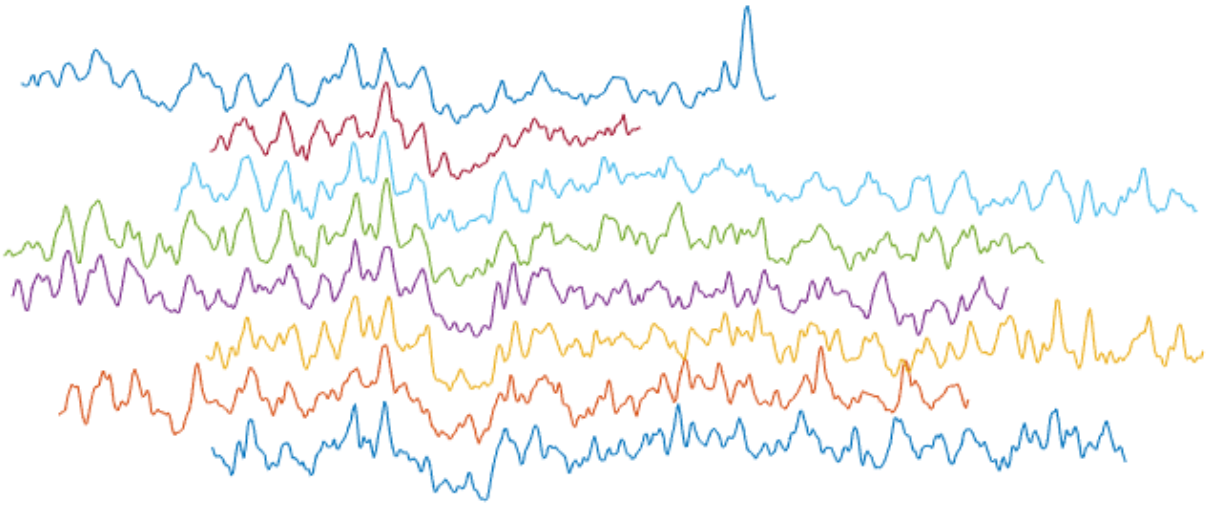

Optical DNA Mapping of the *Saccharomyces cerevisiae* Genome

A First Step Towards *De novo* Assembly of the Genome of *Saccharomyces cerevisiae*

Master's thesis in Biotechnology

HANNA ZACHRISSON

DEPARTMENT OF LIFE SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

Master's thesis 2023

Optical DNA Mapping of the *Saccharomyces cerevisiae* Genome

A First Step Towards *De novo* Assembly of the Genome of *Saccharomyces cerevisiae*

HANNA ZACHRISSON

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Life Sciences
Division of Chemical Biology
Westerlund BioNanoFluidics Research Group
Chalmers University of Technology
Gothenburg, Sweden 2023

Optical DNA Mapping of the *Saccharomyces cerevisiae* Genome
A First Step Towards *De novo* Assembly of the Genome of *Saccharomyces cerevisiae*
HANNA ZACHRISSON

Supervisor: Luis Mario Leal Garza, Department of Life Sciences, Chalmers
Examiner: Fredrik Westerlund, Department of Life Sciences, Chalmers

Master's Thesis 2023
Department of Life Sciences
Division of Chemical Biology
Westerlund BioNanoFluidics Research Group
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Visualisation of a bargroup made by the alignment of several experimental barcodes.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2023

Optical DNA Mapping of the *Saccharomyces cerevisiae* Genome
A First Step Towards *De novo* Assembly of the Genome of *Saccharomyces cerevisiae*
HANNA ZACHRISSON
Department of Life Sciences
Chalmers University of Technology

Abstract

DNA molecules are present in every living organism and contain all the genetic information necessary for an organism's existence. It is of great interest to extract this information from the DNA molecules and one common approach to doing so is through sequencing, in which the sequence of the DNA molecules is obtained with base pair resolution. However, structural variations (SVs), events affecting the chromosome organisation, can be challenging to detect with sequencing techniques as relatively short reads are generated which are hard to localise in the genome. SVs are however important to detect as they could be associated with diseases. Optical DNA mapping (ODM) is a complementing technique for obtaining genomic information, including SVs. In this master thesis project, ODM was performed on the genome of the budding yeast *Saccharomyces cerevisiae* (strain BY4742), with the aim to perform reference-based assembly and explore first steps towards a *de novo* assembly of its genome. This was performed by culturing the yeast strain, extracting the DNA from the yeast cells, and performing ODM based on competitive binding (CB). ODM was performed by labelling the DNA molecules in a sequence-specific manner using YOYO-1 and netropsin, stretching the DNA molecules in nanochannels in a nanofluidic chip, and imaging the molecules using fluorescence microscopy. Kymographs and corresponding barcodes were generated from each DNA molecule and were used for the data analysis. With the generated barcodes, reference-based alignment was performed as well as *de novo* assembly experiments. The results from the reference-based alignment indicate good quality data which could be used to perform the *de novo* assembly. However, especially chromosome I and XII showed unexpected results which could indicate complex or currently unmapped regions within these chromosomes. A complete *de novo* assembly of the genome was not obtained in this project. However, high quality 'bargroups' were generated from the assembly experiments, which indicates that the assembly pipeline could potentially assemble the whole genome. To this date, this project is the first time CB-based ODM using a nanofluidic chip is applied on the genome of *S. cerevisiae*, and is a first step towards a *de novo* assembly of its genome using ODM. By further improving the assembly pipeline, there is a high probability that the data collected in this project is able to generate a reliable *de novo* optical genome assembly of the yeast *Saccharomyces cerevisiae*, which could be used for detecting SVs. A future objective is to expand and apply this technique on the human genome to detect SVs associated with diseases, and thereby utilising it as a diagnostic tool.

Keywords: DNA, Structural variations, Optical DNA Mapping, *Saccharomyces cerevisiae*, Competitive binding, YOYO-1, Netropsin, Reference-based alignment, *De novo* assembly.

Acknowledgements

Firstly, I would like to express my gratitude to my supervisor of this project, Luis Leal Garza. You have been an amazing supervisor and I would like to thank you for all the support and guidance throughout this project. Without you this project would not have been possible.

I would also like to show my gratitude to my examiner Fredrik Westerlund, within whose division and group this project was conducted. Thank you for letting me do this project and having this experience. I would also like to thank all of the members in Fredriks group, Westerlund BioNanoFluidics Research Group, who have been so helpful and made me feel very welcome.

Finally, I must thank my friends and family for all the support, not only throughout my studies but also in my life. I could not have done this without you by my side. To my friends, you have made these five years unforgettable and it would not have been the same without you. And to my family, thank you for always supporting me and believing in me. You are the best. I love you.

Hanna Zachrisson, Gothenburg, June 2023

List of Acronyms

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

BME	β-mercaptoethanol
CB	Competitive Binding
CCA	CleanCut Agarose
CSB	Cell Suspension Buffer
DBM	DNA Barcode Matchmaker
DNMTs	DNA methyltransferases
HCA	Human Chromosome Alignment
ODM	Optical DNA Mapping
OGM	Optical Genome Mapping
SV	Structural Variations
TBE	Tris-Borate-EDTA
WGD	Whole-Genome Duplication
YPD	Yeast extract Peptone Dextrose

Contents

List of Figures

List of Tables

1
Introduction

The genetic information of every known living organism on Earth is stored in their DNA, which is densely compacted into chromosomes and makes up the genome of the organism. Genomic structural variations (SVs) are events in which large regions of genetic material are duplicated, deleted, inserted or rearranged, and thereby affecting the structure of an organism's chromosomes [2],[3]. These structural variations are a natural source of genetic diversity which contribute to evolution since SVs result in variation of the structure of the chromosomes [4]. However, even though SVs are common in genomes and could lead to positive effects in some organisms, they are also associated with diseases [2], and are therefore important to detect.

Optical DNA mapping (ODM) is an approach that can be used for obtaining information about the genome and detecting SVs [2],[5]. ODM is based on stretching labelled DNA molecules with high molecular weight (HMW), typically DNA molecules larger than 100 kbp, and using fluorescence microscopy, sequence-dependent information of these long DNA molecules is visualised and imaged. This approach has in recent years been developed to detect SVs and is now able to detect structural variations which the traditional methods have difficulties detecting [2].

The budding yeast *Saccharomyces cerevisiae* (*S. cerevisiae*) is a widely used organism, which has a genome consisting of 16 chromosomes. The genome of *S. cerevisiae* has been completely sequenced, however, the genetic information comes from different strains. The sequenced genome is thereby an assembly of information from different strains which is problematic. Therefore, it is of high interest to apply the ODM technique on *S. cerevisiae* to perform *de novo* assembly of its genome. In the Department of Life Sciences at Chalmers University of Technology it is also of interest to use ODM to detect SVs in mutants of *S. cerevisiae*. However, a reference map of the genome of the yeast is needed to be generated using ODM to be able to investigate whether it is possible to detect SVs in *S. cerevisiae* using this technique.

1.1 Aim

The aim of the project is to perform reference-based assembly and explore first steps towards a *de novo* assembly of the genome of the yeast *Saccharomyces cerevisiae* using optical DNA mapping.

1.2 Limitations

The project is focused on and limited to the laboratory strain BY4742 of *Saccharomyces cerevisiae*, since the aim of the project is to explore first steps towards creating a reference map of the yeast *S. cerevisiae*.

Other limitations of the project are regarding the optical DNA mapping. ODM can be performed using different strategies for both labelling and stretching the DNA molecules. In this project, the ODM is limited to labelling the DNA molecules based on competitive binding (CB) between the two small molecules YOYO-1 and netropsin, and stretching of the DNA molecules is performed in nanochannels using a nanofluidic chip.

2
Theory

The current chapter of the report presents a theoretical framework for the project which includes theory regarding DNA, optical DNA mapping, structural variations, and the yeast *Saccharomyces cerevisiae*.

2.1 DNA

The DNA molecule is a fundamental component of all living organisms on our planet, including both prokaryotic and eukaryotic cells, as it serves as the carrier of genetic information of the organisms [6]. This genetic information defines the characteristics and traits of the organism and serves as the fundamental blueprint for the synthesis of all the proteins that the organism will produce throughout its lifetime [7]. Essentially, the existence of every known living organism on Earth relies on DNA. The complete set of DNA within an organism, which is found in every living being's cell, is defined as the genome [8].

2.1.1 Structure of DNA

DNA is an abbreviation of the name Deoxyribose Nucleic Acid which is derived from the structural characteristics of the molecule. The DNA molecule is built up by two complementary long polynucleotide chains, also called DNA strands, each of which is composed of four different types of nucleotide subunits that are linked covalently [7]. The structure of the DNA molecule can be seen in Figure 2.1 below. A nucleotide consists of a five-carbon sugar which is attached to one or several phosphate groups, as well as a nitrogen-containing base. In DNA, a single phosphate group is attached to the sugar which is deoxyribose, thereby the name "deoxyribonucleic acid". Deoxyribose has one hydroxyl group (-OH) attached to it compared to ribose, the sugar in RNA, which has an additional hydroxyl group attached to it, making RNA more reactive and thereby less stable.

It is the nitrogen-containing base, shortly called base, that differs between the nucleotide subunits. There are four different bases in DNA: adenine (A), cytosine (C), guanine (G), and thymine (T). These four bases form two purine-pyrimidine base pairs; adenine is paired with thymine (A-T) and cytosine is paired with guanine (C-G) [7]. A simple illustration of the structure of the bases and the base pairs can be seen in Figure 2.1. Adenine and guanine consists of two nitrogen-containing rings and are thereby purines, while thymine and cytosine are pyrimidines, consisting of

one nitrogen-containing ring. It is the linear sequence of the DNA strand, composed of the nucleotides containing the different bases, that encodes for the genetic information [6].

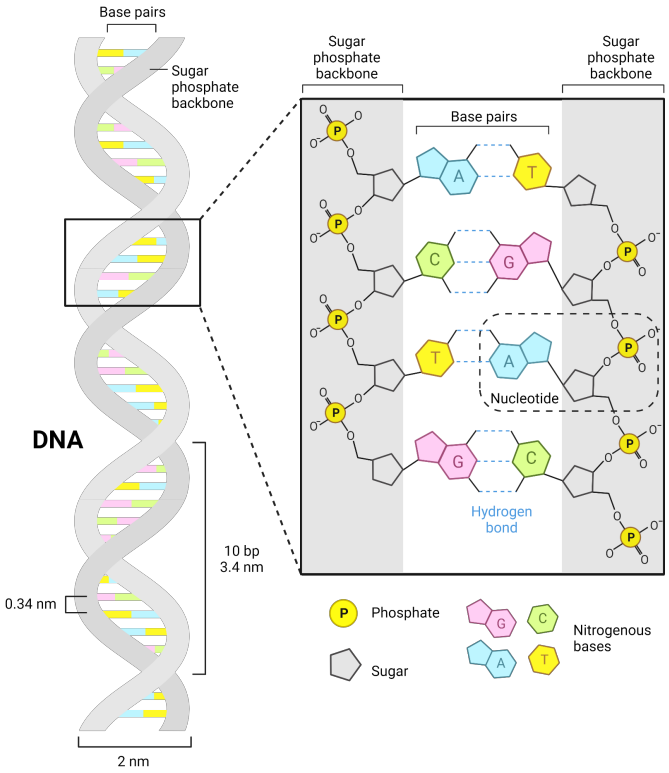

Figure 2.1: Structure of the DNA molecule, including simple structures of the four bases in DNA as well as the dimensions of the DNA molecule. Created in Biorender.com.

The two complementary DNA strands consist of a repetitive sugar-phosphate backbone, in which each sugar molecule is connected to the next one through a phosphate group [6]. The backbone of the DNA strands acts as a scaffold for the bases, from which they protrude from. The two DNA strands are held together by weak hydrogen bonds between the base pairs of the strands. Adenine is paired up with thymine by two hydrogen bonds in between them, while cytosine is paired up with guanine by three hydrogen bonds. The result is a molecule with a helical shape having the bases on the inside of the molecule and on the outside the backbone [7]. Due to the phosphate groups in the backbone of the DNA strands, the DNA molecule is negatively charged.

The helical structure is the most energetically favourable structure of the DNA molecule and is a result of the complementary base-pairing and the base-pair stacking. The base pairs in the double helix of the DNA molecule are arranged in the most energetically advantageous way, due to the complementary base-pairing. The complementary base-pairing ensures similar width of each base pair and thereby

equal spacing between the backbones along the DNA molecule is maintained [7]. In addition, the backbones of the DNA strands intertwine to maximise the efficiency of the stacking of the base pairs, which results in one complete turn of the double helix every ten base pairs.

The dimensions of the DNA molecule can also be seen in Figure 2.1. The DNA molecule has a width of 2 nm and the distance between each base pair, also called rise per base pair, is 0.34 nm [9]. Since each turn of the double helix consists of 10 base pairs, one complete turn has a length of 3.4 nm.

2.1.2 DNA in Eukaryotic cells

The genetic information that is carried by and stored in the DNA is located in every living being's cells. In eukaryotes, DNA is stored inside the membrane-bound nucleus of every cell [7], while in prokaryotes the DNA is not stored inside any nucleus. Instead the DNA is located in the nucleoid region in which the DNA is bundled together. Within eukaryotes, the genetic material is distributed between a distinct set of chromosomes, each consisting of one extremely long DNA molecule being densely packed with help of proteins [10]. Each single linear chromosomal DNA molecule is associated with proteins called histones, which aid in folding, twisting and packing the DNA molecule into a more condensed structure, as well as non-histone chromosomal proteins required for DNA replication, DNA repair and gene expression. The DNA molecules are condensed in an organised manner by the histones which generates a succession of coils and loops, thereby preventing the DNA from tangling and allowing it to be replicated, repaired and expressed. The different levels of structural order result in a tightly packed DNA-protein complex known as a chromatin. The structure of the chromosome including some of the different levels of compaction can be seen in Figure 2.2.

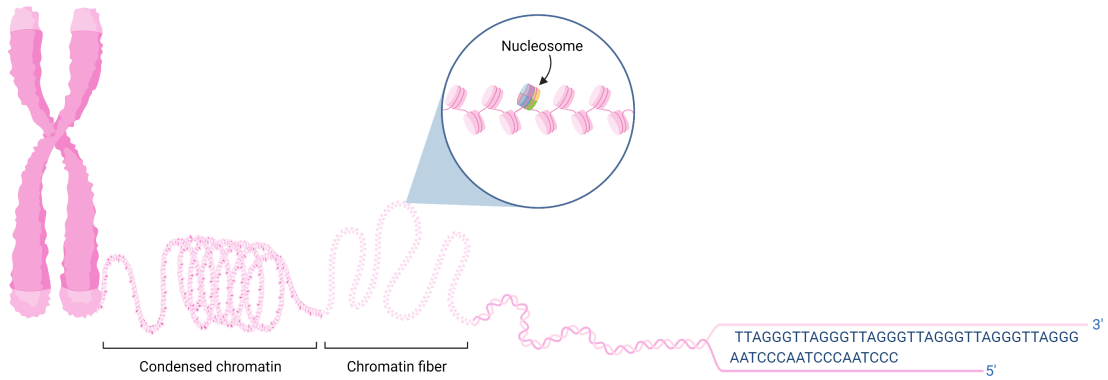

Figure 2.2: Structure of the chromosome and its different levels of compaction, including nucleosomes, chromatin fibre and condensed chromatin. Created in Biorender.com.

The histones are positively charged proteins which fold the DNA by binding tightly to the negatively charged DNA molecule, resulting in complexes called nucleosomes, which are the most fundamental level of chromosomal DNA packing [10],[11]. The nucleosome is composed of 8 histone proteins forming a core around which the DNA double helix is tightly wrapped 1.65 times. The repeating arrangement of nucleosomes and DNA linking the nucleosomes is in turn folded and condensed into a chromatin fibre, a fibre with a length of approximately one-third of the DNA molecule's initial length and a diameter of 30 nm. The chromatin fibre is condensed into a chromatin by forming loops, which is tightly coiled to form the final structure of the chromosome.

2.1.3 DNA Confinement

The long DNA molecules in the cells, with a typical length of several centimetres, are folded, twisted and thereby condensed to fit in the small cell which has a diameter of approximately 10 μm [12]. From this compact structure of DNA, it is challenging to extract information from the DNA molecules. One approach to obtain the information contained in the tightly packed DNA molecules is to confine the molecules within a confining environment, such as a nanochannel. When a DNA molecule is confined within a nanochannel, self-avoidance interactions cause the molecule to stretch out along the channel, and thereby analysis of the genomic content in a linearised form can be performed. Using nanochannels is an ideal approach for achieving high-throughput DNA linearisation. The extended conformation of the DNA molecules in the nanochannels is their equilibrium conformation, which is the result of the passive stretching mechanism in the nanochannels and thereby, no flow is needed to maintain the stretching [12].

The behaviour of the DNA molecule, both its size and shape, is dependent on the degree of confinement, which is determined by the average cross section (D) of the confining environment. Parameters used to describe the behaviour of a DNA molecule are the contour length (L) and persistence length (P) of the DNA molecule as well as the radius of gyration (R_G) of coiled DNA [1],[12]. The contour length of a DNA molecule refers to the length of the molecule when fully extended. It is the end-to-end distance of the molecule when no thermal fluctuations are present, and can thereby be calculated by multiplying the total number of base pairs of the DNA molecule with the distance separating them, which is 0.34 nm. An important parameter describing the DNA molecule's rigidity is the persistence length, which refers to the length over which the DNA molecule resists bending. A DNA molecule free in solution can be described by the radius of gyration, as the molecule in that case coils to minimise its free energy [1]. The radius of gyration thereby describes the radius of the coiled DNA, and is the value of the root mean square of the distance from the segments to the centre of the coil.

When forcing a DNA molecule inside a confining environment, such as a nanochannel which is the case for this project, the different parameters described above; P and R_G, as well as D, are used for determining the degree of confinement of the DNA molecule. The confining environment is described by the parameter D, which is the average cross section based on the width and height of the confining environment [1]. The extension of the nanoconfined DNA molecule is dependent on the nanochannel's cross section (D). Figure 2.3 shows three different nanochannels with different cross sections in which a DNA molecule is confined within, as well as the extension of the molecule which can be explained by different regimes [12].

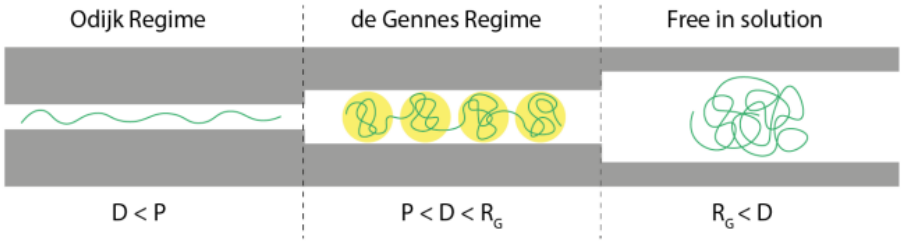

Figure 2.3: Illustration of a DNA molecule confined in three different nanochannels with different cross sections. Reused with permission from [1].

