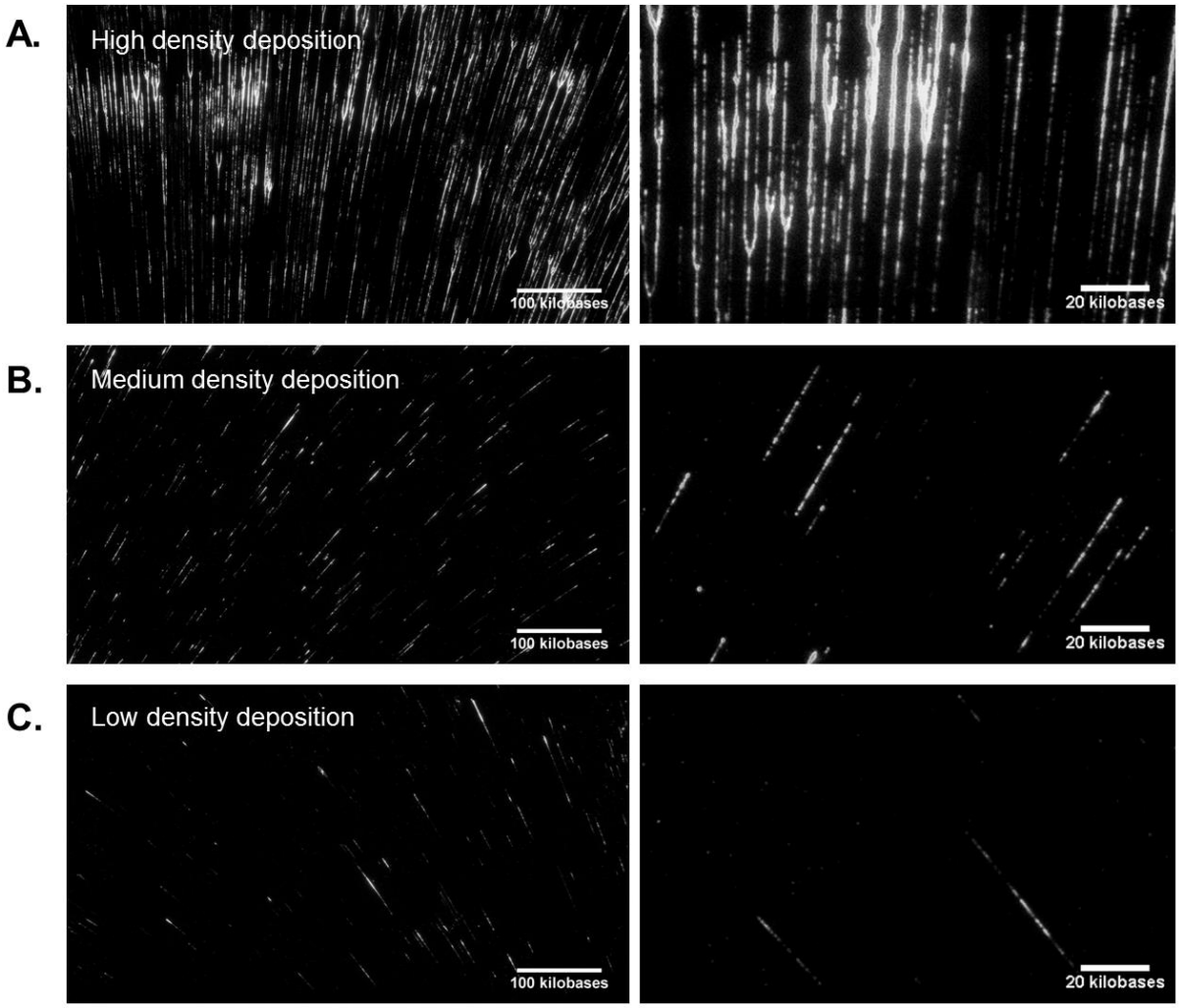


Figure 3.7 **Effect of concentration on DNA deposition. M.TaqI-directed Atto647N-labelled T7 or lambda DNA is combed at 20mm/min, in 20mM MES pH5.7, 2.7M NaCl, onto zeonex-covered glass cover slips. A stitched image of several frames is shown along with a zoom that shows individual DNA molecules. A) High density of deposition, ~1ng/µl labelled lambda DNA. DNA fibres form at high concentrations and overlapping DNA molecules make extracting individual intensity profiles impossible. B) Medium density of deposition, ~100pg/µl labelled T7 DNA. C) Low density of deposition. Unsuitable for high throughput extraction of intensity profiles.**

The amount of salt is important (Supplementary Figure 7.8) as is the pH (Supplementary Figure 7.9). Further optimisation of the combing conditions is important to improve experimental data for alignment and this is perhaps the limiting step in this experiment. However, in general, combing at relatively low concentrations ~50-200 pg/µl and relatively fast speeds (20 mm/min) gave sufficiently good deposition for further analysis (Figure 3.6).

Bleaching movies of deposited DNA molecules can be used to localise individual fluorophores, in a manner analogous to single molecule counting studies of pUC19 in CHAPTER 2. To do this with high throughput, molecules must be automatically extracted from the movies. For pUC19 this was achieved using the YOYO-1 intercalator. Since plasmids appeared as individual spots it was straightforward to use the Localizer plugin[94] to localise the plasmid positions. However, in contrast to small plasmids, long DNA molecules that have been stretched appear as long, linear fragments.

To allow automated single molecule counting for comparison with pUC19 the following procedure was developed and applied using custom Matlab code, for extraction of DNA from combing experiments. The procedure is shown in Figure 3.8, but in short, the steps are:

1. Estimate direction of combing (theta)
    - Using Hough transform
2. Smooth in that direction
    - Using convolution with Gabor filter in direction of theta
3. Detect edges of DNA molecules
    - Using Sobel edge detection

4. Use edges to define intensity profiles
    - Image dilation in direction of theta to fill gaps
    - Group edges by connectivity (i.e. define each edge)
    - Extract ends of edges
    - Merge close ends to define lines along DNA molecules
    - Extract intensity along line
    - Also extract length and convert to estimate of number of base pairs
5. Merge data for multiple images

This procedure takes advantage of the directionality in individual images (Figure 3.8A) to extract lines. The Hough transform[178] is a feature extraction technique that can be used to detect lines or shapes in images. Peaks in the Hough transform (Figure 3.8B) will correspond to lines in the image and here the median angle from the top ten features (theta) is used to give an estimate of the directionality of DNA molecules.

The intensity profile is not uniform, which makes normal feature extraction (e.g. by thresholding) difficult. Therefore, the median angle (theta) is used to blur DNA molecules in that direction to make extraction of intensity profiles more straightforward. This also means an intercalator which uniformly stains along the length of a DNA molecule, e.g. YOYO-1, is not necessary for the extraction of intensity profiles. A Gabor filter with an angle set by the directionality (theta) is used (Figure 3.8C) and convolved with the original image (Figure 3.8D).

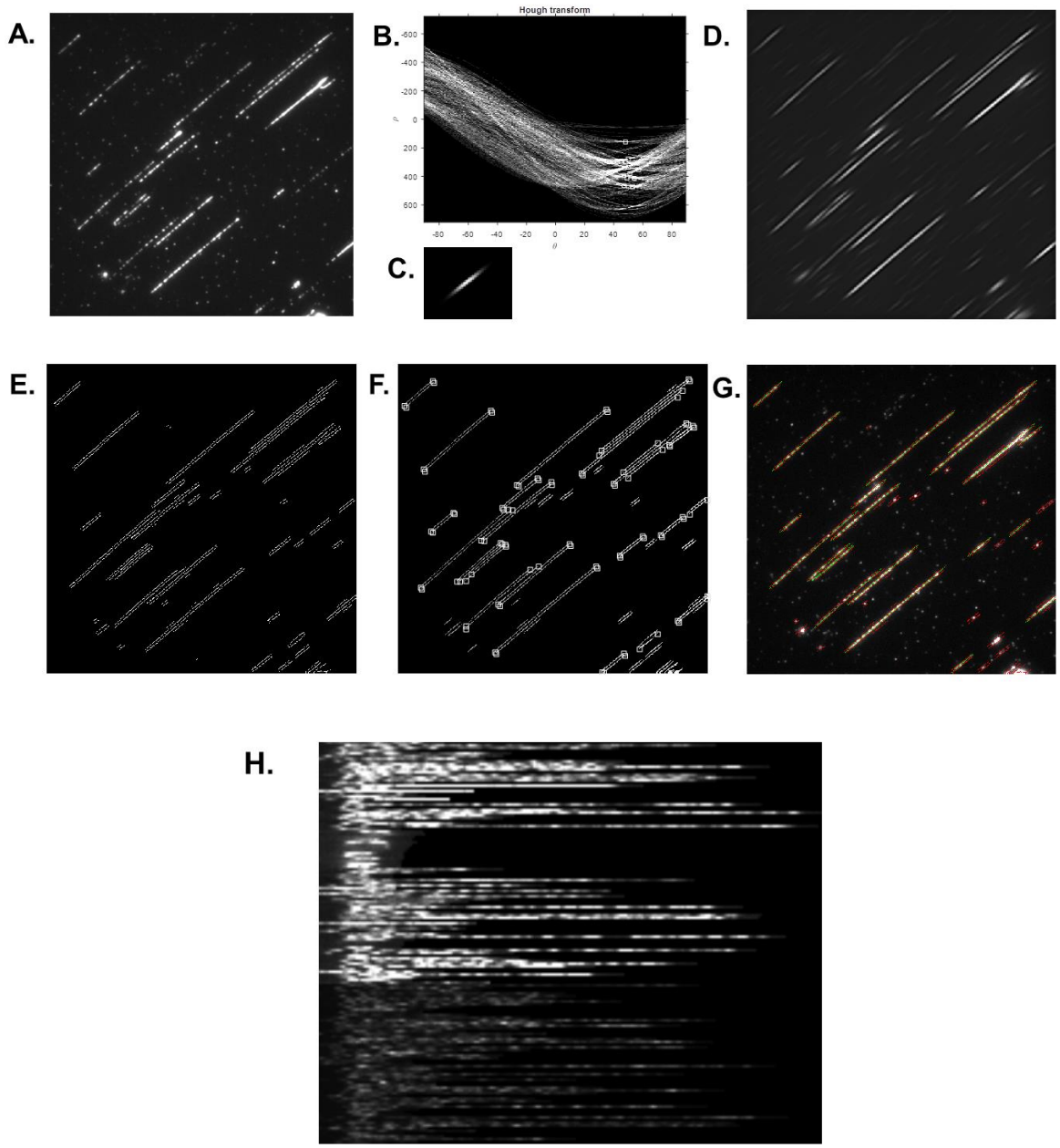


Figure 3.8 **Automated extraction of intensity profiles. A) Typical combing image. B) Hough transform used to estimate direction of combing. 10 peaks are selected, and the median used to define theta. C) A Gabor filter in the direction of theta is created and a convolution with A) gives D). E) Edge detection on D) using the Sobel method and an automatically detected threshold. F) Edges are dilated, and connectivity used to define the ends of lines. When two pairs of end points are close a line is drawn between them. G) This line is shown in green, edges in red. H) This is done for many images and extracted intensity profiles are merged.**

From this, the edges of DNA molecules are detected using the Sobel method and an automatically detected threshold, by a built-in Matlab function (Figure 3.8E). Edges are dilated to fill gaps and grouped by connectivity. If edges are short (i.e. <50 pixels) they are removed, and the ends of each edge are detected. When the ends of a pair of edges are close (within 15 pixels) they are paired (Figure 3.8F). The average position of a pair of ends is used to define the two ends of an individual DNA molecule and used to define a line along the DNA molecule (Figure 3.8G), from which the intensity profile and length of the molecule is extracted. This is repeated for many images to automatically extract the intensity profile along many DNA molecules (Figure 3.8H).

The coordinates of extracted DNA molecules can be used to assign individual fluorophores to a given molecule. The assignment of these individual molecules can then be used for single molecule counting experiments (Figure 3.9). Bleaching movies are taken and individual fluorophore positions estimated using the Localizer plugin[94] in IgorPro. Localisation coordinates are overlaid with DNA molecules extracted from the image (Figure 3.9B). Any localisations within close proximity (e.g. 5 pixels) of the line along the DNA molecule are assigned to it (Figure 3.9C). A ratio of the number of localisations and the length of the DNA fragment is used to calculate a labelling efficiency and is presented as a histogram. For example, for T7 DNA a label is expected every 180 bp on average, so 100 fluorophores along 18 kbp would give 100% labelling efficiency.

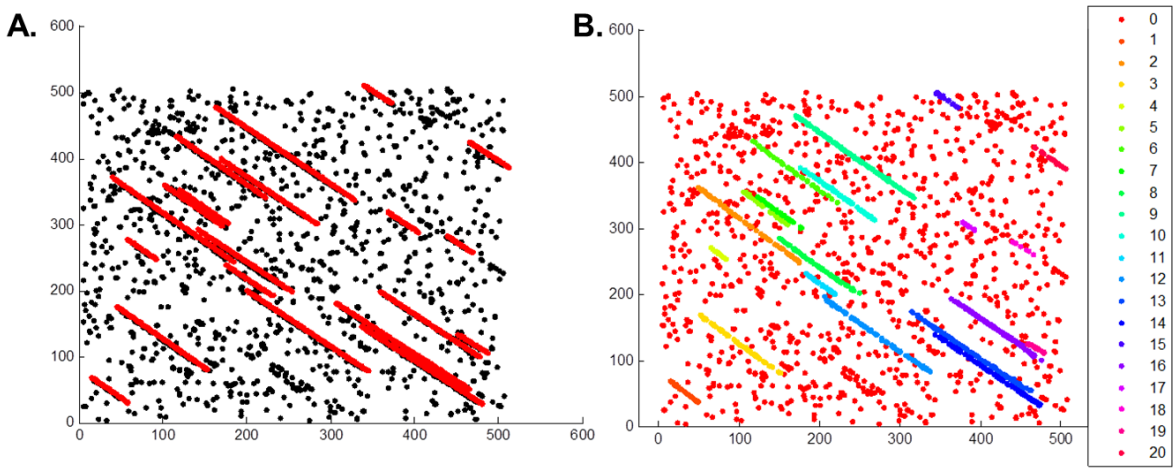


Figure 3.9 **Single molecule counting for combed DNA. The bleaching movie for the typical combing image of Figure 3.8 is used for localisation of individual fluorophores. A) Localisations (black) are overlaid with lines extracted by the process shown in Figure 3.8 (red). Note that coordinates have been flipped. B) Localisations are assigned to DNA fragments, each denoted by a different colour.**

Figure 3.10 shows that labelling efficiency for T7 is consistent with labelling of pUC19, though slightly reduced overall. The extent of non-specific labelling was not tested but is expected to be similar to the results for pUC19. When individual intensity profiles are examined this appears to be the case (Figure 3.11). Careful combing at low concentrations is required for reliable results or efficiency is overestimated due to overlapping DNA molecules (Supplementary Figure 7.10).

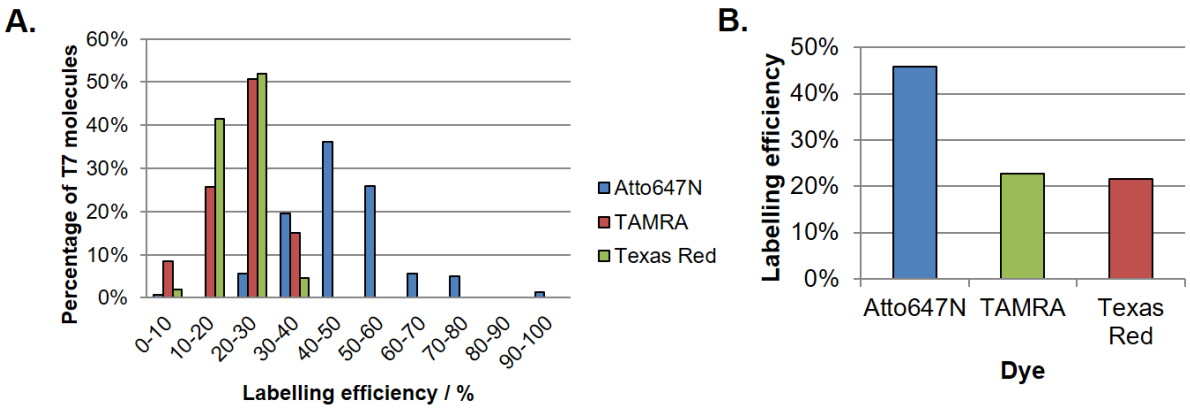


Figure 3.10 **Labelling efficiency of T7 DNA for different commercial dyes. T7 was labelled with M.TaqI and AdoHcy-azide and coupled post-transalkylation. A) Single molecule counting results for each dye. B) Labelling efficiencies for each dye.**

The labelling/combing procedure that has been used throughout the rest of this chapter is as follows:

1. Run protection assay with AdoHcy-azide to confirm minimum amount of required M.TaqI
2. Label 2 µg of DNA with AdoHcy-azide and M.TaqI. Mix the following and incubate at 50°C for 1 hour.

All in µL	
Water	30.0
10x CutSmart	4.0
0.5mg/ml T7	4.0
15mM AdoHcy-azide	2.0
0.3mg/ml M.TaqI	2.0

3. Add 2 µl proteinase K (18 mg/ml)/0.1% Triton X-100 and incubate at 50°C for 1 hour.
4. Purify by ethanol precipitation into 50 µl 1xPBS

5. Mix 2.5 µl 1xPBS, 1.5 µl 50 mM Atto647N NHS Ester and 1.0 µl and 20 mM DBCO-amine. Incubate at 4^{o}C for 1 hour.
6. Add DBCO-Atto647N mixture and 5 µl DMSO to purified DNA. Incubate at room temperature overnight.
7. Purify on silica columns into 200 µl elution buffer. Dilute in elution buffer to around 2 ng/µl.
8. Mix 2 µl labelled DNA, 17 µl 100 mM sodium phosphate buffer (pH 5.7) and 1 µl DMSO. Comb 1.5 µl at 20 mm/min on zeonex-coated glass coverslips.
9. Image using TIRF/widefield, 100/150X, 640 nm excitation.

In Figure 3.11A an individual T7 molecule is shown that has been labelled and combed by this procedure. The intensity profile is automatically extracted and is shown in Figure 3.11B. It has been aligned to a reference intensity profile and shows good visual agreement with the expected profile. The alignment procedure will be described in detail in the next part of the chapter (3.2.3).

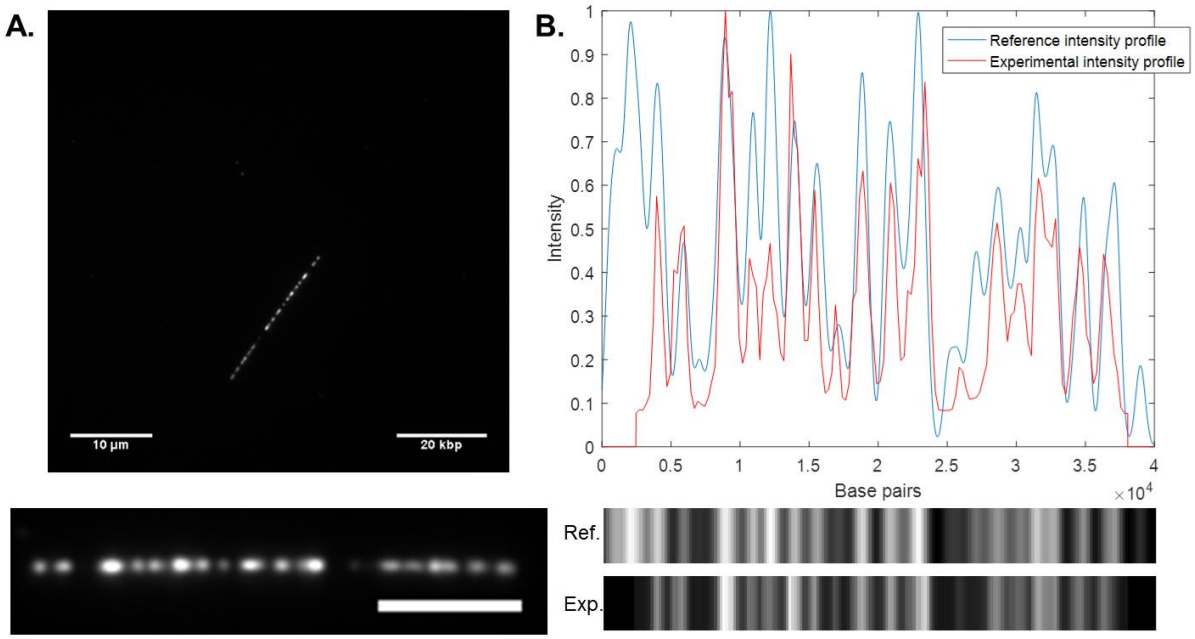


Figure 3.11 Individual labelled molecule of T7 DNA. M.TaqI-directed labelling with AdoHcy-azide is followed by post-transalkylation coupling with Atto 647N. DNA is then combed onto a coated glass coverslip and imaged by excitation at 640nm. A) A typical individual molecule, with a zoom showing the experimental intensity profile (scale bar = 10µm). B) The intensity profile is automatically extracted (red and bottom) and can be aligned to the reference intensity profile (red).

3.2.2 Generation of DNA barcodes *in silico*

The alignment of intensity profiles (henceforth referred to as DNA barcodes) can be tested fully *in silico* to inform the experimental and analytical techniques. The importance of experimental factors e.g. labelling efficiency, can be determined and alignment strategies developed and tested more easily. In particular, the alignment of a 'ground-truth' can inform the development of alignment strategies for more complex experimental samples.

For modelling *in silico,* a series of realistic DNA barcodes are generated. The imperfections of the barcodes that can be obtained experimentally are simulated by considering various experimental parameters: labelling efficiency; non-specific labelling;

variation in fluorescence intensity; fragment length and position in genome; non-uniform stretch; camera pixel size; orientation of fragments; noise; and the effective PSF. These parameters are summarised in Table 3.1, along with typical estimated experimental values.

Variables	Description	Typical values
sequence	Genome sequence	n/a
data_no	No of fragments to be generated	100-10000
meth_efficiency	Labelling efficiency	30-60%
false_methylation	Chance of non-specific labelling (per base pair)	0-0.001
flu_intensity_var	Variation in fluorophore intensity	10-30%
min_fragment_length	Minimum length of fragment (in base pairs)	30,000
max_fragment_length	Maximum length of fragment (in base pairs)	50,000
base_length_var	Variation in stretching	0-10%
sample_freq_mean	Average pixel size (base pairs per pixel)	300-400
sample_freq_distr	Variation in pixel size (will depend on direction)	10-20%
pixel_distr	Variation in pixel sampling (in base pairs)	20-60
noise_mag	Magnitude of noise	10-20%
PSF_ref	[sigma,size] for reference PSF (in base pairs)	n/a
PSF_frag [sigma]	sigma for experimental PSF (in base pairs)	250-500
PSF_frag [size]	size for experimental PSF (in base pairs)	n/a
PSF_frag_var	Variation in experimental PSF	5-20%

Table 3.1 Typical experimental parameters for *in silico* generation of experimental barcodes (procedure shown in Figure 3.13).

The procedure for generating barcodes *in silico* is shown in Figure 3.13 and is as follows:

1. Import labelled genome
    - Known sequence is imported from FASTA and converted to matrix form (A=1, C=2, G=3, T=4)

- Convert to labelled sequence. Labelled bases=2, unlabelled bases=0. E.g. for M.TaqI the A of TCG<u>A</u>=2, all other bases=0.

2. Choose random fragment from genome
    - Random first fragmentation site
    - Second fragmentation site up to max fragment length away e.g. 50,000 bp
    - Only fragments larger than minimum fragment length e.g. 30,000 bp and within ends of genome are carried forward
    - Store size of fragment. Known size of fragment (in bases) altered by random variation in stretch e.g. ±10%
3. Reduce labelling efficiency.
    - Each site gets two random numbers between 0 and 1. If number is greater than desired frequency e.g. 0.45 then label is removed.
4. Add in non-specific labelling.
    - Each site gets a random number between 0 and 1 and if number is less than desired frequency e.g. 0.001 then methylation site is added
5. Allow fluorescence intensity to vary e.g. ±20%
6. Convolve labelled fragment with Gaussian filter
    - Define sigma and length of PSF e.g. 300 and 3000 bp
    - Allow for variation in sigma e.g. ±20%
7. Sample intensity along fragment
    - Sample every e.g. 250 fragments, based on pixel size, with allowed variation e.g. 200 to 300 bp per pixel (since this will depend on orientation of molecule)
    - Don't sample perfectly e.g. sample at 0, 245, 502, 746, etc.

8. Randomly flip fragment (i.e. 50% likelihood fragment is backwards)
9. Add noise. E.g. Intensity varies by e.g. ±15%
10. Repeat for desired number of fragments, merge data and display

Examples of a generated barcode and an experimental barcode are shown in Figure 3.12 and show differences between them that must be considered. In experimental barcodes there will be a background of fluorescence intensity, unless this is subtracted from the image. Also, in generated barcodes the intensity profile is abruptly terminated, whilst in experimental barcodes signal is extracted from slightly beyond the molecule. This will be taken into account during alignment of experimental barcodes.

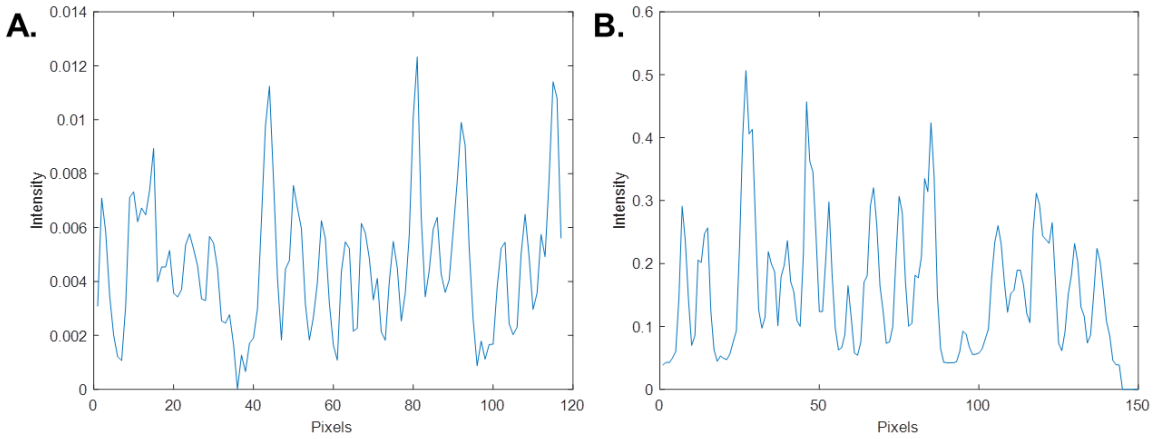


Figure 3.12 Examples of generated and experimental DNA barcodes. DNA is M.TaqI labelled. A) Barcode generated *in silico* using procedure and parameters shown in Figure 3.13. B) Experimental barcode from Figure 3.11.

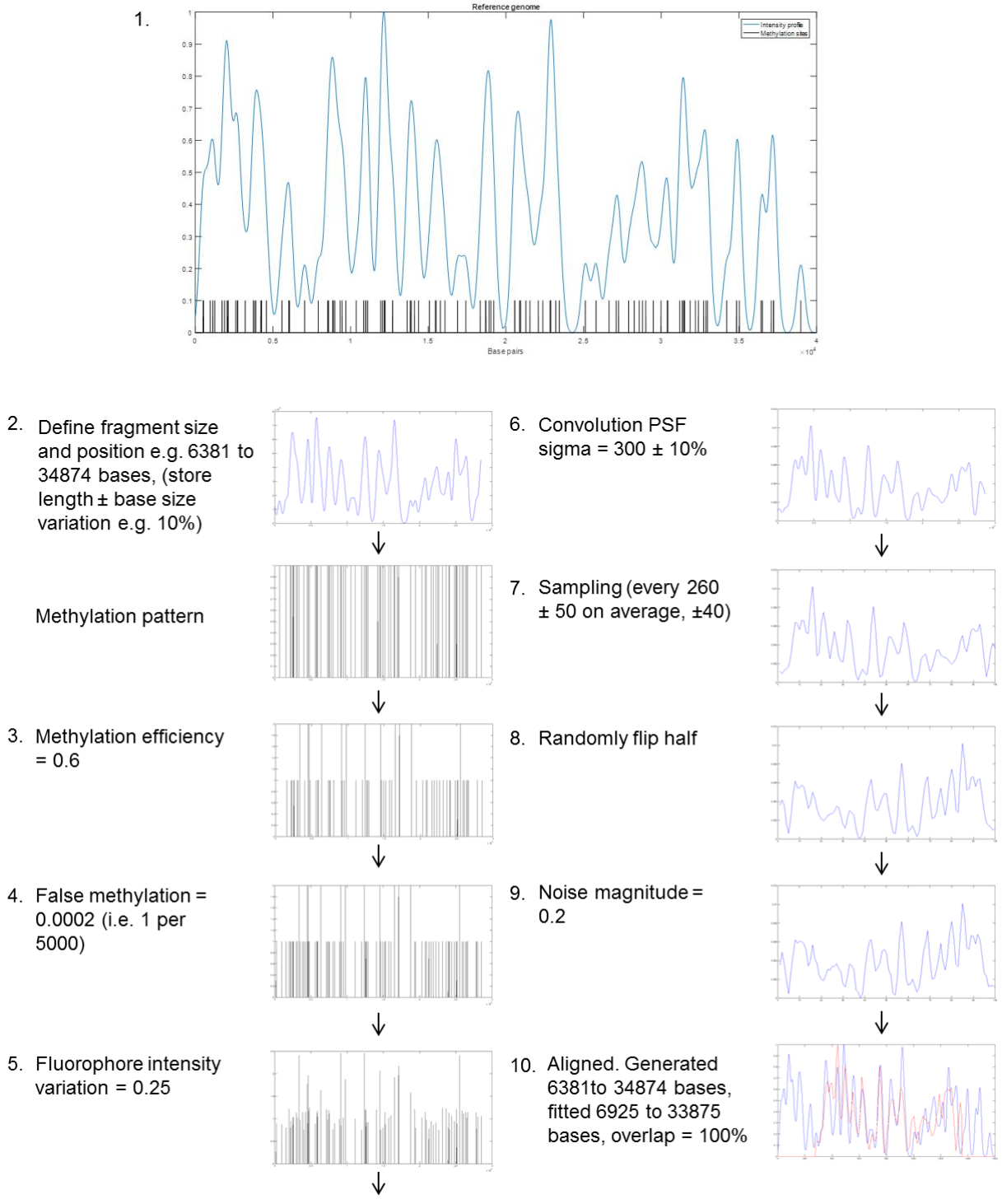


Figure 3.13 **Procedure for *in silico* generation of DNA barcodes. Full procedure is described in main text. 1) Import labelled genome, e.g. M.TaqI-labelled T7 genome. 2) Choose random fragment from genome. 3) Reduce labelling efficiency. 4) Add in non-specific labelling. 5) Allow fluorescence intensity to vary. 6) Convolve labelled fragment with Gaussian filter. 7) Sample intensity along fragment. 8) Randomly flip fragment. 9) Add noise. 10) Alignment of generated barcode to reference**

3.2.3 Alignment of DNA barcodes to a reference

Fragments can be readily aligned to a known reference by maximising the cross-correlation. Cross-correlation is a measure of similarity between two signals as a function of the displacement of one relative to the other. It is commonly used in signal processing, for example to calculate the delay between the same signal collected from different sensors[179]. This is a problem analogous to the alignment of DNA barcodes. The cross-correlation, xcorr, between two signals, x and y (e.g. Figure 3.14A), of length, N, at delay, d, is defined as:

$$xcorr(d) = \sum_i x(i)y(i-d)$$

Where d=-N+1,-N+2,…,N-2,N-1. An example is shown in Figure 3.14B, which shows that the peak in cross-correlation correlation corresponds to a displacement of zero. However, this peak is not significantly higher than values of the cross-correlation at other displacements, since it also depends on the number of values overlapping between signals.

To correct for this the cross-correlation can be normalised, so that for complete overlap between signals the cross-correlation is 1. An example of this is shown in Figure 3.14C. The mean is subtracted from each signal and each signal is divided by its standard deviation, before the final result is divided by the length of the signal:

$$xcorr(d) = \frac{1}{N}\frac{\sum_i (x(i)-\bar{x})(y(i-d)-\bar{y})}{\sqrt{\sum_i (x(i)-\bar{x})^2}\sqrt{\sum_i (y(i)-\bar{y})^2}}$$

Note how the peak cross-correlation is now largely unaffected by overlap and so can be more easily identified. Cross-correlation is equivalent to the convolution of $x^*(-i)$ and

$y(i)$, which allows for exploitation of fast Fourier transform algorithms for efficient computation. This is what the built-in Matlab functions xcorr and xcorr2 use:

$$xcorr(x, y) = x^*(-i) * y(i)$$

$$FT\big(xcorr(x, y)\big) = FT(x)^*.FT(y)$$

Where FT denotes the Fourier transform, and x^* the complex conjugate of x.

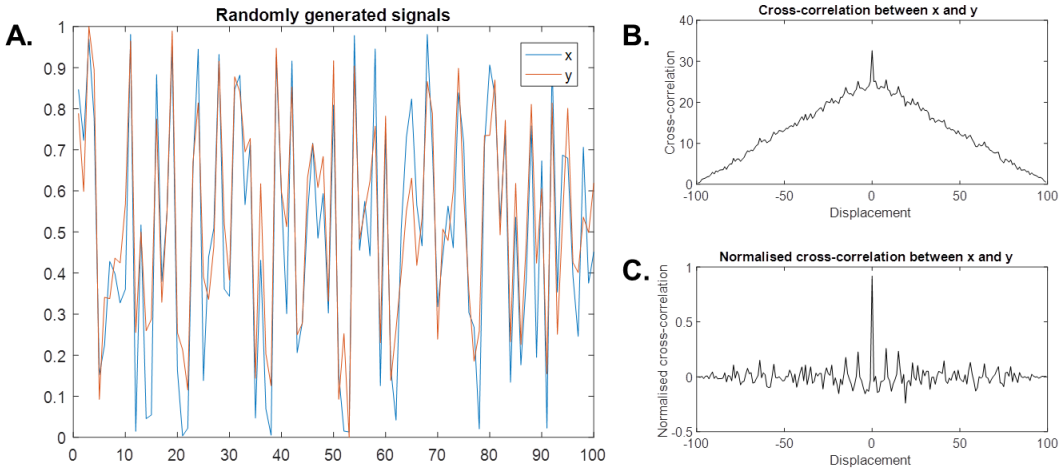


Figure 3.14 Example of cross-correlation. A) Generated signals, x (blue) and y (red). x is randomly generated signal between 0-1, and y is with noise of 20%. B) Cross-correlation between x and y. The peak in cross-correlation is at a displacement of zero, which is when the signals are correctly aligned. C) Normalised cross correlation between x and y. The dependency on overlap is largely removed, making the peak cross-correlation more easily identified.

For fitting fragments to larger genomes (e.g. *E. coli*) there will be a large difference in the length of signals. When signals are of different lengths zeros will be appended to the shorter signal to make the lengths equal (e.g. Figure 3.15A). However, if the global mean and standard deviation are used to normalise the reference signal then the alignment will be biased towards the most intense parts of the signal (Figure 3.15B). Since the absolute intensity of an individual barcode is not necessarily known this is undesirable,

the barcode should have an equal probability of fitting across the genome, independent of intensity. Normalisation using a rolling mean and standard deviation, from the same length as the shorter signal, removes most of the bias (Figure 3.15C).

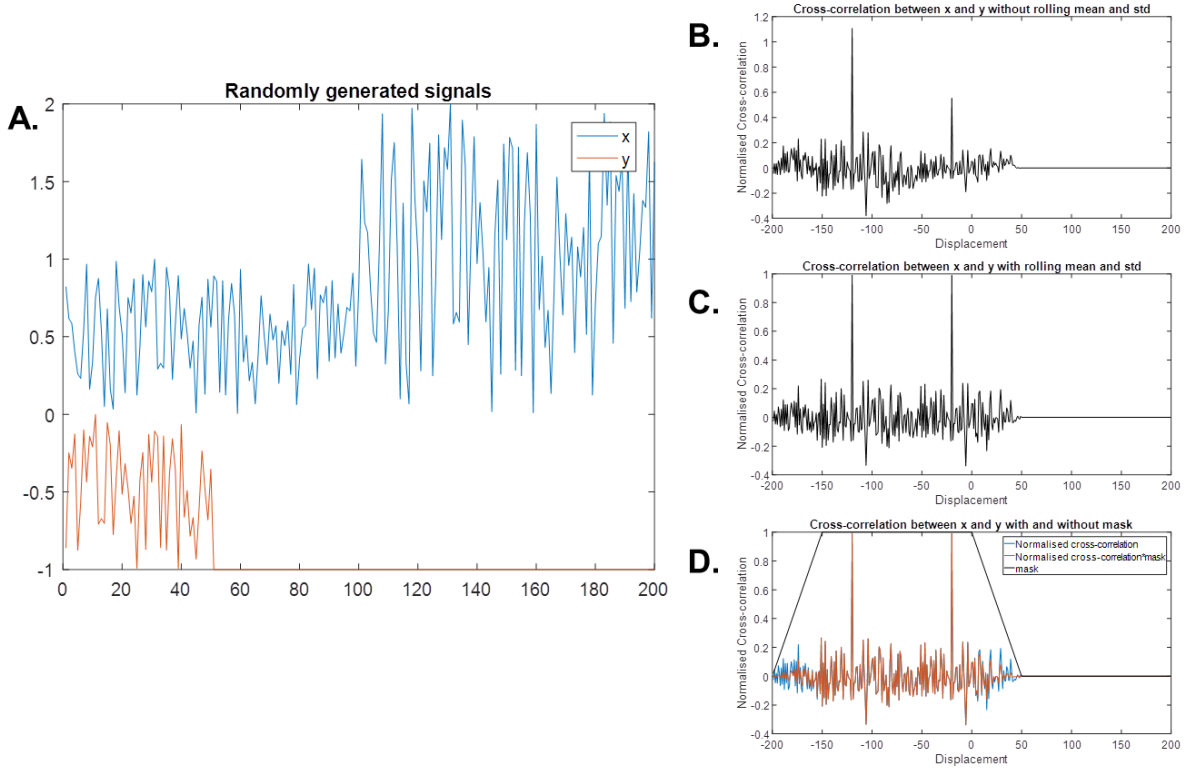


Figure 3.15 **Example of cross-correlation between two signals of different length, x and y. A) Generated signals, x (blue) and y (red). The second half of x is twice the intensity as the first half of x. y is a region taken from the first half of x. Zeros are appended to y to make it the same length as x for cross-correlation. B) Normalised cross-correlation between x and y (black). There is a clear bias towards the intense regions, which is undesirable. C) Normalised cross-correlation using a rolling mean and standard deviation of the length of y, between x and y (black). This removes most of the bias towards intense regions. D) Normalised cross correlation using rolling mean and standard deviation (blue) multiplied by percentage overlap (red) and the percentage overlap mask (black). For short barcodes when large amounts of overlap are expected this reduces spurious cross-correlation for partial overlap.**

Finally, to penalise cross-correlation when there is not complete overlap the normalised cross-correlation can be multiplied by the percentage overlap between signals, henceforth called the mask. This will be used for short genomes, e.g. T7/lambda, where

there should be significant overlap between DNA barcodes. Spurious cross-correlation for only very partial alignment is reduced (Figure 3.15D).

Alignment of experimental DNA barcodes to known reference barcodes uses the cross-correlation between them. Experimental barcodes are stretched to 90-110%, (in intervals of 1%), of the estimated length and the normalised cross-correlation calculated for each stretch and for both orientations (forward or reverse). The maximum value of the normalised cross-correlation is used to give the displacement, stretch and orientation of the experimental barcode and used for alignment to the reference barcode. The full procedure is outlined below:

1. Generate reference barcode
    - Reference sequence is imported from FASTA and converted to matrix form (A=1, C=2, G=3, T=4)
    - Convert to labelled sequence. Labelled bases=2, unlabelled bases=0. E.g. for M.TaqI the A of TCG<u>A</u>=2, all other bases=0.
    - Convolution with PSF, select sigma from 250-400bp
    - Sample every N base pairs, where N is 'sampling'
2. Stretch experimental barcode
    - Length of DNA fragment calculated based on pixel size.
    - Use crystallographic length of bp (0.34 nm) and estimated stretch (1.52) to estimated base pairs per nanometre = 1.93 bp/nm
    - Use estimated length in number of bases to estimate stretch required
    - Use interpolation to 'stretch' fragment to full length in base pairs
    - Sample every N base pairs, where N is 'sampling density'

- Include reverse of fragment

3. Cross correlation with reference barcode to define best stretch
    - Use normalised cross correlation and test for 90%-110% of estimated stretch
    - Maximise normalised cross correlation to define best stretch and orientation
4. Align fragment
    - Maximum normalised cross correlation gives corresponding displacement
    - Use displacement to align stretched and oriented fragment along reference genome
5. Repeat for each fragment

An example of the result of this procedure is shown in Figure 3.16A. 100 barcodes, randomly generated from the T7 genome (Figure 3.16B), can be aligned to the T7 reference barcode (Figure 3.16C) in less than 1 second.

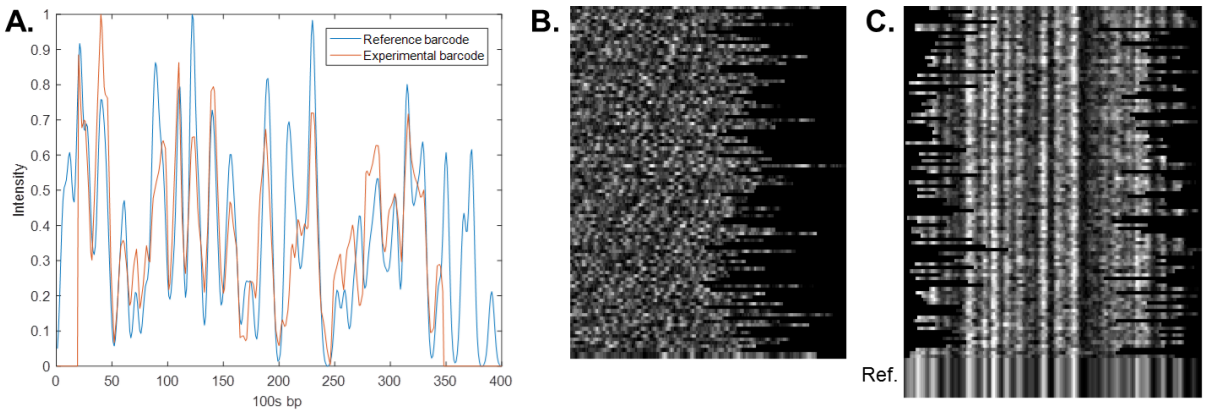


Figure 3.16 Example of alignment procedure. A) T7 reference barcode (blue) and generated barcode (red), generated *in silico* from the T7 genome. The generated barcode has been stretched, oriented and aligned by maximising the normalised cross-correlation. B) 100 generated barcodes, generated *in silico* from the T7 genome. C) Alignment of generated barcodes to the reference barcode (bottom).

3.2.4 Sensitivity analysis and experimental parameters

Using this alignment procedure in tandem with random generation of fragments *in silico* the sensitivity of alignment to each experimental parameter can be determined. This can inform the experimental protocol, for instance, what sort of labelling efficiency is required and what length of DNA fragment?

Here Monte-Carlo simulations are used to assess the sensitivity towards each experimental parameter. In these simulations the parameters for *in silico* generation of barcodes are randomly varied, between reasonable experimental values, shown in Table 3.2. 100 barcodes are generated for each of 5000 sets of parameters and aligned to the reference barcode. The position from which the barcode was generated is compared to the aligned position and where overlap between the generated and fitted position is greater than 98% the barcode is considered to be correctly aligned. This means for each set of parameters the number of correctly aligned barcodes is known.

Variable	Description	max	min
sequence	Genome sequence	n/a	n/a
data_no	No of fragments to be generated	100	100
meth_efficiency	Labelling efficiency	0.1	1
false_methylation	Chance of non-specific labelling (per base pair)	0	0.01
flu_intensity_var	Variation in fluorophore intensity	0	1
min_fragment_length	Minimum length of fragment (in base pairs)	30000	10000
max_fragment_length	Maximum length of fragment (in base pairs)	min+10000	min+10000
base_length_var	Variation in stretching	0	0.2
sample_freq_mean	Average pixel size (base pairs per pixel)	100	500
sample_freq_distr	Variation in pixel size (will depend on direction)	0	0.2
pixel_distr	Variation in pixel sampling (in base pairs)	0	0.2
noise_mag	Magnitude of noise	0	1
PSF_ref	[sigma,size] for reference PSF (in base pairs)	n/a	n/a
PSF_frag [sigma]	[sigma,size] for experimental PSF (in base pairs)	250	500
PSF_frag [sized]	[sigma,size] for experimental PSF (in base pairs)	n/a	n/a
PSF_frag_var	Variation in experimental PSF	0	0.5

Table 3.2 Experimental parameters for *in silico* generation of barcodes for Monte-Carlo simulations.

A 2D histogram can be produced for each parameter, plotting the number of correctly fitted fragments against the value of the parameter. An example is shown for the labelling efficiency parameter, in Figure 3.17A and B. Each point in the scatter plot in Figure 3.17A represents a single Monte-Carlo run for a specific set of parameters, these are then used for a histogram for visualisation in Figure 3.17B (histograms for each parameter are shown in Supplementary Figure 7.11). The correlation between the parameter and the number of correctly fitted fragments can then be used to assess sensitivity of the alignment towards each parameter. For this the Spearman rank correlation coefficient is used (Figure 3.17C). This can be used to assess the dependence between the rankings of the two variables, where a value of 1 is perfect positive correlation, a value of -1 is perfect negative correlation and a value of 0 means there is no correlation. The line of best fit is shown (Figure 3.17A and B) and approximates the correlation of the parameter and the number of correctly fitted barcodes.

The alignment of barcodes is most sensitive to the labelling efficiency, this is clear from both the Spearman rank correlation coefficient and the trend in the 2D histogram. Non-specific labelling, the length of fragments and the amount of noise are also important to consider, however the alignment is less sensitive to other variables, such as fluctuations in fluorophore intensity or changes in the PSF.

Conceptually these results make sense. The most important characteristic of a signal for cross-correlation is the position of peaks i.e. the position of labels, which is most affected by labelling efficiency and non-specific labelling. However, parameters that affect the intensity of the peaks are less important, for example variations in fluorophore intensity or noise. Also, the pixel size and PSF, within these ranges, do not significantly affect the

presence and location of peaks. Shorter signals are less likely to find a unique position along the reference barcode, but problems with non-uniform stretch are less important.

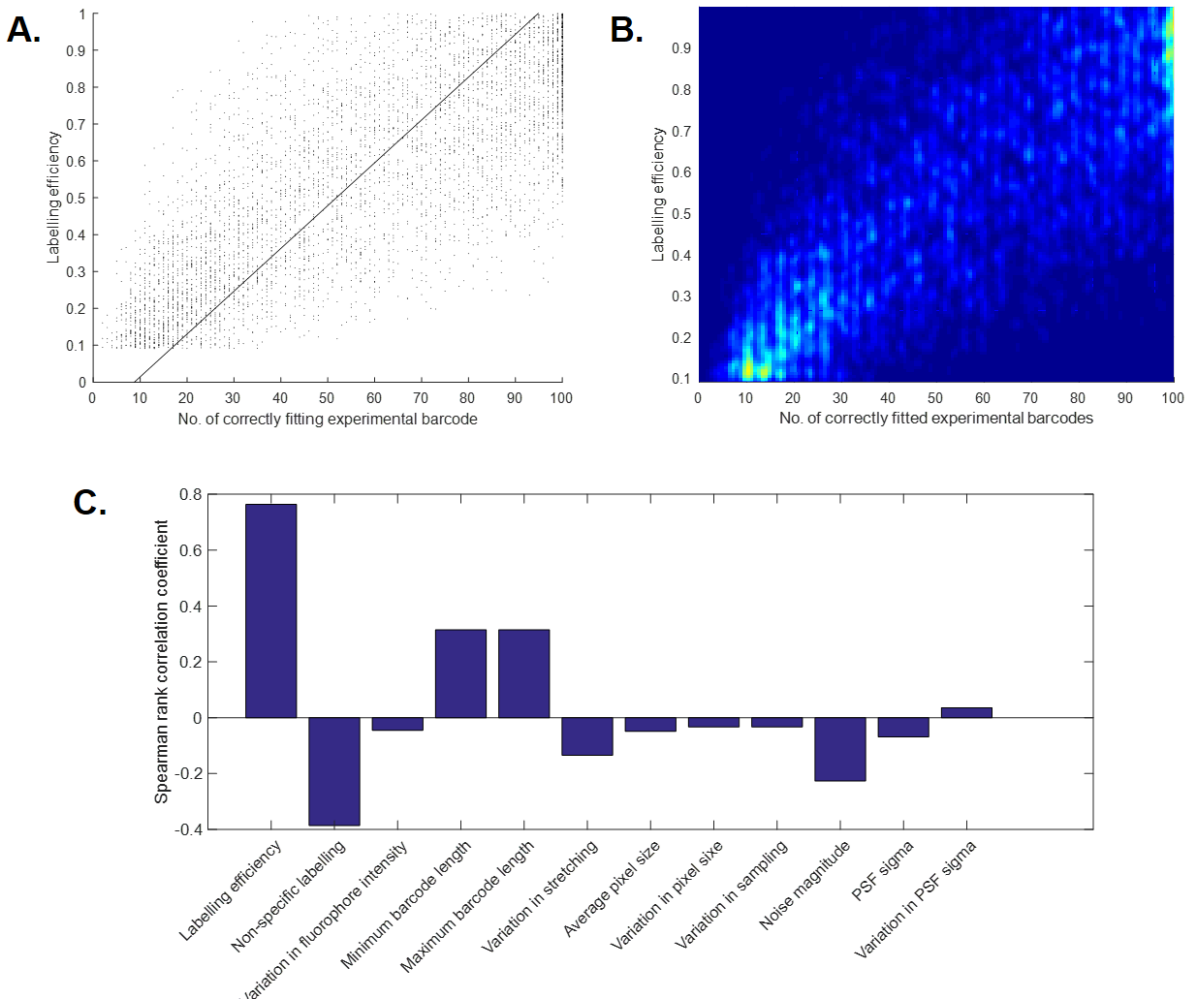


Figure 3.17 **Monte-Carlo simulation to test sensitivity of parameters. 5000 sets of parameters were run for 100 fragments each. Experimental barcodes were generated and aligned from/to the T7 genome. A) Example scatter plot for the labelling efficiency. Each point in the scatter plot in represents a single Monte-Carlo run for a specific set of parameters. B) These can be visualised as a 2D histogram. The line of best fit shows the sensitivity of each parameter on the number of correctly fitted barcodes. C) This can be quantified using the Spearman rank correlation coefficient for each parameter.**

Here the two parameters to which alignment was most sensitive, labelling efficiency and non-specific labelling, are investigated in more detail. This is especially useful as the parameters can be experimentally tested, for example labelling efficiency of T7 is shown in Figure 3.10. This was also optimised in CHAPTER 2, so knowledge of the required labelling efficiency can be used to assess the quality of methyltransferase-directed labelling.

Each parameter is varied systematically, and barcodes generated and aligned to the reference barcode. Figure 3.18A shows a 2D histogram for T7. Each pixel represents a different labelling efficiency and amount of non-specific labelling, whilst the colour of each pixel reports on the proportion of generated barcodes that were correctly fitted. The average dependence on labelling efficiency (for all values of non-specific labelling) is shown in Figure 3.18B and the average dependence on non-specific labelling (for all values of labelling efficiency) is shown in Figure 3.18C.

For T7 these results suggest that a labelling efficiency of around 50% should be sufficient for around 90% of fragments to be fitted, whilst non-specific labelling can approach around 1 label per 1000 bases, before there is a significant drop in correct alignment. Experimentally, M.TaqI-directed labelling of pUC19 with Atto647N achieved 50% labelling efficiency, but a non-specific label around every 1500 base pairs (results in CHAPTER 2). For TAMRA labelling was around 30% efficient, but with no significant non-specific labelling. The proportion of correctly fitted barcodes for these regimes is 93% and 97% respectively. Similar results should be expected for genomes of similar length, however for larger genomes alignment will be less reliable.

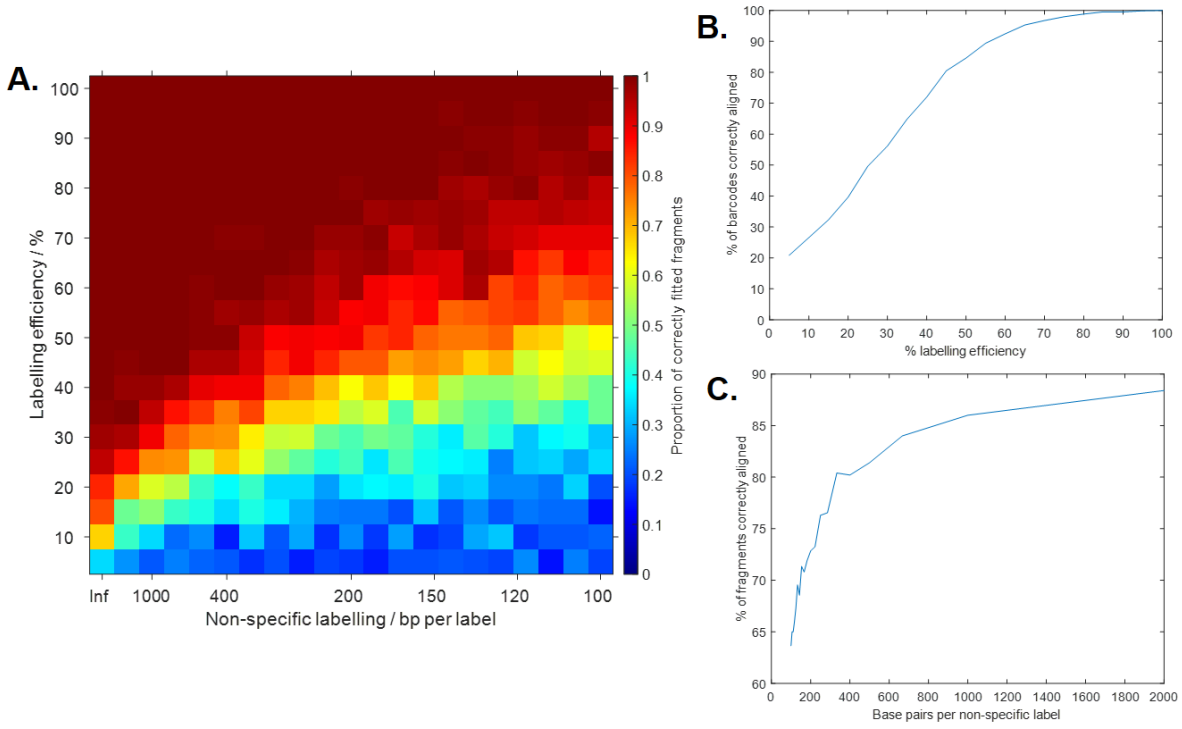


Figure 3.18 Simulating effect of labelling efficiency and non-specific labelling on alignment of DNA barcodes to/from T7 bacteriophage DNA. M.TaqI-directed labelling is simulated, with parameters given in supplementary materials and methods. A) Simulation from 5-100% efficiency and non-specific labelling of 0 to 1 in 100 base pairs. 100 barcodes are generated and aligned per pixel and colour indicates the proportion that was correctly aligned. B) Average number of barcodes correctly fitted against labelling efficiency. C) Average number of barcodes correctly fitted against the frequency of offsite labels.

Figure 3.19 shows the same results for DNA barcodes aligned to and generated from the *E. coli* K-12 genome (4640 kbp). This genome is approximately 100 times longer than the T7 genome (40 kbp), so the search for correct alignment is more difficult. Now around 80% labelling efficiency is required for 90% of barcodes to be correctly aligned, though the effect of non-specific labelling is similar. At 50% labelling efficiency and non-specific labels around every 1500 base pairs the proportion of correctly aligned barcodes is reduced to 60%. There is also a reduction at 30% labelling efficiency (and no non-specific labelling) to 66%.

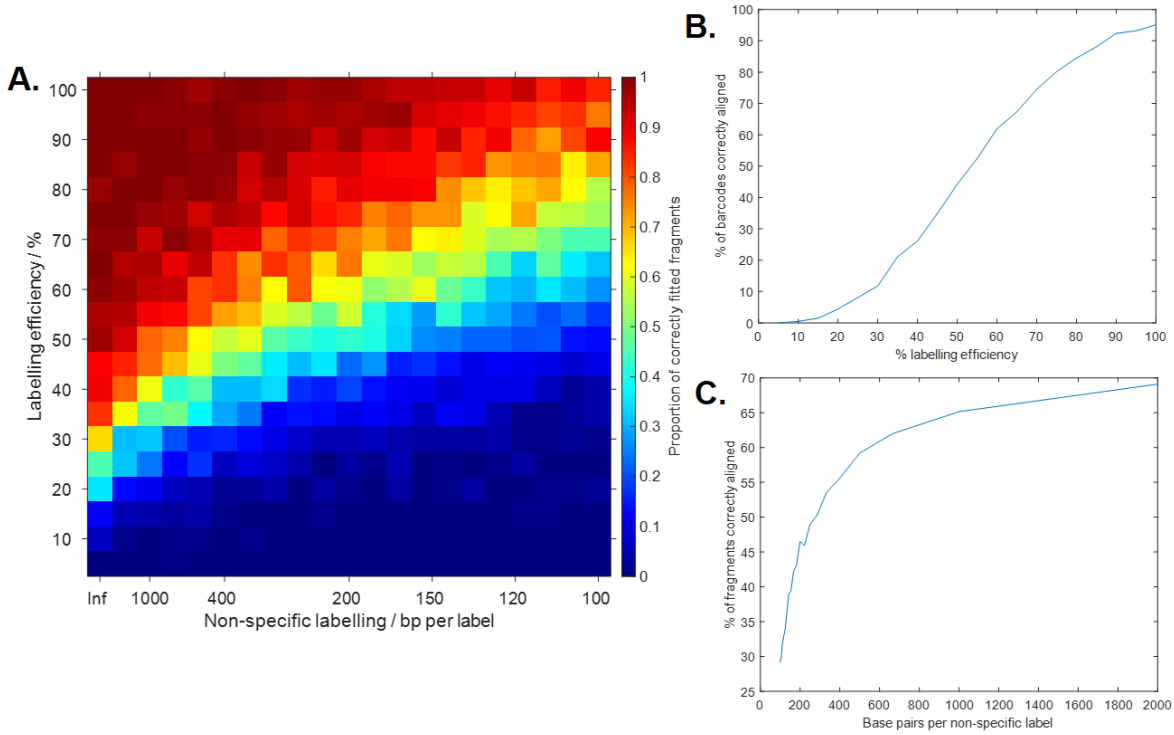


Figure 3.19 **Simulating effect of labelling efficiency and non-specific labelling on alignment of DNA barcodes to/from *E. coli* K-12. M.TaqI-directed labelling is simulated, with parameters given in supplementary materials and methods. A) Simulation from 5-100% efficiency and non-specific labelling of 0 to 1 in 100 base pairs. 10 barcodes are generated and aligned per pixel and colour indicates the proportion that was correctly aligned. B) Average number of barcodes correctly fitted against labelling efficiency. C) Average number of barcodes correctly fitted against the frequency of offsite labels.**

For long genomes such as *E. coli* the computational time for alignment of barcodes becomes apparent, the generation of Figure 3.19 took 6 hours, whilst Figure 3.18 only took 15 minutes. This is because although cross-correlation is relatively fast, (since fast Fourier transform algorithms can be used), long signals still take a long time to process. In cross-correlation the two signals must have the same effective resolution, but if a single intensity value for each base pair is used the time of computation is prohibitive. However, in reality the number of base pairs per pixel is only 200-400, so this is unnecessary and reduced sampling of the intensity profile can be used to speed up cross-correlation.

To reduce the sampling each barcode is initially stretched to the estimated number of base pairs, by interpolation, then a reduced sampling rate is used, i.e. the intensity is taken every 100 base pairs. The same sampling rate is used for the reference genome and this is used for alignment by cross-correlation. The effect of different sampling rates on the time taken for alignment is shown in Figure 3.20A. Increasing the sampling rate from 100 base pairs to 1000 base pairs decreases the computation time by a factor of more than 10. However, the rate of sampling will also affect the quality of the alignment procedure. This is shown in Figure 3.20B, where the effect of labelling efficiency is tested for various sampling rates. The alignment begins to suffer at sampling rates above approximately twice the pixel size (400-600 base pairs). So, for the results in the rest of this chapter a sampling rate of 100-300 bp will be used

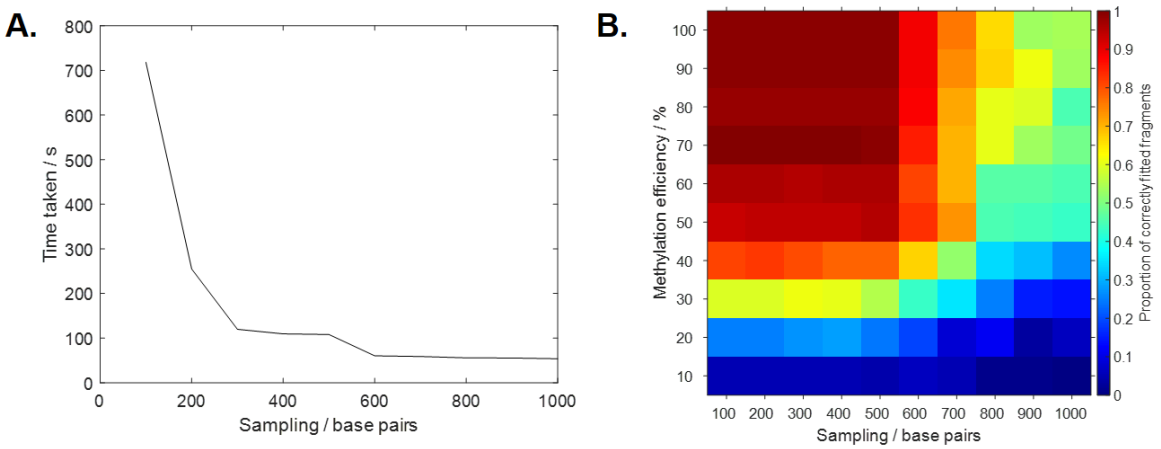


Figure 3.20 **Simulating the effect of sampling rate. Each DNA barcode is initially stretched to the estimated number of base pairs, by interpolation. Then sampling is used, i.e. the intensity is taken every 100 base-pairs. A) Time taken for alignment of 10 fragments to the T7 genome against the sampling rate. B) Histogram showing number of fragments correctly fitted to *E. coli* K-12, depending on the labelling efficiency and sampling rate.**

3.2.5 Measures of alignment accuracy

The normalised cross-correlation is used to align barcodes, as it is a rapid and reliable method to align signals. However, it is not always a good measure of alignment, i.e. whether a barcode is aligned correctly or not since maximising the cross-correlation is not a guarantee of good alignment, it might just be the best of a bad match. Ideally barcodes should be separated into correctly fitted barcodes (positive) and incorrectly fitted barcodes (negative, e.g. barcodes from a contaminating sample) with complete accuracy, for example using a threshold normalised cross-correlation of 0.7 (Figure 3.21A).

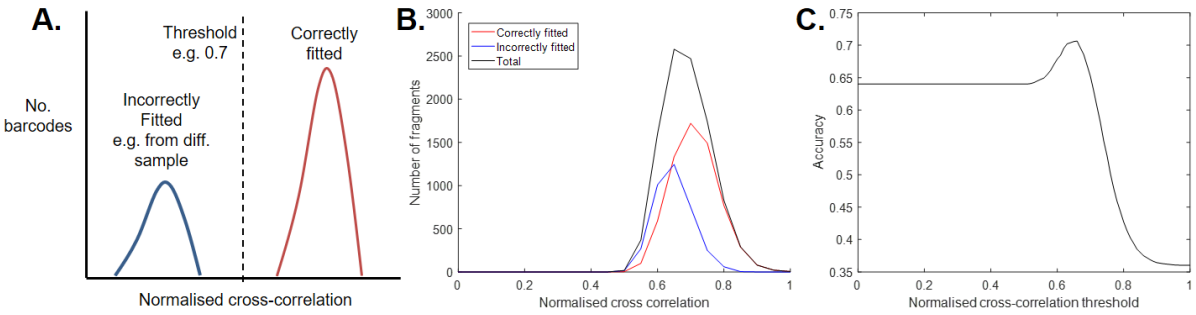


Figure 3.21 Normalised cross-correlation as a measure of alignment accuracy. A) Ideal separation by normalised cross correlation. Correctly fitted barcodes (red) will have a higher normalised cross correlation than incorrectly (blue). A threshold can be used to discriminate with 100% accuracy. B) Example for barcodes generated from and aligned to *E. coli* K-12, with 40% labelling efficiency. There is a large amount of overlap between correctly (red) and incorrectly (blue) fitted barcodes, making discrimination impossible from the total data (black). C) The accuracy of separation at different normalised cross-correlation thresholds, for data in B).

However, for typical experimental parameters (e.g. 40% labelling efficiency) there is not complete separation between correctly and incorrectly fitted barcodes (Figure 3.21B). At any chosen threshold there will be correctly fitted barcodes that are falsely assigned as negative and/or incorrectly fitted barcodes that will be falsely assigned as positive.

The accuracy for any threshold gives the percentage of barcodes that are correctly assigned (i.e. as being incorrectly or correctly fitting) and can be calculated as:

$$Accuracy = \frac{true\ positives + true\ negatives}{true\ positives + false\ positives + true\ negatives + false\ negatives}$$

This can be compared to other commonly used measures such as sensitivity and specificity. Sensitivity is the probability that a correctly fitted barcode will be identified as such, and specificity is the probability that an incorrectly fitted barcode will be identified as such. They can be calculated as:

$$Sensitivity = \frac{true\ positives}{true\ positives + false\ negatives}$$

$$Specificity = \frac{true\ negatives}{false\ positives + true\ negatives}$$

A peak accuracy of 71% is reached for a normalised cross-correlation threshold of 0.66 (Figure 3.21C), which is insufficient for reliable identification of a sample. Therefore, in addition to normalised cross-correlation other measures will be used to assess alignment.

Conceptually the location and height of peaks and troughs is important to align barcodes, therefore two additional measures will be used: the mean difference in the intensities of the generated and reference barcode (Figure 3.22A), which primarily reports on the height of peaks; and the mean difference in the gradient of the intensities (Figure 3.22B), which will primarily report on the location of peaks and troughs. These are subtracted from 1 to give reasonable distributions, where 0=no alignment and 1=perfect alignment.

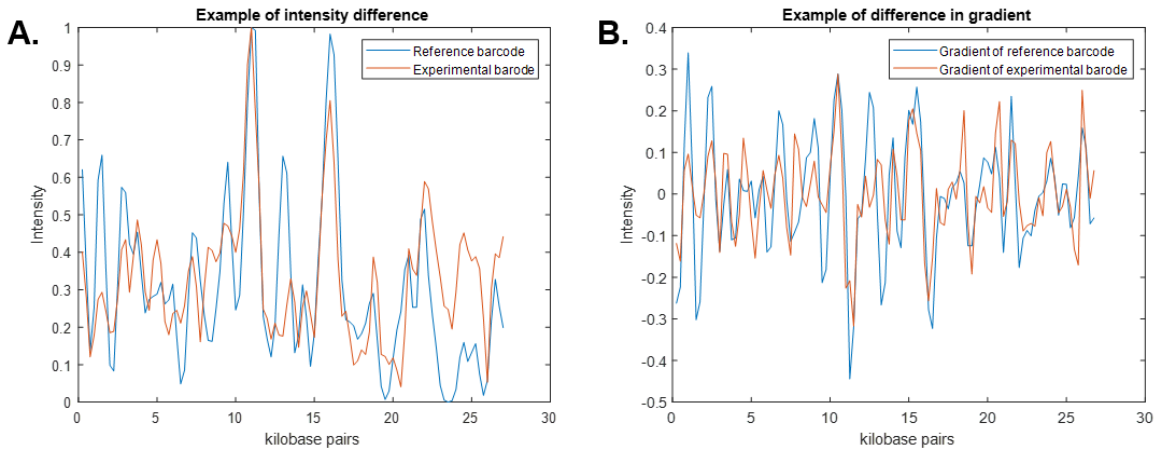


Figure 3.22 Example intensity profile for DNA barcode generated from and aligned to *E. coli* K-12. A) Intensity profiles. Across the fragment the difference between intensities is divided by the average intensity of the two signals, and then the mean used. B) Gradient of the intensity profiles. Across the fragment the difference between intensities is averaged.

