Rapid identification of fluorescently labelled DNA by image analysis

Jack Brazier

April 20, 2021

A masters thesis on the current DNA identification techniques and potential image analysis methods.

There are currently many DNA identification methods such as nanopore sequencing (identifying DNA sequences base by base) and restriction mapping (creating maps of DNA fragmented at certain sites). DNA identification is a tool that can be used for identifying diseases and species. The more frequently these methods are used, the quicker and cheaper they need to be. For example, in a disease outbreak many people may need to be tested. The current issue is that methods may cost over £100 per sample or produce results which do not contain enough data to be matched to a species. There is no method that can quickly (in one day or less) and cheaply ($<$ £100) identify viruses and bacteria in a sample of DNA.

Optical Mapping of fluorescently labelled DNA is a technique which costs no more than £1 per sample. It has the potential to produce enough data to match a variety of DNA in a sample (a swab from someones mouth) to a range of species (bacteria). This is a candidate for the solution to rapid diagnosis in hospitals.

Optical mapping is currently being developed and DNA can be analysed and matched to a species. However, this method only works for short DNA molecules (λ DNA is 48.5k base pairs long) and matches have only been made to 5 different species. My approach is working towards labelling and combing longer DNA, extracting the molecule's Fluorescent Intensity Profiles (FIPs) from the images, and matching a variety of DNA molecules from a species to its reference FIP sequence. The challenge of the current method is that long DNA molecules do not comb in a perfectly straight line and will cover multiple images. The new method aims to identify molecules which may be curved by using a different method of identification and extraction.

UNIVERSITY OF BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Contents

I Literature review **5**

1 Introduction **5**

1.1 DNA identification 5

1.2 Applications of DNA identification 5

2 History of DNA identification **6**

3 Current identification techniques **7**

3.1 Sequencing techniques 7

3.1.1 Sanger 7

3.1.2 Single molecule real time sequencing (SMRT) 8

3.1.3 Nanopore sequencing 8

3.1.4 Illumina sequencing 9

3.2 Optical Mapping 9

3.2.1 Restriction mapping 9

3.2.2 Fluorescent labelling 10

3.2.3 Maps from direct linear analysis 10

3.2.4 Methyltransferase reaction and fluorescent labelling 10

4 Extracting DNA data **11**

4.1 Image analysis 11

4.2 Locating DNA FIPs 11

4.3 Extracting DNA FIPs 13

5 FIP analysis **14**

5.1 Sampling 14

5.2 Alignment 15

5.2.1 Smith-Waterman 15

5.2.2 FASTA 16

5.2.3 Basic Local Alignment Search Tool 16

6 Applications **17**

6.1 Microbial identification 17

6.1.1 Diagnosis and disease 17

6.2 Analysing diverse ecosystems 17

7 Literature review conclusion **18**

II Results **19**

8 Introduction **19**

8.1 Motivation 20

9 Methods **23**

9.1 Labelling 23

9.2 Restriction test 23

9.3 Dye labeling and purification 24

9.4 Combing and visualisation 24

9.5 FIP extraction. 26

9.6 Existing algorithm 26

9.7 New method, locating FIPs 26

9.8 Extracting classified FIPs 34

10 Results and discussion **39**

10.1 Gel electrophoresis 39

10.2 Combing, visualisation 40

10.3 Analysis 43

10.4 Existing and new method comparison 45

10.5 Extracted data 47

10.6 Data analysis 51

10.7 Data sets with higher curvature 52

11 Conclusion **56**

Part I

Literature review

1 Introduction

Single molecule real time sequencing (SMRT) and Nanopore sequencing methods are effective in analysing a DNA strand one base at a time. Optical mapping is a method which examines modified DNA and produces an image with the intensity profiles for strands of a species DNA, all visible in one image, thus providing a map of a species DNA sequence.

1.1 DNA identification

Every species has its own unique DNA code sequence in its cells that acts as instructions for the construction of the organism. This could be described as a "barcode", which contains specific features and is unique for every species. The role of DNA is to store information and transfer genetic properties. In 1952 an experiment was performed by Hershey and Chase which demonstrated that the nucleic acid of a T2 Virus, not its protein, is transferred from the phage into the bacteria, the first proof of nucleic acid being responsible for the transfer of genetic properties.[1] More specifically, DNA contains small sections of base sequences called genes. A gene is a sequence nucleotide bases which are instructions for building the different proteins which make up the organism. DNA differentiates between viruses which means that even though they are not visible by eye they are each still identifiable in a person's DNA sample if a technique can analyse which DNA sequences are present.

1.2 Applications of DNA identification

The applications of DNA identification and mapping are primarily clinical. This is on different scales, from chromosomal down to base-by-base readings.[2] Basic methods such as fluorescent in situ hybridisation (FISH)[3] can identify targeted sequences in chromosomes making these sites visible under the microscope. These results can be screened to identify chromosomal conditions. When observing on a smaller scale, the specific sequence reads can be matched to known sequences or be used to generate genome mappings for a known sample. If specific bacteria can be identified in a sample of a person's DNA (such as from a swab check) then the antibiotics can be tailored to target those bacteria. Specialised treatments prevent bacteria adapting to defend against general antibiotics. If mutating bacteria is not tackled it will cause a societal and financial cost of US$100 trillion by

2050. This costs alongside untreatable infections can be avoided by $30 billion in proposed recommendations.[4] For disease outbreaks, cancer detection or analysing ecosystems, rapid methods need to be carried out many times for minimum cost. The benefits of optical mapping include cheap preparation and the ability to identify multiple sequences in a single set of images.

2 History of DNA identification

The first significant study of DNA in 1929 by Phoebus Levene [5] identified and classified the 4 bases A,C,G,T (Adenine, Cytosine, Guanine and Thymine) found in the DNA structure which provided the sequence with which the DNA is encoded. In addition, DNA is composed of a varying combination of nucleotide sequences, this was confirmed by Erwin Chargaff in 1950.[6] Chargaff also commented on the fact that DNA could and probably is responsible for the transfer of genetic properties. This indicates the key purpose of DNA being a store of information. One of the most significant discoveries was by Rosalind Franklin which included the application of x - ray diffraction methods on Sodium deoxyribose nucleate which indicated that DNA has similar structural properties to other molecules resembling a double helix.[7] Following this, in 1953 James Watson and Francis Crick published that the structure was a double helix polymer where base A is paired with base T and base C is paired with base G.[8] In 1956 an experiment in enzymatic synthesis was conducted by Kornberg. ATP was used in a reaction with a primer to produce long polymer chains, this paved the way for DNA amplification.[9] Amplification is an essential part of a variety of DNA identification techniques. In 1958, experiments indicated that the DNA replication process involves dividing the double helix and one strand going to each of the daughter molecules.[10] This was built upon in 1968 by Chargaff where direct base analysis and indirect analysis of the RNA products synthesized enzymically demonstrated that the transcripts of these single-stranded DNA fractions are indeed faithful copies.[11] This research confirmed that the ratio of A:T and C:G were the same respectively along both and each strand of DNA.

Through 40 years of discovery there is evidence for DNA being the store of genetic information which can be replicated and is composed of 4 nucleotide bases A, C, G and T. Scientists are continuing today to develop new methods to be able to read and identify DNA sequences as quickly and inexpensively as possible.

3 Current identification techniques

3.1 Sequencing techniques

In 1975 DNA sequences were determined in an experiment by primed synthesis. By using a primer AGAAATAAAA and the polymerase from bacteriophage T4 the experiment was able to deduce a sequence of 50 nucleotides in a few days.[12] By today's standards this is an incredibly slow method but determining a sequence revolutionised DNA identification.

Following this experiment was a polymerase chain reaction (PCR) method designed to amplify a small sequence of DNA. This application was able to diagnose sickle cell anemia by identifying the alleles responsible. An allele is a type of gene responsible for a feature like eye colour. In the allele responsible for the disease an A is replaced with a T. Such a small change was identifiable with amplification and sequencing. Amplification can make a species DNA easier to read. This method takes a long time to process but the ability to identify a certain allele in prenatal diagnosis provides not only a useful scientific and medical breakthrough, but also DNA amplification and identification techniques with potential application to other areas of molecular biology.[13]

3.1.1 Sanger

The primed synthesis experiment in 1975 [12] was improved by the use of chain terminating inhibitors which allows the DNA strands of increasing lengths to be identified according to which base pair they end with.[14] Chain terminating inhibitors are 2',3'-dideoxy and arabinonucleoside analogues of the normal deoxynucleoside triphosphates which terminate the polymerase at specific residues on newly synthesised chains. The sequence can be determined by knowing which strands end in which base pair and seeing how long each strand is through the experiment, Figure 1. The Sanger method is an established sequencing method that is still used today for producing exact reads with a maximum length of around 800bps.

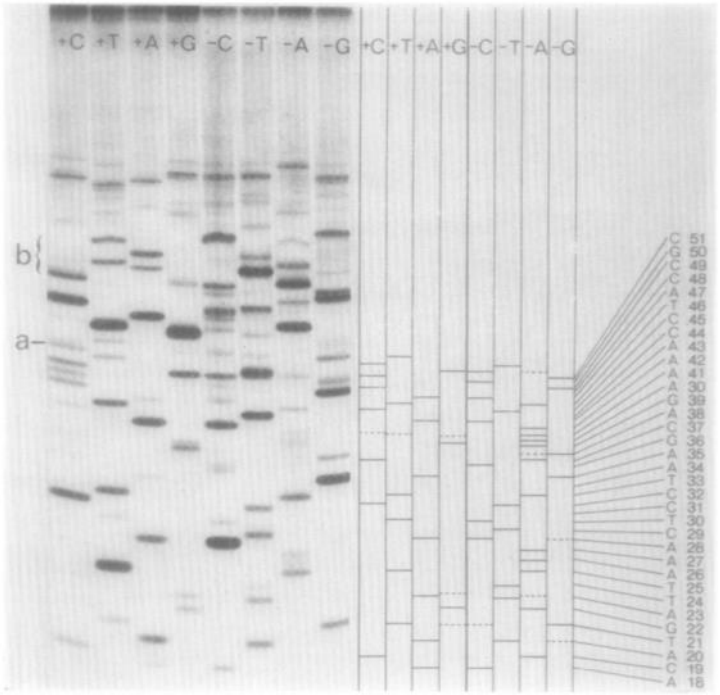


Figure 1: The results from a Sanger sequencing experiment from Sanger *et al* [12]. The code on the right displays the order of base pairs in the DNA.

3.1.2 Single molecule real time sequencing (SMRT)

SMRT makes use of phosphate labelled nucleotides with unique fluorophores for each nucleotide base. In a zero mode waveguide (small illuminated chamber) there is an immobilised polymerase enzyme and free labelled nucleotides.[15] A single strand of DNA is fed through the polymerase enzyme and as the DNA is passed through the labelled nucleotides are added and the fluorophores are cleaved off the phosphate group. These fluorophore signals are read by using an optical microscope in real time as the DNA is amplified by the enzyme.[16] In 2010 the speed was 1-3 bases per second. This is a parallelised method that can be run on thousands of individual molecules simultaneously.

3.1.3 Nanopore sequencing

The concept of identifying DNA sequences through nanopore sequencing involves applying a potential difference to DNA to feed it through a nanopore in a flow cell. The current being read is the movement of negatively charged ions moving through the pore, DNA blocks this flow of ions as it moves through the pore. The level of this current blockade is different for different bases. These can be processed by specific software to produce a genomic sequence for an organism.[17] This method is powerful for identifying DNA as

sequences for whole genomes can be accurately obtained for unmodified DNA samples. However the process faces challenges when dealing with rare secondary structures in DNA. In addition, this method for identifying a species from a sample is expensive costing approximately £100 per experiment in addition to the equipment cost. The data volume collected is many gigabytes per read. If used in mainstream microbial identification it requires storage in standardised useable formats.

3.1.4 Illumina sequencing

The most widely used method of sequence identification is through Illumina sequencing as this method is responsible for over 90% of the world's sequenced DNA data. Double stranded DNA (dsDNA) is cut and the ends are labelled with primers. Then through bridge amplification the DNA strands are fed into a flow cell which is covered with oligos and copies of the DNA are made into forward and reverse strands by polymerase. This amplifies the DNA and leaves forward and reverse strands, where reverse strands can be washed away. Tagged nucleotides for each base pair are added one at a time and the identity of a given base is read from a fluorescent signal. Finally, the reads of overlapping fragments are matched up to form contigs (contiguous DNA fragments) which creates a consensus code (reference 'barcode') for the DNA. Billions of bases can be read per day using the illumina platform. This method of identification is accurate, however, the read time is dependent on the length of the sequence being read. In addition, the experiment is expensive costing over £170 per flow cell sample. For a shotgun sequence analysis method (overlapping reads of lots of DNA strands) the capabilities are accurate to 0.01% error.[18]

3.2 Optical Mapping

Optical mapping uses images of DNA strands to produce maps of the species genome. When DNA strands are measured by lengths the map is called a restriction map, these are usually made by digesting DNA with restriction enzymes. This method can produce results for multiple reads at a time and the fragments can be aligned to form a consensus DNA 'barcode'.

3.2.1 Restriction mapping

The early methods of optical mapping were performed by digesting DNA using enzymes and this would produce a restriction map where fragments are of different lengths depending on where they have been cut. This is done at specific sequences. The images of DNA are produced where the sizes of the strands will indicate which part of the genome was present. The size of the DNA molecule between cut sites is determined using the total fluorescence intensity in the image of the molecule between those sites. This is possible

because the dye used to stain DNA binds non-specifically and uniformly along the length. A longer image can be produced while preserving the pixel registration. This creates a consensus barcode reference without need for PCR or electrophesis. This method is still improving and in 1999 it was said that if Optical Mapping can be done quickly we can identify whole genomes.[19]

3.2.2 Fluorescent labelling

Around 10 years later this method had already undergone significant improvements whereby restriction enzymes are replaced with methyltransferase labelling of DNA. This enables the strands of DNA to be longer and eliminates error from damaged DNA. In addition, certain sites in dsDNA can be labelled with fluorophores. The pattern of these labelled sites will be visible under a microscope which can be matched to a reference genome.[20] This is the drive for my project as this experiment is being developed to be able to detect specific genomes in a sample and be reproducible for many samples at a low cost ($<$ £1 for reagents per sample).

3.2.3 Maps from direct linear analysis

Direct linear analysis (DLA) is suitable for long reads of DNA strands as it measures the distance between motif tags (8 or 7 base long DNA sequence specific motifs). The distances measured for a λ phage genome of 48.5kb pairs have an accuracy of $\pm$0.8kb pairs, while localisation has an accuracy of $\pm$2kb pairs. Strands are fed through a receiver and an intensity profile signal produced shows the distance between the sequence motifs. The strands are tagged with sequence specific motifs that are constructed by peptide nucleic acids. The disadvantage is that each strand needs to be fully stretched to get an accurate reading of distances between sequence motifs.[21]

3.2.4 Methyltransferase reaction and fluorescent labelling

DNA labelling is performed using a methyltransferase reaction to transfer a functional group from a co-factor to a site on a DNA molecule (Figure 4).[22] The labelling in this experiment is performed using the enzyme M.TaqI and the co-factor AdoHcy-6-N3 to transfer the azide functional group to the A base in the sequence 5'-TCG<u>A</u>-3'.

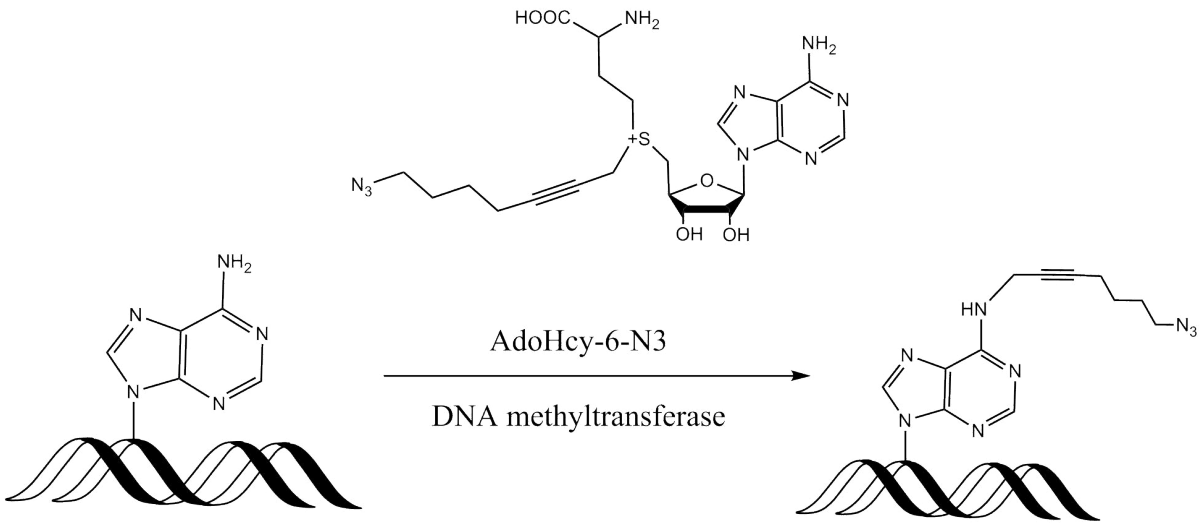


Figure 2: Methyltransferase reaction

4 Extracting DNA data

Moving on from DNA identification methods, this discussion shifts to the different methods available for extracting data from an image.

The method used in my experiment produces images of fluorescently labelled DNA where the data takes the form of an intensity profile. The DNA data in the images will be referred to as a DNA Fluorescent Intensity Profile (FIP).

