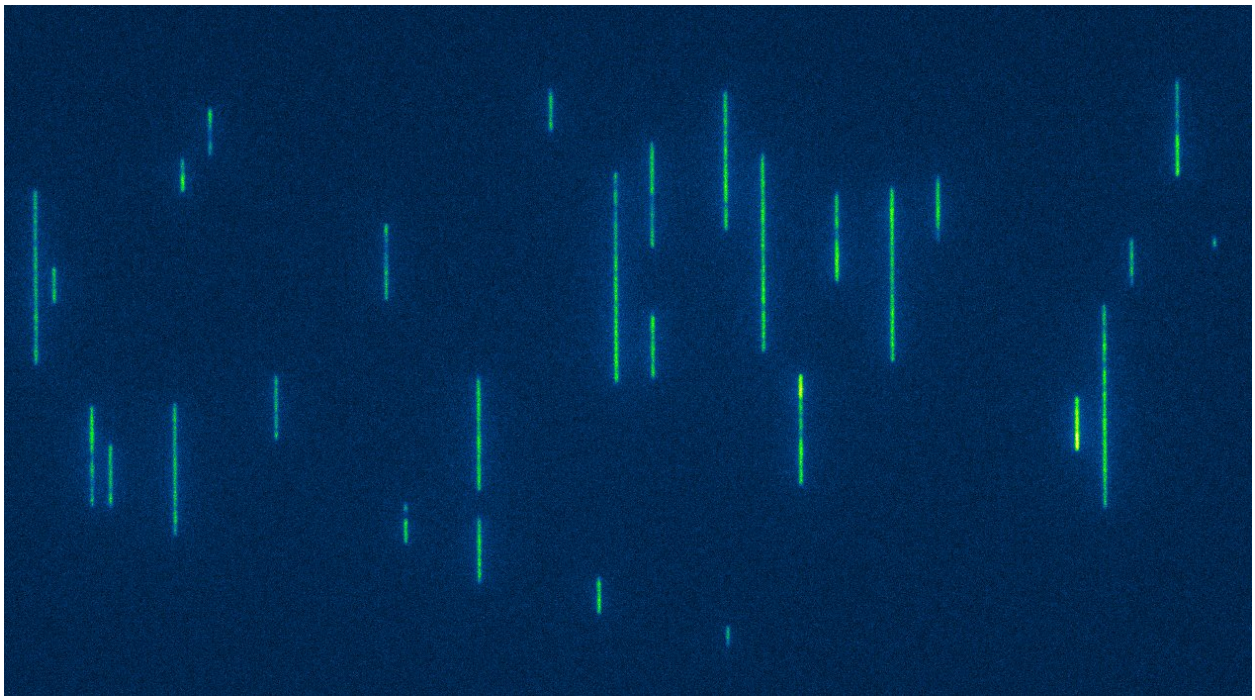
CHALMERS
UNIVERSITY OF TECHNOLOGY



Optical DNA Mapping as a Tool for Determination of EHEC Serotypes

Master's thesis in Biotechnology

HANNA MÅRTENSSON

DEPARTMENT OF LIFE SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2024
www.chalmers.se

Master's thesis 2024

Optical DNA Mapping as a Tool for Determination of EHEC Serotypes

HANNA MÅRTENSSON

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Life Sciences
Division of Chemical Biology
Fredrik Westerlund Research Group
Chalmers University of Technology
Gothenburg, Sweden 2024

Optical DNA Mapping as a Tool for Determination of EHEC Serotypes
HANNA MÅRTENSSON

Supervisor: Radhika Nambannor Kunnath, Division of Chemical Biology, Department of Life Sciences, Chalmers University of Technology
Examiner: Fredrik Westerlund, Division of Chemical Biology, Department of Life Sciences, Chalmers University of Technology

Master's Thesis 2024
Department of Life Sciences
Division of Chemical Biology
Fredrik Westerlund Research Group
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Fluorescence microscopy image of DNA fragments labeled by competitive binding and stretched in nanochannels.

Typeset in LaTeX
Gothenburg, Sweden 2024

Optical DNA Mapping as a Tool for Determination of EHEC Serotypes
HANNA MÅRTENSSON
Department of Life Sciences
Chalmers University of Technology

Abstract

Enterohemorrhagic *Escerichia coli* (EHEC) is a pathogenic subgroup of *E. coli*, responsible for causing severe diarrheal disease though food-borne outbreaks. Infection can further progress to Hemolytic Uremic Syndrome (HUS), a critical condition associated with renal failure and potential fatality. In epidemiological surveillance of EHEC infection, it is of interest to determine the serotype of the EHEC strain, as the risk of severe disease is linked to a few serotypes. Currently, this process involves extensive hands-on involvement while also being time-consuming, suggesting the need to explore alternative methods to identify EHEC serotypes.
In this project, Optical DNA Mapping was employed for bacterial typing. DNA from seven different EHEC samples was labeled through competitive binding and then stretched in nanochannels. Using fluorescence microscopy, the DNA molecules were imaged which due to the labeling resulted in unique fluorescence intensity profiles along the molecules corresponding to the DNA sequence. These intensity profiles were then matched trough reference-based alignment to a reference database in order to identify the bacterial strain associated with the DNA.
The method demonstrated a 100% true positive rate identification at the species level. At the strain level, a majority of matches corresponded to the correct serotype. The results also indicated the potential to assign the matches to strain groups with high taxonomic resolution. In conclusion, the findings provide a positive indication of the possibility in using ODM for determining of EHEC serotypes.

Keywords: DNA, Optical DNA Mapping, Competitive binding, Nanofluidics, Fluorescence, Enterohemorrhagic *Escherichia coli*, Serotype

Acknowledgements

First and foremost, I want to thank my supervisor Radhika Nambannor Kunnath. Your support throughout this project has been invaluable. Thank you for your guidance and all your patient responses to my questions. It has been a pleasure working with you.

I would also like to express my gratitude to my examiner Fredrik Westerlund for letting me do this project in your research group. It has been a great experience and I am thankful for the opportunity to be involved in such interesting research. Additionally, I want to thank the rest of the Westerlund research group for warmly welcoming me and being the most helpful.

I also want to thank my friends and family for your endless support. I am so lucky for having you all around me. And last but not least, thank you David for always being there by my side though my ups and downs over these past five and a half years. You are the best.

Hanna Mårtensson, Gothenburg, January 2024

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

A	Adenine
BME	β-mercaptoethanol
C	Cytosine
CB	Competitive Binding
DBM	DNA Barcode Matchmaker
DNA	Deoxyribonucleic acid
EHEC	Enterohemorrhagic *Escherichia coli*
G	Guanine
HC	Hemorrhagic Colitis
HCA	Human Chromosome Alignment
HUS	Hemolytic Uremic Syndrome
LEE	Locus of enterocyte effacement
ODM	Optical DNA Mapping
SG	Strain group
STEC	Shiga-toxin producing *Escherichia coli*
Stx	Shiga toxin
T	Thymine
TBE	Tris-Borate-EDTA

Contents

List of Figures

List of Tables

1

Introduction

With the recent COVID-19 pandemic in mind, many have been made more aware of the importance of reliable epidemiological surveillance. Identification of pathogens during disease outbreaks is crucial for traceback of contamination and for long-term monitoring of pathogen populations [1]. Today, these diagnostics mostly rely on sequencing-based methods, such as multi-locus sequence typing and whole genome sequencing, which are time- and labor-intensive methods that also require cultured samples.

Recently, it has been demonstrated that Optical DNA Mapping (ODM) can be used for bacterial identification [2],[3]. This single-molecule method is based on fluorescent labeling of DNA in a sequence-specific manner, generating sequence information that can be used to identify bacteria. Previous studies have shown that it is possible to extract DNA directly from uncultured clinical samples and identify bacteria at the strain level. However, for certain subgroups of bacteria, such as Enterohemorrhagic *Escherichia coli* (EHEC), it is of interest to not only identify the strain but also the bacterial serotype as there are certain serotypes that are more frequently linked to severe disease [4].

1.1 Aim

This project aims to investigate the feasibility of optical DNA mapping (ODM) for the accurate detection of Enterohemorrhagic *Escherichia coli* (EHEC) strains and their serotypes. In this study, clinical isolates of seven EHEC serotypes will be investigated. As seven diverse serotypes have been chosen for this study, the results can indicate the utility and clinical relevance of ODM for serotype identification .

1.2 Limitations

In this project, only seven samples of different EHEC serotypes will be investigated. However, it is to be noted that there are over 100 different serotypes of EHEC, many of which could be of clinical interest. Moreover, this study is limited to clinical isolates in which the bacteria have been isolated and cultured in homogeneous samples. Further studies will be needed to establish the technique's performance on heterogeneous, uncultured clinical samples.

2

Theory

The theoretical aspects of this project, including theory on DNA confinement, EHEC bacteria and Optical DNA Mapping, are discussed in this chapter.

2.1 DNA

All living organisms share a common trait; their genetic information is stored within the DNA molecule's chemical code [5]. DNA, short for deoxyribonucleic acid, forms a double-stranded polymer composed of two paired chains. These chains consist of monomers known as nucleotides of which there are four different kinds. The sequence of these four nucleotides determines the genetic information. Similar to a blueprint, DNA encodes all functions and properties necessary for cells to sustain life and to procreate.

2.1.1 Structure

The fundamental building blocks of DNA, nucleotides, are made up of a five-carbon sugar (deoxyribose) attached to a phosphate group and a nitrogen-containing base [6]. As previously noted, there are four different bases in DNA: adenine (A), cytosine (C), guanine (G), and thymine (T). Covalent bonds connect these bases to the sugar and phosphate groups, which make up the pentose-phosphate DNA backbone in a single-stranded DNA molecule (ssDNA). To create the characteristic double-helix structure known as DNA, another single-stranded DNA strand is necessary, resulting in a double-stranded DNA molecule (dsDNA). These two strands are complementary, with bases on one strand matching and binding to specific bases on the other strand. Adenine (A) pairs with thymine (T) through two hydrogen bonds, while cytosine (C) pairs with guanine (G) through three hydrogen bonds. This specific base-pairing, illustrated in Figure 2.1, creates the unique binding between the bases. Consequently, the two complementary ssDNA strands are joined by hydrogen bonding, forming the double-stranded DNA structure.