The accuracy of these measures can be assessed as before (Figure 3.23A). Both measures are an improvement on using normalised cross-correlation: the difference of the intensities gives an accuracy of 76% at a threshold of 0.68; and the difference in gradients gives and accuracy of 80% at a threshold of 0.72. Combining the measures, using a mean, gives an overall accuracy of 81% at a threshold of 0.69, which is a significant improvement over using normalised cross-correlation alone.

Another way to assess the accuracy of each measure is to record the sensitivity and specificity for each threshold and to plot a receiver operating characteristic (ROC) curve. This is shown for each measure in Figure 3.23B. If there was ideal separation (i.e. 100% accuracy) then both sensitivity and specificity would equal one and the area under the curve (AUC) would equal 1. If there was no separation, then the ROC curve would fall along the dashed line and the AUC would equal 0.5. Plots which are closer to the ideal case and have a larger AUC indicate a better measure of discrimination. As before,

combining measures is shown to be the most effective method of discrimination with an AUC equal to 0.88, which can be considered excellent for accurate separation.

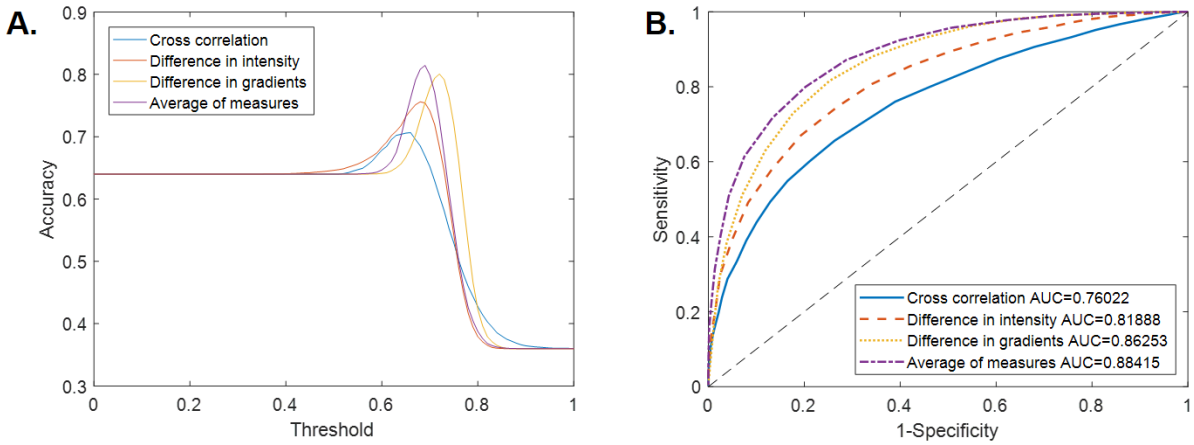


Figure 3.23 **The accuracy of separation using alternative measures. A) Accuracy of separation. B) Receiver operating characteristic (ROC) curves. 10,000 barcodes generated from and aligned to *E. coli* K-12, with 40% labelling efficiency. Correctly and incorrectly aligned barcodes are separated by several measures, using thresholds ranging from 0 (no alignment) to 1 (perfect alignment). Measures include normalised cross-correlation (blue); difference in intensity (red); difference in gradients (yellow); and an average of all three measures (purple).**

The combination of normalised cross-correlation, difference of intensities and difference in gradients will henceforth be referred to as the alignment 'weight'. It is important to note that this measure is only used to check barcode alignment. It is not used during the alignment procedure since an implementation would generally slow computation significantly; the reason cross-correlation is a popular method is because it is fast. However, an alignment procedure which used this measure, or similar, would likely improve accuracy, so implementation could be considered in future work.

3.2.6 Effect of labelling density and resolution

All these simulations have been carried out with M.TaqI (labelling TCGA sites) since it proved to be the most effective and reliable methyltransferase for labelling (see CHAPTER 2). However, labelling by other methyltransferases can easily be simulated, to determine if the barcodes would also be effective for alignment. Here enzymes with varying labelling density are trialled: M.EcoRI (low density, GAATTC sites); M.HhaI (medium density, GCGC sites) and M.MpeI (high density, CG sites).

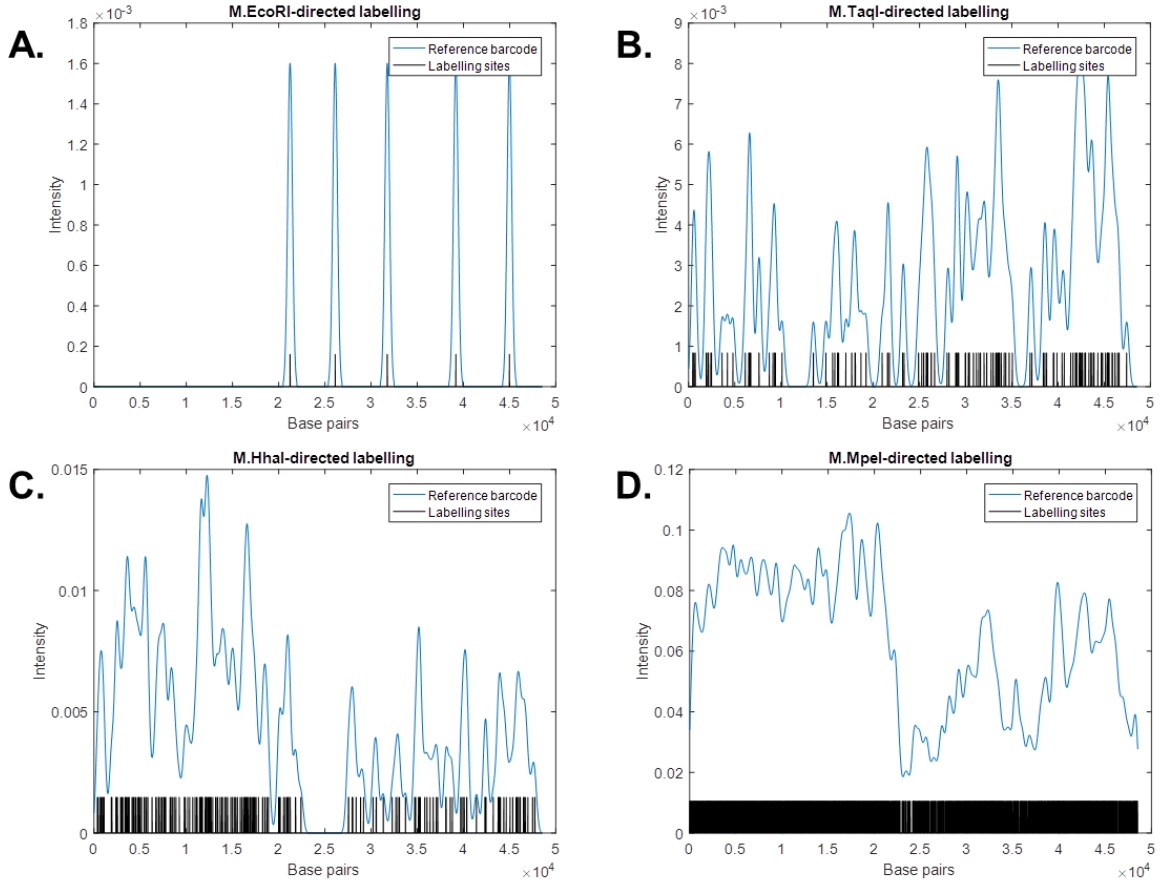


Figure 3.24 Reference barcodes for lambda genome (48.5 kbp), labelled by alternative methyltransferases. Shown in black are the labelling sites, and in blue the reference barcode with a PSF of 250 bp. A) M.EcoRI-directed labelling (GAATTC sites). B) M.TaqI-directed labelling (TCGA sites). C) M.HhaI-directed labelling (GCGC sites). D) M.MpeI-directed labelling (CG sites). These have different densities and highlight the difference in CG content across the lambda genome.

Figure 3.24 shows the reference barcodes generated by each methyltransferase-directed labelling reaction for lambda DNA (~48.5 kbp). There are only 5 EcoRI sites in the whole genome (Figure 3.24A), which is a density similar to traditional optical restriction mapping, whilst there are 121 TaqI sites and 157 HhaI sites, giving intermediate labelling densities, though different patterns. The lambda bacteriophage genome has a large difference in GC content across the genome, which is particularly visible in M.MpeI labelling, for which there are 3113 sites, giving an extremely high-density labelling, similar to the density seen for affinity labelling (since this typically uses GC content to generate the barcode).

Figure 3.25 shows the alignment of generated barcodes generated using these methyltransferases, depending on labelling efficiency. Figure 3.25A shows results of simulations for barcodes generated from and aligned to the lambda genome. It shows that M.MpeI barcodes are aligned well to the genome, due to the difference in GC content. This also means M.HhaI performs better than M.TaqI. However M.EcoRI performs poorly, as the very low density of sites gives insufficient information for robust alignment which is the reason traditional optical restriction mapping has not been widely used for short genomes (Figure 3.1).

Figure 3.25B shows the results for the T7 genome. Here there are no M.EcoRI sites, so the labelling fails completely for alignment. Now M.MpeI does not outperform the intermediate labelling densities (M.TaqI and M.HhaI), since there is less GC content variation across T7. Given the reference barcode we might expect it to perform even worse as the information content (i.e. number of and contrast between peaks and troughs) is rather low (Supplementary Figure 7.12).

If the simulations are run for the *E. coli* K-12 genome this is indeed what happens (Figure 3.25C). M.HhaI and M.TaqI perform as expected, however now M.MpeI-directed labelling is far poorer for alignment of barcodes. Even at 100% labelling efficiency only around 50% of barcodes are correctly aligned, compared to at least 90% for M.HhaI and M.TaqI-labelled barcodes. At 100% labelling efficiency even M.EcoRI-directed labelling is able to align 50% of barcodes correctly.

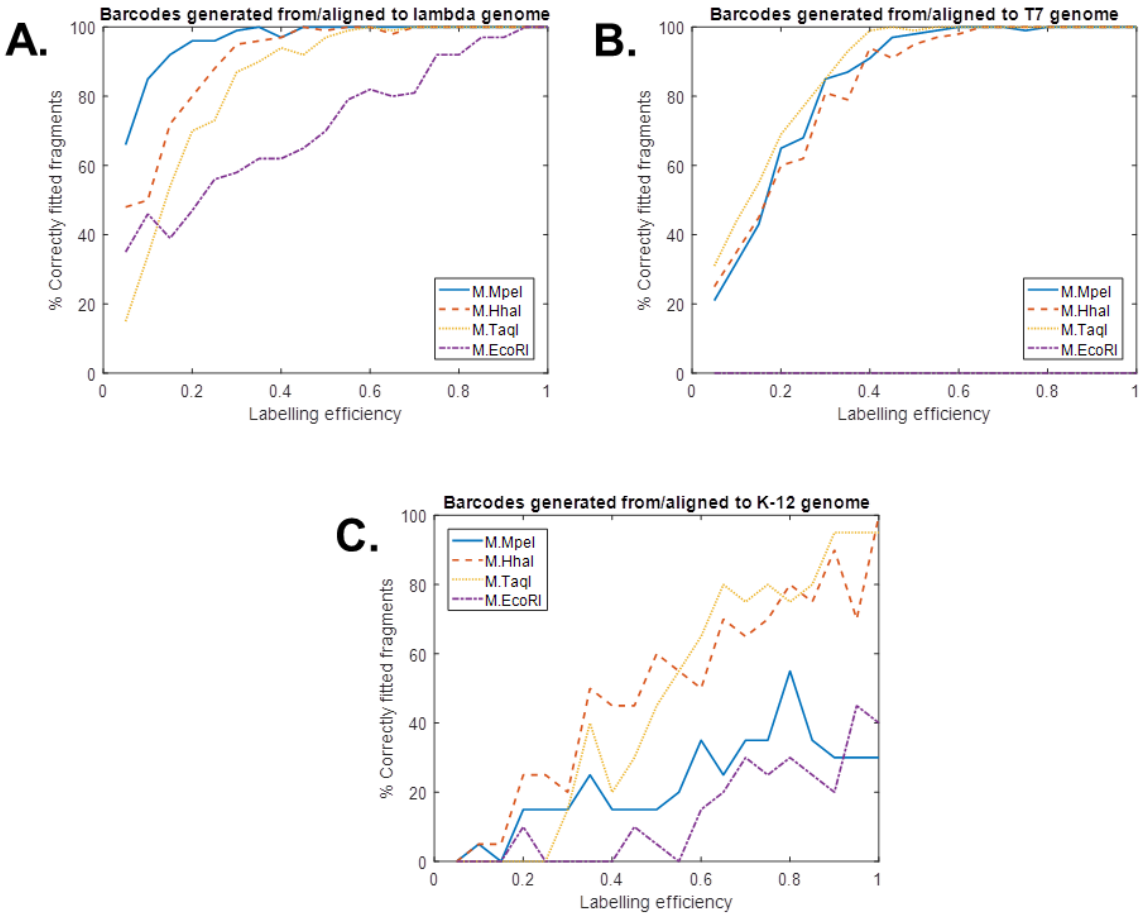


Figure 3.25 **Alignment of barcodes labelled by alternative methyltransferases. For 5-100% labelling efficiency, 100 barcodes were generated from and aligned to: A) lambda genome, B) T7 genome, C) *E. coli* K-12 genome. Barcodes were generated for different methyltransferases and non-specific labelling also varied: M.EcoRI, no non-specific labelling (purple, dash-dot); M.TaqI, 1 per 10 kbp (yellow, dotted); M.HhaI, 1 per 10 kbp (red, dashed); M.MpeI, 1 per 1 kbp (blue, solid).**

In the literature, localisation procedures have been used to improve the resolution of optical mapping[93]. This is equivalent to reducing the size of the PSF and increasing fluorophore localisation precision. The effect of the size of the PSF, on the reference barcode of the T7 genome, is shown in Figure 3.26. Here the PSF is varied from 10 base pairs, where almost all individual fluorophores can be localised with a very high precision, to a PSF of 1000 base pairs, where most information is lost, and only very broad peaks or troughs remain. These extremes are analogous to results in literature, where localisation of labels has been used for alignment[93] through to where barcodes are imaged in nanochannels, where thermal fluctuations reduce the effective PSF[83], see for instance Figure 1.20C and D.

For barcodes that are generated using typical experimental parameters (e.g. M.TaqI-labelled, with 50% labelling efficiency) the number of correctly aligned barcodes does not increase as the PSF width is decreased below around 500 bp, for barcodes generated from the T7 genome (Figure 3.27A), and below around 200 bp for barcodes from the *E. coli* K-12 genome (Figure 3.27C). At values higher than this there is a decrease in the number of barcodes that are correctly aligned, as the detail of the features is lost. It is important to note that for imaging barcodes that have been combed onto a surface the PSF width is typically 300-500 bp, however for nanochannels the reported intensity profiles have a PSF more typically of 500-1000 bp.

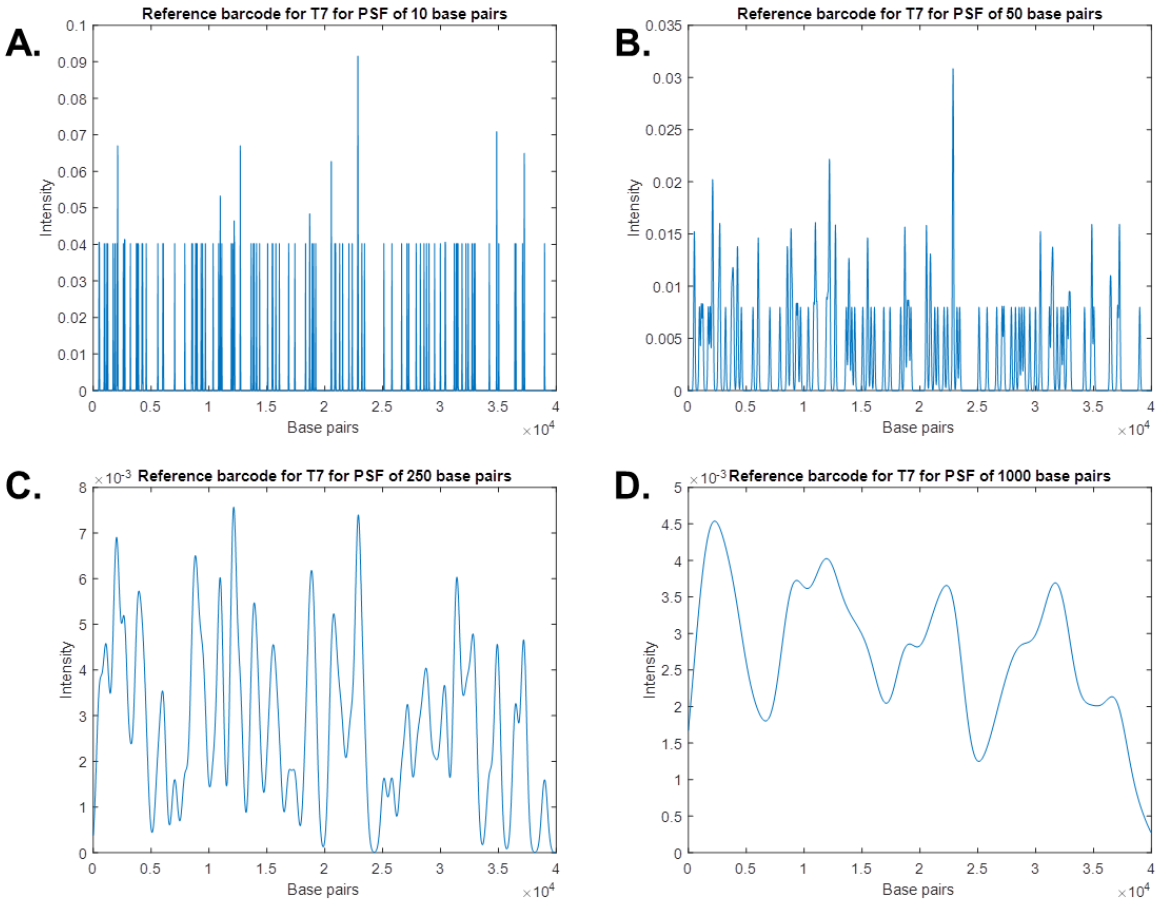


Figure 3.26 Reference barcodes for T7 genome with various PSF widths. Sigma values of: A) 10, B) 50, C) 250 and D) 1000 base pairs.

Below 200-500 bp there is still an advantage for increased localisation precision. The quality of alignment for correctly aligned barcodes improves significantly with increased localisation (Figure 3.27B and D). In effect the separation between correctly and incorrectly fitted barcodes is being increased (Figure 3.21A), which may improve the reliability of alignment overall.

Although this is significant, there are still a number of drawbacks to increasing localisation beyond the diffraction limit (~300-500 bp in a typical combing experiment). For super-resolution localisation methods, many time points are needed, significantly increasing imaging time, in addition to the processing time after imaging. For example, if

1000 frames are needed rather than a single frame, the data acquisition will be inherently 1000 times slower. Also significant is the higher sampling rate that is necessary once intensity profiles are extracted. This should be typically of the order of the pixel size, so increased localisation will lead to a higher sampling rate, which significantly slows computational speed (Figure 3.20). These factors will significantly reduce throughput, without a significant decrease in correct alignment, therefore non-localised, but combed, barcodes will be used.

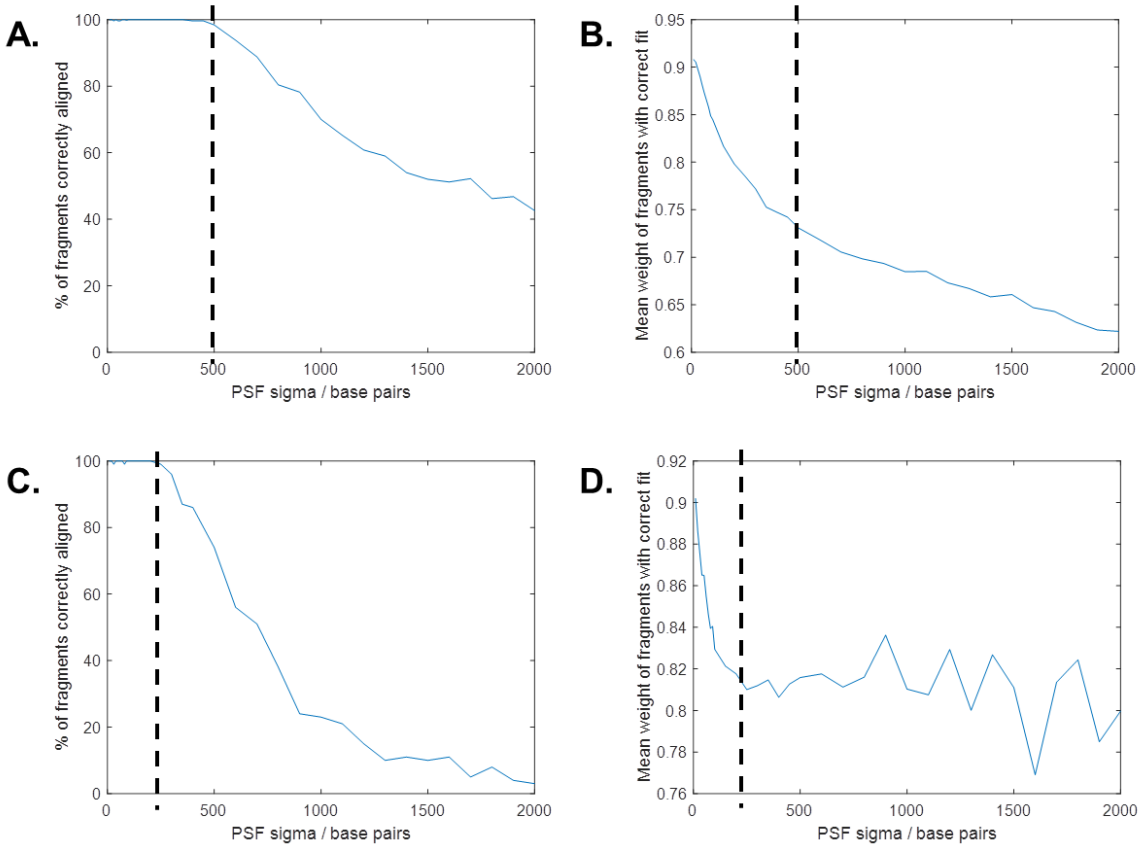


Figure 3.27 **Effect of PSF on alignment of barcodes. For PSF sigma of 10-2000, 100 barcodes, M.TaqI-labelled with 50% labelling efficiency, generated from and aligned to: A-B) T7 genome (dashed line at ~500bp), C-D) *E. coli* K-12 genome (dashed line at ~200bp). A and C) No. of correctly aligned barcodes, depending on PSF. B and D) The mean alignment weight for correctly aligned barcodes, depending on PSF. Although the no. of correctly aligned barcodes does not increase at a PSF width below the dashed line, the quality of alignment continues to increase.**

3.3 Conclusion

In silico generation and alignment of DNA barcodes has been used to investigate many of the important experimental parameters of optical mapping and to inform development of the analytical procedures, in particular the alignment procedure. As well as considering the parameters for the experimental methods used in this research, for example the labelling efficiency, there have been some comparisons with alternative optical mapping techniques (Figure 3.1).

In particular the main aspects to consider when comparing alternative techniques are the width of the PSF and the density of labelling sites. Typically, nanofluidic devices will have a relatively wide PSF (500-1000 bp), due to thermal fluctuations, whilst molecular combing can be easily used in tandem with localisation microscopy to achieve a very narrow PSF (down to around 20 bp). The density of sites varies, typically from around 1 site every 10 kbp, for traditional optical restriction mapping, to a much higher density of labelling along the whole DNA molecule as affinity-based labelling is very dense (for example CG sites occur on average once every 16 bp). The techniques used in this research have intermediate resolution (PSF width of 300-500 bp) and labelling density (around 1 label every 500 bp).

These regimes are compared in Figure 3.28. Genomes 5 Mbp in length, are generated with densities varying from 1 label every 10 bp (10^{-1} labels per base), to 1 label every 10 kbp (10^{-4} labels per base). From these genomes, barcodes are generated and aligned, with PSF widths ranging from 50 to 1000 bp. This confirms that for the identification of microorganisms (e.g. bacteria) alignment is poor at low effective resolutions as well as high or low labelling densities. Alignment is best at very high resolutions, but obtaining

this data reduces the throughput significantly. Therefore a 'sweet spot' lies in the intermediate range of values, at the labelling densities and resolution obtained for M.TaqI-directed labelling, molecular combing, and typical widefield microscopy. This confirms that the primary advantage of using the combination of these techniques is for the rapid identification of microorganisms.

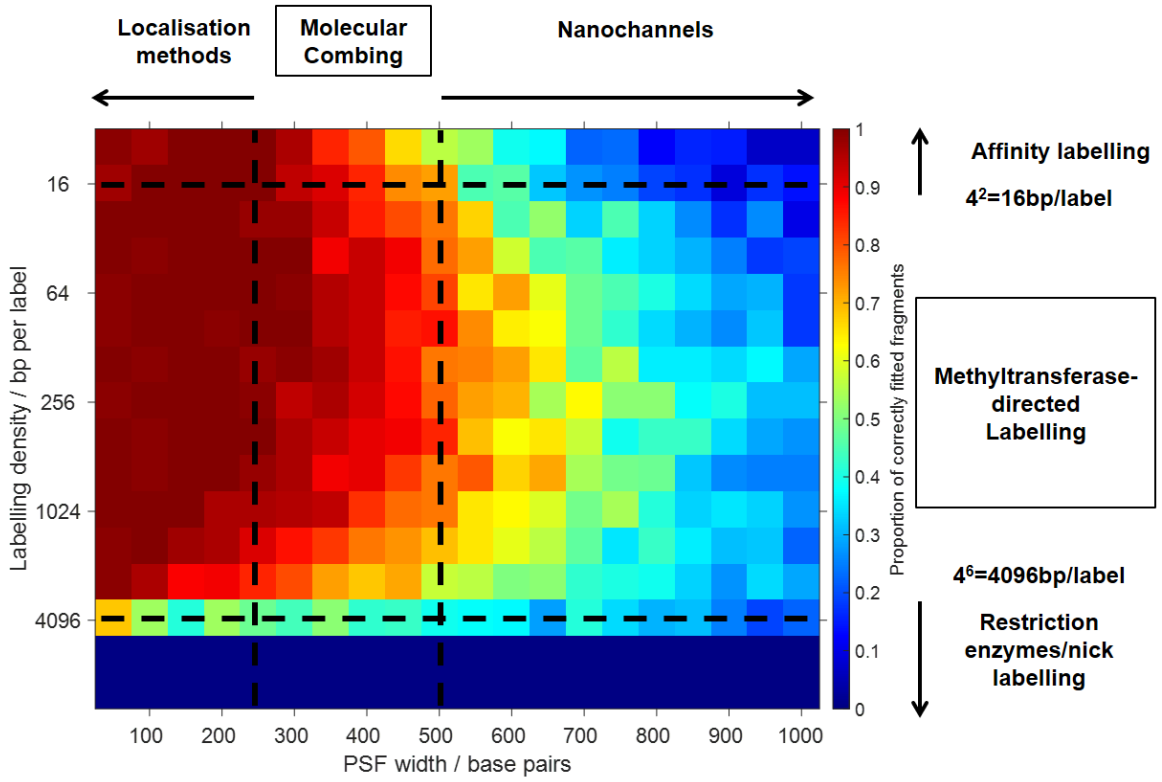


Figure 3.28 **Comparison of optical mapping techniques based on labelling density and PSF width. Random 5Mbp genomes with labelling densities of between 10^{-4} and 10^{-1} per base were generated. 40 barcodes with a PSF width of between 50 and 1000 bp, and 80% labelling efficiency, were generated from, and aligned to, each genome. The colour of the pixel represents the proportion of barcodes that were correctly aligned. Various experimental regimes are shown. The best alignment is at intermediate labelling densities ($10^{-1.6}$ to 10^{-3} labels per base) and high resolution (PSF width of <200 bp). However, obtaining this resolution significantly reduces throughput, therefore an intermediate resolution (PSF width of 300-500 bp) is preferred. This is the 'sweet-spot' that methyltransferase-directed labelling and molecular combing (with no localisation) occupies.**

In CHAPTER 4 this knowledge will be applied to experimental samples and will show the applicability of methyltransferase-directed labelling, molecular combing and alignment of intensity profiles for rapid identification of microorganisms.

3.4 Materials and Methods

3.4.1 Genomic DNA restriction assay (Figure 3.5)

The following were mixed and incubated at 50°C for 1 hour.

All in µL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Water	7.5	7.5	7.5	7.5	7.5	7.5	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	7.5	8.0	8.0
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
0.5mg/ml T7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
15mM AdoMet-azide	0.5	0.5	0.5	0.5	0.5	0.5									0.5		
0.3mg/ml M.TaqI	0.25	0.13	0.06	0.03	0.02	0.01	0.25	0.13	0.06	0.03	0.02	0.01	0.25	0.25			
32mM AdoMet													0.2				

0.5µl R.TaqI was added to samples 1-16 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 18mg/ml proteinase K /0.1% Triton X-100 to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

3.4.2 Labelling of genomic DNA

A 200µl solution containing 1x CutSmart (NEB), 10µg T7 DNA (NEB), 0.9µg M.TaqI (Ashleigh Rushton) and 750µM AdoHcy-azide (Andrew Wilkinson) was incubated at 50°C for 1 hour. 5µl 18mg/ml proteinase K (NEB)/0.1% Triton X-100 (Sigma-Aldrich) is added and incubated at 50°C for 1 hour, before purification by GenElute Bacterial Genomic DNA kit (Sigma-Aldrich) and elution into 200µl 1xTE (Sigma-Aldrich). Meanwhile a 20µl solution containing 0.5x PBS/50% DMSO, 1mM DBCO-amine (Sigma-Aldrich) and 12.5mM Atto647N (Sigma-Aldrich) was incubated at 4°C for 1 hour. The DNA sample was split into 30µl aliquots and either 10µl of the NHS Ester mix was added or 5µl DMSO, 5µl 1x PBS and 2.5µl 50mM TAMRA or Texas Red DBCO (Jena Bioscience) was added and incubated at room temperature overnight, before purification by GenElute Bacterial Genomic DNA kit and elution into 50µl 1xTE.

3.4.3 Molecular combing

Molecular combing of DNA based on procedure described by Deen *et al.*[57]

To prepare Zeonex-coated coverslips: Glass coverslips (20x20 mm, BRAND, no. 1) were cleaned to remove any fluorescent contaminants by incubation in a furnace oven at 450°C for 24 h prior to coating. A few drops of Zeonex solution (Zeon Chemicals, 1.5% w/v solution in chlorobenzene) was deposited onto a stationary coverslip and subsequently spun at 3000 rpm for 90s. Zeonex-coated coverslips were allowed to dry at room temperature overnight and stored in a desiccator.

For molecular combing: mix 2µl Atto647N-labelled DNA (2ng/µl in 1xTE), 17µl 100mM sodium phosphate buffer (pH 5.7) and 1µl DMSO. Deposit 1.5µl droplet on surface of Zeonex-coated coverslip, lower a pipette tip to contact the droplet and drag at 20mm/min across the coverslip. Image using TIRF/widefield, 100/150X, 640 nm excitation.

3.4.4 Automated extraction, *in silico* generation of barcodes and alignment procedures

Custom code was written using Matlab 2016b for automated extraction, *in silico* generation of barcodes and alignment procedures. Unless otherwise stated, barcodes were generated using parameters in Supplementary Table 7.1. Copy of the code is available on request from Robert Neely, University of Birmingham.

CHAPTER 4 OPTICAL MAPPING FOR IDENTIFICATION OF COMPLEX MIXTURES OF VIRAL AND BACTERIAL DNA – EXPERIMENTAL SAMPLES

Robert K. Neely and Iain B. Styles provided supervision and guidance for the research undertaken in this chapter. Nathaniel O. Wand (the author) designed, performed and analysed all labelling experiments including optical mapping experiments. Nathaniel O. Wand (the author) also developed and performed all extraction and alignment procedures unless otherwise stated.

4.1 Introduction

In CHAPTER 3 barcodes were generated *in silico* to test the various experimental parameters and inform analytical procedures.

In terms of optical mapping techniques, M.TaqI-directed labelling and molecular combing provide the 'sweet spot' of labelling density and resolution. It was shown that labelling efficiency is the most important parameter for alignment. A labelling efficiency of 50% is sufficient to align around 80% of barcodes to the T7 genome which is at the upper end of the experimental labelling efficiency (20-50%). In addition, it was shown that the maximum non-specific labelling seen experimentally (1 label per 1500 bp) will not significantly affect alignment, whilst barcodes of the experimentally obtained length (20-50 kbp) are sufficient for alignment. Therefore, it is likely the limiting factor for this optical mapping technique may be the molecular combing and the extraction of experimental barcodes.

In silico results informed development of a straightforward alignment procedure based on normalised cross-correlation. For this alignment procedure, a sampling rate of 100-

500 bp can be used to reduced computation time and the alignment weight, (an average of the normalised cross-correlation, the difference in barcode intensities and the difference in the gradient of barcodes), gives good alignment accuracy.

In CHAPTER 4 barcodes extracted from experimental samples of DNA will be aligned to short reference genomes (10 kbp to 10 Mbp). In addition to straightforward alignment, d*e novo* alignment will be used, for simple and more complex mixtures. Together with CHAPTER 3, these results will show the applicability of methyltransferase-directed labelling, molecular combing and alignment of intensity profiles for rapid identification of microorganisms.

4.2 Results and discussion

4.2.1 Alignment of experimental barcodes to short reference genomes

Ideal molecular combing of DNA molecules would produce optical images with distinct and uniformly stretched individual molecules, from which individual barcodes could be extracted. However, in practice ideal molecular combing is difficult to achieve and the automated extraction of individual DNA barcodes remains challenging.

Consequently, before the barcodes extracted from experimental samples can be aligned, it is important to clean-up the data, to exclude barcodes which did not represent individual, correctly extracted barcodes. This is done in several steps.

The first step is to select barcodes based on size. The longer a barcode is, the better the alignment (3.2.4) and the maximum size may be known based on the sample. For example, for lambda DNA (~50 kbp) barcodes that are 35 kbp to 55 kbp can be selected (Figure 4.1A). A few kbp are extracted at either end of the barcode, so an allowance of 5-10 kbp is used. Once these barcodes are selected they are also selected by average intensity (e.g. 5000-30000 for a 16-bit image, Figure 4.1B). Dim barcodes are likely to be artefacts from the extraction procedure (i.e. background rather than a labelled DNA fragment), whilst bright barcodes are likely to result from overlapping DNA fragments.

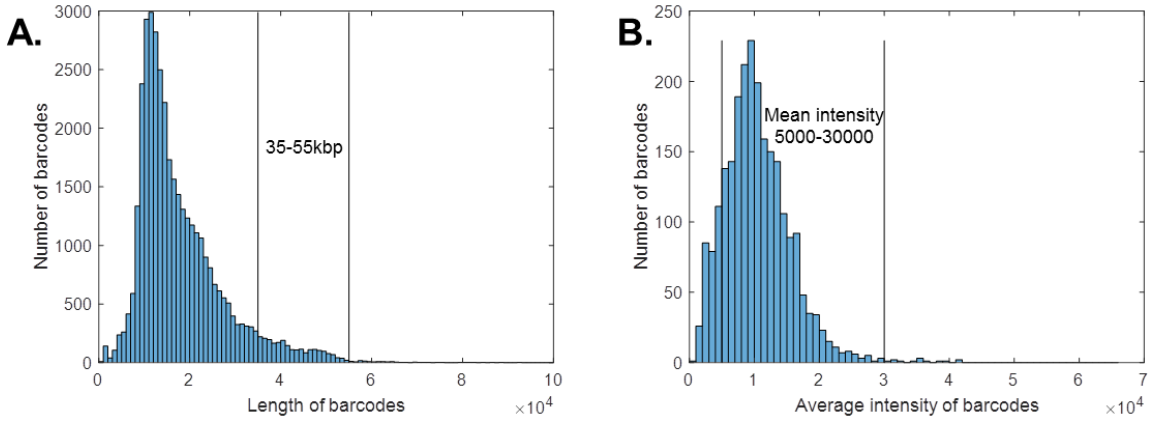


Figure 4.1 **Selecting DNA barcodes based on length and average intensity. An example here is given for Atto647N-labelled lambda DNA. A) Length of barcodes. Barcodes 35-55 kbp in length are selected. B) After selection by length, the average intensity of barcodes. Barcodes with mean intensities between 5000 and 30000 are selected.**

The next step is to select barcodes based on the intensity profile. The barcode is cut at the ends, to remove the extra 5-10 kbp (noted previously, see Figure 3.12). Next, barcodes with very bright and/or very dim regions should be discarded, as they are likely artefacts from the extraction procedure or due to overlapping DNA molecules (Figure 4.2A and B). A rolling mean, a fifth of the size of the barcode, is calculated along the barcode. When the ratio of the minimum to the maximum rolling intensity is very small this typically indicates very bright or very dim regions along the intensity profile, and so the DNA is discarded (e.g. ratio minimum:maximum of 0.2, Figure 4.2C).

Finally, DNA fragments which have saturated the detector are discarded. This is indicative once again of overlapping DNA molecules.

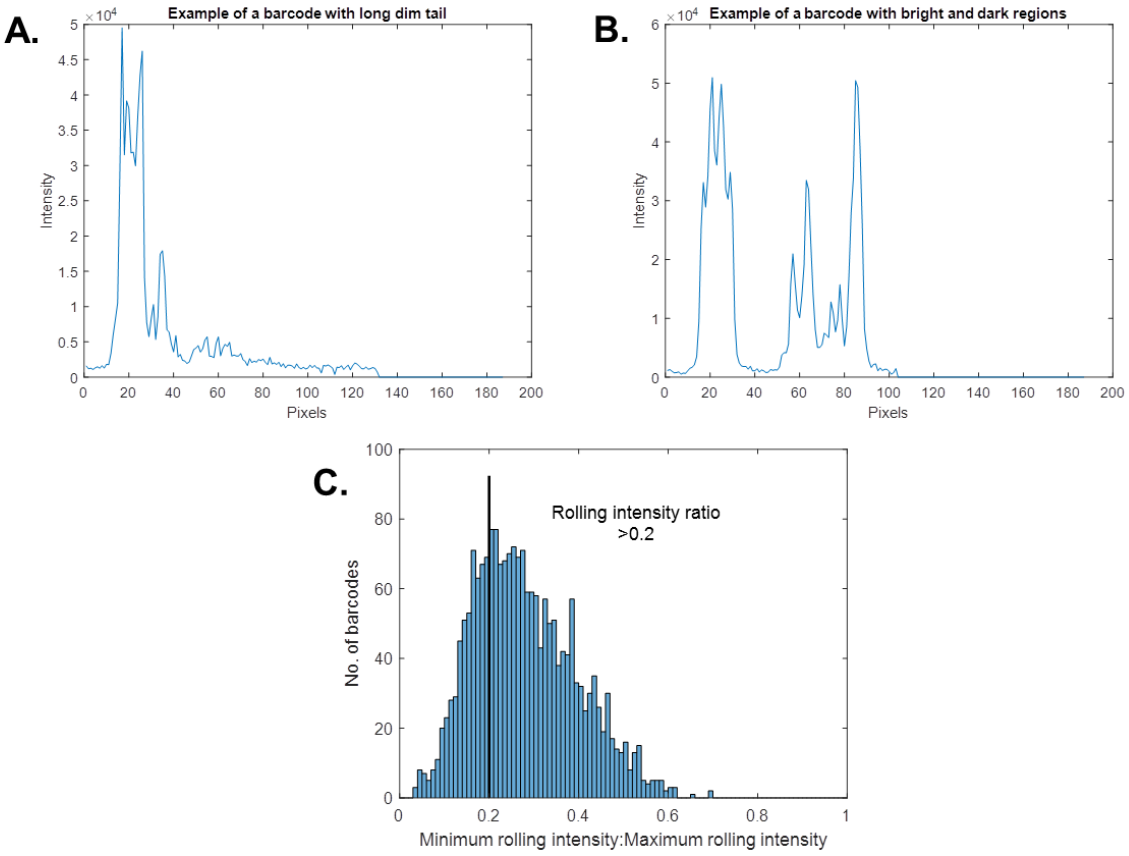


Figure 4.2 **Selecting DNA barcodes are based on the intensity profile. An example here is given for Atto647N-labelled lambda DNA. A, B) Examples of discarded fragments based on the intensity profile. A) This barcode is dim for much of its length, probably as it is not extracted properly. B) This fragment has both very bright and very dim regions, meaning overlapping DNA molecules have probably been extracted. C) A rolling mean of the barcode, a fifth of the size of the barcode, is calculated. The ratio of the minimum:maximum rolling mean is calculated and used to select barcodes (e.g. with a rolling intensity ratio >0.2).**

Using the alignment procedure developed for barcodes generated *in silico*, it is now possible to align extracted and cleaned barcodes to a known reference sequence. For short genomes, i.e. <100kb, 1000 fragments take around 30 seconds to align. The combined alignment weight can then be used to extract those fragments that have aligned well. A cut-off value of 0.7 is used, since this gave a good level of accuracy (~80%) in simulations (CHAPTER 3).

Here the extraction and alignment procedures have been used for samples of lambda and T7 DNA labelled using M.TaqI. Figure 4.3 shows typical results for a pure sample of lambda DNA. Around 38,000 raw barcodes (approximately 200 Megabases) are extracted from around 1,000 images of combed DNA, from which 1077 cleaned barcodes are finally extracted (Figure 4.3A). Each of these is aligned, in turn, to the experimental barcode, with a PSF of 300 bp. An example of a barcode that aligns well (i.e. has a high alignment weight) is shown in Figure 4.3B and a histogram of alignment weights for all 1077 barcodes is shown in Figure 4.3C. A threshold of 0.7 is used to filter 368 well-aligned molecules, which are shown in Figure 4.3D. The mean intensity of these is calculated and shown at the bottom of Figure 4.3D and in Figure 4.3E. The background from this can be removed by using a 'rolling ball' with a suitable diameter (e.g. 50 pixels), this is also shown in Figure 4.3D (second mean barcode from bottom of image).

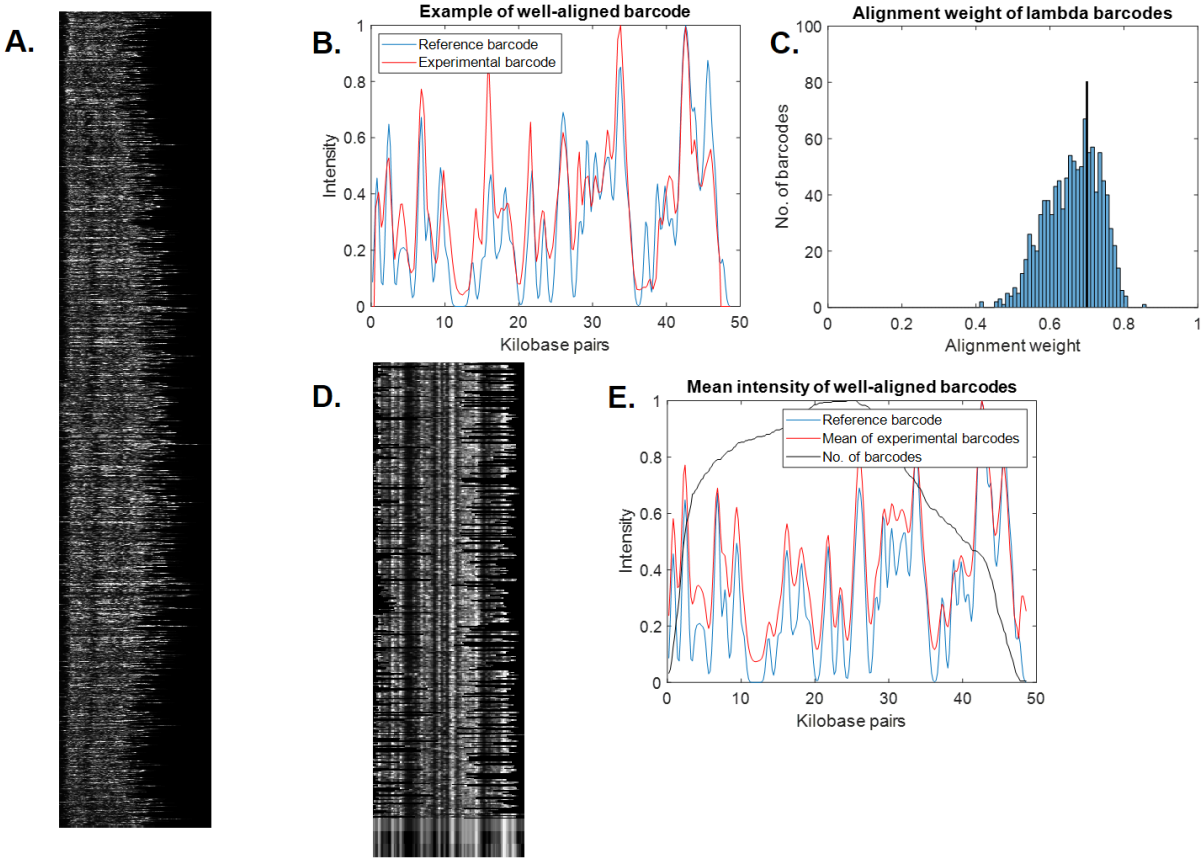


Figure 4.3 **Alignment of pure experimental sample of lambda DNA. Lambda DNA was labelled using M.TaqI with Atto647N, combed and imaged. 38,000 raw barcodes were extracted from the images. A) After cleaning 1077 barcodes are extracted. B) An example of an experimental barcode (red) that aligns well to the reference barcode (blue). C) Alignment weight of all experimental barcodes, with a threshold of 0.7 shown (black). D) Alignment of 368 barcodes with alignment weight greater than threshold. At the bottom of the image is shown the mean experimental barcode, the mean with the background removed and the reference barcode (top to bottom). E) Plot of mean experimental barcode (red) against the reference barcode (blue), with the number of barcodes (black).**

The same results for a pure sample of T7 DNA are shown in Figure 4.4. Here 1166 barcodes are extracted from the sample, of which 174 (15%) have an alignment weight greater than 0.7.

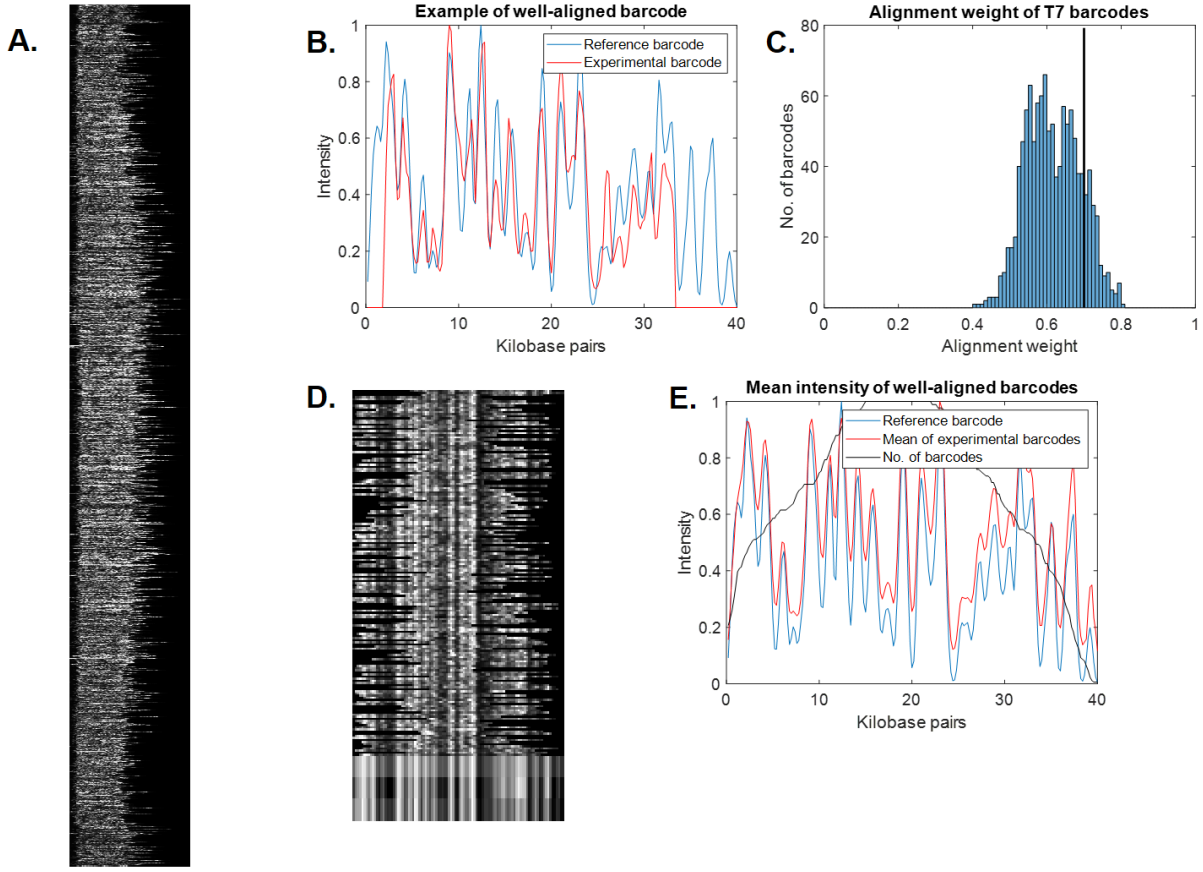


Figure 4.4 **Alignment of pure experimental sample of T7 DNA. T7 DNA was labelled using M.TaqI with Atto647N, combed and imaged. A) After cleaning 1166 barcodes are extracted. B) An example of an experimental barcode (red) that aligns well to the reference barcode (blue). C) Alignment weight of all experimental barcodes, with a threshold of 0.7 shown (black). D) Alignment of 174 barcodes with alignment weight greater than threshold. At the bottom of the image is shown the mean experimental barcode, the mean with the background removed and the reference barcode (top to bottom). E) Plot of mean experimental barcode (red) against the reference barcode (blue), with the number of barcodes (black).**

From the simulations in CHAPTER 4, and the experimental labelling efficiency we would expect greater than 90% of barcodes to be correctly aligned. It is unlikely that large numbers of correctly aligned barcodes are excluded based on the alignment weight threshold, therefore it appears that there are a larger than expected number of barcodes that cannot be aligned well to the reference barcode, for both samples. These will be described henceforth as 'junk' barcodes.

One source of these 'junk' barcodes will be the molecular combing and extraction procedures. Ideal molecular combing is difficult and consequently automated extraction of individual barcodes remains challenging, as discussed above. The artefacts of molecular combing that are retained during barcode extraction, for instance overlapping DNA molecules, may thus be one source of barcodes that do not align to the reference.

Another possible source of 'junk' barcodes is contamination, for instance during the labelling and purification procedure. Even a handful of cells would be capable of introducing foreign DNA to the sample and is a known problem in DNA sequencing for which tools have been developed, for example to discard sequences obtained from contaminating DNA[180]. Contamination may be reduced if greater care is taken during DNA preparation and labelling, for instance by ensuring all reagents, columns and tubes are free of contamination.

4.2.2 Identification of mixtures by alignment to short reference genomes

For many samples this simple alignment is not applicable, since it requires a single known reference, which all experimental barcodes are aligned to. In contrast, the identification of microorganisms would require all experimental barcodes to be aligned to a library of reference barcodes. In reality, a DNA sample will also often contain a mixture of genomes. However, if the alignment of DNA barcodes is reliable, then individual barcodes should align well enough to a library of reference genomes to identify a mixture of DNA.

Here this simple procedure will be applied on a mixed sample containing T7 and lambda DNA. If the data from Figure 4.3 and Figure 4.4 is combined then we have an effective ground-truth, since the source of each barcode is known. When each barcode is aligned to both T7 and lambda reference barcodes we can examine the results.

The alignment weights for each barcode, aligned to both references, are displayed as a joint scatter and histogram in Figure 4.5. There is separation between the samples, although the separation is not complete (due to the issues discussed in Figure 3.21). However, an alignment weight threshold of around 0.65 would be sufficient to separate most molecules successfully, in other words an alignment weight above 0.65 means the DNA is likely to have come from that specific genome.

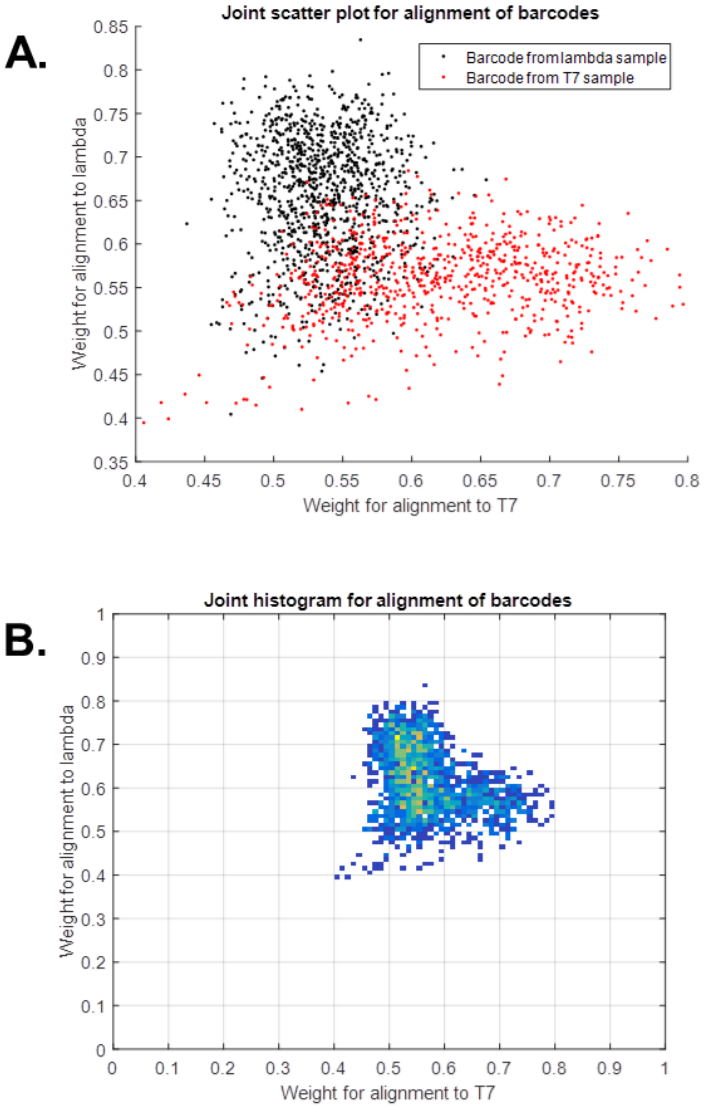


Figure 4.5 **Alignment of artificially combined samples of pure lambda and pure T7 DNA. DNA was labelled using M.TaqI with Atto647N, combed and imaged. A) Scatter plot showing weight of each molecule for alignment to both the T7 and lambda reference barcodes. Molecules from the pure lambda sample (black) and from the pure T7 sample (red) can clearly be separated. B) Joint histogram showing the same information.**

If a real mixture of DNA is used, i.e. a 1:1 mixture of T7 and lambda DNA then the following results can be obtained which agree well with the artificial mixture (Figure 4.6). 1756 barcodes are extracted (Figure 4.6A), of which 136 align well to lambda (Figure 4.6B) and 161 align well to T7 (Figure 4.6C). The distribution of alignment weights is shown in Figure 4.6D and E.

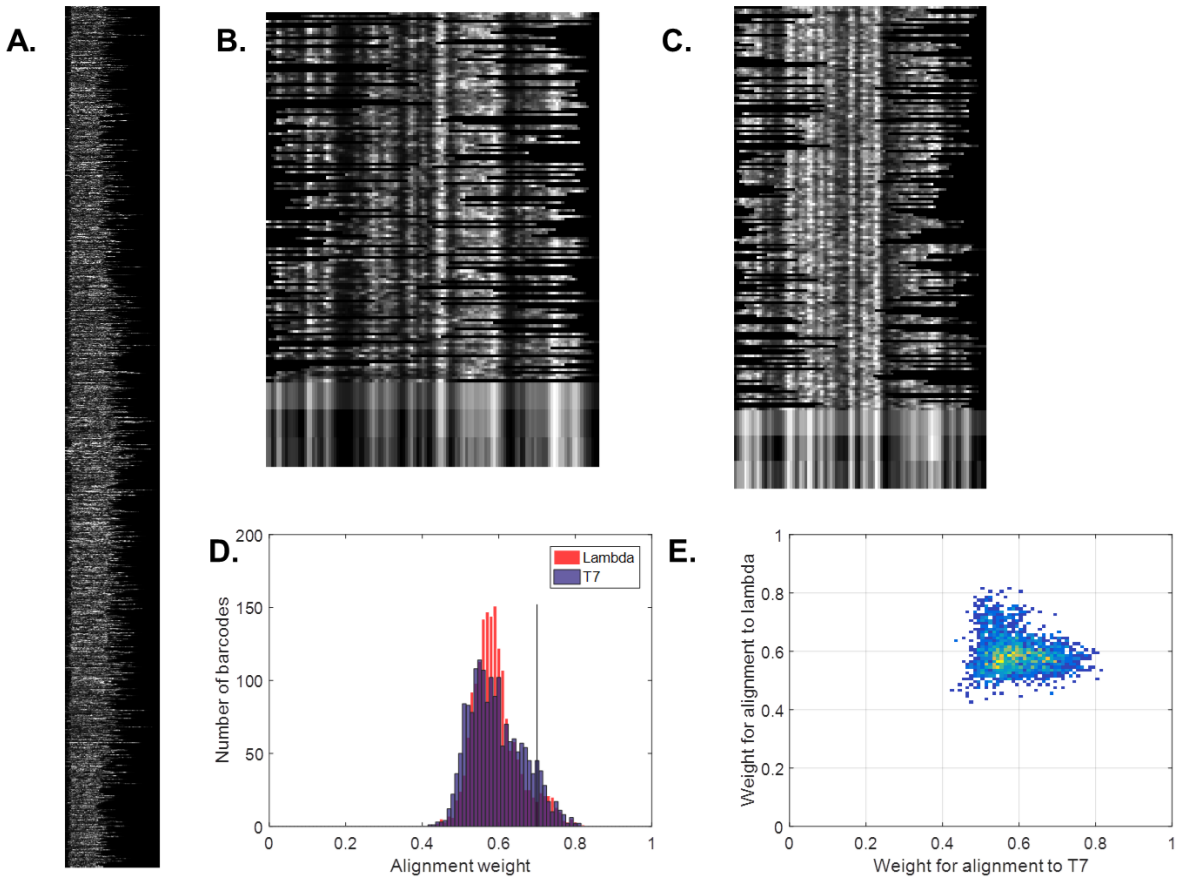


Figure 4.6 **Alignment of experimental sample of mixed T7/lambda DNA. DNA was labelled using M.TaqI with Atto647N, combed and imaged. A) After cleaning 1756 barcodes are extracted. B-C) Alignment of barcodes with alignment weight greater than threshold. At the bottom of the image is shown the mean experimental barcode, the mean with the background removed and the reference barcode (top to bottom). B) For lambda reference 136 barcodes are fitted. C) For T7 reference 161 barcodes are fitted. D) Alignment weight to lambda(red) and T7 (blue) of all experimental barcodes, with a threshold of 0.7 shown (black). E) Joint histogram for weight of each barcode for alignment to both the T7 and lambda reference barcodes.**

These methods can be extended to identify bacteriophage DNA. If the barcodes are aligned to a library of reference barcodes, then the genome from which the DNA originated should be identified. This has already been used on a small number of DNA barcodes labelled using M.TaqI, against a small library of bacteriophages by Grunwald *et al*[83].

Grunwald *et al*. used nanochannels to stretch DNA labelled using M.TaqI labelled with TAMRA, from which intensity profiles were extracted. This method is inherently lower-throughput than DNA combing, since the molecules are moving and a time-lapse must be recorded, therefore only 85 barcodes were aligned. Extracted barcodes were aligned by cross-correlation and assigned to the genome for which the maximum cross-correlation was obtained. The results are shown below in Figure 4.7. Barcodes were either from a sample of pure lambda DNA (red), of from pure T7 DNA (black), which are clearly identified. A similar procedure has also been reported for resistance plasmids, affinity-labelled and mapped in nanochannels[61].

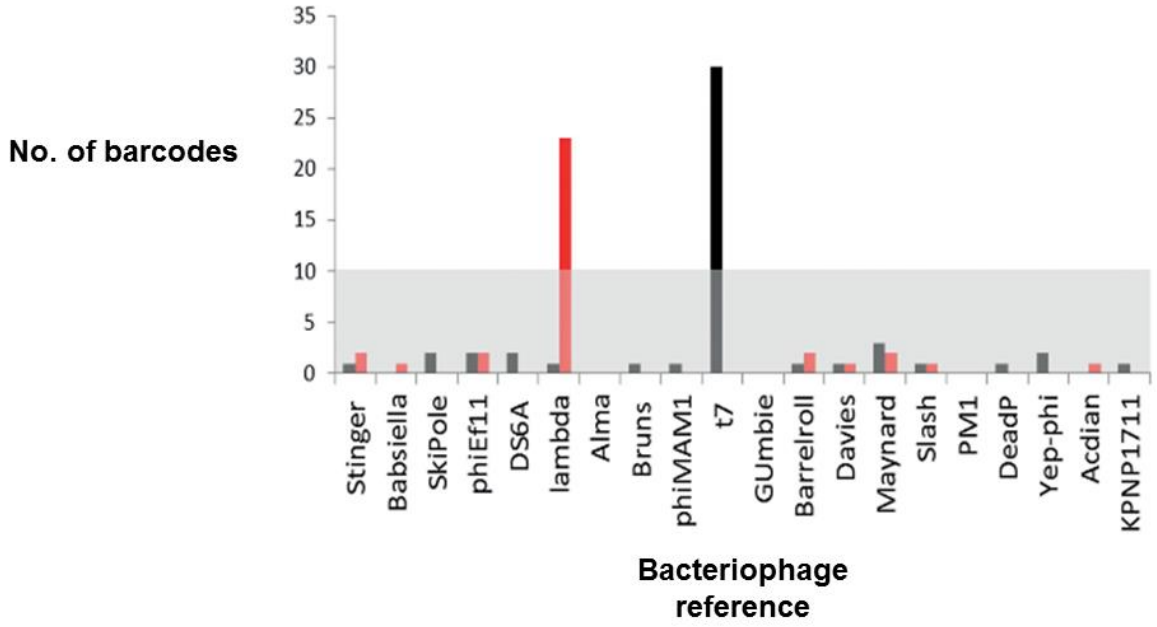


Figure 4.7 **Identification of bacteriophage DNA, adapted from Grunwald et al. Experimental barcodes from 35 lambda molecules (red) and 50 T7 molecules (black), labelled using M.TaqI and TAMRA, were aligned, using normalised cross-correlation, to 20 different phage reference genomes. Each experimental barcode was assigned to the phage to which its alignment yielded the highest normalised cross-correlation. The histogram shows the numbers of experimental barcodes assigned to each phage.**

This type of analysis can be carried out on the samples which were aligned previously (Figure 4.8). In Figure 4.8A similar results are shown, but for over ten times more molecules and using the combined weight measure instead of normalised cross-correlation.

The T7 sample in particular has many barcodes which are assigned to other genomes. This is presumably because molecular combing has introduced the problems mentioned previously: artefacts in barcode extraction and overlapping DNA molecules. Therefore, there are many 'junk' barcodes, which will either be assigned to a reference barcode which happens to share the same pattern, or more generally to a reference barcode with a particularly intense region (since normalised cross-correlation is biased towards bright regions). It is unlikely that the other viruses are present in the sample, therefore barcodes which are assigned to other barcodes are likely to be junk barcodes. Examples of junk barcodes which are assigned to other genomes are shown in Figure 4.9.

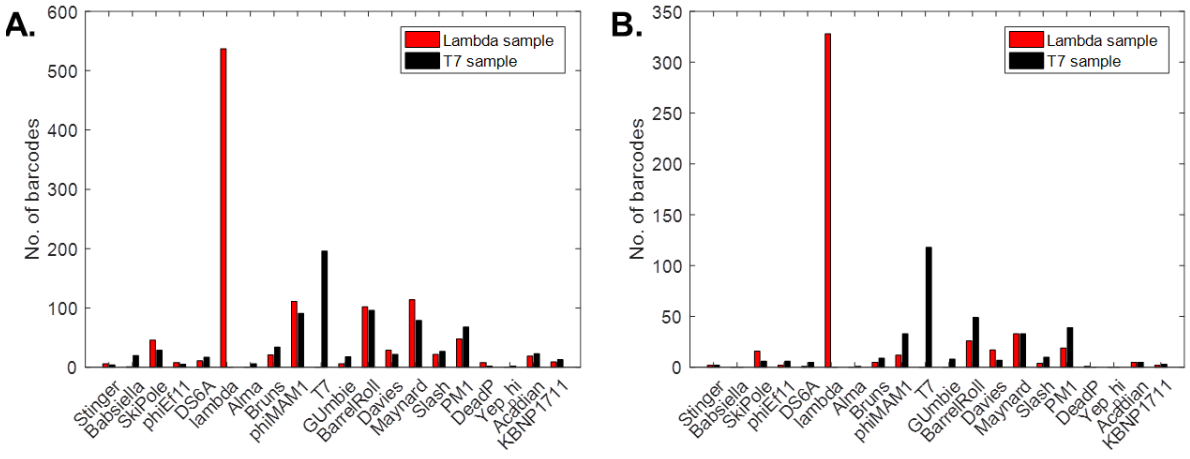


Figure 4.8 **Identification of bacteriophage DNA from pure samples. 1098 experimental barcodes from pure lambda sample (red) and 752 from pure T7 sample (black) aligned and assigned to 20 different phage genomes. A) Each experimental barcode was assigned to the phage to which its alignment yielded the highest alignment weight. B) The number of fragments fitting to each genome with an alignment weight greater than 0.7**

Using a threshold (e.g. alignment weight>0.7) to remove junk barcodes improves the results (Figure 4.8B), the identity of the sample is clear from the alignment of the barcodes. However, a notable feature of this analysis is that the greater the number of phage genomes, the more experimental barcodes are needed, as they become assigned to different genomes by chance.