4.1 Image analysis

The leading technique for image classification is a deep residual learning algorithm, this makes the use of convolutions on the image through 30 to 50 layers to detect the objects in an image.[23] This demonstrates that the use of convolutions is currently the best method we have for object detection in images.

4.2 Locating DNA FIPs

The current mathematical method for detecting shapes in an image is the general Hough transform. This transform converts a 2D image into a graph with the angle on the x axis and the distance on the y. Many lines are drawn through each point in the image. The distances and angles from the lines to the origin are plotted on a graph. The most common angles and distances can be seen from a bright spot in Figure 3 due to the

intersections at common values. This bright intersection indicates the most common angle that the points follow along in the image.

Lines are plotted through the data points and the distance r and angle θ are calculated by drawing a line from the origin to the line plots.

$$r = x\cos(\theta) + y\sin(\theta) \tag{1}$$

This technique was patented in 1972[24] and generalised 10 years later[25], while still being used today.

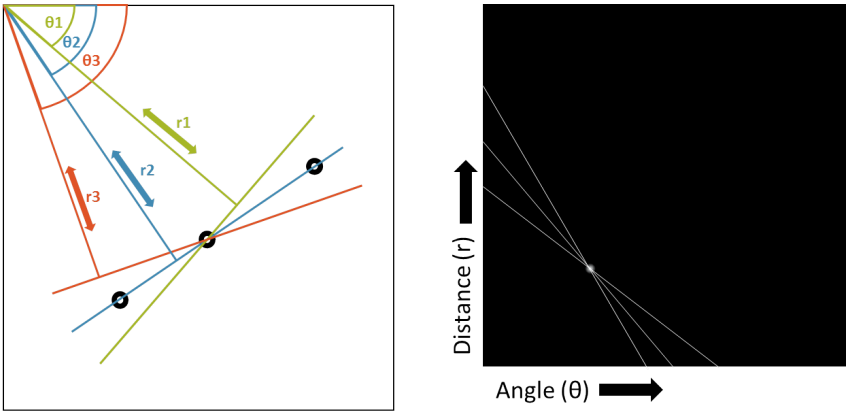


Figure 3: Example of a Hough transform where data points are on the left where lines are plotted, on the right is a map where the high density of points indicate the angle that the data points follow along. The right image displays a point of high intensity which indicates a common angle across a line that the data points are following.

While the Hough transform can indicate the angle and shapes of objects in an image the quality of the image may not be good enough to be able to read the data because of noise. The Gabor filter can be used here to improve clarity of the image, for example a fingerprint.[26] By applying the Gabor filter the shapes in an image following a pattern can be enhanced. The lines in a fingerprint can be highlighted among noise in an image. The 2D Gabor filter $g(x, y)$, can be written as a combination of a 2D elliptic centered Gaussian distribution $w(x, y)$, multiplied by a sinusoidal plane wave function $m(x, y)$.

The 2D elliptic Gaussian is given in Cartesian coordinates as:

$$w(x, y) = \exp\left[-\frac{1}{2}\left(\frac{x^2}{a^2} + \frac{y^2}{b^2}\right)\right] \tag{2}$$

Then by applying spacial offsets x_0, y_0 to get translation t:

$$x_t = x - x_0, \qquad y_t = y - y_0 \tag{3}$$

Then by rotation of the distribution by counter clockwise angle A using offsets x_0, y_0 we have:

$$x_g = x_t\cos(A) - y_t\sin(A), \qquad y_g = x_t\sin(A) + y_t\cos(A) \tag{4}$$

Then finally by substituting x for x_g (a^2, b^2 being variances in x_g and y_g directions) we have the elliptic Gaussian as:

$$w(x, y) = \exp\left[-\frac{1}{2}\left(\frac{x_g^2}{a^2} + \frac{y_g^2}{b^2}\right)\right] \tag{5}$$

Next we define the wave function as:

$$m(x, y) = \cos\left[-2\pi i(U_0 x + V_0 y) - P\right] \tag{6}$$

Where U_0, V_0 are frequencies and the phase shift $P = (U_0 x_m + V_0 y_m)$ where x_m, y_m are displacements from the center of the Gaussian. In Matlab P is defined as ψ.

The Gabor filter can now be written as a product of a 2D elliptic Gaussian and a sinusoidal plane wave:

$$g(x, y) = \exp\left[-\frac{1}{2}\left(\frac{x_g^2}{a^2} + \frac{y_g^2}{b^2}\right)\right] \times \cos\left[-2\pi i(U_0 x + V_0 y) - P\right] \tag{7}$$

The 2D Gabor filter above is described in.[27]

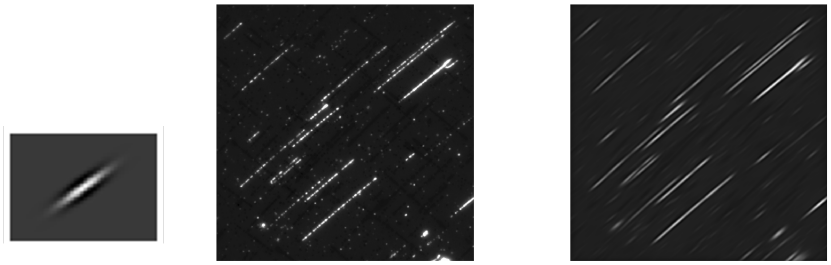


Figure 4: Left is an example of a Gabor filter (theta = 0.8727; lambda = 10; gamma = 0.3; sigma = 2.8; psi = 0;). This is applied to an image of combed fluorescently labelled T7 DNA (middle). Resulting in a filtered image (right).

4.3 Extracting DNA FIPs

One technique [28] which is an alternative to the Hough transform is a feature classification function. This method works by plotting intensity profile lines across the image horizontally. Feature classification is used for locating barcodes in an image, see figure 5. The regions of the image where the barcodes are located are depicted by intensity profiles which have areas of rapidly varying intensities. Then the method will analyse the quality of the image by using resembling barcode measure matrix (RBM). The RBM is a matrix of vectors derived from the intensity profiles. This is followed by a binarisation and then an algorithm will read the barcode (the algorithm chosen is dependent on the quality of the image given by the RBM). This demonstrates a smart way of reading barcodes and gives a new method and alternative to using the Hough transform.[28] This could be

a future method for image analysis where the areas in an image where the FIPs pass through can be identified by changes in an intensity profile across the image.

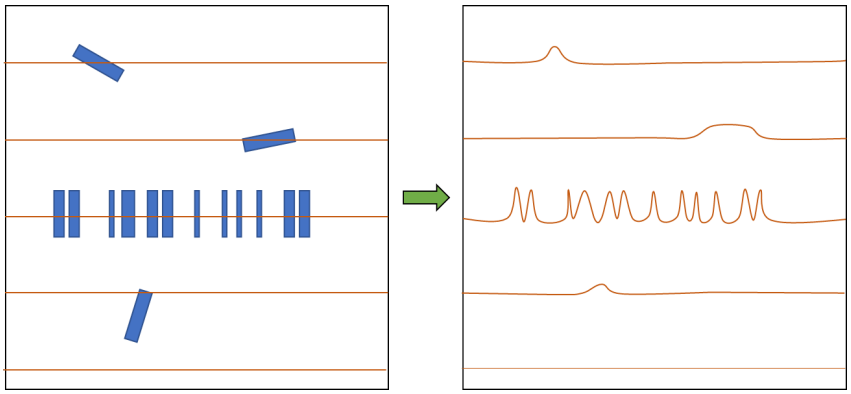


Figure 5: An example of a feature classification method of plotting lines through the image and looking at the intensity profiles to be able to locate a barcode in an image.

5 FIP analysis

After extracting FIPs from a stack of images they need to be matched to a reference DNA sequence, this allows species to be identified by FIPs. There are currently many methods for analysing intensity profiles and signals. In addition, there are algorithms designed to match DNA sequences. These methods are explored below.

5.1 Sampling

Dispersive Fourier transforms are widely used when a fast continuous measurement is needed, making a fast analysis of an incoming signal.[29] In barcode readers (for example in a supermarket) the spectrometer will return an optical pulse in the frequency domain which is converted to a signal in the time domain by Dispersive Fourier Transform (DFT), representative of the barcode. The acquisition time for this type of barcode reader is 25MHz.[30]

A spectrogram is a mapping of a signal which plots a time-amplitude signal in terms of time, frequency and intensity. One application of this method is in speech recognition, it is an effective analysis method for this type of signal as the individual frequencies in the signal are separated to be able to eliminate noise.

One form of robust speech recognition starts with a sampling of 20ms frames with a 10ms overlap. This is followed by a fast Fourier transform on each frame which magnitudes are squared to then produce the power spectrum. The spectra are then warped by a statistical prediction model called all-pole auto-regression (all-pole AR). The AR

model is used to predict outcomes and this can make an approximation of the spectra, while the poles will give indication to where the function is at 0 therefore defining the main peaks in the prediction model.

From the warped spectra the linear coefficients are converted to cepstral coefficients. Cepstral coefficients are obtained from the inverse Fourier transform of the log of the magnitude of the Fourier transform. These coefficients are real and even packets of information providing a robust speech recognition method.[31]

A spectrogram can then be converted into a modulation spectrogram where the peaks can be obtained. This modulation is done by performing a fast Fourier transform on a Hamming window of the spectrogram. From here filters and energy peaks are used in the final image. This gives a clear spectrogram where the noise can be eliminated.[32]

5.2 Alignment

After the DNA FIPs have been extracted they need to be aligned to the correct reference FIP. This needs to be fast and accurate as in a disease outbreak it is critical that the extracted FIP is matched to the cause of a disease as soon as possible.

The alignment analysis of DNA sequences consists of both optimal and heuristic methods. The first being the Smith-Waterman[33] method using matrices to find exact alignments, followed by the FASTA[34] and BLAST[35] programs and the more recent USEARCH[36] algorithm. The newer search methods have a more general approach by finding the most similar matches rather than the exact alignment. This saves an enormous amount of computation time for a result that needs to be quick whilst still providing useful results.

5.2.1 Smith-Waterman

This algorithm is primarily used in the identification of matches between molecular subsequences. When comparing two sequences it can be difficult to know where the sequences align and if there are any deleted nucleotide bases. This algorithm can find where the sequences align and can compensate for gaps in the sequences.

A similarity score is given to the elements in both sequences of bases. The score is a similarity between bases a, b from sequences A, B. $a \in A$ and $b \in B$ where $A = a_1, a_2, ...a_n$ and $B = b_1, b_2, ...b_n$.

A matrix is formed with one sequence along each axis and the similarity score is given to a match between elements a, b in sequences A, B. The matrix is filled in from the top left corner to the bottom right corner, each score is dependent on the scores above and before the score being observed. So as similarities increase a score and differences give

a penalty, if there is a matching sequence it will follow a line of highest values in the matrix. These high scores can be traced back from the bottom right to the top left giving a sequence of similarities between both DNA sequences.[33]

Although the Smith-Waterman method performs well at aligning sequences, the computation time increases exponentially as the sequence/dataset size increases. For sequence length A and B the matrix size is $A \times B$. Double the sequence size and the matrix is 4 times bigger. To emphasise this with 4 times longer sequences: $(4A \times 4B = 16(A \times B))$. Therefore, this algorithm only works best for short alignments of sequences.

5.2.2 FASTA

The FASTA algorithm was developed several years after the Smith-Waterman algorithm to help deal with longer sequence comparisons without sacrificing computational time.

First the protein alignment version called FASTP detects similarity between protein sequences by scoring based on the the Point accepted mutation (PAM) matrix or BLOcks SUbstitution Matrix (BLOSUM). These give similarity scores for proteins based on the difference between the three out of the 20 amino acids needed for each protein.

FASTA is for aligning DNA and as it is based on similarity scores it is a heuristic algorithm. This does not perfectly align sequences as the Smith-Waterman algorithm but does align sequences quicker by finding similarity.

From scoring, the similarities in the sequences are found and then the longest length of high scoring sub-sequences are selected based on a threshold value. These subsequences are then connected to find where the original sequences are aligned. For high density regions of one nucleotide, a Monte Carlo (random sampling) method can be used to shuffle the high density regions and remove the bias.[34] In addition, the Smith-Waterman algorithm can be used on the similarity of the subsequences.

For alignment, a heuristic algorithm can therefore be just as effective while saving huge amounts of computation time.

5.2.3 Basic Local Alignment Search Tool

The Basic Local Alignment Search Tool (BLAST) is a heuristic algorithm designed to match DNA sequences as quickly as possible. The program will optimise a measure of local similarity, this is done by calculating a Maximal Segment Pair (MSP) score. A segment pair is a pair of letters, for example: VALLAR and PAMMAR. There is a similarity between these segment pairs and a MSP is a segment pair of maximum similarity and score. Both sequences are compared and identities are given a high score where mismatches are given a negative score. From the identical length segments the one

with the highest score is named the MSP.[35]

6 Applications

6.1 Microbial identification

Methods for identifying viruses or bacteria in a sample include staining and looking at them under a microscope and looking at the culture of the cells. When these cultures are not distinguishable optical mapping offers a way to observe multiple microbes at a time and match each one to its reference.

6.1.1 Diagnosis and disease

In terms of detection by DNA analysis the main method is 16s ribosomal RNA qPCR sequencing. This is the gold standard where the RNA used for bacterial identification is converted to DNA using reverse transcription This has been particularly useful in outbreaks all over the world, such as detecting Plasmodium falciparum in 2004-2005. The particular system used for this example was GreeneChip, which includes a "panmicrobial" genetic database of over 29,000 probes and protocols.[37]

This literature outlines the importance of early disease detection by using DNA analysis techniques. New diseases need certain methods to combat them so it is important to know what virus or bacteria is causing the problem to combat it directly.

6.2 Analysing diverse ecosystems

Environmental DNA (eDNA) can be extracted in a variety of ways and indicates which species are present in an ecosystem or what the animals are consuming. eDNA is any DNA source that is not intentionally preserved for genetic analysis. One study considered three examples of eDNA analysis. They first took a water sample and through qPCR, Illumina sequencing and by aligning the sequence with BLAST were able to identify Carp being present. The next example took leeches and by using mammal primers was able to amplify the DNA and sequence it using Sanger (chain terminating) sequence. The final example investigated which insects a bat had eaten and needed to use 30 different tags in a PCR reaction, following a sequencing technique called Roche FLX which is used after 2016.[38] This demonstrates the large variety of identification methods being used already.

My optical mapping experiment would be able to identify many varieties of eukaryotes or

prokaryotes in a sample therefore could be applied to many eDNA analysis experiments in blood, water or soil.

7 Literature review conclusion

To conclude the review of the literature there is an extensive history of research over the past 90 years into what DNA is used for, what its structure is and what it is made of, leading up to my research today which will use the properties of DNA to be able to make a visible pattern in an image indicative of the unique species that it belongs to. My research in the following chapters demonstrates the new image analysis methods used to enable localisation and extraction of fluorescently labelled DNA, not only when it is combed in a straight line but also in a curved line.

Part II

Results

8 Introduction

This report describes experimental techniques to label bacteriophage lambda genomic DNA with fluorophores (azide group is used to attach fluorophore dye via click chemistry) at the sequence specific sites. This is followed by combing of a droplet containing the dye labelled DNA onto a glass slide. These sites are visible when exposing the fluorophores to a laser under a microscope producing fluorescence images where the sites are visible as patterns in the image (Figure 6). Primarily, different algorithms have been developed for extracting the fluorophore patterns as DNA Fluorescent Intensity Profiles (FIPs) from fluorescence images of DNA. By extracting the FIPs of sequence specific labelled DNA, the pattern of these sites can be visualised and used for sensitive identification of DNA sequences.

This report describes the experiment for labelling, combing and imaging of DNA to produce fluorescence images. Subsequently, a description of analysis and algorithms are given to demonstrate how to extract the most FIPs from a fluorescence image using MATLAB.

A FIP is the type of data extracted from the images. The name FIP is used because the fluorescence pattern is in the form of a non-binary intensity profile where higher intensity is produced from a higher concentration of fluorophores.

The resolution of a FIP is related to the numerical aperture of the objective being used, this is due to diffraction. Abbe's law states that the diffraction limit 'r' for a wavelength 'λ' and a numerical aperture 'NA' is:

$$r = \frac{\lambda}{2NA} \tag{8}$$

The objective used in the experiment has a numerical aperture of 1.49 and the laser has a wavelength of 640nm, therefore, the limit for two fluorophores to be distinguishable is $\frac{640\text{nm}}{2\times1.49} = 215\text{nm}$.

A DNA base pair has a unit length of 0.34nm and the pixel size used in the experiments for visualising lambda DNA is 160nm/pixel. Due to the force as the molecule leaves the droplet, the DNA is being stretched by 1.60 times so the measured bp length is 1.60×0.34nm = 0.54nm. Therefore, the number of base pairs between resolvable fluorophores is $\frac{214.77\text{nm}}{0.54\text{nm/bp}} = 398\text{bps}$.

The ultimate aim is to detect multiple strains of bacteria and virus in a sample of a pa-

tient's DNA in a hospital, in order to provide effective treatment and specific antibiotics. In addition, it is contributing to the first imaging technique to map the human genome.

So far the work on optical mapping[22] has been able to simulate and identify bacteriophages T7 and lambda in a mixture of 20 bacteriophages. Experimentally, bacteriophages T7 and lambda, and the Ad12 adenovirus have been labelled, extracted and aligned to reference genomes. Lambda DNA has been identified in a mixture of host genomic DNA (*E. coli*i ER2566)[20]. Simulated labelled *E. coli* has been identified in background DNA. In addition, pure sample of M.taqI directed fluorescently labelled *E. coli* DH10B has been combed and fragments have been aligned for almost full coverage of its genome.