In the figure, three different scenarios of nanoconfined DNA molecules are visualised. If the radius of gyration of the DNA molecule is smaller than the nanochannels' cross section ($R_G < D$), the DNA molecule will behave similarly as it does unconfined, as when free in solution. In this scenario, the dimensions of the channels exceed the radius of gyration of the DNA molecule and no confinement of the molecule will be performed [1],[12]. If the case instead is that the cross section is smaller than the persistence length of the DNA molecule ($D < P$), the behaviour of DNA can be explained by the Odijk regime. In this state, the movement of the DNA molecule is restricted to small wave motions and the molecule is fully uncoiled. However, if the cross section of the nanochannel is much larger than the molecule's persistence length but smaller than the radius of gyration ($P \ll D < R_G$), de Gennes regime describes the behaviour of the DNA molecule well. In this case, the DNA molecule is weakly confined resulting in a DNA molecule which has several non-interacting blobs in a series, where each blob has a diameter similar to D, the cross-sectional area of the nanochannel. In the nanochannels used in this project, the persistent length and the cross section of the nanochannels are similar, more specifically $P \leq D$, resulting in a regime referred to as the extended de Gennes regime. The extended de Gennes regime is described as a transition regime between the Odijk regime and de Gennes regime. In all the different regimes described here, the degree of extension of the DNA molecule and its contour length is directly proportional. As a result, this means that sequence-specific information located at a position X% along the sequence of the DNA molecule can be observed on the nanoconfined DNA molecule at the corresponding position X% along the molecule.

The behaviour and stretching of the DNA molecules are also affected by the surrounding solvent and its ionic strength. Low salt concentration, thereby low concentration of ions in the solution, result in an increase in DNA stretching [12]. This is the case as a reduced amount of ions results in a decreased number of electrostatic interactions with the DNA molecule which shields the DNA from its negatively charged backbone. This in turn leads to an increased stiffness of the DNA molecule and consequently, a greater degree of extension. As the stiffness of the DNA molecule is affected by the concentration of ions in the solution, the salt concentration also affects the persistence length of the DNA molecules.

2.2 Optical DNA Mapping

As mentioned previously, the genome of an organism consists of DNA molecules containing its genetic information. It is therefore of high interest to be able to extract information from the DNA molecules to gain information about the organism and their characteristics and features. Today, there are several techniques available and used for extracting information from DNA molecules. Sequencing is one of such techniques which revolutionised our understanding of DNA, as very detailed information can be generated of specific genes. However, the techniques available today for sequencing, such as next generation sequencing (NGS), are limited as the generated reads are relatively short. Short-read sequencing generates reads of a length approximately between 150 and 300 base pairs (bp), while the reads generated by long-read sequences are on an average around 15 kilobase pairs (kbp) [13]. It is a challenge to localise these relatively short reads in the genome and therefore incomplete information about the genome is obtained. For example, repetitive and complex regions are challenging to map with both short-read and long-read sequencing, even though larger fragments are obtained with large-read sequencing. This in turn results in problems with detecting structural variations of the chromosomes, such as deletions, duplications and translocations of genes, that can result in different diseases.

Optical DNA mapping, abbreviated ODM, is another type of technique for obtaining information about the genome and its DNA molecules, in which information of larger fragments of the genome is obtained. ODM is a collection of methods that is used for visualisation of sequence-dependent patterns along elongated, individual DNA molecules, and thereby generating sequence information of single DNA molecules [14]. In general, ODM is performed by labelling the DNA molecules in a sequence-specific manner, followed by stretching and imaging of the DNA molecules using a fluorescence microscope [5]. The individual DNA molecules, visualised and analysed using ODM, are typically larger than 100 kbp and are said to be of high molecular weight (HMW).

2.2.1 Applications of ODM

There are several advantages with performing ODM. One major advantage is the length of the fragments of which information is obtained. Typically, DNA molecules of a size between 100 kilobases (kb) and 1 million bases (Mb) are analysed and visualised by ODM [14], however there is no upper limit. In other words, information of larger fragments of the genome can be obtained with ODM, which is not possible with the sequencing techniques today [5]. As information of long DNA molecule fragments is obtained with ODM, the generated data can be used for identifying large structural variations, as well as assisting in assembly of DNA sequences in complex genomes. ODM also has the ability to effectively cover genomic regions which is a challenge for DNA sequencing techniques, such as repetitive regions [13]. The data obtained from ODM can also be utilised for mapping DNA damage and epigenetic marks, for example DNA methylation, throughout the genome [5].

One of many applications of ODM is to generate reference optical maps of the genome of different organisms by performing *de novo* optical genome assemblies, also called ODM genome assembly. This is of big interest since structural variations can be detected using the generated reference optical maps. This technique, in which the whole genome of an organism is analysed, mapped or assembled, is also called optical genome mapping (OGM). In other words, optical DNA mapping (ODM) and optical genome mapping (OGM) is based on the same principle; visualisation of DNA molecules in a sequence-specific manner using a fluorescence microscope. ODM is any technique that is used for this visualisation, however, when ODM is applied on large eukaryotic genomes, the technique is also called OGM [2].

2.2.2 ODM Strategies

There are several strategies in performing ODM, in which both the stretching and labelling of the DNA molecules differs. DNA molecules can be stretched on either modified glass surfaces or in nanochannels, also called nanofluidic channels, using a nanofluidic chip [5],[14]. Using modified glass surfaces is the traditional way of stretching DNA molecules [15]. However, nanochannel-based ODM results in a more uniform stretching of the molecules as well as high throughput data collection, compared to stretching the molecules on glass surfaces [5],[14].

There are also different types of labelling of the DNA molecules to generate a sequence-specific pattern, which are divided into two main groups; sparse labelling and continuous labelling [16].

2.2.2.1 Sparse Labelling

Sparse labelling, also called enzymatic labelling, is based on enzymatic reactions which label specific sequences of the DNA molecules with fluorescent tags [16]. This type of labelling results in DNA molecules with fluorescent tags that are distributed throughout the length of the DNA molecules and generates a barcode-like pattern with the fluorescent tags being represented as markers that can be precisely localised.

ODM based on enzymatic labelling is the traditional version of ODM, and the most widely used enzymes are nicking enzymes which creates a single-stranded break, also called nick, at specific recognition sites with a length of several bp in the backbone of the DNA molecules [5],[16]. Followed by repair of the nicks by fluorescently labelled nucleotides together with a DNA polymerase and ligase, sequence-specific marks visualised as dots along the DNA molecules are generated. DNA stain in the form of YOYO-1 is also used in combination with the nicking enzymes to detect the contour of the DNA molecules in the microscope [5]. Enzymatic labelling can also be performed using methyltransferases, in which the DNA is labelled without being damaged, resulting in that more denser labelling can be performed. Compared to the labelling method using nicking enzymes, which is a two-step labelling, labelling based on methyltransferases is performed directly in one step by the use of DNA methyltransferases (DNMTs) catalysing methylation of the DNA bases with a synthetic methyl donor analog containing a fluorescent residue [16]. Recently, sparse labelling using Cas9 has also been developed.

2.2.2.2 Continuous Labelling

Continuous labelling, the other main group of labelling, produces a continuous map based on the fluorescence intensity along the DNA molecule and is the result of frequent enzymatic labelling or affinity-based binding [16]. Denaturation mapping is an additional approach for continuous labelling of DNA molecules. If enzymatic labelling is performed in frequently abundant or closely located recognition sites, the emission signals of the fluorescent labels overlap, thereby generating a continuously fluorescent pattern along the DNA molecules. However, a more cost-effective labelling method for continuous labelling is based on competitive binding (CB). In this approach, the labelling is performed using two different DNA binding molecules in combination [5],[16]. One of the molecules has specific affinity for either AT- or CG-rich regions and thereby binding preferentially to these regions, and the other molecule is an intercalating fluorescent molecule which binds nonsequence-specific to DNA. YOYO-1 is commonly used as the intercalating molecule in combination with netropsin which is AT-specific. Competitive binding occurs as the intercalating molecule is hindered to bind by competition at the specific sites, at which the molecule with specific affinity binds. The result is thereby continuous AT- or GC-density profiles.

When performing enzymatic labelling, extensive labelling protocols are required as multiple washing steps and removal of unbound fluorophores are needed [5],[14]. Affinity-based labelling methods, such as competitive binding, are thereby seen as a simpler approach for labelling of DNA molecules.

2.2.3 ODM based on Competitive Binding

In this project, optical DNA mapping based on competitive binding between the two small molecules YOYO-1 and netropsin is performed, as well as using a nanofluidic chip in which the DNA molecules are stretched. A schematic overview of the process of ODM used in this project can be seen in Figure 2.4.

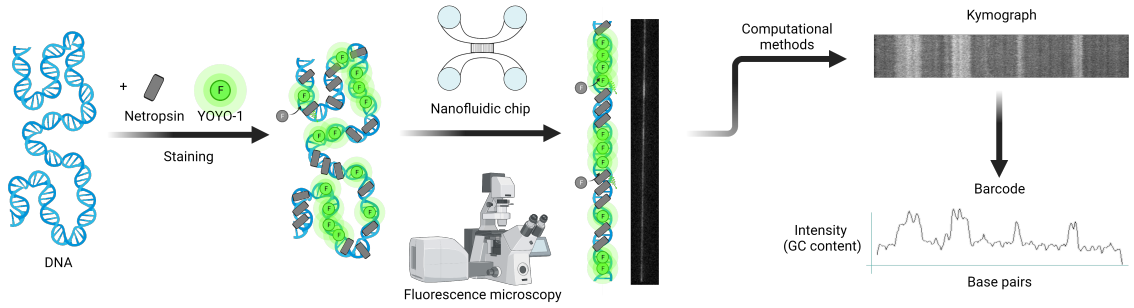

Figure 2.4: Schematic overview of ODM based on competitive binding between YOYO-1 and netropsin and using a nanofluidic chip for stretching of the DNA molecules. Created in Biorender.com.

As previously described, competitive binding is an approach for labelling DNA molecules based on DNA binding affinity of different molecules. The competitive binding is based on that netropsin binds selectively to AT-rich regions of DNA, thereby blocking these regions for YOYO-1, which binds non-specifically to DNA and becomes fluorescent upon binding [5]. The labelled DNA molecules are stretched in nanochannels in a nanofluidic chip and visualised using a fluorescence microscope. Based on the variation of AT/GC content along the DNA molecules, a sequence-specific pattern is thereby created, where the regions rich in AT-base pairs are darker compared to GC-rich regions. For each stretched DNA molecule fragment, a series of images is captured using a fluorescence microscope. A kymograph is then generated for each molecule using computational methods. To generate the kymograph, each time frame in the series of images taken of the DNA molecule is stacked onto each other. The features of the kymographs are thereafter aligned and a barcode is created for each molecule. These barcodes can be used for various applications, such as reference-based assembly or *de novo* assembly.

More detailed information about the two molecules, YOYO-1 and netropsin, used for competitive binding in this project are described below, as well as λ-DNA which is used as a size reference.

2.2.3.1 YOYO-1

YOYO-1 is a molecule commonly used as a fluorescent dye to stain double-stranded DNA [17]. The molecule is a homodimer consisting of two aromatic ring structures called YO (oxazole yellow) [18], which are linked together by a methine bridge [1]. YOYO-1 is a bis-intercalator which binds nonsequence-specific to DNA by inserting

its two ring structures (YO) into the DNA molecule in-between the base pairs. The chemical structure of YOYO-1, as well as an illustration of how it binds to DNA can be seen in Figure 2.5. Also netropsin and its binding to DNA is illustrated in the figure.

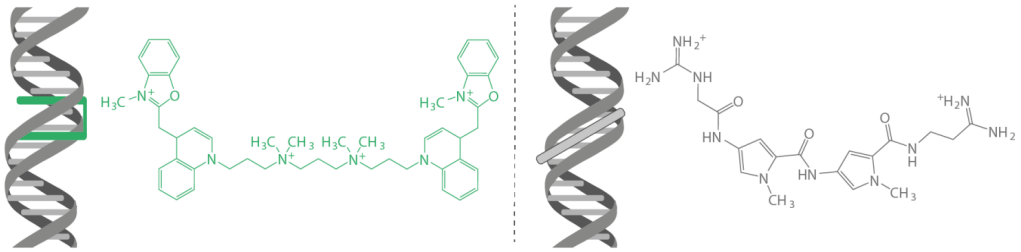

Figure 2.5: Chemical structure of YOYO-1 and netropsin and an illustration of their binding to double-stranded DNA. YOYO-1 is seen in green to the left, and netropsin in grey to the right. Reused with permission from [1].

YOYO-1 is classified as a cyanine dye and belongs to the dimeric-cyanine family which is the most popular family of DNA stains [12]. This is due to the uniformity of the staining as well as the great emission enhancement of the dyes upon binding to DNA. Unbound YOYO-1 is almost non-fluorescent, however, when YOYO-1 is bound to a double-stranded DNA molecule, there is a great fluorescence enhancement of the molecule resulting in a high signal-to-noise ratio [17]. The large enhancement in emission is due to the fact that the rotation around the methine bridge of the YOYO-1 molecule becomes sterically hindered, which results in an increased emission of more than a thousand-fold [1]. When bound to DNA, YOYO-1 has an excitation and emission maximum at 491 nm respectively 509 nm [1],[5].

YOYO-1 has a high binding constant to double-stranded DNA and thereby binds strongly to DNA [17], but there is a limited amount of YOYO-1 molecules that can bind to DNA. When full saturation occurs, there is approximately one YOYO-1 molecule bound per every 4 base pairs of the DNA molecule (1:4 YOYO-1:bp ratio) [1],[19]. When this occurs, the contour length of the DNA molecule is increased with approximately 38% [19].

It is important to note that when the fluorescent dye YOYO-1 binds to DNA, some structural and mechanical properties of the DNA molecule are changed significantly [17]. Both the structure and rigidity of the DNA molecule are affected upon binding to YOYO-1, thereby the persistence length and contour length are affected. More specifically, for every intercalating YOYO-1 molecule, the contour length of the DNA molecule will increase with approximately 0.51 nm. Binding of YOYO-1 to DNA also affects the helical twist of the DNA molecule and results in unwinding of the double helix with approximately 24° for every bound YOYO-1 molecule [1],[17]. Additionally, the YOYO-1 molecule is tetracationic, which can be seen in Figure 2.5, and will therefore result in a reduced negative charge of the DNA molecule when bound to it [1].

2.2.3.2 Netropsin

Netropsin, on the other hand, is a non-fluorescent molecule. It is an oligopeptide antibiotic which binds selectively to AT-rich regions of double-stranded DNA molecules [16],[20]. More specifically, netropsin binds to AT-base pairs located in the minor groove of the DNA helix [1]. Netropsin binds to AT-base pairs by creating hydrogen bonds with the bases A and T. However, the binding of netropsin to GC-base pairs is hindered due to the amino group in the guanine base, resulting in that netropsin is AT-specific. Additionally, due to the deeper structure as well as the higher proportion of potential hydrogen bonding acceptor sites in the minor groove, binding to the minor groove is preferred over the major groove [20]. For netropsin to bind in the minor groove of the DNA molecule, a minimum of three base pairs is necessary. The chemical structure of netropsin as well as an illustration of how it binds to DNA can be seen in Figure 2.5.

2.2.3.3 λ-DNA as Size Reference

When performing optical DNA mapping, an important aspect is to include a size reference used to determine the stretching of the extended DNA molecules. A commonly used standard for experiments using DNA is λ-DNA. λ-DNA, also called λ-phage DNA, is DNA from a bacteriophage from *E. coli* and is a widely used DNA for different studies [21]. It has a known amount of base pairs, which is 48 502 bp often written as 48.5 kbp, and its contour length is 16.5 μm, which is the length when fully extended [12].

2.3 Structural Variations

Structural variations (SVs), also referred to as genomic structural variations or rearrangements, are events that affect the structure of the chromosomes of an organism and include duplications or multiplications, deletions, rearrangements, inversions, and insertions of large regions of genetic material [2],[3]. Deletions, insertions and duplications are also classified as copy number variants (CNVs), which are defined as a large DNA segment, larger than 1 kb, that compared to the reference genome exist at a varying copy number [3]. Structural variations can span millions of nucleotides and are divided into two classes. SVs with a size between approximately 1 kb and 3 Mb are referred to as submicroscopic SVs, while SVs larger than approximately 3 Mb are microscopic [3].

Structural variations can arise from various mutational mechanisms [22]. Mechanisms such as DNA replication, DNA repair and DNA recombination can lead to arrangement of the genetic material and thereby alterations in the structure of the chromosomes, resulting in formation of SVs. These events are common and contribute to evolution as they are a natural source of genetic diversity and result in variation of the chromosomes [4]. Some structural variations are however known to cause disease and genetic disorders, and are associated with disadvantageous phenotypes [2],[22]. The phenotype may be affected due to several mechanisms which

lead to abnormality regarding the gene dosage, gene disruption and gene fusion, and other related mechanisms [22]. When a structural variation is associated with a disease, or believed to be causing a disease, the term structural abnormality is used instead [3].

A variety of diseases, ranging from cancer to rare genetic disorders as well as other related conditions, are associated with structural variations [16]. One example is the cancer disease acute lymphoblastic leukaemia (ALL) [23]. In leukemic cells, there are a multitude of structural variations present which have been observed to be closely linked to drug resistance, as well as the initiation and progression of the disease. Thereby, characterisation of which SVs that are present in the leukemic cells of patients is important to know as they play an essential role in determining the level of risk as well as predicting treatment outcomes of the patients.

2.3.1 ODM for Detecting SVs

The ODM technique applied on genomes of different organisms, optical genome mapping (OGM), has in recent years been developed to detect structural variations [2]. This technique has overcome the challenges and limitations associated with techniques currently used for SV detection, and is today able to detect SVs that are challenging to identify using conventional methods.

Recent advancements in SV detection using ODM have also provided significant benefits in the identification and understanding of genetic disorders. Two genetic disorders that have benefited from the detection of SVs using ODM are Poretti-Boltshauser syndrome and Duchenne muscular dystrophy [16].

The technique of using ODM for detection of SVs is thought to have the capability to replace the current classical methods used, such as karyotyping, FISH (fluorescence *in situ* hybridization), and CNV (copy number variant) microarrays [24]. These techniques have different resolutions and are today used in combination for SV detection in cytogenetic diagnostics.

2.4 *Saccharomyces cerevisiae*

The budding yeast *Saccharomyces cerevisiae* (*S. cerevisiae*), also known as brewer's yeast or baker's yeast, is an unicellular eukaryotic fungus that reproduces asexually through a process called budding, thereby the name "budding yeast" [25],[26]. During the budding process, smaller daughter cells emerge by budding, or pinching, off from the mother cell after nuclear division and passing on a nucleus to the daughter cell [26]. This characteristic cell division, which can occur as frequently as once every 90 minutes, distinguishes *S. cerevisiae* from other yeast strains, such as the fission yeast *Schizosaccharomyces pombe*, which undergoes fission-based cell division.

S. cerevisiae is a widely used model organism, especially used for investigating essential aspects of eukaryotic cell biology, and it is known as the most suitable eukaryotic cell type for conducting biological studies. As *S. cerevisiae* is an eukaryotic organism, it contains a cell nucleus, mitochondria, endoplasmic reticulum, cytoskeleton, and several other organelles that are present in eukaryotic cells [25],[26]. Additionally, several biological pathways are highly conserved in yeast that are associated with ageing and disease in humans, such as DNA repair mechanisms, stress response, protein folding, nutrient signalling, cell cycle regulation and regulated cell death [25].

To use *S. cerevisiae* as a model organism has many advantages. Not only that many pathways are conserved between the yeast and humans. The genome of *S. cerevisiae* is relatively simple and there is a high degree of homology between yeast and mammalian genomes [25], allowing researchers to gain more insight and understand the basis of diverse human diseases and disorders. In other words, findings from studying conserved pathways in the yeast cell can be applied both to humans and higher organisms. One major advantage of using yeast as a model organism is also the available yeast genetics which has an incredible power [25]. The yeast genome is easy to genetically manipulate, in the sense of introducing specific mutations, deletions and/or insertions into the genome of the yeast. This allows investigations of specific genes, their functions, gene interactions and the resulting phenotypic effects [26]. The manipulation can be performed by either introducing plasmids into the yeast cell with the desired gene or by directly manipulating the chromosomes.

Additionally, the simplicity and low cost to work with *S. cerevisiae*, as well as high reproducibility and rapid growth of the cells [25], have made the yeast an attractive model organism for many researchers. The rapid growth rate of the yeast enables fast experiments and thereby high-throughput screenings.