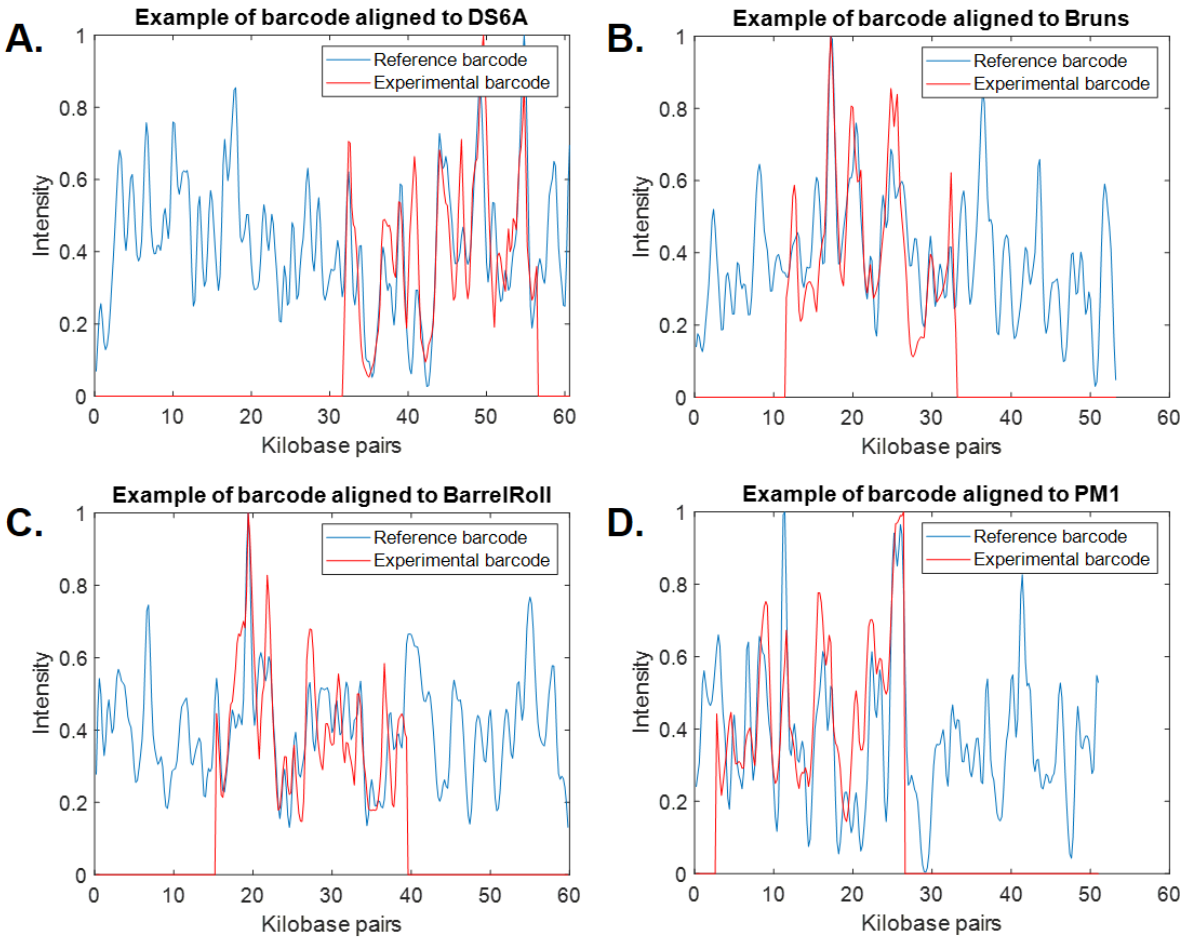


Figure 4.9 **Examples of fragments assigned to genomes other than T7/lambda. 848 experimental barcodes (red) from a mixed T7/lambda sample were aligned to 20 bacteriophage genomes (reference barcodes, blue). Examples of fragments assigned to genomes other than T7/lambda are shown. Aligned to: A) DS6A, B) Bruns, C) BarrelRoll, D) PM1.**

This has also been tested with the T7/lambda experimental mixture (Figure 4.10), which further highlights the problems with this approach for this type of data. From the approach Grunwald *et al*. used (Figure 4.10A) it is very difficult to identify the mixture, although it is easier when the approach of using thresholds is used (Figure 4.10B). This is because there will always be some good alignment to genomes which the DNA is not derived from, particularly for junk DNA. For more complex mixtures the problem would be exacerbated further.

A still more fundamental issue exists with this procedure. It takes around 10 seconds to align 1000 experimental barcodes against each reference, therefore the total analysis takes around 10N seconds in total, where N is the number of genomes in the reference library. For running against a library of 2000 genomes this means the procedure takes around 5 hours. If the DNA is not in the library, then such an analysis fails to produce the correct match.

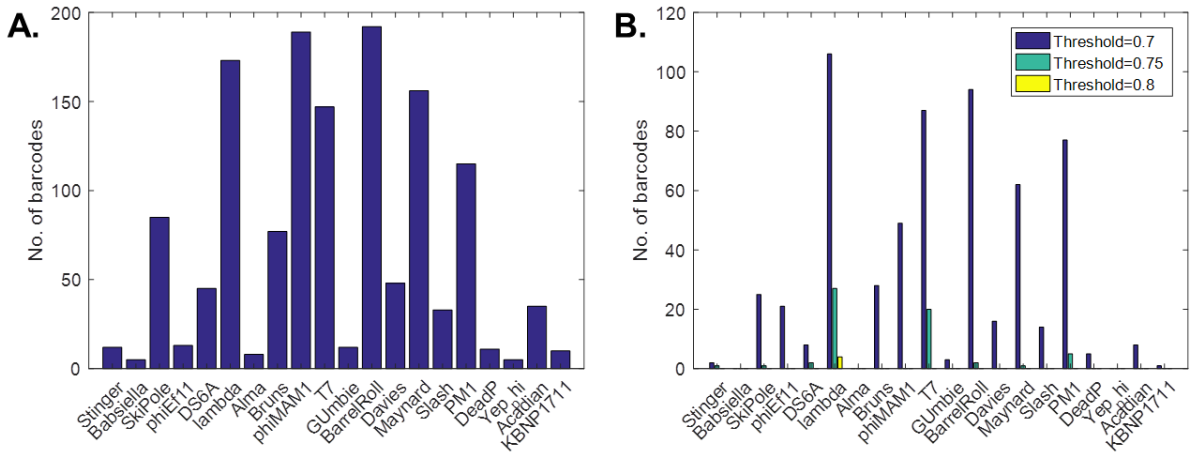


Figure 4.10 Identification of bacteriophage DNA in a mixed sample. 1756 experimental barcodes from a mixed T7/lambda sample. A) Each experimental barcode was assigned to the phage to which its alignment yielded the highest alignment weight. Note that lambda and T7 cannot be readily identified. B) The number of experimental barcodes aligned to each reference genome with an alignment weight greater than 0.7 (blue), 0.75 (cyan) or 0.8 (yellow).

4.2.3 Applications of simple alignment procedure

Although the simple alignment procedure described above has severe limitations when used to identify DNA in a reference library, it is useful for checking the quality, methylation state, or sequence of short genomic DNA, for instance from a sample of known viral DNA. The number of barcodes that align well to the genome is indicative of the quality, whilst the mean alignment may show any discrepancies between the experimental barcode and reference barcode. An example of this is shown in Figure 4.11. Dam methylation (5'-G<u>A</u>TC-3') is known to block restriction by R.TaqI (the sister endonuclease to M.TaqI) at 5'-TCG<u>A</u>TC-3' sites[181]. This is also expected to block M.TaqI-directed labelling, since the adenine in the 5'-TCG<u>A</u>-3' is methylated, however the reverse strand reads 5'-G<u>A</u>TCG<u>A</u>-3'. Whether this M.TaqI site is labelled is not known.

This can be tested by optical mapping using M.TaqI-directed labelling. If all 5'-TCG<u>A</u>TC-3' and 5'-G<u>A</u>TCG<u>A</u>-3' sites are blocked by dam methylation the expected reference barcode is shown in Figure 4.11A. There are only 8 such sites, giving very subtle changes in the overall intensity profile, shown by the difference in the dam-methylated and un-unmethylated reference genomes.

In Figure 4.11B the fragments from Figure 4.3 (dam-negative lambda) are aligned to the dam-methylated reference barcode. Apart from a poorer alignment weight overall (290 barcodes have an alignment weight greater than 0.7 compared to 368 previously), the mean experimental barcode can be compared to the reference and the difference calculated. When plotted with the expected difference (Figure 4.11C), the known discrepancies between the two are clear. When this analysis is repeated for dam-methylated lambda, similar results are seen (Figure 4.11D). This shows that it is unlikely

that M.TaqI-directed labelling is significantly blocked by dam methylation. This result is confirmed by results using the *de novo* alignment which is described later in this chapter and is shown in Supplementary Figure 7.13.

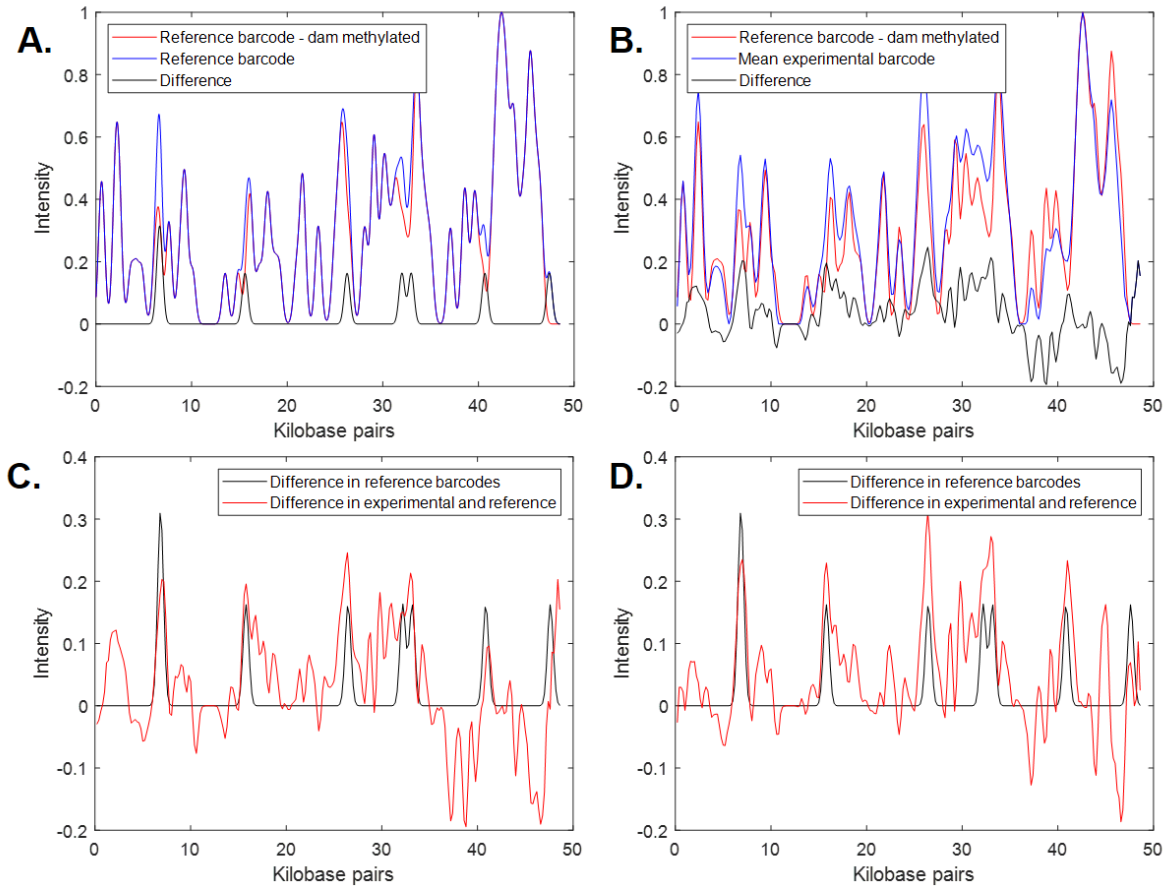


Figure 4.11 **Effect of dam methylation on M.TaqI-directed labelling. Dam methylation occurs at 5'-G<u>A</u>TC-3' sites and M.TaqI-labelling at 5'-TCG<u>A</u>-3' sites. Therefore, for dam-methylated DNA 5'-TCG<u>A</u>TC-3' sites will be unlabelled, although the effect on 5'-GATCG<u>A</u>-3' sites is not obvious. A) The reference barcode for lambda DNA if dam-methylation blocks M.TaqI-labelling (red) or not (blue) and the difference (black). B) Dam-negative DNA, labelled using M.TaqI, is aligned to the dam-blocked reference genome (red). The mean alignment is shown (red) and the difference (black). C) The difference between the expected and experimental results for dam-negative lambda DNA. Good agreement is seen, showing the expected discrepancies. D) The difference between the expected and experimental results for dam-methylated lambda DNA. Similar results are seen to C), suggesting dam-methylation does not block M.TaqI-labelling.**

Another application for this type of alignment is for alignment of a second colour of interest, for example bound fluorescently-tagged proteins (e.g. Kim *et al.*[95]) or incorporated synthetic nucleotides. Here this has been applied to lambda DNA that has been affinity-labelled, in addition to M.TaqI-labelling, to demonstrate the procedure.

Lambda DNA is labelled using M.TaqI with Atto647N as usual, however prior to combing the labelled DNA is incubated with YOYO-1 and netropsin. YOYO-1 is a DNA intercalator which will normally bind uniformly along the DNA (Supplementary Figure 7.14), however netropsin will competitively bind to AT rich regions, therefore generating a barcode which represents the underlying GC:AT ratio[59]. This is combed onto a surface and both the M.TaqI-labelling and affinity barcodes are imaged. A typical field of view is shown in Figure 4.12. For nanofluidic devices affinity-based labelling works well, however for molecular combing there are a number of issues. Ideal combing is difficult, there was a large number of overlapping molecules and very few long individual molecules could be extracted. Additionally, YOYO-1 binding is difficult to maintain during combing, perhaps because the DNA is being stretched, whilst YOYO-1 appears to disrupt the Atto647N barcode.

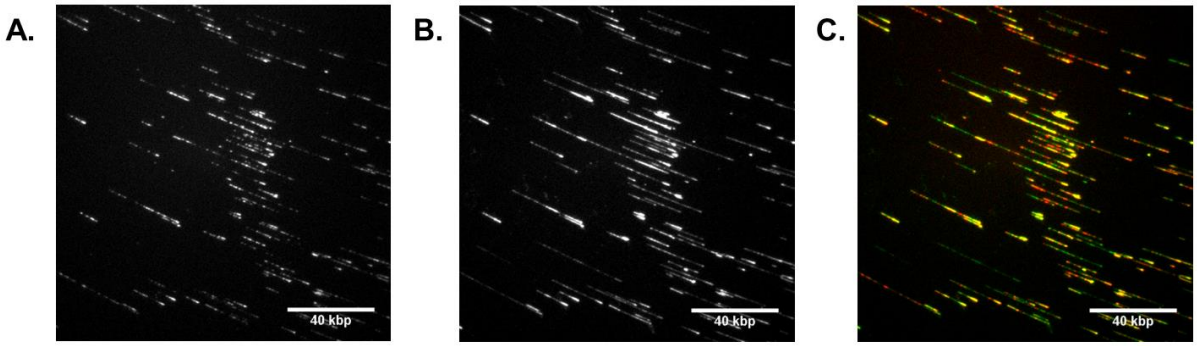


Figure 4.12 Dual colour imaging of lambda DNA labelled using YOYO-1 and M.TaqI-directed labelling. A typical region is shown for lambda DNA that is labelled using M.TaqI with Atto647N, then affinity-labelled using YOYO-1 and netropsin. A) Atto647N, B) YOYO-1, C) Overlay of Atto647N (red) and YOYO-1 (green). Note that there are a large number of overlapping DNA molecules, and the YOYO-1 and Atto647N are not as bright or uniform as is typical in other experiments.

Despite these challenges, because of the high-throughput nature of the experiment a significant number of experimental barcodes can be extracted, using the M.TaqI-barcode as normal (Figure 4.13A and B). The M.TaqI-barcode can then be used for alignment to the reference genome. Figure 4.13C shows an individual molecule, in which the experimental M.TaqI-barcode is well-aligned. The YOYO-1 intensity does not appear to have the affinity-based labelling expected, as there is no difference between GC and AT-rich regions. However, when many barcodes are aligned (Figure 4.13D) the average affinity pattern can be calculated to investigate the ensemble of individual molecules. The results for this are shown in Figure 4.13E and F, with and without background subtraction. From these results the expected affinity barcode appears to be obtained.

As well as aligning the second colour using the M.TaqI-barcode, it can be foreseen that using two colours for the alignment could improve overall reliability of alignment, although this has not been explored here. Conceptually the alignment of both barcodes should allow for a unique alignment position to be more reliably identified.

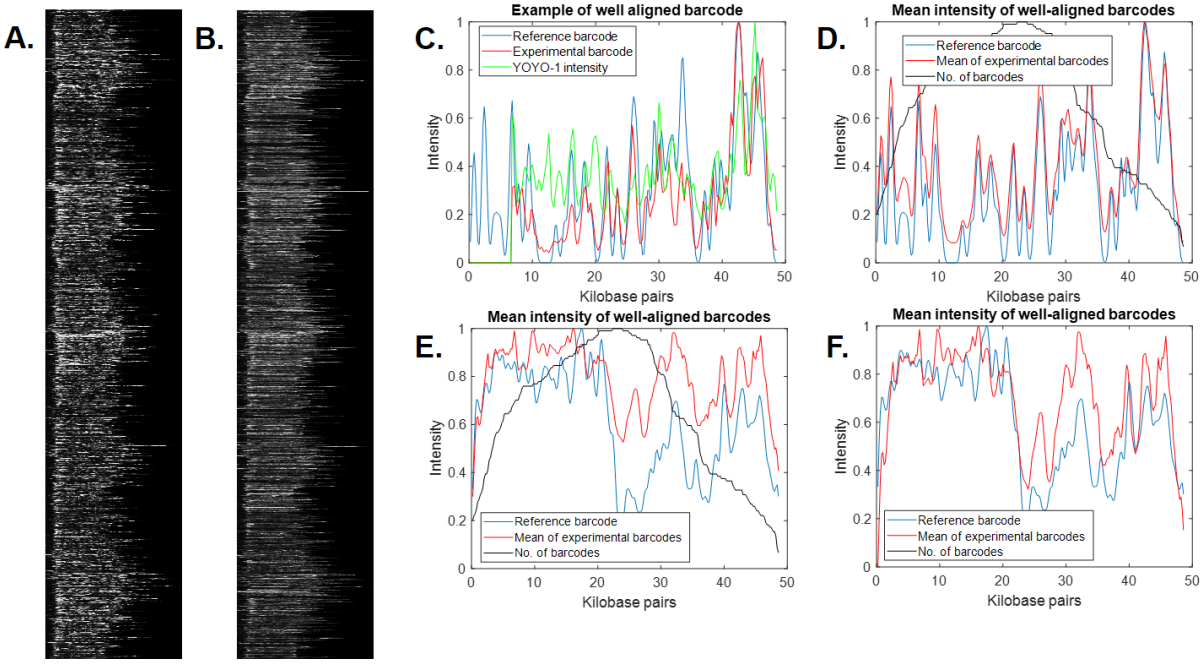


Figure 4.13 **Alignment of a second colour using M.TaqI-directed labelling. Lambda DNA is labelled using M.TaqI with Atto647N, then affinity-labelled using YOYO-1 and netropsin. A) M.TaqI-barcodes extracted as normal. B) The M.TaqI-barcode is used to extract the second colour, YOYO-1. C) An example of an individual barcode. The M.TaqI-barcode (red) is used for alignment to the reference barcode (blue), and for alignment of the affinity-barcode (green). D) Barcodes with an alignment weight greater than 0.65 are used to calculate a mean M.TaqI-barcode (red), which aligns well to the reference barcode (blue). E) The mean affinity-barcode (red) is calculated and aligns to the expected barcode (blue) – generated by CG labelling. F) The background is removed from the mean affinity-barcode (red) and aligns to the expected barcode (blue).**

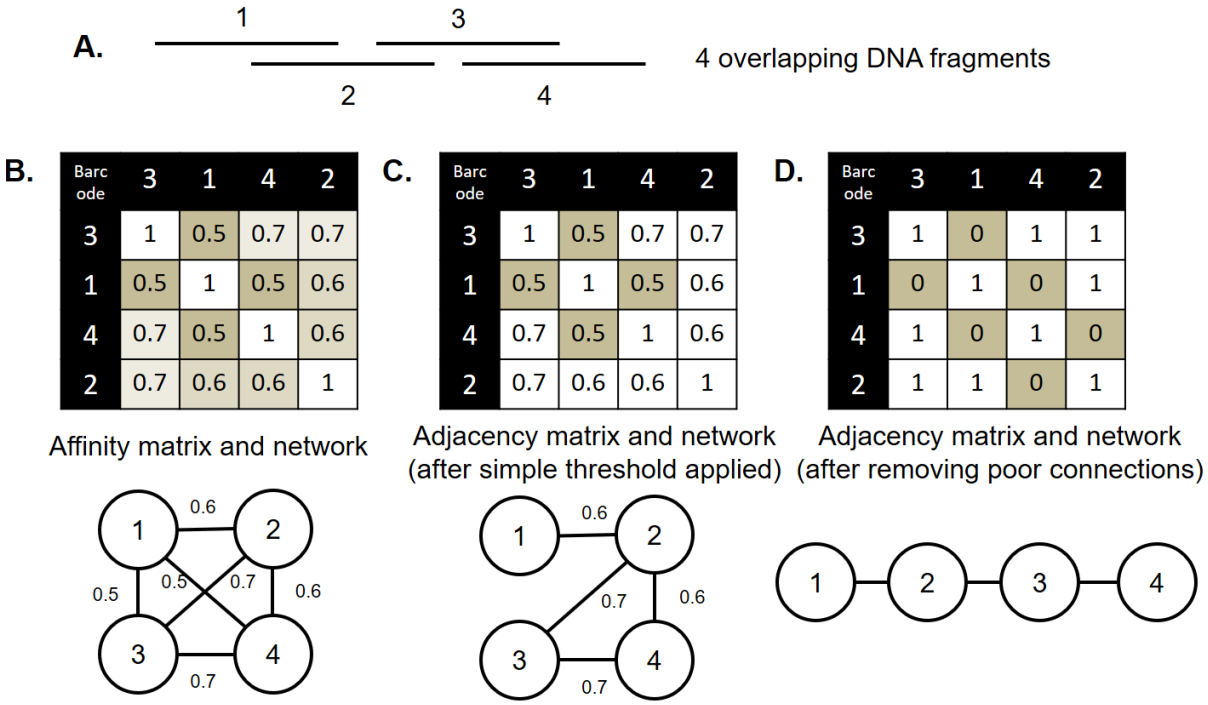
4.2.4 *De novo* separation, alignment and identification of short genomes in complex mixtures

The severe limitations with simple alignment procedures, when applied to identify complex mixtures, have been demonstrated previously. Here an alternative method to separate and then identify short (e.g. viral) genomes will be demonstrated.

The first step in this procedure (after line extraction and cleaning) is to separate the mixture. If different genomes can be separated at this step this can lead to more rapid and reliable identification, since if a consensus barcode can be derived for each genome then that can be used for identification, rather than unreliable individual barcodes. Separation will be achieved by treating barcodes as a network and detecting communities of similar DNA barcodes from within this network. When constructing the network, the 'affinity' of experimental barcodes will be used to define the edges of the network, with the barcodes themselves defining the nodes. This will be assessed by the alignment weight.

Consider for instance a simple network of four overlapping DNA barcodes (Figure 4.14A). An affinity matrix (Figure 4.14B) comparing all experimental barcodes is generated, by aligning all barcodes to all other barcodes and using alignment weight. Overlapping barcodes will, in general, have a higher alignment weight than non-overlapping barcodes. This is converted into an adjacency matrix (Figure 4.14C) using a simple threshold, for instance the two most similar edges for each node are retained above a certain alignment weight. Finally, this network can be refined (Figure 4.14D) by removing edges between nodes that don't share many connections or by adding in edges between nodes that share many connections. These networks can then be visualised as

graphs and the communities and the order of DNA barcodes (for instance from a long genome) should be generated.



Barcode	3	1	4	2
3	1	0.5	0.7	0.7
1	0.5	1	0.5	0.6
4	0.7	0.5	1	0.6
2	0.7	0.6	0.6	1

Barcode	3	1	4	2
3	1	0.5	0.7	0.7
1	0.5	1	0.5	0.6
4	0.7	0.5	1	0.6
2	0.7	0.6	0.6	1

Barcode	3	1	4	2
3	1	0	1	1
1	0	1	0	1
4	1	0	1	0
2	1	1	0	1

Figure 4.14 **Representation of the principles for generating a network from experimental DNA barcodes. A) Consider four overlapping DNA barcodes. B) An affinity matrix is calculated by calculating the similarity (alignment weight) between each barcode. C) This is converted to a simple adjacency matrix by using a threshold. D) Edges can be removed or added based on how many neighbours are shared between nodes, to produce a final adjacency matrix.**

A process known as t-Distributed Stochastic Neighbour Embedding (t-SNE)[182] is used to visualise these networks. t-SNE is a machine learning algorithm for dimensionality reduction, particularly suited to visualising high-dimensional datasets. It models each high-dimensional object as a two or three-dimensional point, in which similar objects are modelled as nearby points and dissimilar objects as distant points. This is a good method to view the results from generating the network. An example of this is shown in Figure 4.15, in which the affinity matrix is calculated by two-dimensional normalised cross correlation. Images that are similar (i.e. have a high maximum normalised cross-

correlation) are clustered together in two-dimensional space, since they will have formed a community within the network.

Figure 4.15 Example of t-Distributed Stochastic Neighbour Embedding (t-SNE) for visualisation of networks. 10 images of members of the Neely group are used to generate 100 noisy and cropped images (padded to the same size with appended zeros). Two-dimensional normalised cross correlation between every image is used to calculate an affinity matrix. t-SNE is used for visualisation of the communities within the network defined by this affinity matrix.

An example of this process applied to DNA barcodes is shown in Figure 4.16. Experimental barcodes can be generated *in silico* from the 20 bacteriophage genomes in the reference library and the affinity (i.e. alignment weight) between barcodes calculated. For 1000 barcodes this takes approximately 1 minute and will scale with the

square of the number of barcodes. This will therefore become prohibitive if large numbers of barcodes are to be aligned, but for example 1000 barcodes is sufficient to give around ten times coverage of an *E. coli* genome. This can then be used to generate an affinity matrix (Figure 4.16A) and an adjacency matrix (Figure 4.16B, refined in Figure 4.16C). The results can be visualised by using t-SNE (Figure 4.16D) and barcodes generated from the same genome are clearly clustered together.

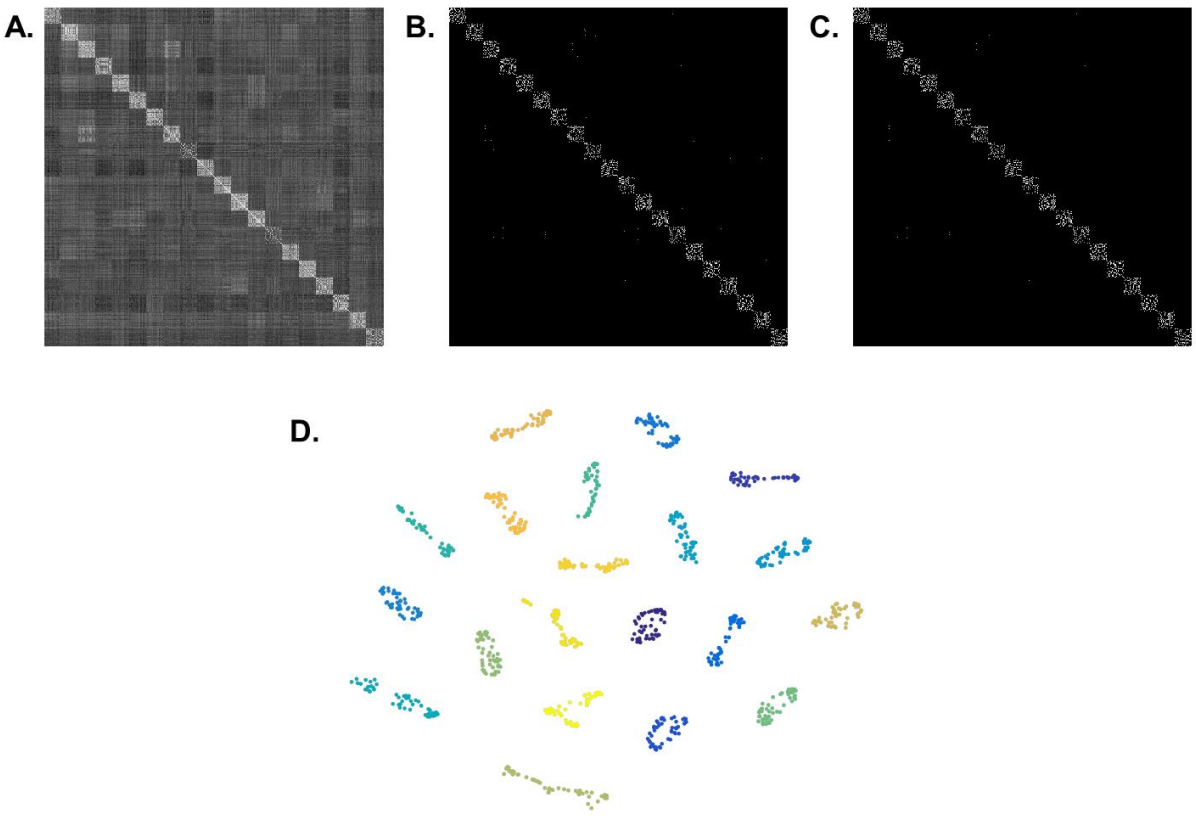


Figure 4.16 **Example of network generation and visualisation for barcodes generated *in silico*. 50 barcodes are generated for 20 phage genomes, with 100% labelling efficiency. A) Affinity matrix. The generated barcodes are ordered, and the diagonal clusters are clear. B) Adjacency matrix. An alignment weight threshold of 0.65 and 10 edges per node is used. C) Refined adjacency matrix. Edges between nodes that share fewer than 3 neighbours are removed. D) t-SNE visualisation. Each point represents a single barcode and each genome is given a different colour. Clusters for each genome are clear.**

To detect clusters of barcodes that are generated from the same genome a community detection algorithm can be used. There are many algorithms available which can detect communities within networks, based on the adjacency matrix. A fast greedy modularity maximization method (by Erwan Le Martelot)[183], which has a Matlab implementation, has been selected for speed, and good unsupervised clustering. The results of this are shown for a realistic simulation in Figure 4.17.

Between 50 and 430 barcodes were generated with 50% labelling efficiency and realistic experimental parameters for each genome (see Supplementary Table 7.1). An adjacency matrix was calculated and a t-SNE visualisation of the results is shown in Figure 4.17B. The communities that are detected are shown in Figure 4.17B and cluster validation gives a Jaccard index of 0.8599 and Rand index of 0.9927. The Jaccard and Rand indexes are measures of clustering defined as:

$$Jaccard\ index = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

$$Rand\ index = \frac{M_{11} + M_{00}}{M_{01} + M_{10} + M_{11} + M_{00}}$$

Where:

M_{11}= barcodes that are clustered in both ground truth and community detection;
M_{00}= barcodes that are not clustered in either ground truth or community detection;
M_{10}= barcodes that are clustered in ground truth but not community detection;
M_{01}= barcodes that are not clustered in ground truth but are in community detection.

The number of clusters gives a good indication of how much faster this method will be for identification of the mixture. For instance, if 1000 fragments are reduced to only 10

clusters, then only 10 barcodes (derived from the clusters) need to be identified. This will equate to a 100-fold decrease in the number of fragments that must be aligned against the reference library and a corresponding decrease in computation time.

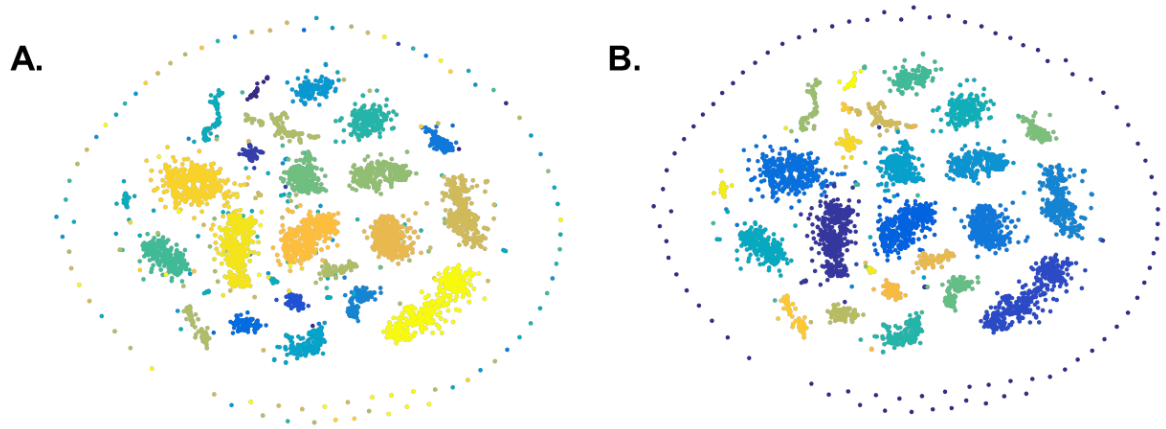


Figure 4.17 Community detection for barcodes generated *in silico*. 50 to 430 barcodes are generated for 20 phage genomes, with 50% labelling efficiency. A) t-SNE visualisation of network generated from adjacency matrix. Each colour represents a different genome. B) Community detection. Each colour represents a community that has been detected.

Once barcodes are separated and clustered, the next step is to align the clusters to produce a consensus barcode, from which the genome can be identified. A method for generation of consensus barcodes has been described previously by Reisner *et al*[53] and a hierarchical method was used by Nyberg *et al*[61] which could be applied to mixtures. The main drawback with these approaches is that barcodes are combined one at a time, which is problematic due to the presence of junk barcodes and for large datasets. By chance barcodes which are not from the original genome (i.e. the genome most of the cluster are from) can be incorporated to the consensus barcodes. As they are incorporated the consensus barcode will be significantly affected, meaning mis-incorporation of another barcode is even more likely in the next step and effectively amplifying the problem.

Here an alternative procedure is described that produces a global alignment of barcodes in one step. Barcodes are aligned to each other to give a value of the normalised cross-correlation at each displacement and for both orientations. The optimum consensus barcode will give a maximum for the overall cross-correlation between barcodes, i.e. a maximum global fit. Finding this optimum is not a trivial problem, since for even 100 barcodes there will be 198 parameters to optimise (99 displacements and 99 orientations). Therefore, to enable rapid alignment only 100 possible combinations of parameters will be tested, by fixing each barcode in turn and maximising the cross-correlation for each other barcode to the fixed barcode. The global fit can be calculated for each fixed barcode and the maximum global fit gives a reasonable consensus barcode. The whole procedure is as follows:

1. Prepare barcodes for normalised cross-correlation
    - Stretch all barcodes using same estimated stretch
    - Store both orientations (forward and reverse)
2. Normalised cross-correlation calculated
    - Calculated for every fragment aligned to every other fragment (forward and reverse)
    - Store value at every displacement and orientation
3. In turn, fix each barcode and use to align every other barcode
    - Position (and orient) every barcode based on maximum normalised cross-correlation to fixed barcode
    - Calculate global fit, i.e. sum normalised cross-correlation for all barcodes
4. Use maximum global fit to align barcodes

- Average to generate consensus barcode

This procedure can be tested on barcodes which are known to fit well to a reference barcode, for instance for experimental data for T7 (Figure 4.4). The top 36 well-aligned barcodes (Figure 4.18A) can be used and give remarkably good results. The *de novo* alignment is shown in Figure 4.18B and can be compared to the known alignment in Figure 4.18C, which shows that 32 out of 36 barcodes are well aligned. The final consensus barcode is shown in Figure 4.18D and is recognisable as the T7 reference barcode.

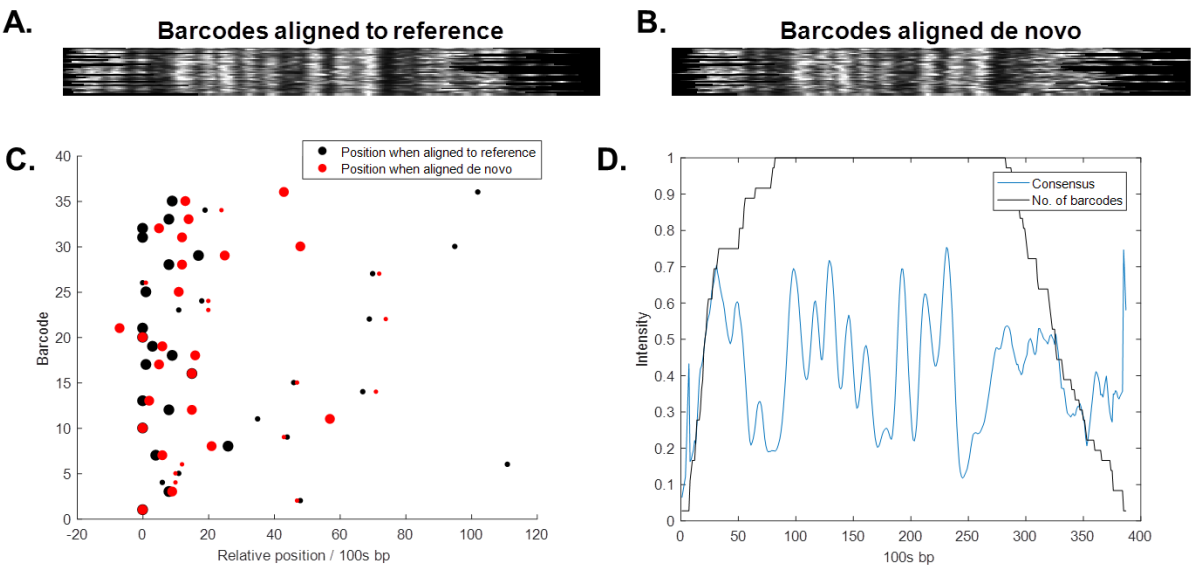


Figure 4.18 ***De novo*** **alignment results for experimental T7 data. T7 DNA was labelled using M.TaqI with Atto647N, experimental barcodes extracted and aligned to the T7 reference genome. A) 36 barcodes are chosen that aligned well (alignment weight>0.72) to the reference genome. Their alignment position is known and displayed. B)** ***De novo*** **alignment of the 36 barcodes, without a known reference. C) Known alignment (black) against** ***de novo*** **alignment (red). Relative displacement is indicated for each barcode and the orientation is indicated by the size of each point. D) Consensus barcode (blue) derived from** ***de novo*** **alignment. Number of barcodes is shown (black).**

The consensus barcode can be aligned against the reference library as normal to identify the genome (Figure 4.19). At this stage the ends are cut, background can be removed, and the consensus barcode can be discarded based on similar criteria to individual barcodes (e.g. length, brightness etc.), as well as how good the global fit is. Importantly as well as being many times quicker, the alignment weight is generally higher than for individual barcodes, making for more reliable identification. Also, if the barcode is not in the reference library, or has discrepancies (e.g. insertions, deletions, rearrangements, mutations) then this will be evident.

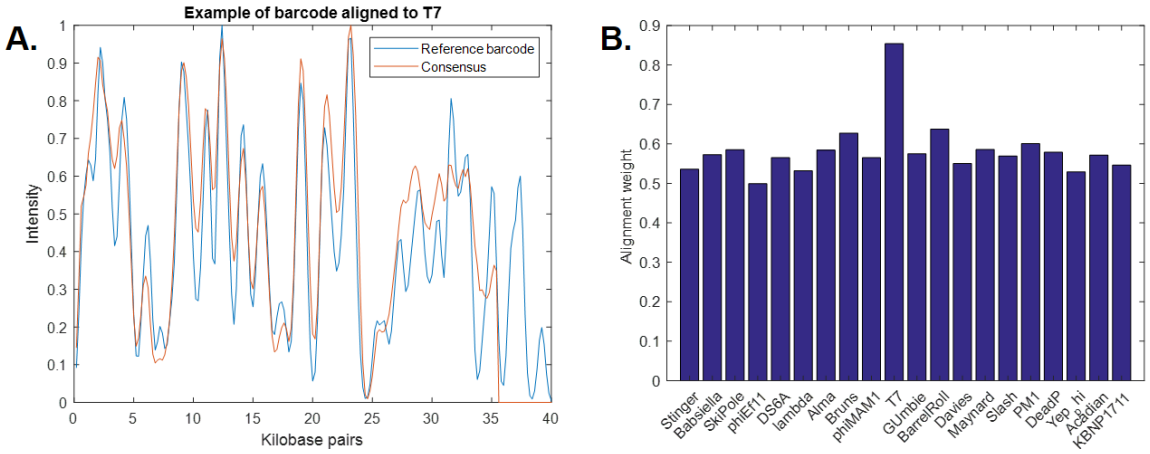


Figure 4.19 Alignment of consensus barcode to library of bacteriophages for identification. A) The ends of the consensus barcode from Figure 4.18 and background are removed. The consensus barcode (red) aligns well to the T7 reference barcode (blue). B) Consensus barcode is aligned to 20 phages and can be clearly identified by the alignment weight.

This procedure can now be applied to the complex simulated mixture of 20 viral genomes. Barcodes are generated *in silico* and clusters identified as shown in Figure 4.17. A consensus barcode for each cluster is calculated prior to identification from a large reference library. For 4800 DNA barcodes, grouped into 23 clusters and aligned against a library of 2000 phage genomes this whole process takes around 1 hour. By comparison the alignment of every barcode takes around 24 hours. The final

identification is shown in Figure 4.20, which shows how successful this procedure has been at rapidly identifying and quantifying the complex mixture.

A.

Genome	No. of barcodes aligned	No. of barcodes generated
NC_023741.1 Mycobacterium phage Stinger	45	50
NC_023697.1 Mycobacterium phage Babsiella	69	70
NC_023748.1 Mycobacterium phage SkiPole	94	90
NC_013696.1 Enterococcus phage phiEf11	107	110
NC_023744.1 Mycobacterium phage DS6A	130	130
NC_001416.1 Enterobacteria phage lambda	148	150
NC_023716.1 Mycobacterium phage Alma	162	170
NC_023687.1 Mycobacterium phage Bruns	189	190
NC_020083.1 Serratia phage phiMAM1	168	210
NC_001604.1 Enterobacteria phage T7	216	230
NC_023746.1 Mycobacterium phage GUmbie	252	250
NC_023747.1 Mycobacterium phage BarrelRoll	283	270
NC_022980.1 Brevibacillus phage Davies	284	290
NC_022768.1 Salmonella phage Maynard	290	310
NC_022774.1 Bacillus phage Slash	309	330
NC_020883.1 Bacillus phage PM1 DN	360	350
NC_023728.1 Mycobacterium phage DeadP	375	370
NC_023715.1 Yersinia phage Yep-phi	375	390
NC_023701.1 Mycobacterium phage Acadian	445	410
NC_023593.1 Escherichia phage KBNP1711	414	430

B.

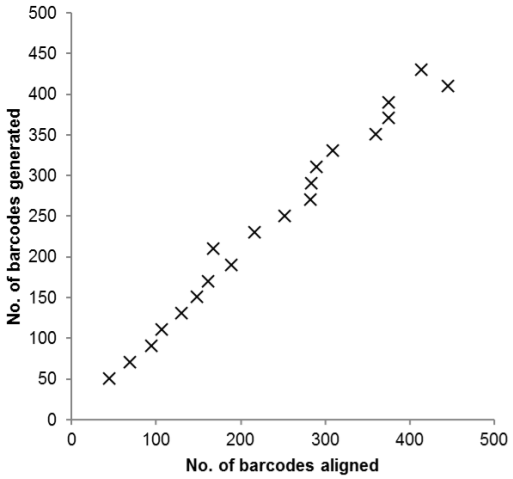


Figure 4.20 Identification and quantification of complex mixture of phages, generated *in silico*. 50 to 430 barcodes are generated for 20 phage genomes, with 50% labelling efficiency. Clustering was shown in Figure 4.17. A consensus barcode was generated for each cluster, aligned to a library of 2000 phages and assigned to the phage for which the maximum alignment weight was obtained. A) Quantification of the result. For each genome the number of barcodes generated *in silico* is shown as well as the number of barcodes contributing to consensus barcodes that were subsequently assigned. B) A plot of these results is shown, showing quantification is good.

The consensuses barcodes and alignment results can be interrogated by either the clusters or genomes to check results. The number and size of clusters is shown in Figure 4.21A. In this example all clusters are used for alignment to the reference library, however in reality clusters which have been discarded (i.e. clusters composed of junk barcodes) can be highlighted at this point. Each cluster can be investigated in more detail, for instance the alignment to the consensus barcode for cluster 8 is shown in Figure 4.21B. This consensus barcode correctly identifies BarrelRoll, but also identifies a number of other genomes which share a high sequence identity, alignments to which are

shown in Figure 4.21C-E. This illustrates the power of this technique, since even when the reference barcode is not present similar genomes would be identified and the differences can be explored.

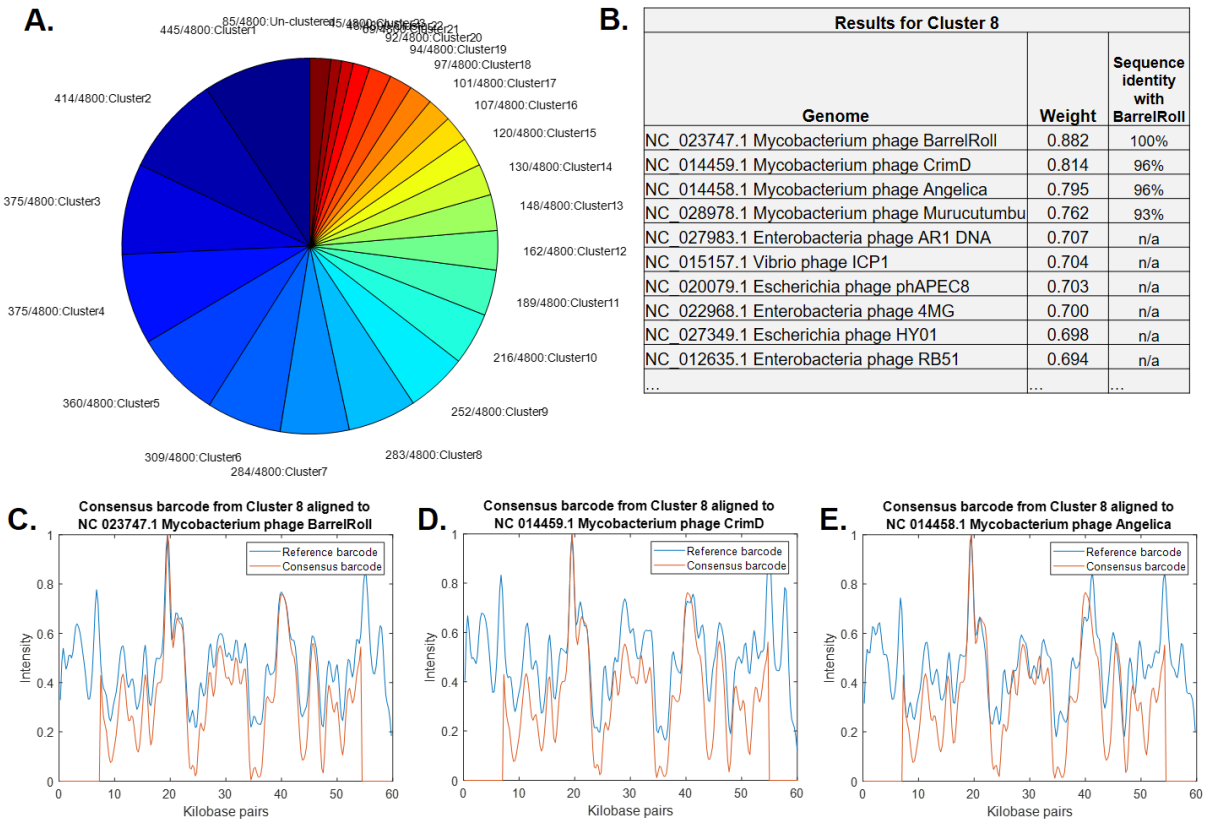
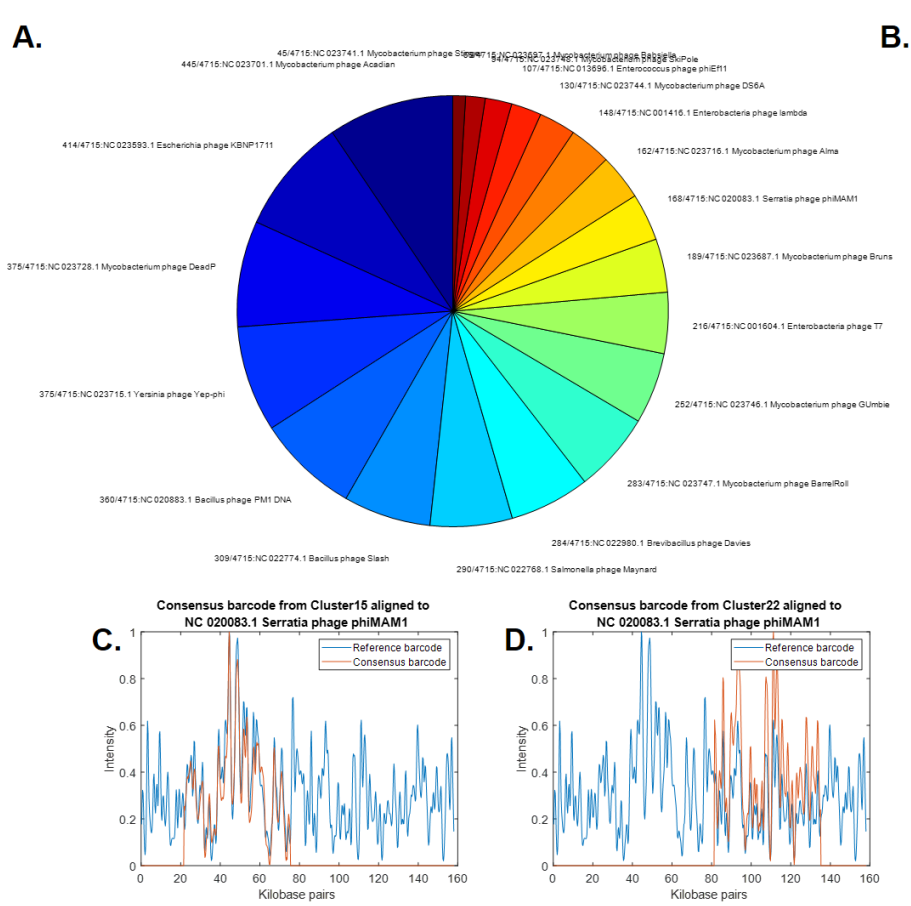


Figure 4.21 **Alignment results for *de novo* alignment of mixture generated *in silico*, by cluster. 50 to 430 barcodes were generated for 20 phage genomes, with 50% labelling efficiency. Clustering is shown in Figure 4.17 and the numbers of barcodes in each cluster is quantified in A). B) Alignment results for consensus barcode generated from cluster 8. The best match (i.e. highest alignment weight) is for the correct phage, BarrelRoll. Also identified are genomes which share significant sequence identity: CrimD, Angelica and Murucutumbu. C-E) Alignment results are shown for consensus barcode (red) to reference barcode (blue) for: C) BarrelRoll, D) CrimD, E) Angelica.**

Another way to interrogate the mixture is by genome (Figure 4.22). Each cluster is assigned to the genome for which the maximum alignment weight is obtained, and the resulting identification of the mixture is shown in Figure 4.22A. An example of the alignments to phiMAM1 is shown in Figure 4.22B. There are two clusters which have very high alignment weights, which are shown in Figure 4.22C and D. These show that two clusters were identified, for barcodes generated from either end of the genome.

This also demonstrates that for many genomes it is difficult to use this procedure to obtain *de novo* barcodes covering the whole genome. If the genome is significantly longer then the experimental barcodes, then the procedure is not sophisticated enough to obtain the full-length barcode. This may be for instance due to are regions of relatively low-density information (i.e. few peaks and troughs) or regions for which relatively few barcodes are obtained (see also Figure 4.40 and accompanying description).



Results for phiMAM1		
Cluster	No. of fragments	Weight
'Cluster15'	120	0.917
'Cluster22'	48	0.915
'Cluster19'	94	0.653
'Cluster21'	69	0.650
'Cluster 1'	445	0.635
'Cluster 9'	252	0.633
'Cluster 5'	360	0.628
'Cluster16'	107	0.626
'Cluster 4'	375	0.618
'Cluster11'	189	0.615
'Cluster 2'	414	0.614
'Cluster17'	101	0.613
'Cluster 6'	309	0.612
'Cluster 7'	284	0.610
'Cluster10'	216	0.599
'Cluster12'	162	0.597
'Cluster 8'	283	0.594
'Cluster14'	130	0.590
'Cluster13'	148	0.590
'Cluster23'	45	0.588
'Cluster 3'	375	0.582
'Cluster20'	92	0.581

Figure 4.22 Alignment results for *de novo* alignment of mixture generated *in silico*, by genome. 50 to 430 barcodes were generated for 20 phage genomes, with 50% labelling efficiency. Clustering is shown in Figure 4.17. Consensus barcodes are generated for each cluster and aligned to a library of reference genomes. A) They are assigned to the genome which has the greatest alignment weight allowing for quantified identification of the mixture. B) Example results for phiMAM1 genome. Cluster 15 and cluster 22 align well to the genome. C-D) Alignment to phiMAM1 reference barcode (blue) of consensus barcode (red) generated from C) Cluster 15, D) Cluster 22.

This procedure can be applied to experimental mixed data. Clustering and t-SNE results for a known mixture of T7/lambda (see Figure 4.5) are shown in Figure 4.22. This takes around 2 minutes for 1500 fragments. Here we can note that the different genomes are separated, but not as well as in simulations (e.g. Figure 4.17), seemingly producing a rather poor clustering result overall. This is likely due to the junk barcodes that have been discussed previously.

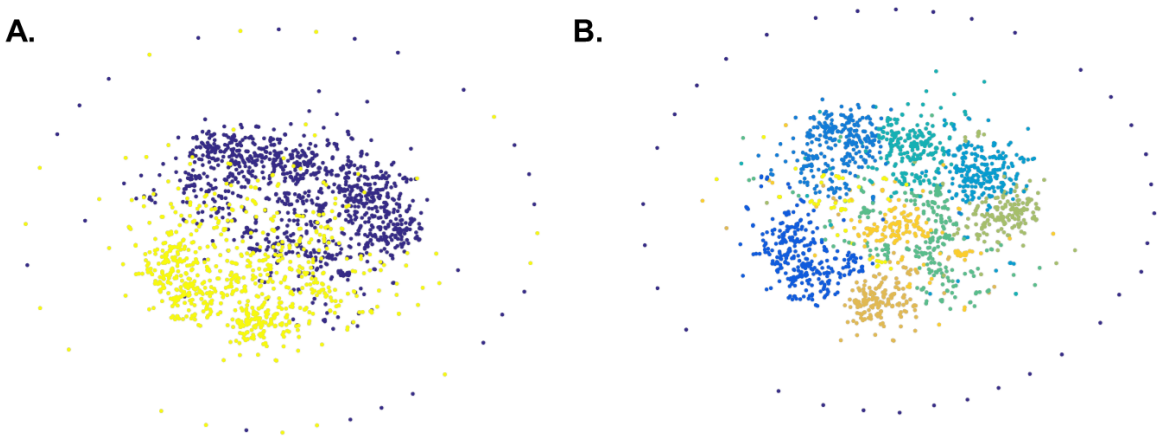
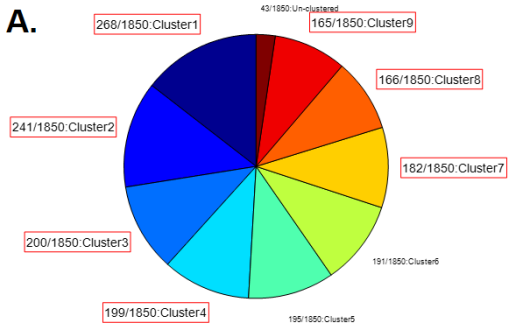
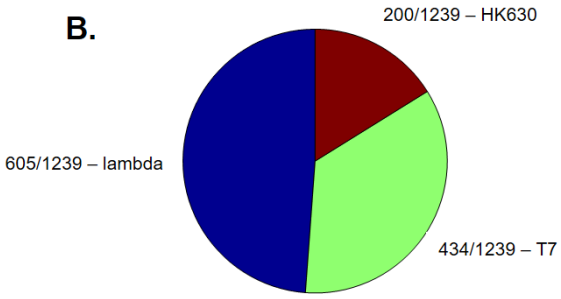


Figure 4.23 Community detection for known mixture of lambda and T7 barcodes. DNA was with Atto647N, combed and imaged. A) t-SNE visualisation of network generated from adjacency matrix. Barcodes are from lambda (blue) and T7 (yellow) samples. B) Community detection. Each colour represents a community that has been detected.

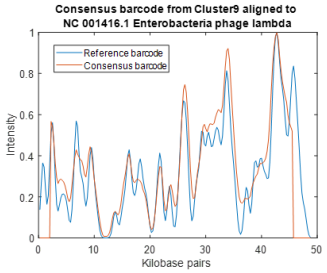
Despite this, consensus barcodes can be derived for each cluster. Clusters are discarded which have poor *de novo* alignments or by the criteria used previously (Figure 4.24A). When the cleaned clusters are aligned to the library of 2000 phage genomes the results in Figure 4.24B are obtained. T7 and lambda are clearly identified within the sample and further interrogation can be used to show the alignment of consensus barcodes (Figure 4.24C and D). Only clusters 5, 8 and 9 do not align well to either genome, therefore are likely constructed from junk barcodes, however these can be discarded for identification, based on poor global alignment, length and intensity.

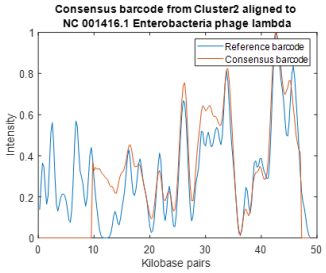




C.

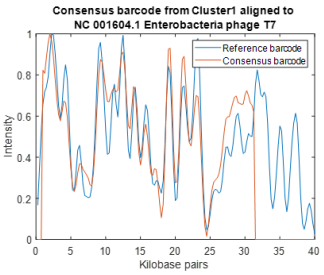
Results for lambda		
Cluster	No. of barcodes	Weight
'Cluster9'	165	0.849
'Cluster2'	241	0.825
'Cluster3'	200	0.823
'Cluster4'	199	0.814
'Cluster7'	182	0.758
'Cluster5'	195	0.586
'Cluster6'	191	0.582
'Cluster1'	268	0.556
'Cluster8'	166	0.542





D.

Results for T7		
Cluster	No. of barcodes	Weight
'Cluster1'	268	0.831
'Cluster8'	166	0.790
'Cluster7'	182	0.563
'Cluster9'	165	0.529
'Cluster4'	199	0.529
'Cluster2'	241	0.518
'Cluster6'	191	0.507
'Cluster3'	200	0.504
'Cluster5'	195	0.474



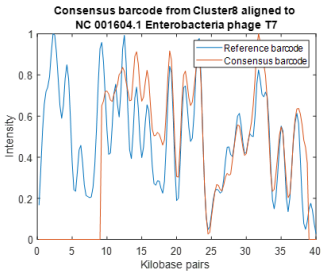


Figure 4.24 **Alignment results for lambda/T7 mixture using *de novo* alignment and assignment of consensus barcodes. A) Experimental barcodes (Figure 4.6A) are separated into 9 clusters and aligned *de novo* to generate consensus barcodes. 7 consensus barcodes are used for assignment of the mixture (highlighted with red boxes), based on global alignment, length and intensity. B) Assignment of consensus barcodes to phage genomes. Consensus barcodes are aligned to a library of 2000 phages and assigned to the genome with the greatest alignment weight. C) Results for lambda reference. Four consensus barcodes align very well to the genome, and the top two alignments are displayed. D) Results for T7 reference. Two consensus barcodes align very well to the genome and are displayed.**

These results can be compared to the previous alignment procedure (Figure 4.8). 920 out of 1850 barcodes are identified as lambda and 462 as T7, compared to 328 and 188, respectively, previously. The previous procedure also identifies many other genomes during the analysis whilst using *de novo* alignment identification is more reliable. These results are for a library that is 100 times larger but only takes twice as long to compute (16 minutes for 2000 genomes against 8 minutes for 20 genomes).

The only issue with the results is that it appears that one cluster has been misidentified as HK630. This is investigated further in Figure 4.25, which shows the consensus barcode also aligns well to lambda. HK630 (and HK629) share a large amount of sequence identity with lambda, meaning the reference barcodes are very similar, hence the apparent misidentification.

Results for cluster 3		
Genome	Weight	Sequence identity with lambda
NC_019723.1 Enterobacteria phage HK630	0.853	99%
NC_019711.1 Enterobacteria phage HK629	0.823	99%
NC_001416.1 Enterobacteria phage lambda	0.823	100%
NC_015274.1 Streptococcus phage Dp-1	0.723	n/a
…	…	…

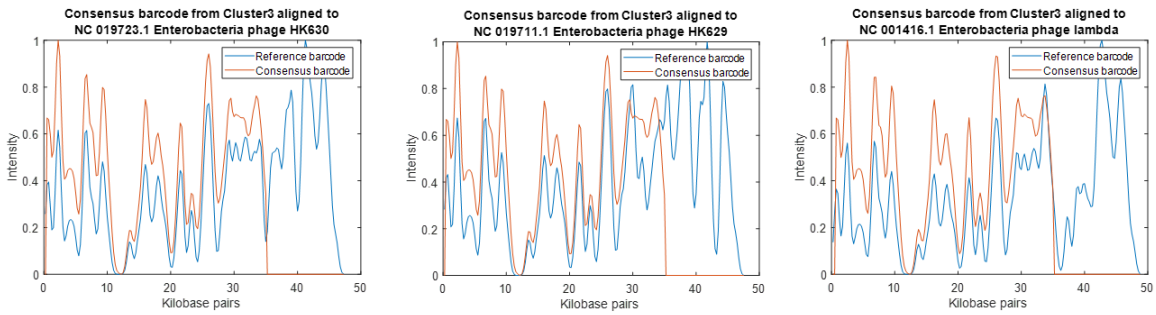


Figure 4.25 Alignment results for consensus barcode from cluster 3 of lambda/T7 mixture of Figure 4.24. Cluster is assigned to HK630, which shares 99% sequence identity with lambda, as does HK629. The alignments are also displayed.

Results for the pure samples of lambda and T7 (Figure 4.3 and Figure 4.4), and a mixed sample (Figure 4.6) are shown in Supplementary Figure 7.15, Figure 7.16 and Figure 7.17. A summary and comparison of the alignment procedures for these samples is shown in Table 4.1. The improved alignment procedure is both more rapid for assignment against a large reference library and more reliable for identification.

		Alignment of each experimental barcode to every genome in reference library (20 phages)				**Separation, d*e novo* alignment and assignment of consensus barcode to reference library (2000 phages)**			
Sample	**No. of barcodes after filtering**	**No. of barcodes assigned (weight>0.7) to:**			**Comput-ation time**	**No. of barcodes assigned to:**			**Comput-ation time**
		lambda	T7	Other genomes		lambda	T7	Other genomes	
Lambda	1077	368 (34%)	0 (0%)	143 (13%)	260 s	751* (70%)	0 (0%)	0** (0%)	585 s
T7	1166	0 (0%)	174 (15%)	453 (39%)	276 s	0 (0%)	394 (34%)	0** (0%)	645 s
Lambda/T7	1756	136 (8%)	161 (9%)	892 (51%)	420 s	329 (19%)	582 (33%)	0** (0%)	712 s

*181 barcodes assigned to HK630 and 45 barcodes to HK629, which both share 99% sequence identity with lambda

**Consensus barcodes are discarded based on global alignment of experimental barcodes, length and intensity.

Table 4.1 Summary and comparison of alignment procedure for samples of lambda and T7.

4.2.5 Identification of Adenovirus A

To illustrate these procedures on a more realistic sample they were applied to identify a sample of Adenovirus A. DNA was extracted from Adenovirus A-infected human cells 72 hours post-infection. Gel electrophoresis of DNA extracted pre-infection and after 48 hours showed that most of the long fragments were viral DNA (Supplementary Figure 7.18). DNA was labelled with Atto647N, combed and imaged as normal.

Around 2000 barcodes were extracted from the sample and were aligned to a reference library of 128 vertebrate viruses, in turn, which took approximately one hour. The results are shown in Figure 4.26A. 276 out of 1986 barcodes (14%) are assigned to Human adenovirus, however a very large number of barcodes are assigned to other viruses, so the sample cannot be reliably identified. For instance, 82 (4%) of barcodes are assigned to Fowl adenovirus A. Examples of experimental barcodes are shown in Figure 4.26B-D.

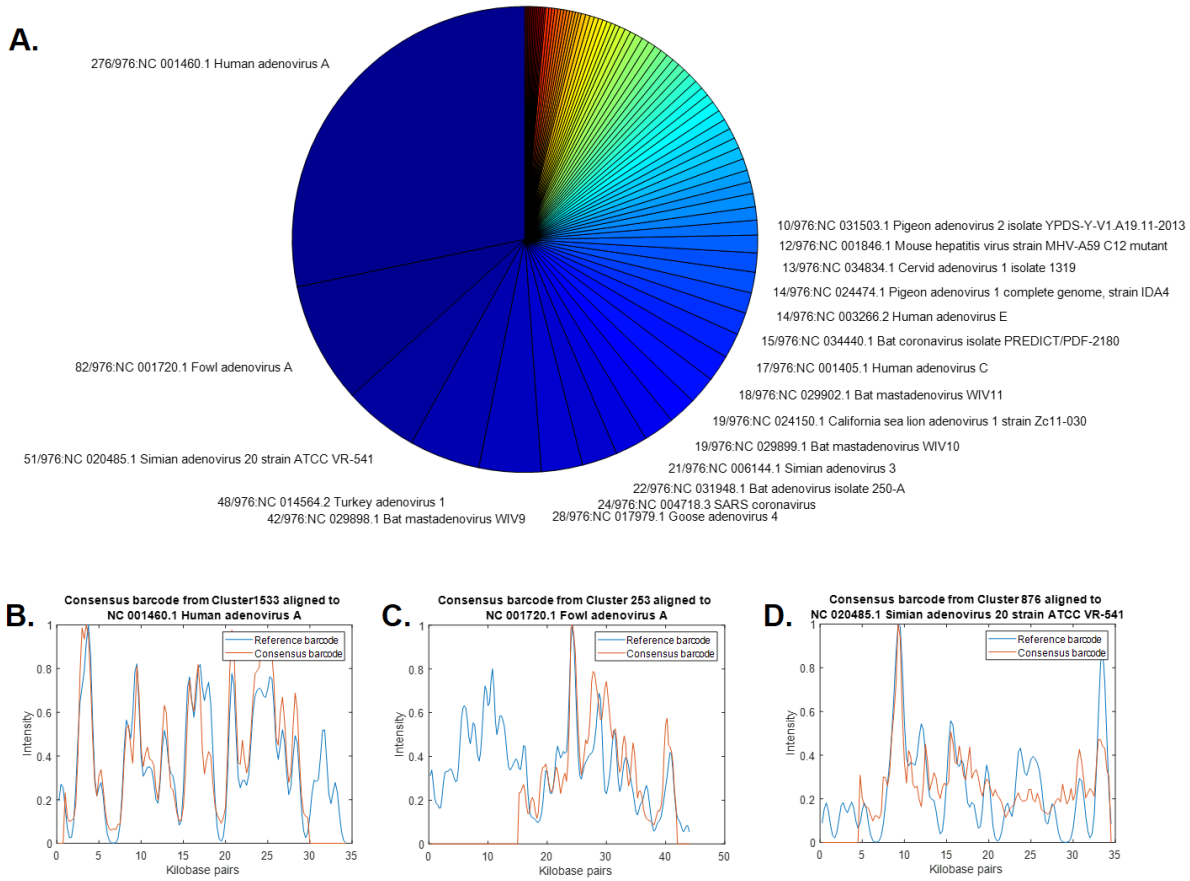


Figure 4.26 Assignment of Adenovirus A DNA sample by alignment of each barcode to each reference genome. 1986 experimental barcodes were extracted and aligned to 128 reference barcodes from a library of vertebrate viruses. A) Barcodes are assigned to the genome with the maximum alignment weight (with a threshold of 0.7). 276 of 1986 barcodes are assigned to Adenovirus A. B-D) Examples of alignments of experimental barcodes to: B) Human adenovirus A; C) Fowl adenovirus A; D) Simian adenovirus 20.