The fluorophore labels are in the same position on either strand of the DNA as the complement to 5'-TCGA-3' is 3'-AGCT-5' which read backwards is 5'-TCGA-3'. Therefore, if both sides of the DNA are fully labelled it would produce the same pattern as if only one side were.

8.1 Motivation

The limitations of the sequencing methods such as nanopore sequencing are that they use expensive flow cells which cost at least £100, and if there is an unusual structure in the DNA it will not fit through the nanopore. In mapping the main equipment required is a microscope (a one time purchase) and then the reagents and glass slides will cost less than £1 per sample. The DNA can be also be combed regardless of the structure, another advantage over sequencing methods.

The FIPs displayed in the fluorescence images are not always straight, especially for the longer DNA molecules. This is because during combing molecules do not adsorb in a straight line and this is intensified with longer molecules. DNA combing is when a droplet of water containing labelled DNA strands is pippeted onto a glass slide and dragged (combed) across the slide depositing the DNA strands as it moves along. As DNA is combed the part of the molecule left in the meniscus will apply a force perpendicular to the movement of the droplet. Longer FIPs are important as they contain more data and give a larger pattern to reference.

Current analysis has only stretched as far as fragment lengths as large as bacteriophages, these DNA molecules (lambda = 48.5kbps), can fit in one 512×512 pixel image. The largest possible molecule in one image would be:

$$\sqrt{512^2 + 512^2} = 724 \text{ pixels}, \ 160\text{nm/pixel} \times 724 \text{ pixels} = 116,000\text{nm},$$
$$\frac{116,000\text{nm}}{1.60 \text{ stretch} \times 0.34\text{nm/bp}} = 213,000\text{bps}.$$

This is limiting to the sizes of FIPs we can analyse, therefore, future methods will impose identification of curved FIPs over multiple images. Figure 6 demonstrates a curved FIP. Before analysis, the FIP is extracted from the image by an existing algorithm which plots

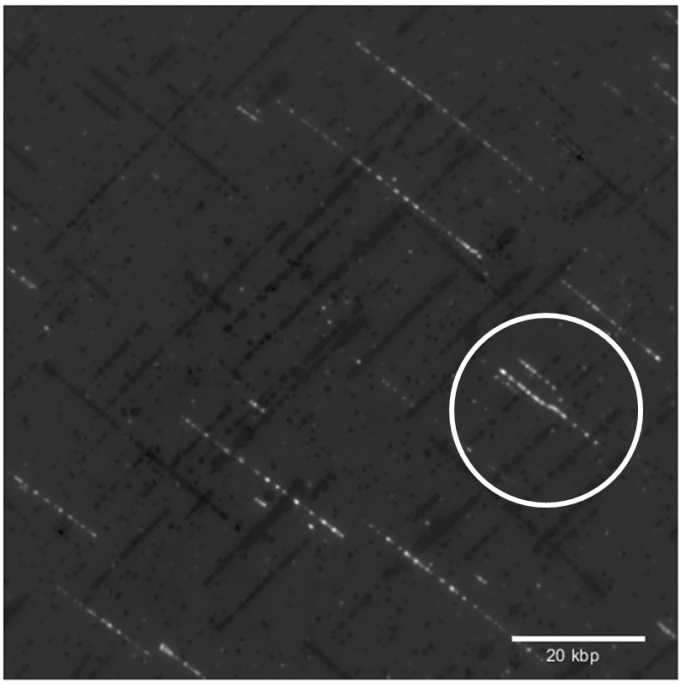


Figure 6: T7 DNA FIPs. Within the white circle is an extreme example of a curved FIP. The DNA FIPs in the image are going from the top left of the image to the bottom right. The circled FIP is a strand which is bound at both ends (top left) and it has forked downwards and does not follow a straight line. Long molecule fragments can have a slight curvature.

straight lines through FIPs in each image. The input is a .tif image stack and the output is an array containing the extracted FIPs (for output array see Figure 19).

The image of a curved FIP (FIP in the centre of Figure 7) indicates the need for an extraction algorithm to be able to extract the FIP along a curved line if all data is to be extracted for analysis.

The motivation for writing a new extraction algorithm is that the existing method extracts the FIP from an image by plotting a straight line from the start to the end, this method works if they are perfectly straight, but otherwise results in a loss of data. Slight FIP curvature with straight line extraction will mean that some of the FIP is lost as the true intensity pattern is through the centre of the FIP. This is demonstrated in Figure 7. It is clear that the algorithm is able to pinpoint the location of fluorophores in an image, however, the absence of a green line over the curved FIP in the centre indicates the need for an algorithm more suited to this kind of pattern.

The reason for the development of a new method is that the current method was unable

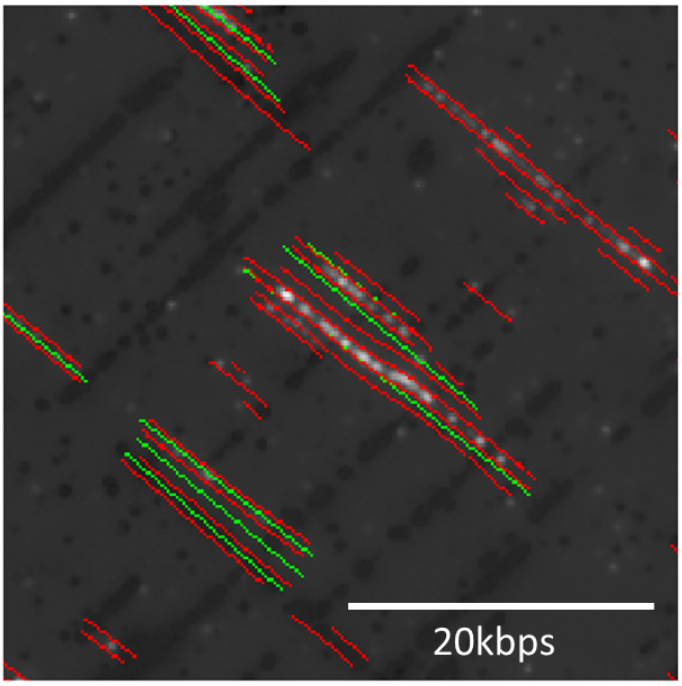


Figure 7: Lambda DNA FIPs with red lines depicting the outlines where the FIPs are located and green lines demonstrating where the data is extracted using existing method. Green lines are drawn in between red lines.

to extract the curved FIPs from the images. The aim is to extract all FIPs from an image so when they are aligned there is no DNA that is not analysed from a sample. All extracted data must be true to the data in the image as lost or incorrect data can misalign fragments to the wrong genome. This would lead to inaccurate diagnosis of a disease in a sample of genetic material.

In summary, optical mapping has huge advantages over sequencing techniques in terms of the simplicity of the data aquired and the economical reproducibility of the experiment. However, after exploring the methods used it is clear that the experiment and algorithm need to be optimised for full fast and accurate identification of viruses and bacteria.

9 Methods

9.1 Labelling

A solution (100 μl) of DNA was prepared from 4 μl bacteriophage Lambda DNA (500 ng/μl), 82 μl water (Sigma W4502 molecular biology reagent, double distilled, deionised, 0.1 μm filtered) , 10 μl CutSmart[39](NEB) buffer, 2 μl of the enzyme M.TaqI (1 mg/μl) and 2 μl of cofactor AdoHcy-6-N_3 (MTC-5) (15 mM). After preparation, the DNA solution was left to incubate at 50°C for two hours. Final concentration of DNA: 20 ng/μl. Next, 2 μl Proteinase K (20 mg/ml) was added to azide labelled DNA solution to digest the M.TaqI enzyme, then incubated for 1 hour at 50°C. The DNA was then purified by ethanol precipitation. To the solution of DNA, 0.1x volume NaCl (3M, 10 μl) and 3x volume EtOH (100% 330 μl) was added, then centrifuged for 16,000 rcf for 20 minutes to form a pellet. The pellet was washed 3 times with 200 μl EtOH (70%), then re-suspended in 50 μl H_2O.

9.2 Restriction test

In the restriction test, digestion with R.TaqI (NEB) was performed on M.TaqI (NEB) directed labelled DNA to determine labelling efficiency.

Four samples were prepared, Table 1, (all measurements in μl).

Restriction test samples				
Sample	(1)	(2)	(3)	(4)
H_2O	8.5	8.5	6.0	4.0
CutSmart (10x)	1.0	1.0	1.0	1.0
Unmodified DNA	0.5	0.5	-	-
Labelled DNA	-	-	3.0	5.0
R.TaqI	-	0.3	-	0.3

Table 1: Components used in four restriction test samples.

After preparation the samples were left to incubate at 50°C for 1 hour.

A 1% agarose gel was prepared with 0.61 g of agarose, 80 ml of TAE buffer and 2 μl of gel red. 2.5 μl of 5× purple loading dye (NEB) was added to each sample. 10 μl of ladder and 10 μl of each sample was pipetted to each of the first 5 lanes in the gel. Lane 1 contained the ladder and lane 2, 3, 4, 5 contained samples 1, 2, 3, 4 respectively. Gel electrophoresis was performed at 120V for 50 minutes while submerged in TAE buffer. The restriction test proved that DNA was partially labelled, yet sufficiently labelled as to continue with the experiment.

9.3 Dye labeling and purification

Dibenzocyclooctyne (DBCO) dye from NHS-ester dye was prepared with 2 μl PBS 1x, 0.68 μl DMSO, 0.60 μl DBCO - amine (20 mM in DMSO), 0.72 μl ATTO647N-NHS-ester dye (50 mM) in DMSO.

5 μl of the DBCO dye was added to the DNA solution where it was left to incubate at room temperature overnight.

Next, the dye labelled DNA was purified using the Sigma-Aldrich NA2110-KT GenElute Bacterial Genomic DNA kit with the following modifications. Wash 2-3x with ethanol wash buffer, warm elution buffer to 50°C.

9.4 Combing and visualisation

To prepare hydrophobic Zeonex-coated coverslips, glass coverslips (20×20 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. The slip was spun at 3000 rpm for 90s. Zeonex-coated coverslips were allowed to dry at room temperature overnight and stored in a desiccator.

DNA is diluted to ($\leq$ 200 pg/μl) with phosphate (combing) buffer. Then a 10 μl droplet containing DNA is dragged (combed) across a zeonex coated coverslip at a constant speed (5mm/minute) from an Arduino motor over a distance of 10mm.

A widefield/Total Internal Reflection Fluorescence (TIRF) microscope was used for imaging. The objective used is a Nikon TIRF 100× magnification with a numerical aperture of 1.49. The setup includes 4 lasers at 405, 488, 561 and 640nm. The camera used is an Evolve 512 Delta, 20MHz 512×512 EM-CCD. The camera's pixel size is 160nm/pixel. Images are captured and processed using ImageJ[40].

The laser was picked according to the dye used, for ATTO647N dye the laser used was 640nm wavelength. The surface of the coverslip was located and then 10 fluorescence pictures were taken of the fluorophores at an exposure time of 10ms. The images displayed the presence of DNA molecules from the FIPs produced from the fluorophores on the coverslip. Microscope was controlled by Micromanager. Images were stored as .tif files at 512×512 pixels at 81,920nm×81,920nm in size.

9.5 FIP extraction.

This section describes automated extraction and alignment algorithms written in MATLAB (R2018a, The MathWorks, Inc., Natick, MA, USA).

The image analysis consists of extracting the maximum number of FIPs from the images followed by aligning them to the reference FIP. Extraction returns a FIP while preserving the shape and length of the fragment in the image. The reference is modelled by applying a point spread function to the sequence of 5'-TCGA-3' sites from the genome on the GenBank DNA database. The current methods include MATLAB codes that extract the FIPs along a straight line in each image, followed by choosing the reference genome from GenBank and using a cross correlation function (a measurement of displacement and similarity) to align the FIP fragments to the reference genome.

In both extraction methods a mask is constructed to overlay the original image where FIPs are located, and then the pixel intensity values along these lines are extracted.

9.6 Existing algorithm

The existing method[1] begins by estimating the combing angle (angle of the FIP's direction in the image) by calculating the angle of the data points in an image. This is done by using a Hough transform. The value for the combing angle is used in a Gabor filter convolution that blurs the image in the direction of the FIPs so that the FIPs become solid lines. These lines are depicted using a Sobel edge detector which then gives an outline of where FIPs are located (red lines). Finally, by merging the close ends of the edges, a straight line (green) is plotted in between the red lines and the pixel intensities along the green lines are extracted FIPs.

This method is very good at identifying where straight lines are located in the image, however, as the same angle from the Hough transform is applied to every FIP, if the FIP is curved then this angle will not be characteristic of part of or all of this FIP.

9.7 New method, locating FIPs

In order to capture data from more DNA molecules, we aimed to develop an approach where DNA deposition angle was not assumed to be uniform across the entire image.

The first step of image analysis is locating the position of the FIPs in each image. By classifying each FIP into an array of coordinates for the brightest points along a FIP its location can be defined. The method for locating FIPs is below:

1. Blur image irrespective of direction
2. Identify location of FIP shaped structures
3. Locate the brightest pixels in these structures
4. Group brightest pixels

The first step in the new algorithm is to make the FIPs more like continuous lines and remove noise that is not a FIP. This step will blur the image similar to the Gabor filter convolution, but in a way that can be applied irrespective of the direction of the FIPs. This is done by using the MATLAB functions *imresize* and *csaps*. The converted image is going to be used as a mask over the original image, no data is lost as the original image is kept separate.

The function *csaps* smooths a matrix which eliminates noise. In addition, there are gaps in between fluorophores in the FIPs, for locating the FIP it helps if the gaps are reduced. The functions *imresize* and *csaps* remove detail from the image and reduce the gaps between fluoprophores. The *csaps* function is computationally expensive, so reducing the size of the image saves a significant (4 times) amount of time. Finally, the height and width of the image is halved because after testing the image loses too much information if shrunk further.

The *imresize* function uses a matrix, a scale and a resizing method as input parameters, the matrix is the fluorescence image, the length scale factor used is 2:1 and 1:2 as the image is halved and then doubled in size along each dimension (half x and y is a quarter area scale factor). 2:1 is chosen as the scale because its the maximum size change which allows the FIPs to preserve their shape. The resizing method is set to the default which is a cubic interpolation, this is chosen as the new data points are calculated over a cubic function which is better fitting than a linear function (Figure 8).

To summarise, the fastest and most effective method to blurring the FIPs together in the image (irrespective of direction) is by reducing it by the largest proportion that doesnt lose too much data (one quarter of the area) in order to quickly perform the *csaps* method. Compared to *csaps*, *imresize* takes a negligable amount of time. During this process the backround noise is also removed which helps the algorithm only pick up DNA FIPs later on.

The csaps function is a cubic smoothing spline that minimises the cubic spline interpolation (left side weighted by p) and straight line fit (right side weighted by $(1-p)$) in a vector or matrix for spline s, equation below. Where each value $x(j)$ approximates the value $y(:,j)$, for $j = 1$:length(x). $Y_i = f(x_i) + \epsilon_i$. ϵ_i is a random independent variable. A higher value of p will produce a curve that fits closer to the y values, and a low p value will be closer to a line of best fit through the y values.

$$p\sum_{i=1}^{n}(Y_i - f(x_i))^2 + (1-p)\int\left(\frac{d^2f}{dx^2}\right)^2 dx. \tag{9}$$

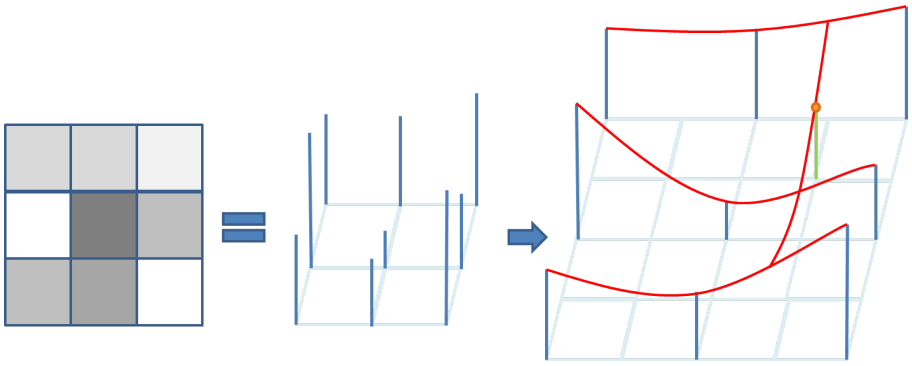


Figure 8: Schematic demonstrating the cubic interpolation used in the imresize function. Left: a grid of pixels where low to high intensity is black to white. Middle: a representation of the grid where each blue line length represents the intensity of the pixels in the image. Right: the curved red lines are cubic lines plotted in between the pixels; the orange circle marks the position of a new pixel where its intensity is given by the length of the straight green line below it.

In two dimensions the function smoothes the image by changing the intensity values at each pixel. When $p = 1$ the function will return the original image. When $p = 0$ the function will produce a 2D image with a least squares line of best fit through the original intensity values (a 2D linear gradient from 0 to 1 in one direction). The equation for 2D csaps function is expressed below. I is the function for the intensity value at a given pixel at coordinate (x_i, y_i).

$$p\sum_{i=1}^{n}(I_i - I(x_i, y_i))^2 + (1-p)\int\left[\left(\frac{\partial^2 I}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 I}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 I}{\partial y^2}\right)^2\right] dx\ dy. \qquad (10)$$

This is applied to an image to blur the points evenly (the sharpness is reduced). A matrix of 1s the size of the image is the input for the x, y values where each pixel intensity value is evaluated at that point. The *csaps* function is computationally expensive so the image is reduced to a quarter of the area to save time.