2.4.1 Culture of *S. cerevisiae*

As mentioned previously, *S. cerevisiae* is easy to grow and work with in the laboratory due to their simple growth conditions [26]. The optimal growth temperature for the yeast *S. cerevisiae* is 30°C, but the acceptable span of temperatures for growth is between 20 and 35°C [27]. An optimal media used for culturing and maintaining cells of *S. cerevisiae* is YPD media. YPD, which is an abbreviation for Yeast extract Peptone Dextrose, is a liquid growth media used for both propagation and maintenance of yeast strains of *S. cerevisiae*. One litre of YPD media contains 10 g yeast extract, 20 g bacto peptone and 20 g dextrose, same as glucose.

The growth of yeast can be described by a growth curve consisting of a lag phase, an exponential phase (also called the log phase), and a stationary phase [28], seen in Figure 2.6. The graph shows the number of cells in the population, in a logarithmic scale, against time. When the cells are set to grow, they first need to adapt to the new environment, and this occurs during the lag phase. When the cells are in this phase, intracellular adaptations of the cells occur and only a minimal cell replication

is performed, if the cells are replicated at all. The next phase is the exponential phase, or log phase, in which reproduction of the cells occur at a rate directly proportional to the existing cell population, resulting in an exponential growth of the cells. In this phase there is no nutrient limitation. When the availability of nutrients becomes limited for the growing cells, or when the cell population becomes restricted due to other environmental conditions, the growth of the cells goes into the stationary phase, at which a plateau is reached regarding the number of cells. A decline in the overall number of cells can thereafter occur due to cellular death if the surrounding conditions of the yeast cells become insufficient to sustain the yeast population. This phase is referred to as the death phase of the growth curve, which is not shown in the figure.

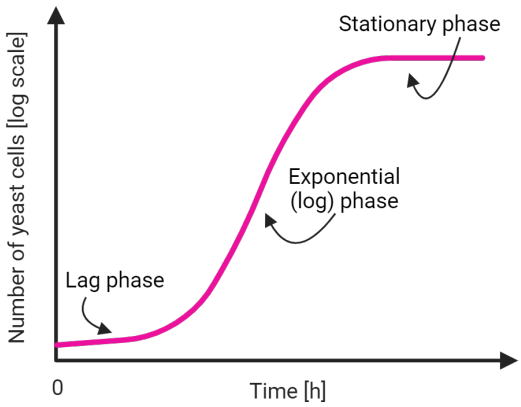

Figure 2.6: Typical growth curve of yeast cells consisting of three different phases; lag phase, exponential (or log) phase, and stationary phase. The y-axis represents the number of yeast cells in the population in a logarithmic scale, while the x-axis represents the time of culturing. Created in Biorender.com.

2.4.2 The Genome of *S. cerevisiae*

A significant milestone was achieved by scientists in 1996, when the genome of *S. cerevisiae* was completely sequenced, making it the first eukaryotic genome to be fully sequenced [25],[26],[29]. The genome of *S. cerevisiae* consists of approximately 12 200 kbp which are distributed between 16 chromosomes. The size of the chromosomes span between approximately 200 kbp to 1.5 Mbp, according to the *Saccharomyces* Genome Database (SGD) [29]. The genome includes approximately 5 800 protein-coding genes out of a total of 6 275 identified genes, and it is estimated that at least 31% of the identified genes in *S. cerevisiae* exhibit homology with genes found in humans [25]. In addition, there is also mitochondrial DNA present in the yeast cells of *S. cerevisiae*, however the size of the mitochondrial DNA is different for different strains [27].

The yeast *S. cerevisiae* can exist both with a haploid and diploid set-up of genes. The haploid yeast cell contains one set-up of the 16 chromosomes, which are thought to have originated from a set of 8 chromosomes through a whole-genome duplication (WGD) event back in the history of *S. cerevisiae* [26]. This WGD has also resulted in repetitive sequences being present throughout its genome [30].

The sequenced genome of *S. cerevisiae* released in 1996, was accomplished due to collaborative efforts of researches and laboratories around the world [29]. More precisely, 94 laboratories located in 19 countries were involved in the sequencing of the genome, which used different sequencing technologies as well as methods. This first released sequenced genome of *S. cerevisiae* is based on the *S. cerevisiae* strain S288C, which is a commonly used laboratory strain due to its minimal nutrient requirements as well as non-flocculent behaviour. However, several different strains were used for the sequencing which are said to be isogenic, genetically identical, with the S288C strain. Since the sequenced genome was published, a large number of corrections have been made to the sequence which now are part of the reference genome. With time, the number of corrections in the reference genome has decreased as there has been a transition towards a more accurate reference sequence. However, there are still changes and corrections to be made.

The reference genome of *S. cerevisiae* has been of great use and still is today. Today, there is a possibility to sequence the whole genome of an organism within a few days, due to the advancements in the sequencing techniques and as they have become much cheaper [29]. The assembly of the sequenced genome is performed by comparison to the valuable guide that the reference genome is.

2.4.3 Strain BY4742

The strain used in this project is the *S. cerevisiae* strain BY4742. This strain is often used in laboratories for research and is a derivative of the S288C laboratory strain. More specifically, BY4742 is a deletion strain which was obtained from S288C and has the genotype MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0, indicating its mating type as well as which genes that have been deleted or disrupted in the strain.

According to the *Saccharomyces* Genome Database (SGD), there is a negligible variation between the strains BY4742 and S288C. Due to this, as well as strain BY4742 being widely used and easy to obtain, as it is used in the laboratories at Chalmers University of Technology, it is a good idea to use the strain BY4742 for this project.

3
Methods & Materials

The following chapter is describing the methodology used to fulfil the aim of the project, as well as providing the materials that have been used.

The methodology of the project is divided into several parts starting with culturing of the yeast *Saccharomyces cerevisiae*, followed by extraction of the DNA from the yeast cells, verification of the extracted DNA using pulsed-field gel electrophoresis (PFGE), and ending with optical DNA mapping in which data collection using a nanofluidic chip and a fluorescence microscope as well as data processing and analysis were performed. The procedure of the different parts are explained more in detail in the following sections.

3.1 *S. cerevisiae* Culturing

The project was initiated by culturing the obtained yeast strain of *S. cerevisiae*, strain BY4742. The procedure of the culturing is illustrated in Figure 3.1, which started with inoculation. The inoculation was performed by transferring a colony of the yeast from an agar plate with the obtained strain to a sterile 50 ml falcon tube containing 10 ml YPD media. Two replicates of the inoculation were prepared to ensure an adequate quantity of culture for the following step, as well as a control containing purely YPD media to confirm sterile handling. After the inoculation, the tubes were left in a shaker in 30°C overnight (12-16 hours) to let the cells grow and propagate.

The day after, a small amount of the culture that grew overnight was transferred to a 500 ml sterile flask containing 100 ml YPD media, to reach an OD_{600} of 0.1. This is in line with the CHEF Yeast Genomic DNA Plug Kit protocol from BIO-RAD which was used for the DNA extraction. To ensure an OD_{600} of 0.1, 2 ml of culture was initially transferred to the flask with YPD media and the OD_{600} was measured. Based on the value, the total amount of culture needed to reach a value of 0.1 was determined. The final amount of culture transferred from the falcon tubes to the flask was 3.333 ml. The flask with the cells was left in a shaker to grow with agitation in 30°C for approximately 6 to 8 hours until an OD_{600} above 1 was reached.

When an OD_{600} of 1 or more was reached, the cell concentration was determined based on the final value of OD_{600} and on empirical data from the group the cells were obtained from. In turn, based on the cell concentration, the culture was transferred

to new 50 ml tubes each containing an amount of culture corresponding to 6 $\cdot 10^8$ number of cells, the number of cells desired for creating 1 ml of plugs according to the CHEF Yeast Genomic Plug Kit. To obtain the cells, the tubes were centrifuged at 5 000 $\times$ g for 10 minutes in 4°C and the supernatants were thrown away. The tubes with the harvested cells were frozen in -20°C to be stored until the creation of plugs was performed.

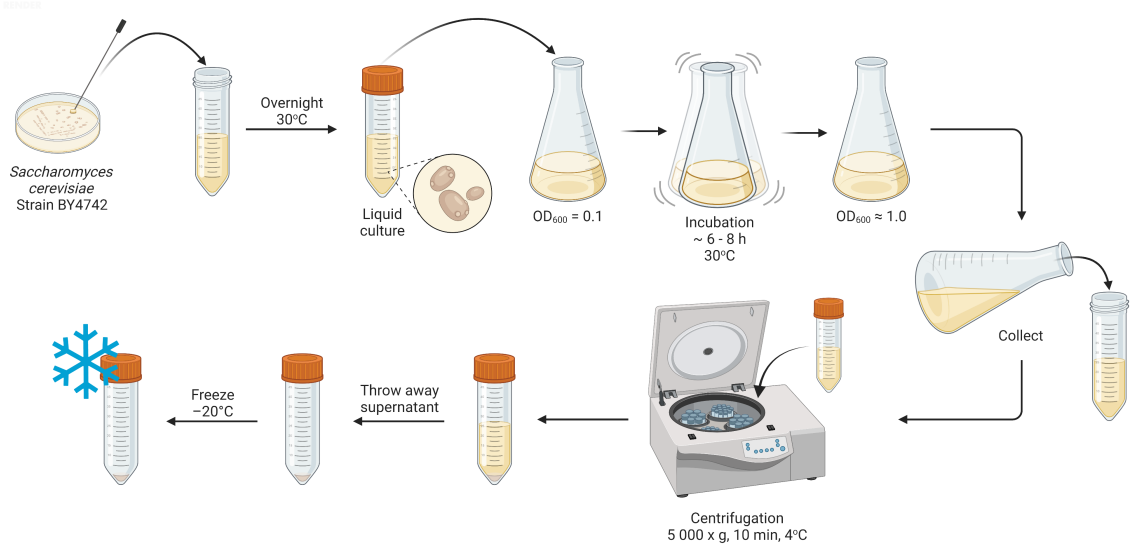

Figure 3.1: Schematic overview of culture and harvest of the yeast *S. cerevisiae* strain BY4742. Created in Biorender.com.

Throughout the process of the yeast cultivation, including all tasks leading up to the collection of the yeast cells, a fume hood equipped with UV light was utilised to ensure sterility.

3.2 DNA Extraction

The extraction of DNA from the yeast cells was performed using the CHEF Yeast Genomic DNA Plug Kit from BIO-RAD and the included protocol. The kit resulted in gel plugs containing the DNA from the yeast cells, from which the DNA could be extracted and used. Details on the procedure are explained in the following paragraphs. In the explanation, all buffers, solutions and chemicals used were generated by BIO-RAD, the kit manufacturer, and the volumes mentioned are the volume required for 1 ml of plugs, which was created.

3.2.1 Creation of Gel Plugs

The DNA extraction of the cultured yeast cells was initiated by thawing the frozen cells. The number of cells required for creating 1 ml of plugs was 6 $\cdot 10^8$, thereby one of the frozen tubes including cells was thawed. The thawed cells were transferred to a 1.5 ml eppendorf tube and additionally one centrifugation was performed at 5 000

× g for 10 minutes in 4°C to get rid of remaining YPD media and the supernatant was thrown away.

Calculation of the amount of Cell Suspension Buffer (CSB) and 2% CleanCut agarose solution (CCA) needed were performed. For 1 ml of plugs, 625 μl CSB and 375 μl CCA, from the CHEF Yeast Genomic Plug Kit, were needed to obtain a final agarose concentration of 0.75%. The 2% CleanCut agarose solution was heated in a microwave 10 seconds at a time until the solution was clear and liquid and thereafter equilibrated in a heating plate to 50°C. The thawed cells were resuspended in the calculated amount of CSB and equilibrated in the heating plate to 50°C. Following, 30 μl of Lyticase stock was added for every ml of plugs to be created to degrade the cell wall of the yeast cells, and immediately after, the agarose (CCA) was added to the cells and gently and thoroughly mixing was performed. Once again, the mixture of the cells and agarose was kept at 50°C in the heating plate. Thereafter, the mixture was transferred to the plug moulds, each with a volume of 100 μl, using a pipette. The plug moulds including the mixture were placed at 4°C for approximately 10 to 20 minutes to let the agarose solidify and add strength to the plugs. The solidified plugs were carefully pushed out of the moulds into conical eppendorf tubes with a volume of 1.5 ml, resulting in each tube containing 3 to 4 plugs. Lyticase solution was prepared by mixing 2.5 ml of Lyticase Buffer with 85 μl of Lyticase stock for every ml of plugs, and was added in equal volume to the tubes. The plugs were incubated in the Lyticase solution at 37°C for 2 hours. After the incubation, the Lyticase solution was removed and the plugs were rinsed with sterile water (Milli-Q (MQ) water). Thereafter, 2.5 ml of Protein K Reaction Buffer followed by 100 μl of Proteinase K stock, for every ml of plugs, were added in equal volumes to the eppendorf tubes containing the plugs to degrade all proteins present and the plugs were incubated overnight at 50°C.

The next day, 1× Wash Buffer was prepared by diluting the 10× Wash Buffer stock from the kit with MQ water 10 times (1:10), and the plugs were washed in the 1× Wash Buffer four times for 1 hour with gentle agitation in room temperature to remove cell debris and proteins. After this step, the plugs were completed and were stored at 4°C in the 1× Wash Buffer.

3.2.2 DNA Extraction from Gel Plug

To be able to visualise and image the DNA molecules, the DNA from the yeast required to be extracted from a gel plug, which was performed by digesting the plug. DNA from one plug was extracted at each time. This was performed by transferring one plug, from the buffer it was stored in, into a new conical eppendorf tube of 1.5 ml. Thereafter, 1× CutSmart solution was prepared by diluting 10× CutSmart solution (New England Biolabs) with MQ water by a factor of 10, and the plugs were washed twice in the 1× CutSmart solution for 15 minutes at room temperature. Following, 78 μl MQ water and 20 μl 10× CutSmart solution were added to the plug and incubation of the plug at 70°C for 10 minutes followed by incubation at 42°C for 10 minutes was performed. Afterwards, 2 μl of agarase (New

England Biolabs) was added to degrade the agarose of the plug and the mixture was gently mixed by pipetting once slowly. Finally, the mixture was incubated at 42°C for 2 hours.

3.2.3 Qubit Fluorometer for DNA Quantification

The DNA concentration of the mixture containing the digested gel plug with the extracted DNA was measured using a Qubit fluorometer and the Qubit dsDNA BR (Broad range) Assay Kit from ThermoFisher.

The Qubit dsDNA BR Assay Kit includes Qubit reagent, Qubit Buffer and two Standards. Qubit Working Solution was prepared by mixing 1 μl Qubit reagent with 199 μl Qubit Buffer for every assay tube that was going to be prepared. In total, 5 assay tubes were prepared for the measuring of the DNA concentration of one plug, 2 assay tubes with the standards and 3 assay tubes with replicates of the sample. The standard assay tubes were prepared by mixing 190 μl of Qubit Working Solution with 10 μl of one of the standards, respectively, resulting in one assay tube containing standard 1 and one containing standard 2. The sample assay tubes were in turn prepared by mixing 195 μl of Qubit Working Solution with 5 μl of the mixture containing the extracted DNA. Three replicates were done of the sample assay tube, with the difference that the extracted DNA was obtained from the top, middle and bottom, of the tube with the digested plug. All 5 assay tubes were vortexed for 2 to 3 seconds before inserted in the Qubit fluorometer, starting with the two standards followed by the three sample assay tubes.

Three results were generated from the Qubit fluorometer, one DNA concentration given in ng/μl for each sample assay tube, from which the average value was calculated.

3.3 Pulsed-Field Gel Electrophoresis

For verification of the DNA, pulsed-field gel electrophoresis (PFGE) was performed. The gel used for the PFGE contained 1% agarose and was prepared by mixing a weight of agarose corresponding to 1% with 0.5× TBE Buffer. The assumption that 1 ml corresponds to 1 g was made when calculation of the amount of agarose was performed. The mixture was heated using a microwave until a clear liquid was obtained, and was cooled down to approximately 55°C. The liquid was poured into the gel caster belonging to the PFGE machine, including a comb for creating the wells, and the gel was let to solidify. The comb was removed and lambda PFG ladder (New England Biolabs) and sample, in form of the gel plugs containing the DNA, was loaded into the wells and the wells were sealed with the agarose solution used for creating the gel.

The PFGE was performed using the CHEF Mapper XA System from BIO-RAD. Before running the PFGE, the tank was filled with 0.5× TBE Buffer, the gel was placed in the tank of the machine, the pump was started and the belonging chiller unit was set to 14°C. Three different attempts of PFGE were performed and the programs for the different attempts of PFGE can be seen in Table 3.1.

Table 3.1: Settings used for the programs of the three different attempts of PFGE.

Attempt	Run time [h]	Switch time	Voltage [V/cm]	Resolution
1	40	36.63 - 126.67 s	Default	Default
2	24	60 - 120 s	6	200 kb - 1.5 Mb
3	24	60 - 120 s	6	200 kb - 1.5 Mb

After the PFGE was run, staining of the gel was performed to be able to visualise the DNA molecules in the gel. SYBR Safe DNA stain (ThermoFisher) was used as staining agent for the PFGE gel and the gel was stained in the staining buffer for 30 to 60 minutes. Thereafter, the gel was destained with distilled water 3 times for approximately 5 min before imaging was performed of the gel.

3.4 Optical DNA Mapping

The final part of the project is the optical DNA mapping, which in this project is based on continuous labelling of the DNA molecules based on competitive binding, and stretching of the DNA molecules using a nanofluidic chip. A schematic overview of the process of ODM used in this project is seen in Figure 3.2, which also was shown in the theory section.

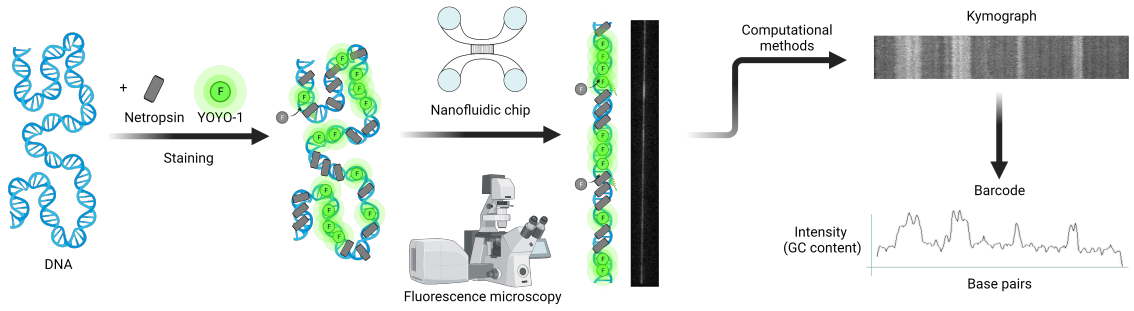

Figure 3.2: Schematic overview of the ODM performed in this project which is based on competitive binding between YOYO-1 and netropsin and using a nanofluidic chip for stretching of the DNA molecules. Created in Biorender.com.

Details on how ODM was performed in this project are described in the following sections, which are divided into two main parts; data collection using fluorescence microscopy as well as data processing & analysis.

3.4.1 Data Collection using Fluorescence Microscopy

Following that the DNA was extracted from a gel plug and the DNA concentration was measured, data collection using a nanofluidic chip and a fluorescence microscope was conducted. This part of the method includes staining the extracted DNA simultaneously with YOYO-1 and netropsin, stretching the DNA in nanochannels using a nanofluidic chip, and taking images of the stretched DNA with a fluorescence microscope. Details on this part of the method are described below.

3.4.1.1 Staining of DNA Molecules

To be able to visualise, capture images, as well as generating sequence-specific barcodes of the extracted DNA molecules from the yeast cells, staining of the DNA molecules was needed to be implemented, which in this project was based on competitive binding. The molecules used for staining and generating sequence-specific barcodes were YOYO-1 and netropsin.

The process of staining was initiated by preparing the reaction mix required for the staining. The reaction mix consisted of YOYO-1 (ThermoFisher), netropsin (ThermoFisher), λ-DNA (ThermoFisher), the DNA extracted from the gel plug and 0.5× TBE buffer. λ-DNA, which has a known size of 48.5 kbp, was included in the reaction mix as an internal size reference [5],[14]. Three different reaction mixes were tested and used which can be seen in Table 3.2, in which details on the staining reaction mixes are provided, including the components and their final concentrations in the reaction mix. Also the total amount of DNA and the proportion between extracted DNA and λ-DNA given in percentage in the staining reaction are shown in the table.