By using clustering and *de novo* alignment to produce consensus barcodes, more reliable identification of the sample is possible. Results are obtained in 10 minutes and are shown in Figure 4.27. 669 (34%) barcodes are assigned to Human Adenovirus A and consensus barcodes show very high similarity to the reference barcode (Figure 4.27B and C). No other references are assigned as all other consensus barcodes are discarded based on poor global alignment, length or intensity.

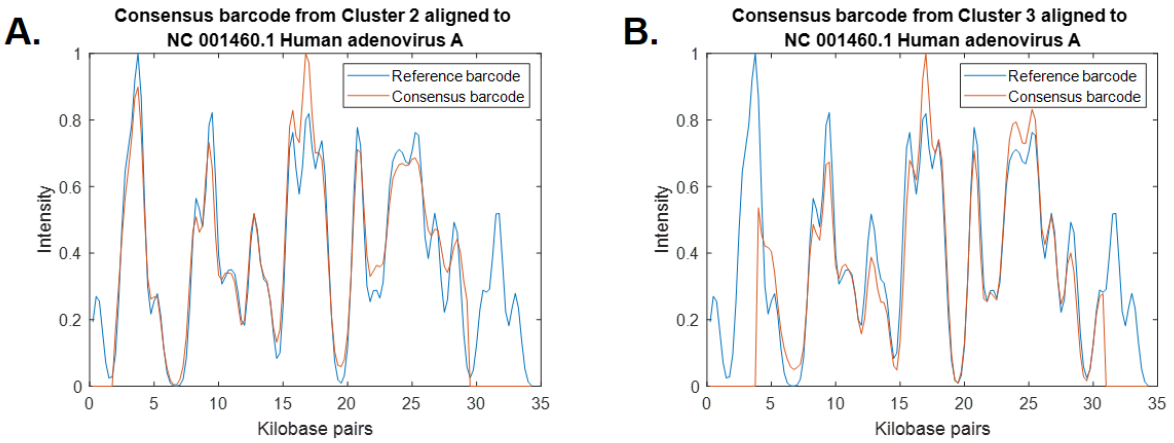


Figure 4.27 Assignment of Adenovirus A DNA sample by separation, *de novo* alignment and assignment of consensus barcode to reference library. 1986 experimental barcodes were extracted and produced 21 consensus barcodes for alignment to 128 reference barcodes from a library of vertebrate viruses. 669 of 1986 barcodes are assigned to Adenovirus A. A-B) Examples of alignment of consensus barcodes.

4.2.6 Separation and identification of viral DNA for complex genomic mixtures

The results for Adenovirus A highlight another aspect of *de novo* separation and alignment of barcodes. As well as junk DNA there will be human genomic DNA present in the sample. In other words, the sample is not pure, but a mixture of viral and genomic DNA. By using separation and clustering it is possible to identify the viral DNA if the copy number is sufficiently high, therefore it is possible that this procedure may also be extended to identify resistance plasmids.

To test this a simple model system was used: DNA was extracted from *E. coli*, with Atto647N as usual and mixed in a ratio of 4:1 with labelled bacteriophage DNA, for example lambda DNA. The bacterial genome is around 100 times longer than the viral DNA, so this is equivalent to a copy number of around 20. More barcodes must be imaged to enable reliable identification (5000 compared to around 1000 before), which will make rapid alignment more important.

A 4:1 mixture of genomic DNA from *E. coli* strain C2566 (henceforth referred to as NEB 'T7 Express') and lambda DNA was imaged, and 5114 barcodes extracted and aligned to the lambda reference barcode. Results are shown in Figure 4.28A and B; 195 (4%) barcodes are assigned to the lambda reference genome (alignment weight greater than 0.7), compared to 368 out of 1077 (34%) in the pure sample.

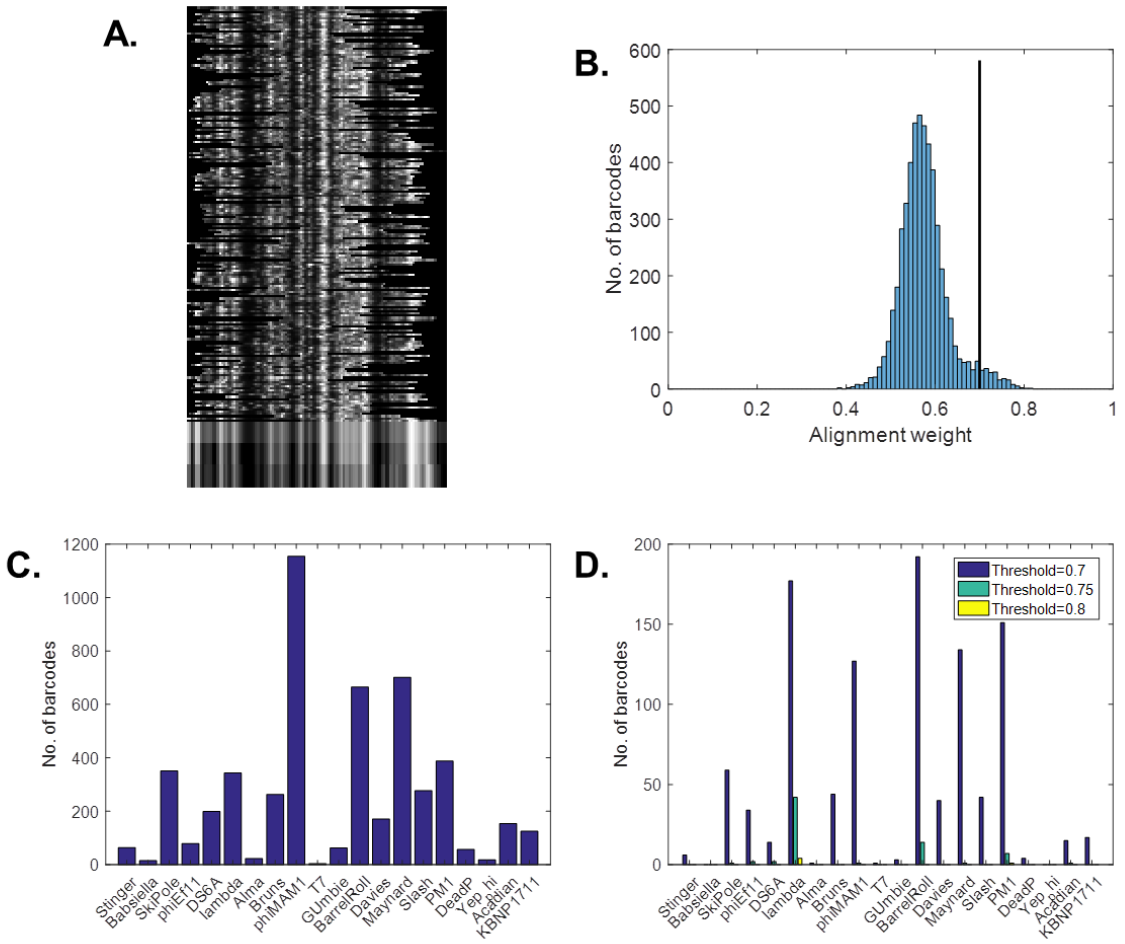


Figure 4.28 **Identification of bacteriophage DNA in genomic mixture by alignment of experimental barcodes to reference barcodes. DNA is with Atto647N and a 4:1 mixture of *E. coli*:lambda is combed and imaged. 5114 barcodes are extracted from the images. A) 195 (4%) of barcodes aligned with weight>0.7 to lambda reference barcode. B) Alignment weight of all experimental barcodes to lambda reference. C-D) Identification of barcodes against a library of 20 phage genomes. C) Each barcode is assigned to the genome for which the largest alignment weight was obtained. D) The number of barcodes assigned to each genome with an alignment weight greater than a threshold (0.7, 0.75 and 0.8).**

If the sample is known, then short genomes can be identified and relatively easily extracted by simple alignment like this. However, it is very difficult to identify a complex sample if the DNA is not known. To obtain suitable coverage, around five times the number of fragments are aligned to the reference library, slowing the procedure by the same amount. In addition, there is a much greater chance that barcodes will be assigned incorrectly, since as well as junk barcodes there are barcodes derived from the genomic DNA, which may, by chance, align well to a reference barcode. Results are shown in Figure 4.28C, for assignment of each barcode to the genome with the maximum alignment weight, and in Figure 4.28D for the assignment of barcodes to genomes based on thresholds of alignment weight. Lambda can be identified in the sample, but not reliably.

In contrast when *de novo* separation and alignment of the sample is carried out lambda is reliably identified (Figure 4.29). The adjacency matrix and clustering is visualised by t-SNE in Figure 4.29A and B. Barcodes that fit well to lambda form a clear cluster which is identified during community detection. When consensus barcodes are derived for each cluster, only two are selected (based on global alignment, length and intensity), which both align well to lambda (Figure 4.29C-E). When compared to the procedure in Figure 4.28 it is clear that *de novo* separation and alignment can be used to give more reliable results. Previously it was difficult to identify lambda against the background of junk and genomic barcodes, however separation and generation of a consensus barcode has made identification trivial.

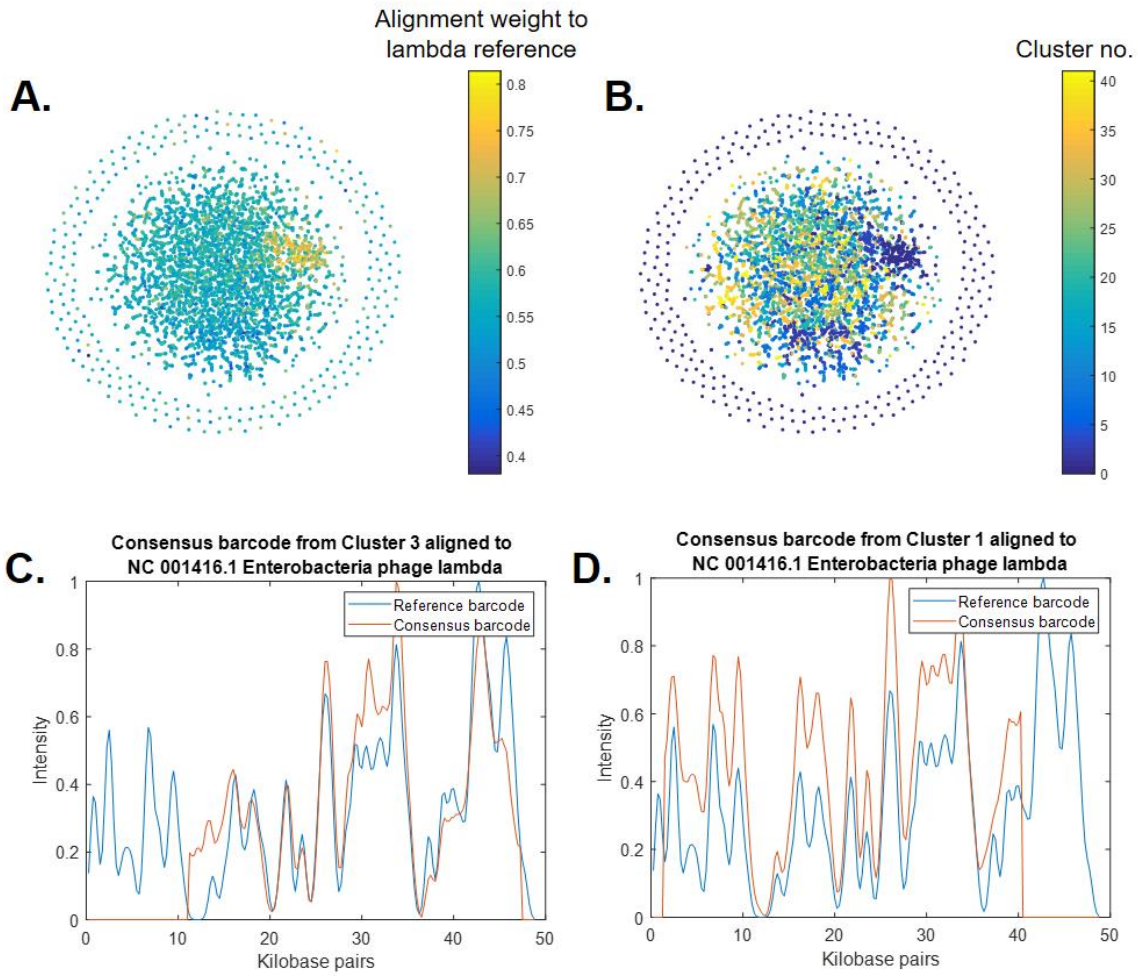


Figure 4.29 **Identification of bacteriophage DNA in genomic mixture by *de novo* separation and alignment of experimental barcodes. A) t-SNE visualisation of network generated from adjacency matrix. Colour is given by alignment weight to lambda reference genome. B) Community detection. Each colour represents a community that has been detected. C-D) Examples of alignment of consensus barcodes. Barcodes are assigned to the genome with the maximum alignment weight. 475 of 5114 barcodes are assigned to Adenovirus A.**

In addition, results have also been obtained for a 4:1 mixture of *E. coli* and T7 (Supplementary Figure 7.19 and Figure 7.20) and for a 4:1:1 mixture of *E. coli*, lambda and T7 (Supplementary Figure 7.21 and Figure 7.22). The results are summarised in Table 4.2, once again highlighting the improvements when using *de novo* alignment, in particular note the number of other genomes to which barcodes are assigned, which make reliable identification problematic.

		Alignment of each experimental barcode to every genome in reference library (20 phages)				**Separation, d*e novo* alignment and assignment of consensus barcode to reference library (2000 phages)**			
Sample	**No. of barcodes after filtering**	**No. of barcodes assigned (weight>0.7) to:**			**Comput-ation time**	**No. of barcodes assigned to:**			**Comput-ation time**
		lambda	T7	Other genomes		lambda	T7	Other genomes	
E. coli: lambda (4:1)	5114	195 (4%)	1 (0%)	883 (17%)	21 mins	475 (9%)	0 (0%)	0* (0%)	47 mins
E. coli:T7 (4:1)	8124	10 (0%)	229 (3%)	4732 (58%)	33 mins	0 (0%)	372 (5%)	0* (0%)	76 mins
E. coli :lambda:T7 (4:1:1)	7963	554 (7%)	208 (3%)	1625 (20%)	32 mins	1306 (16%)	407 (5%)	0* (0%)	90 mins

*Consensus barcodes are discarded based on global alignment of experimental barcodes, length and intensity.

Table 4.2 Summary and comparison of alignment procedures for samples of *E. coli* and lambda and/or T7.

4.2.7 Separation and identification of resistance plasmids in complex genomic mixtures

Given the reliable results for experimental mixtures of viral DNA and *E. coli* genomic DNA it should be possible to extend these approaches to mixtures of resistance plasmids and genomic DNA. The isolation and purification of large, low copy number, plasmids (such as resistance plasmids) is generally time-consuming and difficult, so if it can be avoided then it would aid with rapid diagnosis. Several reference plasmid/bacteria systems will be used experimentally here to test this separation and identification: pCT (94 kbp) in *E. coli* strain EC958 (5,249 kbp); pNDM (89 kbp) in *E. coli* strain DH10B (4,686 kbp); and pKpQIL (114 kbp) in *K. pneumoniae* strain Ecl8 (5,325 kbp).

There are several important differences when considering these samples. The copy number of the resistance plasmids is likely to be very low, perhaps only 1-5 copies per cell, compared to around 20 for the artificial lambda/T7 mixtures. These plasmids will also be circular, rather than the linear viral DNA that has been used to this point. These differences have been investigated *in silico* and an example for pCT/*E. coli* is shown in Figure 4.30.

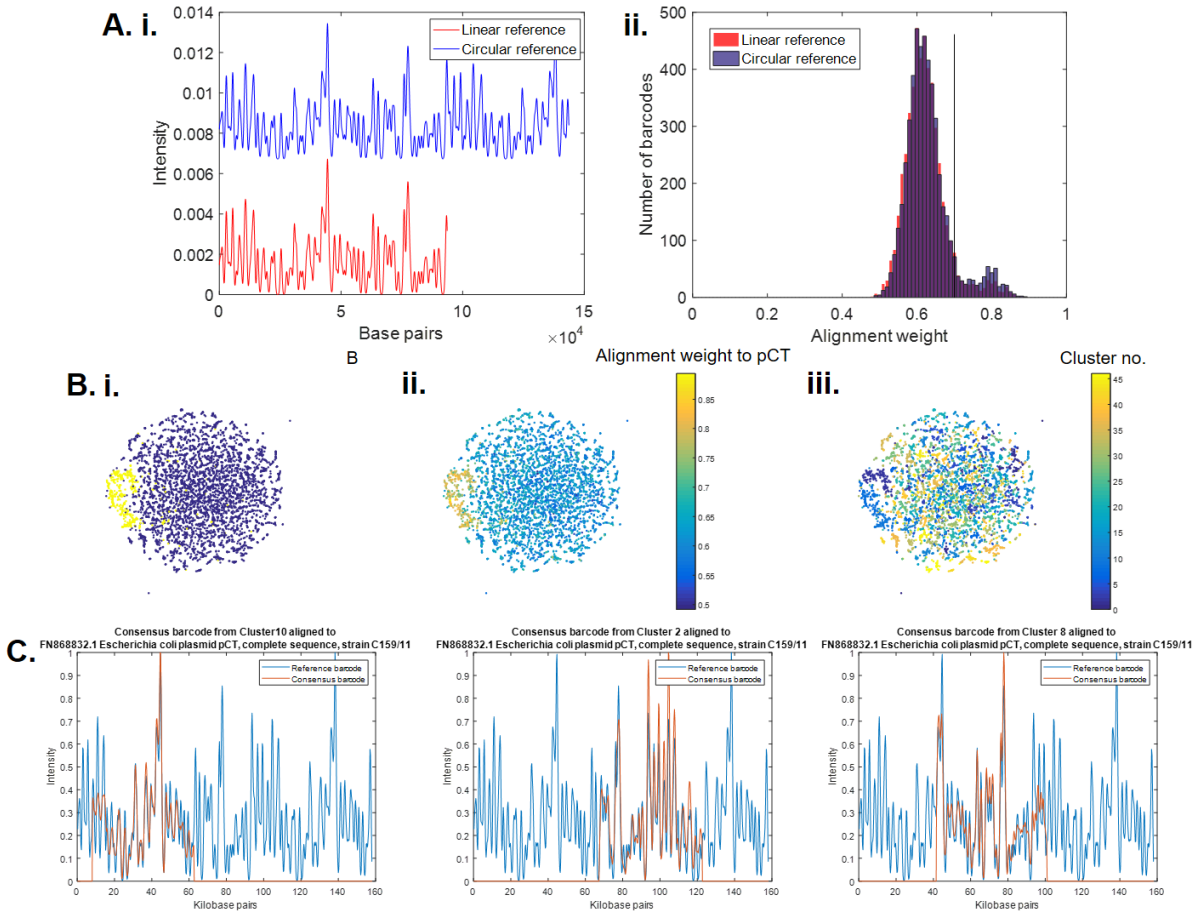


Figure 4.30 **Identification of resistance plasmids by alignment of DNA barcodes, copy number=5. 5000 barcodes were generated *in silico* for a pCT:*E. coli* mixture, in a ratio of 11:1 (copy number of 5). A) Simple alignment of barcodes to pCT reference barcode. i) pCT is circular, so a linear reference barcode (red) is unsuitable for alignment. A region of the barcode, of the length of the largest experimental barcode, is appended to the end to simulate a circular reference (blue). ii) The alignment weight of all fragments is shown and shows that simulating a circular reference improves the alignment results. B) *De novo* separation of barcodes. t-SNE is used to visualise the network generated from the adjacency matrix. i) Barcodes generated from pCT (yellow) and *E. coli* (blue) reference genomes. ii) Weight of alignment to pCT reference genome in A). iii) Community detection. C) *De novo* alignment of barcodes to generate consensus barcodes (red). The three clusters are shown that align well (weight>0.85) to the pCT reference genome (blue).**

5000 barcodes were generated *in silico,* from the reference genomes for *E. coli* strain and pCT, in a ratio which reflects a copy number of 5 (~11:1, *E. coli*:pCT). These were generated with typical experimental parameter values, for example a labelling efficiency of 50%. From the simple alignment of all barcodes to the pCT reference genome 377 barcodes align with a weight greater than 0.7 when the circularity of plasmids is not considered (Figure 4.30A). However, this can be simply altered by appending a copy of the reference genome, of the length of the maximum barcode, to the end of the linear reference. Now 503 barcodes align with a weight greater than 0.7 (Figure 4.30A). In experimental samples, it is likely that the natural shearing of DNA during simple DNA extraction procedures is sufficient to linearise most plasmid molecules. An example restriction assay for an experimental sample is shown in Supplementary Figure 7.5 and shows that the majority of fragments are 30-50 kbp in length, despite no explicit step to shear genomic DNA.

De novo separation and alignment of barcodes can also be carried out (Figure 4.30B-C). There is a clear cluster formed by barcodes generated from the pCT reference genome, which can be separated automatically into several clusters. The consensus barcodes from these clusters aligns well to the reference genome and can be used to reliably identify the resistance plasmid in the sample.

Despite these results the identification of resistance plasmids in experimental samples was unsuccessful (data not shown). This could be for a number of experimental reasons, however perhaps the most likely reason is that the copy number is very low, perhaps only 1-2 per cell. This may have been reduced further during purification, since a commercial genomic DNA extraction kit was used, which may have caused the loss of

most plasmids from the sample or if plasmids have not sheared they will not bind to the surface during molecular combing. A copy number of one plasmid per cell (~55:1 *E. coli*:pCT) is simulated in Figure 4.31 to test this; 90 barcodes out of 5000 are generated from pCT. When compared to a copy number of 5 in Figure 4.30 the results are striking and similar to the experimental results.

Few molecules are reliably identified by alignment weight (Figure 4.31A), but perhaps more importantly the *de novo* separation and clustering of barcodes fails to produce any significant cluster of barcodes generated from the pCT reference (Figure 4.31B). This means there is no possibility to generate a consensus barcode that can be reliably identified. The top three consensus barcode alignments to the pCT reference are shown in Figure 4.31C. These results demonstrate the propensity of cross-correlation to align barcodes to bright regions of the reference genome. This leads to 211 barcodes aligned to the reference genome with a weight greater than 0.7, of which 123 were generated from the *E. coli* reference and only 88 from the pCT reference. The alignment to the bright region is also clear in Figure 4.31Aii.

These results demonstrate the challenges for alignment of barcodes, particularly in complex mixtures. Good experimental data and a relatively high copy number are necessary to reliably identify small DNA fragments from a mixture. To ensure good experimental data both optimisation of the methyltransferase-directed labelling (as in CHAPTER 2) and careful imaging of individual labelled DNA molecules are required. In particular, the molecular combing employed in this research appears to introduce many junk molecules that make identification difficult. Therefore, further development of

imaging techniques (including nanochannels) would improve the reliability of identification.

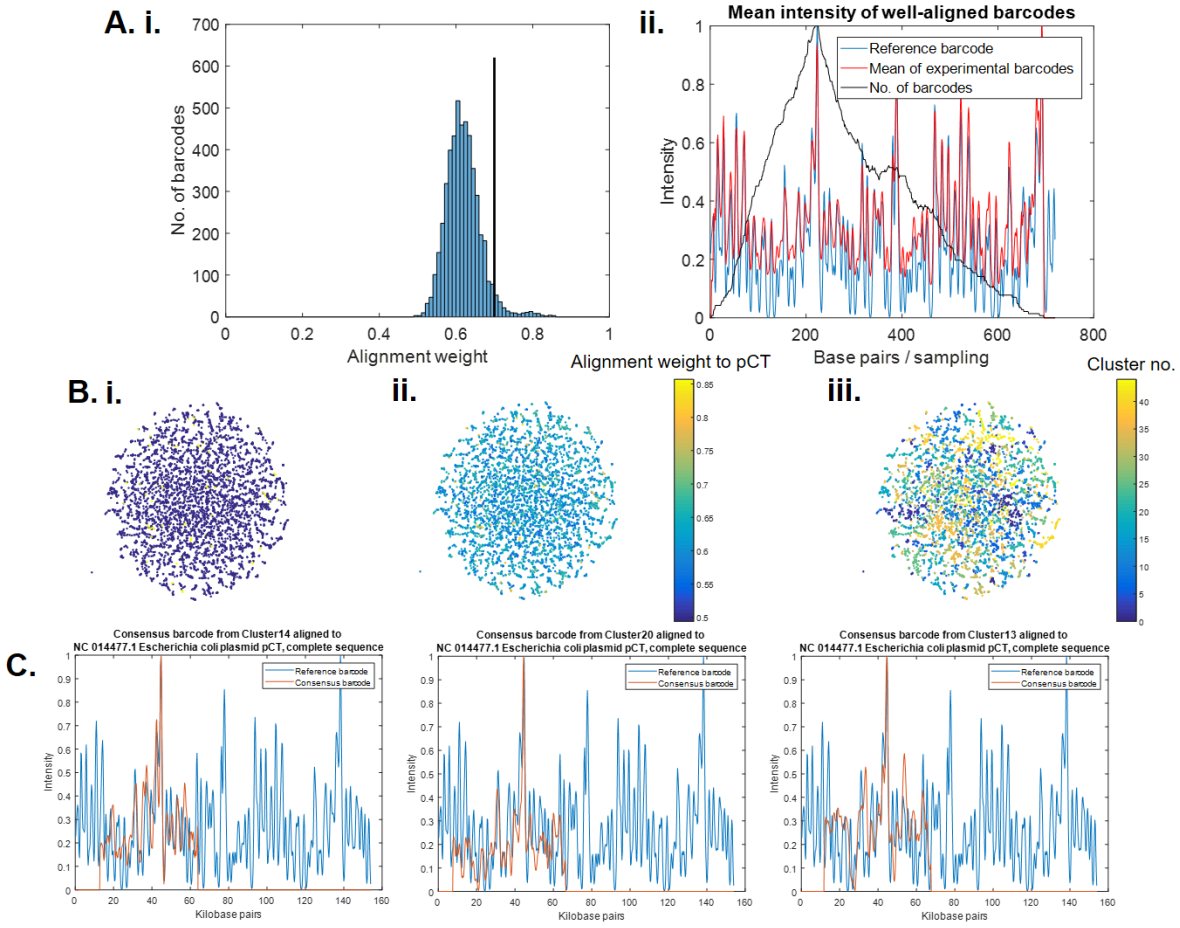


Figure 4.31 **Identification of resistance plasmids by alignment of DNA barcodes, copy number=1. 5000 barcodes were generated *in silico* for a pCT:*E. coli* mixture, in a ratio of 55:1 (copy number of 1). A) Simple alignment of barcodes to pCT reference barcode. i) The alignment weight of all fragments. 211 barcodes align with a weight greater than 0.7, but only 88 of these are originally generated from pCT. ii) The mean alignment of 211 barcodes with alignment weight greater than 0.7. B) *De novo* separation of barcodes. t-SNE is used to visualise the network generated from the adjacency matrix. i) Barcodes generated from pCT (yellow) and *E. coli* (blue) reference genomes. ii) Weight of alignment to pCT reference genome in A). iii) Community detection. C) *De novo* alignment of barcodes to generate consensus barcodes (red). Three clusters are shown aligned to the pCT reference genome (blue), but none align well.**

4.2.8 Identification of bacterial species and strains

In this final section these procedures are applied to the identification of bacterial species and strains. This is a more challenging problem than the identification of short genomes, since previously the experimental barcodes were around the same size as viral genomes but are now around 100 times shorter than bacterial genomes. Consequently, to get the same coverage 100 times more fragments are required, which slows computation speed, whist in addition the alignment to the correct region of the genome is 100 times more difficult. Despite these challenges, identification of genomes by using affinity labelling has been demonstrated by Nilsson *et al.*[63] It was shown previously that a lower density of labelling improves alignment to bacterial genomes (see 3.2.6), therefore using M.TaqI-directed labelling should allow for more reliable identification.

To test this, fragments were generated *in silico* from the *E. coli* strain EC958, with 50% labelling efficiency, and aligned to a library of bacterial genomes. This simple identification procedure is conceptually the same as that applied for bacteriophage genomes (e.g. Figure 4.10), however now alignment is 100 times slower (since the reference genomes are 100 times longer). The results for this alignment are shown in Figure 4.32 and Figure 4.33, and show that the strain can clearly be identified. Barcodes are either assigned to reference genomes based on the maximum alignment weight (Figure 4.32A) or by an alignment weight threshold (Figure 4.32B). In this simulated data the threshold does not identify the species/strain as clearly as the assignment to the maximum alignment weight. This is in contrast to short genomes (e.g. Figure 4.8), since there is now a much greater chance that a fragment will align well, by chance, to a bacterial genome, since there are many more possible sites of alignment.

First, the barcodes are aligned to a library of 23 bacterial genomes that represent a wide range of bacterial species, this is shown in Figure 4.32. Since strains of the same bacterial species will have similar genome sequences this can be used to identify the species before identifying the strain, for example in this case *E. coli* K-12 is clearly identified.

This allows for the barcodes to be aligned to a refined library of *E. coli* genomes for more accurate identification, the results of alignment to 10 *E. coli* genomes is shown in Figure 4.33. *E. coli* strain EC958 is rapidly (~1 hour) and clearly identified based on the generated barcodes. Also evident is the similarity of genomes, which can make it difficult to resolve genomes which share regions of largely identical sequences. For example, JJ1886 is most closely related to EC958 and SE15 is slightly more distantly related, but all three belong to the ST131 specific lineage[184]. CFT073 and 536 strains belong to the same B2 phylogroup, so also share a large amount of sequence similarity, however other genomes are more distantly related.

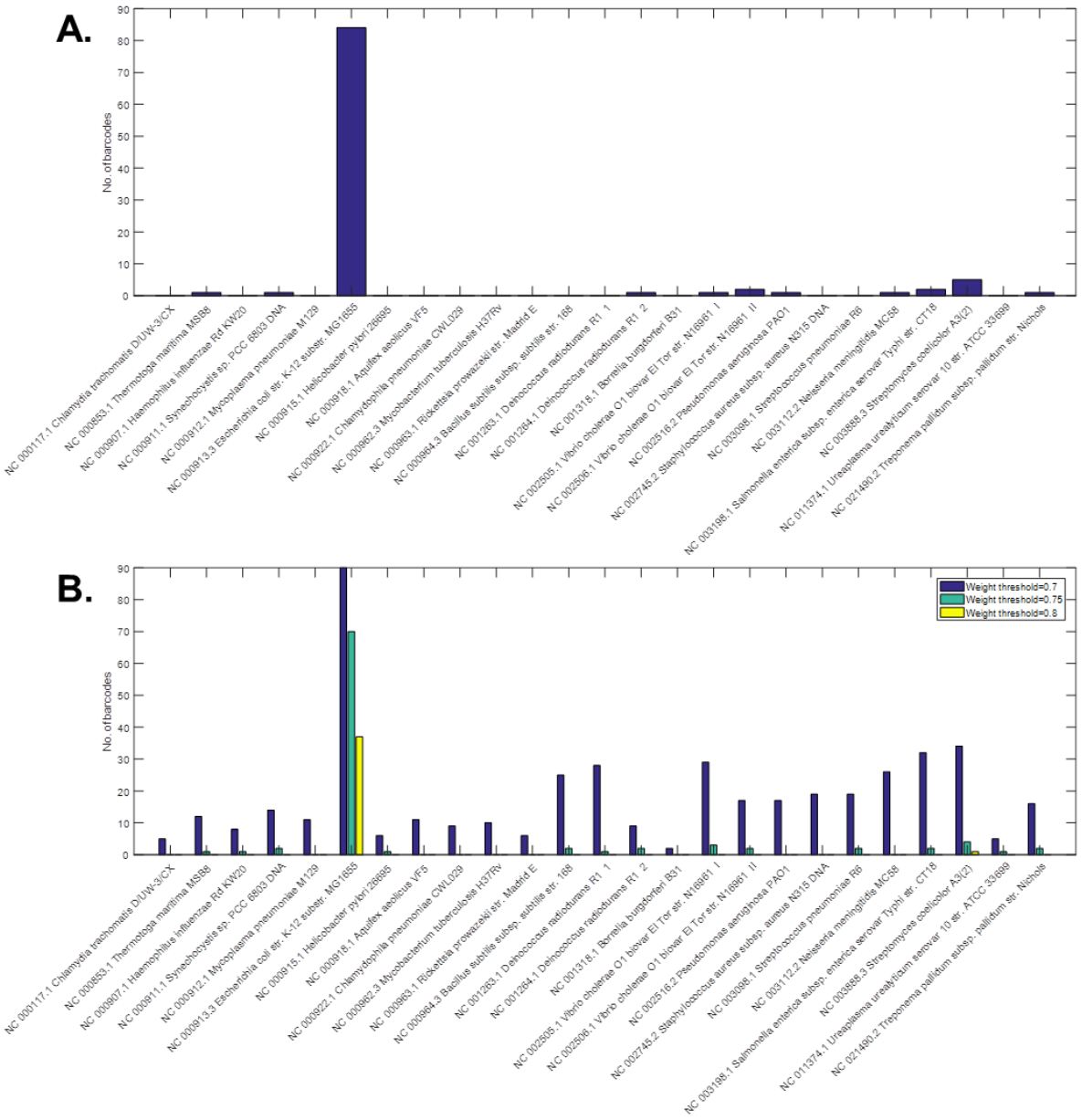


Figure 4.32 **Identification of *in silico* barcodes generated from E*. coli* strain EC958, by species. 100 barcodes were generated *in silico* from the *E. coli* strain EC958, with 50% labelling efficiency. A) Each barcode was assigned to the species to which its alignment yielded the highest alignment weight. B) The number of barcodes aligned to each reference genome with an alignment weight greater than 0.7 (blue), 0.75 (cyan) or 0.8 (yellow).**

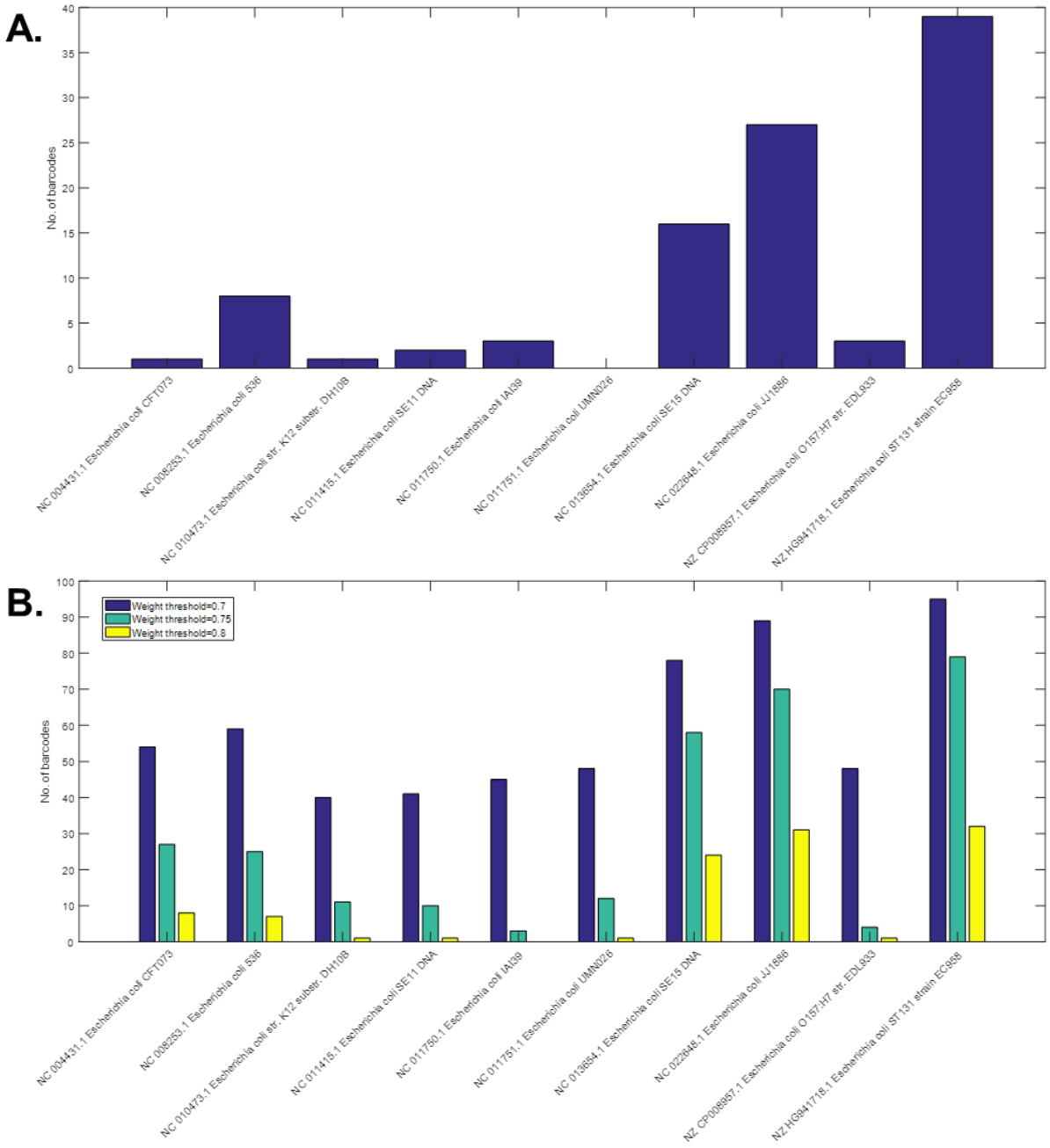


Figure 4.33 **Identification of *in silico* barcodes generated from E*. coli* strain EC958, by strain. 100 barcodes were generated *in silico* from the *E. coli* strain EC958, with 50% labelling efficiency. A) Each barcode was assigned to the strain to which its alignment yielded the highest alignment weight. B) The number of barcodes aligned to each reference genome with an alignment weight greater than 0.7 (blue), 0.75 (cyan) or 0.8 (yellow).**

For experimental barcodes this type of identification is not as reliable. Results are shown for an experimental sample of *E. coli* strain DH10B in Figure 4.34 and Figure 4.35. *E. coli* is identified from the library of bacteria (Figure 4.34), but not as clearly as before since a higher proportion of barcodes also align well to other genomes. These are either junk barcodes that have been discussed previously, or experimental barcodes which do not align as well as those generated *in silico* because of poorer experimental parameters such as labelling efficiency. Regardless, DH10B is also the genome that is identified when the barcodes are run against the library of *E. coli* genomes (Figure 4.35).

Note that although fewer than 1000 barcodes are used for this alignment, giving relatively poor coverage of the genome, the identification is robust. A relatively low coverage must be used to reduce computation time and therefore fragments can be filtered thoroughly (based on length and the intensity profile). Consequently, the fragments that remain are of high quality and identification appears to be reliable. The number of fragments assigned to *E. coli* is over double any other species, and *E. coli* is the only species with a significant number of fragments that have an alignment weight of over 0.8. Similar results are obtained for DH10B when compared to other strains.

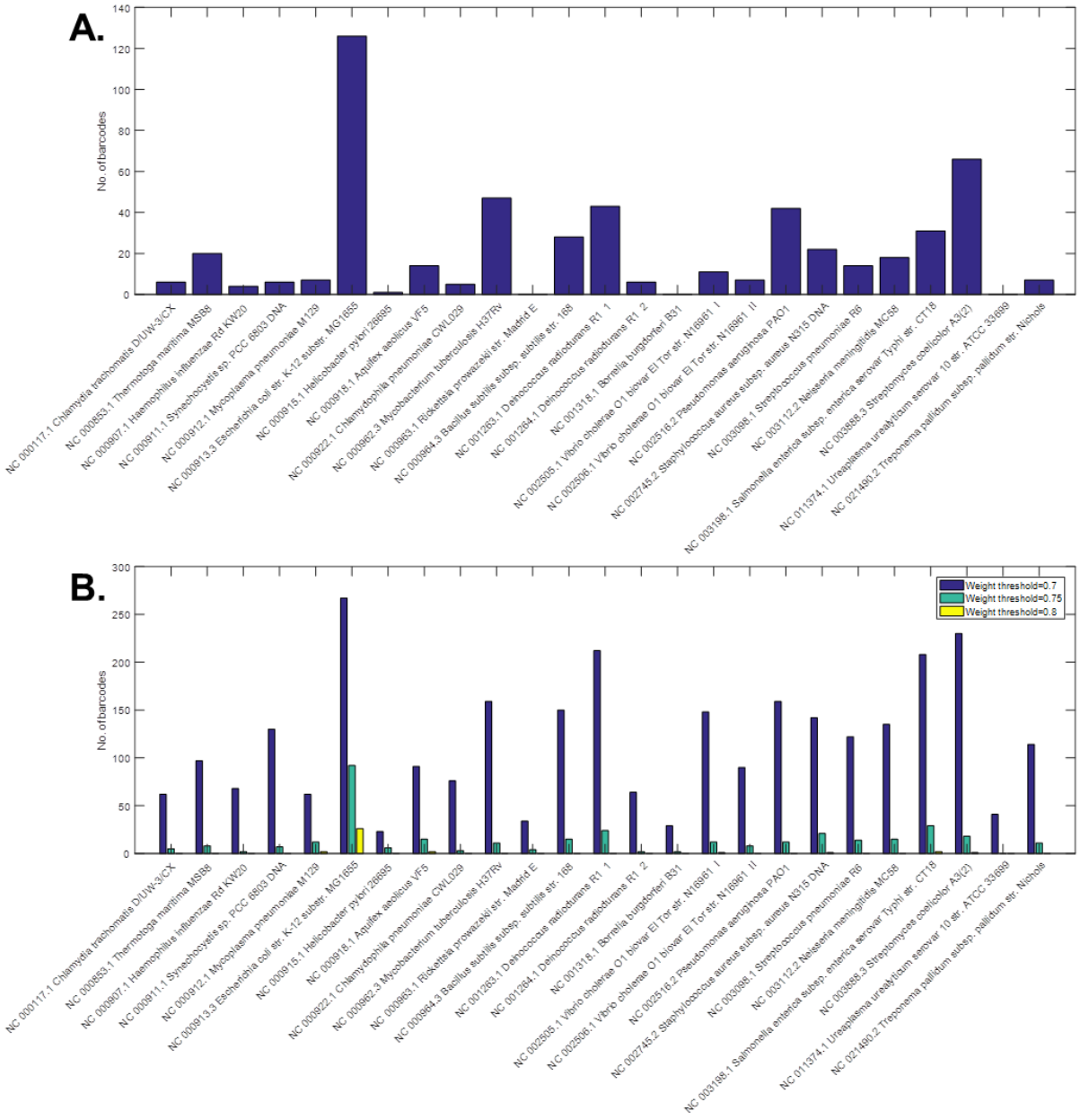


Figure 4.34 **Identification of E. *coli* strain DH10B, by species. Identification of bacterial DNA by species. A sample of *E. coli* strain DH10B was labelled using M.TaqI with Atto47N, combed and images as normal. A) Each barcode was assigned to the species to which its alignment yielded the highest alignment weight. B) The number of barcodes aligned to each reference genome with an alignment weight greater than 0.7 (blue), 0.75 (cyan) or 0.8 (yellow).**

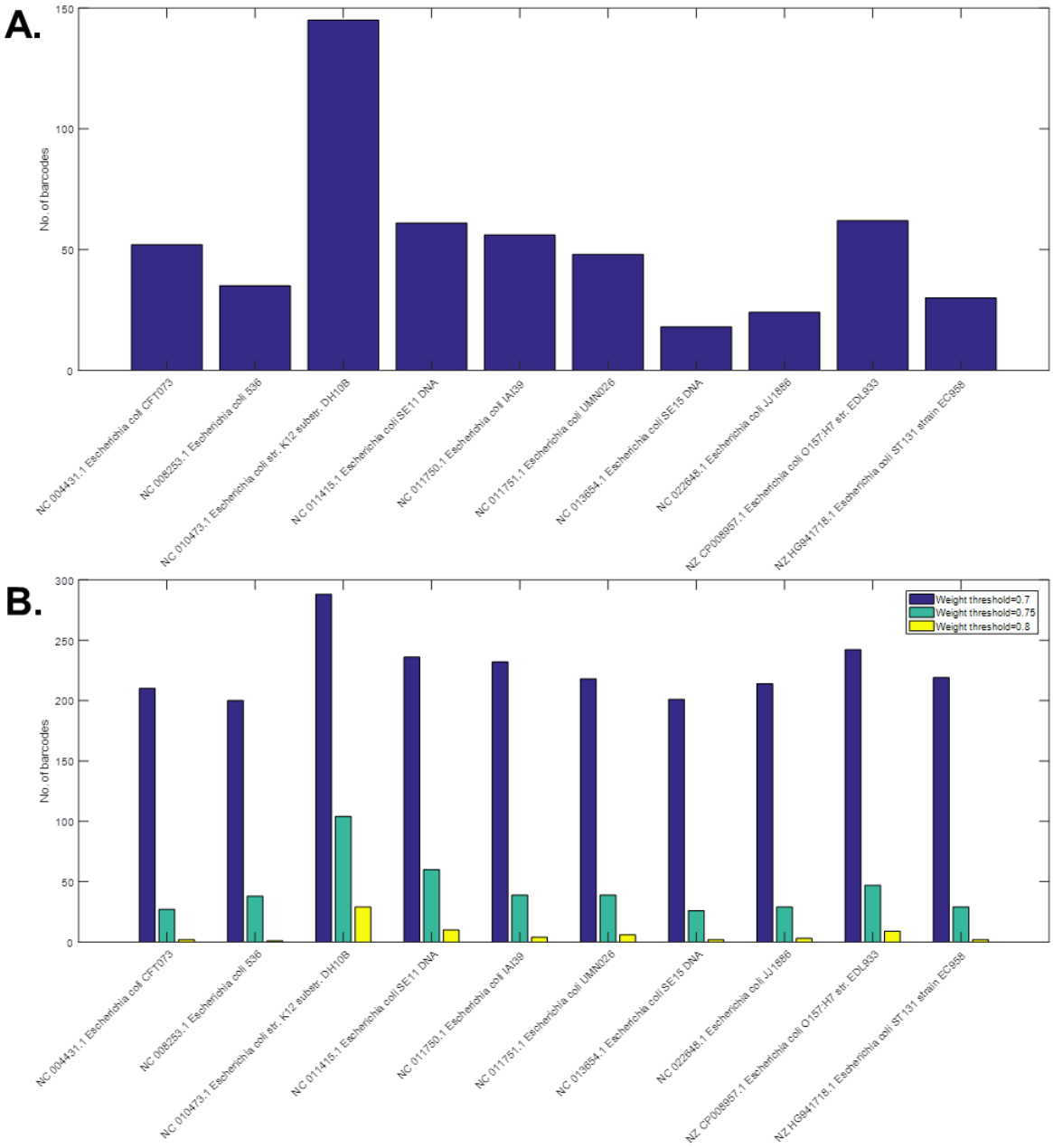


Figure 4.35 **Identification of E*. coli* strain DH10B, by strain. A sample of *E. coli* strain DH10B was labelled using M.TaqI with Atto47N, combed and images as normal. A) Each barcode was assigned to the strain to which its alignment yielded the highest alignment weight. B) The number of barcodes aligned to each reference genome with an alignment weight greater than 0.7 (blue), 0.75 (cyan) or 0.8 (yellow).**

This has also been used to identify the three samples which contained resistance plasmids, and which were from: *E. coli* strain DH10B; *E. coli* strain EC958 and *K. pneumoniae* strain Ecl8. The species can be identified by assigning barcodes to the strain with which the highest alignment weight is obtained (Figure 4.36A). Only a draft complete genome sequence is available for EcI8, however the *E. coli* strains can be identified from a library of *E. coli* genomes (Figure 4.36B). Here barcodes are assigned to a genome if a threshold of 0.75 is reached for the alignment weight. DH10B is clearly identified, whilst EC958 and the other strains from the ST131 lineage are also identified.

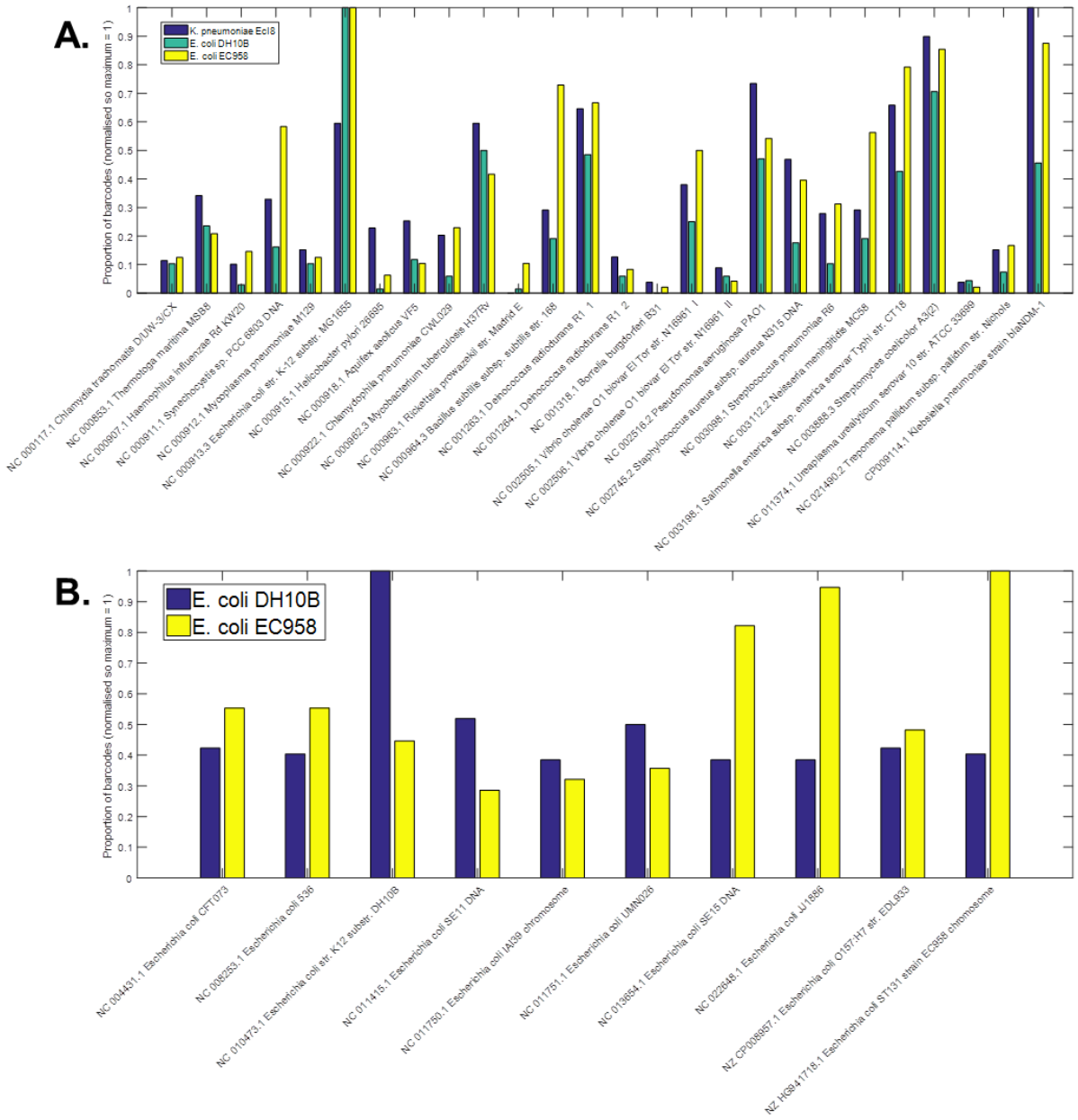


Figure 4.36 **Identification of *E. coli* strain DH10B; *E. coli* strain EC958 and *K. pneumoniae* strain Ecl8. DNA from *E. coli* strain DH10B; *E. coli* strain EC958 and *K. pneumoniae* strain Ecl8 was extracted, labelled using M.TaqI with Atto647N, combed and imaged as normal. Between 100 and 1000 experimental barcodes were used from each sample for alignment. A) Each barcode was assigned to the species to which its alignment yielded the highest alignment weight. DH10B (cyan) and EC958 (yellow) are both identified as *E. coli* and Ecl8 (blue) as *K. pneumoniae*. B) The number of barcodes aligned to each *E. coli* reference genome with an alignment weight greater than 0.75. DH10B (blue) and EC958 (yellow) are both identified.**

As well as identifying bacteria this type of simple alignment can be used if a strain is already known to confirm the genome identity. For example, if barcodes which are generated *in silico* from the DH10B strain are aligned to the reference genome of the same strain then the alignment weight and number of barcodes fitting across the genome is consistent (Figure 4.37A). However, if the same barcodes are aligned to the closely-related strain W3110 then there are several regions where barcodes align poorly (Figure 4.37B). A BLAST analysis can be carried out to compare the genomes and show the regions of difference (larger than 5 kbp). These appear to line up with the regions where poor alignment is occurring. This type of analysis therefore shows the regions in which the genome sequences differ and can be used to trace evolutionary history or to check sequence alignment. Experimental results are shown in Figure 4.37C and D, however the results are not as clear, presumably as a result of junk barcodes but may also be due to other experimental factors which limit correct alignment (e.g. labelling efficiency). Note how there are several regions where alignment is preferred, these are intense regions of the reference barcode, for which cross-correlation has a propensity.

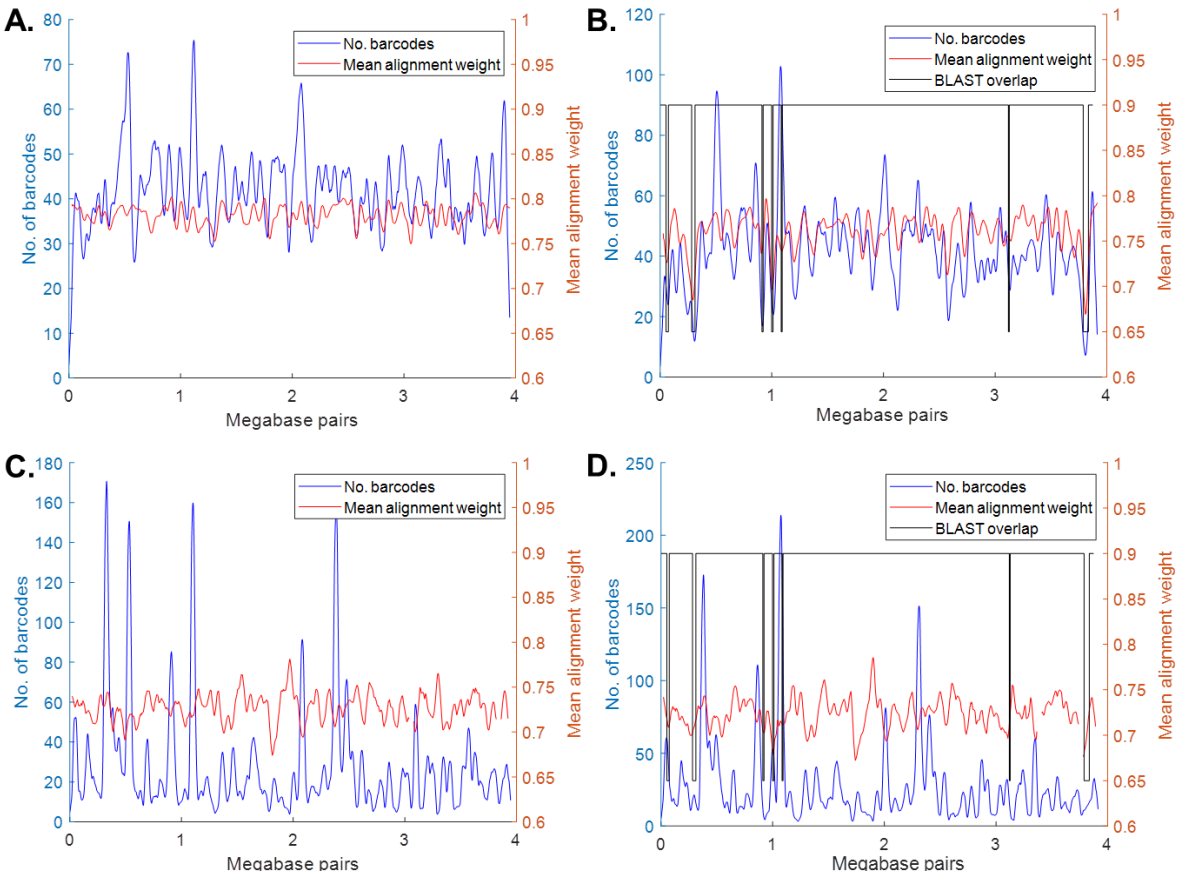


Figure 4.37 Examination of alignment across reference genome sequences. 5000 fragments were generated *in silico* from the DH10B genome, with 50% labelling efficiency for analysis in A) and B). For C) and D) around 5000 fragments were extracted from an experimental sample of DH10B genomic DNA. A and C) The alignment of fragments across the DH10B reference barcode. The no. of barcodes aligned across the reference (blue) and the mean alignment weight across the reference (red) is shown. The alignment across the reference is consistent. B and D) The alignment of DH10B fragments across the W3110 reference barcode. Now the number of barcodes aligning (blue) and the mean alignment weight (red) shows regions of poor alignment. These regions correlate with differences in the genomes which can be shown by BLAST alignment.

As well as this simple alignment procedure for identification and confirmation, the procedure for *de novo* separation and alignment can be used. As before, this provides the advantage of drastically reducing the time for alignment to large reference libraries. In this case since each reference is 100 times larger than a viral genome this is even more important. However, this procedure fails for experimental samples. Cross-correlation for

this experimental data (i.e. at this level of coverage and with this quality) isn't sufficient to identify tight clusters. Therefore, consensus barcodes cannot be used for identification (Supplementary Figure 7.23).

Despite this, to test the validity of this approach, barcodes have been generated *in silico* from the three strains described (*E. coli* DH10B and EC958 and *K. pneumoniae* blaDNM-1), with 60% labelling efficiency and an average length of 40 kbp. An affinity matrix and adjacency matrix are generated as normal and the results are visualised by t-SNE in Figure 4.38. Each genome is shown by a different colour in Figure 4.38A and the cluster detection in Figure 4.38B. It is clear that regions from each genome can be clustered together to derive many reliable consensus barcodes.

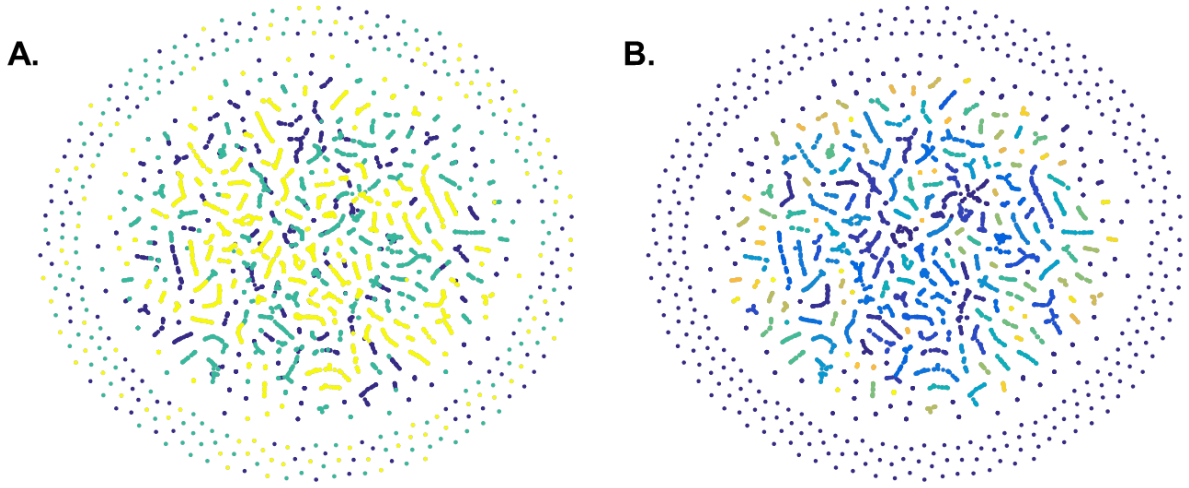


Figure 4.38 **Community detection for barcodes generated *in silico* from bacterial genomes. 1500-2500 barcodes are generated for DH10B, EC958 and blaDNM-1 genomes, with 60% labelling efficiency. A) t-SNE visualisation of network generated from adjacency matrix. Each colour represents a different genome: DH10B (blue), EC958 (cyan) and blaDNM-1 (yellow) B) Community detection. Each colour represents a community that has been detected.**

The consensus barcodes are aligned in turn to a library of approximately 500 bacterial genomes (complete enterobacteria genomes with lengths between 4.5 Mbp and 5.5 Mbp) and the results are shown in Figure 4.39. The identification of the three strains is clear and reliable, similar to the results for complex mixtures of bacteriophages (Figure 4.22) and the number of barcodes assigned to each genome is similar to the number of generated barcodes (Figure 4.39B).

The difference when compared to bacteriophages is that a large number of very similar strains are now identified, the top 19 strains, for which at least 50 barcodes were assigned, are shown in Figure 4.39A. This is because that even if large regions of the genomes are different, the bacterial strains can share 100 kbp regions (i.e. the size of the consensus barcodes) that are identical, so other *K. pneumoniae* and *E. coli* strains are identified. The barcodes are not being mis-assigned (examples of aligned consensus barcodes are shown in Figure 4.39C-E), the genomes themselves are nearly identical, which can make the exact determination of the sample difficult. However, if unique regions (i.e. sequences) can be identified for each genome then these can be used as a reference to identify the exact strain.

The results in Figure 4.38 also suggest a possible route to *de novo* alignment of whole bacteria genomes. It is clear from the shape of clusters that long regions of the genome are being identified, in other words one end of the cluster will contain barcodes from many kbps away from barcodes contained in the other end of the cluster. If the adjacency matrix was perfectly formed then these fragments would form a circle in the t-SNE visualisation of the network (Figure 4.40A), which would represent the ordering of all fragments along the genome. This could be used to generate the whole bacteria

genome *de* novo, although not by the procedure described in this research, since this is optimised for alignment of barcodes which completely overlap.

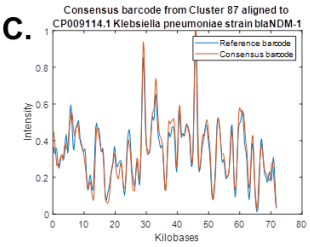
A.

Genome	Number of barcodes	Query coverage of K. pneumoniae blaNDM-1	Query coverage of E. coli EC958	Query coverage of E. coli DH10B
CP009114.1 Klebsiella pneumoniae strain blaNDM-1	881	100%	42%	47%
HG941718.1 Escherichia coli ST131 strain EC958	803	41%	100%	87%
CP006784.1 Escherichia coli JJ1886	483	41%	98%	87%
CP000948.1 Escherichia coli str. K12 substr. DH10B	392	40%	77%	100%
AP009378.1 Escherichia coli SE15 DNA	160	40%	90%	86%
CP003200.1 Klebsiella pneumoniae subsp. pneumoniae HS11286	144	95%	43%	47%
CU928164.2 Escherichia coli IAI39	104	41%	82%	88%
CP006918.1 Klebsiella pneumoniae 30684/NJST258_2	99	95%	42%	46%
CP006659.2 Klebsiella pneumoniae strain ATCC BAA-2146	88	40%	43%	46%
CP001671.1 Escherichia coli ABU 83972	84	40%	89%	86%
CP010392.1 Klebsiella pneumoniae strain 34618	83	95%	42%	46%
CP008797.1 Klebsiella pneumoniae subsp. pneumoniae KPNIH24	77	96%	43%	47%
CP006656.1 Klebsiella pneumoniae JM45	76	94%	42%	46%
CP008827.1 Klebsiella pneumoniae subsp. pneumoniae KPNIH	71	95%	42%	46%
CP000647.1 Klebsiella pneumoniae subsp. pneumoniae MGH 78578	70	92%	42%	47%
CP000802.1 Escherichia coli HS	60	41%	78%	90%
LM993812.1 Escherichia coli genome assembly EcHMS174Chr	55	40%	78%	98%
CP007799.1 Escherichia coli Nissle 1917	51	40%	88%	86%
CP010444.1 Escherichia coli K-12 strain ER3475	50	40%	79%	98%

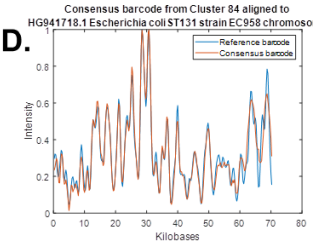
B.

Genome	No. of generated barcodes	No. of assigned barcodes (consensus weight >0.8)
K. pneumoniae blaNDM-1	2500	2153
E. coli EC958	2000	1787
E. coli DH10B	1500	1135

C.



D.



E.

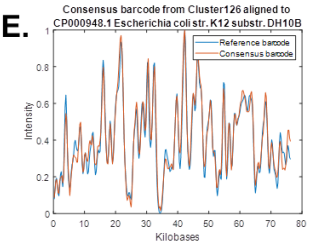


Figure 4.39 Identification and quantification of complex mixture of bacteria, generated *in silico*. 1500-2500 barcodes are generated for DH10B, EC958 and blaDNM-1 genomes, with 60% labelling efficiency. Clustering was shown in Figure 4.38. A consensus barcode was generated for each cluster, aligned to a library of around 500 enterobacteria and assigned to the strain for which the maximum alignment weight was obtained. A) The overall result by genome. A relatively large number of genomes are identified, but all share a large amount of homology (shown by the BLAST query coverage) with one of the genomes from which barcodes were generated. B) Quantification of the result. For each genome, the number of barcodes generated *in silico* is shown as well as the number of barcodes contributing to consensus barcodes that were assigned, based on an alignment threshold of 0.8. C-E) An example alignment for each genome: C) blaDNM-1; D) EC958 and E) DH10B.

A more realistic t-SNE visualisation of barcodes generated from a single genome is shown in Figure 4.40B, which shows that a circular genome would be difficult to obtain. Very large clusters can be formed, which represent fragments from regions hundreds of

kbp in length, however these are separated when, by chance, two adjacent barcodes are not well-aligned, which removes the edges from the network. However, the possibility remains that despite isolated clusters being formed, these could be connected by considering the network formed from the affinity if enough edges can be identified to join clusters together.