When the image is reduced in size the *csaps* function is applied and then the image size is increased again. This results in an image which has FIPs with smaller gaps in between the fluorophores and a reduction in background noise (Figure 9).

The algorithm follows with a minimum pixel intensity threshold (any pixel with an intensity lower than this threshold is set to 0 intensity). The threshold is calculated by the average intensity of pixels over each image.

A convolution grid that highlights areas of the image where there are bright pixels arranged in a line is applied next. The grid is 13×13 pixels to match the diameter of the fluorophore spots (measured by eye), schematic for grid in Figure 10. The grid is centred at each pixel in the image. The convolution will detect brightest and darkest pixels for every row and column within the grid, and then see if these points are arranged at the same angle (figure 11).

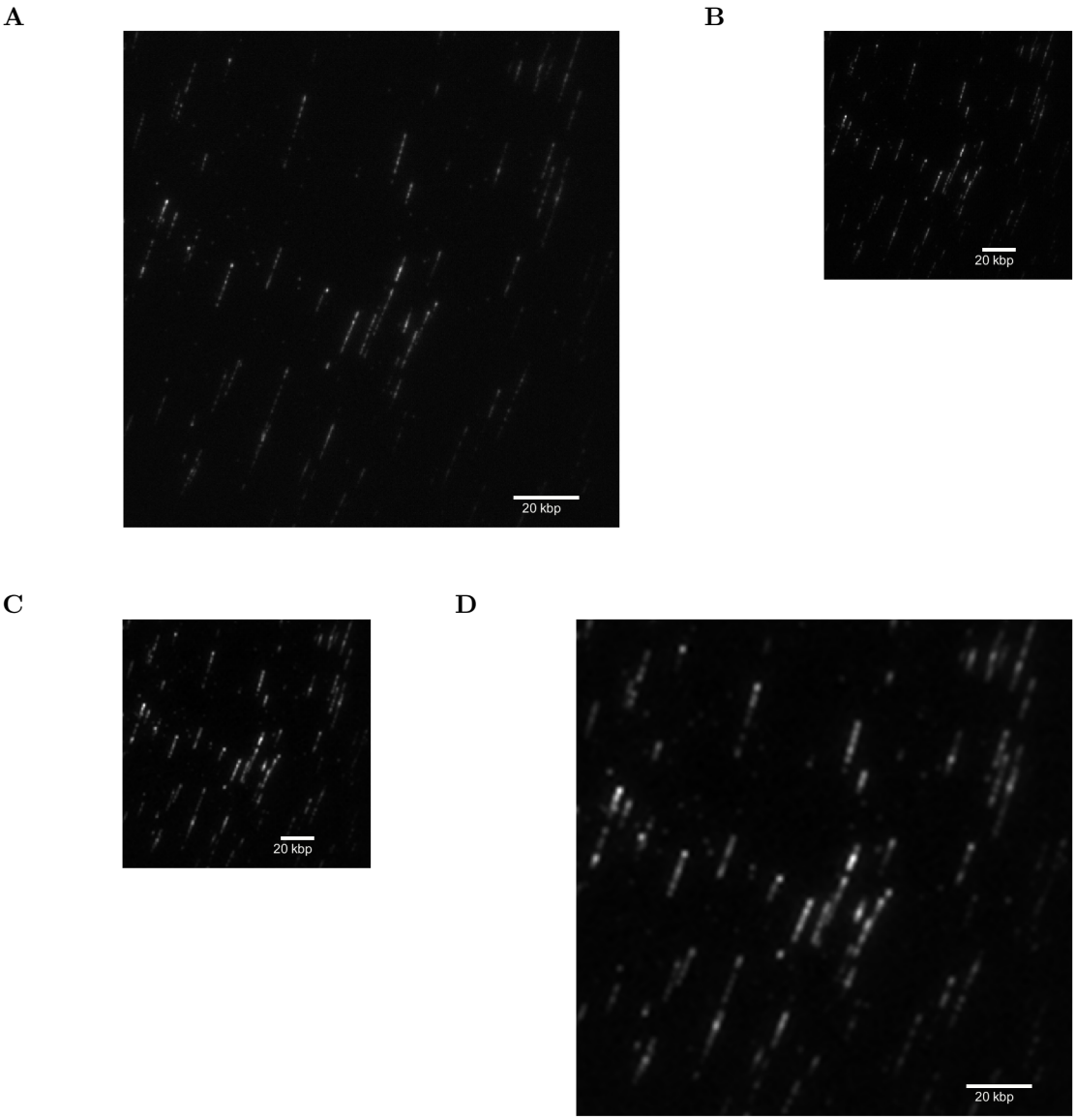


Figure 9: Blurring method. **(A)** Lambda DNA FIPs image (512×512 pixels) followed by **(B)** the size reduction (256×256 pixels) and then **(C)** csaps (256×256 pixels) followed by **(D)** image expansion to produce the image on the bottom (512×512 pixels) where FIPs are smoother.

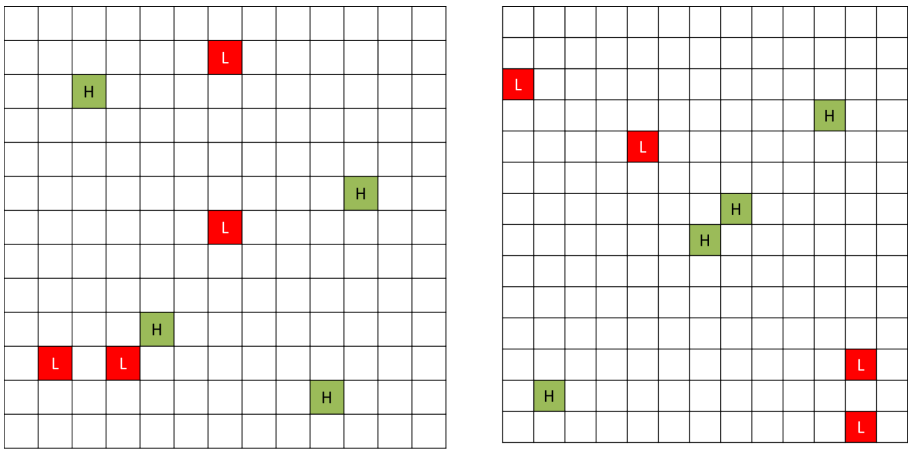


Figure 10: Schematic demonstrating the convolution grid with high intensity points displayed by 'H' and low intensity points identified by 'L'. Left: this point in the image would not classify as a location of a FIP as the angle of the points in the image do not follow the same line. Right: the 'H' points are in a line and the angle of the 'L' points is different by 90° meaning that this is a location of part of a FIP.

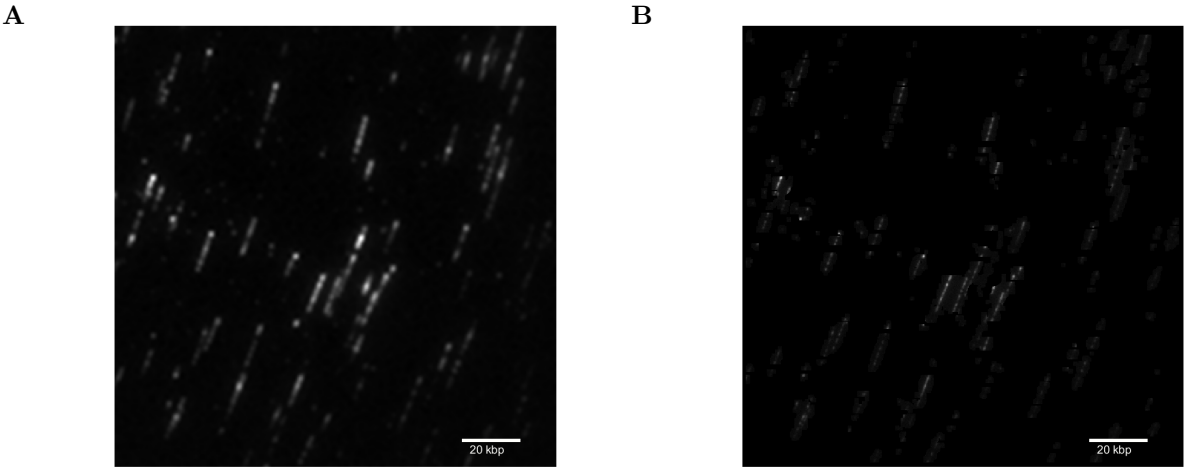


Figure 11: Locating FIP structures. **(A)** Blurred lambda DNA FIPs image followed by **(B)** the image after the convolution picking out the FIPs and leaving the background black.

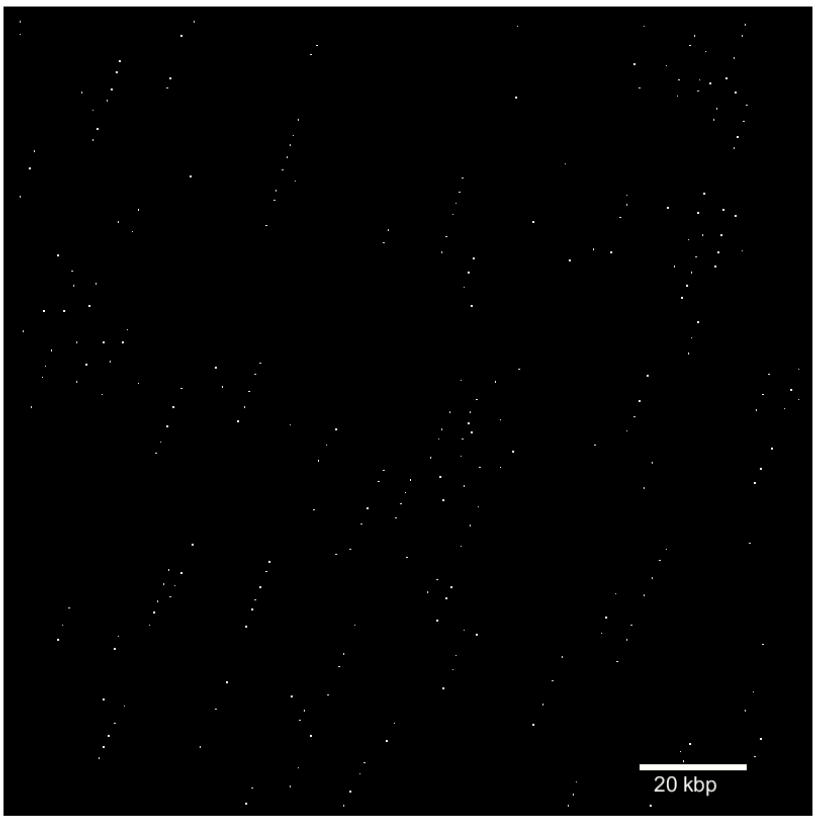


Figure 12: The points of highest intensity in the highlighted FIPs image.

From the highlighted FIPs in the template convolution, the brightest pixel in a smaller template (11×11 pixels) is identified. Figure 12 displays the brightest pixels in the FIPs in the image.

To group the highest intensity points by FIP, at each brightest pixel the algorithm will plot short (15 pixels long) lines centred at these pixels (Figure 13). To merge the lines the image is reduced in size by length scale factor 0.5, followed by applying the csaps function. Finally, the image is increased in size again by a length scale factor of 2 so that the high intensity points are grouped in separate lines in an image.

A binarisation is applied to the final image in Figure 13. For every pixel value v, if $v < 0.5, v = 0$ and if $v > 0.5, v = 1$. A binary image is produced where FIPs are located, demonstrated in Figure 14. 0.5 is the threshold chosen as it filters the groups of blurred FIPs from the background of the image. The background intensity values are close to zero and the blurred FIPs intensity values are close to 1, so the binarization at 0.5 allows identification of these groups at the next step.

From this point the function *bwconncomp* for detecting connected components is used,

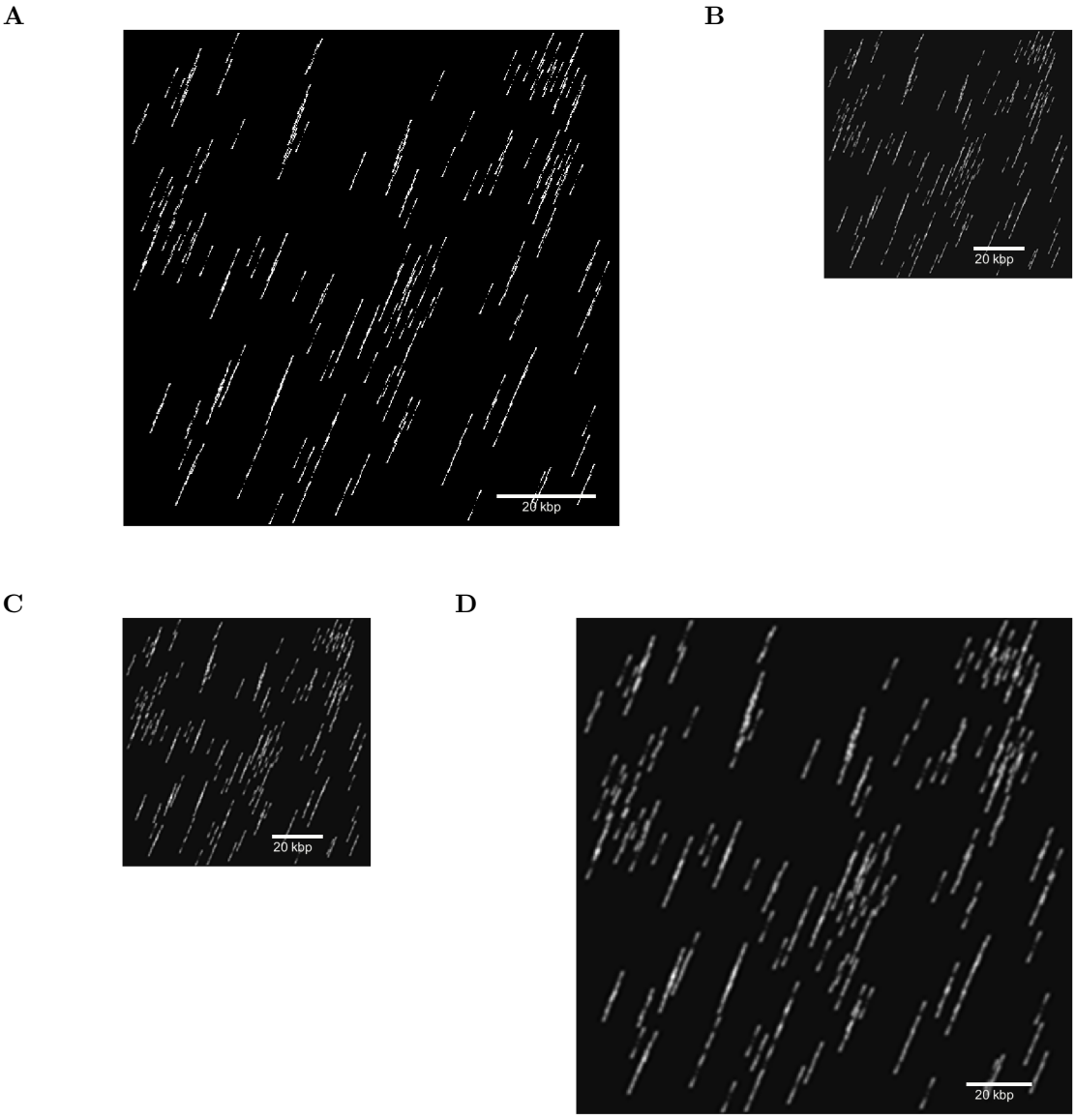


Figure 13: The points of highest intensity being categorised into FIPs. **(A)** Lines drawn at every bright pixel. **(B)** Resizing by length scale factor of 0.5. **(C)** Blurring the lines by applying the csaps function. **(D)** Resizing by length scale factor of 2.

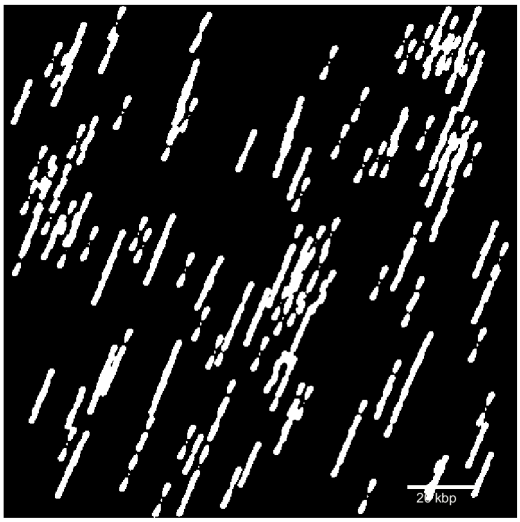


Figure 14: The points of highest intensity being categorised into FIPs. The image of the blurred lines with a threshold of 0.5 intensity applied.

which groups pixels into an array if their edges or corners are touching.

9.8 Extracting classified FIPs

The second step of image analysis is extracting the FIPs from each image. By connecting the points in each categorised FIP a smooth line can be drawn along the FIPs pixels which defines where the strand of DNA is located, traversing along the fluorophores visible in the image. Finally, the intensity values can be extracted from the pixels along the lines drawn in the images. The method is outlined below:

1. Draw a smooth curve for each group of bright points
2. Extract intensity values from original image along where the curves are drawn.

A separate array for each FIP's bright points is taken and each x and y coordinate is used to calculate the distance in between each point. These distances are used in an arc length parameterisation to plot pixels evenly along each FIP, filling in the gaps between the brightest points (schematic in Figure 16). The parameterisation is explained below.