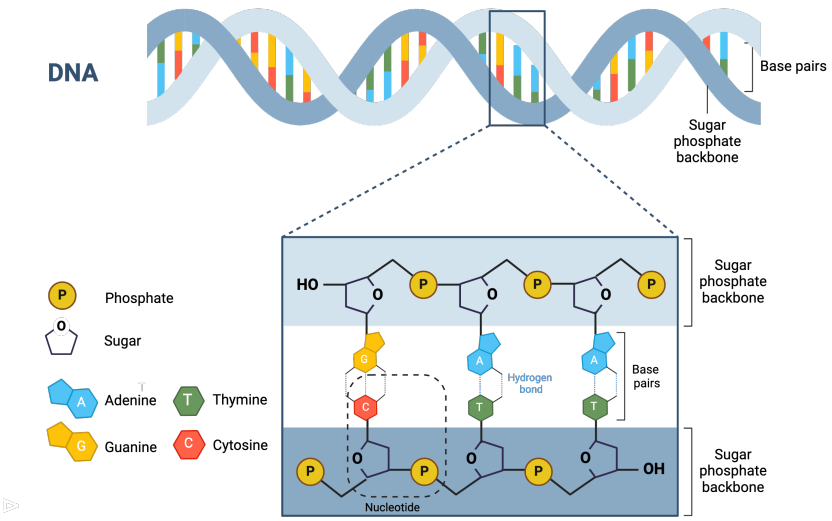


Figure 2.1: *Structure of the DNA-molecule.The double helix with the nitrogeneous bases is illustrated at the top and the sugar-phosphate backbone structure with the base-pairing rules of A pairing to T and C pairing to G is illustrated at the bottom. Created in Biorender.com*

The chemical structure of the two strands, composed of their constituent nucleotides, is responsible for the double helix formation [6]. This structure positions the hydrogen-bonded nucleotides on the inside and the sugar-phosphate backbone on the outside, facilitating the most energetically favorable conformation as a double helix. In neutral pH environment, the bases are uncharged, while the the backbone carries a negative charge causing the dsDNA helix to orient the bases inward while exposing the polar backbone to the solvent resulting in the helical shape.

2.1.2 DNA in confinement

The physical dynamics of free DNA in solution and within confinement can be explained through polymer theory, characterized by three key parameters: the contour length L_C, persistence length L_P and radius of gyration R_G [7]. The contour length represents the length of the DNA unaffected by external forces [8]. It is therefore calculated by multiplying the number of base pairs in the molecule by the height of each pair, typically 0.34 nm for normal B-state dsDNA. However, when visualized with fluorescent labels, such as the intercalating dye YOYO-1 used in this study, each bound molecule adds 0.51 nm to the length of the DNA [7]. The persistence length characterizes the rigidity of the DNA polymer, which is its ability to bend under thermal energy [7],[8]. The persistence length of DNA in high ionic strength buffers is approximately 50 nm, meaning that over a length of 50 nm the DNA cannot be bent [7]. Free in solution, DNA minimizes its free energy by coiling, with the coil size represented by the radius of gyration (R_G), describing the radius of the coiled DNA based on the root mean square of the distance between parts of the coil and its centre.

On the other hand, when confined, such as in the nanochannels used here, the physical behaviour of DNA differs from its free solution state. The average cross section D of a nanochannel of width D_W and height D_H influences the stretching of DNA within the channel and it can be divided into two behavioral regimes [9]. If D exceeds R_G, typical of DNA in free solution, the DNA can fold up on itself and form coils. However, when weakly confined (D smaller than R_G but larger than L_P), the DNA can form local coils while also exhibiting some extension. This is a DNA molecule organized as non-interacting blobs close to D in diameter, as described by the de Gennes regime. In scenarios where D is significantly smaller than L_P, DNA coiling is prevented due to limited space and high energetic costs [7]. This behavior is explained by the Odijk regime. The different regimes described and how DNA is affected as D get smaller is illustrated in Figure 2.2.

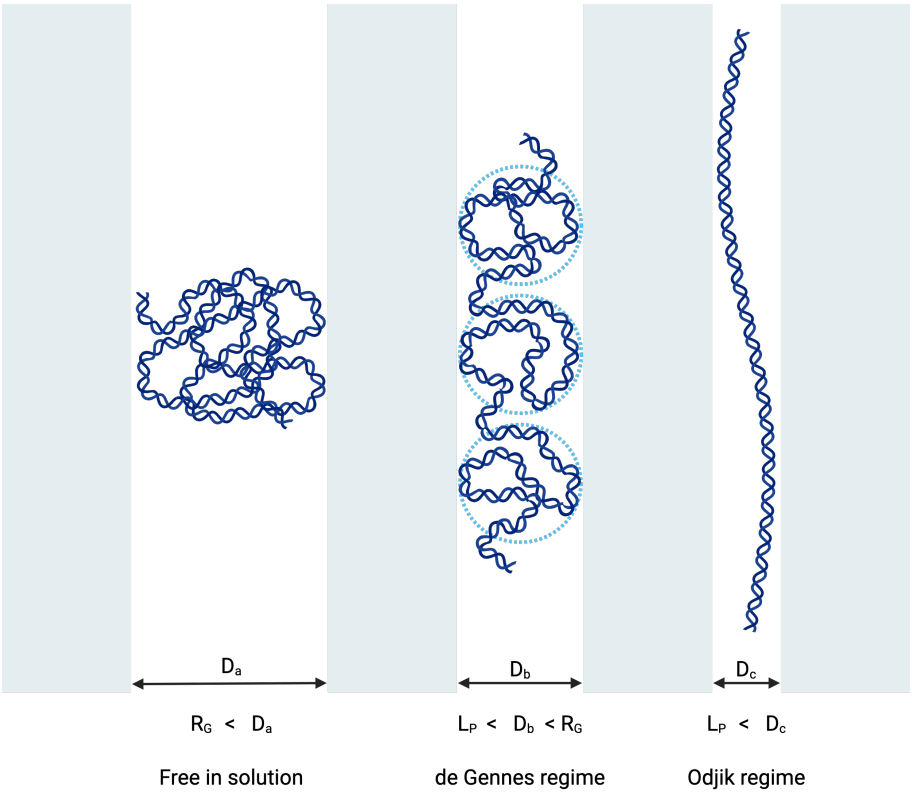


Figure 2.2: *Illustration of how DNA behaves under decreasing level of confinement where D_a>D_b>D_c. Created in Biorender.com*

Using nanochannels where D is near L_P, as done in this study, results in a regime transition between Odijk and de Gennes, referred to as the extended de Gennes regime [8]. Although using channels aligned with the Odijk regime ensures stretching of DNA molecules, it complicates channel fabrication and conducting experiments, due to the small sizes [9]. Opting for nanochannels where D is similar to L_P bridges Odijk and de Gennes behaviors, resulting in a balance between experimental feasibility and DNA stretching.

The behavior of DNA within confinement is also influenced by the ionic strength of its solvent. DNA is a charged molecule and its extension is affected by electrostatic repulsion among its segments and with the channel wall [9]. Lower ionic strength, indicating a reduced concentration of ions in the solution, tends to increase the extension of DNA. This increase in extension can prove advantageous when using stretching in nanochannels for ODM, which will be discussed further on. However, it should be noted that reducing ion concentrations also impacts the performance of several fluorescent dyes commonly used in experiments. This alteration can potentially affect the reliability and accuracy of experimental outcomes.

2.1.3 Bacteria and their genome

Bacteria, single-celled organisms lacking membrane-enclosed organelles, fall under the to the prokaryotic family [5]. They are found in various diverse ecological niches, both extreme environments as well as the human body, exhibiting a wide range of capabilities. Typically measuring only a few micrometers, they exhibit various shapes, from spherical to rod-shaped. The structural components of a bacterium is illustrated in Figure 2.3. The cell's interior is enclosed by a cell wall with the purpose to maintain cell shape and offer mechanical protection [10]. Externally, pili and flagellum are attached to the cell wall. Pili, thin, hair-like protrusions from the cell surface, play a crucial role in adhesion to host cells while the flagellum, a protein filament, allows for cell movement. Within the cell is the cytoplasm housing organelles primarily constituted of ribosomes responsible for synthesizing cellular proteins. Unlike their eukaryotic counterparts, the storage of bacterial DNA is much simpler. Lacking a enveloping nucleus, bacterial DNA exist as a long circular dsDNA in a nucleoid where it is compacted with proteins and RNA, or as plasmids, smaller circular DNA molecules [5]. The nucleoid facilitates contact between the bacterial chromosome, the cell's cytoplasm, and its ribosomes.

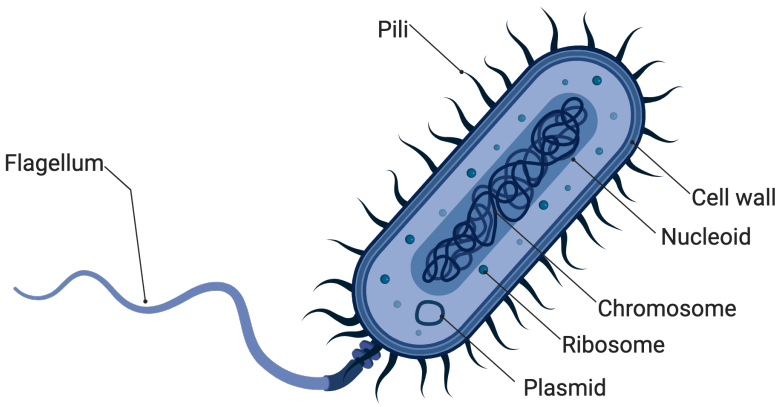


Figure 2.3: *Overview of the structure of a bacterium. In the figure flagellum, pili, cell wall, nucleoid, bacterial chromosome, ribosomes and plasmid are illustrated and marked. Created in Biorender.com*

Bacteria can exhibit pathogenic traits, often distinguished by minimal genetic differences from their non-pathogenic equivalents [11]. Diseases arise from specific genes known as virulence genes encoding for virulence factors. Commonly, these virulence genes are found in clusters on the bacterial chromosome called pathogenicity islands, or on extrachromosomal virulence plasmids.

2.2 Enterohemorrhagic *Escherichia coli*

E. coli are gram negative bacteria that typically reside in the gastrointestinal tract of humans, coexisting with the host without causing harm [12]. However, certain strains of *E. coli* have acquired virulent traits capable of causing a broad range of disease such as diarrheal disease, urinary tract infections and sepsis. These pathogenic variants are grouped based on their virulence factors into categories know as pathotypes. Among them enteropathogenic *E. coli* (EPEC) and uropathogenic *E. coli* (UPEC) can be found. However, the pathotype of interest for this thesis is enterohemorrhagic *E. coli* (EHEC).

EHEC is a subgroup of shiga toxin-producing *E. coli* (STEC), producing one or two different types of shiga toxin (Stx), known as Stx1 and Stx2, respectively [13]. Their genome size typically ranges from 5.5 to 5.9 Mbp depending on the serotype [14]. Infection by EHEC causes bloody diarrhea, or hemorrhagic colitis (HC), by inducing thrombogenic and inflammatory responses in the endothelial cells of the intestine. In 5-15 % of the cases, this progresses further to potentially fatal hemolytic uremic syndrome (HUS) with infection by EHEC being the primary cause of HUS in children[15]. EHEC bacteria are primarily transmitted through food and infections have often been linked to consumption of undercooked meat.