Table 3.2: Staining Reactions for Competitive Binding

	Final conc. $[\mu M]$		
Component	First staining	Second staining	Third staining
YOYO-1	0.6	0.3	0.6
Netropsin	180	90	180
λ-DNA	0.6	0.9	1.8
Extracted DNA	5.4	2.1	4.2
0.5× TBE	-	-	-
Total DNA	6	3	6
% Ex. DNA/%λ-DNA	90/10	70/30	70/30

The reaction mix resulted in a molar ratio of 10:1:300 (DNA base pair: YOYO-1: Netropsin) between the total amount of base pairs in the reaction, YOYO-1 and netropsin, which is the standard ratio used for CB-based ODM in Fredrik Westerlund research group at Chalmers University of Technology. In other words, for every 10 base pairs of DNA in the reaction mix, one YOYO-1 molecule was present as well as

300 molecules of netropsin. The amount of extracted DNA in the reaction mix was based on the concentration measurements using the Qubit fluorometer. An average of the concentrations of the three sample assay tubes were used and conversion from ng/μl to μM was performed, based on the average molecular weight of a DNA base pair (bp) which is 650 g/bp mol.

The differences between the staining reactions are the total amount of DNA, as well as the proportion of extracted DNA and λ-DNA. For the first staining reaction the total amount of DNA was 6 μM in which 90% was extracted DNA and 10% was λ-DNA. In contrast, the second and third staining reaction had a total amount of DNA of 3 and 6 μM, respectively, in which 70% was extracted DNA and 30% was λ-DNA.

Following are some important notes regarding the preparation of the staining reaction mix. YOYO-1 is light sensitive and should thereby be kept away from light. Also the reaction mix, after YOYO-1 had been added, should for the same reason be kept away from light. Additionally, when pipetting DNA, extra caution is needed to minimise breakage as the DNA molecules are fragile. Therefore, a cut pipette tip was used at all times when handling the DNA molecules to decrease the stress on the molecules and slowly pipetting was performed.

The reaction mix was heated at 50°C for 30 minutes and the mixture, which had a total volume of 10 μl, was thereafter diluted in 88 μl of MQ water and at the end 2 μl of BME (β-mercaptoethanol) was added to prevent photodamage to the sample. This resulted in a dilution by a factor of 10 of the mixture and thereby a final buffer concentration of 0.05× TBE. If the λ-DNA molecules were perceived to be too short when looking at the sample in the microscope, additionally 49 μl MQ water and 1 μl BME were added to the reaction mix. 1 or 2 μl of BME could also be added if the DNA molecules were perceived to break a lot, which complicates the process of imaging the DNA molecules.

3.4.1.2 Imaging using Fluorescence Microscope

When the reaction mix for the staining was accomplished, imaging of the DNA molecules were performed using a nanofluidic chip and a fluorescence microscope. The fluorescence microscope used to capture images of the DNA molecules was Zeiss Observer D.1, which was equipped with a 63× (1× optovar) oil immersion objective (NA = 1.46, Zeiss), a Colibri 7 light source (Zeiss) and a Photometrics Evolve EMCCD camera. A total of 30 frames were captured for each DNA molecule with the microscope using the 63× oil immersion objective, in which each time frame had an exposure time of 100 ms. The pixel size of the microscope was 254 nm/px, resulting in that the relation between one pixel of the microscope and one kbp being approximately 1:1. However, this depends on the stretching of the molecules.

Imaging of single stretched DNA molecules was accomplished using a nanofluidic chip, illustrated in Figure 3.3. The nanofluidic chip, which is fabricated using silica, consist of 4 loading wells that are interconnected by microchannels and nanochannels, also called nanofluidic channels. The loading wells are divided into two pairs, with a microchannel interconnecting each pair. The two microchannels, with a dimension of 50 nm $\times$ 850 μm (width $\times$ height), are in turn connected by 120 nanochannels which are 500 μm long and have a cross section of 150 $\times$ 100 nm^2 (width $\times$ height). The dimensions of the chip can also be seen in the figure below.

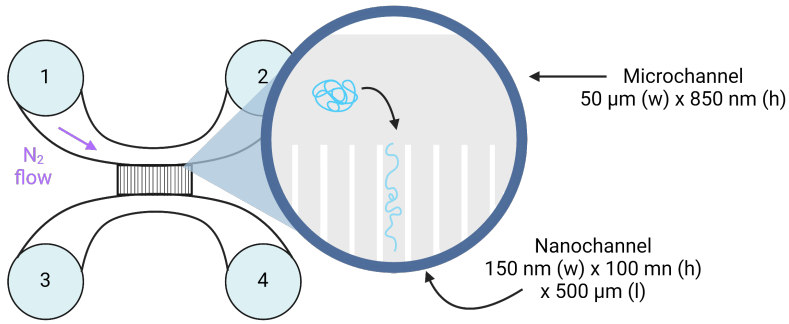

Figure 3.3: Illustration of the nanofluidic chip used for the imaging of DNA molecules, including the dimensions of the microchannels and nanochannels. The zoom-in on the chip visualises a DNA molecule located in the microchannel as well as confined in one of the nanochannels. Created in Biorender.com.

Before imaging, the nanofluidic chip was washed with chlorine to clean the chip, followed by one or two washes with MQ water, and finally one wash with 0.05$\times$ TBE buffer containing 2% BME to pre-wet the chip to ensure uniform conditions before loading the samples. Each wash consisted of loading each well with 10 μl of the washing buffer, connecting the chip to a pressure driven nitrogen (N_2) flow and pushing the fluid for 1 to 2 minutes in different directions. The fluid was first pushed through the microchannels, thereafter through the nanochannels, and additionally one more time through the microchannels in the same direction. After the chip had been washed and pre-wetted with the 0.05$\times$ TBE buffer with 2% BME, the chip was loaded. 10 μl of the prepared staining reaction mix (DNA sample) was loaded in one of the loading wells of the chip, and 10 μl of the 0.05$\times$ TBE buffer containing 2% BME was loaded in the rest of the wells. Using the pressure driven nitrogen flow, the DNA molecules were forced from the loading well, via the microchannels, into the nanochannels, in which they were confined and thereby stretched. Figure 3.3 also shows a zoom-in on the nanofluidic chip that visualises a DNA molecule located in the microchannel, where it behaves as if it was unconfined, as well as confined in one of the nanochannels.

In the majority of cases, one DNA molecule was imaged at a time. After the molecule had entered a nanochannel, there was a 90 second waiting time before imaging the molecule to make sure the DNA molecule found its equilibrium state in the nanochannel. As mentioned above, 30 frames were captured of each DNA molecule to record the molecule's intensity profile. This was done to account for thermal fluctuations of the DNA molecule which lead to variations in intensity along the characteristics in the resulting aligned kymograph, generated from the images.

As mentioned in the theory, the genome of *S. cerevisiae* consists of 16 chromosomes, which all have a size between approximately 200 kbp to 1.5 Mbp. In this project, there was a limit of how large molecules that could be imaged in one field of view of the microscope. For the microscope used, this limit was 512 pixels, resulting in that some of the smaller chromosomes could be imaged in one field of view, due to the approximate 1:1 relation between one pixel and one kbp. However, several DNA molecules that were imaged spanned out of the field of view of the microscope. In those cases, several images were taken of the same DNA molecule to cover the whole molecule.

For every session of imaging, a few molecules of λ-DNA were imaged and the length of the molecules was measured. The average length of the measured λ-DNA molecules for all the sessions was used as a size reference and thereby a reference of the stretching of the DNA molecules in the sample, which was important for the following data analysis.

3.4.1.3 Imaging using Multiplex Device

As a side track, imaging of the extracted DNA molecules from *S. cerevisiae* cells was also performed using a multiplex device, enabling higher throughput data collection compared to the nanofluidic chip shown in Figure 3.3, and another fluorescence microscope called Zeiss AxioObserver.Z1. This fluorescence microscope was equipped with autofocus, an Andor iXon Ultra EMCCD camera and two different light sources; LDI-7 Laser Diode Illuminator (89 North) and Thorlabs Solis-3C (High-Power LED for Microscopy, 5700 K (Day Light White), 3.5 W (Min)).

This was performed to be able to test if automation of the data collection was possible and if the results generated from the data collected using the multiplex device were comparable with those generated from the nanofluidic chip.

3.4.2 Data Processing & Analysis

After the data collection, the collected output data, in the form of movies consisting of a series of images of the DNA molecules, needed to be processed. In general, the data processing and analysis included molecule detection, generation of kymographs, alignment of kymographs, generation of barcodes as well as reference-based alignment and assembly experiments of the generated barcodes. These steps were performed using custom MATLAB scripts developed and provided by Tobias Ambjörnsson's group at the Department of Astronomy and Theoretical Physics at the University of Lund. Details on these steps of the data processing and analysis are described in the following sections.

3.4.2.1 Molecule Detection

Molecule detection and kymograph generation were performed by a MATLAB based tool called DBM, which stands for DNA Barcode Matchmaker. To generate the kymographs, firstly the movies of the imaged DNA molecules in .czi format were converted by the script into .tif format. The .tif files were thereafter loaded into the DBM from which the DNA molecules were individually detected. Kymographs were then created as each time frame of the detected molecule was stacked onto each other, resulting in each row of the kymograph illustrating the intensity along the DNA molecule in a single frame. Following, the kymographs were aligned. This was performed by aligning the characteristics present in the kymographs, and was done to compensate for the minor thermal fluctuations of the DNA molecules confined within the nanochannels.

The settings used in MATLAB for the DBM can be seen in Appendix A.1.

3.4.2.2 Reference-Based Alignment

The following part of the data processing was implemented using another specific MATLAB script encoding for a tool called HCA. HCA stands for Human Chromosome Alignment and is an in-house algorithm for chromosome alignment used for reference-based alignment given any reference chromosome. From the aligned kymographs, a sequence-specific barcode was generated by the script for each molecule detected by condensing the kymographs. The sequence-specific barcode is an 1D intensity profile that is averaged over time along the length of the DNA molecule. Following, reference-based alignment was performed by comparing the generated experimental barcodes to theoretical barcodes or intensity profiles of the genome of *S. cerevisiae*, which were generated based on the sequence of the 16 chromosomes of the yeast.

However, before the reference-based alignment was performed, merging of neighbours was implemented. For the DNA molecules that spanned out of the field of view of the microscope and thereby were imaged several times to cover the whole molecule, merging of the kymographs belonging to the same molecule was performed. The merging was nomenclature based, meaning that the neighbouring kymographs

were named in a specific manner, and by loading them into MATLAB, merging was performed by aligning the 100 most similar consecutive pixels of the barcodes generated from the kymographs. Following this step, the merged barcodes could be used in the data processing.

As mentioned before, the λ-DNA was used as an internal size reference to determine the degree of extension of the DNA molecules in the sample, which is given in nanometers per base pairs (nm/bp). The theoretical barcodes, as well as the generated experimental barcodes, were adapted based on the DNA extension observed in the imaging sessions. In addition, a stretching factor of 15% (0.85 - 1.15) and a step of 1% (0.01) were used when performing HCA. This means that during the reference-based alignment the barcodes were shrunk down to 85% of its original length as well as stretched up to 115%, and each percent of stretching or shrinking of the barcodes within this range were compared to the theoretical barcodes. In Appendix A.2, more information about the different parameters used for the reference-based alignment using HCA can be seen.

Every experimental barcode with every stretching, and both directions of them, was compared to every possible starting position along the theoretical barcodes. Every comparison was evaluated in the form of a cc-score, which is a score of the matching to the reference barcodes based on a Pearson correlation coefficient. A better match is indicated by a higher value of the cc-score, thereby, the maximum cc-score found is the best match to the reference genome. The output from HCA consists of the highest cc-scores for every experimental barcode, as well as the size of the experimental barcode in pixels. Given was also at which of the 16 chromosomes that the best match was found, the start pixel on the chromosome at which the best match was located as well as the extension in pixels of the match, and finally, the stretching of the experimental barcode for each given match.

HCA and reference-based alignment was done on both the barcodes when no merging had been done, as well as when merging had been done. HCA was also performed on the data generated by the multiplex device. The software R was used for analysing the output of the HCA.

3.4.2.3 *De novo* Assembly

In addition to performing reference-based alignment, assembly experiments of the experimental barcodes generated from the genome of *S. cerevisiae* were performed. Instead of the barcodes being compared to theoretical reference barcodes, as in the reference-based alignment, the experimental barcodes were compared to each other, with the goal to obtain 16 groups of barcodes, also called bargroups, representing the 16 chromosomes of the *S. cerevisiae* genome.

In the assembly experiments, each experimental barcode was compared to every other barcode in every possible orientation and the computer algorithm looked for the best overlap, of a specific minimum length, between the barcodes, that also resulted in a good overall overlap. These comparisons were evaluated by the cc-

score, described previously, and were thereafter compared to the rest of the generated scores and using a threshold bargroups were generated. A bargroup is created so that each barcode in the bargroup overlaps with at least another barcode within the same bargroup with a high cc-score. This high cc-score becomes an important threshold for bargroup quality, which we decided to call "thresCC". This means that, if a bargroup is made by barcodes A, B and C, A must have a cc-score over thresCC for its overlap alignment with either B, C, or both. The overlap calculated for this thresCC is not the full overlap, but rather the minimum overlap as described before. As part of the process of parameter optimisation, three different arbitrary levels for thresCC were used: 0.9, 0.85, and 0.8. These levels yielded different results which must be interpreted with caution.

The assembly was performed both with the barcodes generated from the data set without merging as well as with merging. The assembly was also tested with masking of the barcodes, meaning that the predominant features, such as high peaks and low valleys, of the barcodes were removed. In more detail, any pixel that was more than 2.5 standard deviations away from the mean of the bargroup was hidden.

In Appendix A.3, the different parameters that were used and changed during the assembly experiments can be seen. One of the most important parameters in the assembly experiments is the threshold (thresCC). This threshold is an internal score in the MATLAB script for the assembly experiments which is used to filter between good and bad bargroups. As mentioned previously, the values of the threshold used were 0.9, 0.85 and 0.8. Two other parameters that are important and could be changed are the minimum overlap of the barcodes in the assembly experiments as well as the stretching of the barcodes. For all of the assembly experiments performed, the minimum overlap used was 120 px, and the stretching factor used in the assembly experiments was the same as for the HCA, a stretching factor of 15% (0.85 - 1.15) and a step of 1% (0.01).

4
Results & Discussion

The obtained results from the project are presented in the following chapter, combined with a discussion around the results and their significance. The results are divided into four main parts. Firstly, the results from the data collection are presented, followed by the results from the PFGE, the results from HCA and the reference-based alignment, and lastly the results of the assembly experiments.

4.1 Data collection

In the following section, results regarding the data collection are presented and discussed. Both the data collection using the nanofluidic chip as well as the data collection using the multiplex device are discussed in this part.

4.1.1 Pre-processing of Data Output

In total around 600 images were captured on DNA molecules during the project. This resulted in around 720 molecules being detected by the DBM, and the same amount of kymographs were generated. After manually going through the generated kymographs, throwing away not usable kymographs, it resulted in 642 kymographs. These kymographs were fed into MATLAB where a threshold of length 150 px was set, thereby filtering out kymographs being shorter than this length. This resulted in 478 kymographs being used for analysis.

Additionally, as mentioned in section 3.4.1.2 in the method, molecules larger than the field of view of the microscope were imaged. More precisely, there were 82 molecules that spanned out of the field of view of the microscope and thereby were imaged two times to cover the whole molecule. The 164 kymographs belonging to these molecules were merged by aligning the 100 most similar consecutive pixels of the barcodes generated from the neighbouring kymographs, and the generated merged barcodes were used for analysis. This resulted in a total of 396 molecules being used for analysis when merging of neighbours had been performed.

The analysis in this project was performed using both the data when no merging of neighbouring molecules had been performed, as well as the data with the merged neighbours. From now on, when presenting results on non-merged data, the input consisted of 478 barcodes, otherwise the input consisted of 396 barcodes.

4.1.2 Distribution of Molecule Sizes

The size distribution and the mean size of molecules used for analysis can be seen in Figure 4.1 and 4.2. Figure 4.1 shows the size distribution of the data without merging of the neighbouring molecules, i.e. the 478 molecules larger than 150 kbp detected by DBM, as well as the mean size visualised by the vertical line which is 342.3 kbp. Figure 4.2 shows the "true" size distribution of the 396 imaged DNA molecules used for analysis as well as the mean size of the molecules which is 349.8 kbp. The height of the bars represent the number of molecules, referred to as count, within the size interval the bar covers. It should be noted that the scale of the x-axis, which represents the size in kbp, is different in Figure 4.1 and 4.2.

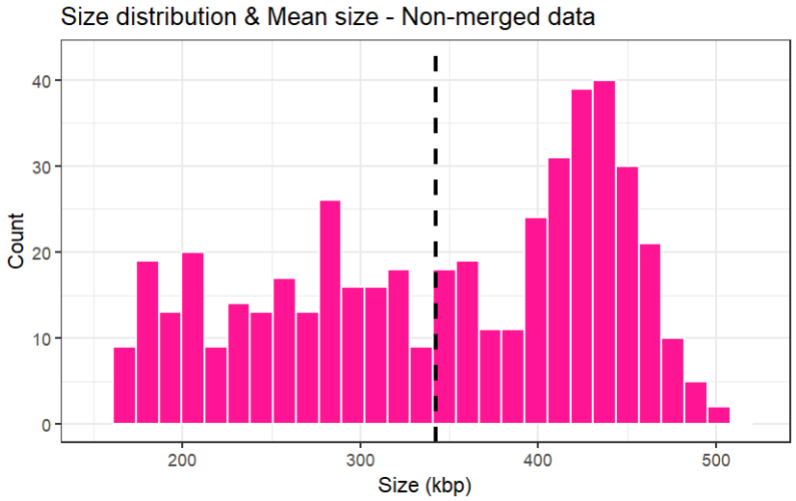

Figure 4.1: Size distribution of the DNA molecules used for analysis when no merging of neighbouring molecules had been performed. The vertical line indicates the mean size of the molecules, which is 342.3 kbp.

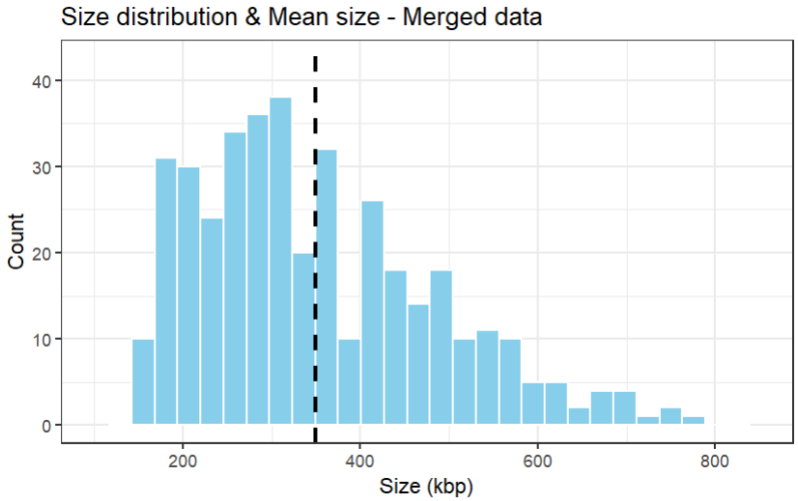

Figure 4.2: Size distribution of the DNA molecules used for analysis when merging of neighbouring molecules had been performed. The vertical line indicates the mean size of the molecules, which is 349.8 kbp.

The limiting length of molecules imaged in one field of view of the microscope used can also be seen in Figure 4.1. The field of view of the microscope used was 512 px, and as can be seen the longest molecule imaged in one picture is around 500 kbp. Due to the nanometer per pixel (nm/px) ratio of the microscope used for collecting the data, one pixel corresponds to approximately 1 kbp. However, it also depends on the stretching of the molecules.

The merging of neighbouring molecules resulted in an increase of the mean size of the molecules by 7.5 kbp to a value of 349.8 kbp. Even though this is a small increase in mean size, Figure 4.2 shows that molecules up to around 800 kbp were imaged. Compared to other techniques analysing the genome of different organisms, such as sequencing techniques, fragments of these sizes are today not possible to analyse.

As mentioned in the theory in section 2.4.2, the 16 chromosomes of the genome of *S. cerevisiae* span between approximately 200 kbp to 1.5 Mbp in length. This indicates that intact chromosomes could have been imaged, as DNA molecules of a size up to approximately 800 kbp were imaged using the fluorescence microscope.

4.1.3 Size of λ-DNA

During the project, images of the extracted DNA were captured during 11 imaging sessions. As mentioned in section 3.4.1.2, few molecules of λ-DNA were imaged and their length were measured for each session. In Table 4.1, the average size of the λ-DNA molecules imaged at each session, as well as the number of λ-DNA molecules captured at each time can be seen.

Table 4.1: Amount and average size of λ-DNA captured at each imaging session.

Session	Date	Number of λ-DNA imaged	Average size of λ-DNA [μm]
0	230213	2	9.8
1	230213	2	11.2
2	230215	2	11.2
3	230221	2	13.0
4	230222	3	12.9
5	230227	3	13.2
6	230228	3	12.5
7	230308	3	13.2
8	230309	5	12.8
9	230314	4	12.9
10	230421	15	12.5

Different staining reactions were used throughout the project which are presented in 3.4.1.1. The second staining was used for the majority of imaging sessions as that staining worked properly and as desired. This staining was used in session 3 to 9 seen in the table, while the first staining was used for the sessions before (0-2), and the third staining was used for the last session (10).