For a line that is not straight, the line can be expressed by a function for that line $\mathbf{r}(t)$, where t is a (not necessarily arc length) parameterisation along this line ($t = a$ to $t = b$). As a vector, this is given in 2D by:

$$\mathbf{r}(t) = (X(t), Y(t)). \tag{11}$$

An arc length $s(t)$ for the curve $\mathbf{r}(t)$ is given below:

$$s(t) = \int_a^t \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt', \tag{12}$$

s can be calculated and expressed as a function of t and then inverted so that t is given in terms of s. The new expression for t in terms of s can be inputted into the equation for $\mathbf{r}(t)$ so to give an arc length parameterisation for $\mathbf{r}(t)$ in terms of s:

$$\mathbf{R}(s) = \mathbf{r}(t(s)) = (x(t(s)), y(t(s))). \tag{13}$$

This demonstrates that a curve can be expressed as a function of arc length. This can be applied to a finite system by numerical integration.

Below is the numerical implementation method for integration, the 2 methods explored below are the midpoint rule (left) and the trapezium rule (right). These methods approximate the area under a graph by a calculating the sum of the area of a finite number boxes at regular intervals along the curve. Schematic in figure 15.

$$\int_a^t f(x)dx \approx \sum_{i=1}^{N}(x_i - x_{i-1}) f\left(\frac{x_i + x_{i-1}}{2}\right), \quad \int_a^t f(x)dx \approx \sum_{i=1}^{N}(x_i - x_{i-1})\frac{f(x_i) + f(x_{i-1})}{2}. \tag{14}$$

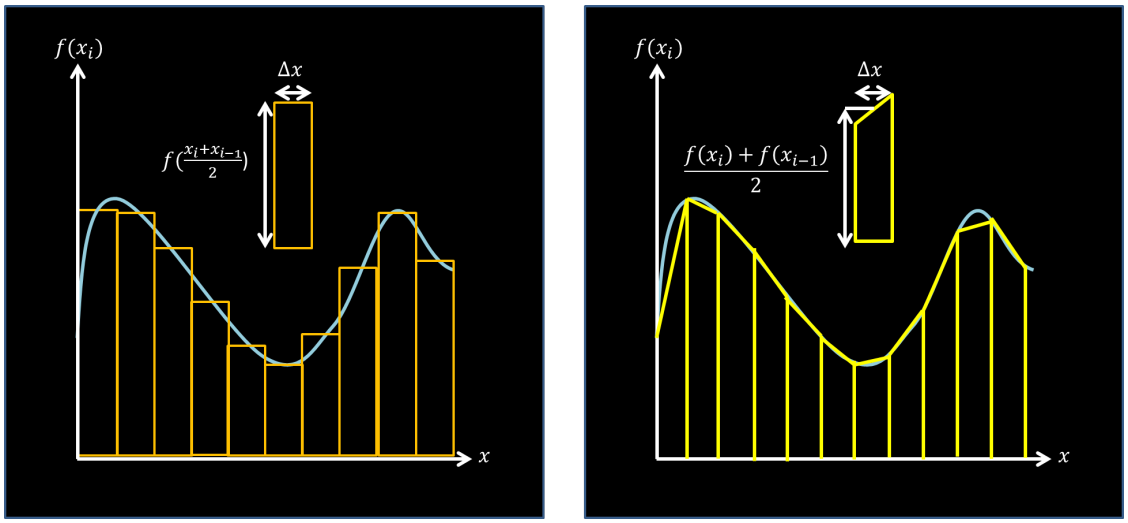


Figure 15: Schematic demonstrating area calculation of a graph. When using integration the number of bars is infinite and Δx tends to 0. Left: Midpoint rule. Right: Trapezium rule.

Numerical integration can be applied to the arc length parameterisation described earlier. By calculating shorter distances along the curve, the length of the curve can be estimated. Firstly, the integral for arc length is expressed as an integral, and as a numerical estimate:

$$s(t) = \int_a^t \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt' \approx \sum_{i=1}^{N} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}. \tag{15}$$

A closer approximation occurs as the distance between coordinates shortens, and the number of points 'N' increases (see Figure 17). If the arc lengths S are calculated then a function of S for the line can be approximated, then by evaluating this function at shorter distances SS along the line, the coordinates along this line can be calculated. This can fill in the pixels in between known pixels along a curved line (Figure 16).

The image mask in MATLAB is plotting lines where the distance between pixels is tending to 1. This will plot a line through adjacent pixels over a FIP, making a solid line. The distance between points x_i, x_{i-1} can be expressed with $|x_{i-1} - x_i| = \delta x_i$, $|y_{i-1} - y_i| = \delta y_i$. An arc length parameterisation for coordinate points in MATLAB to gain an estimate for the points along a line arc $\mathbf{A}$ becomes:

$$s_N = \sum_{i=1}^{N} \sqrt{(\delta x_i)^2 + (\delta y_i)^2}, \quad \mathbf{A}(s) = (x(s), y(s)). \tag{16}$$

For each FIP the highest intensity points along it are given as arrays for the x and y coordinates, (with 'N' number of points).

$$X = \{x_1, x_2 ... x_N\}, \quad Y = \{y_1, y_2 ... y_N\}. \tag{17}$$

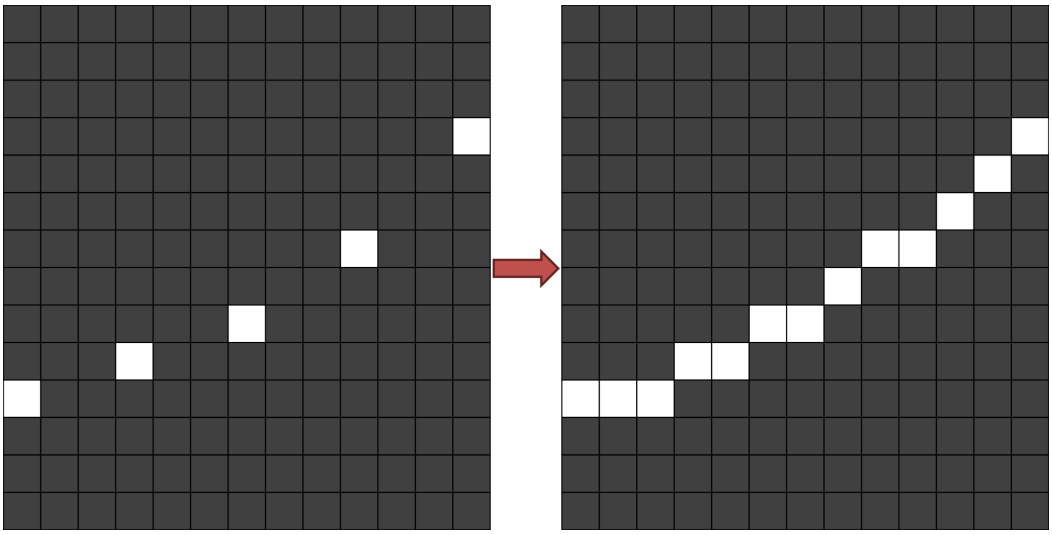


Figure 16: Schematic demonstrating the approximation of the points along a curve, filling in pixel values in between known values using interpolation. Left: brightest pixels in a FIP. Right: curve plotted through pixels.

From the coordinates the lengths between them are calculated and stored as an array S.

$$s_1 = 0, \quad s_n = \sum_{j=2}^{n} \sqrt{(X_j - X_{j-1})^2 + (Y_j - Y_{j-1})^2}, \quad S = \{s_1, s_2, ...s_N\}. \tag{18}$$

From this parameterisation, a Piecewise Cubic Hermite Interpolating Polynomial function (pchip) is used to return a function that separately estimates a polynomial function for x and y values with respect to the lengths S between the points.

$$p_x = \text{pchip}(S, X), \quad p_y = \text{pchip}(S, Y). \tag{19}$$

This function is used because the polynomial is a shape-preserving piecewise cubic interpolation. It is important to preserve the shape so that the line stays true to the shape of the FIP. The pchip function is chosen over the spline function so as to preserve the shape of the curved FIP.

The x and y coordinates are then calculated by evaluating the polynomials p_x, p_y using ppval at shorter distances SS. The line is divided into unit length partitions where the number of points is the length of the curve (s_N) bounded down. In MATLAB this is calculated by $SS = \text{linspace}(s_1, s_N, s_N)$, s_N is the length of the line in pixels. SS will approximate short distances tending to 1 pixel apart, filling in pixels along the line as demonstrated in Figure 16. The construction of the array SS is described below, $\lfloor s_N \rfloor = N$.

$$i = \{1, 2, ..., N\}, \quad ss_i = \frac{(i-1)(s_N - s_1)}{\lfloor s_N \rfloor - 1}, \quad SS = \{ss_1, ss_2, ...ss_N\}. \tag{20}$$

$$X = \text{ppval}(p_x, SS), \quad Y = \text{ppval}(p_y, SS). \tag{21}$$

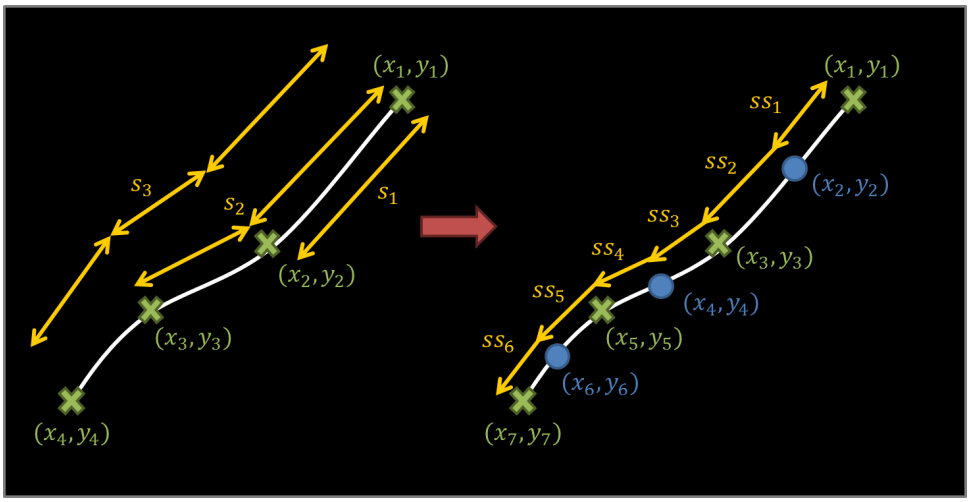


Figure 17: Filling in pixel values from first to last pixel. White lines indicate FIPs. Green 'X's are the brightest intensity points along a FIP. Left: first calculation of 's' distances using $\sqrt{(\delta x)^2 + (\delta y)^2}$. Right: blue circles indicate the interpolated points calculated by using the yellow 'ss' distances.

This gives a smoother line that evenly cuts through every bright point in every FIP.

Finally, the pixel intensity values along each line are extracted and these FIPs are added to a final array containing each FIP side by side. This is the output of the extraction function, the FIPs from Figure 18 are displayed in Figure 19.

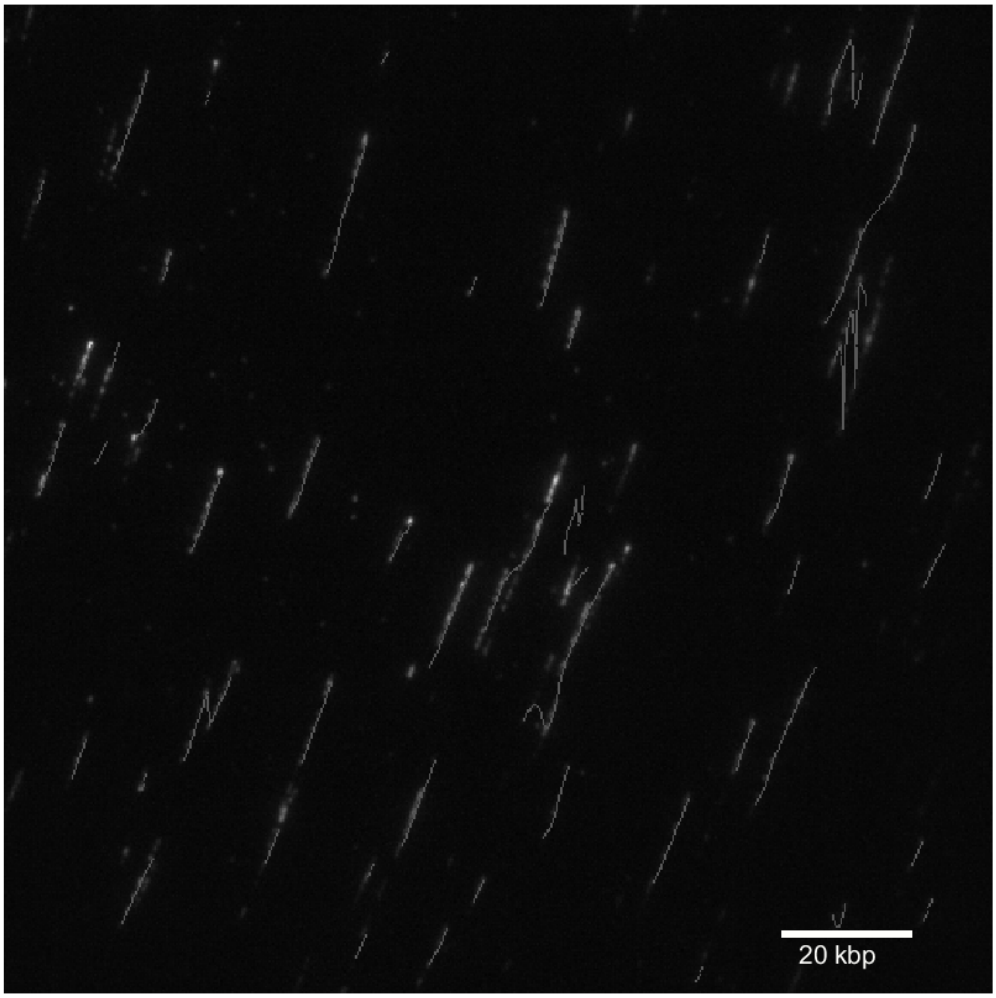


Figure 18: The FIPs location template plotted with an arc length parameterisation through the brightest intensity points.

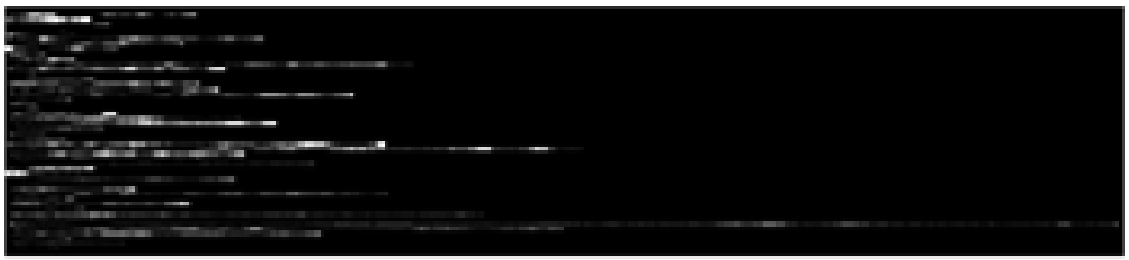


Figure 19: Extracted lambda FIPs from one image using new extraction method.

10 Results and discussion

This section discusses the results of my research gained from implementing the methods from the previous sections. The results will be analysed as the section progresses demonstrating the purpose and outcome of my research.

10.1 Gel electrophoresis

The gel electrophoresis test demonstrates the labelling efficiency of the DNA. If the pattern in the gel for digested modified DNA is different than the digested unmodified DNA then the modified DNA has been labelled sufficiently and the digestion enzyme has been blocked. The digestion enzyme is blocked because of labelled sites. The DNA sample should pass this test in order to continue with the experiment.

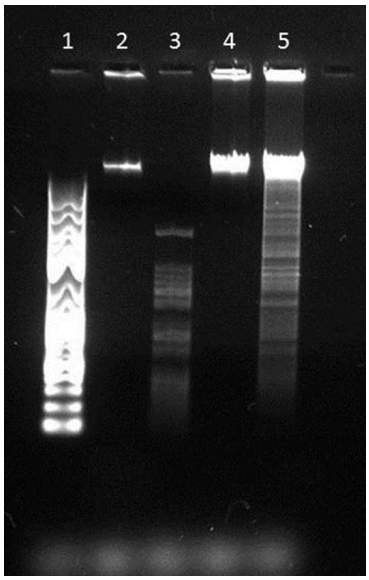


Figure 20: Gel test. Lane 1: ladder, Lane 2: sample1 (unmodified DNA), Lane 3: sample2 (unmodified DNA + R.Taq1), Lane 4: sample3 (labelled DNA), Lane 5: sample4 (labelled DNA + R.Taq1).

The gel electrophoresis is performed on labelled and unlabelled DNA to determine labelling efficiency.

Lane 1 is too bright relative to the sample lanes as too much ladder (Quick-Load Purple 1 kb DNA Ladder NEB # N0552) was added to the lane (Figure 20).

A fraction of labelled DNA had moved through the gel (Lane 5), however the pattern produced was different to the unmodified DNA (Lane 3). This indicates the sizes of the R.TaqI digested DNA molecules were different in the modified and unmodified DNA. The

digestion does not happen at a 5'-TCGA-3' site if the DNA has been labelled there with an azide group. This means that the modified DNA was partially labelled and as most of this DNA did not move through the gel it proves that the labelling was close to 100%.

10.2 Combing, visualisation

Experiments have been carried out in labelling and imaging the lambda bacteriophage genome, data sets for two samples of different concentrations have been produced.