2.2.1 Virulence factors

The shiga toxins produced by EHEC are the two types Stx1 and Stx2 along with various subtypes, where their disease-causing significance varies among the subtypes [13]. Stx2 has been shown to be more frequently associated with severe disease and development of HUS, especially its subtypes Stx2a, Stx2c and Stx2d, while others are related to mild or no disease in humans [16]. Comparatively, Stx1 is less potent than Stx2 and is less likely to induce HUS to the same extent. Stxs causes cell damage by inhibiting protein synthesis through the deactivation of ribosomes, preventing their association with elongation factor 1, such that tRNA cannot bind to rRNA [13]. This induces cellular stress responses that are both pro-inflammatory and pro-apoptotic.

In addition to Stx-producing genes, most EHEC also have a pathogenetic island on their chromosome called *locus of enterocyte effacement* (LEE) that contains the *eae* gene coding for the adhesin intimin [17]. Intimin facilitates the attachment of EHEC to enterocytes and is characteristic for EHEC-associated HC and HUS. Furthermore, if an EHEC strain carries both *Stx2* and the *eae* gene, it is associated with a higher risk of causing severe disease [18].

2.2.2 Serotypes

E. coli strains are divided into distinct serotypes based on two surface antigens found on the cell's outer membrane: the O-antigen located in the lipopolysaccharides and the H-antigen found on the flagellum [4]. These serotypes are named based on the combination of the O-antigen and H-antigen, as seen in names such as O157:H7. With approximately 186 known O-groups (serogroups) and as 53 H-types identified for *E. coli*, the diversity results in a vast amount of serotypes.

Historically, the O157:H7 serotype was the most commonly associated with HUS until the mid-2000s, with other serotypes being infrequently isolated [15]. However, since the early 2010s, a change in epidemiology has been observed where non-O157:H7 EHEC have emerged and been linked to HUS outbreaks, notably in Europe. Among these, six non-O157 serogroups, collectively named the 'big six': O26, O45, O103, O111, O121, and O145, are dominant in human infections and have become notably relevant in clinical contexts [18].

2.2.3 Clinical identification of EHEC serotypes

To verify the presence of EHEC in a patient, fecal samples typically undergo enrichment through cultivation followed by polymerase chain reaction (PCR) analysis to detect the presence of *Stx* and *eae* genes [15]. However, this method does not discern between different EHEC serotypes which is of interest for epidemiological surveillance. While selective media containing sorbitol can identify non-sorbitol-fermenting O157:H7, there are no available selective and differential media for non-O157 strains. Therefore, EHEC serotyping requires molecular methods. Currently, serotyping techniques involve pulsed-field electrophoresis, multi-locus sequence typing, and whole-genome sequencing to distinguish between EHEC serotypes. Alternatively, methods such as PCR for verifying presence of O-antigen gene clusters and using agglutination reactions targeting O- and H-antigens provides another means of identifying serotypes [19]. These methods rely on using specific reagents for each serotype, such as primers and antisera, which can make serotyping time- and labor intensive.

2.3 Optical DNA mapping

ODM is a term that covers various methods used for single molecule visualisation of DNA molecules with the objective to acquire sequence-related information [7]. This technique allows for the analysis of complete DNA molecules, reaching sizes up to millions of base pairs (Mbp), and provides a resolution in the range of thousands of base pairs (kbp). This places the method in the gap between analysis methods on chromosomal level, such as fluorescence *in-situ* hybridization, and base pair level achieved by DNA sequencing. Several strategies exist for conducting ODM, but the fundamental principle involves sequence-specific labeling followed by stretching of the DNA molecule and imaging.

2.3.1 Stretching methods

Various methods are available for stretching of DNA, such as stretching on activated glass surfaces and stretching within nanochannels [9]. In this thesis, ODM is performed on DNA that has been stretched within nanochannels. However, the stretching approach on glass was also used to visualize DNA samples without mapping, thus both approaches are described in greater detail below.

2.3.1.1 Stretching on glass surfaces

To achieve stretching of DNA on glass surfaces, two aspects need to be fulfilled. Firstly, the surface needs to be able to attach to the molecule, i.e. being activated, and secondly a force is necessary to extend the DNA [20]. An activated glass surface refers to a surface that is either positively charged enabling interaction with the DNA's negatively charged backbone, or hydrophobic, allowing tethering of unraveled DNA ends with an exposed hydrophobic core [21]. Hydrophobic surfaces are typically created using silanizing compounds. When attached to positively charged surfaces, additional force from a flowing solution is required to extend the DNA [20]. On the contrary, stretching on a hydrophobic surface relies on shear forces caused by the liquid-air interface between the hydrophobic surface and the hydrophilic solution in which the DNA is deposited.

2.3.1.2 Stretching in nanochannels

An alternative method for stretching DNA involves using nanochannels, eliminating the need to attach the molecule to a surface. Stretching in nanochannels causes the DNA to linearize due to the channel's confinement as described in Section 2.1.2. Studies have demonstrated that nanochannels of dimensions of around 100 nm can be used to uniformly stretch DNA molecules [22]. Consequently, confinement within nanochannels results in the extension of DNA molecules scaling linearly with the contour length and the passive stretching results in molecules being in an equilibrium conformation. This offers reproducibility that is challenging to achieve when stretching on glass surfaces [23].

The chips containing the nanochannels are called nanofluidic chips and are typically fabricated in fused silica [7]. Fused silica is advantageous since in contact with the aqueous buffer in which the DNA is deposited, the silica surface becomes negatively charged thus repelling the negatively charged DNA backbone and preventing DNA molecules from sticking to the narrow channels. There are several possible designs of nanofluidic chips, but the design used in this thesis is comprised of four loading wells connected to two microchannels that linked by nanochannels as seen in Figure 2.4.

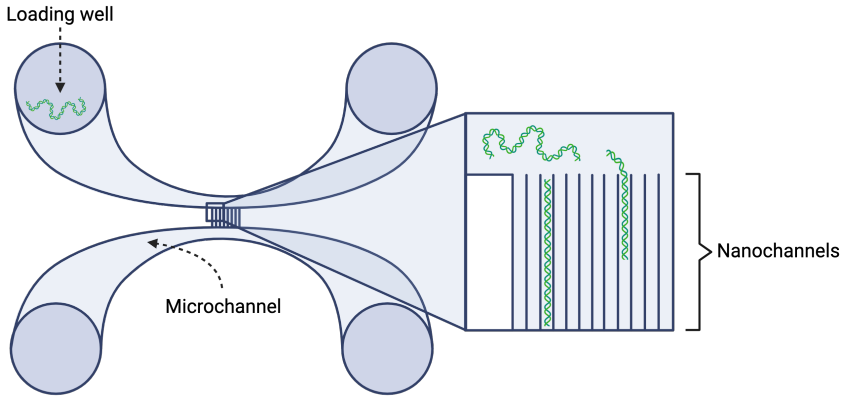


Figure 2.4: *Schematic illustration of the nanofluidic chip. The four loading wells are connected to two microchannels that are joined by the nanochannels. The zoom-in illustrates the transition between the micro- and nanochannel and how the DNA becomes stretched in the nanochannel. Created in Biorender.com.*

In this specific chip, the nanochannels are 100 x 150 nm^2 in cross-section and 500 µm long. Four tubes are attached to the loading wells, allowing connection to a pressure-driven nitrogen gas (N_2) flow. This setup enables for pushing of the fluids inside the chip and thereby allow precise control of the movement of the DNA molecules.

2.3.2 Labeling methods

To acquire sequence-specific information from DNA, it must be labeled in a manner that targets unique properties of the DNA sequence. This is done by different methods that all include fluorescent DNA labels that bind to specific regions of interest in the DNA. These labeling strategies for ODM generally fall into two categories: enzyme-based and affinity-based, both of which are described below.

2.3.2.1 Enzyme-based labeling

Enzymatic labeling generally targets specific DNA sequences, resulting in sparsely distributed fluorescence along the DNA molecules, often achieved through a technique called nick-labeling [24]. Nicking enzymes create site-specific single-strand breaks, or nicks, in the DNA backbone. The nicks are repaired by fluorescently labeled nucleotides with the help of a DNA polymerase accompanied by a ligase that is used to seal the strand. This results in a DNA molecule where the previous nicks in the backbone are displayed by fluorescent dots, allowing for localization of DNA sequence. Additionally, YOYO-1 is commonly used to stain the entire DNA contour, providing visualization of the complete molecule.

However, drawbacks exist as nicks occurring too close on opposing strands can cause double strand breaks and increase DNA fragility [25]. An alternative enzymatic labeling method which does not cause DNA damage involves methyltransferases that methylate DNA by attaching co-factors at specifically recognized sequences

[7]. The co-factors can then undergo a second reaction for fluorescent labeling or can be pre-labeled before attachment. However, enzymatic labeling methods often require extensive sample preparation, particularly washing steps for removal of the unbound fluorescent dyes for subsequent analysis to be possible [25]. Additionally, these methods rely on the presence of specific enzymatic sites along the DNA.

2.3.2.2 Affinity-based labeling

Affinity-based labeling does not rely on enzymatic reactions, instead generating sequence-specific patterns based on the AT and GC base pair content in the DNA [7]. This can be done by principles such as denaturation mapping or competitive binding labeling.

Denaturation mapping is based on the principle that regions of a DNA molecule rich in AT basepairs will, due to the lower amount of hydrogen bonds, melt (i.e. the two strands will detach from each other) at lower temperature compared to regions rich in CG pairs [26]. Consequently, this partial melting roughly correspond to the DNA sequence. Combined with a DNA dye that only remains fluorescent in the non-melted double-stranded regions, this generates a fluorescence profile along the DNA molecule. This profile reflects the AT and GC ratio, visually represented as a unique "barcode" for the DNA sequence. However, this method require extensive manual involvement which makes it unsuitable for high throughput applications [25].

Competitive Binding (CB) is another method that maps the AT versus GC content in DNA. This method utilizes the non-fluorescent molecule netropsin, a natural antibiotic known for its high binding specificity to AT-rich regions of DNA [23]. In combination with the intercalating fluorescent dye YOYO-1, they bind to DNA in a competitive manner, where the binding of netropsin to AT-rich regions blocks YOYO-1 from binding to AT. Consequently, the fluorescent dye mainly binds to GC-rich regions as illustrated in Figure 2.5.