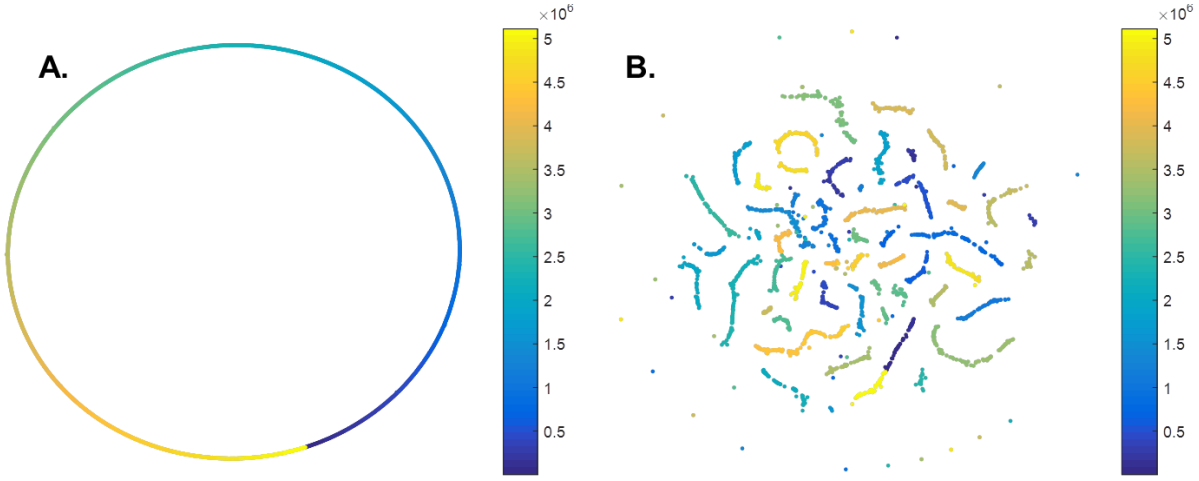


Figure 4.40 **t-SNE visualisations for experimental barcodes generated *in silico* from a bacterial genome. 5000 barcodes are generated for genomes, with 60% labelling efficiency. A) Ideal t-SNE if adjacency matrix connected all overlapping barcodes. The colour represents the position of the barcode in the genome. A circle is formed which orders all fragments and could be used for *de novo* alignment of a whole bacterial genome. B) t-SNE results if adjacency matrix is used, based on alignment of all barcodes to each other. Long clusters are formed which represent regions of the circle, but these are separated when there are gaps present in the adjacency matrix.**

4.3 Conclusion

An optical mapping method using methyltransferase-directed labelling of DNA, molecular combing and widefield microscopy has been developed and demonstrated for the identification of microorganisms. This exploits the unique pattern generated by fluorescent labelling at specific sequences, to identify DNA fragments. In CHAPTER 3 it was shown that this combination of experimental techniques is the 'sweet spot' for accuracy and speed of identification of microorganisms. In CHAPTER 4 a new rapid procedure for separation, *de novo* alignment and identification of microorganisms has been described to exploit the barcodes which are extracted.

These procedures have been applied to identify DNA in a number of systems ranging in the size of genomic DNA and the complexity of the sample. A summary of these applications is shown in Figure 4.41. This shows both the types of samples for which identification has been demonstrated, but also those to which this optical mapping procedure has not yet been applied. At the boundaries of these, barcodes generated *in silico* have been used to demonstrate the applicability of these procedures, but these have not yet been confirmed experimentally.

The easiest samples to identify are pure samples of short genomes, and simple and *de novo* alignment has been applied to reliably identify experimental samples of lambda, T7 and Adenovirus A viral DNA. More complex mixtures increase the challenge of identification, however experimental mixtures of lambda, T7 and *E. coli* genomic DNA have been reliably identified, as have resistance plasmids from mixtures generated *in silico.* These procedures have been pushed to the limit for very complex *in silico* samples containing twenty phages. Apart from increasing the complexity of the mixture,

increasing the size of the genomic DNA makes identification more challenging. Pure experimental samples of *E. coli* and *K. pneumoniae* have been identified and this has been extended to mixtures of bacterial strains *in silico*. As well as identification these methods have been applied to check the DNA sample (e.g. M.TaqI-labelling of dam-methylated lambda DNA) and to enable two-colour mapping of an alternative label (e.g. YOYO-1 affinity labelling).

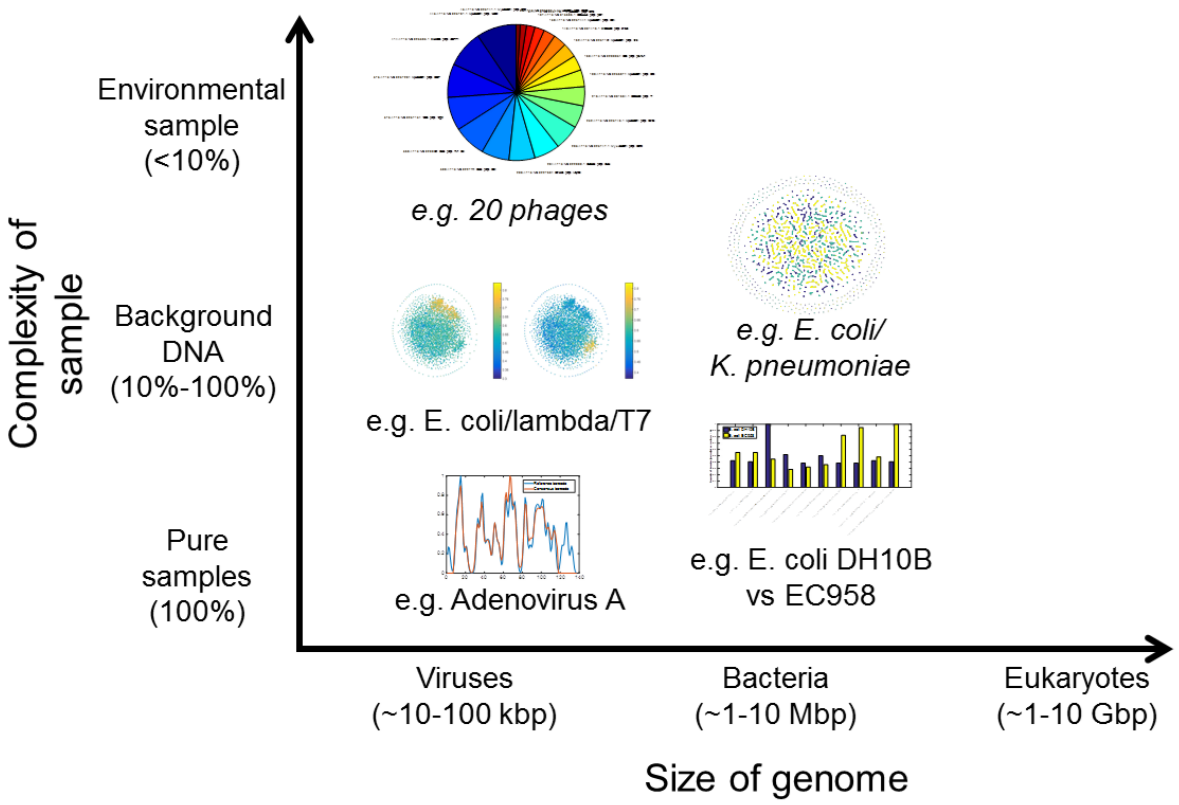


Figure 4.41 **Overview of samples identified in this research. As the size of the DNA and the complexity of the sample increases the identification of the sample by optical mapping becomes more challenging. M.TaqI-directed labelling, molecular combing and widefield microscopy have been applied in this research to relatively small genomes and simple mixtures. Examples from this chapter are shown and italics are used to highlight samples that were generated *in silico*.**

The main issues that appear to be holding back this technique are junk barcodes (artefacts in barcode extraction and/or overlapping DNA molecules). These are primarily due to poor labelling efficiency and difficulties obtaining ideal molecular combing of molecules. Further optimisation of the labelling strategy (described in CHAPTER 2) in tandem with improved molecular combing would reduce these problems, and may therefore allow these procedures to be applied beyond the bounds of this research, towards complex environmental samples, and possibly toward larger genomes (e.g. prokaryotes).

4.4 Materials and Methods

4.4.1 Source and extraction of genomic DNA

Adenovirus A sample was kindly provided by Roger Grant. Overnights of pCT in *E. coli* strain EC958; pNDM in *E. coli* strain DH10B; and pKpQIL in K. pneumoniae strain Ecl8 were provided by Michelle Buckner and DNA was obtained by extraction using a GenElute Bacterial Genomic DNA kit (Sigma-Aldrich).

4.4.2 *In silico* generation of barcodes and *de novo* alignment procedures

Custom code was written using Matlab 2016b for automated extraction, *in silico* generation of barcodes and alignment procedures. Unless otherwise stated, barcodes were generated using parameters in Supplementary Table 7.1. Copy of the code is available on request from Robert Neely, University of Birmingham.

CHAPTER 5 LOCALISATION AND DYNAMICS OF SINGLE PLASMID MOLECULES IN *E. COLI*

Robert K. Neely and Stephen J. W. Busby provided supervision and guidance for the research undertaken in this chapter. Nathaniel O. Wand (the author) designed and analysed all labelling experiments, transformations and imaging experiments. Lara Horne performed all labelling experiments, transformations and imaging experiments under the supervision of Nathaniel O. Wand (the author). Nathaniel O. Wand (the author) also developed and performed all image analysis procedures unless otherwise stated.

5.1 Introduction

5.1.1 Plasmid function and maintenance

Plasmids are small extrachromosomal pieces of DNA that are found naturally in bacteria and have essential roles in metabolism[185], pathogenesis[186] and resistance[187]. They are usually circular, ranging in size from 1 to 100kbp and have copy numbers ranging from a single copy to hundreds[103]. The genes plasmids carry include those which promote replication, maintenance and proliferation of the plasmid, but also genes that help the host to adapt to the environment. These include genes for antibiotic resistance, one of the greatest public health threats we face[109]. However, recombinant plasmids can also carry genes that can be manipulated for industrial processes, such as the production of human insulin[188]. Understanding the mechanisms of plasmid maintenance would allow better control of plasmid retention, whether desirable or not.

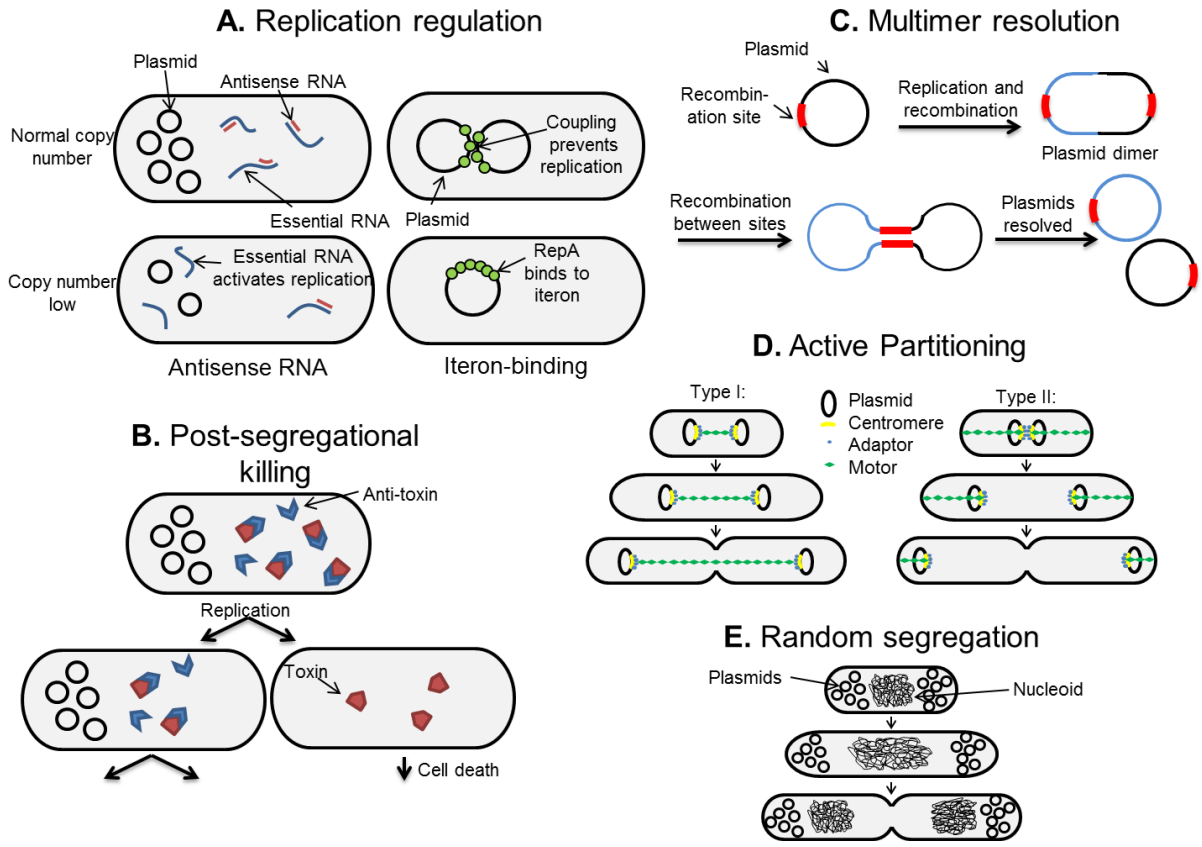


Figure 5.1 **Overview of plasmid maintenance and transfer in bacteria. A) *Replication regulation*. There are two main mechanisms: antisense RNA and iteron-binding. In antisense at high copy numbers the antisense RNA will hybridise to an RNA which is essential for replication, whilst at low copy numbers less antisense RNA is transcribed and replication can proceed. In iteron-binding mechanisms a protein, RepA, binds to a specific site on the plasmid, the iteron. When the plasmid is replicated RepA couples the newly replicated plasmids, preventing further replication until RepA is diluted. B) *Post-segregational killing*. When copies of the plasmid are present a short-lived antitoxin is produced that prevents a longer-lived toxin from killing the cell. However, if during cell division the plasmid is no longer present, the toxin leads to cell death. C) *Multimer resolution*. During replication and recombination plasmids can form dimers, which are resolved by multimer resolution systems. D) *Active partitioning*. Several types of active partitioning system exist in bacteria, to ensure proper segregation of low copy number plasmids. The two most common are Type I and II, which both use adaptor proteins to bind to specific regions on the plasmid, known as centromeres. Motor proteins can bind to the adaptor and either 'push' or 'pull' the plasmids to the cell poles. E) *Random segregation*. For high copy number plasmids, it is thought that plasmids randomly segregate during division, since there is a low probability of daughter cells having no copies of the plasmid.**

Several mechanisms are used by bacteria to ensure plasmids are maintained in a stable copy number[189] and transferred to daughter cells[113,190]. These mechanisms are shown in Figure 5.1 and include regulation of replication, partition mechanisms, multimer resolution and post-segregational killing. Active partitioning mechanisms are used for low copy number plasmids, to ensure proper segregation and transfer to daughter cells. However, it is thought that no such active partition mechanisms are involved for higher copy number plasmids. Instead, it is believed they are randomly segregated during division, meaning by chance each daughter cell should retain at least a single copy[114]. There is some debate about this mechanism, as microscopy of fluorescently-labelled plasmids has shown them clustering, despite the absence of any known active partitioning systems[115].

5.1.2 Plasmid localisation and dynamics in bacteria

It was first shown that high copy number plasmids cluster at specific locations by using DAPI, a nuclear stain, to visualise derivatives of the high copy number plasmids R100 (60kbp, 10-15 copies per cell) and pBR322 (8kbp, 40-60 copies per cell)[191]. This used a non-specific stain to visualise DNA in fixed cells, and showed that the R100 derivative was evenly spaced throughout the cell (Figure 5.2A), whilst the pBR322 derivative was localised to cell poles (Figure 5.2B). The main drawback with this approach is that the staining is non-specific. The nucleoid is stained as well as plasmids, so cannot be differentiated except by location. If, for instance, individual plasmids are residing within the nucleoid region, they are not visible.

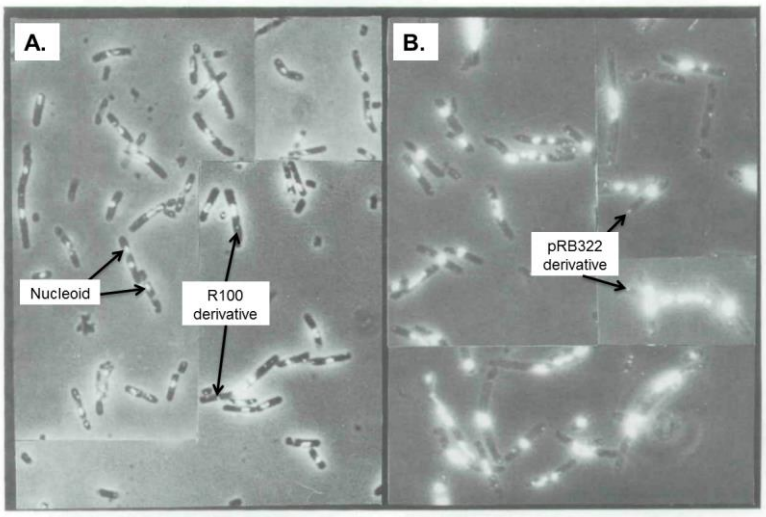


Figure 5.2 Clustering of high copy number plasmids, by staining with DAPI. Adapted from Eliasson et al[191]. A comparison is shown between intracellular distribution of derivatives of A) R100 and B) pRB322. The R100 derivative is evenly spaced throughout the cell, while the pBR322 derivative is localised to cell poles. Note how the DAPI stain is non-specific and has labelled the nucleoid (large, bright region in the centre of cells) as well as plasmids (small spots).

An alternative approach for fixed cells is to use fluorescent in situ hybridisation (FISH, Figure 5.6A), which allows for specific sequences to be fluorescently tagged. In this method short DNA or RNA probes are designed to bind to unique complementary DNA sequences. When these are fluorescently labelled they can be used for visualising the location of plasmids, which contain the unique sequence. This must be done on fixed and permeabilised cells, since cellular uptake of probes is poor, and probes can be unstable. This method has been used since the late 1990s to visualise low copy number plasmids and investigate their subcellular distribution during active partitioning[115,192,193]. For example, the low copy number plasmid R1 (95kbp, 4-5 copies per cell), is known to be segregated by an active partition mechanism. However, microscopy showed that there were fewer plasmid foci in cells than plasmid copies, suggesting clustering occurs. These

results were confirmed by a copy number mutant, which contained the same number of foci as before but now two to three times more plasmids per cell[193].

More recently FISH has been used, in combination with localisation microscopy, to study the distribution of high copy number plasmids[118]. Here a ColE1-like plasmid, pBluescript (3kbp, >300 copies per cell), containing an array of 256 LacO repeats and 96 TetO repeats (~14kbp) was used, and visualised by a 20bp Atto 532-labelled probe that targeted the LacO repeats. The plasmids were shown to aggregate into large clusters, however most plasmids were observed to be located randomly throughout the cell (although excluded by chromosomal DNA). The number of plasmids in each cluster could also be quantified, however the movement of plasmids and changes in distribution cannot be studied using fixed cells.

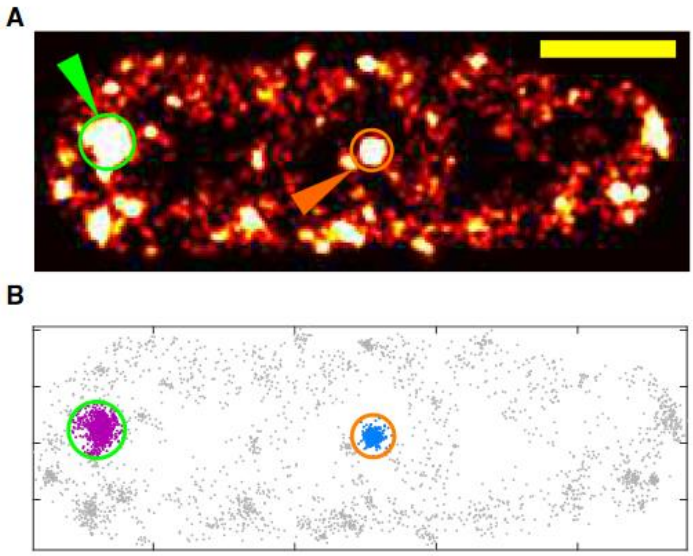


Figure 5.3 **Fluorescent in situ hybridisation (FISH) for quantitative localisation of high copy number plasmids. Taken from Wang et al[118], scale bar = 500 nm. A-B) The location of derivatives of pBluescript plasmid within a single *E. coli* bacterium. A) Pixel brightness indicates the density of individual plasmid localisations, which are shown in B). Two dense clusters are highlighted (green and orange rings). The number of plasmids in each cluster can be estimated based on the hybridisation efficiency of the probe.**

The distribution and dynamics of plasmids in living cells has generally been studied using fluorescent repressor operator systems (FROS, Figure 5.6B). These were used in the late 1990s to visualise low copy number plasmids[194–196], and the first report of FROS to visualise clustering of a high copy number plasmid was by Pogliano *et al.* in 2001 (Figure 5.4)[115]. FISH was used as a comparison. Two FROS systems in common use are the tetracycline (Tet) and the lactose (Lac) operator/repressor systems, which both rely on bacterial repressors fused to a fluorescent protein. The fusion protein is expressed within the cell and subsequently binds to its respective operator system which can be inserted as a tandem array within the plasmid of interest.

Pogliano *et al*. inserted a large tandem repeat of Lac operators (10kbp, ~256 copies) into RK2 (60kbp, 5-8 copies per cell) and pUC19 (2.7kbp, 40-250 copies per cell) derivatives[115]. A GFP-LacI fusion protein was expressed by cells from the arabinose promoter on a compatible plasmid and used to visualise plasmids. The majority of cells containing RK2 had one or two clusters visible with either a single focus localised near the mid-cell, or two foci near the ¼ and ¾ cell positions (Figure 5.4A-C). Cells containing pUC19 also had un-clustered plasmids and plasmids that appeared to be rapidly diffusing through the whole cell (Figure 5.4D-F). The distribution of RK2 clusters was shown to change during cell growth, as time lapse images showed a single focus could split rapidly into multiple foci and the number of clusters was dependent on the cell length. The clustering of RK2 is conserved across bacterial species[197] and similar behaviour has been observed for ColE1 (6.6kbp, 10-15 copies per cell)[116]. ColE1 was found to cluster at cell poles, but also movement to the mid-cell was observed. There is also evidence that as well as replication[197,198], transcription and translation of plasmids

can drive their clustering[119,199], although it is not clear these processes are always required for clustering[200].

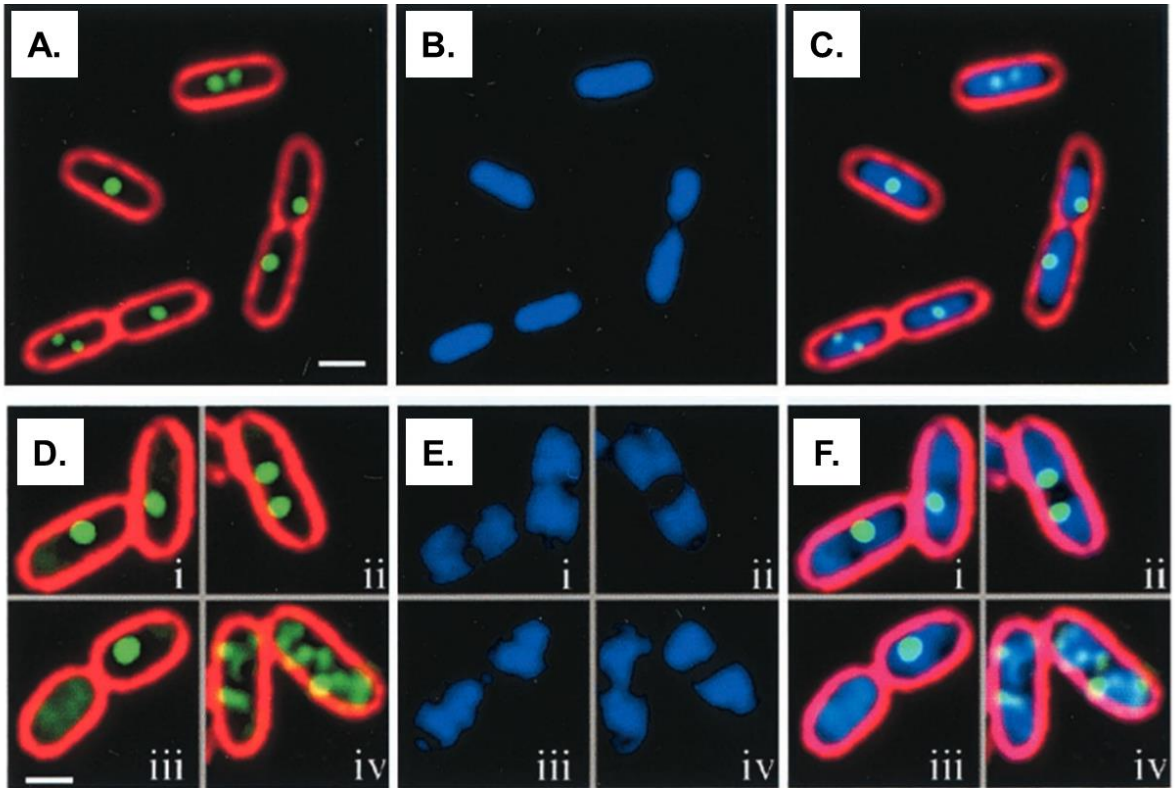


Figure 5.4 **Clustering of high copy number plasmids, using fluorescent repressor operator system (FROS). Adapted from Pogliano *et al*[115], scale bar = 1µm. A comparison is shown between intracellular distribution of derivatives of A-C) RK2 and D-F) pUC19. Cell membranes are stained with FM 4-64 (red), plasmids tagged with LacI-GFP (green) and DNA stained with DAPI (blue). A-C) The RK2 derivative forms one or two clusters, with either a single focus localised near the mid-cell, or two foci near the ¼ and ¾ cell positions D-F) The pUC19 derivative: i) with one focus, ii) with two foci, iii) during division, iv) with multiple foci.**

There remains some debate about the likely explanation for this clustering of plasmids[113,201]. Importantly, multimer formation is unlikely to cause the clustering effect, since, for example, RK2 encodes a potent multimer-resolution system[115]. Reyes-Lamothe et al. attributed localisation to displacement of plasmids by the nucleoid[117]. A TetO operator array (24, 48 or 96 copies) was inserted into the ColE1-type plasmid, pJHCMW1 (11kbp, 20-30 copies per cell) and a fluorescent TetR expressed from the

chromosome. They described how individual plasmids were highly mobile, but tended to be excluded from the nucleoid, therefore localised at cell poles, only occasionally moving along the long axis of the cell.

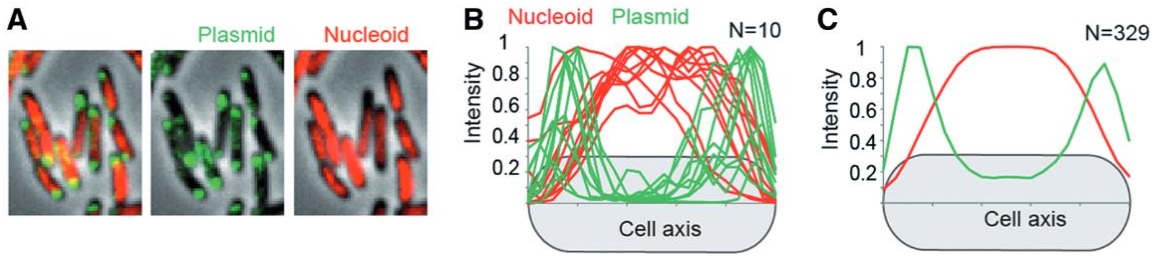


Figure 5.5 **Exclusion of plasmids from the nucleoid, using FROS. Taken from Reyes-Lamothe et al[117], using the ColE1-type plasmid, pJHCMW1. A) Typical localisation of pJHCMW1 (green) and nucleoid (red). B) Ten example, normalised, traces of the variation in plasmid (green) and nucleoid (red) signal, across the long cell axis. C) Average of the normalised traces for 329 examples. This shows that plasmids are clustered and localised to the poles of the cells and therefore tend to be excluded from the nucleoid.**

Studying individual plasmids, using either FISH or FROS, remains challenging. Quantitative localisation microscopy using FISH has been demonstrated only on fixed cells and therefore nothing about the plasmid dynamics, particularly during cell division, can be inferred. However, studying kinetics in live cells experiments using FROS is challenging, since the density of plasmids will be high. One method to overcome this has been to effectively dilute out the plasmid by preventing replication[117]. As bacteria divide the distribution of plasmids will become uneven, with some bacteria having none or very few plasmids. Individual spots can then be assumed to be individual plasmids, although this cannot be definitive.

This reduction in copy number may alter the behaviour of plasmids, as might increasing the size of plasmids several times, by the insertion of tandem arrays. The binding of many repressor molecules may also have an effect, for example GFP is known to form

inclusion bodies (insoluble aggregates of misfolded proteins) at high concentrations[202]. Therefore, it is important to develop other labelling techniques for visualisation of plasmids.

5.1.3 Fluorescent labelling of plasmids

There are a number of strategies for fluorescent labelling of plasmids for live cell imaging (Figure 5.6), an excellent review of these is given by Rombouts *et al*[97]. These broadly fit in to two categories, non-specific and sequence-specific labelling. Non-specific and non-covalent labelling can be achieved by using DNA-binding dyes (Figure 5.6C) that include groove-binding dyes (e.g. DAPI) and intercalators (e.g. YOYO-1). These can be used for staining DNA in live cells, depending primarily on cell permeability and their effect on the structure of DNA[203]. However, their non-specific and non-covalent nature means that staining plasmids with sequence specificity is not feasible.

Non-specific, but covalent attachment can be used to overcome this drawback, since the labelling will be irreversible (Figure 5.6D). Plasmids can be labelled *in vitro* before transfection back in to cells. There are a number of commercial labelling kits available for covalent attachment of fluorophores[204–206] One example is the use of reagents with an aromatic nitrogen mustard, which can be used to covalently alkylate DNA, primarily at the N7 of guanine bases[204]. This can be used to enable the direct coupling of fluorophores (e.g. Cy3, CX rhodamine) to plasmid DNA, prior to transfection and these kits have been used extensively to study plasmids in living mammalian cells. The plasmid DNA is still replicated and transcribed, however high labelling densities or bulky fluorophores can begin to effect the behaviour of the DNA[207,208].

Fluorophores are generally hydrophobic and will increase steric hindrance, whilst cationic dyes will begin to lower the overall negative charge of plasmids. Therefore, the minimum number of fluorophores, should be used to try to prevent this behaviour, however this can be difficult to control with non-specific labelling. The minimum number of fluorophores that can be used will be determined by the nature of the fluorophore and the detection system. For reliable detection a large signal to noise ratio is required, so bright fluorophores and sensitive detectors will require fewer fluorophores than dim fluorophores with insensitive detection.

Specific, covalent labelling of plasmids can be used to control labelling density, as well as label location. Methyltransferase-directed labelling is like some of the non-specific covalent labelling approaches, since organic dyes can be covalently attached to the DNA bases, but in contrast to non-specific approaches labelling is now directed to specific sequences. DNA methyltransferases are naturally occurring enzymes which covalently methylate cytosine or adenine bases. Methylation occurs within a specific DNA target, typically consisting of a palindromic sequence, 2 to 8 base pairs in length. In nature all known classes of DNA methyltransferases use the cofactor *S*-adenosyl-L-methionine (AdoMet) as the methyl donor, but it has been demonstrated that synthetic AdoMet analogues can also be used to catalyse the transfer of more complex chemical groups. Fluorophores and other modifications can be targeted to specific sites in a DNA sequence, efficiently and non-destructively[86], as shown in CHAPTER 2.

Other labelling strategies include the incorporation of modified bases[209], DNA-binding by small molecules[210], as well as by DNA hybridisation (e.g. FISH) and DNA-binding by proteins (e.g. FROS), that have already been mentioned.

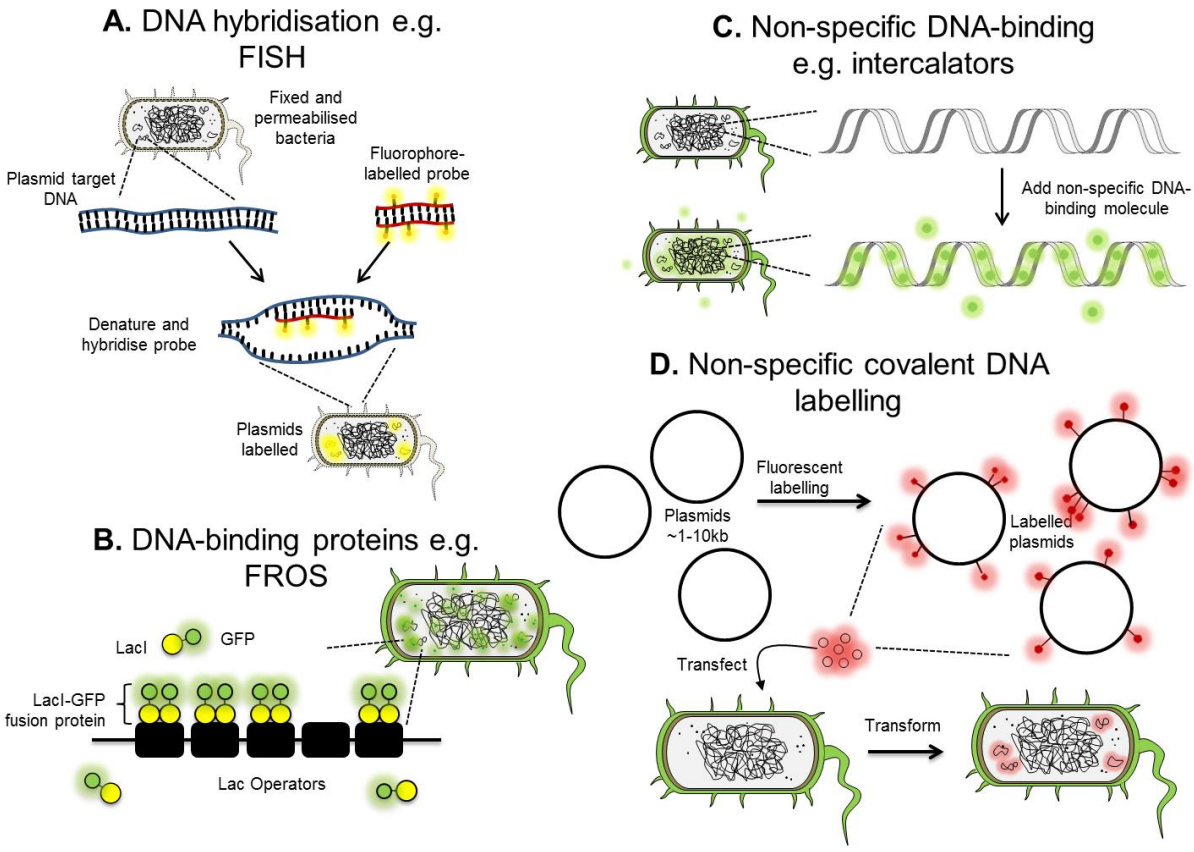


Figure 5.6 **Overview of methods for visualisation of plasmids in bacteria. A) DNA hybridisation methods, e.g. FISH. Short, fluorescent, DNA or RNA probes are designed to bind to unique complementary DNA sequences on the plasmid. Bacteria are fixed and permeabilised, the probe added and allowed to hybridise to the target DNA for visualisation. B) DNA-binding protein methods, e.g. FROS. A tandem array of a specific protein-binding site (e.g. LacO) is inserted into the plasmid. A fusion protein, composed of a repressor that binds to the array (e.g. LacI) and a fluorescent protein (e.g. GFP), is expressed within the cell and subsequently binds to the plasmid for visualisation. C) Non-specific DNA binding, e.g. intercalators. Small fluorescent molecules that bind to DNA, e.g. DAPI, YOYO-1, are added to the cell. These bind to the DNA non-specifically allowing visualisation. D) Non-specific covalent DNA labelling. Plasmids are fluorescently-labelled covalently, e.g. by an aromatic nitrogen mustard. Labelled plasmids are used to transform bacteria and allow visualisation.**

5.1.4 Overview and objectives

There are many complex mechanisms that bacteria use to maintain and transfer plasmids. In particular, there remains debate about the mechanisms of segregation of high copy number plasmids, since clustering of plasmids is found in bacteria, but the current methods of fluorescently-labelling plasmids are unsuited to studying these mechanisms. DNA hybridisation techniques such as FISH are most easily used on fixed and permealised samples, so are unsuitable to study plasmid movement. On the other hand, FROS is ideally suited to live cells, but suffers from difficulties in visualising and quantifying individual plasmids, against the dense background of fluorescent proteins.

Covalently labelling plasmids prior to transformation can overcome these drawbacks. Individual plasmids are easily tracked, and the technique can be used on live cells. Methyltransferase-directed labelling is well suited to covalently-labelling plasmids since it provides easy control of labelling density. In addition, the target sequence could be inserted into distinct locations on DNA, to label specific positions. However, the main drawback with this approach is that there is no commercially available kit. The optimisation of plasmid labelling with methyltransferases was undertaken in CHAPTER 2.

The plasmids labelled in CHAPTER 2 will be used to transform *E. coli* which will subsequently be grown on agarose pads. Ampicillin selection and expression of a fluorescent protein will be used to identify *E. coli* that contain a copy of the plasmid, allowing for visualisation of the fluorescently-labelled plasmid.

5.2 Results and discussion

5.2.1 Plasmids and strains

Two approaches were used to select for transformed *E. coli*: ampicillin resistance and the expression of a fluorescent protein. Selection is important since only around 0.1-1% of bacteria are typically transformed when chemical transformation is used[211], therefore identifying these bacteria, i.e. those that contain the fluorescently-labelled plasmid, is not straightforward.

Expression of genes encoded by the plasmid can be used to select bacteria that contain a copy of it. β-lactam antibiotics, such as ampicillin, are used to kill bacteria, since they inhibit cross-linking of the peptidoglycan in the bacterial cell wall. Eventually this leads to a weakening of the cell wall, particularly during cell growth, which can no longer compensate for the osmotic pressure and will effectively burst, causing cell death. The effect of ampicillin can be seen in Figure 5.7 and is dependent on concentration and time of growth. Ampicillin resistance is conferred to bacteria by a gene that encodes for β-lactamase, which hydrolyses β-lactams.

Alternatively, a gene that encodes for a fluorescent protein, e.g. GFP, was used. As the fluorescent protein is expressed, the fluorescence it produces will effectively light-up bacteria that contain a copy of the plasmid. In practice this is often under the control of an inducible promoter (Figure 5.8A). For example, the GFP gene can be inserted after a T7 promoter, therefore only when T7 RNA polymerase is present will the gene be transcribed and the GFP be expressed. The expression of T7 RNA polymerase, from a gene on the bacterial chromosome, can be controlled by the lac operator. The Lac repressor will bind to the lac operator, preventing expression of T7 RNA polymerase

unless lactose (or a structural isomer, e.g. IPTG) is present. Therefore, addition of IPTG should induce expression of the fluorescent protein, although a basal level of 'leaky' expression is likely, since the T7 RNA polymerase will be expressed at a low level, despite the Lac repressor. The expression of GFP is seen in Figure 5.8B, which shows how much brighter bacteria become when they contain a copy of the plasmid and GFP expression is induced.

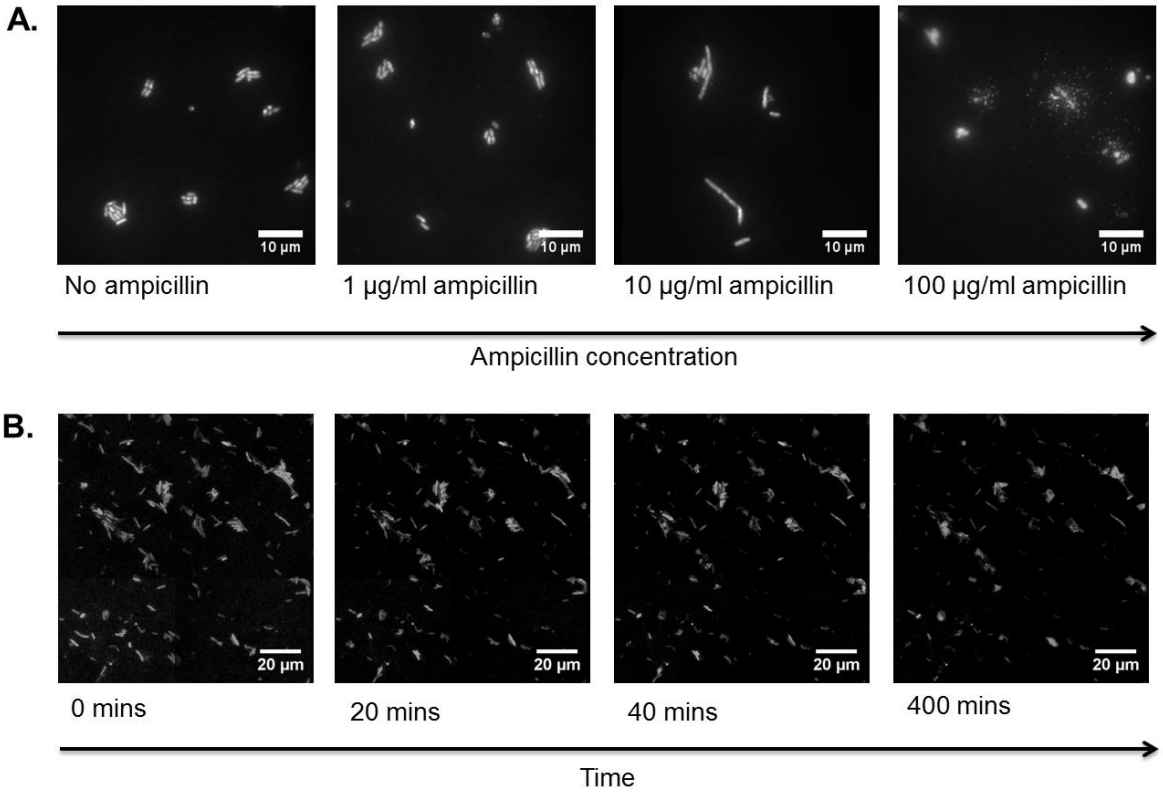


Figure 5.7 **Effect of ampicillin on *E. coli* during growth. Ampicillin inhibits cell wall formation, so during cell division osmotic pressure cannot be compensated and cells will 'burst'. A) Effect of ampicillin concentration. RLG221 strain was grown on agarose pads containing varying concentrations of ampicillin, at 37°C for four hours. As the concentration increased bacteria become stressed and elongated, before bursting at high concentrations. B) Effect of time. RLG221 strain was grown on agarose pads containing 0.1 mg/ml ampicillin overnight. Most bacteria burst within 1 hour, although some expand and eventually burst after around 10 hours.**

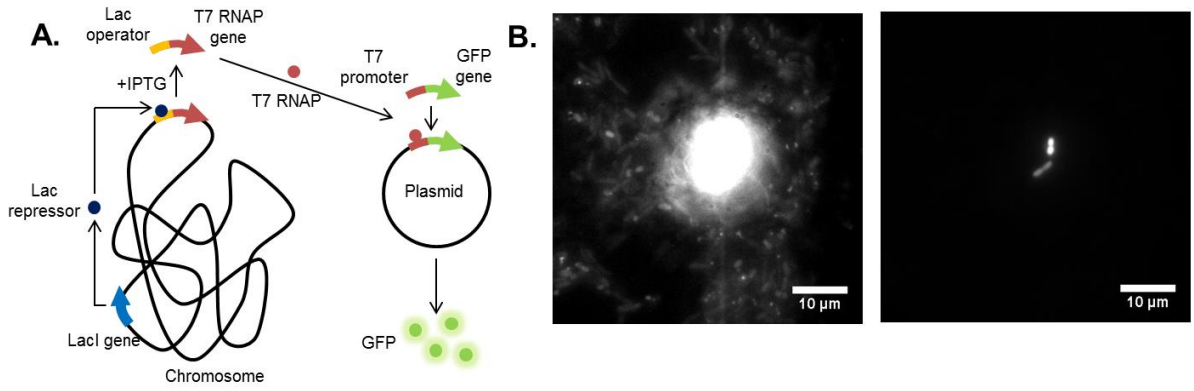


Figure 5.8 **Expression of fluorescent proteins in bacteria. A) Inducible expression system. Lac repressor is expressed from a gene on the bacterial chromosome and usually prevents expression of T7 RNA polymerase. When IPTG is added the Lac repressor falls off the DNA and T7 RNA polymerase is expressed. This is then free to express GFP, the gene for which is present on a plasmid, downstream of the strong T7 promoter. B) Expression of EGFP in *E. coli*. The same image is shown in different contrasts, to show the difference between bacteria that aren't expressing EGFP and those two in the middle of the image that are. Those that are expressing EGFP must contain a copy of the plasmid.**

To take advantage of these selection mechanisms two separate systems were used. The first used ampicillin selection and pUC19, a common, engineered, high copy number plasmid, the distribution of which has been studied previously (Figure 5.4 D-F)[115]. A map of pUC19 is shown in Figure 5.9A. It is 2.7 kbp in size, has a copy number of 40-250, depending on the strain, and contains an origin of replication (ori) and genes encoding β-lactamase (AmpR) and an N-terminal fragment of β-galactosidase (lacZα). The AmpR gene is for ampicillin selection, whilst the multiple cloning site is within the lacZα gene, so the activity of β-galactosidase reports on whether foreign DNA has been inserted. This activity is tested when transformed colonies are grown on IPTG and X-gal, since β-galactosidase will hydrolyse X-gal and colonies will appear blue when no insert is present.

pUC19 has four target sites for the TaqI RM system and can therefore be labelled with up to eight fluorophores by DNA methyltransferase M.TaqI, using the methyltransferase-directed labelling methods described in CHAPTER 2. Here pUC19 has been labelled with an amine AdoMet analogue, AdoHcy-amine, which was coupled pre-transalkylation to Atto 647N NHS ester. This has a maximum absorbance at 646 nm and maximum emission at 664 nm, which is well suited to detection within bacteria, where autofluorescence is a problem for excitation at below around 500 nm[212,213]. Single molecule counting results are shown in Figure 5.9B and representative images of labelled plasmids in Figure 5.9C. There are at least 2.6 fluorophores per plasmid on average, with only 2.8% of plasmids having no fluorophores visible. This means plasmids are bright and suitable for localisation experiments.

This was transformed into *E. coli* K-12 strain RLG221. K-12 is a common laboratory strain first isolated in 1922[214] which has been extensively studied and thousands of mutants produced. RLG221 is one such mutant that contains a deletion of the *lac* gene and the *recA* gene, to reduce plasmid recombination.

The second system studied exploits expression of a fluorescent protein for detection. Enhanced GFP (EGFP) is a mutant of the original GFP that exhibits 100-fold greater fluorescence intensity[215] and is now in widespread usage. Here the protein will be expressed from a pRSET B derivative, for which the key features are shown in Figure 5.10A. It is 3.6 kbp, contains a gene encoding β-lactamase (AmpR), and will be present in a high copy number due to the pUC19 derived origin of replication (ori). The F1 origin of replication (f1 ori) is phage derived and allows for replication of pRSET B into single-stranded DNA. The gene encoding EGFP has been inserted in the multiple cloning site,

downstream of the T7 promoter. This ensures that when T7 RNA polymerase is induced and the plasmid is present, EGFP will be expressed.

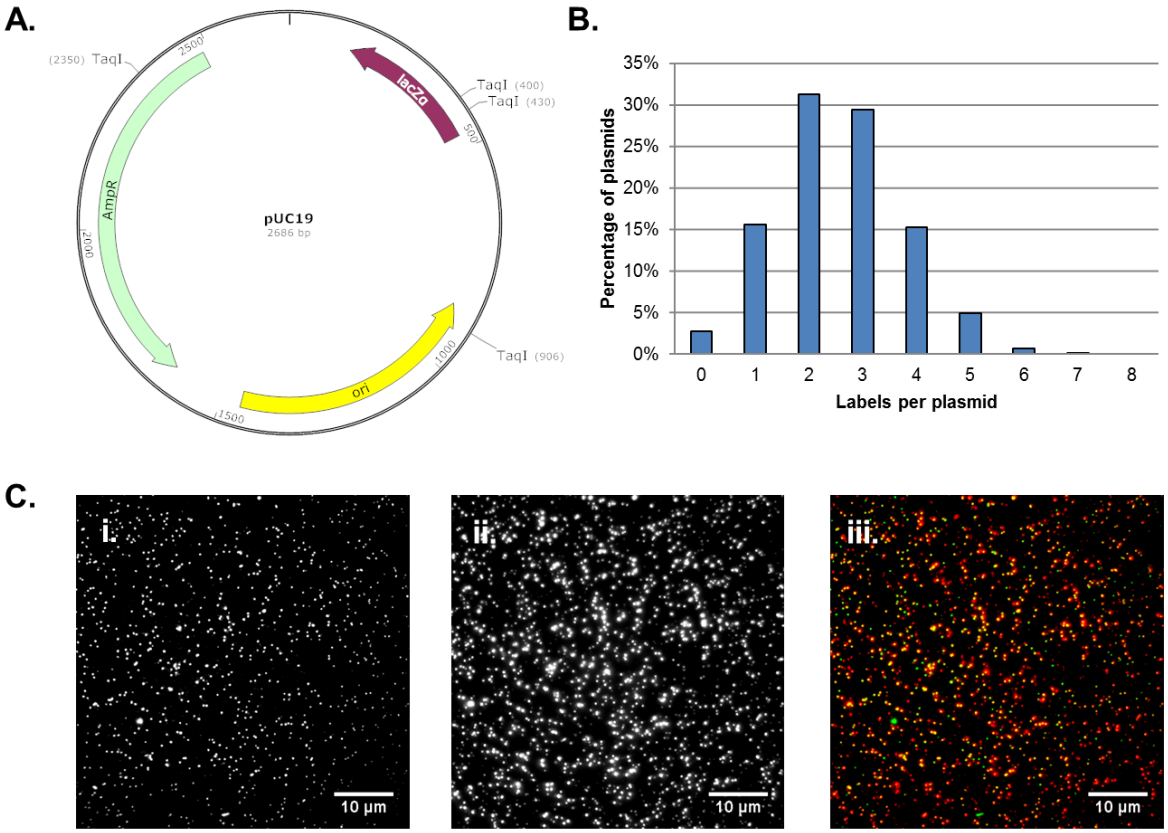


Figure 5.9 **pUC19 plasmid map and labelling. A) Map of pUC19. pUC19 is 2.7 kbp and contains an origin of replication (ori) and genes encoding β-lactamase (AmpR) and an N-terminal fragment of β-galactosidase (lacZα). It contains four TaqI sites and therefore can have a maximum of eight fluorescent labels. B) Single molecule counting results. pUC19 was labelled by M.TaqI with AdoHcy-amine, coupled pre-transalkylation to Atto 647N. There are 2.6 fluorophores per plasmid on average. C) Representative images of labelled pUC19. i) YOYO-1 labelled plasmids, ii) Atto 647N label, iii) overlay: YOYO-1 (green) and Atto 647N (red).**

pRSET B-EGFP has ten target sites for the TaqI RM enzymes (5'-TCGA-3'), however one site overlaps with the site for Dam methyltransferase (5'-GATC-3'), present in many *E. coli* strains, which may block labelling. pRSET B-EGFP was labelled by M.TaqI with Atto 647N, similarly to pUC19 and there was an average of at least 4.1 fluorophores per plasmid. Single molecule counting results are shown in Figure 5.10B and representative

images of labelled plasmids in Figure 5.10C. A restriction assay is shown in Supplementary Figure 7.24.

The strain used with this plasmid was T7 Express, purchased commercially from NEB. This is an enhanced BL21 derivative, in which T7 RNA polymerase is expressed from the chromosome under control of the Lac operator, making it a suitable strain for induced expression of EGFP (Figure 5.8A).

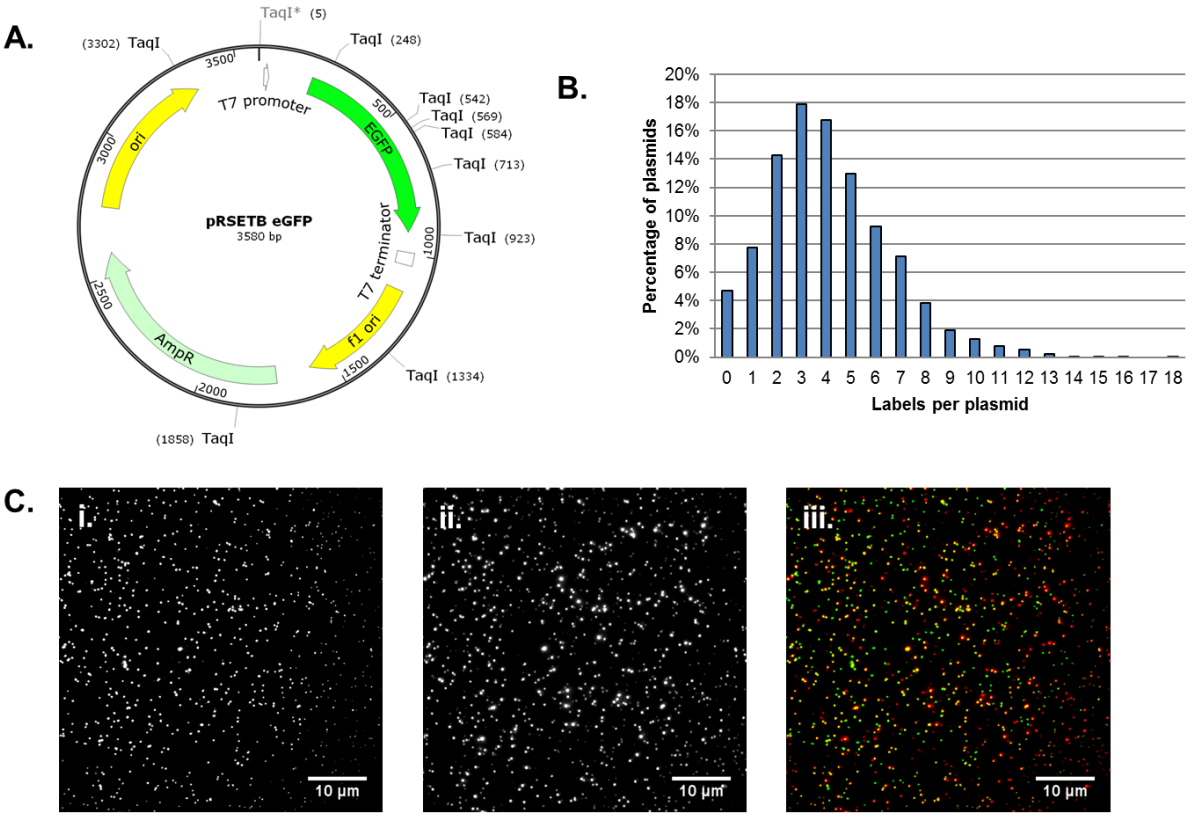


Figure 5.10 **pRSET B-EGFP plasmid map and labelling. A) Map of pRSET B-EGFP. pRSET B-EGFP is 3.6 kbp and contains two origins of replication (ori and f1 ori) and genes encoding β-lactamase (AmpR) and EGFP, under the control of a T7 promoter. It contains 10 TaqI sites, although one overlaps with Dam methylation (*) and therefore can have a maximum of eighteen or twenty fluorescent labels. B) Single molecule counting results. pRSET B-EGFP was labelled by M.TaqI with AdoHcy-amine, coupled pre-transalkylation to Atto 647N. There are 4.1 fluorophores per plasmid on average. C) Representative images of labelled pRSET B-EGFP. i) YOYO-1 labelled plasmids, ii) Atto 647N label, iii) overlay: YOYO-1 (green) and Atto 647N (red).**

5.2.2 Optimising transformations and imaging

Transformation efficiency was tested by a standard transformation protocol. Chemically competent cells were mixed with plasmid DNA, on ice, for 30 minutes. This was followed by heat shock at 42°C, a recovery period and growth overnight, at 37°C, on LB agar plates which contained ampicillin. Bacteria which have been transformed by the plasmid DNA and can replicate and transcribe the plasmid will divide and form visible colonies, whilst bacteria that haven't been transformed will be killed by the ampicillin. Both plasmid/strain systems were tested with labelled and un-labelled plasmids and typical results are shown in Figure 5.11. In both systems there is no significant difference between labelled and un-labelled plasmids, which was expected, as similarly covalently modified DNA has previously been shown to be transcribed and replicated in eukaryotic cells[204].

LB agar plates like these are unsuitable for fluorescence microscopy. The sheer size of them makes them unsuitable for most microscopes and they do not typically have a glass bottom. However more problematic is the background autofluorescence introduced by the rich media. This contains many compounds which will absorb visible light and make imaging problematic[213]. One solution is to adsorb bacteria to poly-L-lysine-coated slides and wash away the media. However there is evidence that using these highly charged surfaces, especially in thick layers, can affect the behaviour of bacteria[216–218]. Indeed such surfaces are known to have anti-bacterial properties[219] and sometimes *E. coli* will not grow on such surfaces.

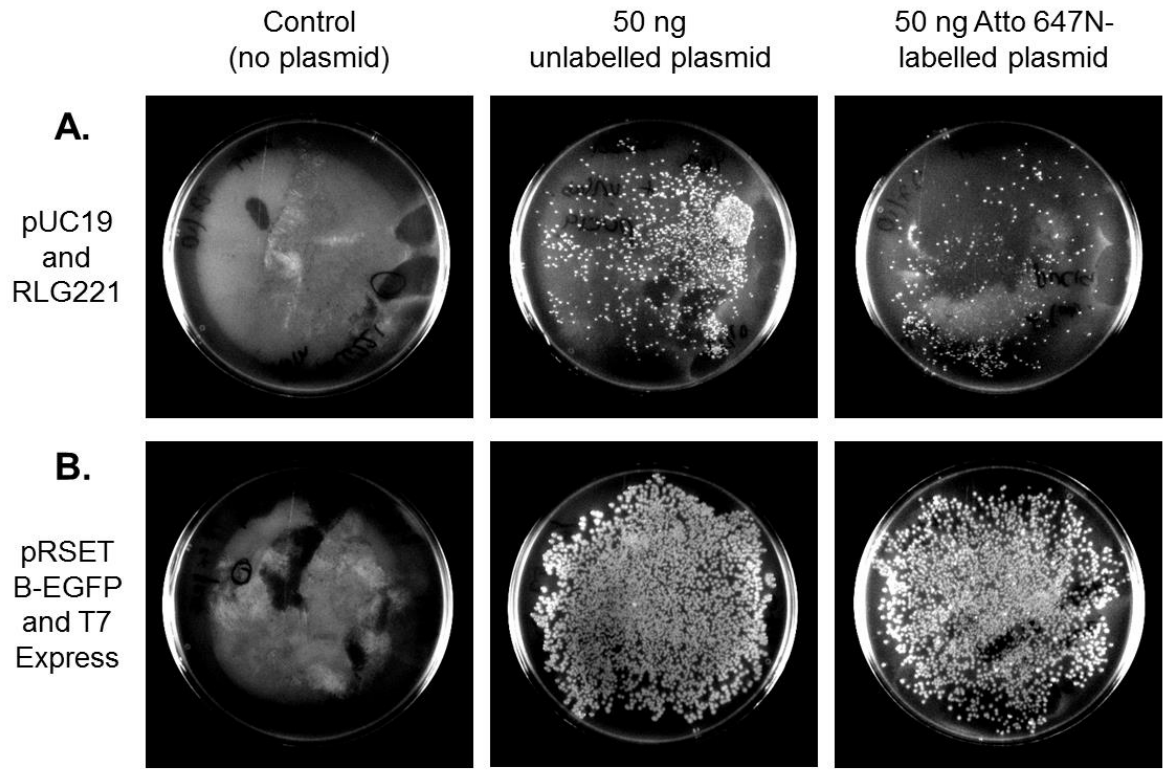


Figure 5.11 Transformation efficiency of labelled and un-labelled plasmids. Typical ampicillin-containing LB agar plates are shown for: A) pUC19/RLG221 and B) pRSET B-EGFP/T7 Express. When no plasmid is transformed no colonies are visible as ampicillin kills all bacteria. When Atto647N-labelled or unlabelled plasmid transforms bacteria colonies grow, as the plasmids contain a gene that expresses β-lactamase. There is no significant difference between labelled and un-labelled plasmids.

Therefore, an alternative method was used in which bacteria are immobilised in agarose pads, which are similar to LB agar plates but composed of a minimal media, to prevent background fluorescence. This method was based on that developed by Young et al[220], which is shown in Figure 5.12. In short, bacteria are loaded on small agarose pads and placed onto sealed glass-bottomed dishes for imaging. It is critical to prepare pads without scratches and to dry pads completely to prevent issues during imaging. Samples will drift as they dry, so sealing the dish is important to minimise this over long time-lapses.

Figure 5.12 Preparation of agarose pads for immobilising and imaging bacteria, taken from Young et al[220]. 1) Melted agarose is pipetted onto a coverslip and 2) sandwiched with another coverslip. 3) Smaller agarose pads are cut from the larger pad and 4) bacteria loaded on top. 5) After drying pads are flipped onto a glass-bottomed dish, 6) many can be loaded onto a single dish. 7) The dish is sealed with parafilm and 8) can be imaged.

Agarose pads can be used to test the difference in transformation efficiency of labelled and un-labelled plasmids in much the same way agar plates are used. Examples are shown in Figure 5.13 for labelled and un-labelled pUC19, transformed into T7 Express. This is used rather than RLG221 since it grows faster and at lower temperatures (particularly on minimal media), which is more suitable to prevent pads drying out over 24 hours. Individual colonies are clearly visible and again there is no significant difference between labelled and un-labelled plasmids.

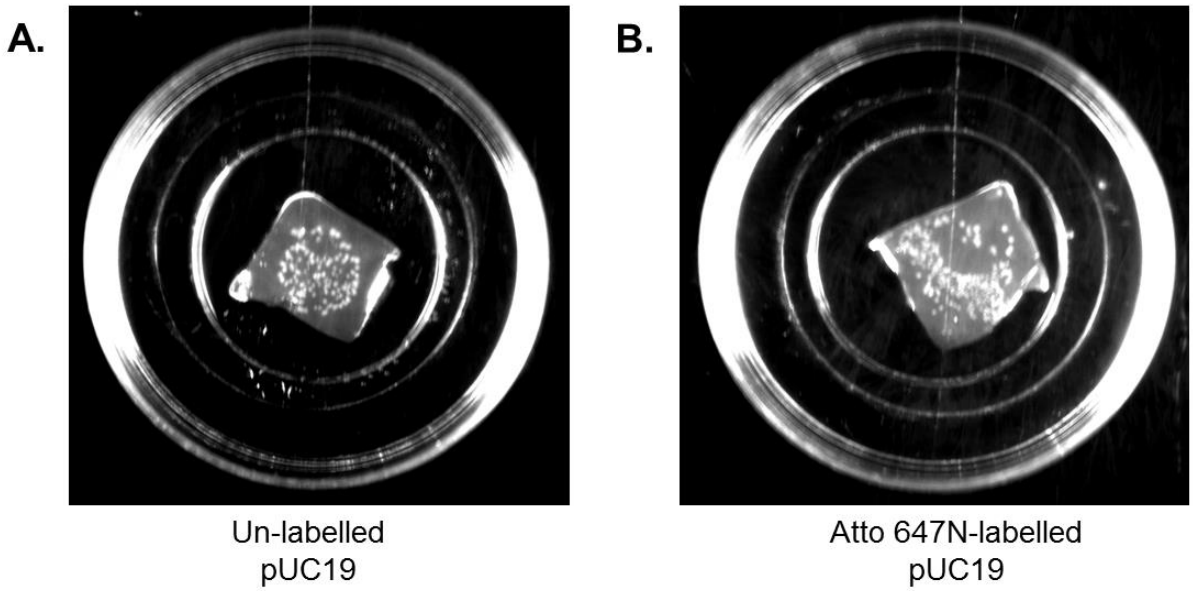


Figure 5.13 Transformation efficiency of labelled and un-labelled plasmids on agarose pads. Typical ampicillin-containing agarose pads are shown for: A) unlabelled pUC19/T7 Express and B) Atto 647N-labelled pUC19/T7 Express. 100ng of DNA was used and pads were incubated overnight at 30°C. There is no significant difference between labelled and un-labelled plasmids.

For imaging labelled plasmids, it is important to optimise the experimental conditions, to ensure a reasonable number of transformed bacteria are present. Too many bacteria on the surface can make imaging individual bacteria impossible, whilst if there are too few transformed bacteria on the pad then they will be impossible to find for imaging ("a needle in a haystack"). Therefore, the correct dilution of the bacterial suspension must be used to place on the pad (typically a 10-100x dilution of an aliquot of competent cells), but also the transformation efficiency must be optimised to make visualisation of the plasmid possible.

The transformation efficiency, for example, is dependent on the concentration of DNA used[211], which is typically between 100 pg and 100 ng for an aliquot of chemically-competent cells. Here the amount of DNA for transformations should be kept to a

minimum, since labelling large quantities of DNA consumes a large amount of cofactor, methyltransferase and dye. Figure 5.14 shows that 10-100 ng of plasmid DNA was sufficient to produce a large number of colonies.

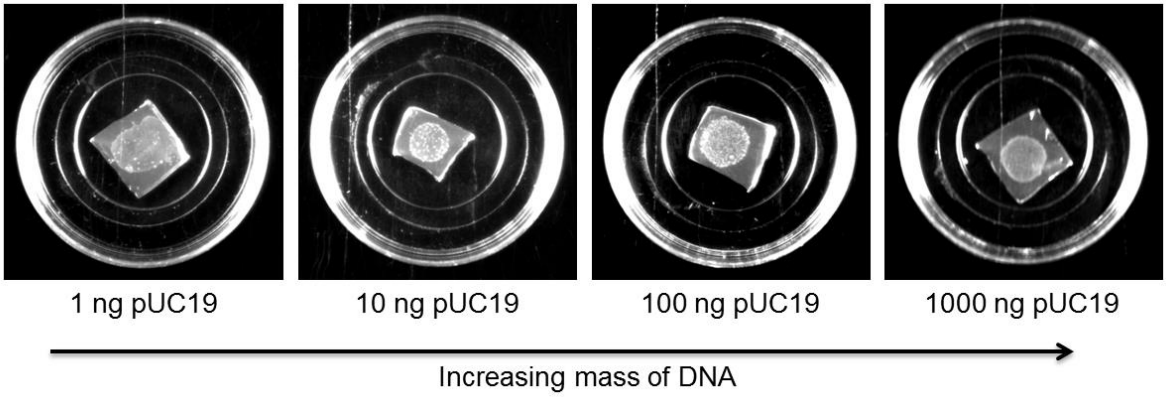


Figure 5.14 Effect of DNA concentration on transformation efficiency. T7 Express was transformed with varying amounts of unlabelled pUC19 and grown on agarose pads, overnight at 30°C. As the amount of DNA used to transform bacteria is increased so does the number of colonies. Between 10 and 100 ng of pUC19 gives a reasonable number of colonies.

Another important factor to optimise was the recovery period, for which results are shown in Figure 5.15. This is the period after heat-shocking but before plating the bacteria when the bacteria should be incubated at 37°C, to promote cell recovery after the heat-shock, but also to promote the expression of β-lactamase. No colonies are seen when there is no recovery period, possibly because the cell membrane will be destabilised during heat-shock, although relatively little is known about the mechanisms controlling this[221]. It is important to use the minimum recovery period possible, since as the bacteria divide the labelled plasmid will be diluted, i.e. will not be present in daughter cells. The buffer the bacteria are stored in (freeze-thaw buffer) was tested but found to give very few colonies for both plasmid/strain systems. Rich LB broth was suitable to give a reasonable number of colonies, within 20-40 minutes. Any longer and

too many divisions are possible, especially for the T7 Express strain. M9 minimal media gives similar results and is preferred, since this is the medium used for the agarose pads.