Figure 21 contains images of combed lambda DNA where the DNA's FIPs from the fluorophores are visible.

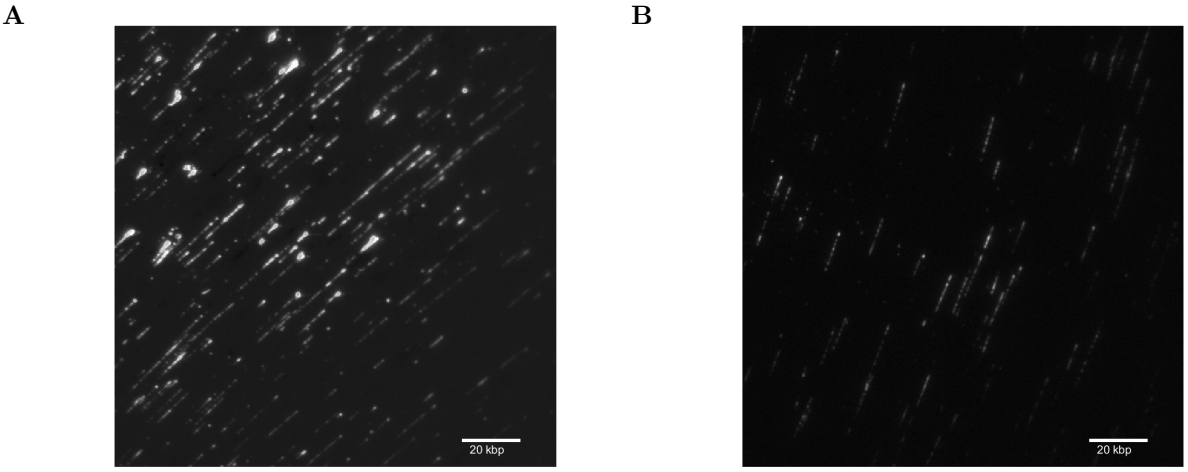


Figure 21: Lambda DNA FIPs indicating the presence of labelled DNA molecules. Scale: height and width = 81.92μm. **(A)** Higher concentration ($\sim$ 200 pg/μl). **(B)** Lower concentration ($\sim$ 100 pg/μl).

Low image quality is an issue as if the images are not clear the FIPs can not be extracted. Examples of low quality images are displayed in Figure 22. Poor image quality can block out areas of an image where FIPs are located or the image can be too blurry to be able to extract a FIP pattern.

When combing the DNA, the images taken are indicative of the quality of the labelling experiment, free dye will be visible in the images as spots dotted randomly (Figure 23). If there is a large amount of free dye then the purification step is not repeated enough for complete purification. Extraction algorithms pick up FIPs along the free dye spots but these FIPs must be discarded as they are not labelled DNA strands.

The concentration of DNA should be within the threshold of 100pg/μl to 200pg/μl. If too low there will be insufficient data collected from each image (Figure 24 **(A)**). If the

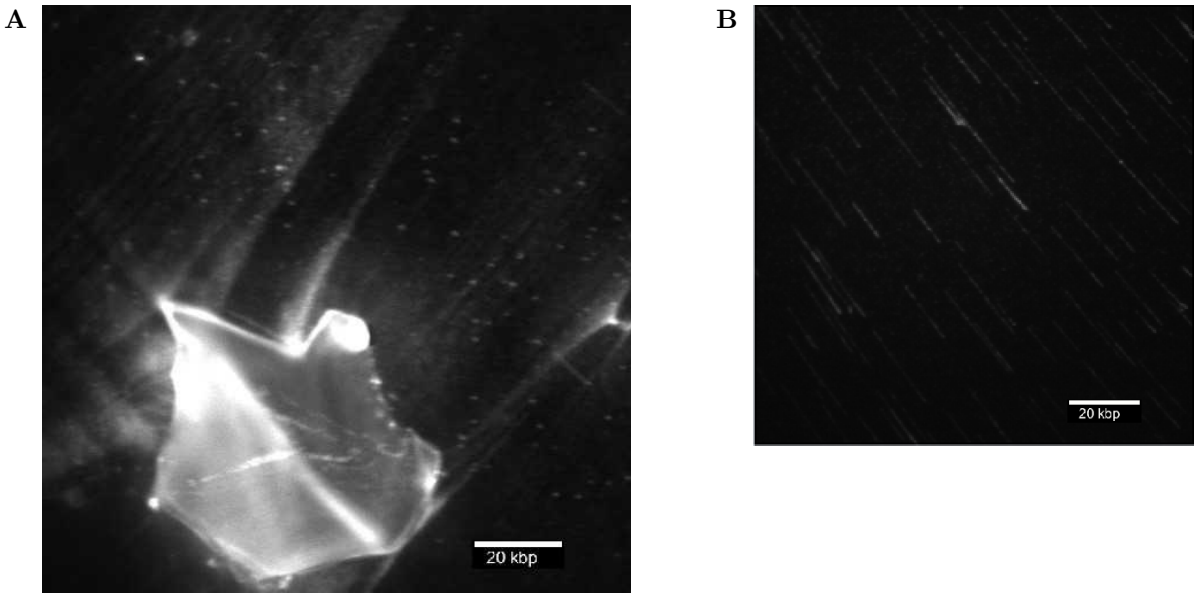


Figure 22: Lambda DNA FIPs indicating the presence of labelled DNA molecules. Scale: height and width = 81.92μm. **(A)** Image with a large blur due to interference. **(B)** Blurry image without clear depiction of FIPs.

concentration is too high then the algorithm will not be able to depict the FIPs as they will be too close together and the fluorophore patterns will interfere with one another (Figure 24 **(B)**).

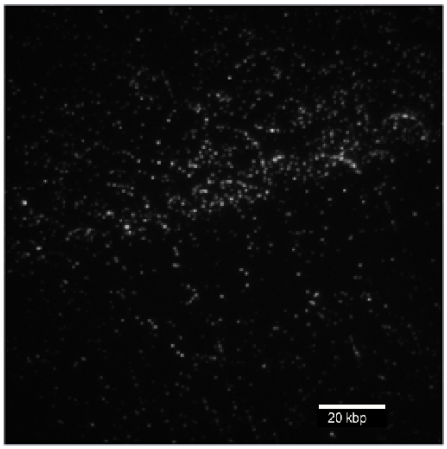


Figure 23: An example of an image displaying no DNA strands (FIPs) but only spots from the free dye deposited on the glass slide.

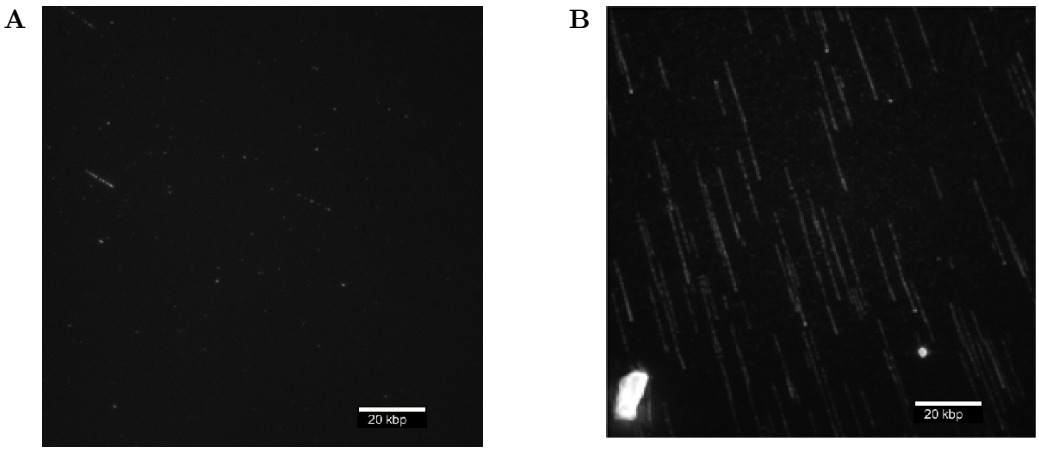


Figure 24: Lambda DNA FIPs indicating the presence of labelled DNA molecules. Scale: height and width = 81.92μm. **(A)** Too low concentration ($<$ 100 pg/μl). **(B)** Too high concentration with merging FIPs ($>$ 200 pg/μl).

10.3 Analysis

Throughout the analysis, a FIP being successfully aligned to a match will be used as a measure of how good an algorithm is. This is done using the current alignment algorithm. First the algorithm uses a filter to exclude extracted data that is not a FIP. Extracted FIPs will be filtered out before alignment based on several parameters (Figure 2). Length measurements remove FIPs which are too small e.g. noise and FIPs that are longer than the reference genome. Mean intensity is considered as FIPs may be too dark meaning the DNA is not fully labelled. A rolling intensity ratio is a moving average calculated across the intensity profile, so if a FIP has an intensity of x and there is a point where the intensity is significantly larger than x then that part of the FIP is unlikely to be that DNA strand's true intensity profile. A rolling intensity ratio will allow the removal of FIPs if they have a spike in the intensity profile which would not match to a genome's FIP. This spike will occur when a FIP is not extracted properly, as a result of the algorithm not extracting a true intensity profile. This can be attributed to dye aggregation, DNA aggregation or contamination of the sample. After testing with simulated data and looking at their normalised cross correlations (which measures displacement and similarity), the threshold for a 'match' is cut off at 0.7. This means that FIPs with a lower similarity (cross correlation) of 0.7 will not be classed as a 'successful' match.

The table of thresholds has two values for each parameter; this is due to having upper and lower bounds in the filter. Each value is also checked against the genome properties in terms of the length in base pairs.

The threshold values are set manually in the code which limits the alignment algorithm to new data sets, for example bacteria. In addition, a genome having a high density of labelled sites may cause FIPs to be discarded as they would not satisfy the rolling ratio, however a unique and high concentration of sites would be more characteristic to a genome. Rejected FIPs may be beneficial to the FIP alignment for one genome, but other genomes will have rejected FIPs which have a higher similarity weighting than those which aren't rejected.

In Figure 25 are the FIPs used in the alignment algorithm. The reference FIP is calculated from the GenBank sequence for lambda phage genome. The FIP is extracted using the new algorithm on the low intensity combed lambda phage DNA (Figure 21). The new extraction algorithm has been performed on multiple data sets, two sets of ten lambda

Table of thresholds.			
Rolling ratio	Minimum length (bp)	Maximum length (bp)	Stretch value
0.05	20,000	36,000	1.1
0.5	32,000	120,000	2.2

Table 2: Table of thresholds for the values used in the alignment algorithm

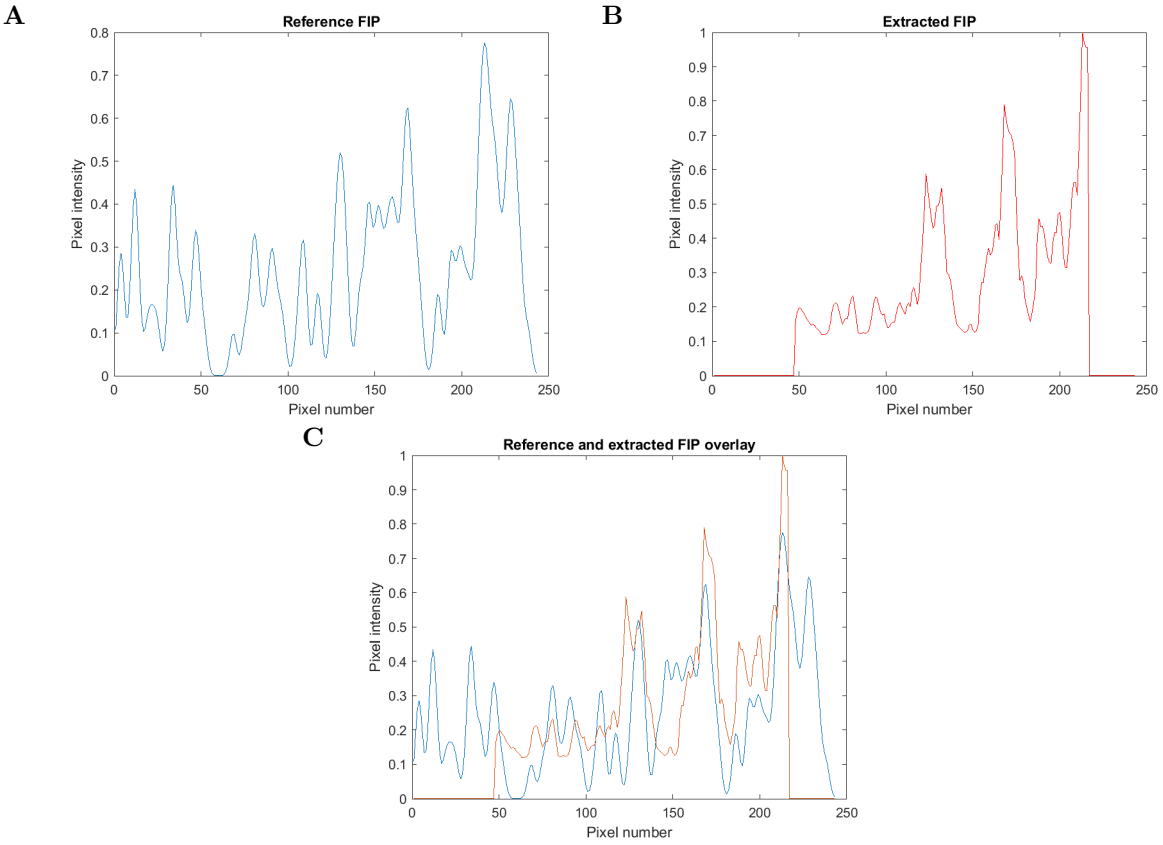


Figure 25: Lambda FIPs. Lambda DNA is 48.5kbps, this example is sampled at 200bps, pixel size is 48,502bps/200 = 243pixels. (A 200bp sample means that each point on the graph/pixel is taken at every 200 base pairs along the genome) **(A)** reference FIP. **(B)** extracted FIP. **(C)** overlay of above FIPs.

DNA fluorescence images and a set of eight T7 bacteriophage DNA fluorescence images.

Overall, new data for the 5'-TCGA-3' fluorescently labelled lambda DNA has been produced and a new extraction algorithm with a higher alignment rate has been developed.

From the 10 images of higher concentration of lambda DNA, 10 images of lower concentration of lambda DNA, and the 8 images of T7 DNA, the number of successfully aligned FIPs extracted using the new algorithm for each is 8, 3 and 10 respectively. From the images of Lambda FIPs, on average, there were fewer than one aligned FIP per image. This is low as there are images being analysed that have no useful data, which is a waste of computing time and demonstrates that the data in the images of lambda FIPs is not a good enough quality.

10.4 Existing and new method comparison

This section will compare the existing extraction method using straight lines to extract FIPs with the new method of extracting FIPs along curved lines. The lambda and T7 DNA are all 5'-TCGA-3' labelled. Figure 26 displays images of the old and new methods demonstrating where the FIPs are extracted by the lines plotted.

For each lambda image stack the algorithms are performed on 10 images. For the data set of the T7 DNA there are only 8 images.

Before alignment, the majority of FIPs (from 74% to 93%) are filtered and discarded from the extracted FIPs. This is because both existing and new algorithms extract FIPs which are not clear enough or are not DNA. This is due to overlapping DNA, strands which are too short, insufficiently labelled DNA or even extracted data that is just free dye that has been deposited on the slide in a visible pattern in image. An example of this error can be seen in Figure 26 **(B)** towards the right where there is a black line which spiked around noise and is not a strand of labelled DNA. Judging a FIP by eye includes looking for a regular pattern of fluorophore spots at least 20kbps in an image. It is assumed that the filter in the alignment treats all extracted data from both algorithms the same despite FIPs being curved or not.

The existing alignment algorithm is used as a measure of the number of FIPs aligned to the reference genome. This method works by comparing the FIPs to the reference genome FIP first by using the MATLAB function *xcorr*. The function *xcorr* measures the correlation coefficient between two FIPs (from 0 to 1). The point of highest correlation identifies the displacement of the extracted FIP to the reference FIP. Following this, the difference in intensities and difference in gradients is measured. The quality of a match is an average of the three comparisons scoring from 0 to 1. FIPs which score higher than 0.7 on the algorithm are considered a match to the reference FIP.