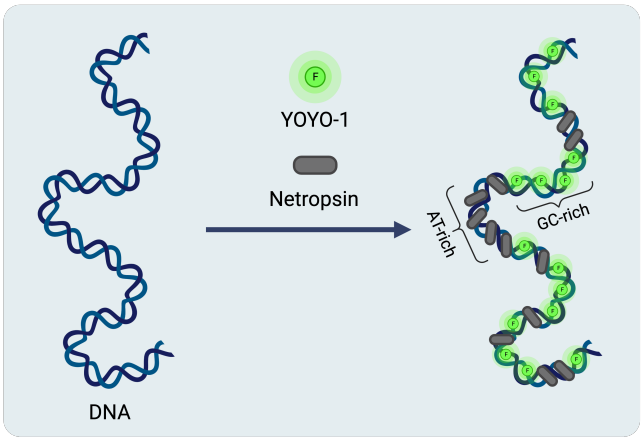


Figure 2.5: *Illustration of CB and how YOYO-1 and netropsin bind to GC- and AT-rich regions of DNA, respectively. Created in Biorender.com.*

YOYO-1 is a cyanine dye that is close to non-fluorescent in solution, but upon binding to dsDNA its fluorescence increases significantly, making it suitable for DNA studies [27],[28]. The YOYO-1 molecule is a dimer consisting of two fluorescent aromatic ring structures of oxazole yellow (YO) [27]. The two monomers are connected by a linker which allows the two monomers to intercalate into the DNA, meaning that they bind in between the base pairs [28]. The molecule can intercalate with both monomers, so called bis-intercalation, but it can also be mono-intercalated where only one monomer has bound to DNA. Its binding to DNA lacks specificity and when bound it has an absorption and emission maximum of 489 nm and 509 nm respectively [27]. As previously mentioned, the performance of the dye is influenced by ionic strength, with higher ion concentrations leading to more homogeneous staining among all DNA molecules in the sample [29]. This effect is due to a higher dissociation rate caused by increased ionic strength, allowing YOYO-molecules to rearrange between DNA-molecules and spread evenly. The equilibrium of YOYO-1 on DNA is also accelerated by higher temperatures.

In contrast to YOYO-1, netropsin lacks fluorescence and intercalating properties [30]. Instead, it strongly interacts with the minor groove of AT-rich dsDNA primarily through hydrogen bonding to A and T [31]. The amino group present on G provides sterical hindrance for netropsin to form any hydrogen bonds to GC, thus explaining its selective binding to AT.

2.3.3 λ-DNA

When performing ODM for reference-based alignment, it is important to include a standardized internal size reference for determination of the extension of the stretched DNA molecule. One commonly used reference is λ-DNA, a 48502 bp long dsDNA-molecule from the λ-phage [32]. λ-DNA exhibits a unique composition ideal for CB applications as approximately half of the molecule is AT-rich while the other is GC-rich. This composition results in a characteristic fluorescence intensity profile where half the molecule fluoresces brightly and the other half appears fainter as shown in Figure 2.6. This makes it easy to identify during ODM experiments.

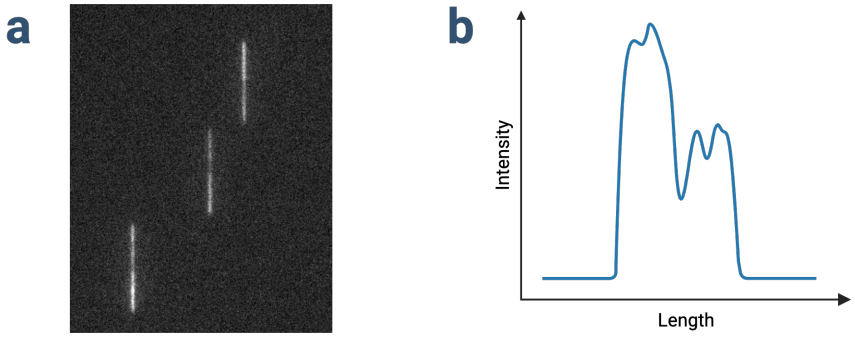


Figure 2.6: *a) Fluorescence microscopy image of CB labeled λ-DNA. b) Graph representing the fluorescence intensity profile obtained from CB labeled λ-DNA. Created in Biorender.com.*

2.3.4 Fluorescence microscopy

This thesis employs fluorescence microscopy for visualizing and imaging DNA molecules for ODM. This is done by a epifluorescence microscope with six main components as imaged in Figure 2.7: a light source, an excitation filter, an emission filter, a dichroic beamsplitter, an objective lens and a detector [33].

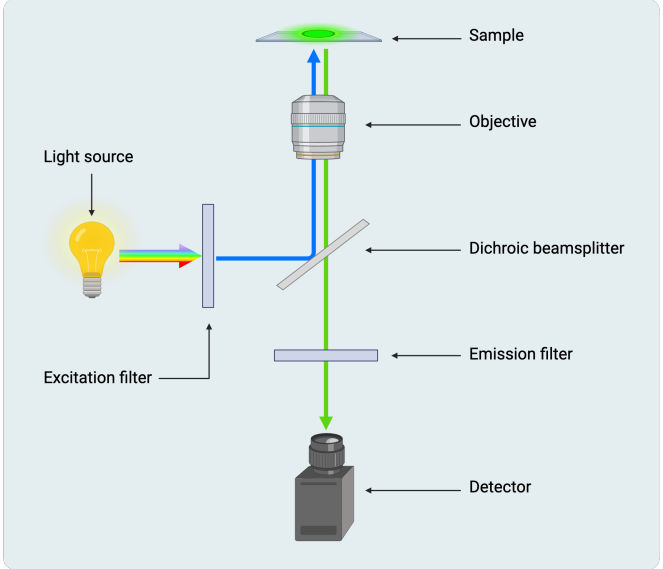


Figure 2.7: *Schematic image of the main components in an epifluorescence microscope and how light travels through and is affected by the different parts. Created in Biorender.com.*

The purpose of the light source is to provide the excitation light, which can be either monochromatic or broad-spectrum depending on the light source. The two filters and the dichroic beamsplitter are adjusted to the purpose of the experiment, where the excitation filter only lets through wavelenghts corresponding to the excitation spectrum of the fluorophore of interest. The beamsplitter directs the filtered excitation light towards the objective and sample while at the same time directing the emitted fluorescence from the sample towards the detector. The emission filter then only lets through the emitted light from fluorescence according to the emission spectrum of the fluorophore while blocking unwanted background light. The emitted light is then detected though a detector, in this setup a camera.

2.3.5 ODM for bacterial typing

The long range genetic information acquired when using ODM based on CB in combination with nanofluidic devices has been proven useful for bacterial diagnostics [2],[3]. In short, as illustrated in Figure 2.8, this method involves isolating chromosomal DNA from bacteria by lysing them in gel plugs. Subsequently, the DNA is extracted and labeled with YOYO-1, competing with netropsin for binding. These labeled DNA molecules are then stretched within nanochannels and imaged

by fluorescence microscopy, generating kymographs for individual molecules. Kymographs are barcode-like images made up of several imaged frames of the same DNA molecule aligned next to each other. These experimental kymographs are then compared against a database comprising theoretical intensity profiles generated from sequence data of bacterial genomes. This results in either a discriminative or non-discriminative match, indicating the closest match to a bacterial genome, which can be used to identify bacteria in the sample.

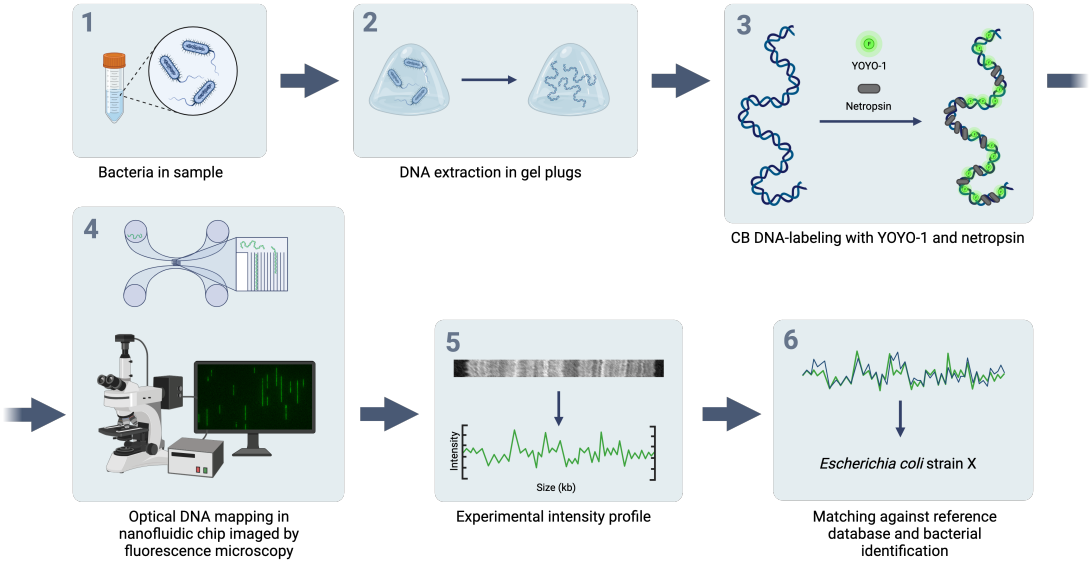


Figure 2.8: *The workflow of ODM used for bacterial identification. It is based on the CB approach for labeling and the usage of a nanofludic chip for stretching of DNA. Created in Biorender.com.*

This method have so far been used to successfully differentiate between bacteria on a species and strain level directly from uncultured, polymicrobial samples [3]. DNA was extracted directly from urine samples from patients with urinary tract infections and by using ODM, it was possible to distinguish between different pathotypes of *E. coli*.

3

Methods & Materials

This chapter covers the methods and materials used to reach the aim of this thesis including the preparation of DNA samples for ODM, the data collection in form of nanofluidic experiments and the data analysis.

3.1 Bacterial samples

The bacterial samples used in the project were cultured from fecal samples that had been collected from EHEC/STEC-infected pediatric patients in Finland between the year 2000 and 2016 [18]. The samples were selected based on their clinical relevance by being associated with HUS and with the objective to have a diverse collection of serotypes, resulting in an array of 7 samples of 6 different serotypes. The samples chosen and their respective serotype can be seen in Table 3.1.

Table 3.1: *The samples provided for the project and their respective serotype.*

Sample	Serotype
EHEC-01	O157:H7 clade 8
EHEC-02	O157:H7 non-clade 8
EHEC-03	O26:H11
EHEC-04	O145:H28
EHEC-05	O103:H2
EHEC-06	O121:H19
EHEC-07	O111:H8

3.2 DNA isolation

DNA isolation was done by Christian G. Giske's group at Karolinska Institute in a Biosafety Level 2-laboratory. The cultured bacteria were enclosed in gel plugs, lysed and treated with RNase A and proteinase K in order to extract long DNA molecules. The gel plugs were stored in buffer in 1.5 mL Eppendorf tubes and sent to Chalmers.

3.3 Sample preparation

Before being used for ODM, the isolated DNA molecules had to be free in solution, which was acquired by extraction from the gel plugs. The detailed procedures are described below.

3.3.1 DNA extraction from gel plugs

The isolated DNA was extracted by digestion of the gel plugs. This was done by first removing the buffer from the Eppendorf tubes and transferring the plugs to new 1.5 mL Eppendorf tubes. Then the plugs were washed using 300 µL 1x rCutSmart Buffer (New England Biolabs) at room temperature for 30 minutes. After that, the buffer was removed and 78 µL Milli-Q (MQ) water was added together with 20 µL 10x rCutSmart buffer. Then the tubes were incubated at 70°C in a heat block for 10 minutes to melt the agar gel followed by slow mixing by pipetting once using wide bore pipette tips. After mixing, another incubation followed at 42°C heat block for 10 minutes. Then 2 µL agarase enzyme was added to digest the agar and the solution was slowly mixed again with wide bore tips before incubation at 42°C in a heat block for 1 hour. The incubation was followed by concentration measurements using Invitrogen Qubit 3.0 benchtop Fluorometer together with Qubit™ dsDNA Quantification Assay Kits from ThermoFisher Scientific.