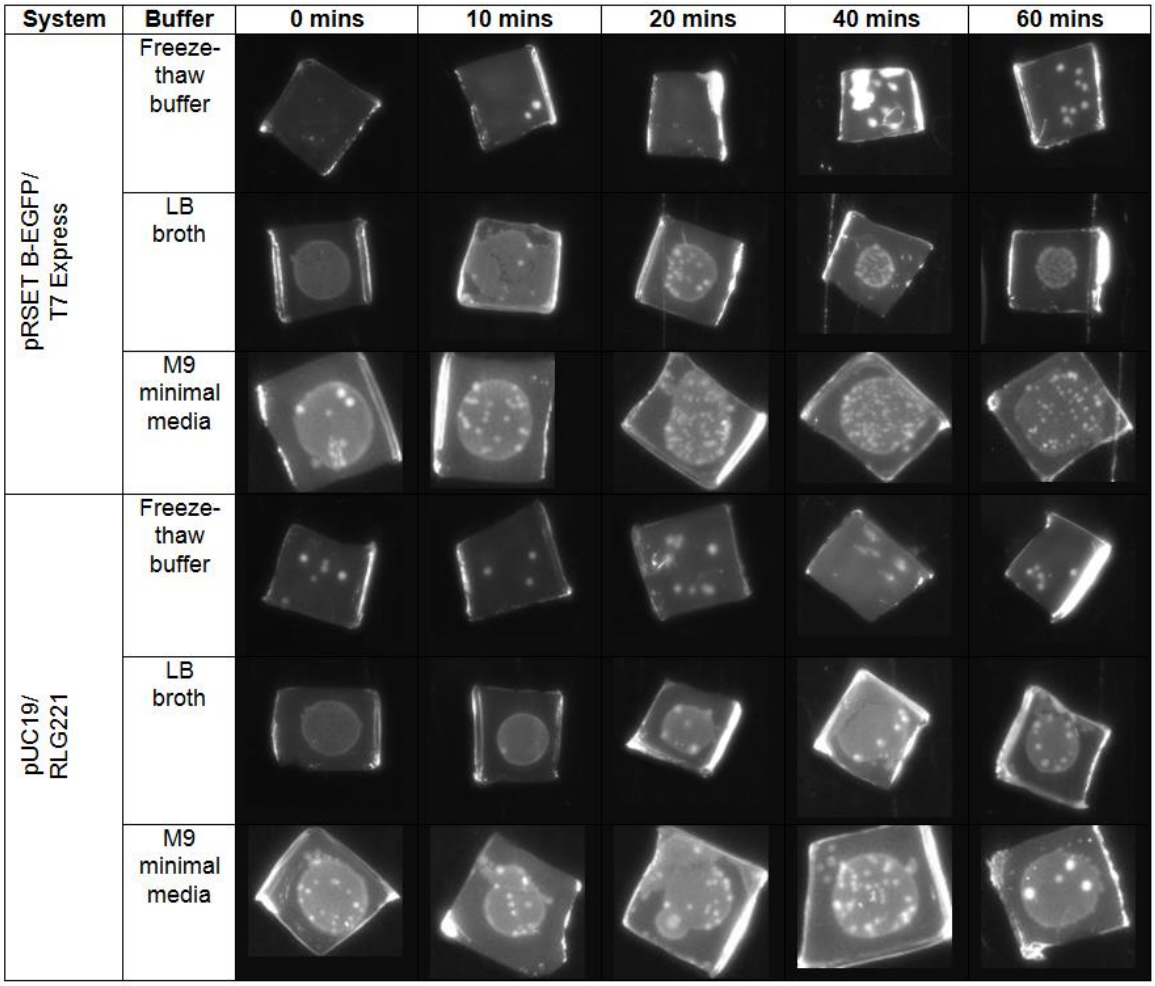


Figure 5.15 **Effect of recovery period and recovery buffer on transformation efficiency. T7 Express and RLG221 were transformed with 100ng of unlabelled pRSET B-EGFP or pUC19 respectively. This was followed by a recovery period, and growth on agarose pads, overnight at 30°C. The recovery period was 0, 10, 20, 40 or 60 minutes and was carried out in freeze-thaw buffer, LB broth or M9 minimal media at 37°C . Generally the longer the recovery period the higher the apparent transformation efficiency, however for long periods cell division is possible. Freeze-thaw buffer was a poor recovery media, whilst LB broth and M9 minimal media gave high apparent transformation efficiency.**

The number of washes was also considered, to ensure DNA that isn't transformed into cells is removed from the background. The final optimised procedure is given in the

methods but in summary, 50 µl aliquots of bacteria were transformed with 50 ng of plasmid DNA. After heat-shocking 0.5 ml of M9 minimal media was added and bacteria incubated at 37^{o}C for 20 minutes. Bacteria were pelleted and washed with 0.5 ml M9 minimal media, four times, before final resuspension in M9 minimal media (volume dependent on dilution required, typically 10-100x). 5 µl of bacterial suspension was placed on agarose pads, allowed to dry and sealed in a glass-bottomed dish for imaging.

5.2.3 pUC19 localisation and dynamics

The optimised procedure for transformations and imaging was applied over two different time periods: a short growth period and a long growth period. After a short growth period of around 1 hour at 30^{o}C, on agarose pads which contain ampicillin, transformed bacteria are visible. Transformed bacteria will be resistant to ampicillin and begin micro-colony growth, whilst bacteria which have no copies of the plasmid will lyse (Figure 5.7). Therefore, during imaging, after a short growth period, any small micro-colonies are indicative of bacteria that originally contained the labelled plasmid. T7 Express was used for these experiments, since it had a faster growth rate than RLG221, making identification of micro-colonies easier.

Results are shown in Figure 5.16A-C, for Atto647N-labelled pUC19, transformed into T7 Express and grown at 30^{o}C for 1 hour. Micro-colonies are clearly seen, although they were not straightforward to find (Figure 5.16 i). There are foci within the micro-colonies that are likely single plasmids (Figure 5.16 ii), however there are also a large number of plasmid or dye aggregates in the background. Washing the labelled plasmid away completely is difficult, even after multiple washes. The behaviour of the plasmids can help address whether they are single, transformed, plasmids or likely not. 15 second

time traces were taken of each micro-colony, allowing the number of bleaching steps for the plasmid to be compared to the expected number of labels (Figure 5.9). The results are consistent with well-labelled plasmids. The tracking of plasmids over time is also shown (Figure 5.16 iii), which allows the diffusion of plasmids to be investigated. For example, the plasmids in Figure 5.16A and C are less confined than most background plasmids, which behave more typically like the plasmid identified in Figure 5.16B.

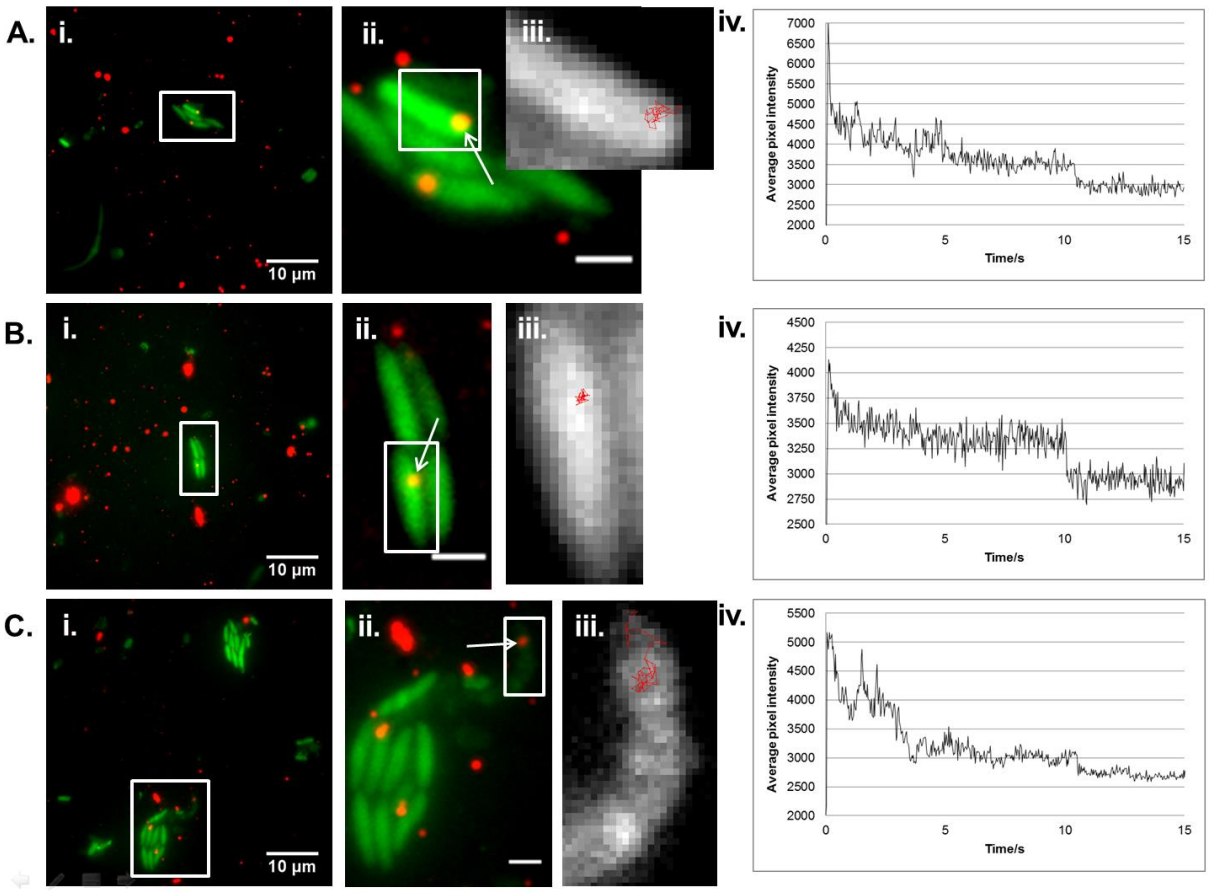


Figure 5.16 **Localisation and dynamics of pUC19 after short term growth. T7 Express (green) was transformed with Atto 647N-labelled pUC19 (red) and grown on agarose pads for 1 hour at 30°C. A-C) Examples of single plasmid molecules. i) Microcolony growth can be used to identify transformed bacteria. ii) Zoom of microcolonies highlighted by white box in i). Plasmids are indicated by a white arrow. iii) Single plasmid tracking (red), in region highlighted in white box in ii), over 15 seconds, each step is 0.6s. iv) Intensity profiles for plasmids highlighted by arrow in ii). The number of bleaching steps is consistent with plasmid labelling (Figure 5.9)**

Whether plasmids are inside bacteria or diffusing over the surface is still difficult to prove, although since these bacteria are transformed they must contain copies of the plasmid. *E. coli* cells are only around 1μm thick, so it was not possible experimentally to take a z-stack to laterally localise the plasmid.

An alternative experiment to this is to image bacteria over long time scales and in large scans. This allows plasmids to be followed during cell growth and division, for example whether plasmids are transmitted during micro-colony growth. An example of this is shown in Figure 5.17. Here RLG221 was transformed with Atto 647N-labelled pUC19 and imaged overnight at room temperature. The slow growth of RLG221 at room temperature allows for large scans to be taken (10x10 grid), since images can be taken every 5 minutes, without excessive growth. The growth of colonies is clear over the 9 hours of imaging. At the centre of each colony a single transformed bacterium is expected, however in most cases there was no overlapping signal from labelled plasmids with these bacteria.

An example of overlap between plasmid and a bacterium is shown in Figure 5.18 (region highlighted in Figure 5.17). Here there is a plasmid located between two dividing bacteria, which appears to be retained by a single bacterium after division. The plasmid can be followed over the course of 9 hours as the colony grows, and a time trace is shown. After initial drift of the sample (due to drying of the agarose pad), the plasmid moves a small amount for several hours, before moving rapidly across the image in the last hour of imaging, likely as the bacterium is moved within the colony as cells continue to divide (although the individual bacterium is no longer visible). Although there were no more examples of this type of behaviour, this time lapse highlights the advantage of

following single plasmids over long timescales. For instance, this technique could be used to follow resistance plasmids, to investigate transmission mechanisms.

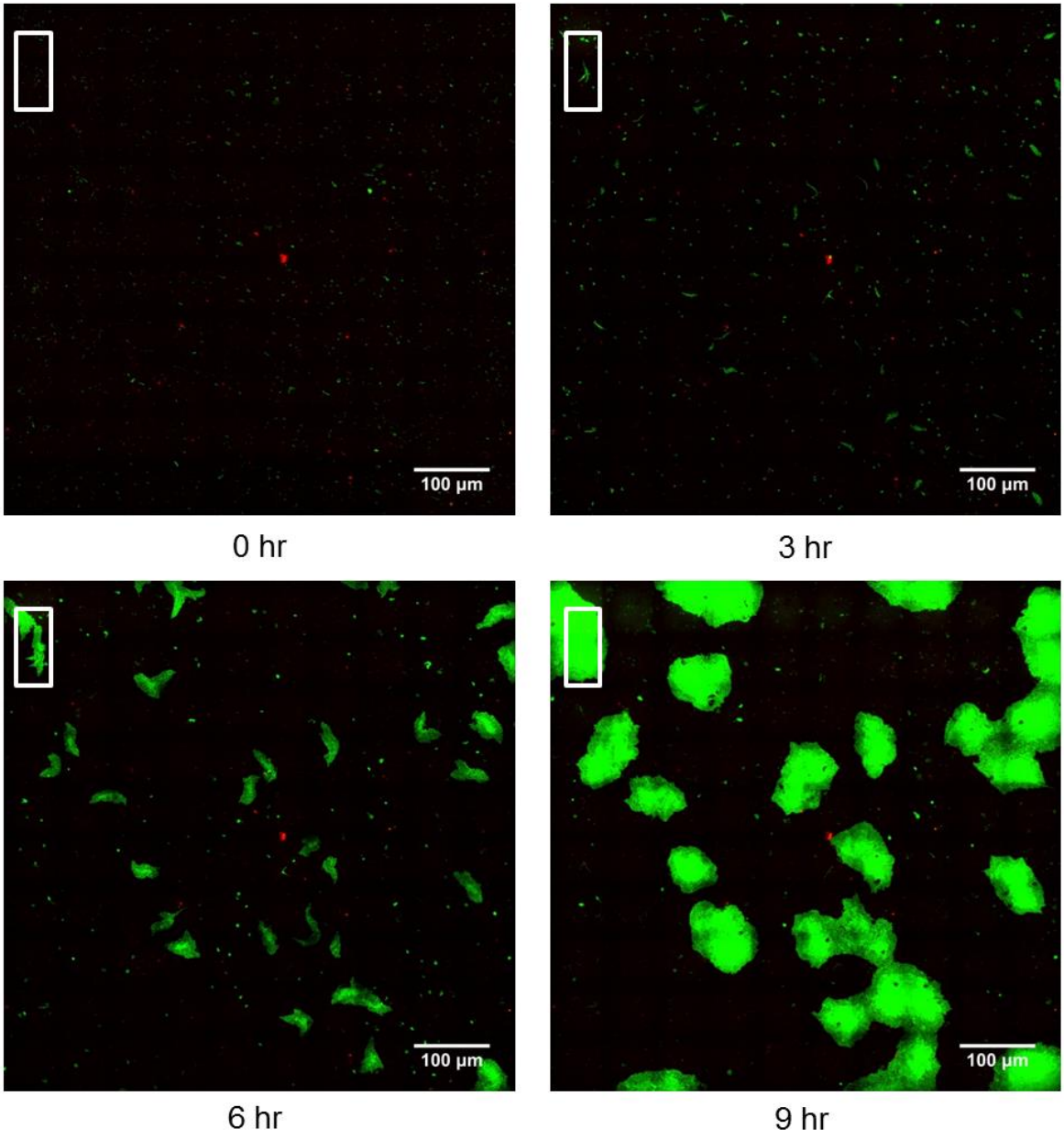


Figure 5.17 **Localisation and dynamics of pUC19 during long term growth. RLG221 (green) was transformed with Atto 647N-labelled pUC19 (red) and grown on agarose pads for 9 hours at room temperature. The growth of colonies is clearly visible. These can be traced back to an individual bacterium that was transformed. The white box highlighted is shown in Figure 5.18.**

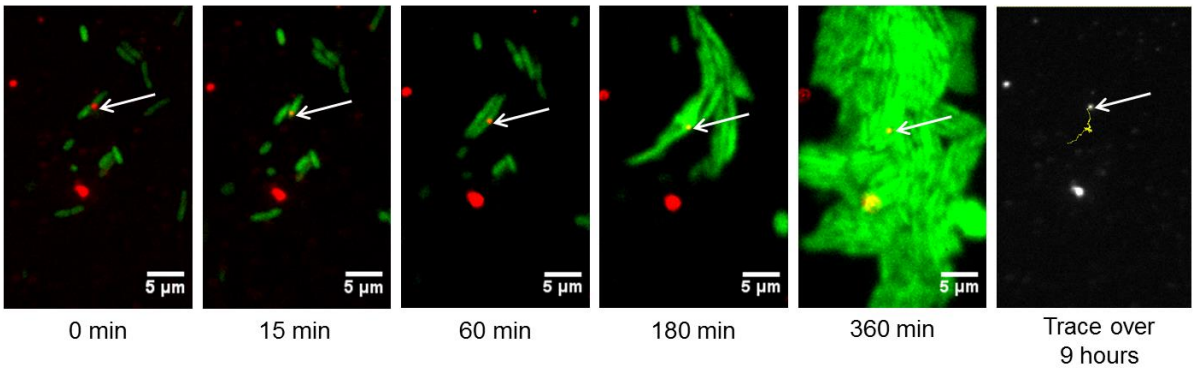


Figure 5.18 Localisation and dynamics of pUC19 during long term growth, zoom. RLG221 (green) was transformed with Atto 647N-labelled pUC19 (red) and grown on agarose pads for 9 hours at room temperature. This is a zoom of a section in Figure 5.17. A single transformed bacterium grows into a colony over 9 hours. Highlighted with an arrow is a single plasmid. A trace of the location of the plasmid is also shown.

The main drawback with identifying transformants by ampicillin resistance alone is that identifying micro-colonies is difficult. Large areas of the agarose pad must be scanned, but identification is difficult to automate since micro-colonies are not well-defined against the background of dead bacteria. Few examples could be imaged and there were many examples of micro-colonies with no labelled plasmids visible. Automated methods for extracting bacteria which overlap with plasmids will be discussed later in this chapter.

5.2.4 pRSET-B localisation and dynamics

One method to overcome some of the experimental challenges, without resorting to automation, is to use expression of a fluorescent protein, as well as ampicillin resistance. This can be used again after a short growth period, for example 1 hour at 30°C. Bacteria transformed with pRSET B-EGFP will not only be resistant to ampicillin and grow microcolonies as before but will also begin to express EGFP if expression is induced by

addition of IPTG to the agarose pads. Bacteria will effectively 'light-up' and make searching for transformants far more straightforward.

Results are shown in Figure 5.19, for T7 Express transformed with Atto 647N-labelled pRSET B-EGFP. Transformants are now easily visualised and short time lapses (~15 seconds) can be recorded to follow plasmid diffusion and bleaching. Examples of likely transformed plasmids are shown in Figure 5.19A and B. These have the correct number of bleaching steps for single, well-labelled plasmids (see Figure 5.10) and clearly diffuse over several pixels across the course of the time lapse. This contrasts with more confined plasmids, which are generally in the background, an example that overlaps an EGFP-expressing bacterium is shown in Figure 5.19C. The difference in behaviour between confined and more freely diffusing plasmids lends further evidence that the plasmids visualised in Figure 5.19A and B are indeed transformed.

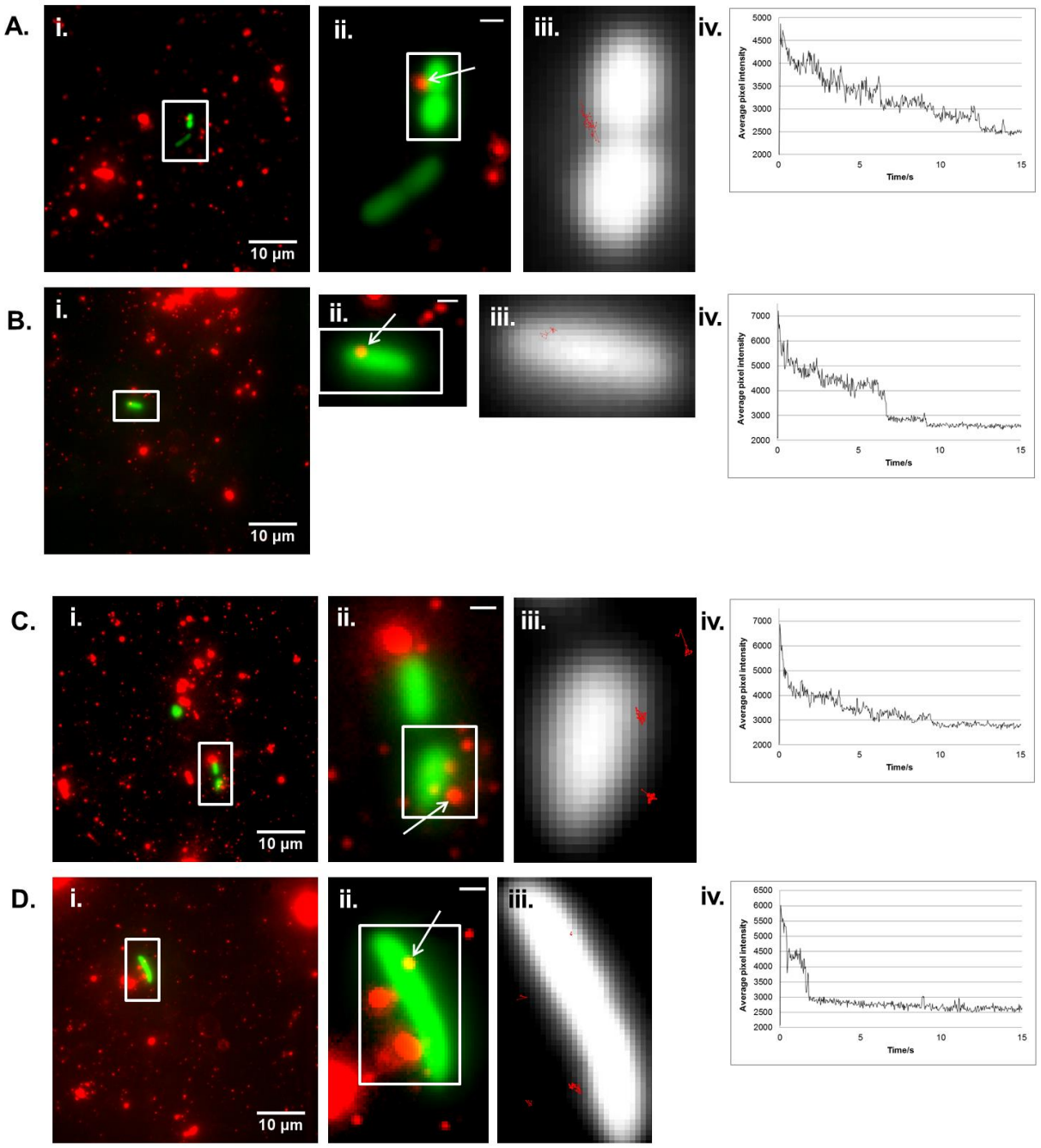


Figure 5.19 **Localisation and dynamics of pRSET B-EGFP after short term growth. T7 Express (green) was transformed with Atto 647N-labelled pRSET B-EGFP (red) and grown on agarose pads (with IPTG) for 1 hour at 30°C. A-D) Examples of single plasmid molecules. i) Expression of EGFP can be used to identify transformed bacteria. ii) Zoom of bacteria highlighted by white box in i). Plasmids are indicated by a white arrow. iii) Single plasmid tracking (red), in region highlighted in white box in ii), over 15 seconds, each step is 0.6s. iv) Intensity profiles for plasmids highlighted by arrow in ii). The number of bleaching steps is consistent with plasmid labelling (Figure 5.10).**

Overnight time lapses can also be used for this plasmid/strain system. An example is shown in Figure 5.20, for T7 Express incubated at room temperature, overnight, on agarose pads containing ampicillin and IPTG. There is no growth of microcolonies seen, in contrast to results with pUC19, and some bacteria that express EGFP die, though others continue to get brighter as EGFP expression is continuing. It seems likely that EGFP expression is preventing normal growth. Normally during protein expression bacteria are grown to maximum density, before induction of the protein, in other words cell division is not necessary, so the effect of protein expression on cell growth is not normally an issue.

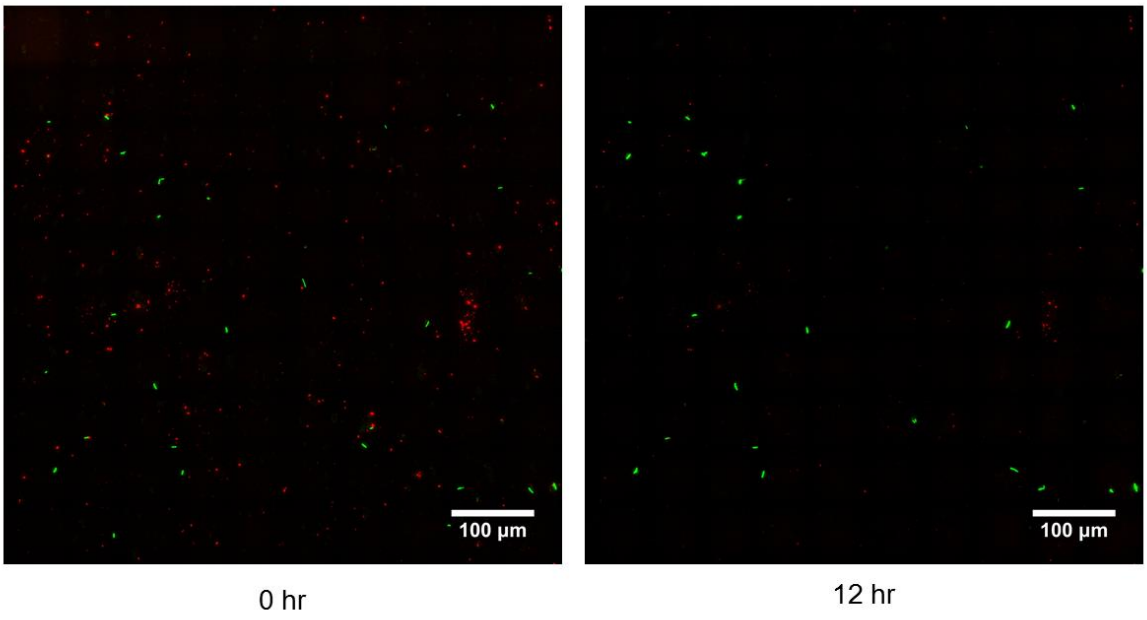


Figure 5.20 **Localisation and dynamics of pRSET B-EGFP during long term growth. T7 Express (green) was transformed with Atto 647N-labelled pRSET B-EGFP (red) and grown on agarose pads (with IPTG) for 12 hours at room temperature. Individual transformants are clearly visible by EGFP expression, however microcolonies do not grow, likely due to the stress of high levels of protein production.**

When IPTG is removed from pads there is still leaky expression of EGFP, allowing transformants to be identified, however microcolonies now grow allowing for plasmid dynamics to be followed as before. An example is shown in Supplementary Figure 7.25. The main issue remains however, that very few bacteria, which apparently contain a copy of the plasmid, have overlapping signal from labelled plasmids. Possible explanations for this will be discussed in the conclusion to this chapter.

5.2.5 Image segmentation

If many transformants (with overlapping signal from labelled plasmids) were present, then another difficulty would present itself. Imaging individual bacteria would be possible (as has been shown here e.g. Figure 5.19), but statistics to describe the behaviour of plasmids in the whole bacterial population would be difficult to obtain. High-throughput imaging analysis is necessary to rapidly identify transformants and to obtain statistics for the whole population. To allow this custom Matlab code has been used to segment bacteria and visualise those of interest (e.g. with overlapping signal from labelled plasmids).

Image segmentation is a classic image processing problem, for which many different strategies exist[222]. The most effective strategy is one that will identify and segment the objects of interest within an image, with the highest possible accuracy. However, there is no general solution to image segmentation problems, so in practice the experimental images must be carefully considered, and custom code is often used to generate the best practical solution.

For segmentation of *E. coli* on agarose pads there are a number of software packages available[220,223–225], however these are generally unsuited to the images obtained here,

since they are applied to brightfield images rather than using the autofluorescence of the bacteria. This means that the contrast in the images is generally better, particularly the edges between bacteria within clumps or microcolonies, which makes for an easier image segmentation problem.

An example region from a large scan is shown in Figure 5.21A. Note how the contrast between individual bacteria or clumps of bacteria, and the background is generally good. This means a simple threshold can be used to segment bacteria from the background, seen in Figure 5.21B. The size of regions identified by a threshold can then be used to identify individual bacteria, for example small regions (e.g. <40 pixels in area) can be removed (Figure 5.21C) and remaining regions can be separated between individual bacteria (e.g. <120 pixels in area, Figure 5.21D) and clumps (e.g. >=120 pixels in area, Figure 5.21E). The final results for this first stage of segmentation are shown in Figure 5.21F.

The next stage is to detect individual bacteria within clumps (shown in Figure 5.22A), first the contrast between bacteria must be improved. For this a contrast filter was used, the "magic contrast filter" used by the SuperSegger software package[225]. At each pixel the minimum intensity of pixels within a radius of interest (e.g. 3 pixels) is subtracted from the pixel. This will mean pixels located at the edges of bacteria, but within clumps, will be darker and increase the contrast between bacteria (Figure 5.22B).

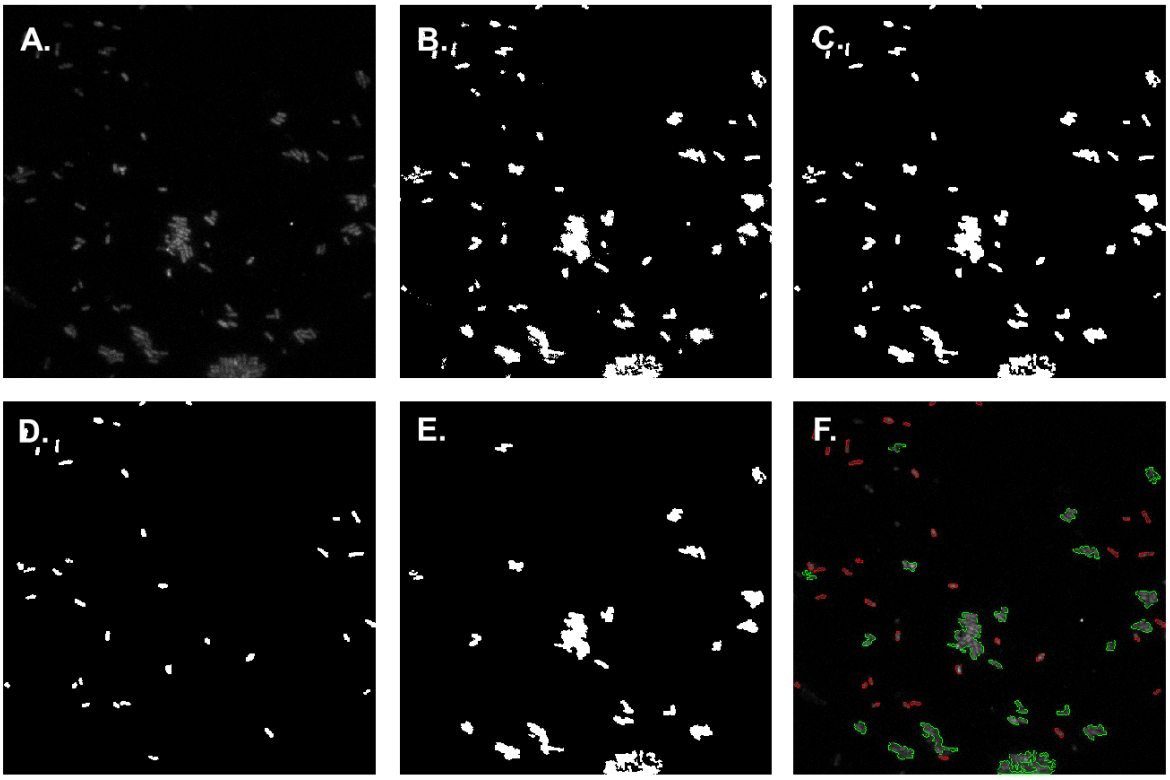


Figure 5.21 **Image segmentation procedure. A) Typical raw image of *E. coli* from a large scan of an agarose pad. Individual bacteria and clumps of bacteria are both visible, but contrast is low. B) A thresold can be used to segment most bacteria from the background. C) Small regions (e.g. <40 pixels) are removed from the threshold image. D) Medium regions (e.g. 40<pixels<120) are segmented as individual bacteria. E) Large regions (e.g. >120 pixels) are segmented as clumps of bacteria. F) Overlay showing segmentation of image into indivudal bacteria (red) and clumps (green).**

This can be used to generate a new threshold image, now of the clumps of bacteria, which should largely separate individual bacteria (Figure 5.22C). After image dilations and erosions and selection by size again many individual bacteria can be successfully segmented (Figure 5.22D). The segmentation is not perfect, particularly in clumps, however this is sufficient to identify large numbers of bacteria and therefore should be sufficient to generate good population statistics. The raw microscopy image is shown in Figure 5.22E and the segmentation and overlap with plasmids (selected by thresholding and size) is shown in Figure 5.22F.

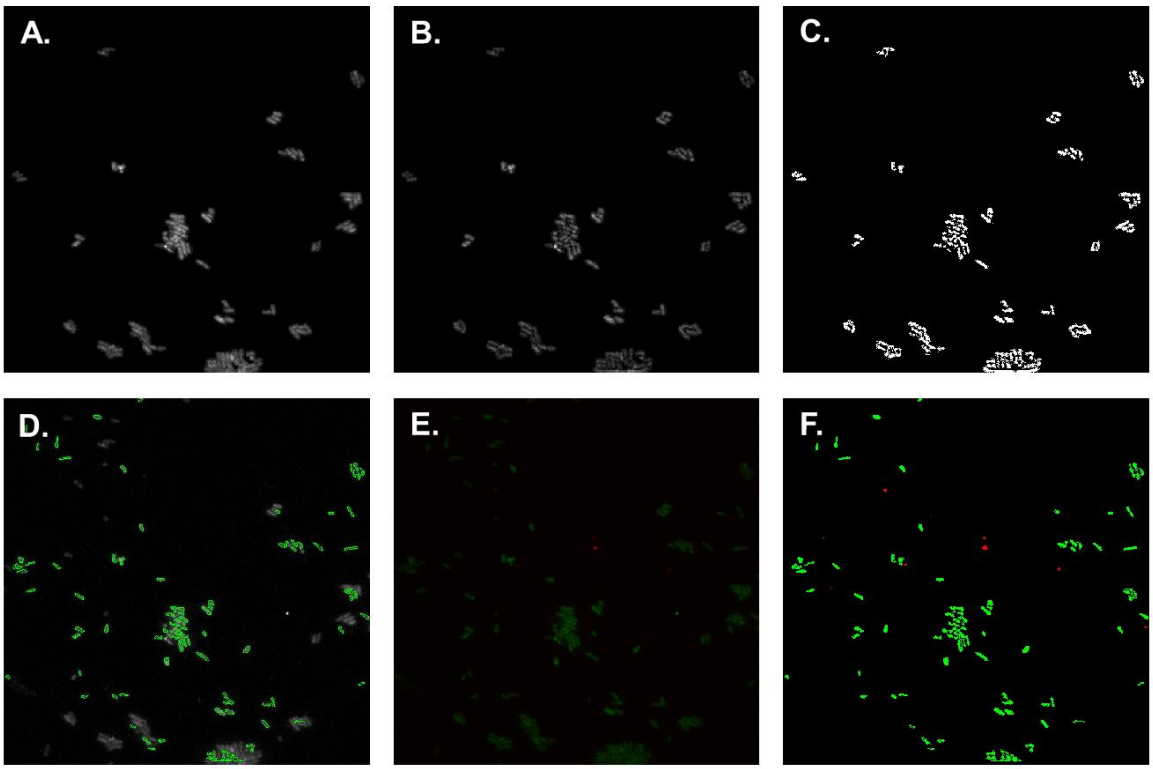


Figure 5.22 **Image segmentation procedure for segmentation of clumps of bacteria. A) Clumps segmented from Figure 5.21. B) A contrast filter is used to increase the edge contrast between bacteria. C) Thresholding can be used to identify bacteria. D) Regions which contain a suitable number of pixels can be identified as individual bactera. E) Original overlay of bacteria (green) and plasmids (red). F) Overlay of segmented bacteria (green) and segmented plasmids (red).**

When this segmentation is carried out on hundreds of images from large scans, simple statistics can be generated, for example the number of bacteria on the surface and the number with overlapping plasmids. For example, for the RLG221/pUC19 image in Figure 5.17 there are 982 individual bacteria found, of which 99 overlap with labelled plasmids. A montage of these is shown in Figure 5.23. In this each overlap between an individual bacterium and a plasmid is displayed, so if there are multiple bacteria for a single plasmid, or vice versa, the same regions may be shown several times, centred on the plasmid in each example.

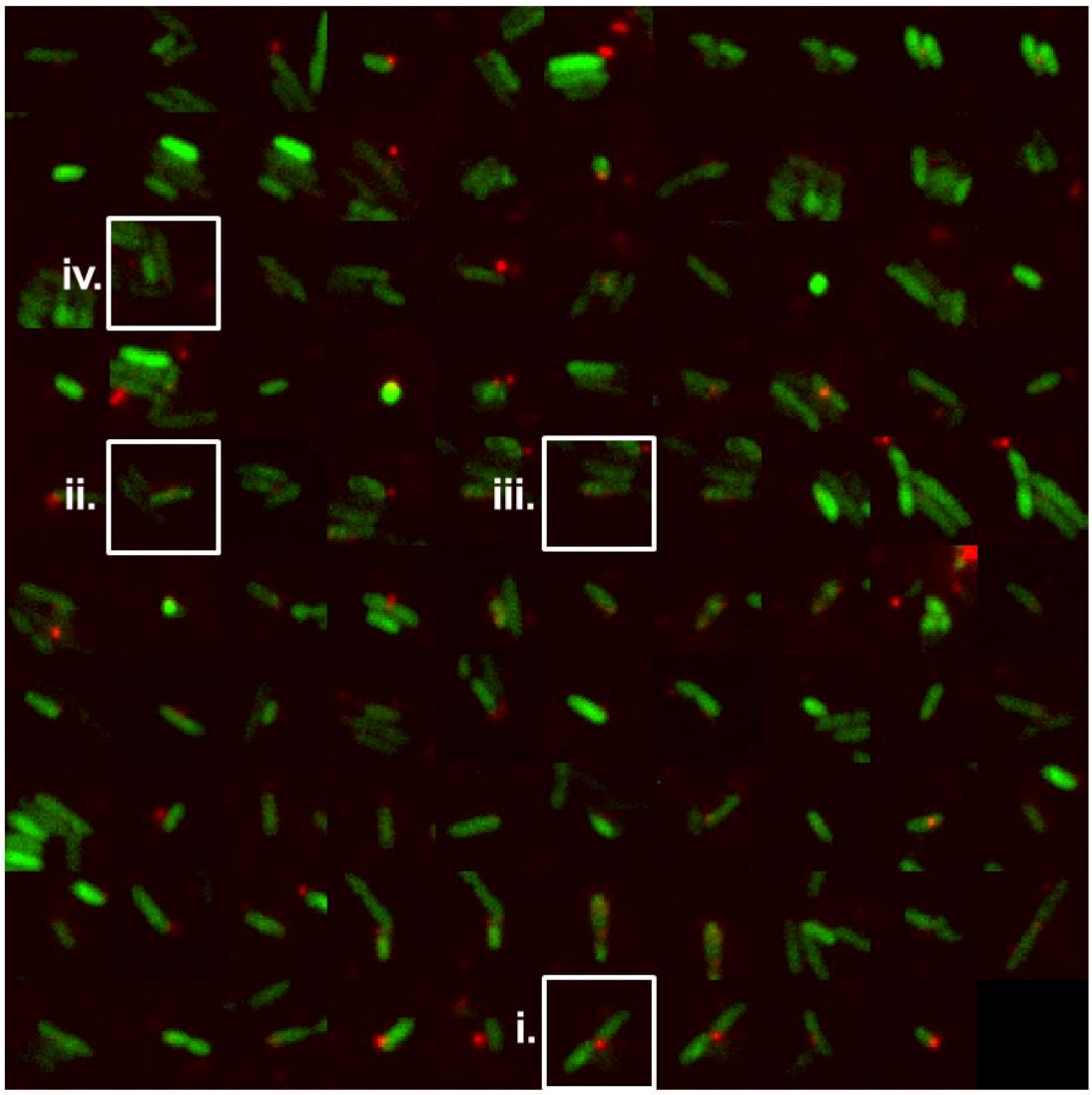


Figure 5.23 **Bacteria of interest automatically identified by image segmentation. Regions of overlap between bacteria (green) and plamsids (red) are identified by image segmentation (Figure 5.22F). Areas are around each plasmid are extracted. Highlighted regions i-iv are shown in Figure 5.23.**

Furthermore, this montage can be used to follow plasmids throughout the whole time-lapse. Four regions from Figure 5.23 have been selected to highlight this. The first region (i) is the same region that was identified in Figure 5.18, but here it has been automatically extracted. Other behaviours are also shown, including: bacteria that grow initially but then stop (ii); bacteria that do not grow normally (iii); and bacteria that divide but do not appear to initially contain the plasmid (iv).

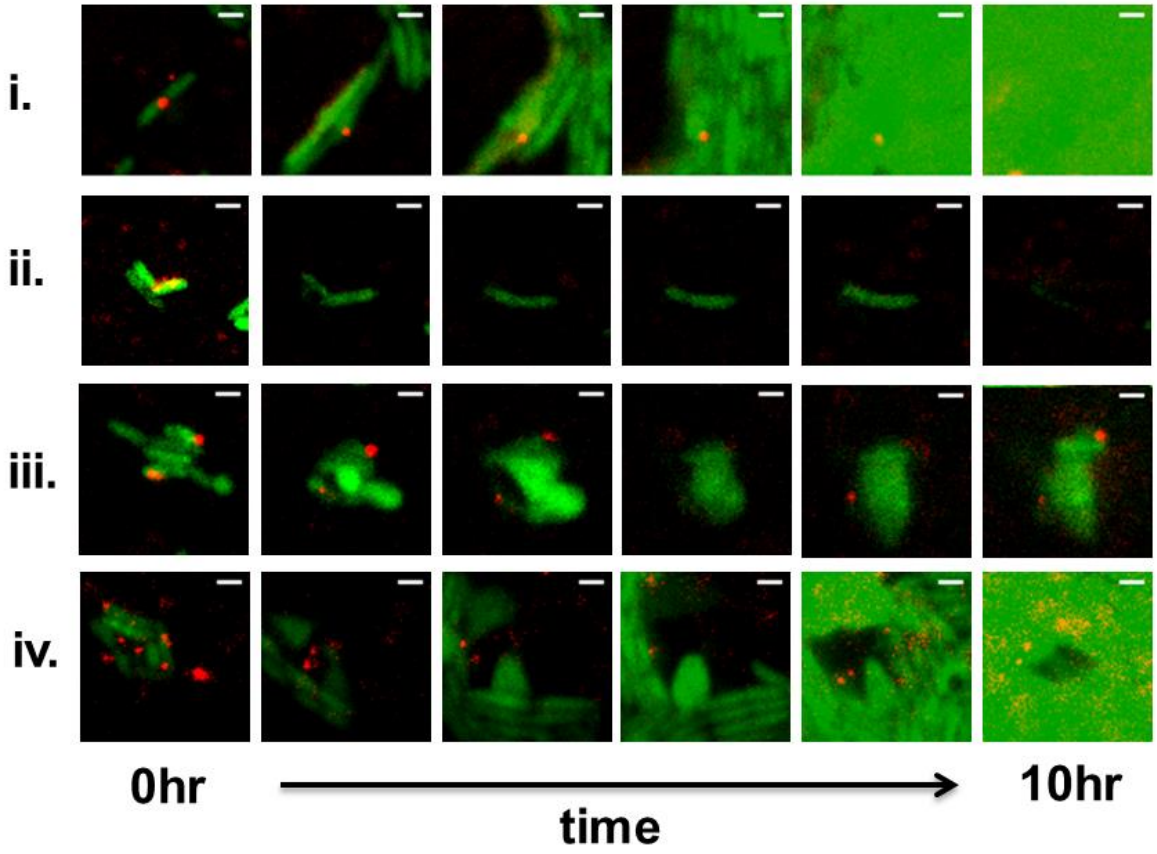


Figure 5.24 **Time lapses of regions of interest identified in Figure 5.23. Regions are automatically identified, and time lapses extracted for each. i) Region identified in Figure 5.18, showing colony growth from a transformant. ii) Bacterium that grow initially but then stops. iii) Bacterium that does not grow normally. iv) Bacterium that divides but does not appear to initially contain the plasmid.**

These examples show how powerful this type of image analysis can be for automatically selecting features of interest, for further investigation and the generation of statistics for the population as a whole. However, here they have only been used sparingly, since experimentally few examples of overlapping signal from bacteria and plasmid were seen.

5.3 Conclusion

Several methods have been developed which should prove valuable for investigating the localisation and dynamics of single plasmids in bacteria. The preliminary results presented are also the first known example of the transformation and imaging of bacteria using plasmids, which have been fluorescently-labelled, covalently, *in vitro* (previous studies have either used eukaryotic cells or small oligonucleotides[221]). Using methyltransferase-directed labelling, plasmids can be fluorescently labelled, with controlled densities, at specific sites. Labelling efficiencies of around 30% are typical and only a small proportion of plasmids have no fluorescent labels visible (<5%).

Labelled plasmids can be used for transformation of *E. coli* strains and the labelling does not affect transformation efficiency. Similar results have been observed previously for non-specifically labelled plasmids, transfected into mouse liver cells[204], in which there was minimal difference in gene expression between labelled and un-labelled plasmids. Either labels have no effect on transcription (i.e. the RNA polymerase can read DNA regardless of the covalent labels) or alternatively repair of modified bases could be taking place (i.e. nuclear excision repair).

In these experiments individual transformants were identified by either resistance to ampicillin and/or expression of EGFP. However, in general, very few of the transformed bacteria had overlapping fluorescent foci that could be attributed to labelled plasmids. Since there is no difference in transformation efficiency, but at least 95% of plasmids are fluorescently labelled, then most transformants should have been originally transformed by labelled plasmids. Cell division has not occurred more than 1-2 times before placing

bacteria on agarose pads, due to the short recovery period, so this also does not explain this observation.

Atto 647N is unlikely to be quenched, (Atto 647N-labelled oligonucleotides have been imaged in cells[221]), and the amide coupling is also unlikely to be hydrolysed. However, to test the factors TAMRA-labelled plasmids, using DBCO coupling were also tested, but showed the same behaviour (results not shown). Also, higher labelling densities using M.MpeI-labelled plasmids did not yield more fluorescent foci (results not shown).

Therefore it seems likely that the fluorescent label is being removed, for instance by nucleotide excision repair [226]. In *E. coli* repair is carried out by the UvrABC system, which recognises a wide range of DNA damage, including alkylation. Therefore, for future work it is advised that strains containing a knock-out of this system should be used, since although they may be 'sick', it is expected that the fluorescent labels would be retained, and single plasmids could be visualised.

5.4 Materials and Methods

5.4.1 Plasmid labelling and single molecule counting (Figure 5.9 and Figure 5.10)

For plasmid labelling, a 10 µl solution containing 2x PBS, 3 mM AdoHcy-amine and 10 mM Atto647N NHS Ester is incubated at 4°C for 1 hour. Next a 40 µl solution containing 1x MES CutSmart, pH 5.75, 1 µg pUC19 (NEB) or pRSET B-EGFP (engineered by Muhammed Rassul), 0.6 µg M.TaqI and 4 µl of the AdoHcy-Atto647N mixture is incubated at 50°C for 1 hour. 1 µl 20 mg/ml proteinase K is added and incubated at 50°C for 1 hour, before purification by GenElute PCR Clean-up kit (Sigma-Aldrich) and elution into 50 µl 1xTE (Sigma-Aldrich).

For imaging, a 50µl mixture of 50% DMSO, 0.5xTE, 0.2 µM YOYO-1 and ~5 ng DNA was incubated at 37°C for 30 minutes. 450 µl 1xTE was added and 100 µl was placed on poly-l-lysine coated coverslips for ~30 seconds. Subsequently, the sample was washed with 3 ml molecular grade water and dried. Samples were imaged using an Olympus IX81 inverted widefield/epifluorescent/TIRF microscope equipped with 491, 560, 640 lasers, and a Hamamatsu CCD camera (Orca R2). All images were analysed using the Localizer plugin for IgorPro[94] and custom Matlab software[141].

5.4.2 Preparation of competent cells

Strains: RLG221, *E. coli* K-12 Δlac ΔrecA, R. Gourse, supplied by Rita Godfrey. T7 Express, enhanced BL21 derivative, New England Biolabs.

Strains were grown in 20 ml LB broth overnight at 37°C, shaking. 50 ml of fresh LB broth was inoculated with 0.5 ml of overnight culture and grown at 37°C, shaking, until mid-exponential phase (OD_{650} = 0.3-0.5). Cells were harvested by centrifugation, ~3700g, for

15 minutes, at 4°C and the cell pellet resuspended in 20 ml ice cold 50 mM calcium chloride and left on ice for 20 minutes. Cells were harvested by centrifugation (~3700g, 15 minutes, 4°C) and the cell pellet resuspended in 5 ml ice cold freeze-thaw buffer (100 mM calcium chloride, 15% glycerol) and left on ice overnight. 50 µl aliquots were stored at -80°C.

5.4.3 Transformations on agar plates

50µl aliquots of competent cells were thawed on ice. 10-100 ng of DNA was added, and cells incubated on ice for 30 minutes. Cells were heat-shocked at 42°C for 2 minutes, then placed back on ice for 2 minutes. 0.5 ml LB broth was added, and cells incubated at 37°C for 1 hour. Cells were harvested by centrifugation (>6000g, 2 minutes) and resuspended in 50 µl LB broth. Culture was spread on agar plates, containing 100 µg/ml ampicillin (Sigma-Aldrich) and grown overnight at 37°C.

5.4.4 Preparation of agarose pads

50 ml M9 minimal media: mix 25 ml 2x M9 salts (Thermofisher), 1 ml 20% glucose, 1 ml 0.1 M magnesium sulfate, 0.1 ml 50 mM calcium chloride, 0.5 ml 10 mg/ml thiamine hydrochloride, 0.5 ml 10 mg/ml biotin, 2.5 ml 20 mg/ml casamino acids, 50 µl trace elements (Sigma-Aldrich) and top up with molecular grade water.

0.15 g low melting point agarose was mixed in 10 ml M9 minimal media and dissolved by microwave. 0.1 ml 100 mg/ml ampicillin and 0.1 ml 100 mM IPTG (as necessary) was added to agarose mix. 1 ml of solution was pipetted onto glass coverslip, sandwiched by another coverslip and left to set. Small pads (approx. 0.5x0.5 cm) were cut out.

5.4.5 Typical transformation onto agarose pads

50 µl aliquots of competent cells were thawed on ice. 100 ng of unlabelled plasmid was added, and cells incubated on ice for 30 minutes. Cells were heat-shocked at 42°C for 2 minutes, then placed back on ice for 2 minutes. Nothing 0.5 ml M9 minimal media was added and cells incubated at 37°C for 20 minutes. Cells were harvested by centrifugation (>6000g, 2 minutes) and washed with 0.5 ml M9 minimal media 3 times. The final resuspension was in 0.5 ml M9 minimal media. 5 µl of suspension was placed on agarose pads and allowed to dry for 15 minutes. Agarose pads were place into glass-bottomed dishes, sealed with parafilm and incubated at 30°C.

For short term growth samples were incubated for 1 hour and imaged using an Olympus IX81 inverted widefield/epifluorescent/TIRF microscope equipped with 491, 560, 640 lasers, and two Hamamatsu CCD cameras (Orca R2 and high-speed ImageEm). Images were analysed using Fiji and the Localizer plugin for IgorPro.

For long term growth samples were incubated at room temperature overnight and imaged on a custom inverted widefield/epifluorescent/TIRF microscope, equipped with 405, 491, 560, 640 lasers and a CCD camera. Images were analysed using Fiji.

CHAPTER 6 CONCLUSIONS AND FUTURE PERSPECTIVE

6.1 Conclusion

In this doctoral thesis methyltransferase-directed labelling of DNA has been optimised and investigated, in particular with the methyltransferase M.TaqI and fluorescent organic dyes (CHAPTER 2). This has enabled robust and reproducible labelling of DNA at intermediate labelling densities, which has been applied for the identification and visualisation of DNA.

The identification of resistance plasmids, bacterial populations and other complex mixtures of DNA remain challenging, despite the emergence and development of a variety of DNA sequencing technologies (CHAPTER 3 and CHAPTER 4). Similarly, the identification of specific DNA fragments (e.g. plasmids) in live cells remains technically challenging (CHAPTER 5). The unique properties and advantages of methyltransferase-directed labelling have enabled the development and application of new methods to address these challenges.

6.2 Future Perspectives

It is important to consider the future perspectives of methyltransferase-directed labelling in light of the unique advantages it provides over alternative technologies. I consider that there are three primary properties that methyltransferase-directed labelling has, which will continue to lend it advantages over other emerging technologies: low resolution information content; multi-channel information; and targeting of DNA specifically, efficiently and with no damage.

6.2.1 Advantage one: lower resolution information content

For rapid identification of bacterial populations or individual human genome sequences DNA nanopores are a particularly exciting emerging technology that is perhaps most comparable to the methods of DNA identification described in this thesis. For example, both methods effectively read the sequence of native DNA, without the need to synthesise DNA, unlike traditional sequencing technologies. Indeed the intensity traces which are derived by both techniques bear a striking similarity[227] (Figure 6.1).

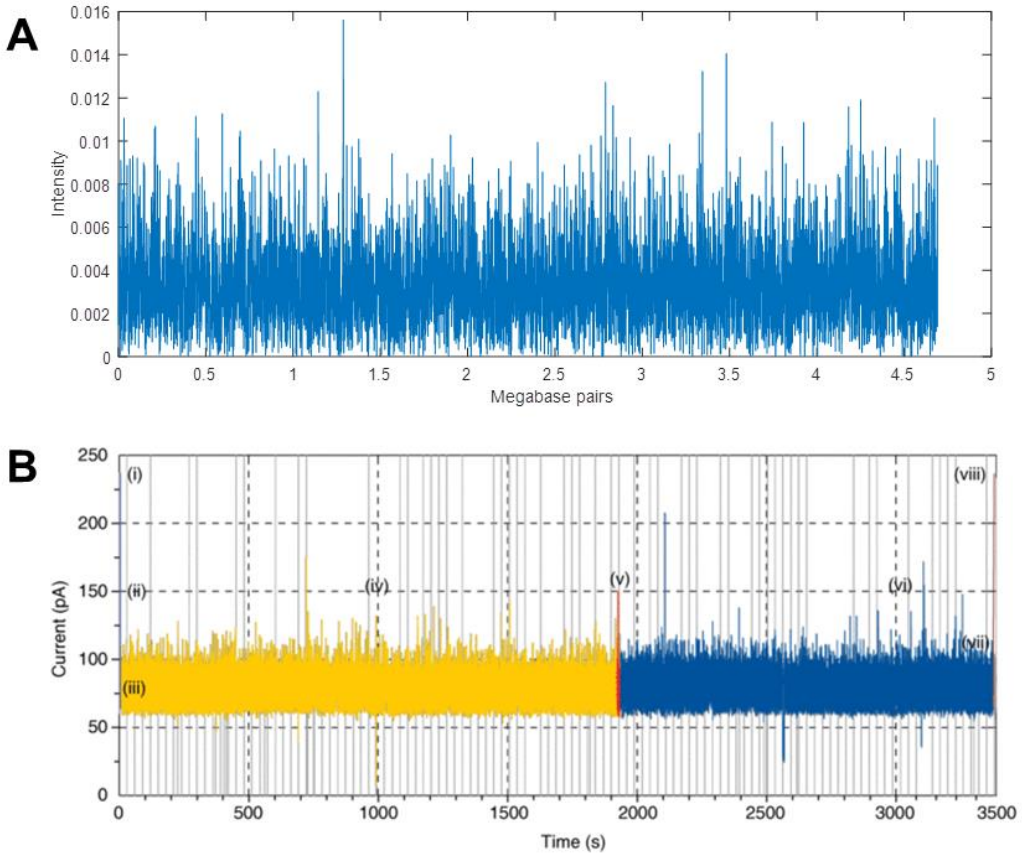


Figure 6.1 Comparison of intensity traces obtained from optical mapping of DNA barcodes and nanopore sequencing. A) Typical intensity trace for optical mapping using M.TaqI-directed labelling. Theoretical intensity trace for E. coli K-12, labelled with M.TaqI and with a PSF width of 250 nm. B) Typical intensity trace for nanopore sequencing using a MinION. Raw current trace for the passage of the single 48 kbp lambda DNA molecule (adapted from Jain et al.[227]).

The primary difference between the two traces is resolution. For optically mapped DNA the intensity profile records a signal at every pixel, approximately every 250 bp (Figure 6.1A), whilst the Oxford Nanopore MinION (Figure 6.1B), can record the current across the nanopore at many time points for even a single base pair. Similarly, the labels that are being detected in DNA labelled by M.TaqI (5'-TCGA-3') are on average every 256 bp, whilst the MinION detects every base pair.

One of the main challenges associated with sequencing DNA in nanochannels is the requirement to achieve single base pair resolution[228]. However, this resolution is not necessary for a simple identification of DNA, as has been shown in this thesis. An intermediate resolution of 100-1000 bp is sufficient to give unique signals for DNA identification, and therefore the challenge of resolution is effectively bypassed entirely. The challenge is simply to obtain high quality signals as rapidly as possible. In addition, lower resolution data is far easier to handle than the incredibly large and unwieldy datasets that are generated using current nanopore technology.

The main limitations with the methods used in this thesis appear to be with the optical mapping procedure. It is difficult to parallelise the detection methods to achieve rapid identification. For molecular combing a single snapshot only contains around 1 Mbp of DNA and takes around 100 ms to acquire. For nanochannels this imaging is even less rapid since kymographs must be recorded. Molecular combing has also proved difficult to apply in practice, and there appears to be a large amount of junk data, due to overlapping DNA molecules.

To overcome these limitations, it may be possible to exploit nanopores, in combination with methyltransferase-directed labelling. Rather than labelling DNA with fluorophores

for optical mapping, labels can be attached to DNA, at sequence-specific sites, which affect the current across a nanopore as the DNA is translocated. This would effectively scale-up the optical mapping procedure described in this thesis and perhaps provide more robust data, given the limitations of optical mapping. The methods applied in CHAPTER 4 could then be used reliably, since this would generate the same type of data as optical mapping. This would allow extremely rapid and robust identification of DNA samples, without the need for single-base resolution.

6.2.2 Advantage two: multi-channel information

The other advantage over current nanopore technology is that methyltransferase-directed labelling allows multiple channels of information. In nanopores only a single channel of information is obtained, the electrical current. However, for methyltransferase-directed labels there is the possibility to use multiple methyltransferases with multiple different dyes which can allow multiplexing of information.

Most obviously this can be applied to ensure that the identification of DNA is as robust as possible. If DNA fragments are labelled with two independent barcodes, then both can be used to maximise the cross-correlation. This would increase the robustness of fit many times since it is less likely the dual barcode will fit accidentally to a different region of the genome. Although this has not been considered in detail in this thesis, an example is shown in Figure 6.2.

In Figure 6.2 two barcodes are simulated for the same stretch of DNA (bases 5,001 to 35,000 of T7). The barcode shown is Figure 6.2A is generated using M.TaqI-directed labelling and the barcode in Figure 6.2B is generated using M.MpeI-directed labelling,

both with only 10% labelling efficiency. On their own, neither barcode can be used to reliably align the fragment to the reference barcode. A displacement of 350 (i.e. 35,000 bp / sampling of 100 bp) is expected, however, the maximum normalised cross-correlation is 0.36 at a displacement of 354 and 0.31 at a displacement of 347 for the M.TaqI-labelled and M.MpeI-labelled fragments respectively (Figure 6.2C). This is rather unreliable alignment (and identification), and is consistent with the results obtained in Figure 3.25, in which these labelling efficiencies would give correct alignment only around a third of the time.

In contrast, when the product of the cross-correlations is used the correct displacement is more reliably obtained (Figure 6.2D); even for only 10% labelling a displacement of 351 is obtained. This simple example demonstrates the advantage for identification and alignment that using two barcodes could provide over using single barcodes only. Of course, using even more channels is possible in theory and would provide robust identification of small fragments from even very large genomes, such as the human genome.

Perhaps a more exciting application is to use other channels to provide a completely new layer of information. It has proved possible to use nanopores to investigate methylation and protein binding sites on DNA[228]. However, this still uses a single channel, the current, and therefore will never be as inherently reliable as using an independent second channel.

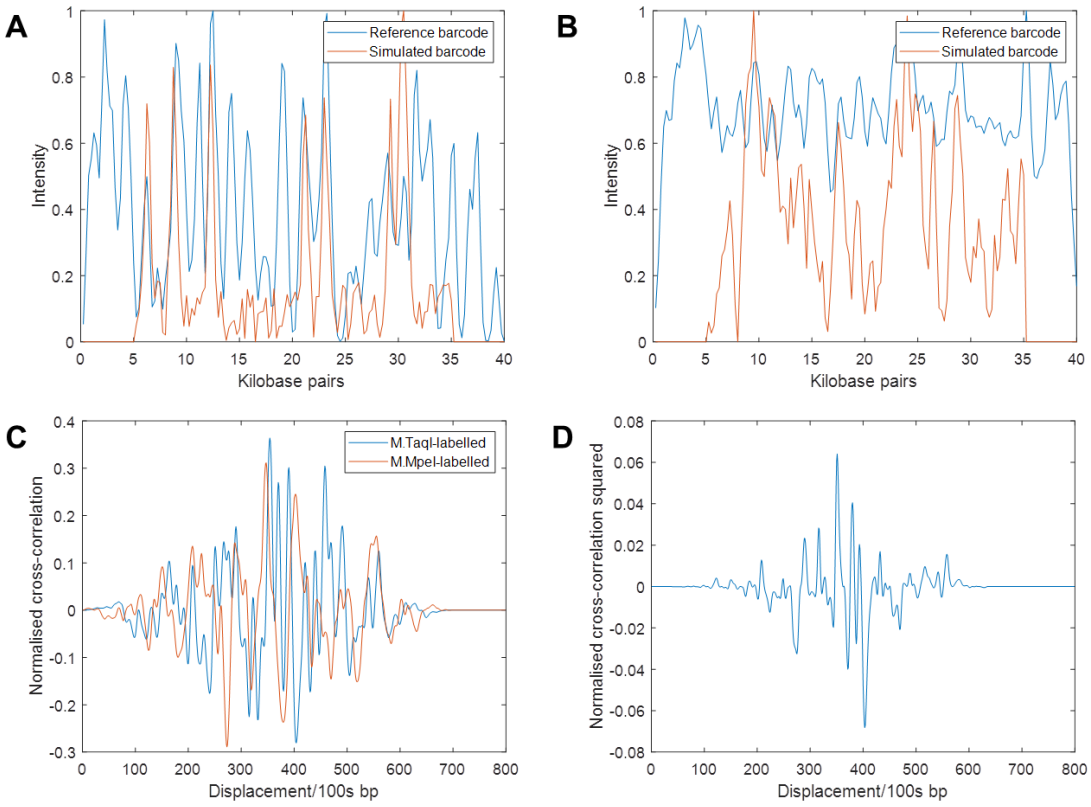


Figure 6.2 **Using two channels for DNA identification. A-B) Simulated barcodes (red) generated from 5,001 to 35,000 bp of reference T7 genome (blue) with 10% labelling efficiency for channels obtained by: A) M.TaqI-directed labelling and B) M.MpeI directed labelling. C) The normalised cross-correlation as a function of displacement for both channels: M.TaqI-labelled (blue) and M.MpeI-labelled (red). This is used to align the simulated fragments, but the results obtained are unreliable. D) The product of the normalised cross-correlation for both channels gives a more reliable alignment result.**

DNA methylation is an intriguing target for epigenetic sequencing. DNA methylation is involved in genome regulation in eukaryotes but in prokaryotes is thought to primarily associated with the restriction-modification system, for protection of native DNA and destruction of foreign DNA. However, there is emerging evidence that DNA methylation in prokaryotes may also have further roles, including control of gene expression[229]. Therefore, the study of the methylation state of prokaryotic genomes, in particular in response to environmental changes, is a growing area of study. Methyltransferase-

directed labelling of DNA could be applied to allow rapid identification of methylation patterns.

For example, one methyltransferase (e.g. M.TaqI) could be used to label the DNA for identification as described in CHAPTER 4. However, a second methyltransferase could be used to label possible methylation sites e.g. Dam methyltransferase. Where sites have been methylated a fluorescent label will not be transferred. A fluorescent signal will only be seen when a site has not been methylated in the bacteria and the methylation pattern could therefore be readily identified. This would allow rapid screening for different environmental conditions, to investigate the role methylation plays in prokaryotic gene regulation.

An alternative is to use the second channel to target DNA-binding proteins. Again, one methyltransferase (e.g. M.TaqI) could be used to label the DNA for identification as described in CHAPTER 4. Now, the second channel is associated with a fluorescently-tagged DNA-binding protein. This kind of method has already been described by Kim *et al.*[95] This could be used to give a static snapshot of bound DNA proteins, similar to current techniques such as ChIP-Seq. However perhaps a more intriguing possibility is to use the DNA barcode in tandem with techniques which investigate the diffusion of DNA-binding proteins.

DNA binding proteins that bind to specific regions of DNA do so with remarkable accuracy and speed, which forms the foundation for all genetic processes, including transcriptional regulation. This is illustrated by the classic lac operon system in which the LacI repressor prevents the transcription of genes required for lactose metabolism

whilst lactose is not present[230]. To be effective it must do so in a timely and accurate manner.

In fact, DNA binding proteins are able to bind far more rapidly to their target sequences than predicted by random diffusion alone[231]. Hence the search for the target sequence is said to proceed by 'facilitated diffusion', where the protein binds randomly to a non-specific site on the DNA and then uses the DNA to direct its subsequent search. The facilitated diffusion of DNA binding proteins has been well characterised from a theoretical perspective, but many aspects are yet to be examined experimentally. There is also some debate whether the rapidity of diffusion is primarily due to electrostatic interactions and not the reduction in dimensionality introduced by facilitated diffusion[232].

Single molecule techniques provide an ideal way to examine the diffusion of DNA binding proteins and address these questions. Techniques such as 'DNA curtains'[233] and 'DNA tightropes'[234] have been used to image protein diffusion along DNA, but only along short, known fragments. The methyltransferase-directed labelling of DNA in these types of experiment would allow the DNA to be identified and the experiment to be scaled up to a whole bacterial genome, rather than small known fragments of DNA, that are currently used. This would provide novel insights into how DNA-binding proteins bind so rapidly to their target sequences.

6.2.3 Advantage three: targets DNA specifically, efficiently and without damage

Finally, the main advantage of methyltransferase-directed labelling of DNA, over other labelling strategies, is that is specific, therefore controllable, efficient and non-damaging.

This means it is a useful labelling strategy for any techniques that require labelled DNA, for instance FISH or DNA capture, as well as for generating a unique intensity profile. For example, in this thesis methyltransferase-directed labelling was exploited to fluorescently label plasmid DNA *in vitro* in addition to optical mapping.

In particular, the ease of using the labelling strategy lends itself to more speculative techniques which require the labelling of DNA. Very large labels with a variety of interesting properties can be readily attached to DNA and cause fundamental changes to the properties of the DNA. For instance, hydrophobic labels such as long polymers, or nanoparticles, including magnetic beads, can be attached to DNA and will completely alter its properties. DNA has already been exploited in a range of applications such as DNA origami and methyltransferase-directed labelling offers a tool to further expand its utility.

6.2.4 Summary

Although nanopore sequencing may currently be a more attractive technology for the rapid identification of microorganisms, optical mapping of DNA using methyltransferase-directed labelling retains unique advantages. The methods developed and applied in this research have utilised, optimised and exemplified these unique advantages and should provide a useful basis for the continued development of associated technologies.