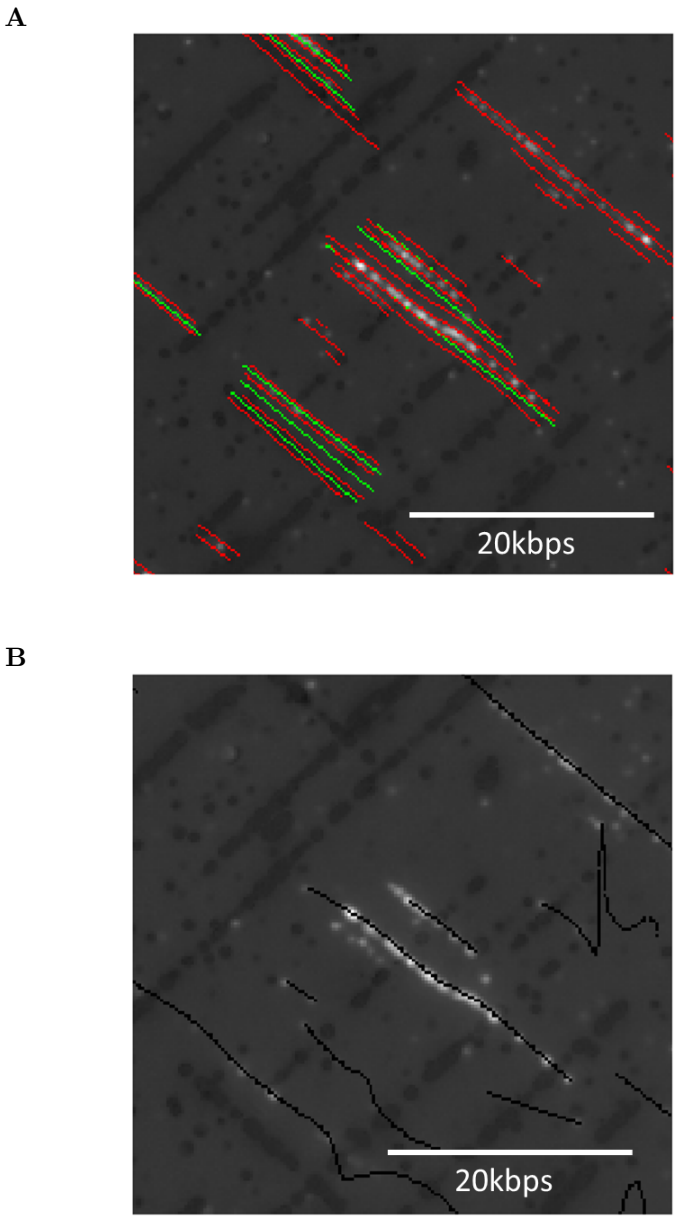


Figure 26: Lambda FIPs with a curved FIP in the centre. **(A)** Existing extraction method on a curved FIP. **(B)** New extraction method on a curved FIP. The black lines are the extracted FIPs, some error is visible here where the algorithm has detected noise as a FIP and not DNA.

10.5 Extracted data

Results from the existing and new extraction methods.

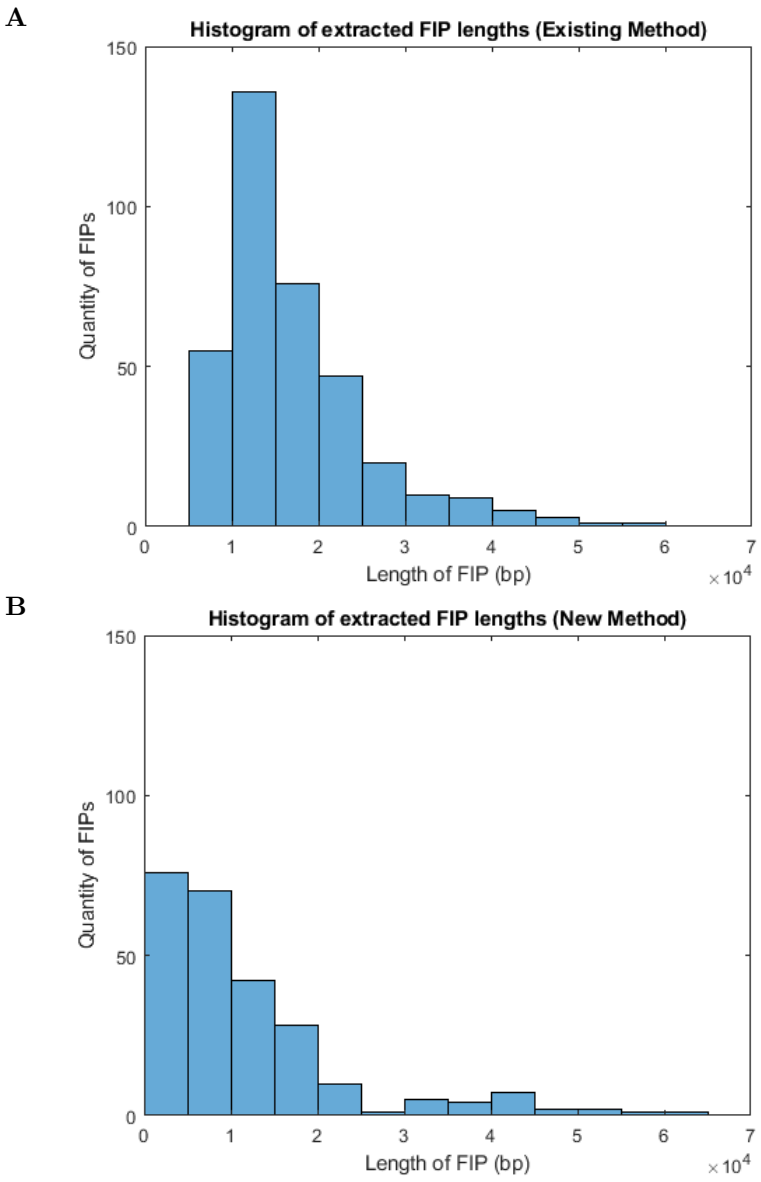


Figure 27: For 10 images of Low concentration Lambda DNA. Histograms displaying the difference between extracted FIP lengths from the existing and new extraction methods. **(A)** Extracted FIPs Existing method. **(B)** Extracted FIPs New method.

Existing method			
	Extracted FIPs	FIPs remaining after filtering	Fips successfully aligned
Lambda higher %	362	78	8
Lambda lower %	363	34	1
T7	154	40	7
Average	293	50	5
New method			
	Extracted FIPs	FIPs remaining after filtering	Fips successfully aligned
Lambda higher %	569	70	8
Lambda lower %	381	27	3
T7	360	46	10
Average	436	47	7

Table 3: Charts displaying the difference between extraction methods.

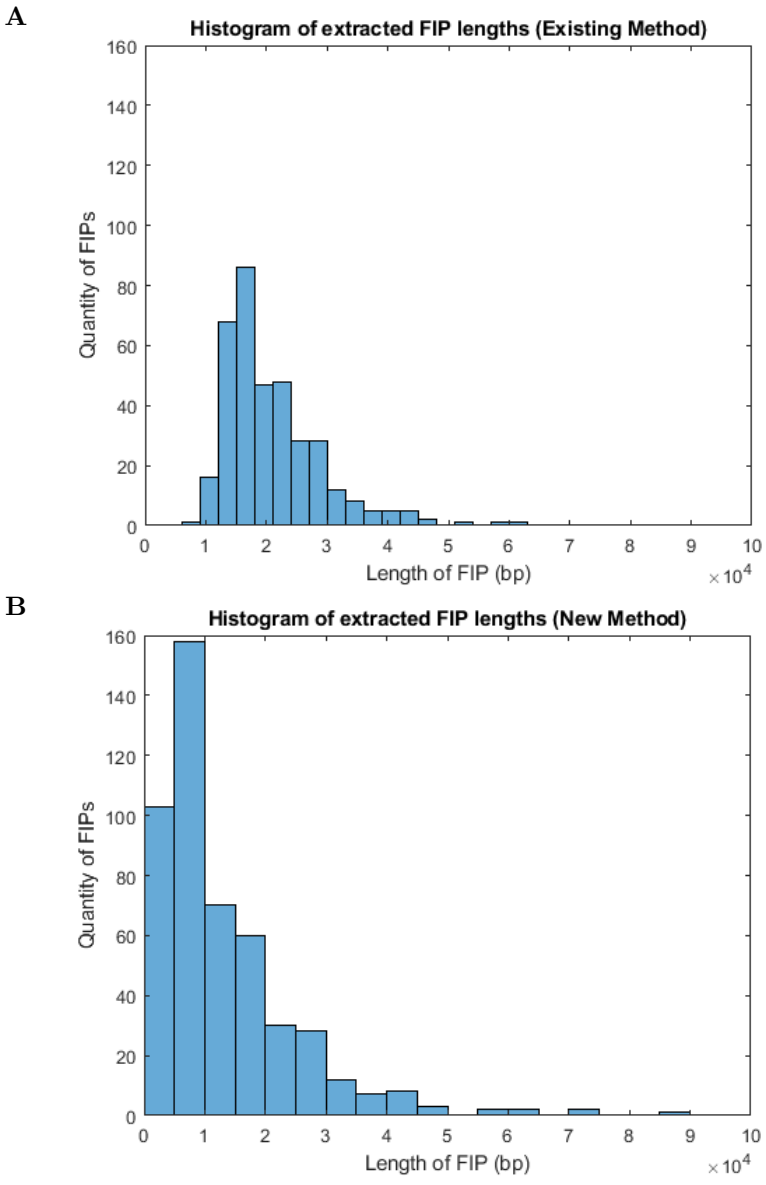
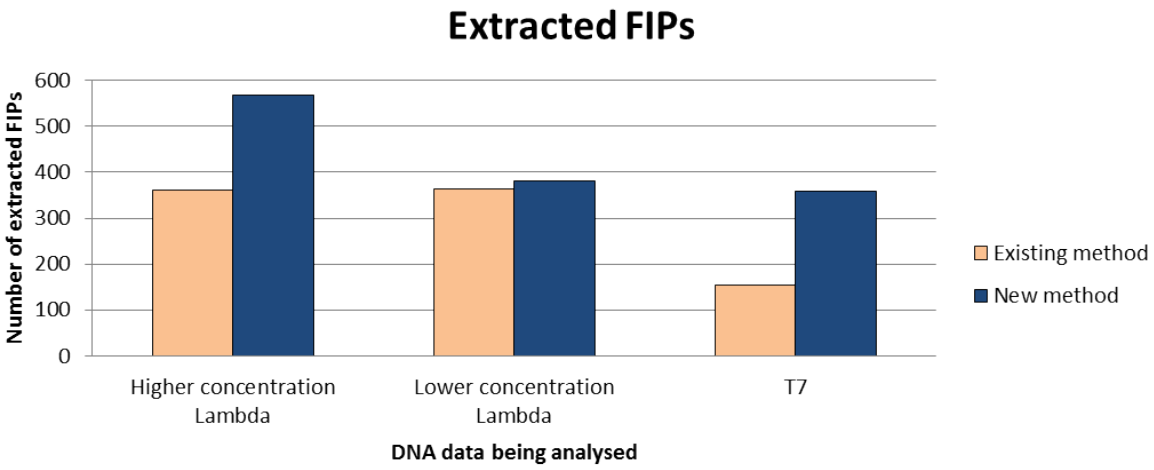


Figure 28: For 10 images of High concentration Lambda DNA. Histograms displaying the difference between extracted FIP lengths from the existing and new extraction methods. **(A)** Extracted FIPs Existing method. **(B)** Extracted FIPs New method.

A



B

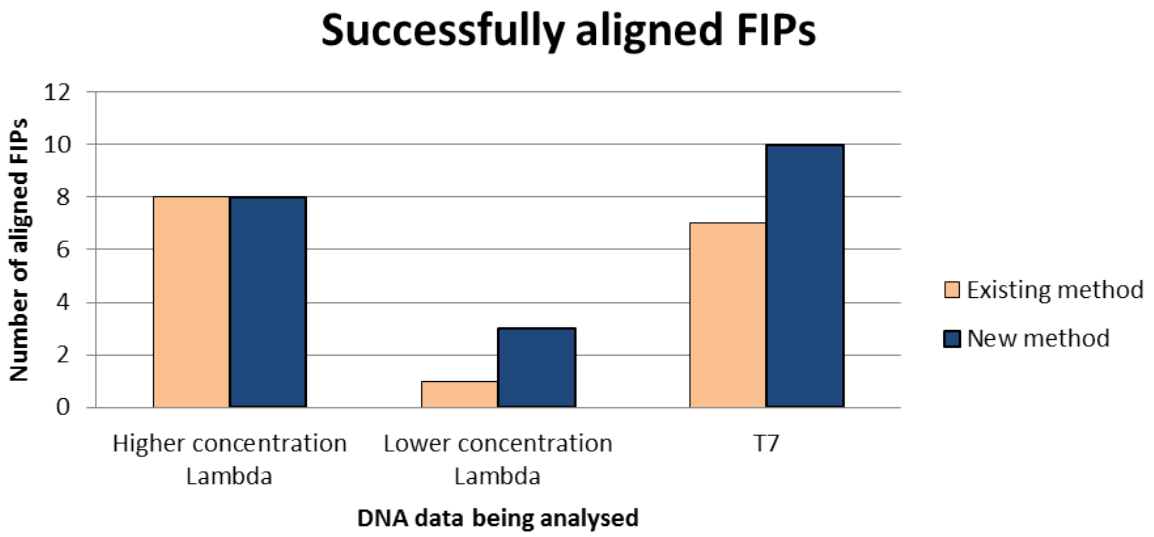


Figure 29: Column charts displaying the difference between data from the existing and new extraction methods. **(A)** Extracted FIPs comparison. **(B)** Aligned FIPs comparison.

10.6 Data analysis

The histograms in Figures 27 and 28 demonstrate the lengths of DNA molecules that are extracted from images of both lower and higher concentrations of combed labelled Lambda DNA. A common pattern with all four of the graphs is that more than half of the extracted FIPS for both methods are smaller than 20kbps. The minimum length for filtering extracted FIPs is at least 20kbps (Figure 2). Therefore the majority of FIPs are discarded because they are too small. This emphasises the fact that during the DNA preparation extra care should be taken to not break the DNA as most of the data in the images taken is discarded. Furthermore, the histograms produced from the new method had a higher proportion of FIPs lower than 1kbp, this indicates that the new method is extracting a lot of FIPs which are much too small to be analysed, and are likely to be incomplete segments of FIPs or just extracted noise from connecting free dye. Finally, there is a peak as expected for the new extraction method around 45kbps which is close to the length of Lambda DNA (Lambda is 48.5kbps long).

To conclude the histogram analysis, the new method extracts more FIPs which are below the useful threshold of 20kbps than the existing method. However the new method is also extracting FIPs of an expected length for these two data sets. This new algorithm can be improved by reducing the number useless FIPs, this can be done by improving the part of the algorithm which identifies where FIPs are located as to reduce FIPs made from noise.

From the results (Figure 29) for the Lambda and T7 DNA data sets, the new algorithm has the same or a higher number of 'successful' matches (a 'successful' match is when the alignment weight is above 0.7).

The results demonstrate a higher number of FIPs extracted and aligned for the new method. The percentage of FIPs aligned from the remaining FIPs after filtering is 10.52% for the existing method and 14.68% for the new method. However, on average, the existing algorithm aligned 1.82% of the extracted FIPs, while the new algorithm aligned 1.60% of extracted FIPs. This demonstrates that the new algorithm is extracting more noise from the images than the existing method; however, the FIPs remaining after filtering are aligning better for the new method, demonstrating that the localisation strategy for the existing method is better at picking up less noise, but the mask extraction step is aligning more FIPs for the new algorithm.

The new method on all three data sets extracted more FIPs from the images but the remaining number of FIPs after filtering is always very similar for both the existing and new methods. On average over 3 data sets, the new method extracted 49% more FIPs compared to the existing method. However, over these 3 data sets the existing method had 6% more FIPs remaining than the new method after the filtering. This illustrates (assuming the filter is not biased to straight FIPs) that the new algorithm is extracting data that is not actually a DNA FIP.

The new algorithm aligned 32% more FIPs than the existing method on average over all data sets.

In respect to the data acquired from the labelling, combing and microscopy, the fact that the T7 data had a much higher percentage of extracted FIPs aligned would indicate that the lambda DNA images are not containing the same quality of FIPs. Therefore, greater care will be taken when pipetting DNA and taking high quality images. In addition, more data sets will be analysed in the future to ensure the new method performs better on all kinds of data sets (a different form of background noise for example).

To address the algorithm's limitations, it is clear from the FIPs template plot image (Figure 18) that the data extracted from an image is not always a DNA FIP, sometimes the high intensity points are not grouped together properly which leads to a skewed plot. This part of the method will be improved by editing the step which connects the bright points into components (Figure 13). In addition, still not every FIP is extracted from the image.

10.7 Data sets with higher curvature

The images from Figures 30 below are examples of the Lambda DNA data set which contains FIPs which have a higher curvature than the ones above. The images are indicative of the combing experiment where the edges of the droplet being combed have dragged the DNA across as its being combed. Figure 31 demonstrates how the curvature is produced from the combing process.

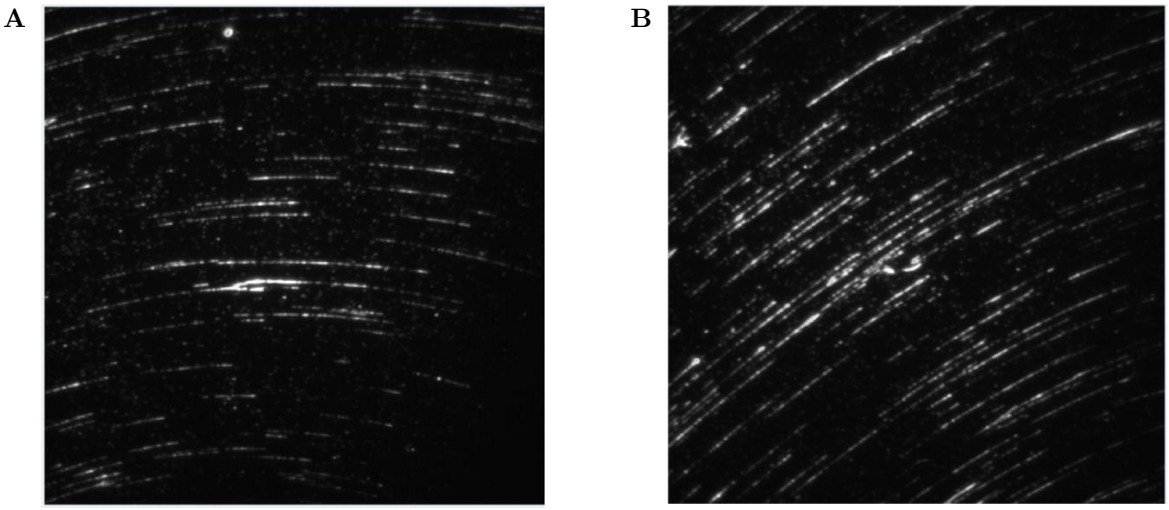


Figure 30: Lambda DNA FIPs indicating the presence of labelled DNA molecules. Scale: height and width = 81.92μm. **(A)** High curvature combed DNA. **(B)** High curvature combed DNA.

Below are the histograms (Figures: 32 and 33 for the lengths of the FIPs extracted from

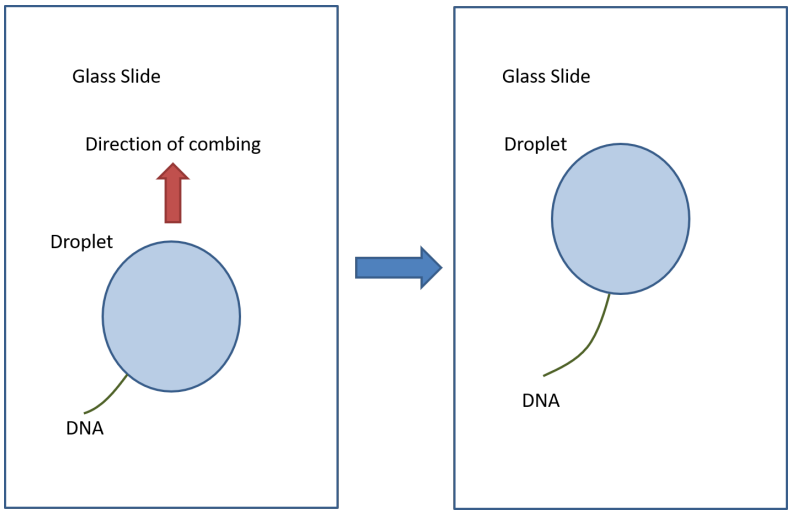


Figure 31: Demonstrative image indicating the curvature of DNA increasing as it is combed across a glass slide.

the Lambda DNA with high curvature for two data sets. It is visible that for these data sets the new method's histogram has a higher proportion of FIP lengths above 20kbps in contrast to the existing method peaking at this point for the data sets of straighter combed DNA FIPs.

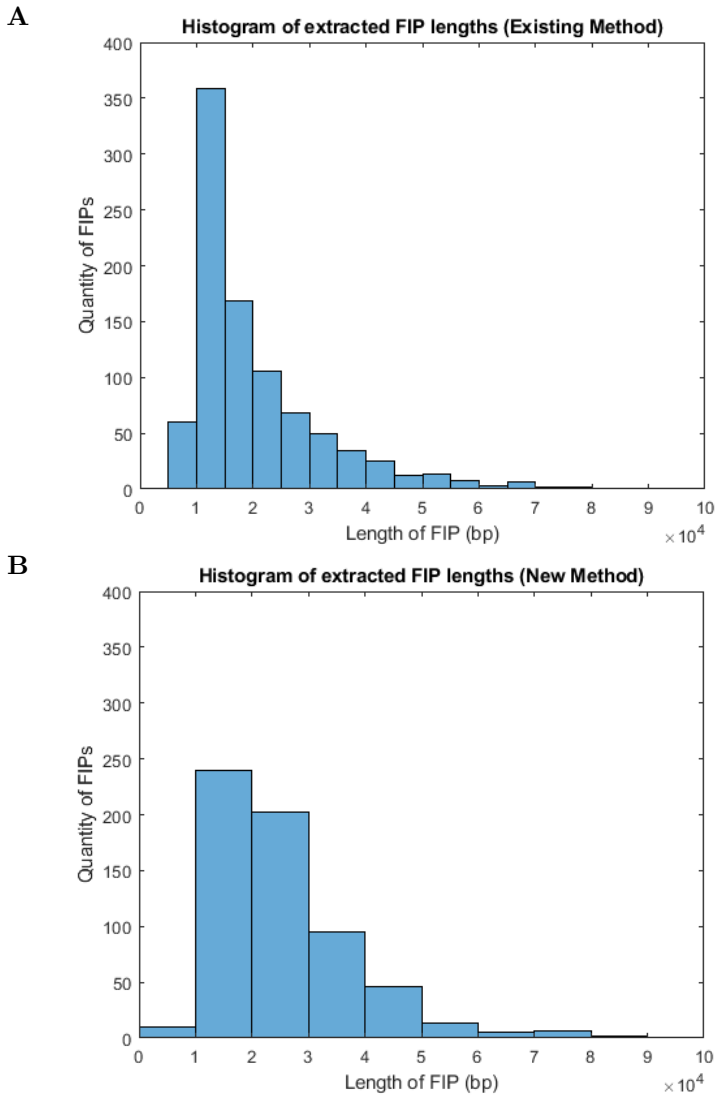


Figure 32: Histograms displaying the difference between extracted FIP lengths from the existing and new extraction methods. For 17 images (images 43-60) of curved Lambda DNA. **(A)** Extracted FIPs Existing method. **(B)** Extracted FIPs New method.

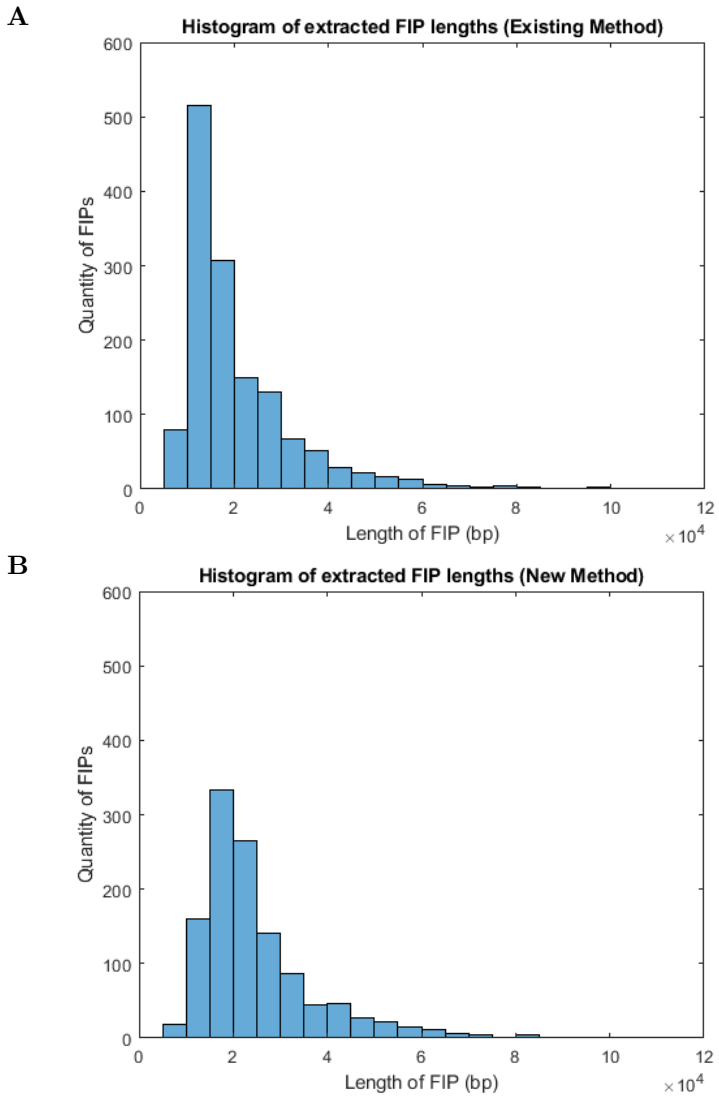


Figure 33: Histograms displaying the difference between extracted FIP lengths from the existing and new extraction methods. For 33 images (images 108-141) of curved Lambda DNA. **(A)** Extracted FIPs Existing method. **(B)** Extracted FIPs New method.

11 Conclusion

We have combed two different concentrations of methyltransferase directed labelled lambda DNA and produced ten fluorescence images from each experiment. I wrote a new automated extraction algorithm that locates and obtains curved FIPs from fluorescence images. On existing and new data we see an improvement by aligning 32% more FIPs than the existing method over three data sets (2×lambda and 1×T7).

Results demonstrate that care should be taken in experiments and imaging for high and accurate throughput. When labelling DNA the concentration should be 100pg/μl to 200pg/μl and the purification step should be repeated three times to remove free dye. When imaging the microscope should be handled so the images are clear and the resolution does not differ between images. This will be essential for obtaining enough data to identify genomes larger than a virus'. For 10× coverage of a 5Mbps bacteria, at the current rate of around 50kbps per image, 1000 images would be needed. Therefore; labelling efficiency, purification, combing and imaging still need optimising for this experiment.

In addition, the experiment for combing DNA will be optimised for longer strands of DNA and the highest concentration of DNA per image for extraction. This would lead to the most data per image being extracted (meaning less images would need to be analysed) and long strands can be constructed over fewer images. Assembling long FIPs is beneficial to the alignment, if there is a region in a genome where the fragments do not align uniquely (discarded in filtering for example) then a FIP that stretches either side of this region may align because of the similarity with other areas in the genome.

The filtering method will also be adjusted so as to filter FIPs based on the characteristics of the reference genome. For example, if the reference has a large peak, do not remove FIPs with a poor rolling ratio as the ones that align to this pattern would be discarded.

Finally, this project aims to be the first to fully map a bacterium genome from visualising labelled DNA, and presented steps towards mapping the human genome via methyltransferase directed labelling.

References

1. A. Hershey, M. Chase, Independent functions of viral protein and nucleic acid in growth of bacteriophage, *The Journal of General Physiology.*, 1952, **36**, 39-56.
2. D. Bogas, L. Nyberg, R. Pacheco, N. F. Azevedo, J. P. Beech, M. Gomila, J. Lalucat, C. M. Manaia, O. C. Nunes, J. O. Tegenfeldt and F. Westerlund, Applications of optical DNA mapping in microbiology, *Biotechniques*, 2017, **62**, 255-267.
3. P. R. Langer-Safer, M. Levine and D. C. Ward, Immunological method for mapping genes on Drosophila polytene chromosomes, *Proceedings of the National Academy of Sciences*, 1982, **79** 4381-4385.
4. L. Piddock, Reflecting on the final report of the O'Neill Review on Antimicrobial Resistance, *The Lancet Infectious Diseases*, 2016, **16**, 767-768.
5. P. Levene, E. London, The structure of thymonucleic acid, *Journal of Biological Chemistry*, 1929, **83**, 793-802.
6. E. Chargaff, Chemical specificity of nucleic acids and mechanism of their enzymatic degradation, *Experientia*, 1950, **6**, 201-209.
7. R. E. Franklin, R. G. Gosling, Molecular Configuration in Sodium Thymonucleate, *Nature*, 1953, **171**, 740-741.
8. J. D. Watson; F.H.C.Crick, Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids, *Nature*, 1953, **171**, 738-740.
9. M. J. Bessman, A. Kornberg, I. R. Lehman, E. S. Simms, Enzymic synthesis of deoxyribonucleic acid, *Biochimica et biophysica acta*, 1956, **21**, 197-198
10. Stahl and B. Y. M. Meselson, The replication of DNA in Escherichia coli, *Proceedings of the national academy of sciences*, 1958, **44**, 671-682.
11. R. Rudner, J. Karkas, E. Chargaff, Separation of B. subtilis DNA into complementary strands. 3. Direct Analysis, *Proceedings of the National Academy of Sciences of the United States of America*, 1968, **60**, 921-922.
12. F Sanger and AR Coulson, A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase, *Journal of Molecular Biology*, 1975, **94**, 441-448.
13. G. Horn, R. Saiki, F. Faloona, K. Mullis, H. Erlich, S. Scharf and N. Arnheim, Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia, *Science*, 2006, **230**, 1350-1354.
14. F. Sanger, S. Nicklen and A. R. Coulson, DNA sequencing with chain-terminating inhibitors., *Proceedings of the National Academy of Sciences of the United States of America*, 1997, **74**, 5463 - 5467
15. M. J. Levene, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra and G. M. Church, Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations, *Science*, 2003, **299**, 682-686.
16. T. A. Clark, E. C. Olivares, K. J. Travers, D. R. Webster, J. H. Lee, S. W. Turner, J. Korlach and B. A. Flusberg, Direct detection of DNA methylation during single-molecule, real-time sequencing, *Nature Methods*, 2010, **7**, 461-465.
17. D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V Pershin, J. M. Ramsey, R. Riehn, G. V Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin and J. A. Schloss, The potential

and challenges of nanopore sequencing, *Nature Biotechnology*, 2008, **26**, 1146-1153.
18. J. J. Kozich, S. L. Westcott, N. T. Baxter, S. K. Highlander and P. D. Schloss, Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform, *Applied and Environmental Microbiology*, 2013, **79**, 5112-5120.
19. J. Lin, R. Qi, C. Aston, J. Jing, T. S. Anantharaman, B. Mishra, O. White, M. J. Daly, K. W. Minton, J. C. Venter, D. C. Schwartz, 'Whole-Genome Shotgun Optical Mapping of Deinococcus radiodurans, *Science*, 1999, **285**, 1558-1562
20. N. O. Wand, D. A Smith, A. A. Wilkinson, A. E. Rushton, S. J. W. Busby, I. B. Styles, R. K. Neely, DNA barcodes for rapid, whole genome, single-molecule analyses, *Nucleic Acids Research*, 2019, **47**, Page e68, https://doi.org/10.1093/nar/gkz212
21. E. Y. Chan, N. M. Goncalves, R. A. Haeusler, A. J. Hatch, J. W. Larson, A. M. Maletta, G. R. Yantz, E. D. Carstea, M. Fuchs, G. G. Wong, S. R. Gullans and R. Gilmanshin, DNA mapping using microfluidic stretching and single-molecule detection of fluorescent site-specific tags, *Genome Research*, 2004, **14**, 1137-1146.
22. R. K. Neely, J. Deen and J. Hofkens, Optical mapping of DNA: Single-molecule-based methods for mapping genomes, *Biopolymers*, 2011, **95**, 298-311.
23. R. C. O'Reilly, D. Wyatte, S. Herd, B. Mingus and D. J. Jilk, Deep Residual Learning for Image Recognition Kaiming, *Frontiers in Psychology*, 2013, **4**, 428-429.
24. R. O. Duda and P. E. Hart, Use of the Hough transform to detect lines, *Communications of the ACM* 1972, **15**, 11-15.
25. D. H. Ballard, Generalizing the Hough Transform To Detect Arbitrary Shapes, *Pattern Recognition*, 1981, **11**, 20111-20123.
26. W. Wang, J. Li, F. Huang and H. Feng, Design and implementation of Log-Gabor filter in fingerprint image enhancement, *Pattern Recognition Letters*, 2008, **29**, 301-308.
27. J. P. Jones and L. A. Palmer, An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex, *Journal Of Neurophysiology*, 1987, **58**, 1233-1258.
28. X. Lu, G. Fan and Y. Wang, A robust barcode reading method based on image analysis of a hierarchical feature classification, *IEEE International Conference on Intelligent Robots and Systems*, 2006, 3358-3362.
29. K. Goda and B. Jalali, Dispersive Fourier transformation for fast continuous single-shot measurements, *Nature Photonics*, 2013, **7**, 102-112.
30. K. Goda, K. K. Tsia and B. Jalali, Amplified dispersive Fourier-transform imaging for ultrafast displacement sensing and barcode reading, *Applied Physics Letters*, 2008, **93**, 131109.
31. B. E. D. Kingsbury, N. Morgan and S. Greenberg, Robust speech recognition using the modulation spectrogram, *Speech Communication*, 1998, **25**, 117-132.
32. S. Greenberg and B. E. D. Kingsbury, in Presented at ICASSP-97, Munich, 2002, vol. **3**, pp. 1647-1650.
33. M. S. Waterman, Identification of Common Molecular Subsequences, *Journal of Molecular Biology*, 1981, **147**, 195-197.
34. W. R. Pearson and D. J. Lipmant, Improved tools for biological sequence comparison (amino acid/nucleic acid/data base searches/local similarity), *Proceedings of the National Academy of Sciences of the United States of America*, 1988, vol. **85**, 2444-2448.
35. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, Basic local

alignment search tool, *Journal of Molecular Biology*, 1990, **215**, 403-410.
36. R. C. Edgar and A. Baterman, Search and clustering orders of magnitude faster than BLAST, *Bioinformatics applications note*, 2010, **26**, 2460-2461.
37. G. Palacios, P. L. Quan, O. J. Jabado, S. Conlan, D. L. Hirschberg, Y. Liu, J. Zhai, N. Renwick, J. Hui, H. Hegyi, A. Grolla, J. E. Strong, J. S. Towner, T. W. Geisbert, P. B. Jahrling, C. Büchen-Osmond, H. Ellerbrok, M. P. Sanchez-Seco, Y. Lussier, P. Formenty, S. T. Nichol, H. Feldmann, T. Briese and W. I. Lipkin, Panmicrobial Oligonucleotide Array for Diagnosis of Infectious Diseases, *Emerging Infectious Diseases*, 2007, **13**, 73-81.
38. K. Bohmann, A. Evans, M. T. P. Gilbert, G. R. Carvalho, S. Creer, M. Knapp, D. W. Yu and M. de Bruyn, Environmental DNA for wildlife biology and biodiversity monitoring, *Trends in Ecology and Evolution*, 2014, **29**, 358-367.
39. New England Biolabs, https://international.neb.com/products/b7204-cutsmart-buffer#Product%20Information, (accessed July 2019).
40. W. S. Rasband, ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, https://imagej.nih.gov/ij/, 1997-2018.