3.3.2 Filtering

One attempt to improve DNA stretching in the nanochannels was done by filtering the DNA samples after gel plug extraction. This to filter away excess salt remaining from extraction that could interfere with the nanofluidic experiments. It was done by placing a 0.1 µm pore sized MF-Millipore™ MCE Membrane Filter (Merck Millipore) on top of a beaker filled to the brim with MQ water, ensuring the filter was saturated with water. Between 50 and 100 µL of DNA sample was added on top of the filter and left to filtrate for 30 minutes. After, the sample was pipetted back into an Eppendorf tube.

3.4 Optical DNA mapping

After completing DNA extraction, the samples were ready for ODM which is the main part of this project. This was done by first labeling the molecules by competitive binding, then performing data collection by fluorescence microscopy of DNA stretched in nanochannels before the final step of ODM which is the data processing and analysis. These steps are described in detail below.

3.4.1 CB staining

In order to visualize the DNA molecules by fluorescence microscopy the DNA had to be stained with YOYO-1 based on a competitive binding approach along with

netropsin. Following below are the two different approaches that were taken and adjustments implemented to improve stain quality in order to improve the results.

For both approaches, the staining mixture was assembled by stocks of 5 µM YOYO-1 (Invitrogen), 500 µM netropsin (Sigma), 10 µM λ-DNA (Roche Diagnostics) and EHEC DNA sample. These were mixed in a manner which resulted in a concentration ratio of 10:1:300 between total amount of DNA base pairs (sample DNA and λ-DNA), YOYO-1 and netropsin, which is the previously established standard ratio used for CB in the research group. The total DNA concentration varied usually between 7 µM and 8.5 µM among experiments, but concentrations of 4 µM and 6 µM were also used. The concentrations of total DNA that were used in all the stains created during this project can be seen in Appendix A.1. What differed between the two approaches was that one in addition contained MQ water while the other contained 5x Tris-Borate-EDTA (TBE) buffer.

3.4.1.1 First approach - Low salt stain

As high salt concentration in the samples decreases the extension of DNA molecules when stretched in nanochannels, the first approach of staining had the objective to contain as low amount of salt as possible. For these reasons MQ water was used together with the previously mentioned filtered DNA samples.

A 10 µL stain was mixed by sample DNA, λ-DNA, netropsin, YOYO-1 and MQ water in volumes such that the concentration ratio mentioned above was achieved. The final concentration of DNA varied among samples depending on the stock concentration of the DNA in order to obtain the 10 µL stain. An example of how a stain was composed can be seen in Table 3.2.

Table 3.2: *Example composition for a low-salt stain consiting of sample DNA, λ-DNA, netropsin, YOYO-1 and MQ water.*

2023/10/25	**Stock conc.[µM]**	**Final conc.[µM]**	**Volume added[µL]**
EHEC-03 DNA	28	6	2.14
λ-DNA	10	1	1.00
Netropsin	500	210	4.20
YOYO-1	5	0.7	1.40
MQ water	-	-	1.26

The solution was prepared by first adding sample DNA, λ-DNA, netropsin and MQ water, allowing it to rest for 10 minuted before adding YOYO-1. The staining mixture was incubated in a 50°C heat block before being stored in 4°C fridge until imaging. On the day of the nanofluidic experiments the stain was diluted 10-fold to 100 µL by adding 88 µL MQ water and 2 µL β-mercaptoethanol (BME) (Sigma-Aldrich). The BME was used to prevent the DNA from breaking due to photodamage.

3.4.1.2 Second approach - High salt stain

For the second approach 5x TBE buffer was added instead of MQ water in order to have a final concentration of 0.5x TBE buffer in the stain. Due to high concentrations of DNA in the stocks, the total volume of the stain was increased to 20 µL in order to minimize effects that could be caused by pipetting small volumes. Many DNA samples were of high concentration which caused the solution to be viscous and difficult to pipette in volumes smaller than 1 µL. In addition, higher concentration of λ-DNA was used to ensure that enough λ-DNA could be imaged during data collection. The stain preparation was done following the same protocol as for the low-salt stain by firstly adding sample DNA, λ-DNA, netropsin and TBE before resting for 10 minutes. Then YOYO-1 was added followed by incubation for 30 minutes in 50°C heat block. Example final concentrations used in the stain can be seen in Table 3.3.

Table 3.3: *Example composition for a high-salt stain consiting of sample DNA, λ-DNA, netropsin, YOYO-1 and TBE buffer.*

2023/12/02	**Stock conc.[µM]**	**Final conc.[µM]**	**Volume added[µL]**
EHEC-01 DNA	36	5.5	3.06
λ-DNA	10	1.5	3.00
Netropsin	500	210	8.40
YOYO-1	5	0.7	2.80
TBE buffer	5x	∼0.5x	2.74

To compensate for eventual effects on the stretching caused by the higher salt concentration, the stain was directly after incubation diluted 15-fold instead of 10-fold to 300 µL by 274 µL MQ water and 6 µL BME. Another adjustment to the staining protocol was that the stain after dilution was stored in room temperature overnight until the experiment the following day.

3.4.2 Nanofluidic experiments

This section covers the experimental setup and procedure for the ODM data collection done through nanofluidic experiments.

3.4.2.1 Setup

In order to collect the ODM data, DNA was stretched in nanochannels and visualized by fluorescence microscopy. The nanofludic chip used was described in Section 2.3.1.2. The microscope used was Zeiss AXIO Observer.Z1 epi-fluorescence microscope equipped with a Photometrics sCMOS camera with a resolution of 110 nm/px and a 100x oil immersion objective. The lightsource was a Colibri LED and image acquisition was done with the software Zeiss ZEN 3.7 Pro.

3.4.2.2 Preparation of the nanofluidic chip for experiments

Prior to sample addition, the nanofluidic chip had to be cleaned and prepared for the experiment which was done by flushing the chip with various fluids controlled by the N_2 flow. For each round of flushing of the chip, 10 µL of fluid was added to each loading well and the fluid was flushed in firstly the microchannels for around 5 minutes followed by flushing of the nanochannels for 5 minutes. After the previous experiment, the chips was first flushed with chlorine to wash away remaining molecules in the micro- and nanochannels. Then the chlorine was replaced with MQ water and the chip was flushed and left for storage until the next use. To prepare the chip for experiment, the chip was first flushed with a buffer of 0.05x TBE and 2% BME. Thereafter the buffer was removed and sample was added to one loading well while adding new buffer to the other wells.

3.4.2.3 Experimental procedure

To start the experiments, the DNA molecules were flushed from the loading well into the microchannel where the molecules were visualized by exciting YOYO-1 using 496 nm light. The movement of the molecules was controlled by the N_2 pressure from the different loading wells which allowed molecules to flush into the field of view. When molecules of interest, i.e. long DNA and λ-DNA, appeared, the pressure could be decreased and channeled from both lower wells to push the molecules into the nanochannels. Once stretched in the nanochannels the molecules were allowed to relax for around 60 seconds before they were imaged in movies comprised of a series of images where each frame was captured using an exposure time of 100 ms and between 10-20 % light intensity. The light intensity was adjusted depending on the contrast achieved in the images or if the DNA molecules broke due to photodamage. A minimum of 10 frames were captured for each molecule, but the majority was imaged for 20 frames. It was also made certain that λ-DNA molecules were imaged. For the first round of experiments over 70 molecules were imaged, but when the staining protocol was changed it was prioritized that the new approach was quality assured, hence a minimum of 30 molecules were imaged to save time during both experiment and analysis.

3.4.3 Data processing and analysis

In order to obtain results from ODM, the microscopy data had to be processed and analysed. This included detection of molecules, kymograph generation, matching against reference database and finding discriminative matches. The majority of the data handling was done through MATLAB scripts provided by Tobias Ambjörnsson's group at Lund University.

3.4.3.1 DNA molecule detection and generation of kymographs

The movies obtained from the fluorescence microscopy were in .czi format and in order to perform further processing the images were converted to .tif format using a MATLAB script. Thereafter preprocessing was done using the software ImageJ

where frames containing molecules that were breaking were deleted from the movie. In ImageJ some images also required cropping to facilitate further analysis.

The .tif files were then loaded into DNA barcode matchmaker (DBM), a MATLAB-based program which detected each DNA molecule individually and created kymographs. The kymographs were generated by stacking each time frame of the detected molecule onto each other, resulting in an image that displayed the molecule's fluorescence intensity over all time frames. The settings used determined among other things a minimum molecule length of 110 pixels for detection, and all settings used in DBM can be seen in Appendix A.2.

To determine the extension of the molecules in the nanochannels, the Lambda pipeline in DBM was used. In this pipeline, λ-DNA in the images were identified and their length measured. The pipeline calculated the extension of all the DNA in the nanochannels by dividing the average λ-DNA length by the number of basepairs in λ-DNA (48 502). This results in an extension factor in the unit of nm/bp which was needed for further analysis of the data.

After DBM, a quality check was performed where each kymograph was observed individually and so called "bad kymographs" were discarded. This included kymographs where the molecule broke during imaging and kymogrpahs generated by wrongly detected artefacts in the image. The kymographs were also filtered by a cut-off of a minimum length of 150 kB. Since the data collection was done manually, it was made certain that only single molecules longer than 150 kB were imaged. However, the procedure of imaging λ-DNA resulted in other DNA molecules also being imaged. These were usually shorter DNA fragments that rarely were long enough to be of value for analysis. However after DBM, kymographs of these molecules remained in the data thus making filtering required to only have the specific molecules of interest in the dataset.

3.4.3.2 Matching against reference intensity profiles

The raw kymographs were thereafter loaded into another MATLAB script called Human Chromosome Alignment (HCA), used for reference-based alignment of kymographs to theoretical CB intensity profiles.

The input in HCA was the kymographs and a reference in form of a database consisting of 20293 theoretical intensity profiles created from sequencing data of bacterial genomes. The settings used for HCA analysis was the experimental stretch factor along with a varying stretch factor and a step length. The experimental stretch factor was calculated from the length of the λ-DNA size reference in the nanochannels and was used to adjust the reference genomes in order to match the extension of the experimental intensity profiles. A varying stretch factor of 10% and a step length of 2.5% was used to cover variations in the extension, meaning that stretches from 90% of the molecule length to 110% separated by steps of 2.5% were compared to the references.

First, the raw kymographs were aligned which meant that drifting of molecules in

the nanochannels and the resulting movement in the kymographs was adjusted for by leveling the ends of the molecules in each time frame with each other. Thereafter from the kymographs, the script generated experimental intensity profiles based on the average fluorescence intensity of all time frames along the length for each molecule. The intensity profiles were then matched against the theoretical database.

The output from HCA was a .txt file containing information on all the experimental intensity profiles analyzed compared to all the genomes in the database. The match between experimental and theoretical intensity profiles was evaluated by a CC-score. The CC-score is calculated based on the Pearson correlation coefficient and ranges between -1 and 1, where a higher CC-score indicates a better match.