CHAPTER 7 APPENDIX

7.1 Chapter 2 Supplementary Figures

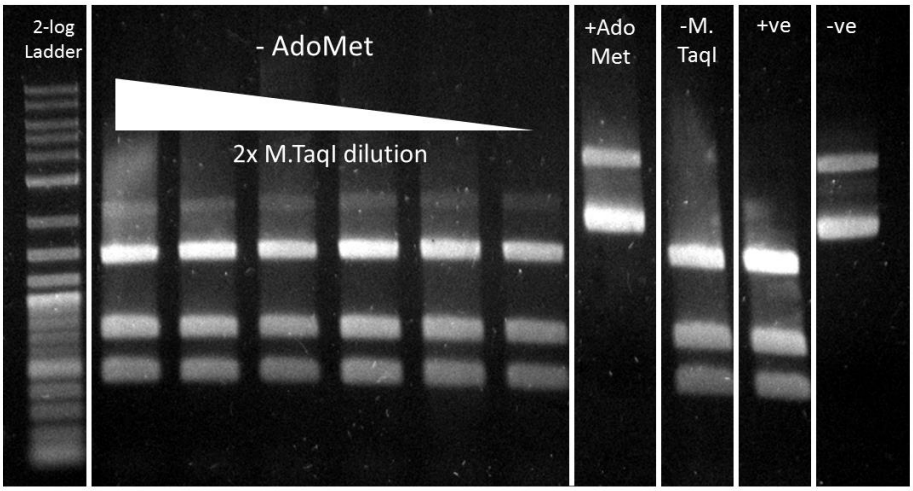


Figure 7.1 **Restriction assay for M.TaqI labelling of pUC19 without cofactor, with purification after incubation and before restriction. Lane 1 = 2 log ladder; lanes 2-7, = No cofactor, 2x dilution of M.TaqI; lane 8 = AdoMet control; lane 9 = no M.TaqI control; lane 10 = restricted pUC19; lane 11 = unrestricted pUC19.**

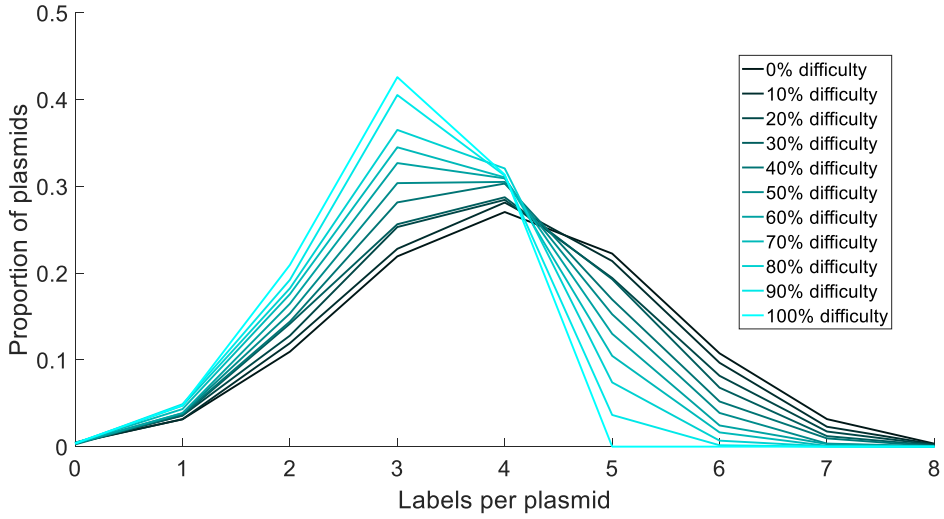


Figure 7.2 **Modelling the effect of hemi-methylation slowing labelling reaction. Expected single molecule counting results for 50% labelling efficiency, where the probability of labelling the palindromic site is reduced by 0 to 100%.**

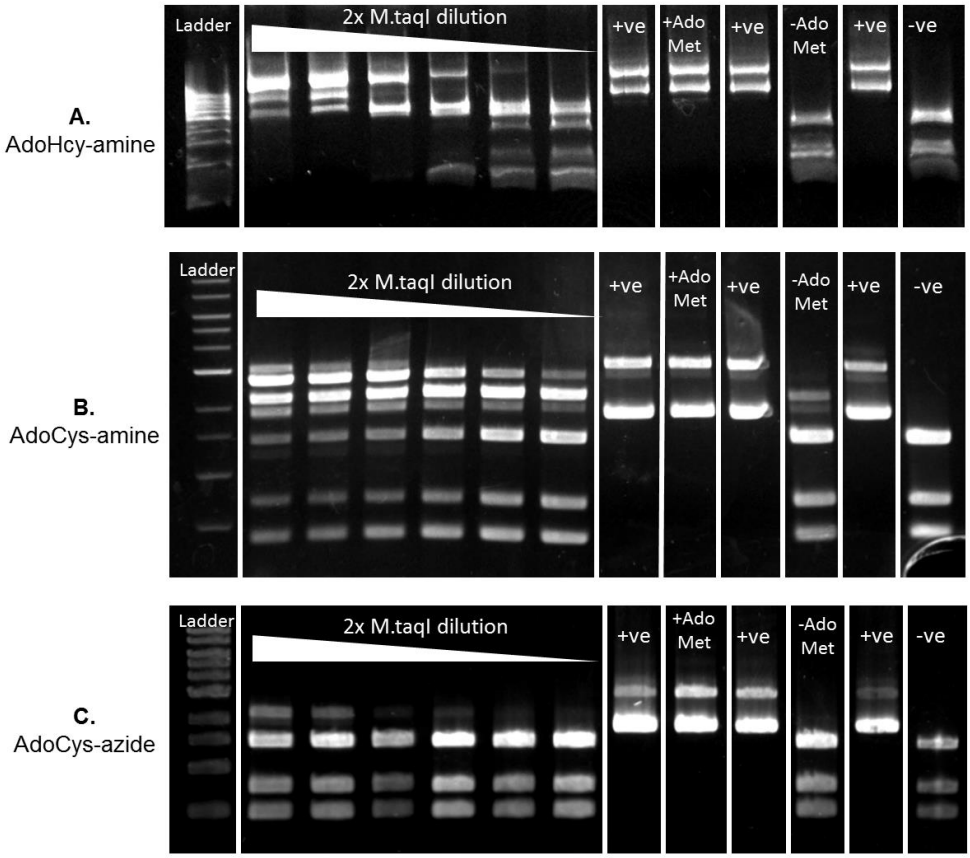


Figure 7.3 **Restriction assays for other AdoMet analogues: A) AdoHcy-amine, B) AdoCys-amine and C) AdoCys-azide. Lane 1= 2 log ladder; lanes 2-7 = AdoMet analogue, 2x M.TaqI dilution; lanes 8, 10, 12 = unrestricted pUC19; lane 9 = AdoMet control; lane 11 = no AdoMet control; lane 11 = restricted pUC19.**

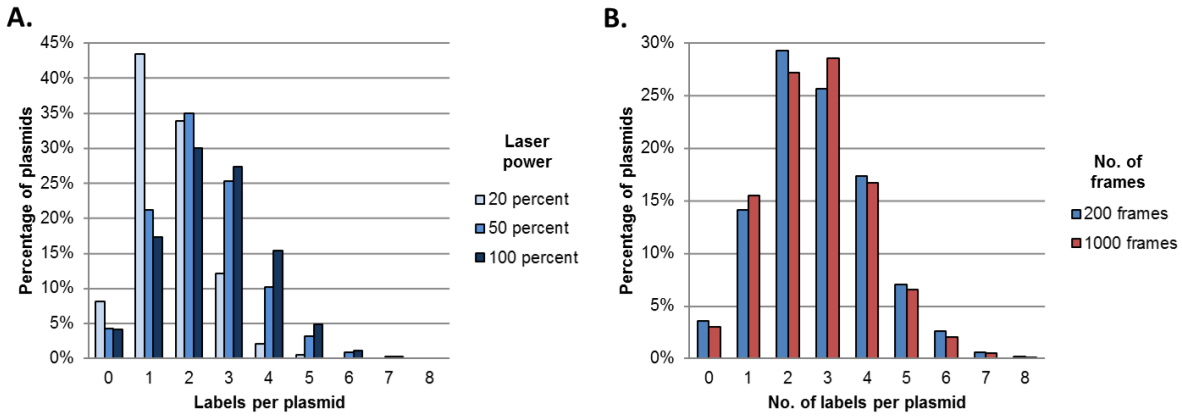


Figure 7.4 **Effect of incomplete bleaching on single molecule counting results. Sample 4 from Figure 14B was used for single molecule counting results. A) Effect of laser power, 500 frames, 20%/50%/100% (blue/red/green) 561nm laser power. B) Effect of number of frames (length of bleaching), 50% laser power, 200/1000 frames (blue/red).**

7.2 Chapter 2 Supplementary Materials and Methods

7.2.1 Restriction assay - Figure 2.6A

The following were mixed and incubated at 50°C for 1 hour.

All in µL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Water	8.0	8.0	8.0	8.0	8.0	8.0	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.0	8.5	8.5
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
15mM AdoMet-azide	0.5	0.5	0.5	0.5	0.5	0.5									0.5		
0.3mg/ml M.TaqI	0.25	0.13	0.06	0.03	0.02	0.01	0.25	0.13	0.06	0.03	0.02	0.01	0.25	0.25			
32mM AdoMet													0.2				

0.5µl R.TaqI was added to samples 1-16 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 18mg/ml proteinase K /0.1% Triton X-100 to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.2 Restriction assay - Figure 2.7A

Oligonucleotides from IDT, both 100µM in water:

T7_1cF 5'-CGTCTATCGGTGAACCTCCGGTATCTCGATCGATCGATCGACCCC-3'

T7_1cR 5'-GGGGTCGATCGATCGATCGAGATACCGGAGGTTCACCGATAGACG-3'

A 1:1 ratio was mixed and heated to 90°C for 15 minutes, before slow cooling overnight. Stored at -20°C.

The following were mixed and incubated at 50°C for 15 minutes, followed by addition of pUC19 and cofactor and incubation at 50°C for 1 hour.

All in µL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Water	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	9.0	9.0	9.0	9.0	9.0	9.0	9.0	9.0
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
0.3mg/ml M.TaqI	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5								
50µM oligos	0.50	0.25	0.13	0.06	0.03	0.02	0.50	0.25	0.13	0.06	0.03	0.02	0.50	0.25	0.13	0.06	0.03	0.02		
Incubate @50oC for 15 minutes																				
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
15mM AdoMet-azide							0.5	0.5	0.5	0.5	0.5	0.5								

0.5µl R.TaqI (100,000units/ml, NEB) was added to samples 1-19 and all samples incubated at 65°C for 2 hours, before adding 0.5µl 18mg/ml proteinase K/0.1% Triton X-100 to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.3 Restriction assay - Figure 2.8A

A 500µl solution containing 1x CutSmart, 6µg M.TaqI and 1µM oligonucleotides was incubated overnight at 37°C. Pierce strong anion exchange spin columns (Thermo-scientific) were used to remove oligonucleotides and the M.TaqI was concentrated to around 120µl in 1x CutSmart using Amicon Ultra 10K centrifugal filter units (Sigma-Aldrich). 6µg of pUC19 was added and 10µl was used for samples 1, 7 and 13 below. Other samples were prepared by serial dilution and incubated at 50°C for 1 hour.

All in µL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Water	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Pre-incubated M.TaqI	1.60	0.80	0.40	0.20	0.10	0.05	1.60	0.80	0.40	0.20	0.10	0.05	1.60	0.80	0.40	0.20	0.10	0.05		
32mM AdoMet							0.25	0.25	0.25	0.25	0.25	0.25								
15mM AdoMet-azide													0.50	0.50	0.50	0.50	0.50	0.50		

0.5µl R.TaqI was added to samples 1-19 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 18mg/ml proteinase K /0.1% Triton X-100 to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.4 Restriction assay - Figure 2.9

The following samples were prepared and incubated at 37°C for 10 minutes. Samples were diluted 10x in 1x CutSmart and concentrated with Amicon Ultra 10K centrifugal filter units 3 times to a final volume of 10µl to remove excess Sinefungin (Cambridge BioScience). pUC19 and cofactors were then added and incubated at 50°C for 1 hour.

All in µL	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Water	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.5	8.5
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
40mM Sinefungin	2.5	1.3	0.6	0.3	0.2	0.1	2.5	1.3	0.6	0.3	0.2	0.1	2.5	1.3	0.6	0.3	0.2	0.1		
0.3 mg/ml M.TaqI	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50		
Incubate @ 37oC for 10 minutes, wash with 1x CutSmart																				
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
32mM AdoMet							0.25	0.25	0.25	0.25	0.25	0.25								
15mM AdoMet-azide													0.50	0.50	0.50	0.50	0.50	0.50		

0.5µl R.TaqI was added to samples 1-19 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 18mg/ml proteinase K /0.1% Triton X-100 to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.5 Restriction assay - Figure 2.10

An 8µl solution containing 2.8mM AdoMet-amine, 15.6mM Atto647N NHS ester (Sigma-Aldrich) and 5x PBS (Sigma-Aldrich) was incubated at 4°C for 1 hour. The following were mixed and incubated at 50°C for 1 hour. (MES CutSmart buffer – 50mM potassium acetate, 20mM MES, 10mM magnesium acetate, 0.1mg/ml BSA, all from Sigma-Aldrich)

All in µL	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
Water	7.5	7.5	7.5	7.5	7.5	7.9	7.9	7.9	7.9	7.9	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.5	8.5
pH	6.75	6.50	6.25	6.00	5.75	6.75	6.50	6.25	6.00	5.75	6.75	6.50	6.25	6.00	5.75	6.75	5.75	6.75	6.75
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
AdoMet-Atto647N	0.5	0.5	0.5	0.5	0.5						0.5	0.5	0.5	0.5	0.5				
0.3mg/ml M.TaqI	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5						0.5	0.5		
32mM AdoMet						0.1	0.1	0.1	0.1	0.1									

0.5µl R.TaqI was added to samples 1-18 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 20mg/ml proteinase K to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.6 Restriction assay - Figure 2.11A

A 10µl solution containing 3mM AdoMet-amine, 10mM Atto647N NHS ester and 1x PBS was incubated at 4°C for 1 hour. The following were mixed and incubated at 50°C for 1 hour.

All in µL	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Water	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.4	8.4	8.5	8.5	8.0	8.0	8.5	8.5
pH	7.2	7.2	7.2	7.2	7.2	5.7	5.7	5.7	5.7	5.7	7.2	5.7	7.2	5.7	7.2	5.7	7.2	5.7
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
AdoMet-Atto647N	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5					0.5	0.5		
0.3mg/ml M.TaqI	0.25	0.13	0.06	0.03	0.02	0.25	0.13	0.06	0.03	0.02	0.25	0.25	0.25	0.25				
32mM AdoMet											0.1	0.1						

0.5µl R.TaqI was added to samples 1-16 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 20mg/ml proteinase K to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.7 Restriction assay - Figure 2.13A and Figure 2.13B

The following were mixed and incubated at 50°C for 1 hour, either at pH 5.7 or pH 7.2.

All in µL	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Water	8.4	8.4	8.4	8.4	8.4	8.4	8.5	8.4	8.5	8.4	8.5
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
15mM AdoMet-azide	0.1	0.1	0.1	0.1	0.1	0.1				0.1	
NEB M.TaqI	2.00	1.00	0.50	0.25	0.13	0.06		2.0	2.0		
32mM AdoMet								0.1			

0.5µl R.TaqI was added to samples 1-6 and 8-11 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 20mg/ml proteinase K to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.8 Restriction assay - Figure 2.16

A 5µl solution containing 3mM AdoMet-azide, 25mM TAMRA-DBCO, 0.05% formic acid and 50% DMSO was incubated at room temperature for 1 hour. The following were mixed and incubated at 50°C for 1 hour.

All in µL	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Water	6.5	7.5	7.5	7.5	6.9	7.9	7.9	7.9	7.0	8.0	8.0	8.0	7.0	8.0	8.0	8.0	8.5	8.5
Good's Buffer	PIPES	MOPS	HEPES	Tris	PIPES	MOPS	HEPES	Tris	PIPES	MOPS	HEPES	Tris	PIPES	MOPS	HEPES	Tris	Tris	Tris
10x* CutSmart, pH 7.2-7.4	2.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
AdoMet-TAMRA	0.5	0.5	0.5	0.5					0.5	0.5	0.5	0.5						
0.3mg/ml M.TaqI	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5					0.5	0.5	0.5	0.5		
32mM AdoMet					0.1	0.1	0.1	0.1										

All in µL	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Water	7.5	6.5	7.5	7.5	7.9	6.9	7.9	7.9	8.0	7.0	8.0	8.0	8.0	7.0	8.0	8.0	8.5	8.5
Good's Buffer	MES	PIPES	MOPS	HEPES	MES	PIPES	MOPS	HEPES	MES	PIPES	MOPS	HEPES	MES	PIPES	MOPS	HEPES	HEPES	HEPES
10x* CutSmart, pH 6.4-6.8	2.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
AdoMet-TAMRA	0.5	0.5	0.5	0.5					0.5	0.5	0.5	0.5						
0.3mg/ml M.TaqI	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5					0.5	0.5	0.5	0.5		
32mM AdoMet					0.1	0.1	0.1	0.1										

*PIPES was 5X

0.5µl R.TaqI was added to samples 1-17 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 20mg/ml proteinase K to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.9 Restriction assay - Figure 2.17

A 5µl solution containing 3mM AdoMet-azide, 25mM TAMRA-DBCO, 0.05% formic acid and 50% DMSO was incubated at room temperature for 1 hour. The following were mixed and incubated at 50°C for 1 hour.

All in µL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Water	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.5	7.4	7.5	8.0	8.5
10x MES CutSmart, pH 5.75	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.1M NaOH	1.00	0.50	0.25	0.13	0.05	0.0											
0.1M NaCl							1.00	0.50	0.25	0.13	0.05	0.0					
AdoMet-TAMRA	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5				0.5	
NEB M.TaqI	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		1.0	1.0		
32mM AdoMet														0.1			

0.5µl R.TaqI was added to samples 1-16 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 20mg/ml proteinase K to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.10 Restriction assay - Figure 2.18

The following were mixed and incubated for 1 hour at the indicated temperature.

All in µL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Water	8.2	8.2	8.2	8.2	8.2	7.9	7.9	7.9	7.9	7.9	8.4	8.4	8.4	8.4	8.4	8.5	8.5
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
32mM AdoMet	0.3	0.3	0.3	0.3	0.3												
15mM AdoMet-azide						0.5	0.5	0.5	0.5	0.5							
0.3mg/ml M.TaqI	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1		
Mix and incubate @																	
Temperature	30°C	40°C	50°C	60°C	70°C	30°C	40°C	50°C	60°C	70°C	30°C	40°C	50°C	60°C	70°C	70°C	70°C

0.5µl R.TaqI was added to samples 1-16 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 18mg/ml proteinase K /0.1% Triton X-100 to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.11 Restriction assay - Figure 2.19

The following were mixed and incubated at 50°C for the time indicated.

All in µL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Water	8.0	8.0	8.0	8.0	8.0	7.8	7.8	7.8	7.8	7.8	8.3	8.3	8.3	8.3	8.0	8.5	8.5
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
32mM AdoMet	0.3	0.3	0.3	0.3	0.3												
15mM AdoMet-azide						0.5	0.5	0.5	0.5	0.5							
0.3mg/ml M.TaqI	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3		
Mix and incubate @ 50°C for																	
Time	5 mins	10 mins	20 mins	40 mins	60 mins	5 mins	10 mins	20 mins	40 mins	60 mins	5 mins	10 mins	20 mins	40 mins	60 mins	60 mins	60 mins

0.5µl R.TaqI was added to samples 1-16 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 18mg/ml proteinase K /0.1% Triton X-100 to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.12 Restriction assay - Figure 2.20

The following were mixed and incubated at 37°C for 3 hours. AdoHcy hydrolase was purchased from Sigma-Aldrich, 0.1 to 1.0 mg/ml.

All in µL	1	2	3	4	5	6	7	8	9	10	11	12	13
Water	8.4	8.4	8.4	8.4	8.4	8.4	8.4	8.0	8.0	8.3	8.3	8.3	8.3
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
15mM AdoHcy-amine	0.1	0.1	0.1	0.1	0.1	0.1	0.1						
NEB M.TaqI	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1		
3.5mM AdoMet								0.25	0.25				
AdoHcy hydrolase	1.00	0.50	0.25	0.13	0.06	0.03		0.10	0.10	0.10	0.10	0.10	0.10

0.5µl R.TaqI was added to samples 1-7, 9, 11, 13 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 20mg/ml proteinase K to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.13 Restriction assay - Figure 2.23

The following were mixed and incubated at 37°C for 1 hour, followed by 65°C for 15 minutes.

All in µL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Water	8.0	8.0	8.0	8.0	8.0	8.0	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.0	8.5	8.5
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
15mM AdoHcy-azide	0.5	0.5	0.5	0.5	0.5	0.5									0.5		
32mM AdoMet													0.2				
NEB M.HhaI or 10mg/ml M.I	2.50	1.25	0.63	0.31	0.16	0.08	2.50	1.25	0.63	0.31	0.16	0.08	2.5	2.5			

0.5µl R.HhaI was added to samples 1-16 and all samples incubated at 37°C for 1 hour, before adding 0.5µl 18mg/ml proteinase K /0.1% Triton X-100 to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.14 Restriction assay - Figure 7.1

The following were mixed and incubated at 50°C for 1 hour.

All in µL	1	2	3	4	5	6	7	8	9	10
Water	17.0	17.0	17.0	17.0	17.0	17.0	17.0	17.0	17.0	17.0
10x CutSmart	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
1mg/ml pUC19	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
0.3mg/ml M.TaqI	0.50	0.25	0.13	0.06	0.03	0.02	0.50			
32mM AdoMet							0.4	0.4		

0.5µl 18mg/ml proteinase K/0.1% Triton X-100 was added and incubated at 50°C for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 18µl water. 2µl 10x CutSmart was added to each sample and 0.5µl R.TaqI was added to samples 1-9, followed by incubation at 65°C for 1 hour. 0.5µl 18mg/ml proteinase K/0.1% Triton X-100 was added and incubated at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.15 Restriction assay - Figure 7.3

The following were mixed and incubated at 50°C for 1 hour.

All in µL	1	2	3	4	5	6	7	8	9	10	11	12
Water	7.5	7.5	7.5	7.5	7.5	7.5	8.3	8.3	8.5	8.5	8.5	8.5
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
15mM AdoHcy-amine	1.0	1.0	1.0	1.0	1.0	1.0						
NEB M.TaqI	2.00	1.00	0.50	0.25	0.13	0.06	0.25	0.25	0.25	0.25		
3.5mM AdoMet							0.25	0.25				

All in µL	1	2	3	4	5	6	7	8	9	10	11	12
Water	6.0	6.0	6.0	6.0	6.0	6.0	8.3	8.3	8.5	8.5	8.5	8.5
10x MES CutSmart, pH 5.75	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
8.1mM AdoCys-amine	2.5	2.5	2.5	2.5	2.5	2.5						
0.3mg/ml M.TaqI	0.100	0.050	0.025	0.013	0.006	0.003	0.10	0.10	0.10	0.10		
3.5mM AdoMet							0.25	0.25				

All in µL	1	2	3	4	5	6	7	8	9	10	11	12
Water	8.4	8.4	8.4	8.4	8.4	8.4	8.3	8.3	8.5	8.5	8.5	8.5
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1mg/ml pUC19	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
55mM AdoCys-azide	0.1	0.1	0.1	0.1	0.1	0.1						
NEB M.TaqI	2.00	1.00	0.50	0.25	0.13	0.06	0.25	0.25	0.25	0.25		
3.5mM AdoMet							0.25	0.25				

0.5µl R.TaqI was added to samples 1-6, 8, 10, 12 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 20mg/ml proteinase K to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

7.2.16 Labelling for single molecule counting assay - Figure 2.6B

Conditions as for samples 7 to 12 scaled up 4 times. Mixed and incubated at 50°C for 1 hour.

All in µL	1	2	3	4	5	6
Water	34.0	34.0	34.0	34.0	34.0	34.0
10x CutSmart	4.0	4.0	4.0	4.0	4.0	4.0
1mg/ml pUC19	2.0	2.0	2.0	2.0	2.0	2.0
0.3mg/ml M.TaqI	2.00	1.00	0.50	0.25	0.13	0.06

2µl 20mg/ml proteinase K was added and incubated at 50°C for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 50µl 1xTE. 10µl DMSO and 0.3µl 50mM TAMRA-DBCO were added to each sample and incubated at room

temperature for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 50µl 1xTE.

7.2.17 Labelling for single molecule counting assay - Figure 2.7B

Conditions as for 8,10, 12 and two further dilutions. The following were mixed and incubated at 50°C for 15 minutes, followed by addition of pUC19 and cofactor and incubation at 50°C for 1 hour.

All in µL	**1**	**2**	**3**	**4**	**5**
Water	38.25	38.25	38.25	38.25	38.25
10x CutSmart	4.50	4.50	4.50	4.50	4.50
0.3mg/ml M.TaqI	2.25	2.25	2.25	2.25	2.25
50µM oligos	1.13	0.28	0.07	0.02	0.00
Incubate @50oC for 15 minutes					
1mg/ml pUC19	2.25	2.25	2.25	2.25	2.25
15mM AdoMet-azide	2.25	2.25	2.25	2.25	2.25

2µl 20mg/ml proteinase K was added and incubated at 50°C for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 50µl 1xTE. 10µl DMSO and 0.3µl 50mM TAMRA-DBCO were added to each sample and incubated at room temperature for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 50µl 1xTE.

7.2.18 Labelling for single molecule counting assay - Figure 2.8B

A 500µl solution containing 1x CutSmart, 6µg M.TaqI and 1µM oligonucleotides was incubated overnight at 37°C. Pierce strong anion exchange spin columns were used to remove oligonucleotides and the M.TaqI was concentrated to around 120µl in 1x CutSmart using Amicon Ultra 10K centrifugal filter units. 6µg of pUC19 was added and 30µl was used for samples below. Other samples were prepared by serial dilution and incubated at 50°C for 1 hour.

All in µL	**1**	**2**	**3**	**4**	**5**	**6**
Water	25.5	25.5	25.5	25.5	25.5	25.5
10x CutSmart	3.0	3.0	3.0	3.0	3.0	3.0
1mg/ml pUC19	1.5	1.5	1.5	1.5	1.5	1.5
Pre-incubated M.TaqI (estimated)	4.80	2.40	1.20	0.60	0.30	0.15
15mM AdoMet-azide	1.50	1.50	1.50	1.50	1.50	1.50

2µl 20mg/ml proteinase K was added and incubated at 50°C for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 50µl 1xTE. 10µl DMSO and 0.3µl 50mM TAMRA-DBCO were added to each sample and incubated at room temperature for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 50µl 1xTE.

7.2.19 Labelling for single molecule counting assay - Figure 2.11B

Conditions as for samples 1 and 6 Figure 6A scaled up 4 times. Mixed and incubated at 50°C for 1 hour.

All in µL	**1**	**2**
Water	32.0	32.0
pH	7.2	5.7
10x CutSmart	4.0	4.0
1mg/ml pUC19	2.0	2.0
AdoMet-Atto647N	1.5	1.5
0.3mg/ml M.TaqI	1.0	1.0

2µl 20mg/ml proteinase K was added and incubated at 50°C for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 50µl 1xTE.

7.2.20 Labelling for single molecule counting assay - Figure 2.12

An 8µl solution containing 2.8mM AdoMet-azide, 9.4mM TAMRA-DBCO, 0.05% formic acid and 50% DMSO was incubated at room temperature for 1 hour. The following were mixed and incubated at 50°C for 1 hour.

All in µL	**1**	**2**
Water	33.3	33.3
pH	7.2	5.7
10x CutSmart	4.0	4.0
1mg/ml pUC19	1.0	1.0
AdoMet-TAMRA	1.5	1.5
0.3mg/ml M.TaqI	0.25	0.25

2µl 20mg/ml proteinase K was added and incubated at 50°C for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 50µl 1xTE.

7.2.21 Labelling for single molecule counting assay - Figure 2.22

A 200µl solution containing 1x CutSmart, 10µg pUC19, 0.45µg M.TaqI and 750µM AdoHcy-azide was incubated at 50°C for 1 hour. 5µl 18mg/ml proteinase K/0.1% Triton X-100 is added and incubated at 50°C for 1 hour, before purification by GenElute PCR Clean-up kit (Sigma-Aldrich) and elution into 200µl 1xTE (Sigma-Aldrich). Meanwhile a 20µl solution containing 0.5x PBS/50% DMSO, 1mM DBCO-amine (Sigma-Aldrich) and 12.5mM Atto647N or Atto565 NHS Ester (Sigma-Aldrich) was incubated at 4°C for 1 hour. The DNA sample was split into 30µl aliquots and either 10µl of the NHS Ester mix was added or 5µl DMSO, 5µl 1x PBS and 2.5µl 50mM TAMRA or Texas Red DBCO (Jena Bioscience) was added and incubated at room temperature overnight, before purification by GenElute PCR Clean-up kit and elution into 50µl 1xTE.

7.2.22 Labelling for single molecule counting assay - Figure 2.24

The following were mixed and incubated at the indicated temperature for 1 hour. Additionally, M.HhaI and M.MpeI reaction were incubated at 65°C for an additional 15 minutes.

All in µL	**No Mtase**	**M.TaqI**	**M.HhaI**	**M.MpeI**
Water	34.0	30.0	20.0	26.0
10x CutSmart	4.0	4.0		4.0
10x M.HhaI buffer			4.0	
1mg/ml pUC19	2.0	2.0	2.0	2.0
15mM AdoHcy-azide		4.0	4.0	4.0
0.3mg/ml M.TaqI		0.25		
NEB M.HhaI			10.0	
10mg/ml M.MpeI				4.0
Mix and incubate for 1 hour @				
	50°C	50°C	37°C	37°C

2.5 µl 18mg/ml proteinase K /0.1% Triton X-100 was added and incubated at 50°C for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 50µl 1xTE. 10µl DMSO and 0.5µl 50mM TAMRA-DBCO were added to each sample and incubated at room temperature for 1 hour, before purification by GenElute PCR Clean-up kit and elution into 50µl 1xTE.

7.3 Chapter 3 Supplementary Figures

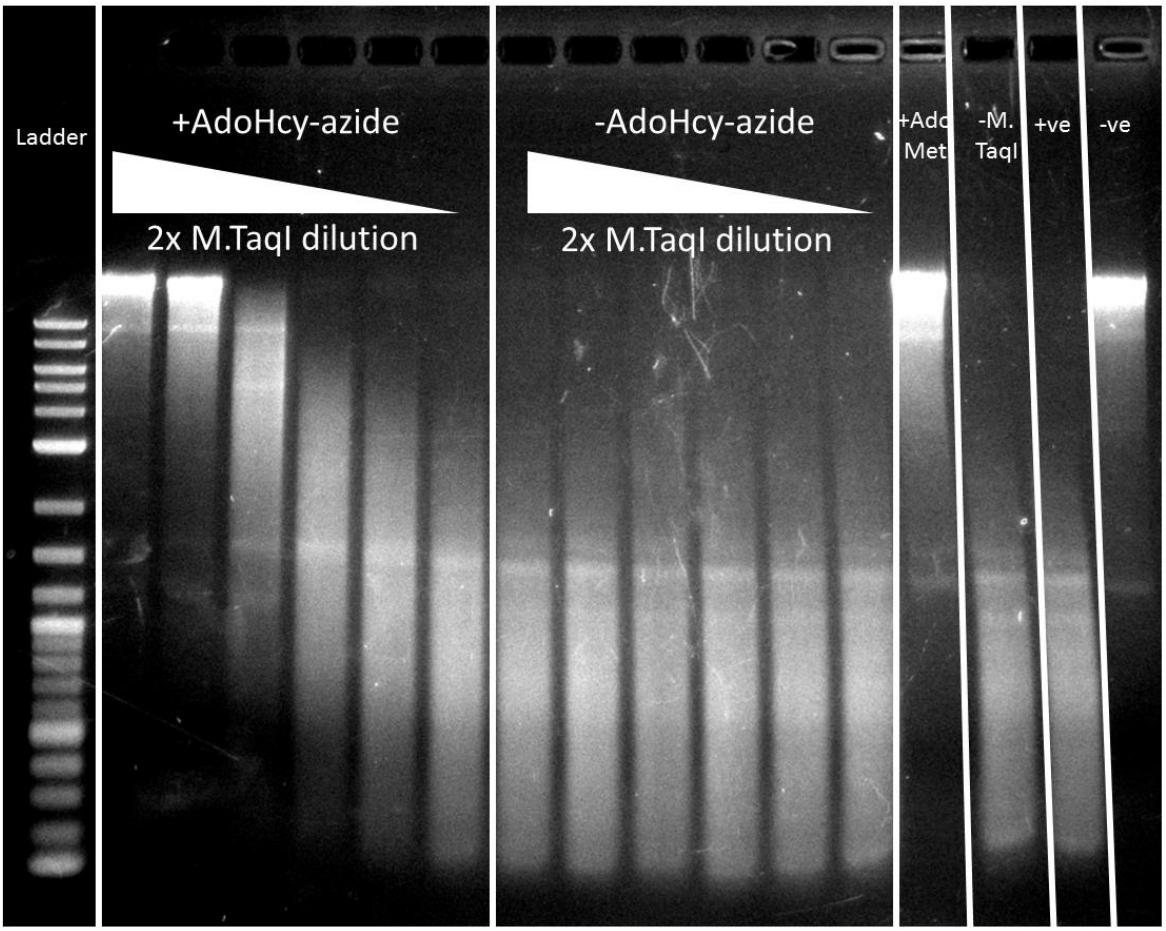


Figure 7.5 **Restriction assay for genomic DNA. A Restriction assay for M.TaqI labelling of pCT/EC958 with AdoHcy-azide and without added cofactor. Lane 1, 2 log ladder; lanes 2-7, AdoHcy-azide, 2x dilution of M.TaqI; lanes 8-13, no cofactor, 2x dilution of M.TaqI; lane 14, AdoMet control; lane 15, no M.TaqI control; lane 16, restricted pUC19; lane 17, unrestricted pUC19.**

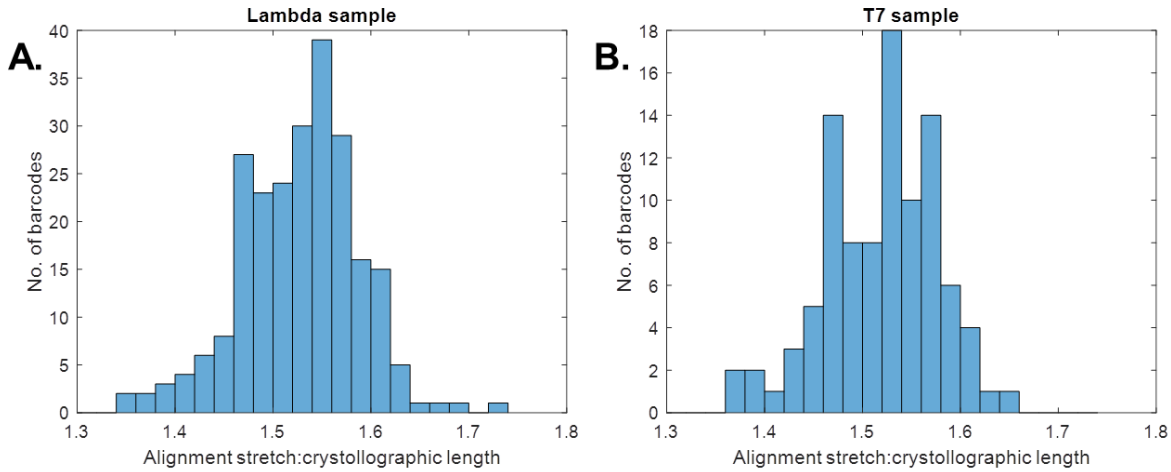


Figure 7.6 **Stretching of DNA during molecular combing. Pure samples of lambda and T7 DNA were M.TaqI-labelled with Atto647N, combed and imaged. Barcodes were extracted and aligned to the respective reference barcodes (see Figure 4.3 and Figure 4.4). The stretch of aligned fragments is displayed here as a histogram for: A) lambda and B) T7. The average and standard deviation for both samples is 1.52±0.06 times the crystallographic length.**

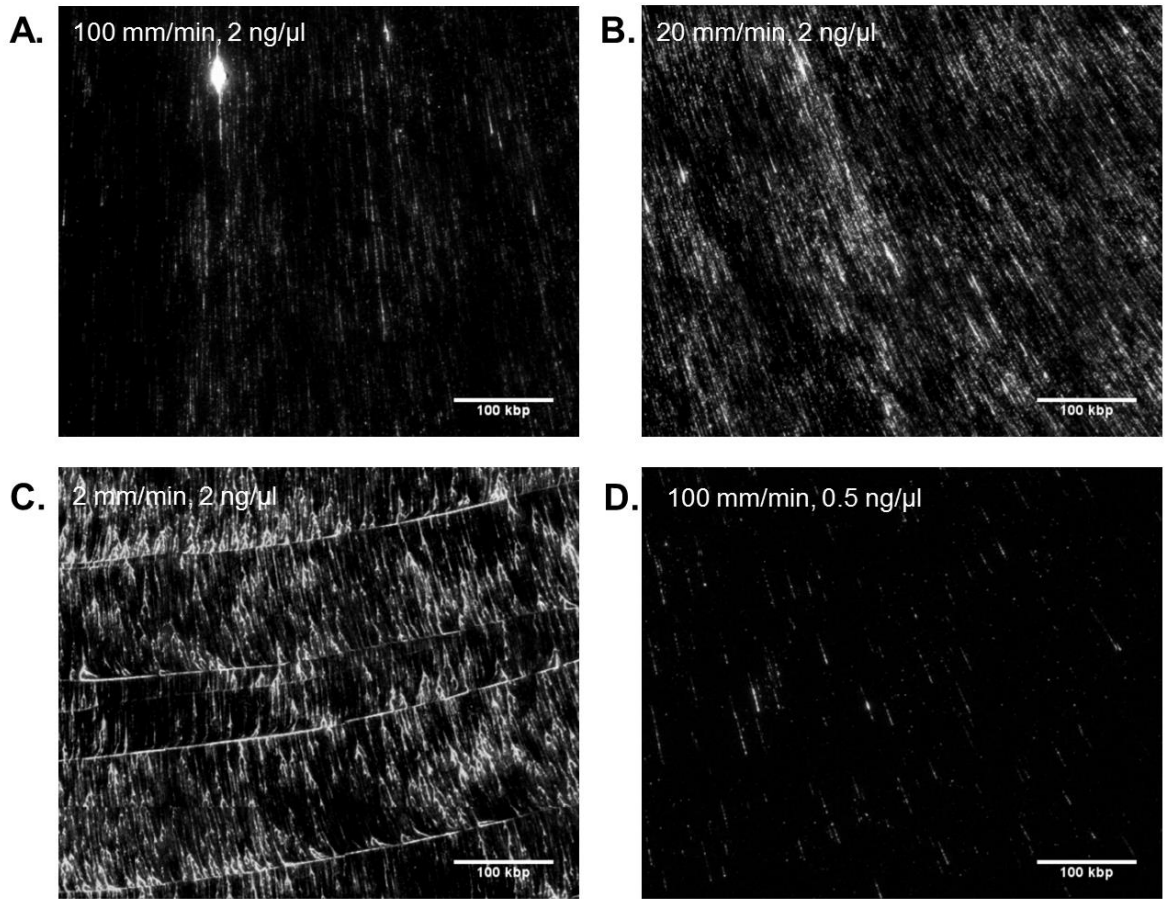


Figure 7.7 **Effect of concentration and combing speed on DNA deposition. M.TaqI-directed Atto647N-labelled T7 DNA is combed in 20mM MES pH5.7, onto zeonex-covered glass cover slips. Several stitched frames are shown A) Low density deposition: 100 mm/min, 2 ng/µl. B) Medium density deposition: 20 mm/min, 2 ng/µl. C) High density deposition: 2 mm/min, 2 ng/µl. D) Very low-density deposition: 100 mm/min, 0.5 ng/µl.**

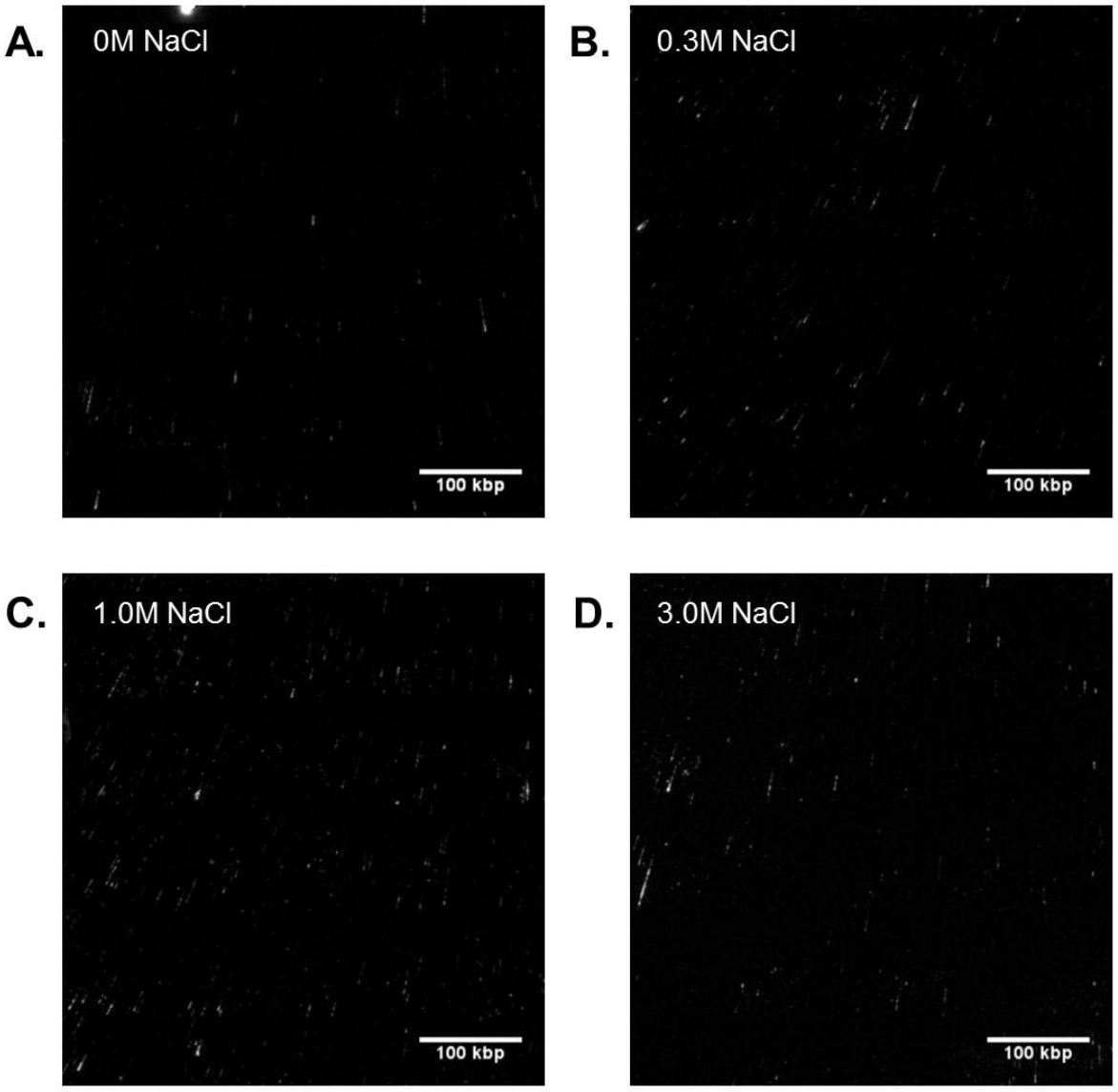


Figure 7.8 **Effect of salt concentration on DNA deposition. M.TaqI-directed Atto647N-labelled T7 DNA is combed in 50mM MES pH5.7, onto zeonex-covered glass cover slips. Several stitched frames are shown for varying concentrations of salt: A) 0 M NaCl; B) 0.3 M NaCl; C) 1.0 M NaCl; D) 3.0 M NaCl**

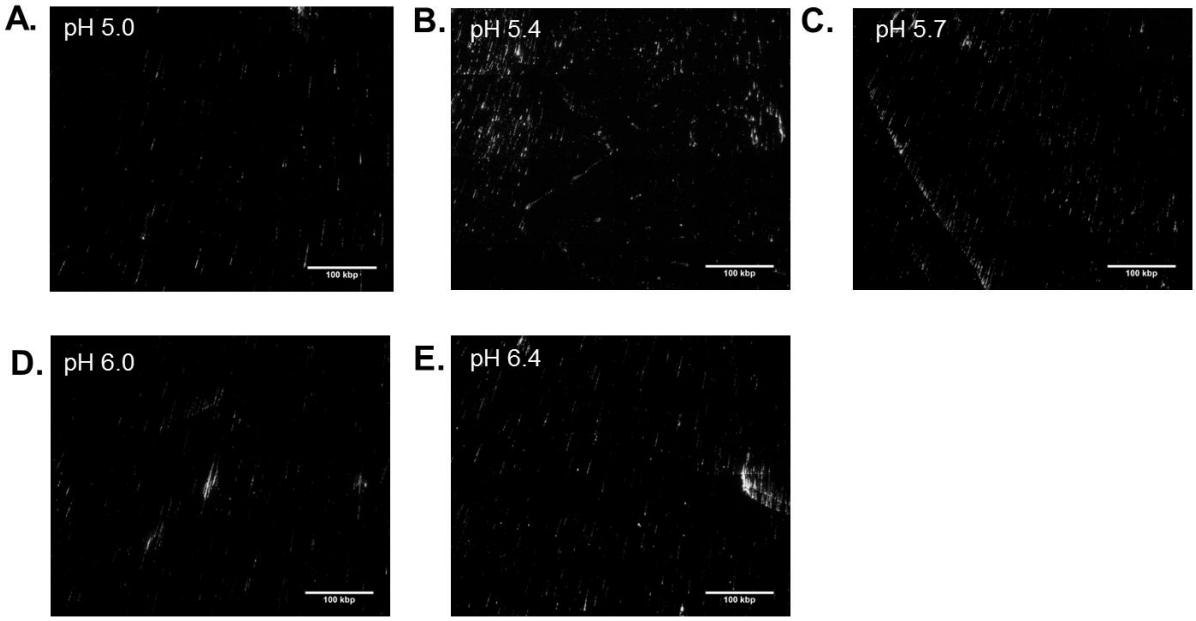


Figure 7.9 **Effect of pH on DNA deposition. M.TaqI-directed Atto647N-labelled T7 DNA is combed in 50mM MES, onto zeonex-covered glass cover slips. Several stitched frames are shown for varying pH: A) pH 5.0; B) pH 5.4 C) pH 5.7; D) pH 6.0; E) pH 6.4.**

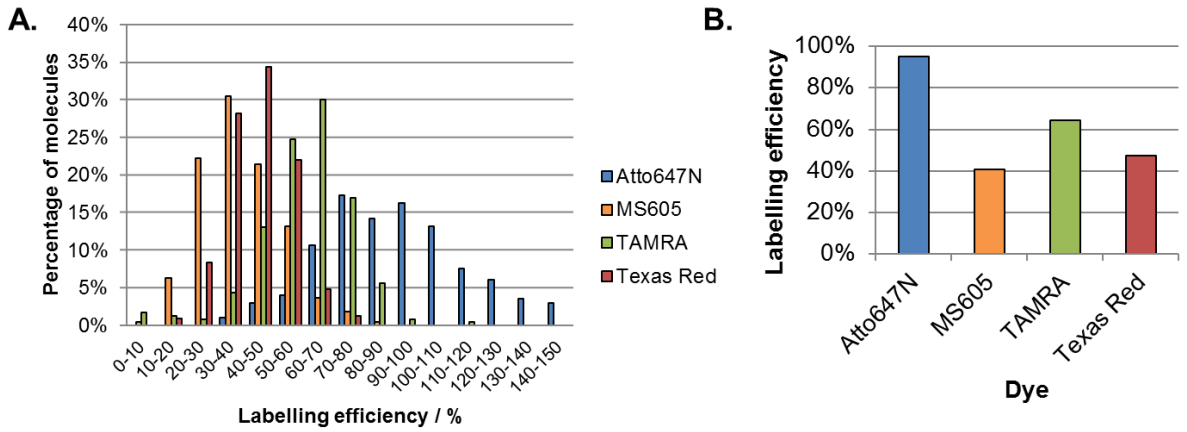


Figure 7.10 **Labelling efficiency of T7 DNA for different commercial dyes, but with high density molecular combing. T7 was labelled with M.TaqI and AdoHcy-azide and coupled post-transalkylation. A) Single molecule counting results for each dye. B) Labelling efficiencies for each dye. The high density of combing meant many DNA molecules were overlapping, which increased the apparent labelling efficiency, compared to Figure 3.10.**

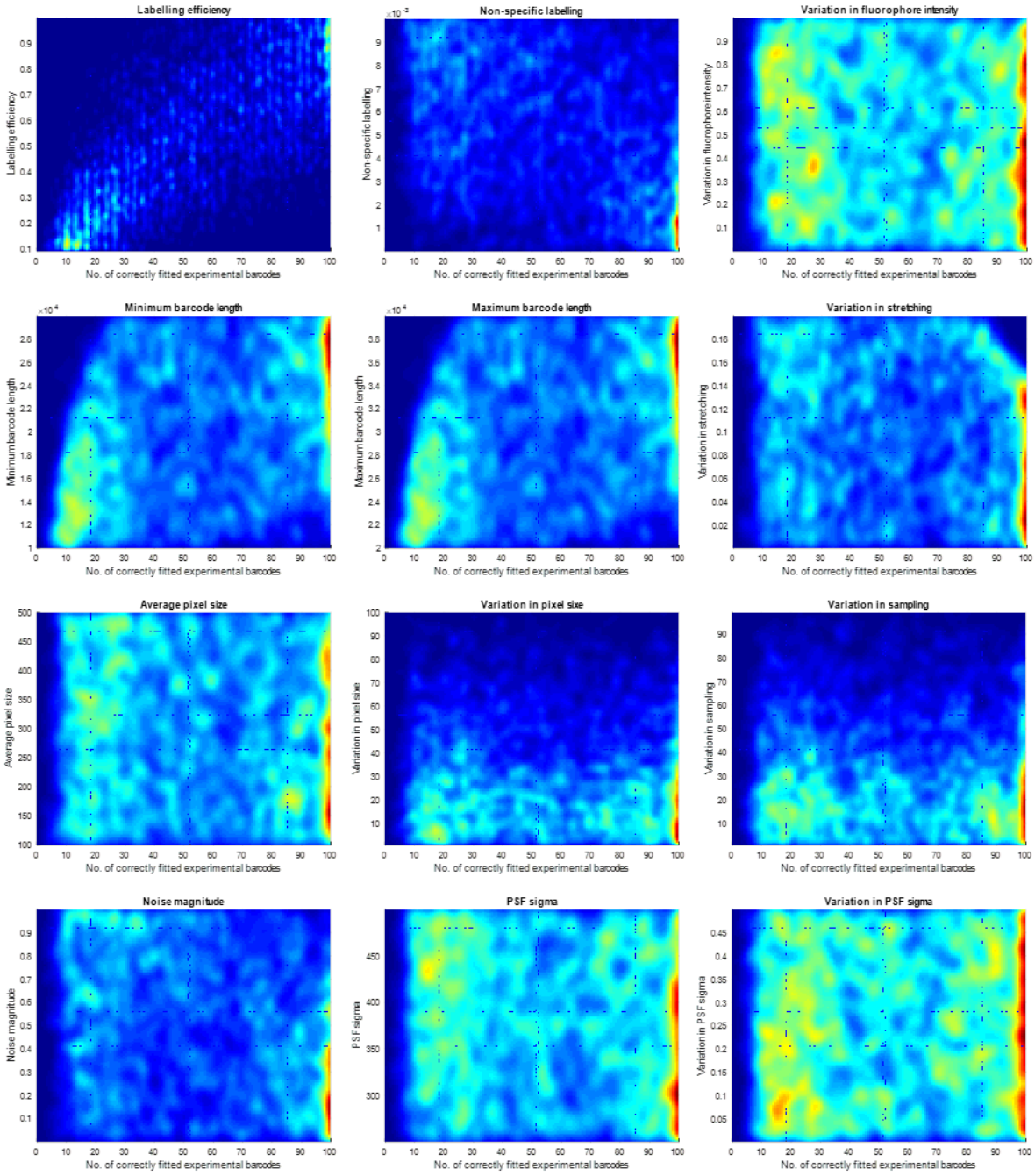


Figure 7.11 **Monte-Carlo simulation to test sensitivity of parameters. 5000 sets of parameters were run for 100 fragments each. Experimental barcodes were generated and aligned from/to the T7 genome. Shown is a 2D histogram for each parameter, comparing the number of correctly-aligned barcodes to the parameter.**

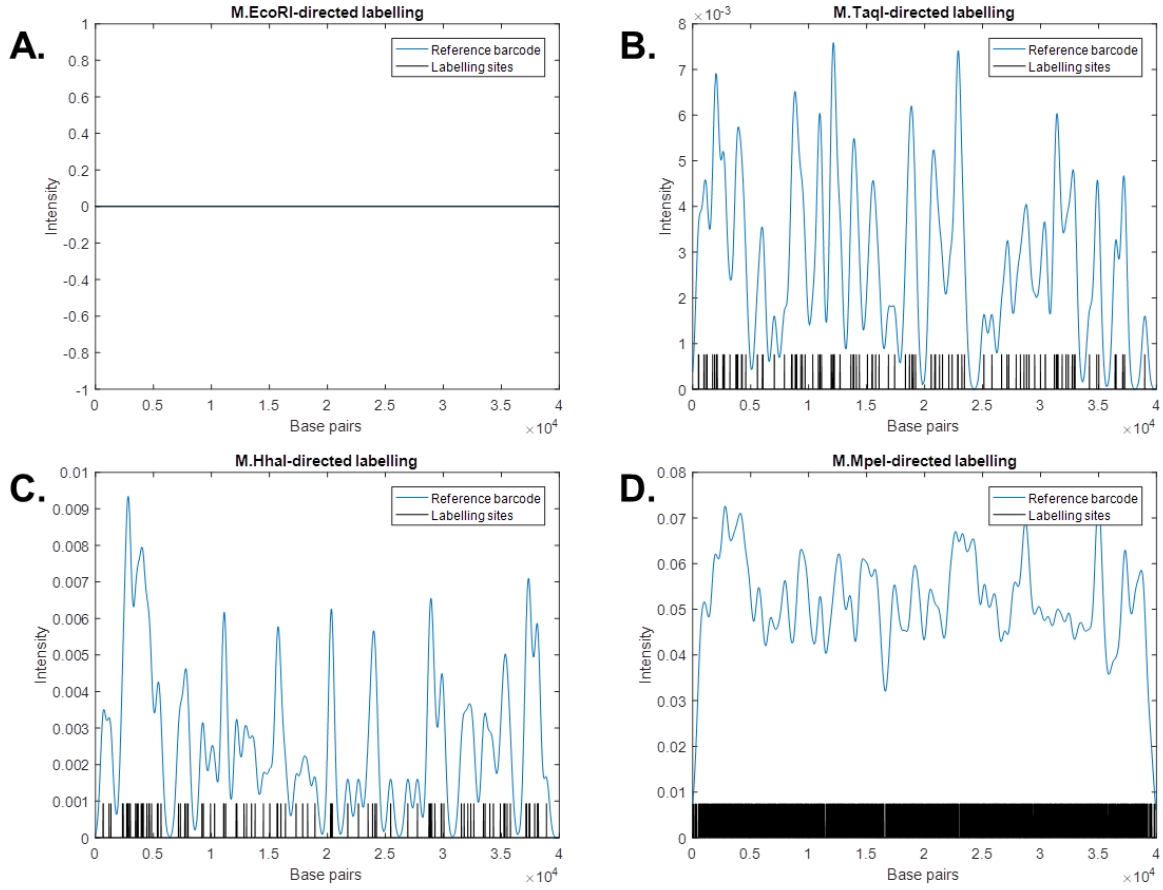


Figure 7.12 **Reference barcodes for T7 genome, labelled by different methyltransferases. Shown in black are the labelling sites, and in blue the reference barcode with a PSF of 250 bp. A) M.EcoRI-directed labelling (GA<u>A</u>TTC sites). B) M.TaqI-directed labelling (TCG<u>A</u> sites). C) M.HhaI-directed labelling (G<u>C</u>GC sites). D) M.MpeI-directed labelling (<u>C</u>G sites).**

7.4 Chapter 3 Supplementary Tables

Variables	Description	Typical values
meth_efficiency	Labelling efficiency	50%
false_methylation	Chance of non-specific labelling (per base pair)	0.0001
flu_intensity_var	Variation in fluorophore intensity	20%
min_fragment_length	Minimum length of fragment (in base pairs)	30,000
max_fragment_length	Maximum length of fragment (in base pairs)	50,000
base_length_var	Variation in stretching	2.5%
sample_freq_mean	Average pixel size (base pairs per pixel)	350
sample_freq_distr	Variation in pixel size (will depend on direction)	50
pixel_distr	Variation in pixel sampling (in base pairs)	40
noise_mag	Magnitude of noise	20%
PSF_frag [sigma]	sigma for experimental PSF (in base pairs)	400
PSF_frag_var	Variation in experimental PSF	10%

Table 7.1 Experimental parameters for generating realistic barcodes

7.5 Chapter 4 Supplementary Figures

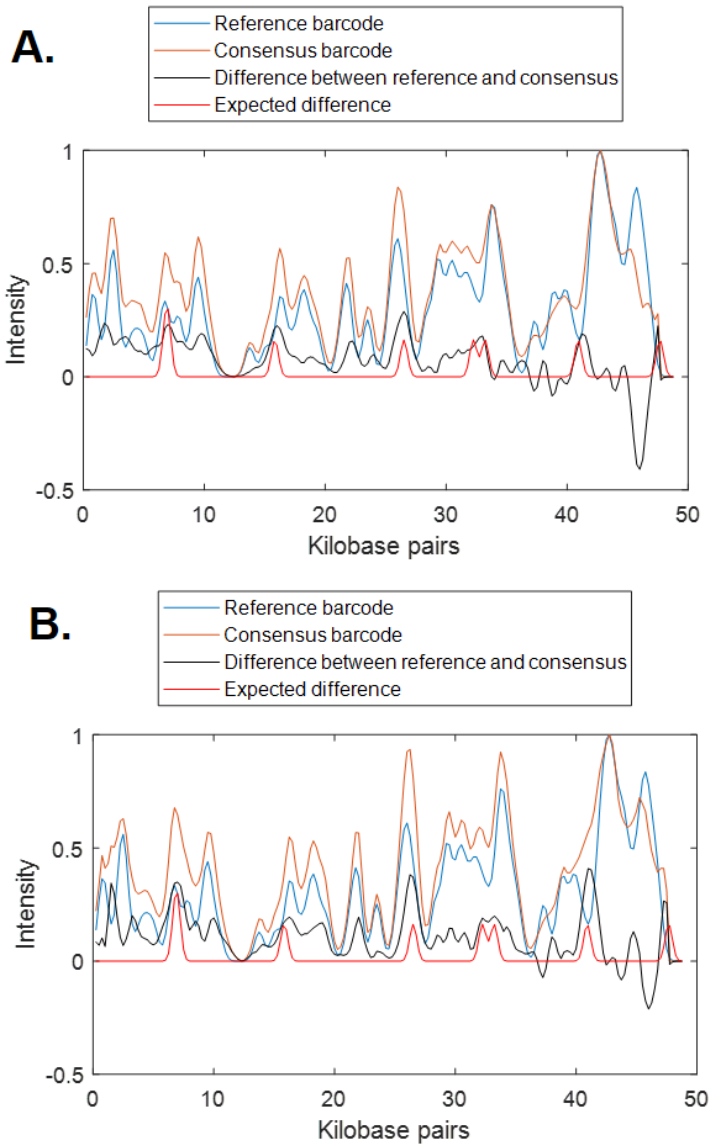


Figure 7.13 **Effect of dam methylation on M.TaqI-directed labelling confirmed by *de novo* alignment. Dam methylation occurs at 5'-G<u>A</u>TC-3' sites and M.TaqI-labelling at 5'-TCG<u>A</u>-3' sites. Therefore, for dam-methylated DNA 5'-TCG<u>A</u>TC-3' sites will be unlabelled, although the effect on 5'-GATCG<u>A</u>-3' sites is not obvious. *De novo* alignment can be used to confirm results for A) Dam-negative lambda DNA and B) Dam-positive DNA. Dam-blocked reference barcode for lambda DNA (blue), *de novo* consensus barcode (brown), difference (black) and expected difference if dam-methylated sites are labelled. This shows that both consensus barcodes are similar, and that the expected and experimental differences are also similar. Both suggest that dam-methylation does not block M.TaqI-labelling.**

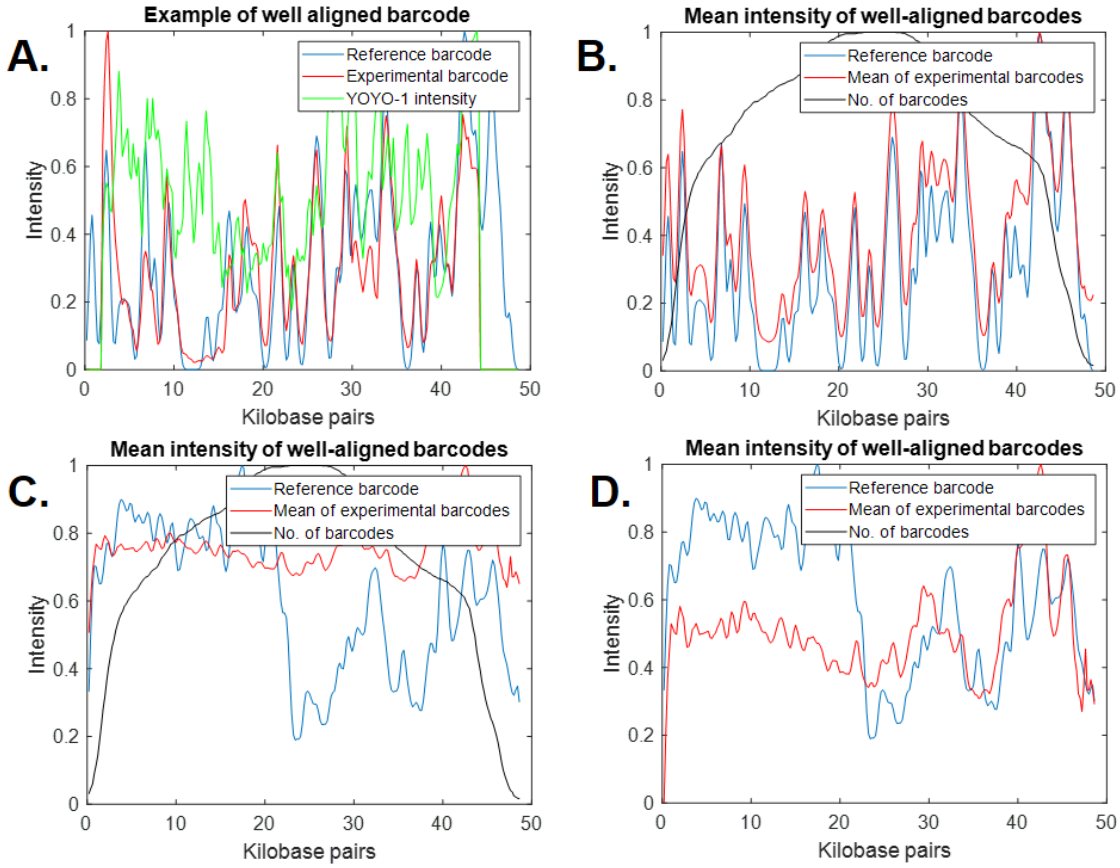
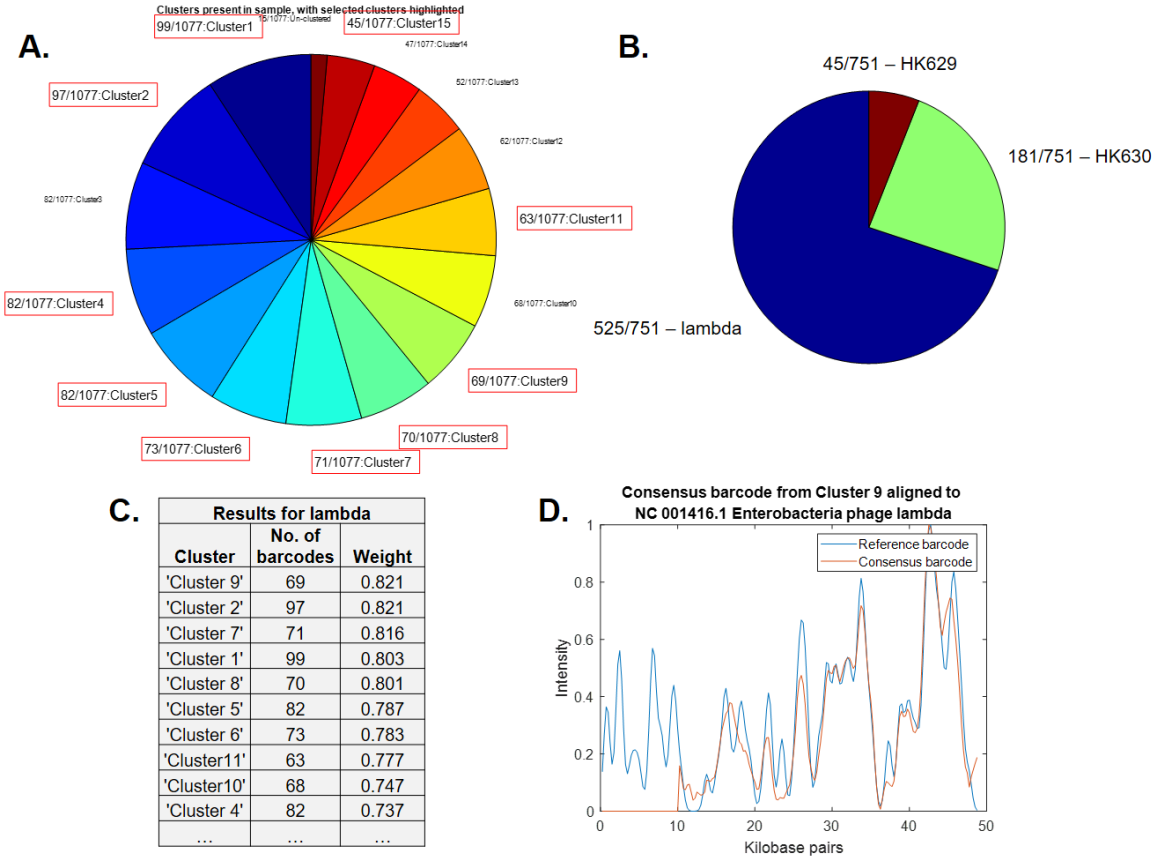
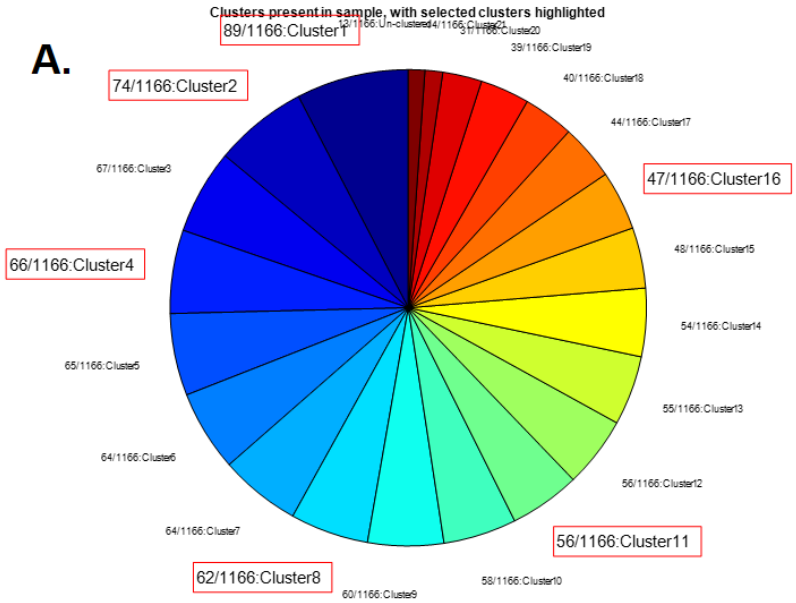


Figure 7.14 **Alignment of a second colour using M.TaqI-directed labelling. Lambda DNA is labelled using M.TaqI with Atto647N, then labelled using YOYO-1. A) An example of an individual barcode. The M.TaqI-barcode (red) is used for alignment to the reference barcode (blue), and for alignment of the affinity-barcode (green). B) Barcodes with an alignment weight greater than 0.65 are used to calculate a mean M.TaqI-barcode (red), which aligns well to the reference barcode (blue). C) The mean affinity-barcode (red) can be calculated and aligned to the expected barcode (blue) – generated by CG labelling. D) The background is removed from the mean affinity-barcode (red) and aligned to the expected barcode (blue).**



Results for lambda		
Cluster	No. of barcodes	Weight
'Cluster 9'	69	0.821
'Cluster 2'	97	0.821
'Cluster 7'	71	0.816
'Cluster 1'	99	0.803
'Cluster 8'	70	0.801
'Cluster 5'	82	0.787
'Cluster 6'	73	0.783
'Cluster11'	63	0.777
'Cluster10'	68	0.747
'Cluster 4'	82	0.737
…	…	…

Figure 7.15 **Assignment of lambda DNA sample by separation, *de novo* alignment and assignment of consensus barcode to reference library. A) 1077 experimental barcodes were extracted and produced 15 consensus barcodes for alignment to around 2000 reference barcodes from a library of phages. B) Barcodes are assigned to the genome with the maximum alignment weight. C) 751 of 1077 barcodes are assigned to lambda with a weight greater than 0.7. D) Example of alignment of consensus barcode.**



B.

Results for T7		
Cluster	**No. of barcodes**	**Weight**
'Cluster 1'	89	0.866
'Cluster 2'	74	0.846
'Cluster 8'	62	0.831
'Cluster11'	56	0.803
'Cluster 4'	66	0.753
'Cluster13'	55	0.752
'Cluster16'	47	0.743
'Cluster21'	14	0.704
'Cluster20'	31	0.695
'Cluster12'	56	0.688
…	…	…

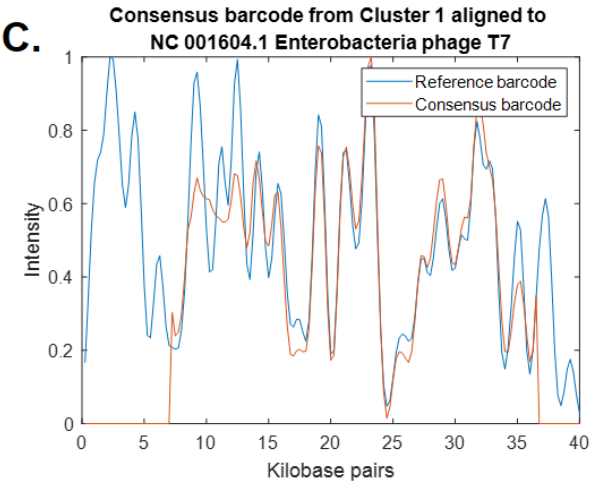
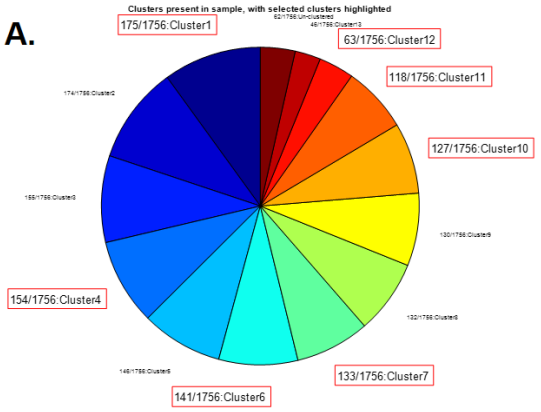
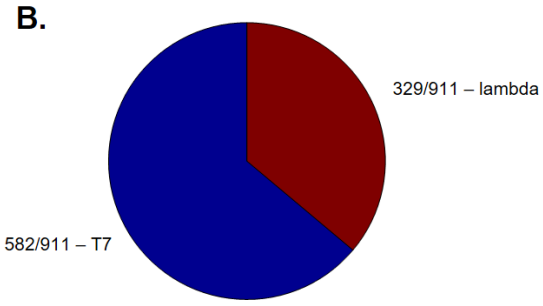


Figure 7.16 **Assignment of T7 DNA sample by separation, *de novo* alignment and assignment of consensus barcode to reference library. A) 1166 experimental barcodes were extracted and produced 21 consensus barcodes for alignment to around 2000 reference barcodes from a library of phages. B) 394 of 1166 barcodes are assigned to T7 with a weight greater than 0.7. C) Example of alignment of consensus barcode.**





C.

Results for lambda		
Cluster	No. of barcodes	Weight
'Cluster 1'	175	0.858
'Cluster 4'	154	0.852
'Cluster13'	46	0.695
'Cluster 2'	174	0.607
'Cluster 6'	141	0.605
'Cluster 5'	146	0.600
'Cluster 7'	133	0.592
'Cluster10'	127	0.586
'Cluster 9'	130	0.582
'Cluster 3'	155	0.566
…	…	…

Results for T7		
Cluster	No. of barcodes	Weight
'Cluster10'	127	0.877
'Cluster 7'	133	0.870
'Cluster11'	118	0.822
'Cluster 6'	141	0.816
'Cluster12'	63	0.768
'Cluster13'	46	0.596
'Cluster 2'	174	0.586
'Cluster 8'	132	0.539
'Cluster 1'	175	0.511
'Cluster 9'	130	0.501
…	…	…

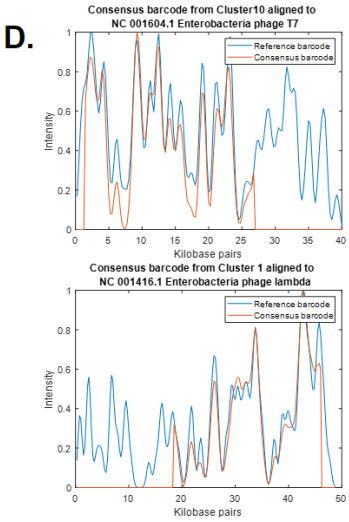


Figure 7.17 **Assignment of lambda/T7 DNA sample by separation, *de novo* alignment and assignment of consensus barcode to reference library. A) 1756 experimental barcodes were extracted and produced 13 consensus barcodes for alignment to around 2000 reference barcodes from a library of phages. B) Barcodes are assigned to the genome with the maximum alignment weight. C) 329 of 1756 barcodes are assigned to lambda and 582 assigned to T7. D) Example of alignments of consensus barcodes.**

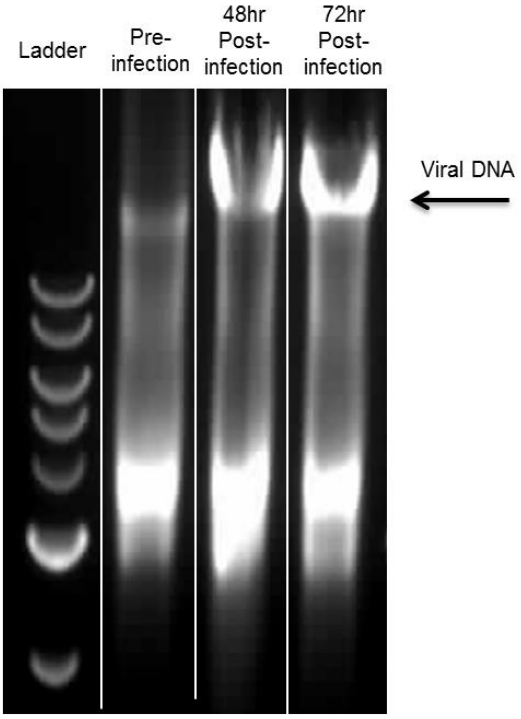


Figure 7.18 **Gel electrophoresis showing Adenovirus A DNA. DNA is extracted from cells pre-infection, 48hr post-infection and 72hr post-infections. A clear band appears that is predominantly Adenovirus A DNA.**

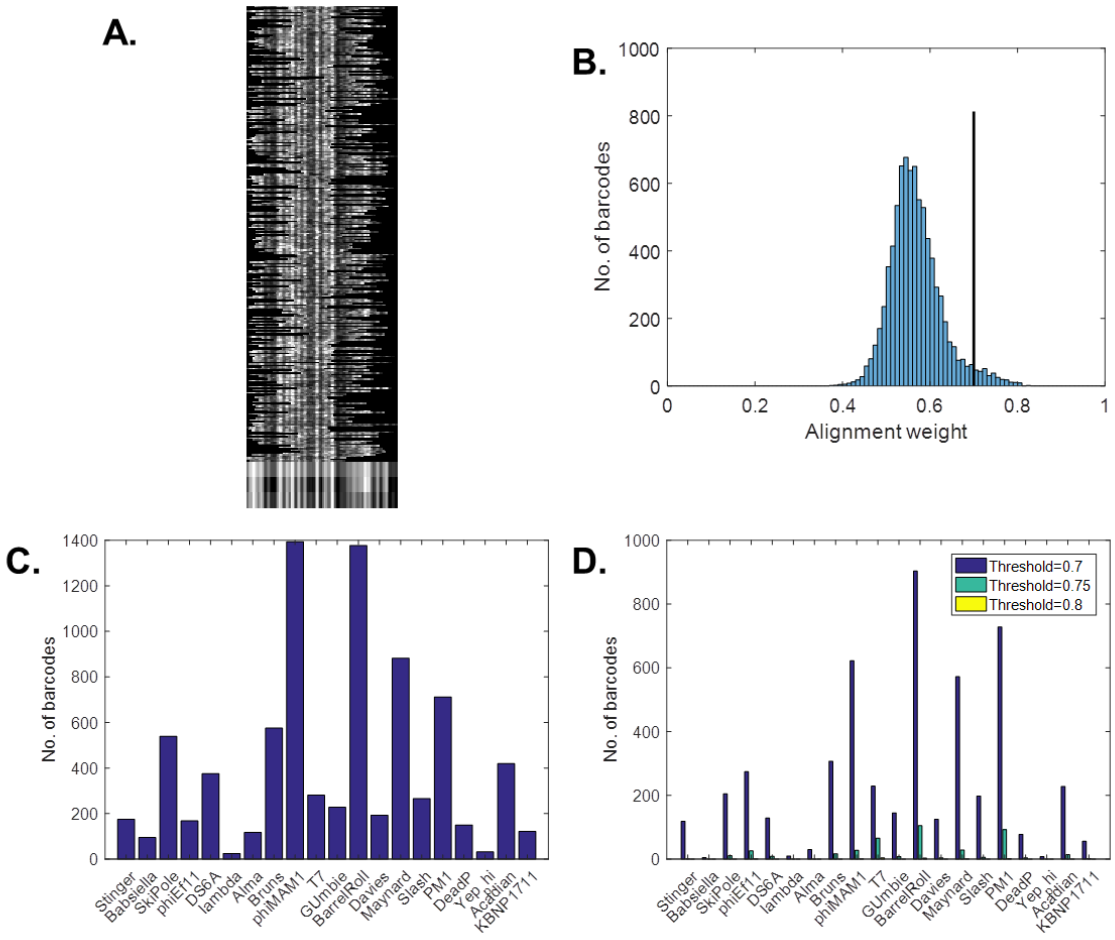


Figure 7.19 **Identification of bacteriophage DNA in genomic mixture by alignment of experimental barcodes to reference barcodes. DNA is labelled using M.TaqI with Atto647N and a 4:1 mixture of *E. coli*:T7 is combed and imaged. 8124 barcodes are extracted from the images. A) 229 (3%) of barcodes aligned with weight>0.7 to T7 reference barcode. B) Alignment weight of all experimental barcodes to T7 reference. C-D) Identification of barcodes against a library of 20 phage genomes. C) Each barcode is assigned to the genome for which the largest alignment weight was obtained. D) The number of barcodes assigned to each genome with an alignment weight greater than a threshold (0.7, 0.75 and 0.8).**

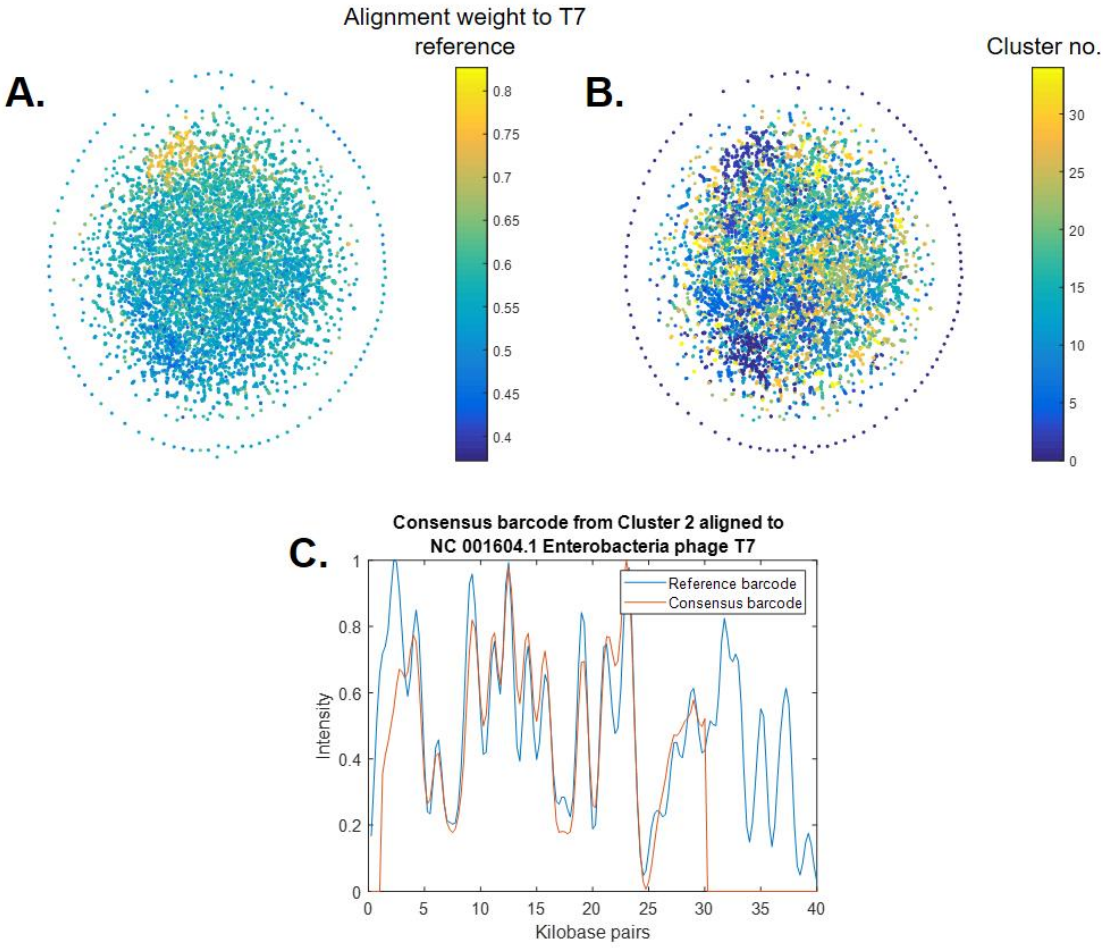


Figure 7.20 **Identification of bacteriophage DNA in genomic mixture by *de novo* separation and alignment of experimental barcodes. A) t-SNE visualisation of network generated from adjacency matrix. Colour is given by alignment weight to T7 reference genome. B) Community detection. Each colour represents a community that has been detected. C) Example of alignment of consensus barcode. Barcodes are assigned to the genome with the maximum alignment weight. 372 of 8124 barcodes are assigned to T7.**

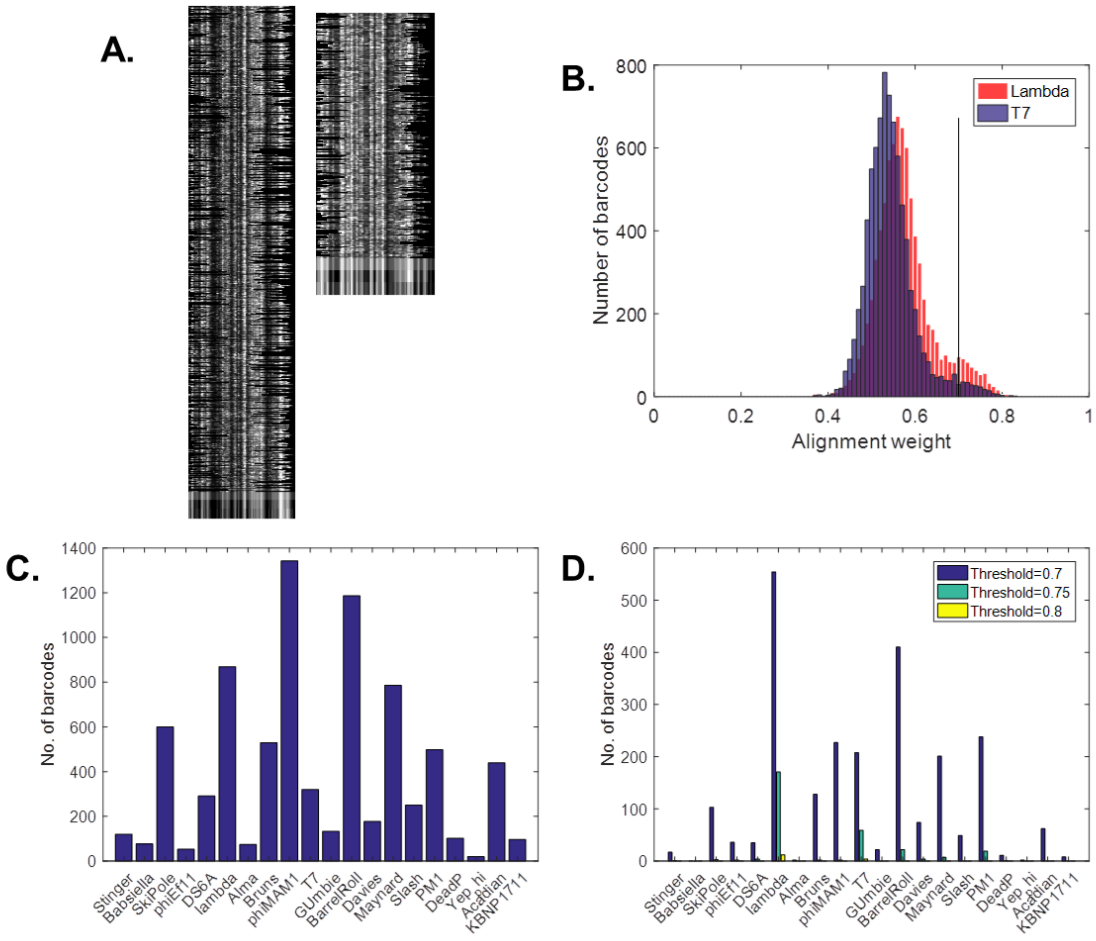


Figure 7.21 **Identification of bacteriophage DNA in genomic mixture by alignment of experimental barcodes to reference barcodes. DNA is labelled using M.TaqI with Atto647N and a 4:1:1 mixture of *E. coli*:lambda:T7 is combed and imaged. 8124 barcodes are extracted from the images. A) 554 (7%) of barcodes aligned with weight>0.7 to lambda reference barcode and 208 (3%) of barcodes aligned with weight>0.7 to T7 reference barcode. B) Alignment weight of all experimental barcodes to T7 and lambda references. C-D) Identification of barcodes against a library of 20 phage genomes. C) Each barcode is assigned to the genome for which the largest alignment weight was obtained. D) The number of barcodes assigned to each genome with an alignment weight greater than a threshold (0.7, 0.75 and 0.8).**

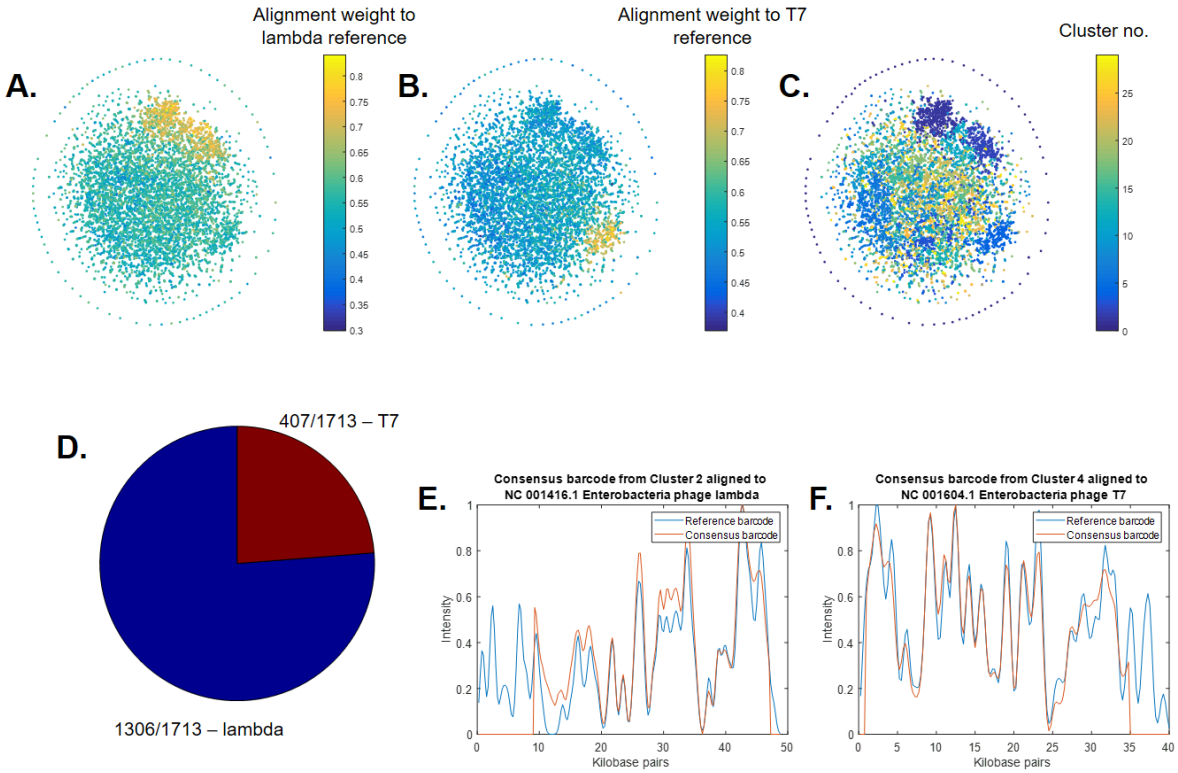


Figure 7.22 **Identification of bacteriophage DNA in genomic mixture by *de novo* separation and alignment of experimental barcodes. A-B) t-SNE visualisation of network generated from adjacency matrix. Colour is given by alignment weight to: A) lambda and B) T7 reference genomes. C) Community detection. Each colour represents a community that has been detected. D) Barcodes are assigned to the genome with the maximum alignment weight. 1306 (16%) and 407 (5%) of barcodes are assigned to lambda and T7 respectively. E-F) Examples of alignment of consensus barcodes.**

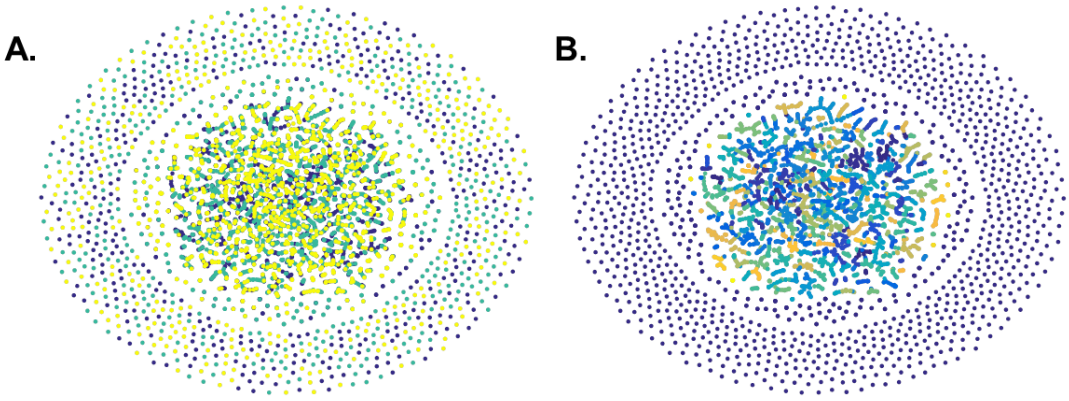


Figure 7.23 **Community detection for experimental barcodes from samples of DH10B, EC958 and blaDNM-1. A) t-SNE visualisation of network generated from adjacency matrix. Each colour represents a different genome: DH10B (blue), EC958 (cyan) and blaDNM-1 (yellow). B) Community detection. Each colour represents a community that has been detected. Note how communities have not be correctly identified by genome compared to *in silico* results (Figure 4.38).**

7.6 Chapter 5 Supplementary Figures

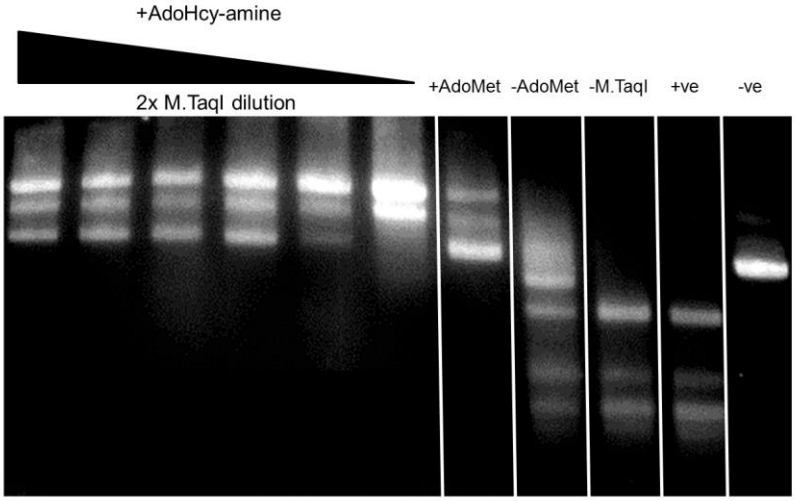


Figure 7.24 **Restriction assay for pRSET B-EGFP. M.TaqI labelling of pRSET B-EGFP with AdoHcy-amine. Lanes 1-6 = AdoHcy-amine, 2x dilution of M.TaqI; lane 7 = AdoMet control; lane 8 = no cofactor control; lane 9 = no M.TaqI control; lane 10 = restricted pRSET B-EGFP, lane 11 = unrestricted pRSET B-EGFP.**

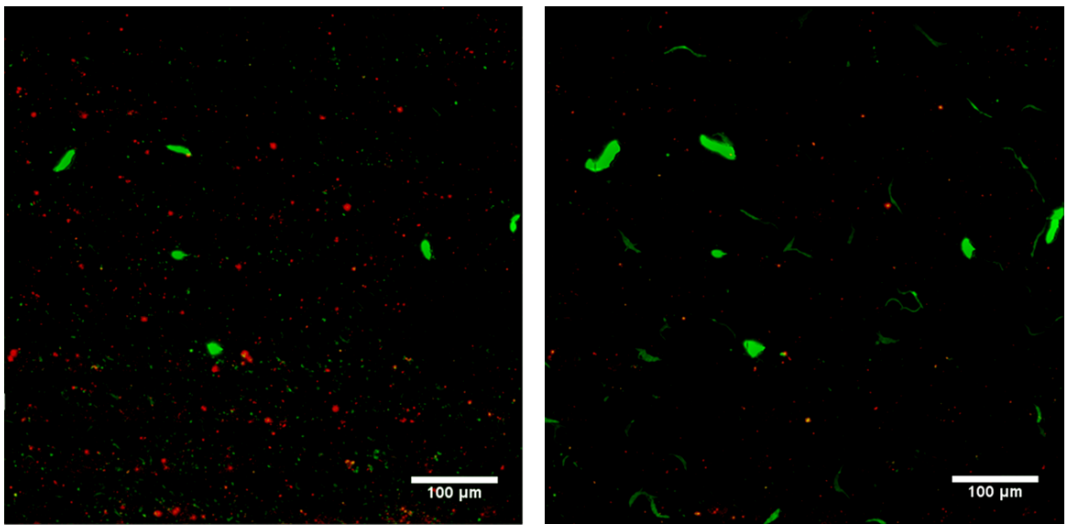


Figure 7.25 **Localisation and dynamics of pRSET B-EGFP during long term growth. T7 Express (green) was transformed with Atto 647N-labelled pRSET B-EGFP (red) and grown on agarose pads (without IPTG) for 12 hours at room temperature. Individual transformants are clearly visible by EGFP expression and microcolonies grow.**

7.7 Chapter 5 Supplementary Materials and Methods

7.7.1 Restriction assay - Figure 7.24

The following were mixed and incubated at 50°C for 1 hour.

All in µL	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Water	3.5	3.5	3.5	3.5	3.5	3.5	3.9	4.0	3.5	4.0	4.0
10x CutSmart	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
0.1mg/ml pRSET B-EGFP	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
15mM AdoHcy-amine	0.5	0.5	0.5	0.5	0.5	0.5			0.5		
0.3mg/ml M.TaqI	0.50	0.25	0.13	0.06	0.03	0.02	0.50	0.50			
3.5mM AdoMet							0.10				

0.5µl R.TaqI was added to samples 1-10 and all samples incubated at 65°C for 1 hour, before adding 0.5µl 18mg/ml proteinase K /0.1% Triton X-100 to all samples and incubation at 50°C for 1 hour. DNA was analysed by gel electrophoresis.

REFERENCES

1. Centenary of Mendel's Paper. *Br Med J* **1,** 368–374 (1965).
2. Griffith, F. The Significance of Pneumococcal Types. *J Hyg (Lond)* **27,** 113–159 (1928).
3. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of Pheumococcal types. *J. Exp. Med.* **79,** 137–158 (1944).
4. Hershey, A. D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36,** 39–56 (1952).
5. Dahm, R. Friedrich Miescher and the discovery of DNA. *Dev Biol* **278,** 274–288 (2005).
6. Levene, P. A. The Structure of Yeast Nucleic Acid IV. Ammonia Hydrolysis. *J. Biol. Chem.* **40,** 415–424 (1919).
7. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171,** 737–738 (1953).
8. Chargaff, E., Zamenhof, S. & Green, C. Human Desoxypentose Nucleic Acid: Composition of Human Desoxypentose Nucleic Acid. *Nature* **165,** 756–757 (1950).
9. Drew, H. R. *et al.* Structure of a B-DNA dodecamer: conformation and dynamics. *PNAS* **78,** 2179–2183 (1981).
10. Crick, F. Central Dogma of Molecular Biology. *Nature* **227,** 561–563 (1970).
11. Nirenberg, M. W. & Matthaei, J. H. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *PNAS* **47,** 1588–1602 (1961).
12. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13,** 303–314 (2012).
13. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43,** D789–D798 (2015).
14. Theisen, A. & Shaffer, L. G. Disorders caused by chromosome abnormalities. *Appl Clin Genet* **3,** 159–174 (2010).
15. Luria, S. E. & Human, M. L. A Nonhereditary, Host-Induced Variation of Bacterial Viruses. *J. Bacteriol.* **64,** 557–569 (1952).
16. Bertani, G. & Weigle, J. J. Host Controlled Variation in Bacterial Viruses. *J. Bacteriol.* **65,** 113–121 (1953).
17. Smith, H. O. & Wilcox, K. W. A Restriction enzyme from Hemophilus influenzae: I. Purification and general properties. *J. Mol. Biol.* **51,** 379–391 (1970).
18. Danna, K. & Nathans, D. Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae. *PNAS* **68,** 2913–2917 (1971).

19. Loening, U. E. The fractionation of high-molecular-weight ribonucleic acid by polyacrylamide-gel electrophoresis. *Biochem J* **102,** 251–257 (1967).
20. Jeffreys, A. J., Wilson, V. & Thein, S. L. Individual-specific 'fingerprints' of human DNA. *Nature* **316,** 76–79 (1985).
21. Danna, K. J., Sack, G. H. & Nathans, D. Studies of Simian virus 40 DNA: VII. A cleavage map of the SV40 genome. *J. Mol. Biol.* **78,** 363–376 (1973).
22. Southern, E. M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98,** 503–517 (1975).
23. Gall, J. G. & Pardue, M. L. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *PNAS* **63,** 378–383 (1969).
24. Bauman, J. G. J., Wiegant, J., Borst, P. & van Duijn, P. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. *Exp. Cell Res.* **128,** 485–490 (1980).
25. Raap, A. K. *et al.* Fiber FISH as a DNA Mapping Tool. *Methods* **9,** 67–73 (1996).
26. Jackson, S. A., Cheng, Z., Wang, M. L., Goodman, H. M. & Jiang, J. Comparative Fluorescence in Situ Hybridization Mapping of a 431-kb Arabidopsis thaliana Bacterial Artificial Chromosome Contig Reveals the Role of Chromosomal Duplications in the Expansion of the Brassica rapa Genome. *Genetics* **156,** 833–838 (2000).
27. Southern, E. M. DNA Microarrays. in *DNA Arrays* 1–15 (Humana Press, 2001).
28. Saiki, R. K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239,** 487–491 (1988).
29. Higuchi, R., Dollinger, G., Walsh, P. S. & Griffith, R. Simultaneous Amplification and Detection of Specific DNA Sequences. *Nat Biotech* **10,** 413–417 (1992).
30. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550,** 345 (2017).
31. Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science* **147,** 1462–1465 (1965).
32. Jou, W. M., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. *Nature* **237,** 82–88 (1972).
33. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260,** 500–507 (1976).
34. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).
35. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *PNAS* **74,** 560–564 (1977).
36. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phiX174 DNA. *Nature* **265,** 687–695 (1977).
37. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *PNAS* **74,** 5463–5467 (1977).
38. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6,** 2601–2610 (1979).

39. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Anal. Biochem.* **242,** 84–89 (1996).
40. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17,** 333–351 (2016).
41. Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet* **16,** 627–640 (2015).
42. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323,** 133 (2009).
43. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nano* **4,** 265–270 (2009).
44. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517,** 608–611 (2015).
45. Schwartz, D. C. *et al.* Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science* **262,** 110–114 (1993).
46. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotech* **30,** 771–776 (2012).
47. Hastie, A. R. *et al.* Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate De Novo Sequence Assembly of the Complex Aegilops tauschii Genome. *PLOS ONE* **8,** e55864 (2013).
48. Teague, B. *et al.* High-resolution human genome structure by single-molecule analysis. *PNAS* **107,** 10848–10853 (2010).
49. Müller, V. & Westerlund, F. Optical DNA mapping in nanofluidic devices: principles and applications. *Lab Chip* **17,** 579–590 (2017).
50. Tegenfeldt, J. O. *et al.* The dynamics of genomic-length DNA molecules in 100-nm channels. *PNAS* **101,** 10979–10983 (2004).
51. Noble, C. *et al.* A Fast and Scalable Kymograph Alignment Algorithm for Nanochannel-Based Optical DNA Mappings. *PLoS ONE* **10,** e0121905 (2015).
52. Riehn, R. *et al.* Restriction mapping in nanofluidic devices. *PNAS* **102,** 10012–10016 (2005).
53. Reisner, W. *et al.* Single-molecule denaturation mapping of DNA in nanofluidic channels. *PNAS* **107,** 13294–13299 (2010).
54. Bensimon, D., Simon, A. J., Croquette, V. & Bensimon, A. Stretching DNA with a Receding Meniscus: Experiments and Models. *Phys. Rev. Lett.* **74,** 4754–4757 (1995).
55. 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).
56. Benke, A., Mertig, M. & Pompe, W. PH- and salt-dependent molecular combing of DNA: experiments and phenomenological model. *Nanotechnology* **22,** 035304 (2011).

57. Deen, J. *et al.* Combing of Genomic DNA from Droplets Containing Picograms of Material. *ACS Nano* **9,** 809–816 (2015).
58. Bensimon, A. *et al.* Alignment and sensitive detection of DNA by a moving interface. *Science* **265,** 2096–2098 (1994).
59. Nyberg, L. K. *et al.* A single-step competitive binding assay for mapping of single DNA molecules. *Biochem. Biophys. Res. Commun.* **417,** 404–408 (2012).
60. Friis Østergaard, P., Matteucci, M., Reisner, W. & Taboryski, R. DNA barcoding via counterstaining with AT/ GC sensitive ligands in injection-molded all- polymer nanochannel devices. *Analyst* **138,** 1249–1255 (2013).
61. Nyberg, L. K. *et al.* Rapid identification of intact bacterial resistance plasmids via optical mapping of single DNA molecules. *Scientific Reports* **6,** 30410 (2016).
62. Müller, V. *et al.* Direct identification of antibiotic resistance genes on single plasmid molecules using CRISPR/Cas9 in combination with optical DNA mapping. *Scientific Reports* **6,** 37938 (2016).
63. Nilsson, A. N. *et al.* Competitive binding-based optical DNA mapping for fast identification of bacteria - multi-ligand transfer matrix theory and experimental applications on Escherichia coli. *Nucleic Acids Research* **42,** e118–e118 (2014).
64. Müller, V. *et al.* Rapid Tracing of Resistance Plasmids in a Nosocomial Outbreak Using Optical DNA Mapping. *ACS Infect. Dis.* **2,** 322–328 (2016).
65. Xiao, M. *et al.* Rapid DNA mapping by fluorescent single molecule detection. *Nucleic Acids Res* **35,** e16–e16 (2007).
66. Jo, K. *et al.* A single-molecule barcoding system using nanoslits for DNA analysis. *PNAS* **104,** 2673–2678 (2007).
67. Das, S. K. *et al.* Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res* **38,** e177–e177 (2010).
68. McCaffrey, J. *et al.* CRISPR-CAS9 D10A nickase target-specific fluorescent labeling of double strand DNA for whole genome mapping and structural variation analysis. *Nucleic Acids Res* **44,** e11 (2016).
69. Neely, R. K. *et al.* DNA fluorocode: A single molecule, optical map of DNA with nanometre resolution. *Chem. Sci.* **1,** 453–460 (2010).
70. Cheng, X. Structure and Function of DNA Methyltransferases. *Annu. Rev. Biophys. Biomol. Struct.* **24,** 293–318 (1995).
71. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **43,** D298–299 (2015).
72. Pignot, M., Siethoff, C., Linscheid, M. & Weinhold, E. Coupling of a Nucleoside with DNA by a Methyltransferase. *Angew. Chem. Int. Ed.* **37,** 2888–2891 (1998).
73. Pljevaljčić, G., Schmidt, F. & Weinhold, E. Sequence-specific Methyltransferase-Induced Labeling of DNA (SMILing DNA). *ChemBioChem* **5,** 265–269 (2004).

74. Comstock, L. R. & Rajski, S. R. Conversion of DNA methyltransferases into azidonucleosidyl transferases via synthetic cofactors. *Nucleic Acids Res* **33,** 1644–1652 (2005).
75. Weller, R. L. & Rajski, S. R. DNA Methyltransferase-Moderated Click Chemistry. *Org. Lett.* **7,** 2141–2144 (2005).
76. Dalhoff, C., Lukinavicius, G., Klimasauskas, S. & Weinhold, E. Direct transfer of extended groups from synthetic cofactors by DNA methyltransferases. *Nat Chem Biol* **2,** 31–32 (2006).
77. 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. *Nat. Protocols* **1,** 1879–1886 (2006).
78. Lukinavičius, G. *et al.* Targeted Labeling of DNA by Methyltransferase-Directed Transfer of Activated Groups (mTAG). *J. Am. Chem. Soc.* **129,** 2758–2759 (2007).
79. Willnow, S., Martin, M., Lüscher, B. & Weinhold, E. A Selenium-Based Click AdoMet Analogue for Versatile Substrate Labeling with Wild-Type Protein Methyltransferases. *ChemBioChem* **13,** 1167–1173 (2012).
80. Peters, W. *et al.* Enzymatic Site-Specific Functionalization of Protein Methyltransferase Substrates with Alkynes for Click Labeling. *Angew. Chem. Int. Ed.* **49,** 5170–5173 (2010).
81. 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).
82. Lee, B. W. K. *et al.* Enzyme-Catalyzed Transfer of a Ketone Group from an S-Adenosylmethionine Analogue: A Tool for the Functional Analysis of Methyltransferases. *J. Am. Chem. Soc.* **132,** 3642–3643 (2010).
83. Grunwald, A. *et al.* Bacteriophage strain typing by rapid single molecule analysis. *Nucleic Acids Res* **43,** e117–e117 (2015).
84. Zhang, J. & Zheng, Y. G. SAM/SAH Analogs as Versatile Tools for SAM-Dependent Methyltransferases. *ACS Chem. Biol.* **11,** 583–597 (2016).
85. Struck, A.-W., Thompson, M. L., Wong, L. S. & Micklefield, J. S-Adenosyl-Methionine-Dependent Methyltransferases: Highly Versatile Enzymes in Biocatalysis, Biosynthesis and Other Biotechnological Applications. *ChemBioChem* **13,** 2642–2655 (2012).
86. Deen, J. *et al.* Methyltransferase-Directed Labeling of Biomolecules and its Applications. *Angew. Chem. Int. Ed.* **56,** 5182–5200 (2017).
87. Goedecke, K., Pignot, M., Goody, R. S., Scheidig, A. J. & Weinhold, E. Structure of the N6-adenine DNA methyltransferase M.TaqI in complex with DNA and a cofactor analog. *Nat Struct Mol Biol* **8,** 121–125 (2001).
88. Klimasauskas, S., Kumar, S., Roberts, R. J. & Cheng, X. Hhal methyltransferase flips its target base out of the DNA helix. *Cell* **76,** 357–369 (1994).

89. Gerasimaitė, R., Vilkaitis, G. & Klimašauskas, S. A directed evolution design of a GCG-specific DNA hemimethylase. *Nucleic Acids Res* **37,** 7332–7341 (2009).
90. 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).
91. Artyukhin, A. B. & Woo, Y.-H. DNA extraction method with improved efficiency and specificity using DNA methyltransferase and 'click' chemistry. *Anal. Biochem.* **425,** 169–174 (2012).
92. Kriukienė, E. *et al.* DNA unmethylome profiling by covalent capture of CpG sites. *Nat Commun* **4,** 2190 (2013).
93. Vranken, C. *et al.* Super-resolution optical DNA Mapping via DNA methyltransferase-directed click chemistry. *Nucleic Acids Res* **42,** e50–e50 (2014).
94. Dedecker, P., Duwé, S., Neely, R. K. & Zhang, J. Localizer: fast, accurate, open-source, and modular software package for superresolution microscopy. *J Biomed Opt* **17,** 126008 (2012).
95. Kim, S. *et al.* Enzymatically Incorporated Genomic Tags for Optical Mapping of DNA-Binding Proteins. *Angew. Chem. Int. Ed.* **51,** 3578–3581 (2012).
96. Jeffet, J. *et al.* Super-Resolution Genome Mapping in Silicon Nanochannels. *ACS Nano* **10,** 9823–9830 (2016).
97. Rombouts, K., Braeckmans, K. & Remaut, K. Fluorescent Labeling of Plasmid DNA and mRNA: Gains and Losses of Current Labeling Strategies. *Bioconjugate Chem.* **27,** 280–297 (2016).
98. Schmidt, F. H.-G. *et al.* Sequence-specific Methyltransferase-Induced Labelling (SMILing) of plasmid DNA for studying cell transfection. *Bioorganic Med. Chem.* **16,** 40–48 (2008).
99. Cabeen, M. T. & Jacobs-Wagner, C. Bacterial cell shape. *Nat Rev Micro* **3,** 601–610 (2005).
100. Gitai, Z. The New Bacterial Cell Biology: Moving Parts and Subcellular Architecture. *Cell* **120,** 577–586 (2005).
101. Lederberg, J. Cell Genetics and Hereditary Symbiosis. *Physiol Rev* **32,** 403–430 (1952).
102. Lederberg, J. Plasmid (1952–1997). *Plasmid* **39,** 1–9 (1998).
103. Thomas, C. M. & Summers, D. Bacterial Plasmids. in *eLS* (John Wiley & Sons, Ltd, 2001).
104. Bosch, F. & Rosich, L. The Contributions of Paul Ehrlich to Pharmacology: A Tribute on the Occasion of the Centenary of His Nobel Prize. *Pharmacology* **82,** 171–179 (2008).
105. Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. *Br J Exp Pathol* **10,** 226–236 (1929).
106. Chain, E. *et al.* Penicillin as a chemotherapeutic agent. *Lancet* **236,** 226–228 (1940).

107. Abraham, E. P. & Chain, E. An enzyme from bacteria able to destroy penicillin. 1940. *Rev. Infect. Dis.* **10,** 677–678 (1988).
108. Munita, J. M. & Arias, C. A. Mechanisms of Antibiotic Resistance. *Microbiol Spectr* **4,** (2016).
109. Piddock, L. J. V. Reflecting on the final report of the O'Neill Review on Antimicrobial Resistance. *The Lancet Infectious Diseases* **16,** 767–768 (2016).
110. Evans, S. R. *et al.* Rapid Molecular Diagnostics, Antibiotic Treatment Decisions, and Developing Approaches to Inform Empiric Therapy: PRIMERS I and II. *Clin Infect Dis* **62,** 181–189 (2016).
111. Livermore, D. M. & Wain, J. Revolutionising Bacteriology to Improve Treatment Outcomes and Antibiotic Stewardship. *Infect Chemother* **45,** 1–10 (2013).
112. Sørensen, S. J., Bailey, M., Hansen, L. H., Kroer, N. & Wuertz, S. Studying plasmid horizontal transfer in situ: a critical review. *Nat. Rev. Microbiol.* **3,** 700–710 (2005).
113. Million-Weaver, S. & Camps, M. Mechanisms of plasmid segregation: have multicopy plasmids been overlooked? *Plasmid* **75,** 27–36 (2014).
114. Summers, D. Timing, self-control and a sense of direction are the secrets of multicopy plasmid stability. *Mol. Microbiol.* **29,** 1137–1145 (1998).
115. Pogliano, J., Ho, T. Q., Zhong, Z. & Helinski, D. R. Multicopy plasmids are clustered and localized in Escherichia coli. *PNAS* **98,** 4486–4491 (2001).
116. Yao, S., Helinski, D. R. & Toukdarian, A. Localization of the Naturally Occurring Plasmid ColE1 at the Cell Pole. *J Bacteriol* **189,** 1946–1953 (2007).
117. Reyes-Lamothe, R. *et al.* High-copy bacterial plasmids diffuse in the nucleoid-free space, replicate stochastically and are randomly partitioned at cell division. *Nucleic Acids Res* **42,** 1042–1051 (2014).
118. Wang, Y., Penkul, P. & Milstein, J. N. Quantitative Localization Microscopy Reveals a Novel Organization of a High-Copy Number Plasmid. *Biophys. J.* **111,** 467–479 (2016).
119. Sánchez-Romero, M.-A., Lee, D. J., Sánchez-Morán, E. & Busby, S. J. W. Location and dynamics of an active promoter in Escherichia coli K-12. *Biochem. J.* **441,** 481–485 (2012).
120. Stokes, G. G. On the Change of Refrangibility of Light. *Phil. Trans. R. Soc. Lond.* **142,** 463–562 (1852).
121. Wiedemann, E. Ueber Fluorescenz und Phosphorescenz I. Abhandlung. *Ann. Phys.* **270,** 446–463 (1888).
122. Valeur, B. & Berberan-Santos, M. N. *Molecular fluorescence: principles and applications*. (John Wiley & Sons, 2012).
123. Song, L., Hennink, E. J., Young, I. T. & Tanke, H. J. Photobleaching kinetics of fluorescein in quantitative fluorescence microscopy. *Biophys. J.* **68,** 2588–2600 (1995).

124. Song, L., Varma, C. A., Verhoeven, J. W. & Tanke, H. J. Influence of the triplet excited state on the photobleaching kinetics of fluorescein in microscopy. *Biophys. J.* **70,** 2959–2968 (1996).
125. Terai, T. & Nagano, T. Small-molecule fluorophores and fluorescent probes for bioimaging. *Pflugers Arch - Eur J Physiol* **465,** 347–359 (2013).
126. Chudakov, D. M., Matz, M. V., Lukyanov, S. & Lukyanov, K. A. Fluorescent Proteins and Their Applications in Imaging Living Cells and Tissues. *Physiol Rev* **90,** 1103–1163 (2010).
127. Resch-Genger, U., Grabolle, M., Cavaliere-Jaricot, S., Nitschke, R. & Nann, T. Quantum dots versus organic dyes as fluorescent labels. *Nat Meth* **5,** 763–775 (2008).
128. Herschel, J. F. W. No. I. On a Case of Superficial Colour Presented by a Homogeneous Liquid Internally Colourless. *Philosophical Transactions of the Royal Society of London* **135,** 143–145 (1845).
129. Duan, Y. *et al.* Recent Progress on Synthesis of Fluorescein Probes. *Mini Rev Org Chem* **6,** 35–43 (2009).
130. Beija, M., Afonso, C. A. M. & Martinho, J. M. G. Synthesis and applications of Rhodamine derivatives as fluorescent probes. *Chem. Soc. Rev.* **38,** 2410–2433 (2009).
131. Kowada, T., Maeda, H. & Kikuchi, K. BODIPY-based probes for the fluorescence imaging of biomolecules in living cells. *Chem. Soc. Rev.* **44,** 4953–4972 (2015).
132. Mishra, A., Behera, R. K., Behera, P. K., Mishra, B. K. & Behera, G. B. Cyanines during the 1990s: A Review. *Chem. Rev.* **100,** 1973–2012 (2000).
133. Combs, C. A. Fluorescence Microscopy: A Concise Guide to Current Imaging Methods. *Curr Protoc Neurosci* **Chapter 2,** Unit2.1 (2010).
134. Abbe, E. Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung. *Archiv f. mikrosk. Anatomie* **9,** 413–418 (1873).
135. Wegel, E. *et al.* Imaging cellular structures in super-resolution with SIM, STED and Localisation Microscopy: A practical comparison. *Scientific Reports* **6,** 27290 (2016).
136. Thorn, K. A quick guide to light microscopy in cell biology. *Mol Biol Cell* **27,** 219–222 (2016).
137. Yamanaka, M., Smith, N. I. & Fujita, K. Introduction to super-resolution microscopy. *Microscopy (Oxf)* **63,** 177–192 (2014).
138. Borgaro, J. G., Benner, N. & Zhu, Z. Fidelity Index Determination of DNA Methyltransferases. *PLOS ONE* **8,** e63866 (2013).
139. Clark, T. A. *et al.* Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res* **40,** e29 (2012).
140. McClelland, M. Purification and characterization of two new modification methylases: MClaI from Caryophanon latum L and MTaqI from Thermus aquaticus YTI. *Nucleic Acids Res.* **9,** 6795–6804 (1981).

141. Lauer, M. H. *et al.* Methyltransferase-directed covalent coupling of fluorophores to DNA. *Chem. Sci.* **8,** 3804–3811 (2017).
142. Klykov, O. & G. Weller, M. Quantification of N -hydroxysuccinimide and N -hydroxysulfosuccinimide by hydrophilic interaction chromatography (HILIC). *Analytical Methods* **7,** 6443–6448 (2015).
143. Agard, N. J., Prescher, J. A. & Bertozzi, C. R. A strain-promoted [3 + 2] azide-alkyne cycloaddition for covalent modification of biomolecules in living systems. *J. Am. Chem. Soc.* **126,** 15046–15047 (2004).
144. 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.TaqI. *J. Mol. Biol.* **265,** 56–67 (1997).
145. Dong, A. *et al.* Structure of the Q237W mutant of HhaI DNA methyltransferase: an insight into protein-protein interactions. *Biol Chem* **385,** 373–379 (2004).
146. Bergerat, A., Guschlbauer, W. & Fazakerley, G. V. Allosteric and catalytic binding of S-adenosylmethionine to Escherichia coli DNA adenine methyltransferase monitored by 3H NMR. *PNAS* **88,** 6394–6397 (1991).
147. Barbés, C., Sánchez, J., Yebra, M. J., Robert-Geró, M. & Hardisson, C. Effects of sinefungin and S-adenosylhomocysteine on DNA and protein methyltransferases from Streptomyces and other bacteria. *FEMS Microbiology Letters* **69,** 239–243 (1990).
148. 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
149. Huber, T. D. *et al.* Functional AdoMet Isosteres Resistant to Classical AdoMet Degradation Pathways. *ACS Chem. Biol.* **11,** 2484–2491 (2016).
150. Good, N. E. *et al.* Hydrogen Ion Buffers for Biological Research*. *Biochemistry* **5,** 467–477 (1966).
151. Good, N. E. & Izawa, S. [3] Hydrogen ion buffers. in *Methods in Enzymology* **24,** 53–68 (Academic Press, 1972).
152. Ferguson, W. J. *et al.* Hydrogen ion buffers for biological research. *Analytical Biochemistry* **104,** 300–310 (1980).
153. Wenner, J. R. & Bloomfield, V. A. Buffer Effects onEcoRV Kinetics as Measured by Fluorescent Staining and Digital Imaging of Plasmid Cleavage. *Anal. Biochem.* **268,** 201–212 (1999).
154. Hülsmann, K. H., Bergerat-Coulaud, A. & Hahn, U. E. coli Dam activity in Hepes buffer asks for a new unit definition. *Nucleic Acids Research* **18,** 7189 (1990).
155. Grunwald, A. *et al.* Bacteriophage strain typing by rapid single molecule analysis. *Nucleic Acids Research* **43,** e117–e117 (2015).
156. Unruh, J. R., Gokulrangan, G., Wilson, G. S. & Johnson, C. K. Fluorescence properties of fluorescein, tetramethylrhodamine and Texas Red linked to a DNA aptamer. *Photochem. Photobiol.* **81,** 682–690 (2005).

157. Ravdin, P. & Axelrod, D. Fluorescent tetramethyl rhodamine derivatives of α-bungarotoxin: Preparation, separation, and characterization. *Analytical Biochemistry* **80,** 585–592 (1977).
158. Gordon, M. P., Ha, T. & Selvin, P. R. Single-molecule high-resolution imaging with photobleaching. *PNAS* **101,** 6462–6465 (2004).
159. Das, S. K., Darshi, M., Cheley, S., Wallace, M. I. & Bayley, H. Membrane protein stoichiometry determined from the step-wise photobleaching of dye-labelled subunits. *Chembiochem* **8,** 994–999 (2007).
160. Zhang, H. & Guo, P. Single molecule photobleaching (SMPB) technology for counting of RNA, DNA, protein and other molecules in nanoparticles and biological complexes by TIRF instrumentation. *Methods* **67,** 169–176 (2014).
161. Liesche, C. *et al.* Automated Analysis of Single-Molecule Photobleaching Data by Statistical Modeling of Spot Populations. *Biophys. J.* **109,** 2352–2362 (2015).
162. Cabral, J. P. S. Water Microbiology. Bacterial Pathogens and Water. *Int J Environ Res Public Health* **7,** 3657–3703 (2010).
163. Torsvik, V. & Øvreås, L. Microbial diversity and function in soil: from genes to ecosystems. *Current Opinion in Microbiology* **5,** 240–245 (2002).
164. MacVane, S. H., Hurst, J. M. & Steed, L. L. The Role of Antimicrobial Stewardship in the Clinical Microbiology Laboratory: Stepping Up to the Plate. *Open Forum Infect Dis* **3,** (2016).
165. Neely, R. K., Deen, J. & Hofkens, J. Optical mapping of DNA: Single-molecule-based methods for mapping genomes. *Biopolymers* **95,** 298–311 (2011).
166. 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).
167. Zhou, S. *et al.* A Single Molecule Scaffold for the Maize Genome. *PLoS Genetics* **5,** e1000711 (2009).
168. 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).
169. Mayjonade, B. *et al.* Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *BioTechniques* **61,** 203–205 (2016).
170. Reslewic, S. *et al.* Whole-Genome Shotgun Optical Mapping of Rhodospirillum rubrum. *Appl Environ Microbiol* **71,** 5511–5522 (2005).
171. Valouev, A., Schwartz, D. C., Zhou, S. & Waterman, M. S. An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proc Natl Acad Sci U S A* **103,** 15770–15775 (2006).
172. Ananiev, G. E. *et al.* Optical mapping discerns genome wide DNA methylation profiles. *BMC Mol Biol* **9,** 68 (2008).
173. Baday, M. *et al.* Multicolor Super-Resolution DNA Imaging for Genetic Analysis. *Nano Lett.* **12,** 3861–3866 (2012).

174. Mendelowitz, L. & Pop, M. Computational methods for optical mapping. *Gigascience* **3,** (2014).
175. Valouev, A. *et al.* Alignment of optical maps. *J. Comput. Biol.* **13,** 442–462 (2006).
176. Nagarajan, N., Read, T. D. & Pop, M. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* **24,** 1229–1235 (2008).
177. Shavit, U., Lowe, R. J. & Steinbuck, J. V. Intensity Capping: a simple method to improve cross-correlation PIV results. *Exp Fluids* **42,** 225–240 (2007).
178. Duda, R. O. & Hart, P. E. Use of the Hough Transformation to Detect Lines and Curves in Pictures. *Commun. ACM* **15,** 11–15 (1972).
179. Azaria, M. & Hertz, D. Time delay estimation by generalized cross correlation methods. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **32,** 280–285 (1984).
180. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one* **6,** e17288 (2011).
181. Backman, K. A cautionary note on the use of certain restriction endonucleases with methylated substrates. *Gene* **11,** 169–171 (1980).
182. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9,** 85 (2008).
183. Martelot, E. L. & Hankin, C. Fast Multi-Scale Detection of Relevant Communities. *arXiv:1204.1002 [physics]* (2012).
184. Forde, B. M. *et al.* The Complete Genome Sequence of Escherichia coli EC958: A High Quality Reference Sequence for the Globally Disseminated Multidrug Resistant E. coli O25b:H4-ST131 Clone. *PLoS One* **9,** (2014).
185. Silva, F., Queiroz, J. A. & Domingues, F. C. Evaluating metabolic stress and plasmid stability in plasmid DNA production by Escherichia coli. *Biotechnology Advances* **30,** 691–708 (2012).
186. Vivian, A., Murillo, J. & Jackson, R. W. The roles of plasmids in phytopathogenic bacteria: mobile arsenals? *Microbiology (Reading, Engl.)* **147,** 763–780 (2001).
187. Bennett, P. M. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *Br J Pharmacol* **153,** S347–S357 (2008).
188. Johnson, I. S. Human insulin from recombinant DNA technology. *Science* **219,** 632–637 (1983).
189. Del Solar, G. & Espinosa, M. Plasmid copy number control: an ever-growing story. *Mol. Microbiol.* **37,** 492–500 (2000).
190. Sengupta, M. & Austin, S. Prevalence and Significance of Plasmid Maintenance Functions in the Virulence Plasmids of Pathogenic Bacteria ▿. *Infect Immun* **79,** 2502–2509 (2011).
191. Eliasson, Å., Bernander, R., Dasgupta, S. & Nordström, K. Direct visualization of plasmid DNA in bacterial cells. *Molecular Microbiology* **6,** 165–170 (1992).

192. Niki, H. & Hiraga, S. Subcellular Distribution of Actively Partitioning F Plasmid during the Cell Division Cycle in E. coli. *Cell* **90,** 951–957 (1997).
193. Weitao, T., Dasgupta, S. & Nordström, K. Plasmid R1 Is Present as Clusters in the Cells of Escherichia coli. *Plasmid* **43,** 200–204 (2000).
194. Gordon, G. S. *et al.* Chromosome and Low Copy Plasmid Segregation in E. coli: Visual Evidence for Distinct Mechanisms. *Cell* **90,** 1113–1121 (1997).
195. Jensen, R. B. & Gerdes, K. Mechanism of DNA segregation in prokaryotes: ParM partitioning protein of plasmid R1 co-localizes with its replicon during the cell cycle. *The EMBO Journal* **18,** 4076–4084 (1999).
196. Lawley, T. D. & Taylor, D. E. Characterization of the Double-Partitioning Modules of R27: Correlating Plasmid Stability with Plasmid Localization. *J. Bacteriol.* **185,** 3060–3067 (2003).
197. Ho, T. Q., Zhong, Z., Aung, S. & Pogliano, J. Compatible bacterial plasmids are targeted to independent cellular locations in Escherichia coli. *EMBO J.* **21,** 1864–1872 (2002).
198. Bignell, C. R., Haines, A. S., Khare, D. & Thomas, C. M. Effect of growth rate and incC mutation on symmetric plasmid distribution by the IncP-1 partitioning apparatus. *Molecular Microbiology* **34,** 205–216 (1999).
199. Yao, S., Toukdarian, A. & Helinski, D. R. Inhibition of protein and RNA synthesis in Escherichia coli results in declustering of plasmid RK2. *Plasmid* **56,** 124–132 (2006).
200. Johnson, E. P., Yao, S. & Helinski, D. R. Gyrase Inhibitors and Thymine Starvation Disrupt the Normal Pattern of Plasmid RK2 Localization in Escherichia coli. *J. Bacteriol.* **187,** 3538–3547 (2005).
201. Nordström, K. & Gerdes, K. Clustering versus random segregation of plasmids lacking a partitioning function: a plasmid paradox? *Plasmid* **50,** 95–101 (2003).
202. Iafolla, M. A. J. *et al.* Dark proteins: effect of inclusion body formation on quantification of protein expression. *Proteins* **72,** 1233–1242 (2008).
203. Martin, R. M., Leonhardt, H. & Cardoso, M. C. DNA labeling in living cells. *Cytometry A* **67,** 45–52 (2005).
204. Slattum, P. S. *et al.* Efficient in vitro and in vivo expression of covalently modified plasmid DNA. *Molecular Therapy* **8,** 255–263 (2003).
205. Daniel, S. G., Westling, M. E., Moss, M. S. & Kanagy, B. D. FastTag Nucleic Acid Labeling System: a versatile method for incorporating haptens, fluorochromes and affinity ligands into DNA, RNA and oligonucleotides. *Biotechniques* **24,** 484–489 (1998).
206. Van Gijlswijk, R. P. *et al.* Universal Linkage System: versatile nucleic acid labeling technique. *Expert Rev. Mol. Diagn.* **1,** 81–91 (2001).
207. Rombouts, K. *et al.* Effect of Covalent Fluorescence Labeling of Plasmid DNA on Its Intracellular Processing and Transfection with Lipid-Based Carriers. *Mol. Pharmaceutics* **11,** 1359–1368 (2014).

208. Gasiorowski, J. Z. & Dean, D. A. Postmitotic Nuclear Retention of Episomal Plasmids Is Altered by DNA Labeling and Detection Methods. *Mol Ther* **12,** 460–467 (2005).
209. Xu, W., Chan, K. M. & Kool, E. T. Fluorescent nucleobases as tools for studying DNA and RNA. *Nature Chemistry* **9,** 1043–1055 (2017).
210. Dervan, P. B. Molecular recognition of DNA by small molecules. *Bioorganic & Medicinal Chemistry* **9,** 2215–2235 (2001).
211. Weston, A., Brown, M. G., Perkins, H. R., Saunders, J. R. & Humphreys, G. O. Transformation of Escherichia coli with plasmid deoxyribonucleic acid: calcium-induced binding of deoxyribonucleic acid to whole cells and to isolated membrane fractions. *Journal of Bacteriology* **145,** 780–787 (1981).
212. Mihalcescu, I., Gateau, M. V.-M., Chelli, B., Pinel, C. & Ravanat, J.-L. Green autofluorescence, a double edged monitoring tool for bacterial growth and activity in micro-plates. *Phys. Biol.* **12,** 066016 (2015).
213. Billinton, N. & Knight, A. W. Seeing the Wood through the Trees: A Review of Techniques for Distinguishing Green Fluorescent Protein from Endogenous Autofluorescence. *Analytical Biochemistry* **291,** 175–197 (2001).
214. Bachmann, B. J. Pedigrees of some mutant strains of Escherichia coli K-12. *Bacteriol Rev* **36,** 525–557 (1972).
215. Cormack, B. P., Valdivia, R. H. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173,** 33–38 (1996).
216. Colville, K., Tompkins, N., Rutenberg, A. D. & Jericho, M. H. Effects of Poly(l-lysine) Substrates on Attached Escherichia coli Bacteria. *Langmuir* **26,** 2639–2644 (2010).
217. Goldberg, S., Doyle, R. J. & Rosenberg, M. Mechanism of enhancement of microbial cell hydrophobicity by cationic polymers. *J Bacteriol* **172,** 5650–5654 (1990).
218. Strahl, H. & Hamoen, L. W. Membrane potential is important for bacterial cell division. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 12281–12286 (2010).
219. Shima, S., Matsuoka, H., Iwamoto, T. & Sakai, H. Antimicrobial action of epsilon-poly-L-lysine. *J. Antibiot.* **37,** 1449–1455 (1984).
220. Young, J. W. *et al.* Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nat. Protocols* **7,** 80–88 (2012).
221. Crawford, R. *et al.* Long-Lived Intracellular Single-Molecule Fluorescence Using Electroporated Molecules. *Biophys. J.* **105,** 2439–2450 (2013).
222. Zaitoun, N. M. & Aqel, M. J. Survey on Image Segmentation Techniques. *Procedia Computer Science* **65,** 797–806 (2015).
223. Sliusarenko, O., Heinritz, J., Emonet, T. & Jacobs-Wagner, C. High-throughput, subpixel-precision analysis of bacterial morphogenesis and intracellular spatio-temporal dynamics. *Mol Microbiol* **80,** 612–627 (2011).
224. Chowdhury, S., Kandhavelu, M., Yli-Harja, O. & Ribeiro, A. S. Cell segmentation by multi-resolution analysis and maximum likelihood estimation (MAMLE). *BMC Bioinformatics* **14,** S8 (2013).

225. Stylianidou, S., Brennan, C., Nissen, S. B., Kuwada, N. J. & Wiggins, P. A. SuperSegger: robust image segmentation, analysis and lineage tracking of bacterial cells. *Molecular Microbiology* **102,** (2016).
226. Truglio, J. J., Croteau, D. L., Van Houten, B. & Kisker, C. Prokaryotic nucleotide excision repair: the UvrABC system. *Chem. Rev.* **106,** 233–252 (2006).
227. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* **17,** 239 (2016).
228. Feng, Y., Zhang, Y., Ying, C., Wang, D. & Du, C. Nanopore-based Fourth-generation DNA Sequencing Technology. *Genomics, Proteomics & Bioinformatics* **13,** 4–16 (2015).
229. Vasu, K. & Nagaraja, V. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiol Mol Biol Rev* **77,** 53–72 (2013).
230. Lewis, M. The lac repressor. *C. R. Biol* **328,** 521–548 (2005).
231. Berg, O. G. & von Hippel, P. H. Diffusion-Controlled Macromolecular Interactions. *Annual Review of Biophysics and Biophysical Chemistry* **14,** 131–158 (1985).
232. Halford, S. E. An end to 40 years of mistakes in DNA-protein association kinetics? *Biochemical Society Transactions* **37 (2),** 343 – 348 (2009).
233. Greene, E. C., Wind, S., Fazio, T., Gorman, J. & Visnapuu, M.-L. Chapter 14 - DNA Curtains for High-Throughput Single-Molecule Optical Imaging. in *Methods in Enzymology* (ed. Nils G. Walter) **472,** 293–315 (Academic Press, 2010).
234. Wang, J., Barnett, J. T., Pollard, M. R. & Kad, N. M. Integrating Optical Tweezers, DNA Tightropes, and Single-Molecule Fluorescence Imaging: Pitfalls and Traps. *Meth. Enzymol.* **582,** 171–192 (2017).