As can be seen in the table, the average size of the λ-DNA molecules got longer to begin with. From the first to second imaging session (0 to 1), the size increased as a result of dilution of the sample. As mentioned in the theory, salt concentration of the solution affects the stretching of the DNA molecules. A decrease in salt concentration, which occurs when diluting a sample, results in an increased stretching of the DNA molecules. The desired length of λ-DNA, which is based on recommendations, is a minimum of around 70% of its contour length. This corresponds to a value of 11.55 μm. As can be seen in Table 4.1, the desired length of λ-DNA was fulfilled at most imaging sessions. A length of 11.2 μm was determined to be acceptable which was the average size for session 1 and 2, as this extension is close to the recommended approximate minimum-value, and the generated barcodes of the λ-DNA at these sessions matched good with the reference of λ-DNA. However, the images captured as session 0 were excluded from the analysis as the length of 9.8 μm was seen as too short, and thereby a dilution of the sample was performed. It can also be seen that the average size per session was approximately the same, as only a variation of approximately 0.5 μm occurs between the majority of imaging sessions.

The average length of the λ-DNA molecules of session 1 to 10 was 12.54 μm, based on the average of each session, and was used as a reference for the degree of extension of the DNA molecules in the sample. This corresponds to a value of 0.259 nm/bp, which is based on the number of base pairs λ-DNA consist of, and was used for the data analysis.

The table also shows the amount of λ-DNA molecules imaged and measured at each session, and as can be seen, there was an increase in λ-DNA molecules with the sessions. From the first to the second staining reaction, the proportion of λ-DNA was increased from 10% to 30%, which helped in the process of imaging λ-DNA molecules, thereby an increase in λ-DNA imaged and measured. However, only a few λ-DNA molecules were needed at each session to determine the stretching of the molecules in the sample. Thereby, when a few λ-DNA molecules were measured, no active search after more λ-DNA was performed.

4.1.4 Challenging Parts with Data Collection

There were some challenging parts with the data collection. Degradation of the DNA molecules was one. During the different imaging sessions it was seen that the DNA molecules got shorter with time using the same sample. This was noted as fewer long DNA molecules, spanning out of one field of view of the microscope, were seen. This indicates a DNA degradation that occurs with time in the prepared sample. However, this does not indicate that DNA degradation occurs in the gel plugs from which the DNA is extracted. To make sure this is not the case, extraction as well as imaging of DNA from a new gel plug could be performed. Breakage of DNA molecules can also occur during pipetting of the DNA molecules as this increases the stress on the molecules leading to breakage. However, to minimise this breakage, a cut pipette tip was used and slowly pipetting was performed at all times when handling the DNA molecules.

Another challenging part of the data collection was detection of λ-DNA molecules. Especially during the first imaging sessions, it was hard to find any λ-DNA molecules. It can be problematic if no λ-DNA molecules are imaged for a session, as no reference of the stretching of the DNA molecules are given leading to that the imaged DNA molecules are unusable in principle. Without a size reference there is no reliability of the matching to the reference of the barcodes generated from the DNA molecules imaged at the same session. This is the case for assembly of the barcodes as well, as no reliable comparison of the barcodes can be made. However, as mentioned above, the proportion of λ-DNA was increased from 10% to 30% from the first to the second staining, which led to a smaller challenge in finding λ-DNA molecules.

Also, as mentioned in the method, three different staining reactions were made during the project, which differed in both total amount of DNA as well as the proportion of extracted and λ-DNA. The change in total amount of DNA between the first to second staining was made due to a lot of DNA molecules in the first staining which lead to some formation of invisible clumps in the nanofluidic chip making the data collection a little more challenging.

4.1.5 Pre-processing of Multiplex Device Data Output

In order to evaluate automation of the data collection, a multiplex device was used instead of the nanofluidic chip, as well as a microscope equipped with autofocus, for imaging the DNA molecules. A comparison was made between the data collected using the nanofluidic chip and the data collected using the multiplex device, with approximately the same amount of experimental barcodes.

Around 3 400 molecules were detected from the data output from using the multiplex device, however only 375 kymographs passed the threshold set to 150 kbp. The size distribution and the mean size of these 375 molecules imaged using the multiplex device can be seen in Figure 4.3. The mean size of the molecules is represented by the vertical line in the figure, which is 199.2 kbp.

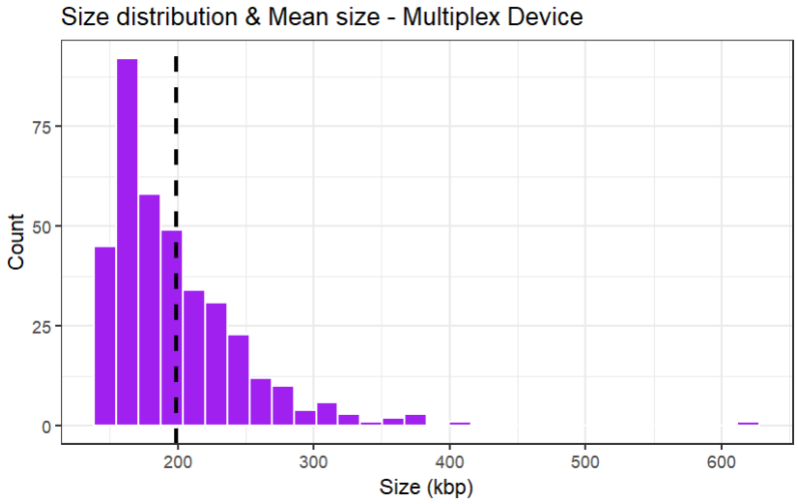

Figure 4.3: Size distribution of the DNA molecules imaged using the multiplex device and fluorescence microscope equipped with autofocus. The vertical line indicates the mean size of the molecules, which is 199.2 kbp.

The majority of the imaged molecules using the multiplex device had a size smaller than 150 kbp, as only around 10% of the molecules were left after filtering the data based on the size threshold set to 150 kbp. It could also be seen in the figure that the majority of molecules imaged by the multiplex device were shorter compared to the data collected using the nanofluidic chip, seen in Figure 4.2. However, one very long molecule with a size of approximately 620 kbp was imaged.

Since larger molecules are preferred, both for the reference-based alignment, but especially for the assembly experiments, the data collected by the multiplex device does not seem like a good candidate for the following analysis. Therefore, the following results are focused on the data collected using the nanofluidic chip, both the non-merged and merged data.

However, it is important to note that the data collection using the multiplex device was performed between session 9 and 10, using the third staining. As DNA degradation was seen with time, this with high probability affected the results of the sizes of the molecules imaged. To establish a fair comparison with the data collected using the nanofluidic chip, new data collection with DNA extracted from a new gel plug using the multiplex device would be desired, and also to collect the data around the same time as with the nanofluidic chip.

For the data collection using the multiplex device, no λ-DNA molecules were measured manually. Instead, a pipeline in the DBM tool was used that automatically recognised and measured the λ-DNA molecules. The average size of the detected λ-DNA molecules were 9.12 μm, which corresponds to a value of 0.188 nm/bp, and was used for the analysis of the barcodes generated from the data collected by the multiplex device.

4.2 Pulse-Field Gel Electrophoresis

Three different PFGE attempts were performed during the project, however, none of them were successful. For the first attempt of PFGE, wide bands of DNA could be seen in the gel, however, no ladder could be seen. For the second attempt, it was the other way around as the DNA ladder could be seen but no bands of DNA from the gel plugs. Figure 4.4 shows the result of the third attempt of PFGE performed. In the figure, the λ-DNA ladder is seen to the left, followed by three lanes of extracted DNA from the gel plugs with an increasing concentration from left to right. As can be seen in the figure, no clear distinct bands can be seen for the ladder or any of the lanes with the extracted DNA. However, with an increasing concentration of the DNA some bands are starting to show.

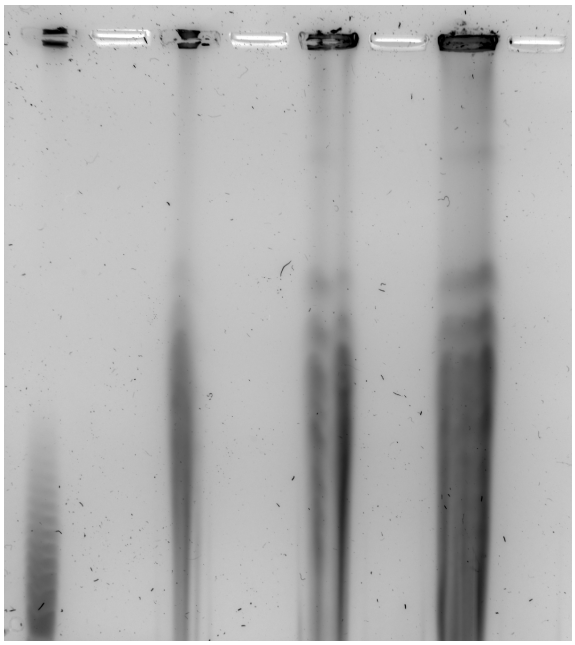

Figure 4.4: Result of the third attempt of PFGE. From left to right: λ-DNA ladder followed by increasing concentration of extracted DNA.

A major theory on the unexpected result of the PFGE is that not only intact chromosomes were present in the gel plugs, resulting in fragments of different sizes of the chromosomes. This would affect the result of PFGE as the sample loaded in the wells contains DNA molecules with a wide variety of sizes, resulting in that no distinct bands are seen. The reason for this is that DNA molecules will then be present in-between where the expected bands would be. In other words, from each chromosome there could be fragments which would travel further in the gel and smudge out the expected bands.

Breakage of DNA could be due to partial digestion by nucleases, which might have not been removed by the proteinase K during the creation of the plugs, or other endogenous nucleases. Another possible reason for DNA-breakage could be that the harvested cells were frozen before the creation of the gel plugs, which may have affected the cells and the DNA molecules.

What also could have contributed to the result is that the program used may not have been optimal for the DNA molecules in the gel plugs, and also for the ladder. Another ladder could also have been used to easier interpret the results of the gel.

4.3 Reference-Based Alignment

The results of the reference-based alignment performed by HCA, the in-house algorithm for chromosome alignment, are presented in this part of the results. This part of the result presents and compares the result of the analysis using both the non-merged data as well as merged data.

4.3.1 Molecule Size vs cc-score

As explained in the method, the output of HCA contains the size and best cc-score, among other things, for every experimental barcode. Figure 4.5 and 4.6 shows plots of the molecule size compared to the cc-score of the experimental barcodes from both the non-merged and merged data, in which every dot represents one experimental barcode. The y-axis of the plots represents the value of the cc-score, which can be any value between 0 and 1, where a value of 1 would be a perfect match, and on the x-axis the size of the experimental barcode is seen, which is the same as the molecule size.

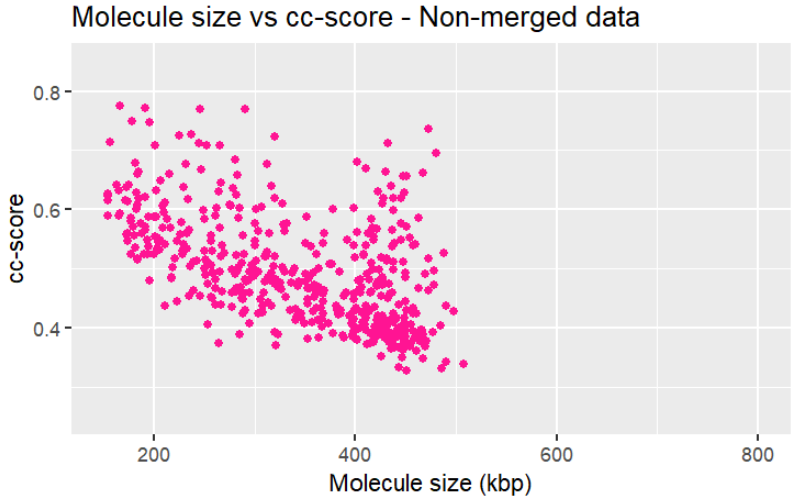

Figure 4.5: The cc-scores plotted against the size in kbp of the DNA molecules used for analysis without merging of neighbouring molecules.

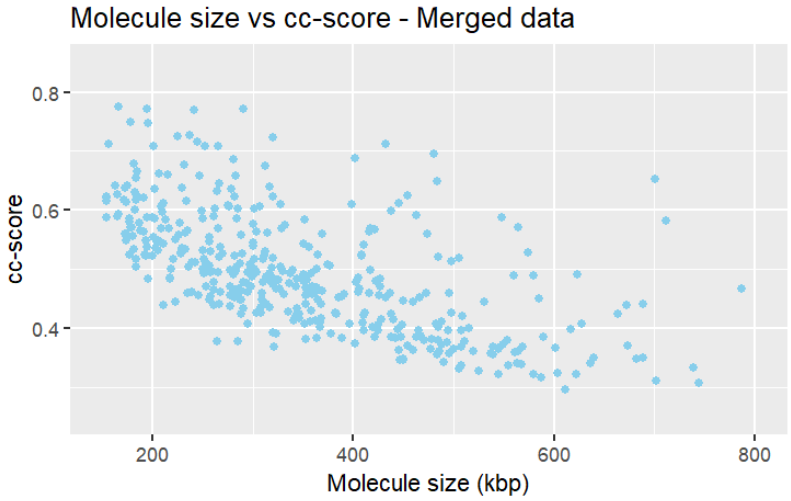

Figure 4.6: The cc-scores plotted against the size in kbp of the DNA molecules used for analysis when merging of neighbouring molecules had been performed.

A trend can be seen in both figures that shorter molecules tend to have a higher cc-score, indicating a better match to the reference barcodes, which are based on the reference genome of *S. cerevisiae* available at SGD. However, a higher cc-score does not always indicate a better match. The cc-score also depends on the size of the experimental barcode or molecule being matched to the reference and therefore, the size must be taken into consideration when evaluating cc-scores. The cc-score is based on the deviation of every pixel of the experimental barcode compared to the reference barcode. Each of the pixels deviated from the reference are summed up, resulting in that the longer the experimental barcode, the more pixels and more deviations are added up.

The trend that is seen indicates that longer barcodes tend to have a lower cc-score. This is as expected since a longer barcode has more features; peaks and valleys, which are compared to the reference barcodes. It is unlikely that each of these features match perfectly to the reference barcode, and therefore, a lower cc-score is expected compared to shorter barcodes with less features. This is the case since all of the small deviations of every pixel compared to the reference add up, resulting in a decreased cc-score with an increasing number of pixels.

As an example, the features of the short λ-DNA molecule include a high peak followed by a lower peak. This feature or pattern is located many times in the reference barcodes, and the λ-DNA molecule matches to each of these patterns with a high cc-score. Also, the second and third best match of the λ-DNA tend to have a high cc-score. Since the λ-DNA molecule is not part of the reference these are false matches to the reference and need to be handled with caution. However, all λ-DNA molecules were filtered out due to the 150 kbp size threshold, and λ-concatemers, which are DNA molecules containing several copies of the λ-DNA sequence in a series, were manually removed.

It could also be seen, when comparing the figures, that the clump of barcodes having a size larger than approximately 400 kbp in Figure 4.5, dissolves and spreads out in Figure 4.6 towards the left corner. These are the merged molecules having a larger size, but generally lower cc-scores compared to shorter molecules. However, there are a few very long barcodes with a size around 800 kbp in Figure 4.6 with higher cc-scores deviating from the trend that larger molecules have lower cc-scores.

What was seen in the output data from HCA was that for some barcodes, the best and second best match had a similar cc-score. The explanation for this could be that there are regions on different chromosomes that are similar to each other. This could be the case due to the whole-genome duplication that occurred in the evolution of the yeast *S. cerevisiae*, in which 8 chromosomes became 16. Thereby, there may be similar regions on the different chromosomes resulting in that a barcode with this region may match good to both chromosomes. This is however problematic, since it could be hard to map these regions correctly.

4.3.2 Merging of Neighbouring DNA Molecules

There are several reasons for why merging of the neighbouring molecules should be done. One reason is because both HCA, the reference-based alignment, and bargrouping by the assembly experiments benefit from having larger molecules, to avoid incorrect matching. Additionally, when no merging of neighbouring molecules is performed, noise to the coverage depth is added. A more detailed explanation of the benefits with merging is explained in section 4.3.3.

Figure 4.7 and 4.8 shows an example of how the merging of neighbouring molecules was performed. The merging of the neighbouring kymographs, as explained previously, was performed by aligning the 100 most similar consecutive pixels of the barcodes generated from the kymographs.

In Figure 4.7, the two neighbouring kymographs that should be merged are presented in the upper part of the figure, as the red and black barcodes. Below, the 100 most similar consecutive pixels of the barcodes that were aligned are shown, as well as the cc-score for this local alignment. The full overlap of the neighbouring barcodes is seen in the lower part of the figure, which in this case is approximately 260 pixels, as well as the cc-score for the overall alignment. As can be seen, this is a good example of merging in which the cc-score for the local alignment is 0.82372 and the cc-score for the full overlap of approximately 260 pixels is 0.63727. As mentioned in the previous section (4.3.1), the cc-score depends on the size of the alignment of two barcodes, and tends to be lower for a longer alignment. This is the case for this alignment. The local alignment has a higher cc-score than the full overlap, which is longer.

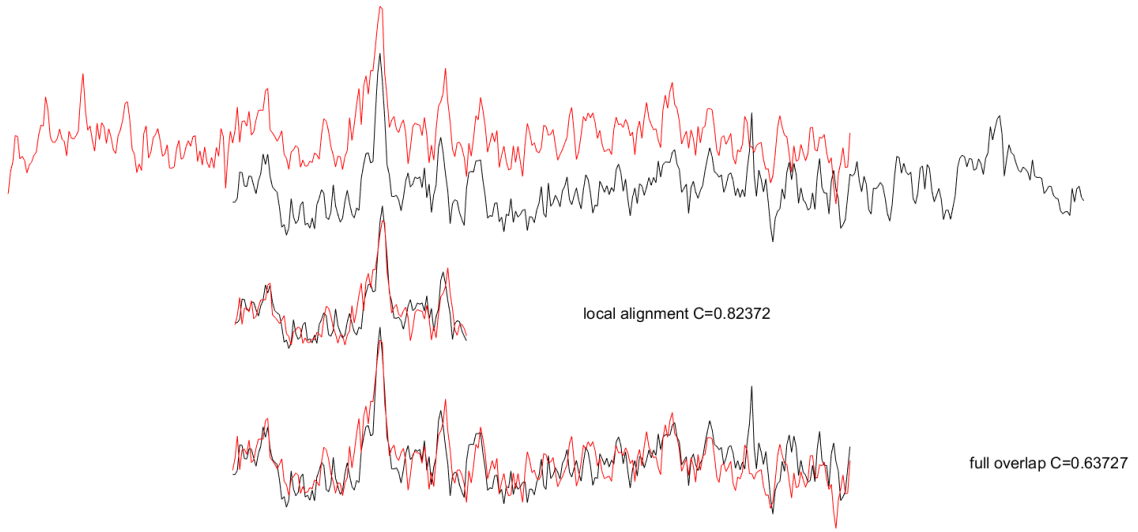

Figure 4.7: An example of merging of two neighbouring barcodes, shown as red and black. The full length of the barcodes are shown in the upper part of the figure. Shown are also the 100 most similar consecutive pixels that were aligned and the full overlap, as well as the belonging cc-scores for the local alignment and full overlap.

Figure 4.8 shows the same example of merging as in Figure 4.7. The blue barcode is the same as the red barcode in Figure 4.7 and the orange barcode is the same as the black one. In this figure, the barcode generated from the merging is also shown, which is the yellow barcode. This generated merged barcode is an average of each pixel for the two barcodes. The final length of the merged barcode in this example has a length of approximately 475 pixels, shown as the x-value in the figure.

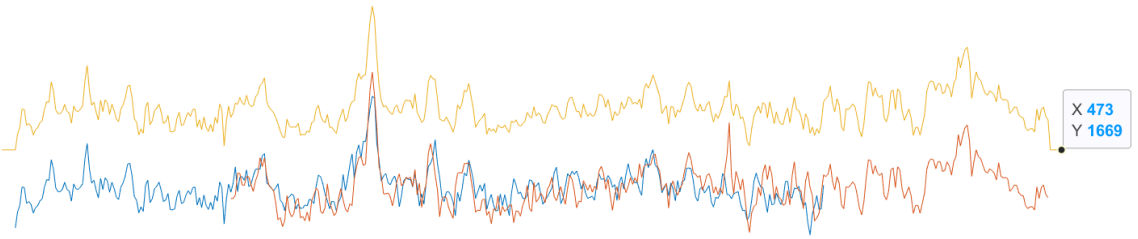

Figure 4.8: An example of two neighbouring barcodes, shown as blue and orange, as well as the generated merged barcode, shown as the yellow barcode. The length of the generated merged barcode is shown by the x-value in the label.

These two figures are only one example out of the 82 merged molecules. This is one of the best merging, based on the cc-scores of the alignment, however, there was no merging of low quality that was performed. Overall, the merging worked really well for the data collected with the nanofluidic chip.

4.3.3 Coverage

The total coverage for the whole genome of *S. cerevisiae*, as well as the coverage of each chromosome both for the non-merged and merged data, are shown in Table 4.2 and 4.3. The coverage is presented as the amount of times (x) the genome or chromosome is covered with the imaged molecules matched to the reference, also referred to as coverage depth, as well as the percentage of pixels covered by the imaged molecules, which could also be referred to as the extent of the coverage.

In Table 4.2, the total coverage of the *S. cerevisiae* genome with the non-merged and merged data is seen. It can be seen that there is a difference in the number of times the genome is covered when the analysis was performed with the non-merged data compared to when the analysis was performed when merging of the neighbouring molecules had been performed. Analysis with the non-merged data resulted in a coverage depth of approximately 13x, compared to approximately 11x with the merged data. The coverage depth is based on the total amount of kbp of the imaged molecules compared to the amount of kbp in the reference genome, which is approximately 12 200 kbp. Therefore, with the molecules used for analysis, it can be said that with these captured molecules, the genome should be covered a certain amount of times. A higher value of this coverage for the non-merged data is therefore expected as the molecules spanning out of the field of view of the microscope were captured twice, resulting in a higher amount of kbp of the imaged molecules. In the merged data, the overlapping regions of these molecules are merged, and thereby, a lower amount of kbp of the imaged molecules.

Table 4.2: The total coverage depth of the genome of *S. cerevisiae* as well as the total coverage extent, given in percentage of pixels of the reference genome covered, for the non-merged and merged data.

Data	Total coverage depth [x]	Total coverage extent [%]
Non-merged	13.4x ($\sim$ 13x)	98.24119
Merged	11.4x ($\sim$ 11x)	98.20861

However, the total coverage extent given in percentage, which represents the amount of pixels covered by the imaged molecules, should be approximately the same for the non-merged and merged data if HCA worked properly. This is the case as the molecules which are merged should map to the same pixels in the reference genome. As can be seen in Table 4.2, this is the case. The total coverage for both the non-merged and merged data is approximately 98.2%. If this would not be the case, it could be an indication that the individual molecules that were merged and used in analysis with the merged data, were mapped to different pixels or regions when the analysis was performed with the non-merged data.

In Table 4.3, the coverage for each chromosome is presented for both the non-merged and merged data. In this table, it can be seen if the molecules of the non-merged and merged data map to different chromosomes, as it would result in a difference of

both the coverage depth and extent between the different data sets. In the table, it can be seen that there is no difference in coverage depth or extent between the different data sets for chromosomes 1, 3, 6 and 9, which indicates that the barcodes in the different data sets are mapped to the same regions of these chromosomes. For the rest of the chromosomes, some of the barcodes were mapped to different reference chromosomes, resulting in different values of the different coverage.

Table 4.3: The coverage depth and coverage extent of each of the 16 chromosomes of *S. cerevisiae* for the non-merged and merged data. Coverage depth represents the times the chromosome is covered, while the coverage extent represents the percentage of pixels of the reference chromosome covered.

	Non-merged data		**Merged data**	
Chromosome	Coverage [x]	Coverage [%]	Coverage [x]	Coverage [%]
I	1.67	91.03	1.67	91.03
II	7.37	90.93	6.94	93.59
III	13.70	97.83	13.70	97.83
IV	12.78	99.68	11.87	98.72
V	15.08	99.66	12.10	99.32
VI	4.92	96.36	4.92	96.36
VII	11.88	99.10	8.58	99.19
VIII	10.96	99.30	7.17	99.30
IX	3.77	98.21	3.77	98.21
X	17.43	98.68	13.51	98.95
XI	20.49	98.38	19.11	99.26
XII	34.52	99.00	28.65	99.00
XIII	8.74	98.09	7.72	97.98
XIV	15.99	99.25	13.13	99.25
XV	8.89	98.56	7.24	97.48
XVI	10.85	99.79	9.72	99.27

Since it is known for sure that the neighbouring molecules, which span out of one field of view of the microscope, should be merged, they should also be mapped to the same region of the genome, the same chromosome. As this is not the case for some of the molecules of the non-merged data, seen in Table 4.3, it is an indication that the merging of neighbours works properly and should be done, as the neighbours then match to the right region of the genome with higher probability. This is one reason why the merging of neighbouring molecules should be done, as the neighbouring molecules should map to the same chromosome, as the merged molecule of these. The following results regarding the reference-based alignment that are presented are thereby result generated with the merged data set.

Visualisation of the coverage depth and coverage extent for each chromosome generated by the merged data will be presented in section 4.3.5.

4.3.4 Distribution of Matches along the Chromosomes

Figure 4.9 shows a barplot presenting the expected as well as the real number of barcodes matched to each of the 16 chromosomes of *S. cerevisiae* for the merged data set. The horizontal coloured bars represent the amount of barcodes matched to each of the chromosomes, while the black dots represent the expected distribution of matches to the chromosomes, which would be the case if a totally random distribution occurred. The same pattern was seen when analysing the data without the merging of the neighbouring molecules.

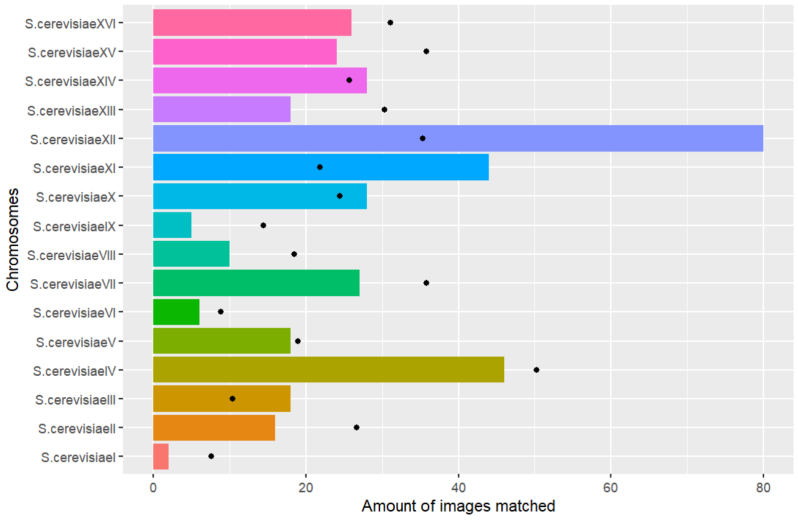

Figure 4.9: The amount of barcodes matched to each of the 16 reference chromosomes of *S. cerevisiae*, which are represented by the coloured horizontal bars, and the expected distribution represented by the black dots for the merged data set.

What can be seen in the figure is that the distribution of matches to the chromosomes follow the expected distribution overall. However, there are some deviations. One chromosome that stands out from the rest is chromosome XII, represented by the blue-purple bar. This chromosome got roughly the double amount of matches than expected, which is alarming. Also chromosome XI got a lot more matches than expected, approximately the double amount. The rest of the chromosomes got approximately the same or less amount of matches as expected, which could be due to chromosome XI and especially chromosome XII stealing matches from the rest of the chromosomes, which is seen as problematic. Another alarming thing that is seen is that chromosome I got very few matches, compared to the expected amount of matches, but also overall.

4.3.5 Coverage plots of the 16 Chromosomes

To look more deeper into how the different barcodes match to the reference genome composed of the 16 chromosomes, and to explore the alarming things from Figure 4.9, coverage plots of the chromosomes were generated. Figure 4.10 shows the coverage plots of each of the 16 reference chromosomes generated by the merged data

set. In the figure, the barcodes are represented as grey lines, spanned between a green and a blue dot which marks the ends of the barcode. The reference chromosomes are represented by the thicker black horizontal lines with pink parts which are located below the barcodes. The black part of the reference chromosomes indicate parts of the chromosome being covered by the barcodes matched to it, while the pink parts represent the uncovered parts of the reference chromosomes. The black lines over the reference chromosomes represent the amount of times each pixel of the reference chromosome is covered by the matched barcodes. If there are many barcodes matched to a chromosome, the steps of the black line will be smaller, as each step represents coverage of one barcode.

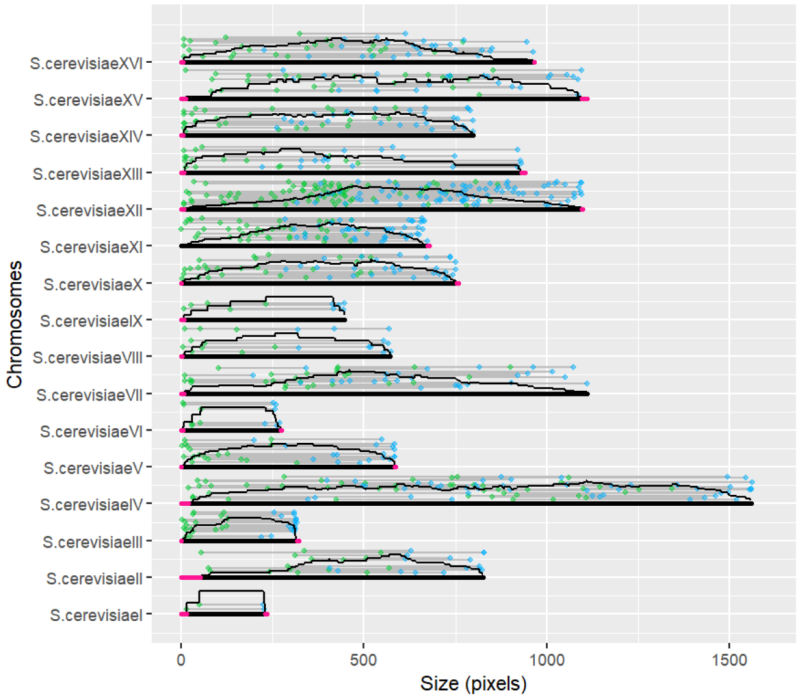

Figure 4.10: Coverage plot of the 16 reference chromosomes. The reference chromosomes are represented by the thicker black horizontal lines with pink parts and every barcode is represented by a grey line, spanned between a green and a blue dot, marking the ends of the barcode. The black part of the reference chromosomes indicate covered parts while the pink parts represent the uncovered parts of the reference chromosome. The black lines over the reference chromosomes represent the amount of times each pixel of the reference chromosome is covered by the matched barcodes.

The regions covered by each of the 16 reference chromosomes can be seen in Figure 4.10. As can be seen, the majority of all the chromosomes are covered by at least one barcode, which is the case as the major part of the reference chromosomes are black. However, there are a few regions that are uncovered, represented by the pink regions of the reference chromosomes. These regions are located at the ends of the reference chromosomes, and it can be seen that the majority of the ends are uncovered regions.

There can be several reasons why especially the ends of the chromosomes are uncovered. One reason is that the reference genome of *S. cerevisiae* is not perfect, and the sequence especially at the ends of the chromosomes could be wrong and potentially longer. It is a high probability that this is the case as the ends, also called telomeric regions, of chromosomes are challenging to sequence. One reason for this is that the end regions tend to be repetitive and have GC-rich regions. As mentioned in the theory in section 2.2, it is challenging for sequencing techniques to sequence and map repetitive regions of genomes, resulting in that the sequence of these regions being less reliable. Also, when sequencing is performed, the DNA going to be sequenced is amplified prior to the sequencing which is performed by PCR. PCR is performed using primers which are selective to certain sequences, and thereby works better on these sequences. For example, PCR primers tend to not work well on GC-rich regions. Thereby, a selective amplification is performed leading to a challenge in determining the sequence of the ends of the chromosomes. This is another reason why the ends could be wrongly sequenced.

Another reason the ends of the chromosomes are not covered could be a result of the HCA. When the kymographs were loaded into MATLAB and HCA, the first and last predominant characteristic, seen as a dark or bright pixel, of the kymograph was detected as the ends of the kymographs. In other words, the ends of some of the kymographs could not have been detected, resulting in a shorter barcode in which the ends are excluded.

In Figure 4.11 the coverage plot of the shortest chromosomes are visualised, which are chromosome I, III, V, VI, VIII and XI. For these chromosomes, which all are shorter than 600 kbp, it can be seen that almost intact chromosomes are imaged, since there are barcodes spanning almost the whole chromosome and only few pixels at some of the ends are not covered. For the larger chromosomes, such as chromosome 12, it was seen that there was separate barcodes covering approximately half of the chromosomes.

However, it could be the case that intact chromosomes were extracted and imaged. This could be analysed by performing *de novo* assembly of the barcodes. A *de novo* assembly of the barcodes could solve the problem with the ends of the chromosomes not being covered. This could be the case since the assembly is not based on any reference, allowing barcodes to be matched to each other without the hindrance that the barcodes can not be aligned expanding out over the ends of the reference chromosomes. This is the case for the version of HCA and the reference-based alignment used in this project which uses global alignment. In other words, HCA is not perfect since the theory, which is based on the reference genome, is not perfect.

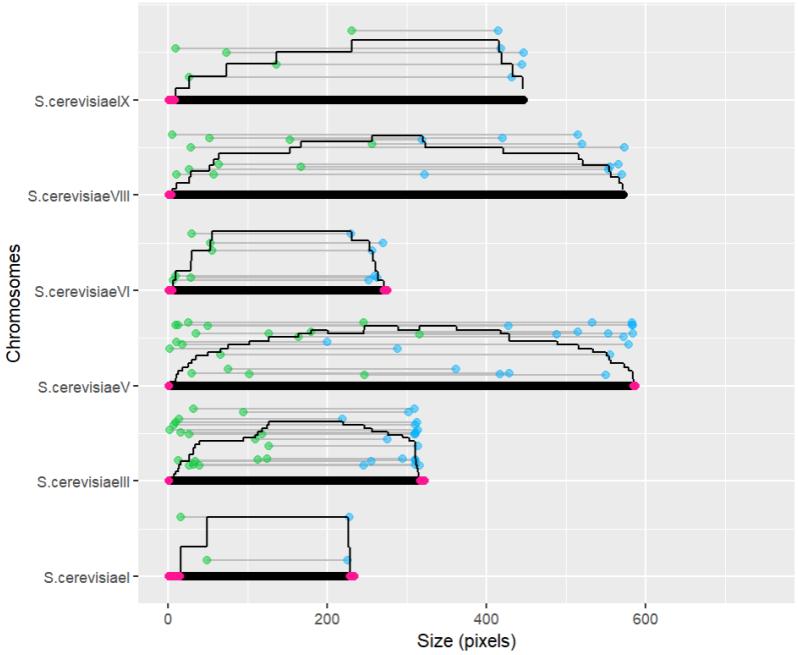

Figure 4.11: Coverage plot of the reference chromosomes I, III, V, VI, VIII and XI, being shorter than 600 kbp. The reference chromosomes are represented by the thicker black horizontal lines with pink parts and every barcode is represented by a grey line, spanned between a green and a blue dot, marking the ends of the barcode. The black part of the reference chromosomes indicate covered parts while the pink parts represent the uncovered parts of the reference chromosome. The black lines over the reference chromosomes represent the distribution of the amount of times each pixel of the reference chromosome is covered by the matched barcodes.

It can also be seen in Figure 4.11, that chromosome I, which is the shortest one with a length of approximately 230 kbp, only has two barcodes matching to the reference chromosome. As mentioned in the previous section, it is alarming that it got few matches overall but also compared to the expected amount of matches. The reason for this could also be due to the drawback of this version of HCA, that it does not allow experimental barcodes to expand out over the end of the reference chromosomes. Therefore, experimental barcodes being longer than the reference chromosome are not even compared to the reference, even if the majority of the barcode matches good to the reference chromosome. This is problematic since once again, the reference genome may be wrong at the ends of the chromosomes, and a result could be that barcodes that should be mapped to chromosome I are mapped to another chromosome instead. This may also be the case for chromosomes VI and IX, which have 6 and 5 barcodes matching to them respectively. This could potentially be solved by allowing experimental barcodes to expand longer than the reference, which can be performed by using local alignment instead. However, performing HCA with local alignment instead of global alignment is computationally intensive, but is used for the assembly experiments in which bargrouping occurs and could therefore result in the ends being covered.

One may think that changing the stretching of the experimental barcodes may solve the problem that the experimental barcodes of chromosome I are too long to be compared to the reference chromosome. However, this is not the case. The λ-DNA molecules which are used as size reference help us determine the right stretching of the experimental barcodes. If an experimental barcode is reduced in size by decreasing the stretching to the same length as the experimental barcode, comparison to the reference could occur. This would however result in that the features of the experimental barcode would not match with a good cc-score to the reference as the stretching of the experimental barcode would not be correct.

The coverage plot of chromosome V (seen in Figure 4.11) is a good example of how a good and even coverage of the whole chromosome can look like, where the coverage along the reference chromosome is similar. In other words, the middle of the chromosome is covered by approximately the same amount of barcodes as the ends. This can be seen as the black line representing the number of barcodes covering each pixel, is approximately at the same level throughout the whole chromosome. However, this is not the case for several of the chromosomes. For example, chromosome II, XIII, XV and XVI, as well as chromosome XI and XII, have a different distribution of the coverage along the chromosome.

Chromosome II, XIII, XV and XVI are shown in Figure 4.12. For all of these chromosomes, one of the ends of each chromosome is less covered. This is especially seen for chromosome II, in which the first 300 kbp seen in the figure are less covered, or not covered at all. This could show a mismatch between the actual barcode and the reference barcode at this position, and one reason for why the start of the chromosome is less covered could be due to a faulty reference sequence.

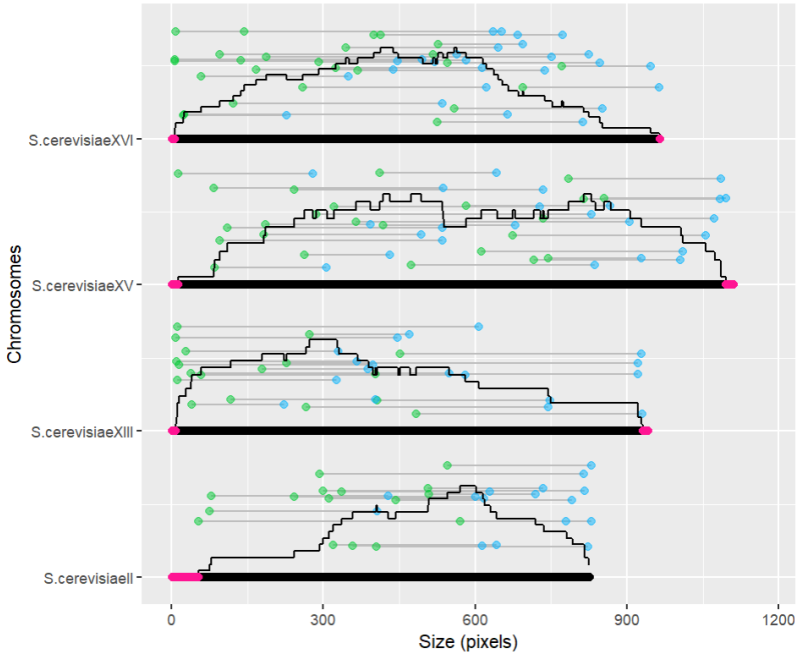

Figure 4.12: Coverage plot of chromosome II, XIII, XV and XVI. The reference chromosomes are represented by the thicker black horizontal lines with pink parts and every barcode is represented by a grey line, spanned between a green and a blue dot, marking the ends of the barcode. The black part of the reference chromosomes indicate covered parts while the pink parts represent the uncovered parts of the reference chromosome. The black lines over the reference chromosomes represent the distribution of the amount of times each pixel of the reference chromosome is covered by the matched barcodes.

Two of the chromosomes that were seen as alarming in the barplot in Figure 4.9, were chromosome XI and XII which had a lot of matches compared to the rest of the chromosomes. The coverage plots of chromosome XI and XII are seen in Figure 4.13. As can be seen, chromosome XII has many barcodes matched to it and the majority of barcodes are mapped to the middle of the chromosome. From literature, it is known that chromosome XII has a repetitive region consisting of around 100 to 200 copies of an unit with a size of 9.1 kb [31]. This region of chromosome XII could be one of the reasons that the large amount of barcodes align with this chromosome, and especially matched to the middle of the chromosome. The case may be similar with chromosome XI, since more barcodes than expected were matched to this chromosome. However, no literature has been found mentioning a known repetitive region of chromosome XI as for chromosome XII. Additionally, the coverage is also greater in the middle and towards one end of the chromosome, indicating that this region is more prone to get matches.

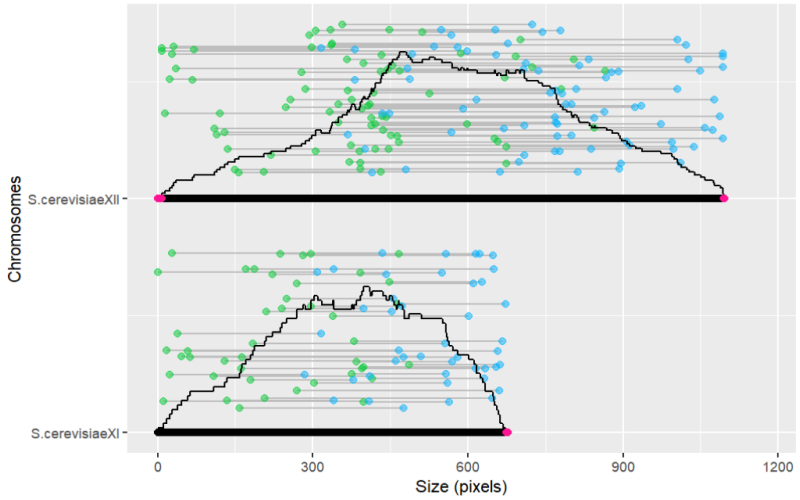

Figure 4.13: Coverage plot of chromosome XI and XII. The reference chromosomes are represented by the thicker black horizontal lines with pink parts and every barcode is represented by a grey line, spanned between a green and a blue dot, marking the ends of the barcode. The black part of the reference chromosomes indicate covered parts while the pink parts represent the uncovered parts of the reference chromosome. The black lines over the reference chromosomes represent the distribution of the amount of times each pixel of the reference chromosome is covered by the matched barcodes.

A possible explanation for the great extension of matches to these two chromosomes (XI and XII) could be due to the whole-genome duplication (WGD) of the genome of *S. cerevisiae*. The case could possibly be that these regions of chromosomes XI and XII that many barcodes are mapped to are similar to regions of other chromosomes, leading to the stealing of barcodes. According to literature, it is also thought that the WGD is one of the reasons for the occurrence of repetitive regions in the genome of *S. cerevisiae*.

Figure 4.14 shows a kymograph of a repetitive region found when analysing the imaged molecules, possibly the repetitive region known to be a part of chromosome XII. As stated above, this region is thought to consist of a 9.1 kb unit copied around 100 to 200 times. However, according to the reference genome, chromosome XII of *S. cerevisiae* consists of approximately 1.1 Mbp. If the repetitive region mentioned in the literature consists of 200 copies of the repetitive unit of 9.1 kb, this region would have a length of 1.82 Mbp which is longer than what the sequence of chromosome XII is according to SGD. It is however possible that chromosome XII consists of more base pairs. This could be the case since it is challenging for sequencing techniques to map repetitive regions, and if the major part of chromosome XII contains this large consecutive repetitive region, it could be that the chromosome has been sequenced wrongly.

On the other hand, it is also possible that the repetitive region is present in several locations in the genome, which would indicate that the repetitive region is not consecutive. Also, repetitive regions tend to be located at the ends of the chromosomes, in the telomeric regions, which also tend to be high in GC-rich regions and hard to sequence. If this would be the case, a major part of this region could not have been sequenced or correctly mapped, and thereby not a part of the reference sequence obtained from SGD.

Figure 4.14: A kymograph showing one repetitive region present in the genome of *S. cerevisiae*.

However, this kymograph seen in the figure gives an indication of the size of the repetitive region, since it can be seen that the molecule from which this kymograph is generated spans almost over the whole field of view of the microscope which has a size of 512 px. Nevertheless, there is nothing that excludes that this repetitive region is even longer. Imaging of longer molecules including the whole repetitive region in-between other sequences could possibly help to determine the size of the repetitive region. Another option to determine the length of the repetitive region could be *de novo* assembly of the data generated through ODM.

To round off the result and discussion of the reference-based alignment, the reference-based alignment performed by HCA gives an indication if the collected data is ready for assembly or not. This since it lets us know if we have enough and good quality data to start with the assembly. A high percentage of coverage of the total reference genome as well as for each chromosome provides a good foundation to get a reliable result of an assembly experiment, such as *de novo* assembly of a genome. In addition, also the coverage depth of the genome and its chromosomes gives an indication of the outcome of the assembly.

For example, an assembly of few barcodes would not give as reliable and trustworthy results compared to an assembly with a greater amount of barcodes used. However, even if the HCA is not perfect, and some chromosomes have a low coverage depth, assembly experiments could be the problem solver and generate more accurate results. As was discussed, chromosome I had very few matches and had a coverage of approximately 1.67x. However, an assembly may solve this problem, as HCA is not perfect and is based on a reference genome which is not known to be completely correct. In other words, things that are mismatched in the reference-based alignment may be fixed with the assembly experiments.

4.3.5.1 Multiplex Device Data

In Figure 4.15, the coverage plots of the 16 chromosomes generated by the multiplex device data can be visualised. As mentioned previously (section 4.1.5), based on the small size of the DNA molecules imaged with the multiplex device, the data generated by the multiplex device was not seen as a good candidate for the following analysis. The coverage plots seen in the figure strengthens this statement as there are a lot of uncovered parts of the chromosomes (pink parts), compared to the coverage plots generated by the data collection by the nanofluidic chip. For this data, the uncovered parts are not only located at the ends of the chromosomes as seen in the other case with the data collection by the nanofluidic chip, also regions located in the middle of some of the chromosomes are uncovered.

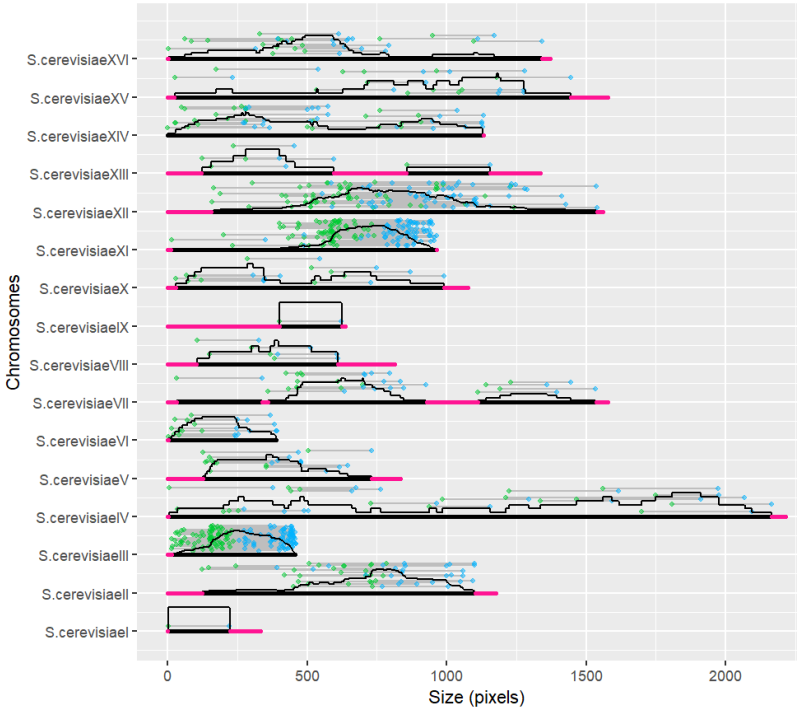

Figure 4.15: Coverage plot of the 16 reference chromosomes generated by the multiplex device data. The reference chromosomes are represented by the thicker black horizontal lines with pink parts and every barcode is represented by a grey line, spanned between a green and a blue dot, marking the ends of the barcode. The black part of the reference chromosomes indicate covered parts while the pink parts represent the uncovered parts of the reference chromosome. The black lines over the reference chromosomes represent the amount of times each pixel of the reference chromosome is covered by the matched barcodes.

Additionally, no uniform distribution of barcodes covering the chromosomes are seen for the 16 chromosomes. Most of the chromosomes have a specific region to which many of the barcodes are matched. It can also be seen that some of the molecules got a lot of matches, such as chromosome III, XI and XII while chromosomes I and IX only got one match each.

Due to these reasons, the following assembly experiments were not performed using the data generated by the multiplex device.

4.4 Assembly Experiments

After the reference-based alignment was performed, assembly experiments on the generated data using the nanofluidic chip were implemented. In this last part of the result and discussion, results from different assembly experiments performed on both the non-merged and merged data are shown and discussed.

As mentioned in the method (section 3.4.2.3), one of the most important parameters that was used and changed in the assembly experiments is the threshold called thresCC. The values of thresCC used in the different assembly experiments were 0.9, 0.85 and 0.8, while the rest of the parameters were unchanged between the different experiments excluding the data set used.

The assembly experiments resulted in a varying number of so-called bargroups, which are groups of barcodes being grouped together based on their similarity. In Figure 4.16, an example of a high quality bargroup is seen, which was generated with the non-merged barcodes using the thresCC value of 0.9. The lower part of the figure visualises the different barcodes belonging to the bargroup individually, and above the barcodes in the bargroup are visualised laying on top of each other. The x-axis of the figure represents the size in pixels.

As can be seen in the figure, the barcodes belonging to this bargroup match well to each other as they have a high similarity, and it is a high probability that these barcodes belong to the same chromosome. The whole barcode spans over a length of approximately 600 px and could for example be the whole chromosome V, which according to the reference genome is approximately 580 kbp long. This could be said since the relation between one pixel and kbp is approximately 1:1 for the microscope used, but it also depends on the stretching. This group of barcodes was seen in several of the assembly experiments, either as a bargroup by itself, or as a part of another bargroup.

Figure 4.16: A high quality bargroup generated using the non-merged data and thresCC 0.9. Upper part: Visualises the barcodes in the bargroup laying on top of each other. Lower part: Visualises the different barcodes belonging to the bargroup individually. The x-axis represents the size in pixels of the bargroup.

However, this was not always the case for the assembly experiments. Figure 4.17 shows another example of a bargroup that was generated for the same data (non-merged), with the same settings (thresCC 0.9). What can be seen in the figure is that the bargrouping is based on the predominant peak that is part of the majority of the barcodes belonging to this bargroup. However, when looking closer at the different barcodes, it can be seen that all the barcodes do not match to each other, as they are not similar beyond the high peak. Some of the barcodes seem to match each other while others do not.

For example, it can be seen that the blue barcode in the middle of the lower part of the figure, spanning from 0 to approximately 700 on the x-axis and the red barcode below match each other. However, since this bargroup was generated with the non-merged barcodes, they could with high probability come from neighbouring molecules, thereby being the same DNA molecule. But even if this is the case or not, it can be seen that they together make up a bargroup being almost 1 000 pixels long. With high probability, this could be one of the chromosomes of *S. cerevisiae*, which according to the reference genome is around 1 000 kbp, or be part of a longer chromosome. According to the reference genome, chromosomes VII, XII, XIII, XV and XVI are around 1 000 kbp long, while only chromosome IV is longer.

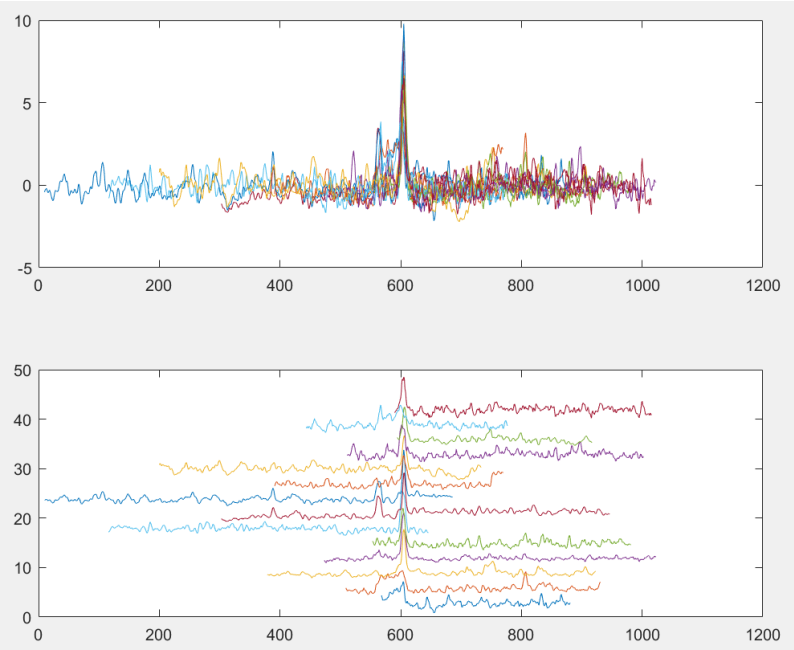

Figure 4.17: A bargroup generated using the non-merged data and thresCC 0.9. Upper part: Visualises the barcodes in the bargroup laying on top of each other. Lower part: Visualises the different barcodes belonging to the bargroup individually. The x-axis represents the size in pixels.

A bargroup with even lower quality is seen in Figure 4.18, which was generated from another assembly experiment using the non-merged data, but the value of thresCC was lowered to 0.85. As can be seen in the upper part of the figure, this bargroup contains a high predominant peak, as well as a lot of barcodes, which seems to not be similar to each other to a high extent. This bargroup is with high probability the same bargroup as seen in Figure 4.17, but expanded by more barcodes. The lower part of the figure is zoomed in on some of the barcodes in the bargroup. In this part of the figure, the high quality bargroup seen in Figure 4.16 can be seen at the bottom. However, this group of barcodes does not seem to be similar to the rest of the barcodes in this bargroup, indicating that they should be their own bargroup. In other words, when lowering the value of the threshold (thresCC) to 0.85, this high quality bargroup disappeared and became a part of a larger one, which it does not belong to.

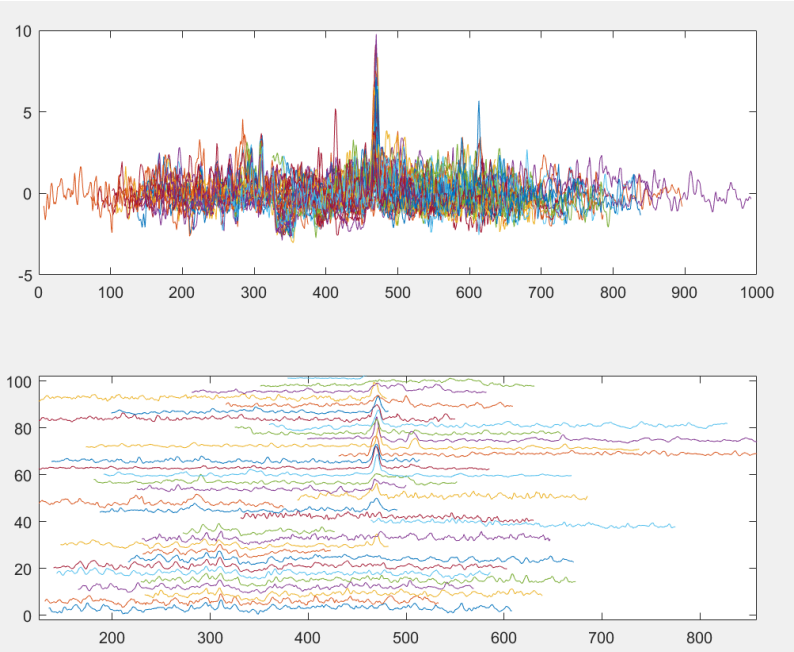

Figure 4.18: A bargroup generated using the non-merged data and thresCC 0.85. Upper part: Visualises the barcodes in the bargroup laying on top of each other. Lower part: Visualises the different barcodes belonging to the bargroup individually. The x-axis represents the size in pixels.

When the threshold was lowered even more to 0.8, this bargroup with the predominant peak, seen in both Figure 4.17 and 4.18, increased in size as even more barcodes were included in this bargroup. However, lowering the threshold did not only result in low quality bargroups. The bargroup seen in Figure 4.19 and 4.20 was generated in the assembly experiment using the non-merged data and thresCC 0.8. These bargroups were not generated using thresCC 0.9.

As mentioned above, since these bargroups were generated using the non-merged barcodes, there is a possibility that the high quality bargroups could contain neighbouring barcodes. However, this is not a problem as they in that case are assembled together, and also with other similar barcodes.

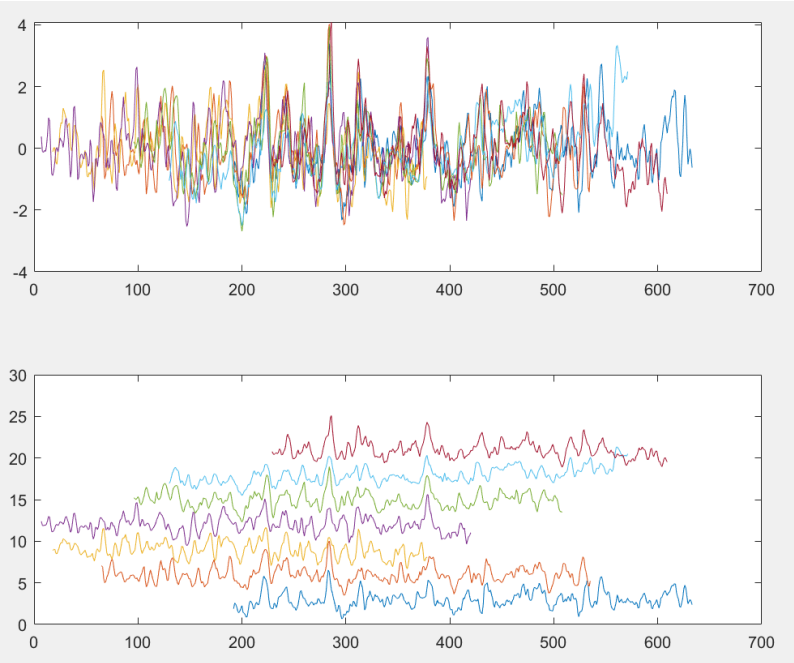

Figure 4.19: A bargroup generated using the non-merged data and thresCC 0.8. Upper part: Visualises the barcodes in the bargroup laying on top of each other. Lower part: Visualises the different barcodes belonging to the bargroup individually. The x-axis represents the size in pixels.

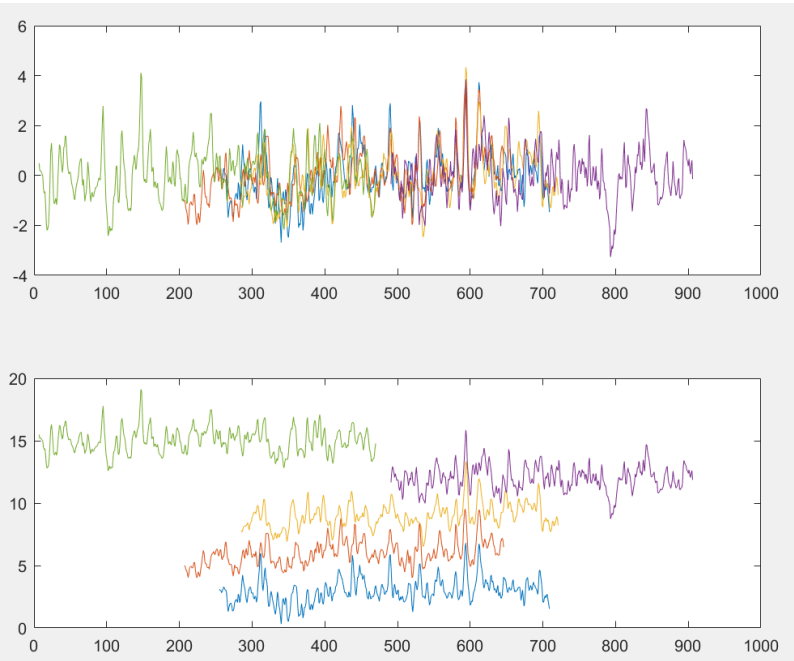

Figure 4.20: A bargroup generated using the non-merged data and thresCC 0.8. Upper part: Visualises the barcodes in the bargroup laying on top of each other. Lower part: Visualises the different barcodes belonging to the bargroup individually. The x-axis represents the size in pixels.

Assembly experiments were also performed using the merged data set, and with the same thresholds. Overall, the same outstanding bargroups were generated compared to the assembly experiments with the same thresholds using the non-merged data. However, a general trend that could be seen was that fewer bargroups and also smaller bargroups, regarding the number of barcodes in the bargroup, were generated for the same thresholds.

For example, the assembly experiment performed using the merged data and thresCC 0.9 generated the same two bargroups as observed in the corresponding assembly experiment for the non-merged data, seen in Figure 4.16 and 4.17. However, the bargroups consisted of less barcodes for the merged data, which could be expected since some of the barcodes belonging to these bargroups may have been merged. Also the bargroup corresponding to the one in Figure 4.18, with thresCC 0.85, was generated with the merged data.

The reason for why fewer barcodes were seen in the assembly experiments generated with the merged data, compared to the non-merged, is due to that the majority of the bargroups for the non-merged data not shown for the merged data consisted of neighbours. One example of such a bargroup is seen in Figure 4.21, in which two neighbouring barcodes are assembled into one bargroup. The figure shows a good example of neighbouring of barcodes.

Figure 4.21: A bargroup consisting of two neighbouring molecules generated using the non-merged data and thresCC 0.9. Upper part: Visualises the barcodes in the bargroup laying on top of each other. Lower part: Visualises the different barcodes belonging to the bargroup individually. The x-axis represents the size in pixels.

4.4.1 Assembly Experiments with Masked Barcodes

As mentioned in the last section of the method, assembly experiments with masked barcodes were performed. The masking was performed by removing the predominant features, such as high peaks and low valleys, of the barcodes and was thought to improve the assembly by avoiding generating bargroups such as the one shown in Figure 4.18. This bargroup contains one high peak to which many barcodes with this predominant peak was assembled to, even if the other regions of the barcodes were not similar and thereby should not be grouped.

What was seen when masking of the barcodes was performed was that the bargroup with the predominant peak disappeared or was divided into mainly two other bargroups. One example on this and that the masking gave a slightly better result was seen when comparing the non-masked and masked barcodes for non-merged data with thresCC 0.85. In Figure 4.22 and 4.23, two bargroups generated from the assembly experiment using the masked non-merged data with thresCC 0.85 are shown. In both bargroups it can be seen that some of the barcodes are non-continuous, indicating that masking of the features of these barcodes was performed. These bargroups were not shown for the corresponding assembly experiment using the non-masked data, indicating that the masking removed the problematic bargroup with the high predominant peak. Instead, these two bargroups were generated out of the barcodes belonging to the problematic bargroup.

For the merged data, the results were the same and bargroups corresponding to the ones seen in Figure 4.22 and 4.23 were also generated with thresCC 0.85.

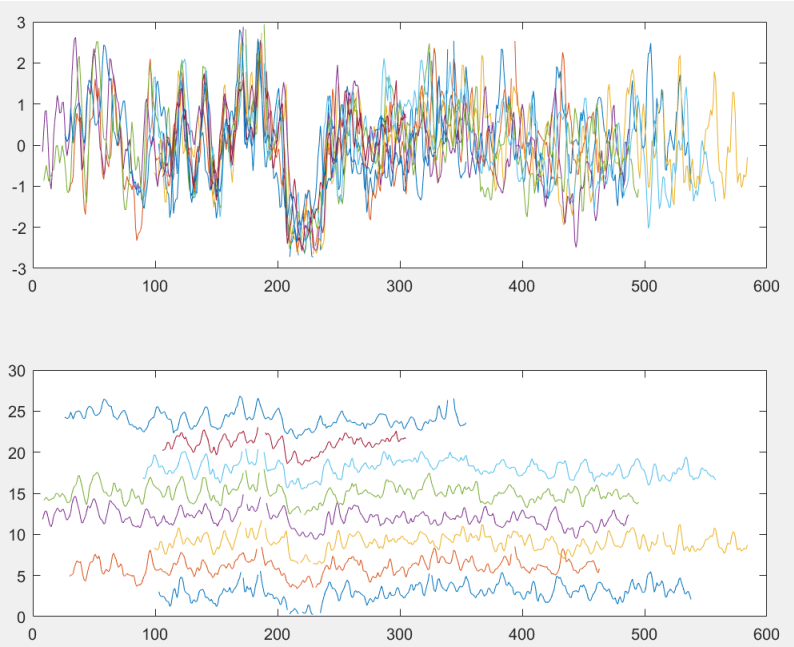

Figure 4.22: A bargroup generated using the masked non-merged data and thresCC 0.85. Upper part: Visualises the barcodes in the bargroup laying on top of each other. Lower part: Visualises the different barcodes belonging to the bargroup individually. The x-axis represents the size in pixels.

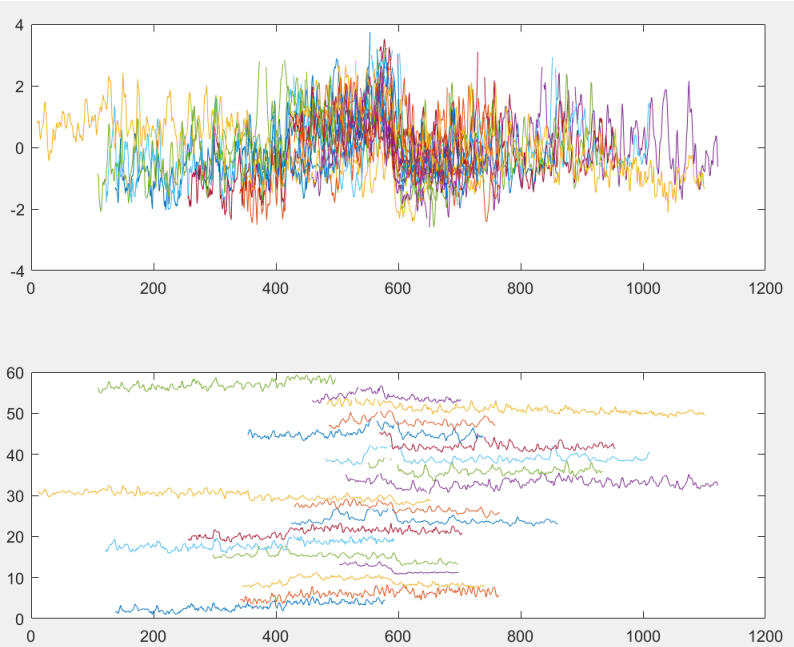

Figure 4.23: A bargroup generated using the masked non-merged data and thresCC 0.85. Upper part: Visualises the barcodes in the bargroup laying on top of each other. Lower part: Visualises the different barcodes belonging to the bargroup individually. The x-axis represents the size in pixels.

By comparing Figure 4.22 with the high quality bargroup in Figure 4.16, it can be seen that they are the same bargroup. However, the generated bargroup seen in 4.23, which was generated by masking and using thresCC 0.85, is not as good. In the lower part of the figure, it can be seen that some of the barcodes match to each other in pairs, probably being the neighbouring barcodes. This was confirmed by comparing this bargroup to the corresponding one generated by the masked merged data, not shown in this report. Otherwise, it is hard to see any similarities between the different barcodes in the bargroup.

To conclude the assembly experiments, the same bargroups were generated from the assembly experiments with the different thresholds using both the non-merged and merged data, only with a small difference regarding the barcodes in the bargroups. Also, when lowering the threshold, different bargroups were generated. However, lowering the threshold resulted in a larger bargroup containing the predominant peak to which every barcode with a high peak could be wrongly assembled to. What could be seen is that masking could be a potential solution for improving the assembly of the barcodes, as it resulted in an increase in number of bargroups both for non-merged and merged data, indicating that the problematic bargroup with the high peak was divided into smaller bargroups. In other words, after analysing the barcodes generated from the masked data, it could be said that the masking improved the bargrouping to an extent.

4.5 Outlook

As could be seen in the result, the assembly experiments did not generate 16 bargroups representing the chromosomes of *S. cerevisiae*, which is the goal with the *de novo* assembly. To reach this goal, future work around parameterisation needs to be performed. The different bargroups that are generated in the assembly experiments depend on the settings of different parameters. Currently, it is not known what parameters and especially which threshold to use to generate and obtain several high quality bargroups at one assembly experiment.

Additionally, future work regarding the assembly experiments is to perform the reference-based alignment (HCA) with the obtained high quality bargroups from the assembly experiments performed with different thresholds and data. If this would be done, the bargroups would be merged into one barcode based on the average of each pixel. The barcode representing the bargroup would then be compared to the reference chromosomes to see if there is a good match, and thereby find out which chromosome the bargroup represents or belong to. A next step could also be to perform a refining assembly in which each of the obtained bargroups are merged into an average barcode, and a second assembly with the merged barcodes are performed.

However, the assembly is challenging and problematic since it is done from scratch and there are no standards to compare with. In any case, when a successful *de novo* assembly has been performed generating a complete reference map of the genome of *S. cerevisiae*, further studies are to investigate mutants of the same strain of the yeast. By conducting ODM of mutants and comparing the results with the reference map of the strain BY4742 generated through assembly, an investigation of structural variations in the mutants could be performed. The findings could in turn detect SVs and provide significant knowledge regarding the underlying mechanisms of different diseases.

5
Closing Remarks

To our knowledge, and to this date, this project is the first time optical DNA mapping performed using competitive binding and a nanofluidic chip is applied on the genome of *S. cerevisiae*. The results from the reference-based alignment, especially the coverage plots of the 16 chromosomes and the coverage depth and extent of coverage, indicate that the data collected is of good quality and enough to be able to perform *de novo* assembly of the genome of *S. cerevisiae*. However, some unexpected results were seen in the reference-based alignment, especially regarding chromosome XII and for chromosome I, which could indicate complex regions located in these chromosomes or currently unmapped regions.

Results from the assembly experiments indicate that the assembly pipeline works, since high quality bargroups were obtained with certain parameters, and could potentially assemble the whole genome divided in 16 bargroups. However, we are not there yet. Improvements on the assembly are needed to obtain 16 bargroups corresponding to the 16 chromosomes and get extended information and understanding about the genome of *S. cerevisiae*. The idea is that a *de novo* assembly of the generated barcodes could provide the intact chromosomes and thereby assist sequencing techniques in finding the correct sequence of the intact chromosomes.

An assembly of the genome would be highly valuable and of great use. Not only for detecting structural variations which could cause disease in individuals, but also for assisting in assembly of sequences of the genome, and thereby cover regions of the genome hard to map with the current sequencing techniques. ODM could in other words be a major driving force to find the correct sequence of the genome.

This project is a first step towards a *de novo* assembly of the genome of *S. cerevisiae* using optical DNA mapping. With more work improving and automating the assembly pipeline, there is a great chance that the data collected in this project is able to generate a reliable *de novo* optical genome assembly of the yeast *Saccharomyces cerevisiae*. In turn, this could contribute to deeper knowledge of the *S. cerevisiae* genome and be of great use in the research field regarding both yeast and genetics.

Bibliography

[1] Müller V. Optical Mapping of Bacterial Plasmids: Method Development and Applications. Department of Biology and Biological Engineering, Chalmers University of Technology; 2017.

[2] Dremsek P, Schwarz T, Weil B, Malashka A, Laccone F, Neesen J. Optical Genome Mapping in Routine Human Genetic Diagnostics—Its Advantages and Limitations. Genes. 2021;12(12):1958.

[3] Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature Reviews Genetics. 2006;7(2):85-97.

[4] Weckselblatt B, Rudd MK. Human structural variation: mechanisms of chromosome rearrangements. Trends in Genetics. 2015;31(10):587-99.

[5] Müller V, Dvirnas A, Andersson J, Singh V, Kk S, Johansson P, et al. Enzyme-free optical DNA mapping of the human genome using competitive binding. Nucleic acids research. 2019;47(15):e89-9.

[6] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. The Universal Features of Cells on Earth. In: Molecular Biology of the Cell. 4th edition. Garland Science; 2002. .

[7] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. The structure and function of DNA. In: Molecular Biology of the Cell. 4th edition. Garland Science; 2002. .

[8] Goldman AD, Landweber LF. What is a genome? PLoS genetics. 2016;12(7):e1006181.

[9] Shibata JH. Nucleic Acids: Structures, Properties, and Functions (Bloomfield, Victor A.; Crothers, Donald M.; Tinoco, Ignacio, Jr.; contributions from Hearst, John E.; Wemmer, David E.; Kollman, Peter A.; Turner, Douglas H.). ACS Publications; 2001.

[10] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Chromosomal DNA and its packaging in the chromatin fiber. In: Molecular Biology of the Cell. 4th edition. Garland Science; 2002. .

[11] Annunziato A. DNA packaging: nucleosomes and chromatin. Nature Education. 2008;1(1):26.

[12] Reisner W, Pedersen JN, Austin RH. DNA confinement in nanochannels: physics and biological applications. Reports on progress in physics. 2012;75(10):106601.

[13] Yuan Y, Chung CYL, Chan TF. Advances in optical mapping for genomic research. Computational and Structural Biotechnology Journal. 2020;18:2051-62.

[14] Muller V, Nyblom M, Johnning A, Wrande M, Dvirnas A, Kk S, et al. Cultivation-free typing of bacteria using optical DNA mapping. ACS infectious diseases. 2020;6(5):1076-84.

[15] Michaeli Y, Ebenstein Y. Channeling DNA for optical mapping. Nature biotechnology. 2012;30(8):762-3.

[16] Jeffet J, Margalit S, Michaeli Y, Ebenstein Y. Single-molecule optical genome mapping in nanochannels: multidisciplinarity at the nanoscale. Essays in Biochemistry. 2021;65(1):51-66.

[17] Günther K, Mertig M, Seidel R. Mechanical and structural properties of YOYO-1 complexed DNA. Nucleic acids research. 2010;38(19):6526-32.

[18] Larsson A, Carlsson C, Jonsson M, Albinsson B. Characterization of the binding of the fluorescent dyes YO and YOYO to DNA by polarized light spectroscopy. Journal of the American Chemical Society. 1994;116(19):8459-65.

[19] Kundukad B, Yan J, Doyle PS. Effect of YOYO-1 on the mechanical properties of DNA. Soft matter. 2014;10(48):9721-8.

[20] Berman HM, Neidle S, Zimmer C, Thrum H. Netropsin, a DNA-binding oligopeptide structural and binding studies. Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis. 1979;561(1):124-31.

[21] Murray NE, Gann A. What has phage lambda ever done for us? Current Biology. 2007;17(9):R305-12.

[22] Burssed B, Zamariolli M, Bellucco FT, Melaragno MI. Mechanisms of structural chromosomal rearrangement formation. Molecular Cytogenetics. 2022;15(1):1-15.

[23] Gao H, Xu H, Wang C, Cui L, Huang X, Li W, et al. Optical Genome Mapping for Comprehensive Assessment of Chromosomal Aberrations and Discovery of New Fusion Genes in Pediatric B-Acute Lymphoblastic Leukemia. Cancers. 2023;15(1):35.

[24] Neveling K, Mantere T, Vermeulen S, Oorsprong M, van Beek R, Kater-Baats E, et al. Next-generation cytogenetics: Comprehensive assessment of 52 hematological malignancy genomes by optical genome mapping. The American Journal of Human Genetics. 2021;108(8):1423-35.

[25] Sudiyani Y, Prastya ME, Maryana R, Triwahyuni E, et al. The Budding Yeast Saccharomyces cerevisiae as a Valuable Model Organism for Investigating Anti-Aging Compounds. In: Saccharomyces. IntechOpen; 2021. .

[26] Duina AA, Miller ME, Keeney JB. Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system. Genetics. 2014;197(1):33-48.

[27] Parapouli M, Vasileiadis A, Afendra AS, Hatziloukas E. Saccharomyces cerevisiae and its industrial applications. AIMS microbiology. 2020;6(1):1.

[28] Yates GT, Smotzer T. On the lag phase and initial decline of microbial growth curves. Journal of Theoretical Biology. 2007;244(3):511-7.

[29] Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, et al. The reference genome sequence of Saccharomyces cerevisiae: then and now. G3: Genes, Genomes, Genetics. 2014;4(3):389-98.

[30] Nijkamp JF, van den Broek M, Datema E, de Kok S, Bosman L, Luttik MA, et al. De novo sequencing, assembly and analysis of the genome of the laboratory

strain Saccharomyces cerevisiae CEN. PK113-7D, a model for modern industrial biotechnology. Microbial Cell Factories. 2012;11:1-17.

[31] Yue JX, Li J, Aigrain L, Hallin J, Persson K, Oliver K, et al. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nature genetics. 2017;49(6):913-24.

A

Appendix

A.1 MATLAB Settings for DBM

Figure A.1 presents the settings that were used for the DBM tool when processing the data collected with the nanofluidic chip. The custom MATLAB script generating the DBM tool was provided by Tobias Ambjörnsson's group at the Department of Astronomy and Theoretical Physics at the University of Lund.

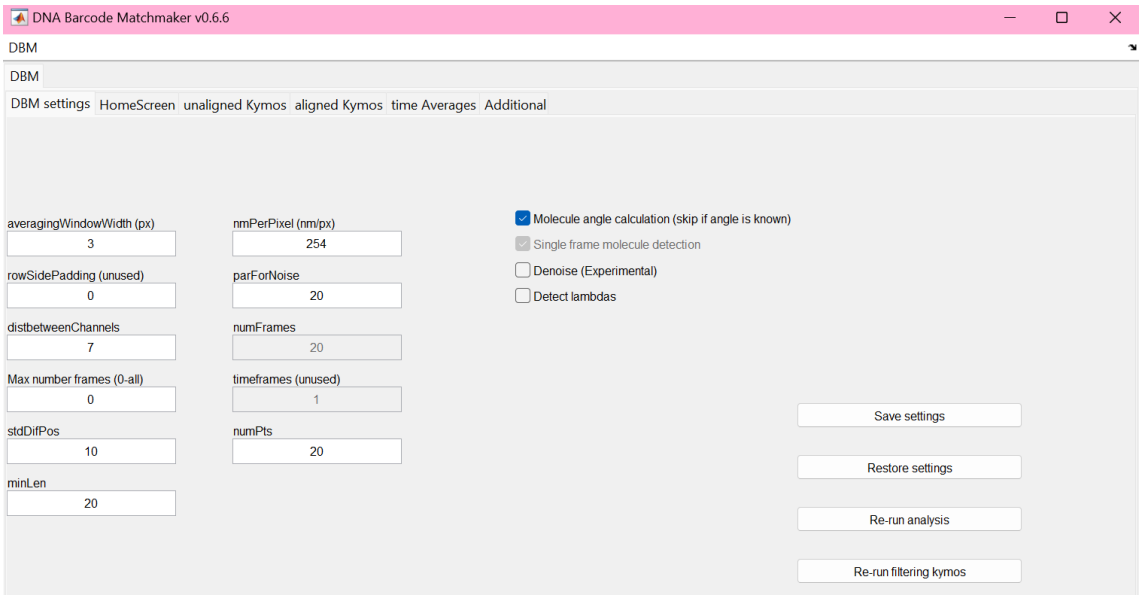

Figure A.1: Settings used in MATLAB for the DBM tool when processing data collected with the nanofluidic chip.

The pixel size (nmPerPixel in the figure), given in nm/px, was set to 254 nm/px based on the fluorescence microscope used in the project for collecting data using the nanofluidic chip. However, when DBM was performed on the data collected with the multiplex device using the Zeiss AxioObserver.Z1 microscope, the pixel size was set to 130 nm/px.

The other parameters seen in the figure were set to their default values, except the parameter numPts which was set to 20.

A.2 MATLAB Settings for HCA & Reference-Based Alignment

Table A.1 presents the settings that were used for the HCA tool, the in-house algorithm for chromosome alignment, generated by a custom MATLAB script provided by Tobias Ambjörnsson's group at the Department of Astronomy and Theoretical Physics at the University of Lund.

Table A.1: Settings used in MATLAB for the HCA tool.

Variable	**Setting**
nmPerPixel (nm/px)	254
Number of time frames	Default
Alignment method	nralign
Minimum length (px)	150
Overlap for merging (px)	100
nm/bp ratio for experiments	0.259
Stretch factor	0.15
Step	0.01

The pixel size (nmPerPixel in the figure), was set to 254 nm/px based on the fluorescence microscope used in the project for collecting data using the nanofluidic chip. The number of time frames was set to default, indicating that all time frames of each molecule were used for the data processing and analysis. The alignment method used for the reference-based alignment was nralign, the default setting. A size threshold of the kymographs was used, resulting in kymographs being shorter were filtered out. The minimum length, given in pixels, that was used as the size threshold was set to 150 px, and the overlap used for merging of neighbours was set to 100 px. The nm/bp ratio for the experiments that was used was 0.259, which was based on the average length of the measured λ-DNA molecules in every imaging session. Finally, a stretching factor of 15% (0.15), as well as a step of 1% (0.01) were used when performing HCA and the reference-based alignment, indicating that every percentage of stretching or shrinking of the barcodes between 85% and 115% was used for the analysis.

Other parameters not mentioned in the table were set to their default values.

A.3 Settings of Parameters in the MATLAB script for Assembly experiments

Table A.2 presents the settings used for the MATLAB script performing the assembly experiments, generated and provided by Tobias Ambjörnsson's group at the Department of Astronomy and Theoretical Physics at the University of Lund.

Table A.2: Settings for the parameters used in the MATLAB script for assembly experiments.

Parameter	**Setting**
ThresCC	0.8, 0.85, 0.9
Minimum overlap (px)	120
Stretch factor	0.15
Step	0.01

Three different values were used for the threshold called thresCC in the assembly experiments, which is an internal score in the MATLAB script that filters between good and bad bargroups. The values used were 0.8, 0.85 and 0.9. The minimum overlap used for the assembly experiments was set to 120 px and the stretching factor used was 15% (0.85 - 1.15) and a step of 1% (0.01).

Other parameters for the assembly experiments not mentioned in the table were set to their default values.

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

CHALMERS
UNIVERSITY OF TECHNOLOGY