3.4.3.3 Finding discriminative matches

For analysing the output from HCA, the software R was used. The .txt file was loaded into a discriminative matches program where the output was the discriminative matches, meaning that a molecule matched significantly to a reference genome on the species level. This was done by comparing the highest CC-scored match for each molecule to the CC-score of the second best matched species. If these two CC-scores differed by more than 0.05, the match was considered discriminative. Another criteria for a discriminative match was that the CC-score of the match must be over 0.5. This resulted in an output with a table that for each discriminative match stated which kymograph had matched to which bacterial species strain, the CC-score of the match and the molecule's length.

3.4.3.4 Discriminative resolution

A further approach to evaluate the discriminative matches is to divide the matches into taxonomic groups at a higher resolution than species, so called strain groups (SGs). This was done using the same method as in [3]. The approach is based on phylogenetic trees where bacterial strains are divided into SGs based on core genes that indicate a common ancestor. The so called resolution of the SG is determined by where the SG can be placed in the phylogenetic tree.

The data obtained from HCA was processed using a Python-based script which assigned (if possible) each discriminative match to a SG on the basis on the set resolutions (h): SG_{low} h=0.05, SG_{medium} h=0.015, SG_{high} h=0.008 and $SG_{ultra\ high}$ h=0.003.

3.5 Stretching DNA on glass slides

In addition to ODM, sample EHEC-05 was also stretched and imaged on glass slides during the last part of the project to evaluate the amount and length of DNA in the sample. This was done to evaluate whether there still was usable DNA in the sample. The steps required for DNA imaging on glass are described below.

3.5.1 Preparation of silanized coverslips

In order to be able to stretch the DNA on glass, the coverslips used needed to be activated which was achieved through treatment with silanizing compounds as previously mentioned. Coverslips were placed in a small rack and placed using tweezers in a 100 mL beaker containing 50 mL acetone, 500 µL (3-Aminopropyl)triethoxysilane (APTES) (Sigma-Aldrich) and 500 µL allyltrimethoxysilane (ATMS) (Sigma-Aldrich). The beaker was covered with parafilm with a small ventilation hole and the coverslips were incubated in room temperature for 2 hours. After incubation the solution was replaced with acetone while making sure that the coverslips did not dry out during the exchange of liquids.

3.5.2 Sample visualization

For sample visualization, stains of a final volume of 30 µL were prepared with a final concentration of 10 µM DNA and 1 µM YOYO-1 according to the previously mentioned ratio of 10:1 DNA:YOYO-1. This was done by adding DNA, YOYO-1 and MQ water to 0.5 mL Eppendorf tubes and incubating the stain for 10 min in 37°C heat block. After incubation, 1 µL of BME was added. To prepare the glass slides for DNA stretching, two lines of vaseline were spread on a glass slide using a cotton swab as seen in Figure 3.1. One coverslip was taken from the acetone using tweezers, dipped in MQ water followed by removal of all water using pressurized N_2. Then the coverslip was placed on top of the vaseline and attached by pressing it down. The sample was added to the glass slide by pipetting 1 µL of sample onto the slide from the side of the coverslip.

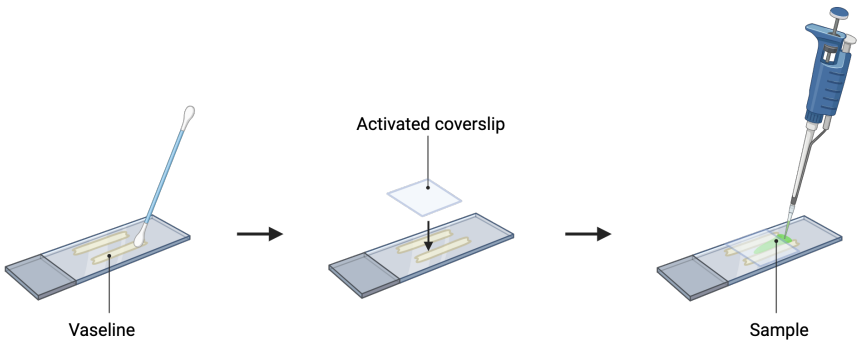


Figure 3.1: *Illustration of the procedure for sample addition to glass slides with activated coverslip for stretching of DNA. The figure illustrates from the left how vaseline is added with a cotton swab to create a channel leading the DNA in one direction. Then the activated coverslip is attached to the vaseline and the sample is added from the side of the coverslip. Created in Biorender.com*

The glass slides were imaged using the same microscope setup as for the nanofluidic experiments. Images were taken through the snapshot function in Zeiss ZEN 3.7 Pro resulting in single-frame .czi files ready for analysis.

3.5.3 Data analysis

The .czi images of glass slides were converted into .tif using a MATLAB script. Thereafter the images were processed using ImageJ. The software was used to adjust brightness and contrast of the images and to measure the length of DNA molecules captured. From previous experiments done in the research group, it had been established that DNA has a stretch factor between 0.3 and 0.4 nm/bp when stretched on glass. Assuming that the stretching experiment worked as expected, the largest stretch factor of 0.4 was used when measuring the DNA length. Combined with the microscope camera resolution of 110 nm/pixel, this allowed the measurements to be converted from pixels into basepairs length.

4

Results & Discussion

In this chapter, the results from the project are presented and discussed. The results obtained from HCA are first analyzed in terms of the experimental conditions that affect the fraction of discriminative matches at the species level. Then, the determination of serotypes and strain sub groups using ODM are discussed. Lastly, results from DNA stretched on glass are presented.

4.1 Discriminative matches

The first part of this project consisted of testing two different staining approaches in order to obtain the highest fraction of discriminative matches from ODM. Below, the results regarding the amount of discriminative matches are presented and discussed on the basis of ion concentration during the experiments as well as the length of the molecules captured.

4.1.1 Effect of ionic strength on discriminative matches

At the beginning of the project, all experiments were conducted with the objective to achieve a high stretch factor. As stated in [3], it is beneficial to have a stretch factor of at least 0.20 nm/bp to achieve good results. This because the more the DNA is stretched, the better the fluorescent signal representing the sequence is able to be captured. Therefore, the CB stains were prepared with low concentration of salt in the solution. However, this led in most cases to uneven and faint staining of the DNA molecules, making it difficult to perform imaging experiments due to varying fluorescence, and resulted in fewer discriminative matches for all the samples as shown by the blue bars in Figure 4.1. The percentage of discriminative matches for a sample is the proportion of discriminative matches to the total number of DNA molecules in the sample that are longer than 150 kbp.

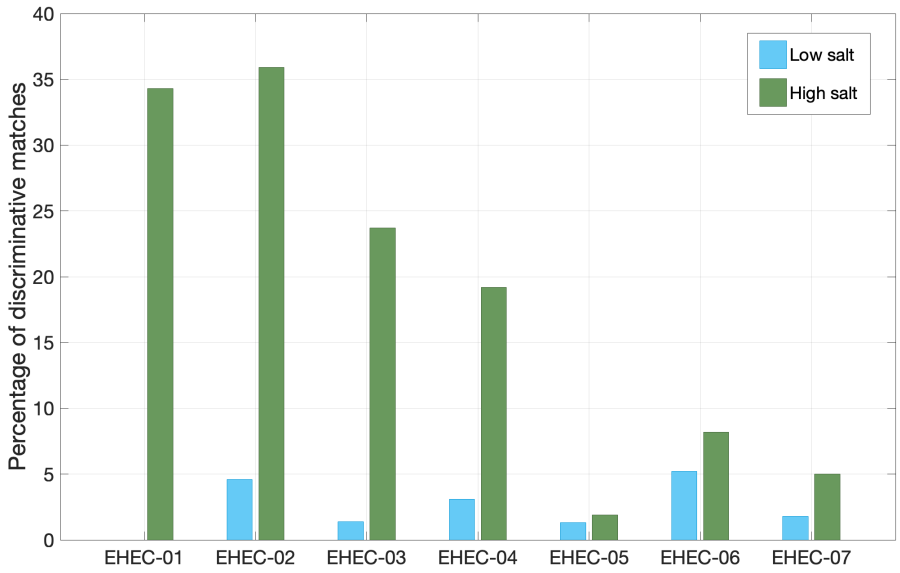


Figure 4.1: *Bar chart illustrating the fraction of discriminative matches on a species level for DNA molecules longer than 150 kbp for both low-salt and high-salt stains. Low- and high-salt matches are illustrated by blue and green bars respectively.*

For low-salt stains, all samples resulted in less than 5% discriminative matches on a species level, and for EHEC-01 there were no matches at all. The low fraction of matches could be explained by the bad quality of stain caused by the low concentration of ions. As shown in [29], lower ionic strength affects the dissociation rate constant of YOYO-1 from DNA meaning that it takes longer time to reach an equilibrium and thus a uniform stain. So in this case, it is unlikely that the samples that were analysed had been allowed to reach stain equilibrium. In addition, these samples were stored in the fridge after staining and since temperature also is a factor that affects the equilibrium, the cold temperature in the fridge made this process too slow. The resulting uneven, faint fluorescence displayed by these molecules were thus not representative for the "true" intensity profile and could not discriminatively be matched to any reference genome.

In order to improve the fraction of discriminative matches, several changes were made to the staining protocol, as previously described in the methods. First, the ion concentration in the stain was increased by using TBE buffer instead of MQ water. Second, the samples were allowed to incubate in room temperature overnight instead of in a fridge so as to provide a warmer temperature for the YOYO-1 and DNA to reach equilibrium. Third, to compensate for the decrease in extension caused by higher salt concentration, the samples were diluted 15 times instead of 10 times after staining which would mean that staining is done in high salt concentration while the subsequent dilution decreases the fraction of salt for nanofluidic experiments. The resulting proportions of discriminative matches are shown as the green bars in Figure 4.1.

The improvements implemented resulted in an increase in discriminative matches for all samples. Especially EHEC-01, -02, -03 and -04 had a large increase in fraction of discriminative matches, spanning from around 20% up to 35%. This indicates that the high-salt stain had the desired effect on the proportion of discriminative matches. However, there were still a low fraction of matches for samples 05-07. This can for EHEC-05 be explained by that only one experiment with high-salt stain was performed before the sample was deemed unusable, as discussed later in Section 4.3. Therefore, it was not possible to perform additional experiments to evaluate this sample further.

Another aspect to take into consideration when comparing high- and low salt stains is how the stretch factor was affected by changing the staining protocol. The stretch factors for the experiments with the best results in regard to fraction of discriminative matches done in high and low ionic strength are presented in Table 4.1.

Table 4.1: *Stretch factors obtained for example experiments of each sample in low-salt and high-salt stain.*

Sample	Stretch factor low salt	Stretch factor high salt	Difference
EHEC-01	0.190	0.215	+0.025
EHEC-02	0.236	0.250	+0.014
EHEC-03	0.200	0.219	+0.019
EHEC-04	0.252	0.232	-0.020
EHEC-05	0.200	0.198	-0.002
EHEC-06	0.243	0.203	-0.040
EHEC-07	0.230	0.228	-0.002

It can be seen in Table 4.1 that for samples 04-07 the stretch factor is decreased by staining in buffer of higher ionic strength, which would be in line with previous studies [29]. However, the extension for the majority of the samples remained around at least 0.20 nm/bp, which indicates that the increased dilution had the wanted effect in preventing too large a decrease in stretch factor. Nevertheless, it is important to note that the DNA extraction from gel plugs results in a certain ion concentration in the original DNA stocks, since the rCutSmart buffer used for extraction contains salt. This salt concentration is not known, and depending of sampling from the DNA stock, there will be different salt concentrations in different samples. This can explain the varying stretch factors obtained among experiments.

To conclude this section, the experiments resulting in the most discriminative matches are presented in Table 4.2. The highest amount of discriminative matches were obtained when staining in high salt, with the exception of EHEC-05.

Table 4.2: *The highest amount of matches for each sample, EHEC-05 was low-salt stain, the others high-salt.*

Sample	>150 kB	Matches	Percentage of matches
EHEC-01	35	12	34.3
EHEC-02	39	14	35.9
EHEC-03	38	9	23.7
EHEC-04	26	5	19.2
EHEC-05*	158	2	1.3
EHEC-06	49	4	8.2
EHEC-07	40	2	5.0

From this it can be concluded that the high-salt stain was the approach that worked the best for a majority of the samples. This correlates with literature stating that salt improves fluorescent dye performance, and as seen in Table 4.1 the negative effects on stretching can be minimized. For EHEC-05, as mentioned earlier, only one unsuccessful experiment was performed with high-salt stain. Even though that high-salt stain resulted in a higher fraction of discriminative matches, the best experiment from low-salt stain still yielded a higher number of discriminative matches (1 versus 2 discriminative matches), which is why the best low-salt result is presented in the table instead.

4.1.2 Effect of molecule length on discriminative matches

Another factor that has been shown to affect the fraction of discriminative matches is the length of the fragments [2]. In Figure 4.2, the size range for the discriminative matches for all samples is visualized as a box plot.

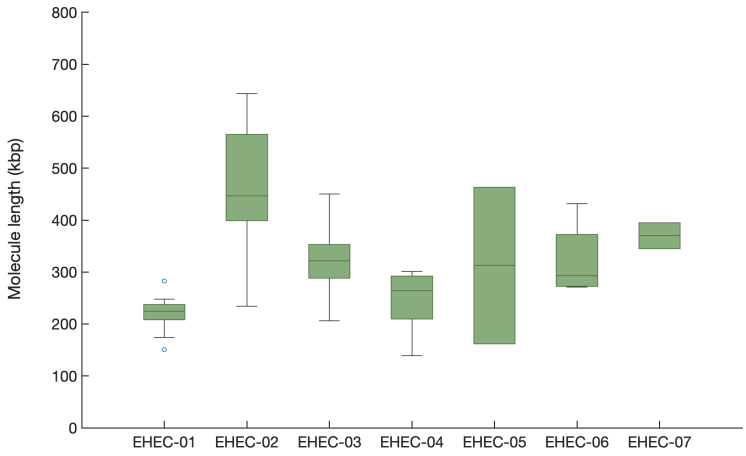


Figure 4.2: *Box chart showing the size range of molecules that had discriminative matches. The top and bottom edges of each box are the upper and lower quartile and the line inside is the median. The extended vertical lines above and below the boxes indicate the maximum and minimum values for each sample.*

The size ranges are varying where EHEC-01 that has the shortest fragments and the smallest range and EHEC-02 has the longest fragments and also the widest range as seen in Figure 4.2. Sample EHEC-03, 04 and 06 fall somewhere between 01 and 02 with sizes between 150 and 400 kbp. For EHEC-05 and -07, the size ranges in the box chart are numbers are based on a low amount of molecules as both these samples only had 2 discriminative matches. This explains the unusual appearance of the boxes representing these samples seen in the chart.

In [2], it was shown that the fraction of discriminative matches increased with longer DNA fragments, where a fragment size of 300 pixels corresponded to around 20% distinct matches. The cut-off set to 150 kbp correlates approximately to 250-300 pixels depending on extension, meaning that a minimum fraction of 20% discriminative matches could be expected. As shown in Figure 4.1, this was achieved for samples 01-04, but not for 05-07.

An interesting thing to observe is that in combination with the results seen in Figure 4.1, the highest proportion of matches were achieved for EHEC-01 and -02, they both had around 35% discriminative matches. However, the size range of these discriminative matches has a vast difference as EHEC-01 had shorter molecules over a small range of sizes while EHEC-02 covers a wider range with longer molecules. Still these two samples had similar results, which indicates that for EHEC-01 it was not necessary to have much longer molecules than 200 kbp in order to obtain discriminative matches.

4.2 Evaluation of matches

When discussing discriminative matches it is important to note that it means, as mentioned in the methods, that a DNA molecule has matched to a certain bacterial genome with a CC-value at least 0.05 higher than the second best species match. It does not take into consideration what species and subsequent subspecies of bacteria it matched to. It is therefore of interest to evaluate the matches with regard to them matching to the correct species and subspecies. In this section the nature of the matches are firstly presented and discussed on the basis of species and serotype matches. Furthermore, a previously used approach of evaluating the matches resolution based on strain groups is discussed.

4.2.1 Matches to correct serotype

As previously described, the discriminative matches are obtained from the HCA output processed in R. This results in a list of bacterial IDs that state the strain the analyzed molecule matched to and it is necessary to assess whether this was a serotype match, a species match or a false positive match to another bacterial species. In most of these matches, it is stated in the ID which serotype a particular strain belongs to, in some only the O-type is stated and for others the strain has another name that does not involve the serotype. In those cases it is required to perform literature search to confirm the serotype. For this project, a match where

only information about the strain's O-type was available also counted as a serotype match. The results from this is summarized as a confusion matrix in Table 4.3.

Table 4.3: *Confusion matrix visualizing the nature of the discriminative matches.*

	Matched Serotype						
Sample	O157: H7	O26: H11	O145: H28	O103: H2	O121: H19	O111: H8	Other E.coli strain
EHEC-01 O157:H7	11	0	0	0	0	0	1
EHEC-02 O157:H7	13	0	0	0	0	0	1
EHEC-03 O26:H11	0	7	0	0	0	0	2
EHEC-04 O145:H28	0	0	3	0	0	0	2
EHEC-05 O103:H2	0	0	0	0	0	0	2
EHEC-06 O121:H19	0	0	0	0	3	0	1
EHEC-07 O111:H8	0	0	0	0	0	1	1

This table shows how many discriminative matches were to the correct serotype of *E. coli* or if they were not, if they matched to the correct species. All samples except 05 had at least one match to correct serotype and for all samples the remaining matches were all to other *E. coli* strains, meaning that there were no false positives and a 100% true positive rate was achieved at a species level.

Another aspect of the discriminative matches is the distribution between serotype and species matches. Figure 4.3 shows the distribution of the discriminative matches as pie charts where serotype matches are denoted by dark blue color and other *E.coli* strain matches are in light blue color.

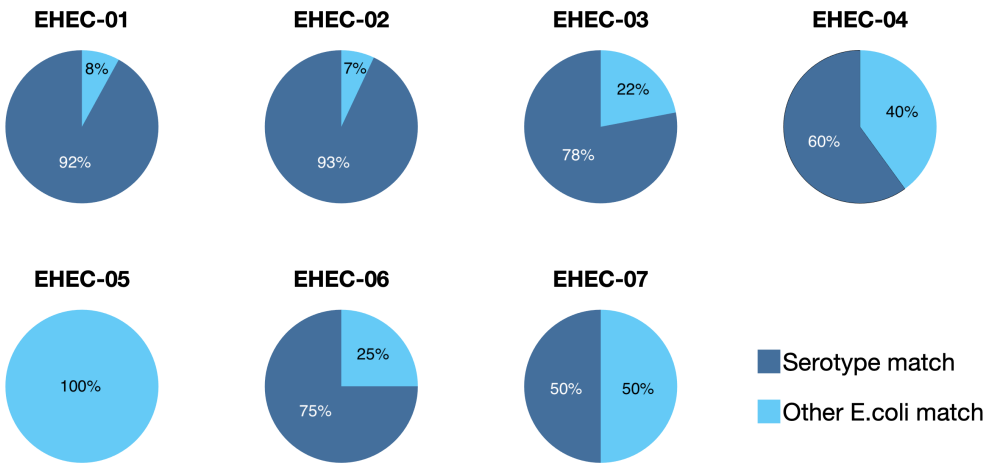


Figure 4.3: *Matches distribution between serotype matches and species matches to other E.coli strains visualized in the form of pie charts for each sample. Matches to the correct serotype are illustrated by dark blue color and macthes to other E.coli strains are illustrated by light blue color.*

As seen in the figure, of all discriminative matches to *E. coli*, a majority of those matches were to the correct serotype with the exception of EHEC-05. The high amount correct seroype matches suggest that there are certain features in the genome of these bacteria that are unique for their serotype, otherwise there would be a higher amount of other *E.coli* matches.

It is important to bear in mind that when utilizing databases, as in this case for matching an "unknown" sample to what is known, a certain level of bias can be expected. For EHEC serotypes, O157:H7 has historically been the predominant serotype in EHEC infections, resulting in it being more well studied compared to other serotypes. Consequently, the database used likely contains a larger number of O157:H7 strains than strains from other serotypes, potentially impacting the obtained results. As previously presented, EHEC-01 and -02 exhibited both the highest fraction of discriminative matches and the largest proportion of serotype matches. These samples also have in common that they share the same serotype, namely O157:H7. This commonality suggest that during matching, there were more genomes in the database that these samples could potentially match to, making it easier to find a match. This provides a possible explanation to why 01 and 02 performed so well when analyzed using ODM.

4.2.2 Matches resolution based on strain groups

In Table 4.4, it is presented how many of the discriminative matches could be assigned a SG at the four set resolutions: SG_{low}, SG_{medium}, SG_{high}, $SG_{ultra\ high}$. The 867 *E. coli* reference genomes in the database had been divided into 10 SG_{low}, 34 SG_{medium}, 88 SG_{high} and 180 $SG_{ultra\ high}$ SG's, respectively.

Table 4.4: *Discriminative resolution of the discriminative matches to SG's on four different levels of resolution. The columns headed by # indicates the number of matches assigned to that particular SG.*

		Discriminative resolution							
		SG_{low}		SG_{medium}		SG_{high}		$SG_{ultra\ high}$	
Sample	# of discriminative matches	**SG**	#	**SG**	#	**SG**	#	**SG**	#
EHEC-01	12	**5**	9	**13**	5	**40**	5	**96**	5
		7	1	**23**	1	**56**	1	**141**	1
EHEC-02	13	**5**	6	**13**	1	**40**	1	**96**	1
EHEC-03	9	**2**	9	**5**	1	**21**	1	**58**	1
				6	6	**27**	6	**79**	6
EHEC-04	5	**5**	4	**16**	3	**46**	3	**103**	3
EHEC-05	2	**2**	2	**6**	1	**27**	1	**79**	1
EHEC-06	4	**2**	2	**7**	3	**33**	3	**89**	3
EHEC-07	2	**2**	2	**6**	1	**27**	1	**79**	1

It can be seen that many of the discriminative matches can be assigned a SG at a high resolution, indicating utility for additional studies of this manner. However, these results are based on the set resolution from [3], where bacteria were distinguished based on strain groups in a particular phylogenetic tree that was created based on many different genes. For the aim of this project, to extract serotype information which only is based on the O- and H-antigens, further bioinformatic analysis and development of the method is required.

4.3 Visualization of EHEC-05 stretched on glass slides

After several rounds of nanofluidic experiments on the EHEC-05 sample, where only very short DNA molecules unsuitable for ODM were observed, the sample was stretched on glass slides to visualize the length and fragmentation of the molecules. Two resulting images are displayed in Figure 4.4.

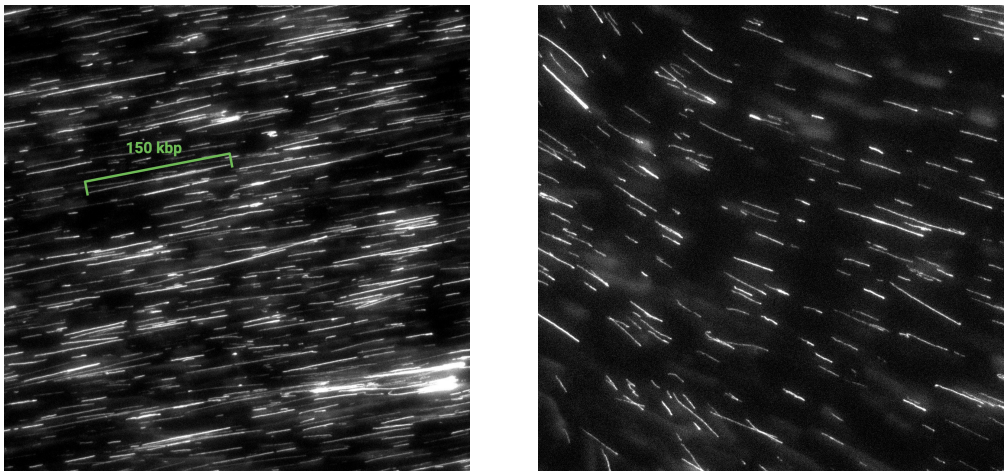


Figure 4.4: *Fluorescence microscopy images of EHEC-05 stretched on glass slides, with length measurement in green of one 150 kbp long molecule in the left image.*

As seen in the figure, the images show fragmented pieces of DNA stretched on glass. DNA molecules longer than 150 kb could not be observed. One reason for this could be that the DNA molecules were sheared by the pipetting steps. It could also be due to degradation of the DNA over time. The extraction of the DNA sample from its gel plug was done in the beginning of September while these images were taken in late November. As the sample was fragmented, we decided not to perform any more experiments on it, explaining the lack of results obtained from this sample.

5

Conclusion

The results suggest promising possibilities for ODM to be used for EHEC serotype determination. Consistent with previous studies, ODM has demonstrated effectiveness in species identification of *E. coli*. As presented in the results, the majority of these species matches were also to subspecies strains of the correct serotype. This suggest that ODM efficiently identifies properties unique for each serotype and can thus be used to differentiate between different serotypes. Assigning the discriminative matches to strain groups also provided promising indications of achieving matches with high taxonomic resolution. However, further data collection is required to extensively evaluate ODM performance. The results achieved for samples EHEC-05, -06 and -07 are inconclusive and require new experiments.

5.1 Future perspectives

To take this research further, it is of interest to extract additional bioinformatic information from the ODM results. Depending on the basis which the phylogenetic trees used for determination of resolution of the discriminative matches are created, genetic relationships between samples can be established, which could be valuable for epidemiological surveillance purposes.

In addition, it is necessary to evaluate ODM for clinical relevance by also applying the method on uncultured samples. If proven stable and reliable, ODM for determination of EHEC serotypes could be an alternative to current methods for serotyping that with further development can be both quicker and require less hands-on involvement.

Bibliography

1. Ståhl K et al. Surveillance of Infectious Diseases in Animals and Humans in Sweden 2021. Tech. rep. Uppsala: National Veterinary Institute (SVA), 2021
2. Müller V et al. Cultivation-Free Typing of Bacteria Using Optical DNA Mapping. ACS infectious diseases 2020; 6:1076–84
3. Nyblom M et al. Strain-level bacterial typing directly from patient samples using optical DNA mapping. Communications Medicine 2023; 3:31
4. Fratamico PM, DebRoy C, Liu Y, Needleman DS, Baranzoni GM, and Feng P. Advances in molecular serotyping and subtyping of Escherichia coli. Frontiers in Microbiology 2016; 7
5. Alberts B et al. Cells and Genomes. *Molecular Biology of the Cell*. 6th ed. New York: Garland Science, 2015. Chap. 1:1–42
6. Alberts B et al. DNA, Chromosomes, and Genomes. *Molecular Biology of the Cell*. 6th ed. New York: Garland Science, 2015. Chap. 4:175–236
7. Frykholm K, Müller V, Kk S, Dorfman KD, and Westerlund F. DNA in nanochannels: theory and applications. Quarterly Reviews of Biophysics 2022; 55:1–27
8. Reisner W, Pedersen JN, and Austin RH. DNA confinement in nanochannels: Physics and biological applications. Reports on Progress in Physics 2012; 75
9. Persson F and Tegenfeldt JO. DNA in nanochannels-directly visualizing genomic information. Chemical Society Reviews 2010; 39:985–99
10. Salton MRJ and Kim KS. Structure. *Medical Microbiology*. Ed. by Baron S. 4th ed. Galveston (TX): University of Texas Medical Branch at Galveston, 1996. Chap. 2
11. Alberts B et al. Pathogens and Infection. *Molecular Biology of the Cell*. 6th ed. New York: Garland Science, 2015. Chap. 23:1263–96
12. Kaper JB, Nataro JP, and Mobley HL. Pathogenic Escherichia coli. Nature Reviews Microbiology 2004; 2:123–40
13. Melton-Celsa AR. Shiga Toxin (Stx) Classification, Structure, and Function. Microbiology Spectrum 2014; 2
14. Ogura Y et al. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. PNAS 2009 Oct; 106
15. Bruyand M et al. Hemolytic uremic syndrome due to Shiga toxin-producing Escherichia coli infection. Medecine et Maladies Infectieuses 2018; 48:167–74

16. Fuller CA, Pellino CA, Flagler MJ, Strasser JE, and Weiss AA. Shiga toxin subtypes display dramatic differences in potency. Infection and Immunity 2011; 79:1329–37
17. Joseph A, Cointe A, Kurkdjian PM, Rafat C, and Hertig A. Shiga Toxin-Associated Hemolytic Uremic Syndrome: A Narrative Review. Toxins 2020; 12
18. Bai X et al. Comparative Genomics of Shiga Toxin-Producing Escherichia coli Strains Isolated from Pediatric Patients with and without Hemolytic Uremic Syndrome from 2000 to 2016 in Finland. Microbiology Spectrum 2022; 10
19. DebRoy C, Roberts E, and Fratamico PM. Detection of O antigens in Escherichia coli. 2011
20. Deen J et al. Combing of genomic DNA from droplets containing picograms of material. ACS Nano 2015; 9:809–16
21. Allemand JF, Bensimon D, Jullien L, Bensimon A, and Croquette V. pH-Dependent Specific Binding and Combing of DNA. Biophysical Journal 1997; 73:2064–70
22. Tegenfeldt JO et al. The dynamics of genomic-length DNA molecules in 100-nm channels. PNAS 2004; 101:10979–83
23. Nyberg LK et al. A single-step competitive binding assay for mapping of single DNA molecules. Biochemical and Biophysical Research Communications 2012; 417:404–8
24. Müller V and Westerlund F. Optical DNA mapping in nanofluidic devices: principles and applications. Lab on a Chip 2017; 17:579–90
25. Müller V et al. Enzyme-free optical DNA mapping of the human genome using competitive binding. Nucleic Acids Research 2019; 47
26. Reisner W et al. Single-molecule denaturation mapping of DNA in nanofluidic channels. PNAS 2010; 107:13294–9
27. Rye HS et al. Stable fluorescent complexes of double-stranded DNA with bis-intercalating asymmetric cyanine dyes: properties and applications. Nucleic Acids Research 1992; 20:2803–12
28. Kucharska K, Pilz M, Bielec K, Kalwarczyk T, Kuźma P, and Hołyst R. Two intercalation mechanisms of oxazole yellow dimer (Yoyo-1) into dna. Molecules 2021; 26
29. Nyberg L, Persson F, Åkerman B, and Westerlund F. Heterogeneous staining: A tool for studies of how fluorescent dyes affect the physical properties of DNA. Nucleic Acids Research 2013; 41:184
30. Berman HM, Neidle S, Zimmer C, and Thrum H. Netropsin, a DNA-binding oligopeptide structural and binding studies. Biochimica et Biophysica Acta 1979; 561:124–31
31. Bailly C and Chaires JB. Sequence-Specific DNA Minor Groove Binders. Design and Synthesis of Netropsin and Distamycin Analogues. Bioconjugate Chemistry 1998; 9:514–38
32. Casjens SR and Hendrix RW. Bacteriophage lambda: Early pioneer and still relevant. Virology 2015; 479-480:310–30
33. Lakowicz JR. Instrumentation for Fluorescence Spectroscopy. *Principles of Fluorescence Spectroscopy*. 3rd ed. New York: Springer US, 2006. Chap. 2

A

Appendix 1

A.1 In-stain DNA concentrations

The in-stain concentrations of DNA used for att the stains created during the project can be seen in Table A.1. This includes both the amount of sample DNA, λ-DNA and the resulting total amount of DNA.

Table A.1: *In-stain DNA concentrations.*

Sample	Stain	DNA [µM]	λ-DNA [µM]	Total DNA [µM]
EHEC-01	1	6	1	7
	2	6	2	8
	3	6.5	1.5	8
	4	5.5	1.5	7
EHEC-02	2	5	1	6
	3	6	2	8
EHEC-03	1	6	1	7
	2	5.5	1.5	7
EHEC-04	1	6	1	7
	2	6	2	8
	3	5.5	1.5	7
EHEC-05	1	6	1	7
	2	6	1	7
	3	6	2	8
EHEC-06	1	3	1	4
	2	5.5	1.5	7
EHEC-07	1	6	1	7
	2	7	1.5	8.5
	3	7	1.5	8.5

A.2 DBM Settings

Settings used for DNA molecule detection and kymograph generation in DBM can be seen in Figure A.1.

Figure A.1: *Screenshot of the settings used for DBM.*

DEPARTMENT OF LIFE SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY