Estimation and Modelling of Errors in the Library Preparation Stage of Next Generation Sequencing

Nathan Beka

Submitted to the University of Hertfordshire in partial fulfilment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

University of Hertfordshire

June, 2020

Acknowledgements

I would like to express my profound gratitude and thanks to my supervisors for their continuous guidance and support throughout my PhD studies. Dr Rene te Boerkhorst for being ever present and sharing his vast knowledge which truly guided me on this journey. Professor Rod Adams for his excellent suggestions which always helped me improve the quality of my work. Dr Maria Schilstra and Dr Neil Davey for all their brilliant contributions and support.

I must also thank my brilliant friends from the UH Biocomputation group who were truly kind for offering their friendship and support during my studies. Also, I would like to thank the staff of the Department of Computer Science for their immense support and the opportunities they have provided me over the years.

On a personal note, I am ever grateful to my parents, Dr Francis Beka and Mrs Shawn Beka, who have provided great support and guidance to me throughout my life. I thank my siblings, Jason, Patrice, Francis and especially Sylvia, for their immense love, care, and support all through my life.

Most importantly, I am thankful to God for the boundless grace and mercy that has been offered to me during the years of my research and throughout my life.

Abstract

Next-generation sequencing has empowered genomics by making it possible to sequence genomes at a lower cost and less time compared to the traditional Sanger method. However, these improvements suffer from reduced accuracy when compared with the Sanger method. During the library preparation stage of sequencing, artefacts can be introduced that affect the reliability of a read. These artefacts can arise from biases due to the structure of the genome, such as preferential splitting of DNA between specific nucleotides, bias of adapter ligation towards certain base pair identities, and temperature dependent denaturation due to nucleotide composition. To investigate these issues a library preparation model was developed to simulate the occurrences and investigate effects of such artefacts. The implemented model simulates the DNA fragmentation, adapter ligation and PCR amplification stages of the library preparation process. A set of parameters characterizing these steps and a DNA sequence are used as input and the output is an array of values representing the number of DNA fragments that cover each position of the input sequence ("coverage"). To validate the model a Genetic Algorithm (GA) was used to find parameters that would lead to coverage values that are closely similar to what is found in empirical sequencing data. The GA was able to acquire such parameters for a subsection of the *Mycobacterium tuberculosis* and *Plasmodium falciparum* genomes but failed when applied to the TP53 gene of the *Homo sapiens* genome. From this it was deduced that the model was better at predicting coverage when applied to genomes with subregions of nucleotide repeats. To find the effects of parameters representing each step of the library preparation process the model was applied to a set of *in silico* generated DNA that represent different sequence structures (GC-rich, AT-rich, neutral composition and a sequence with specific areas of GC and AT rich repeats). My study found that the parameters for the fragmentation, adapter ligation and PCR steps affected coverage. I also found that a combination of parameters between consecutive steps further affected coverage. In the fragmentation step, large fragment size had a negative effect on coverage ($p = 0.0$), in the adapter ligation step, coverage of AT-rich sequences was affected by a terminal bias ($p = 0.0$). Modifying parameters for the PCR step affected the coverage of both GC and AT rich sequences due to a temperature dependent bias. Finally, an interaction between the parameters of fragmentation and other steps were found to further reduce coverage. This simulation was able to suggest parameters that need to be fine-tuned to improve coverage.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Motivation

Next generation sequencing (NGS) of DNA has dramatically transformed approaches to genomic and genetic research (Oyola *et al.*, 2012). DNA sequencing refers to a laboratory method used to determine the sequence of a DNA molecule. Some of the well-known technologies that are applied in this process include the Roche GS-FLX 454 Genome Sequencer (originally 454 sequencing), the Illumina Genome Analyser (originally Solexa technology), the ABI SOLiD analyser, Polonator G.007, Helicos HeliScope, Pacific Biosciences Single Molecule Realtime (SMRT) sequencing and the Oxford Nanopore Technologies sequencing platforms.

These technologies (also referred to as massively parallel sequencing technologies) have enabled the sequencing of DNA at unprecedented speeds (Zhang *et al.*, 2011) compared to the "original" sequencing methodology known as the Sanger method (Sanger, Nicklen & Coulson, 1977). Although NGS has revolutionised biology by increasing current understanding of many genes and mutations involved in the pathogenesis of human diseases (Zhang *et al.*, 2011), there are still challenges associated with the use of these new technologies.

For example, the sequencing of parts of a genome characterized by extremely biased base composition is still a great challenge to the currently available NGS platforms (Oyola *et al.*, 2012). The genomes of certain important pathogenic organisms like *Plasmodium falciparum* and *Escherichia coli* for instance are characterised by noticeable high-AT content and high-GC content respectively. The degree with which a sequencing technology covers such regions of the genome (called "coverage") and hence the reliability of the output, can be affected by a number of artefacts introduced at various stages of the sequencing process.

In sequencing, coverage refers to how much of the sequenced or targeted genomic region is covered by "reads". A read is the fundamental unit of output of the sequencing process, and refers to the base identity (A, T, G or C) that corresponds to a single nucleotide position. During the first stage of the sequencing process, a DNA sample is broken (ideally at random) into several fragments, which at a later stage can be reassembled computationally. The number of fragments covering a read position quantifies the reliability of that read. This number is referred to as depth of coverage and is measured for a single genomic position as the number of reads aligned to that position. Thus, coverage for each position in a DNA sample is equal the number of reads aligned to that position.

A perfect sequencing method should provide an end to end reading of a genome, and accurately identify variant structures of interest such as polymorphisms and mutations. However, in reality, the length of reads are short and contain errors which can be misidentified as sequence variants. These errors can be introduced at different stages of the sequencing process. For example, as will be reported in Section 2.6.4, during the PCR stage of sequencing (the stage at which DNA fragments are "cloned" to ensure a sufficiently large sample size of copies), fragments with a high content of A and T can be destroyed by the increase in temperature essential for this process. This results in low coverage of AT-rich regions of the genome and may lead to situations where errors are misidentified as sequence variants, polymorphisms and mutations, resulting in false conclusions in studies. Therefore, it is important to assess the uniformity of coverage by calculating the variance in coverage across a sequence. This will ensure even coverage, and detection of regions with low coverage, thus leading to the production of higher quality reads (Sims, et al., 2014). Furthermore, in DNA resequencing where genetic variations are explored in relation to diseases in humans, accurate detection of variants is essential. This accuracy is affected by low quality reads and non-uniform coverage. Increased coverage can counteract these effects and improve the detection of variants.

Coverage can be improved by increasing the amount of DNA to be sequenced, but this would lead to higher sequencing costs. Arguably, it is more economical to invest in studies into sequencing errors that cause reads of poor quality because of low and uneven coverage. Such research may yield improved error detection and methods for the differentiation between such errors and actual DNA variants.

To investigate the effects of sequencing errors, a sequence of systematic and controlled experiments in which a single sequencing parameter is varied while keeping the others constant can be carried out. Followed by a statistical analysis of the outcomes to check the individual effect of each parameter on coverage. But the problem with this approach would be that it ends up being more cost-intensive than increasing the volume of DNA used for a sequencing run due to the repetition involved. An alternate but cost-effective solution to this would be to simulate the sequencing process by implementing a virtual laboratory that carries out the process *in silico*.

This kind of simulation can be helpful in studying the interaction between sequencing steps without worrying about shortcomings from the hardware and errors from the experimenter. It can be seen as an “agent-based model” where each sequencing step is characterised by a set of properties which represent their real-world parameters. Agent-based models (ABMs) have their origins in artificial life which is the study of man-made systems that model the behavior of natural living systems (Aguilar *et al.*, 2014; Boden, 1996). An ABM is a model that is composed of agents. Each agent is an autonomous individual element with its own properties and actions that exist in a computer simulation. A system can be modeled using agents, an environment, the interactions between the agents as well as interactions between the agents and the environment. In artificial life models of living systems are translated to computational algorithms using ABMs as they are able to characterise the properties of these living systems through computation (Langton, 1997). They have been extensively used to model different phenomena, for instance in computational biology to model gene regulatory networks (Wang *et al.*, 2009), in business to model consumer behaviour (Huiru *et al.*, 2018) and in ecological studies to model population dynamics (Arifin *et al.*, 2014). However, these models are not a perfect representation of an original process and would not capture all of the original complexities involved but the results derived from them can serve as an advisory for changes that can be made in an empirical process.

The main advantage to using this approach here is that if the model is able to reliably simulate the real sequencing method, the cost constraint of increased DNA volume is solved and it is now easier to investigate how varying parameters that represent different sequencing steps affect coverage. In my research, I focus on the initial stage of the sequencing process called library preparation. This stage is a step-by-step process that prepares a DNA sample for the sequencing hardware. It includes the following steps: Fragmentation, End-repair, A-tailing, Adapter ligation and PCR amplification. Previous studies indicate that errors and biases due to parameters from the PCR step (van Dijk, et al., 2014) and Ligation step (Seguin-Orlando, et al., 2013) among others, can negatively affect uniform coverage of a sequence. Using a simulation of the library preparation process it becomes possible to vary the parameters for each step to find their consequential effects on coverage. The discoveries from this process can inform wet-lab researchers of what steps of the library preparation procedure need to be investigated in more detail. Thus, a study of problems that can arise from the library preparation stage of the sequencing workflow and affect final sequencing output will form the basis of my work.

1.2 Objectives

The aim of this research is to analyse how artefacts that may occur during DNA library preparation affect sequencing coverage. I will first identify the stages of library preparation as used in the Illumina NGS platform. This will provide an in-depth understanding of the procedure and the necessary knowledge I will need to model the identified stages.

Subsequently, artefacts that can occur due to biological effects (biases) and experimental design during library preparation will be identified. This will give insight to how, why, and when these artefacts occur. With this knowledge the identified artefacts are modelled and introduced at their related stage during library preparation.

Finally, with the completion of a simulated library preparation platform, integrated models of these artefacts will be applied to track their individual and combined effects on sequencing output. Research questions of this study include:

1. What artefacts can occur at the different stages of library preparation?
2. How do these artefacts affect sequence coverage?
3. Are these effects additive, if not how do they combine?

1.3 Contributions to Knowledge

The contributions to knowledge in this research include:

1. A virtual platform that simulates the fragmentation, ligation, and PCR stages of library preparation in Illumina sequencing, which allows researchers to study the effects of the parameters representing these library preparation steps and their affiliated artefacts on coverage.
2. A genetic algorithm validating that the coverage resulting from a library preparation model can match those from actual DNA sequencing.

3. Identification of artefacts and values of parameters representing steps of library preparation and statistical confirmation of their effects on the uniformity of coverage. These effects depend on the base composition and degree of serial dependency (e.g. nucleotide repeats) of the sequences considered.
4. A demonstration of how preceding library preparation steps combine with the succeeding steps affect coverage. The parameters of the fragmentation step were found to interact with those of the ligation and PCR steps. I present suggestions on the cause of this occurrence and a possible solution to reduce its effect on coverage.

1.4 Thesis Outline

The structure of this thesis is as follows:

Chapter 2: Background – An overview of the structure and function of DNA is provided along with a summary of DNA sequencing technologies. This is followed by a step by step walkthrough of the different stages of Illumina NGS and a review of possible artefacts that can occur during these stages. Finally, existing tools which can be used to simulate sequencing reads are evaluated. The information in this chapter provides the background knowledge required to understand the rest of this thesis.

Chapter 3: Modelling the Library Preparation of NGS – This chapter provides a description of the model for simulating the library preparation process and its implementation along with the metrics and tools that will be used to measure its output.

Chapter 4: Matching Model Outcomes with Results of Real Sequencing using a Genetic Algorithm – Background of genetic algorithms and the implementation of a genetic algorithm to establish optimal parameter values for fitting model output to the real coverage found for DNA samples of *Plasmodium falciparum, Mycobacterium tuberculosis* and *Homo sapiens*.

Chapter 5: Effects of Library Preparation – Statistical analysis of the effects of the implemented parameters. The first part informs about the statistical effects of each single implemented parameter on coverage uniformity for *in silico* generated DNA sequences with different nucleotide composition and sequential dependency. The second part deals with the effects of preceding library preparation stages on subsequent stages and their combined effect on coverage. The chapter concludes with a validation of the effect of the model parameters applied to the real DNA sequences introduced in Chapter 4.

Chapter 6: Conclusions – The conclusions drawn from the validation of the model and the results of my analysis and possible extensions of the work done are outlined.

Chapter 2 Background

This chapter provides the background needed to understand the development of the proposed library preparation model and results derived from experiments carried out using it. After a short overview of the structure and function of DNA (section 2.1), the second section (2.2) covers the basics of DNA sequencing. Next generation sequencing with focus on the Illumina sequencing platform is the topic of sections 2.3 and 2.4. A step by step explanation of the stages of the sequencing process is given in section 2.5 followed by a description of artefacts and biases that can occur during the library preparation stage (section 2.6). At the end of the chapter software tools that have been used to simulate the sequencing process are reviewed (section 2.7).

2.1 Deoxyribonucleic acid

Deoxyribonucleic acid (DNA) is a nucleic acid that encodes the genetic information in all organisms. A DNA molecule consists of two polynucleotide chains that form a spiral called a double helix and is made up of units called nucleotides. These units come in two types: purines and pyrimidines. The purines are Adenine ("A") and Guanine ("G"), and the pyrimidines are Cytosine ("C") and Thymine ("T"). Nucleotides are linked sequentially to each other by phosphor-sugar bridges (see Figure 1).

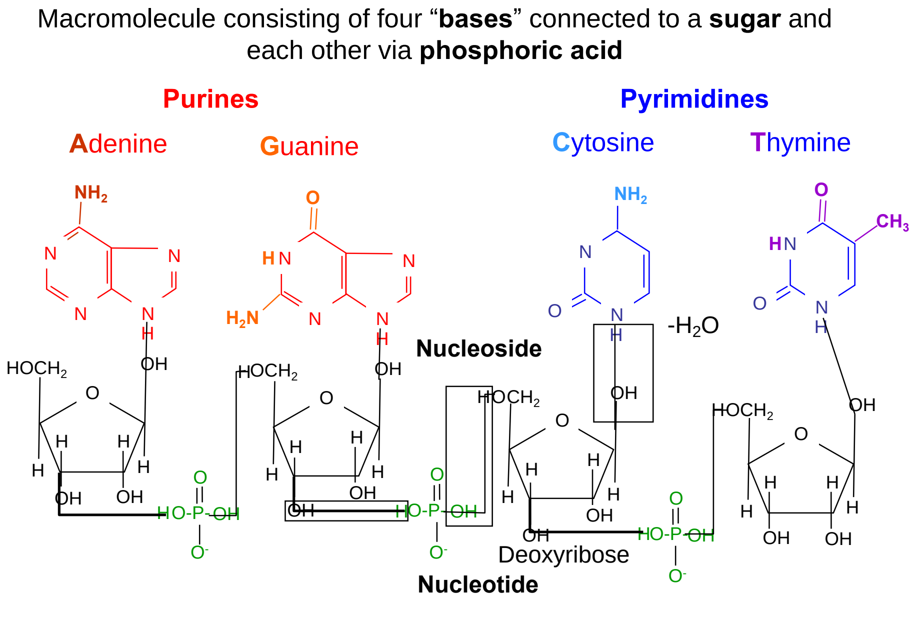

Macromolecule consisting of four "bases" connected to a sugar and each other via phosphoric acid
Purines
Pyrimidines
Adenine
Guanine
Cytosine
Thymine
Nucleoside
-H2O
Deoxyribose
Nucleotide

Figure 1: DNA structure

The double helix is held together by complementary base pairing of A to T and C to G by means of hydrogen bonds, consequently forming units called base pairs.

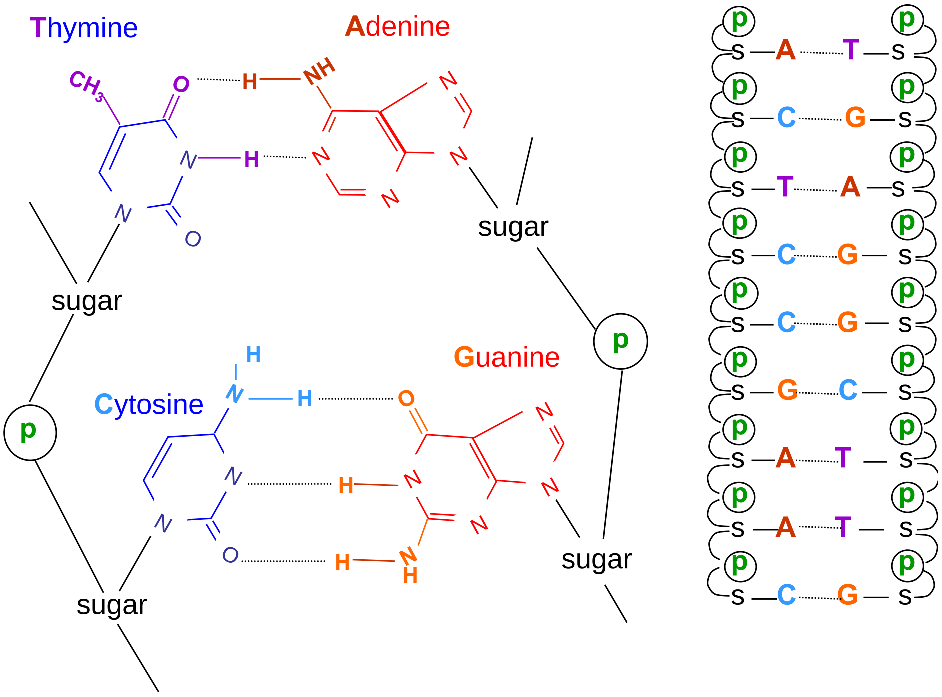


Figure 2: Complementary base pairing

DNA molecules are organized into a thread-like structure called chromosomes. A chromosome is made up of a DNA molecule, which is tightly wrapped around histones, a particular type of protein.

A gene is a segment of DNA that acts as a unit of hereditary information and is situated at a locus, a specific position at a chromosome. Genes are essential for the synthesis of proteins, which are very important in the structure and functioning of all living organisms. Just as DNA is a macromolecule composed of a particular sequence of four types of building blocks, proteins are a build-up of twenty types of basic units called amino acids.

Genes are composed of protein-coding parts (exons) separated by non-coding structures (introns); in turn, exons consist of a series of three nucleotides (called codons). To form a protein, codons are "transcribed" into a complementary strand called messenger RNA. During this transcription, thymine is replaced by uracil. The messenger RNA acts as a template to which another type of RNA, transfer RNA connects temporarily. Transfer RNA contains two attachment sites. The first one is called the anticodon; it consists of three nucleotides that are complementary and therefore attach to a messenger RNA codon. Amino-acids, the type of which is specified by the particular nucleotide sequence of the anticodon, dock on to the second site.

Subsequently, the bonds between the messenger RNA and transfer RNA are broken and the amino acids are linked together to form a polypeptide chain. These are then folded into a protein.

The function and structure of a protein are determined by the way it is folded in three dimensions, which in turn depends on the sequence of its amino-acids and hence on the sequence of the nucleotides in the exons. Changes in the latter, for instance by the transformation of a particular nucleotide into another, are called mutations. Mutations may cause disruption of the structure and function of a protein.

Several mutations can occur in DNA sequences, including, but not limited to, point mutations (Single Nucleotide Polymorphisms or SNPs), insertions and deletions. SNPs are the most common type of mutation and a number of them are known to be associated with particular diseases. With the help of DNA sequencing, mutations can be detected in a genome. This allows for the prediction of, among other things, the susceptibility for certain diseases.

2.2 DNA Sequencing

The first revolution in DNA sequencing occurred in the 1970s with the development of the Maxam-Gilbert chemical degradation method (Maxam & Gilbert, 1977) and the Sanger enzymatic dideoxy method (Sanger, Nicklen & Coulson, 1977). The majority of DNA sequencing technology today relies on variations of the Sanger method (Sanger, Nicklen & Coulson, 1977). The Sanger method also known as the chain-termination method involves the use of dideoxynucleotides (ddNTPs) in combination with deoxynucleotides (dNTP’s). The key difference between ddNTP’s and dNTP’s is the presence of a hydrogen group on the 3’ carbon rather than a hydroxyl group (OH). When these modified ddNTPs are integrated into a sequence, they inhibit the addition of further nucleotides (Obenrader, 2003). This is caused by the inability of the ddNTP to form a phosphodiester bond with the next nucleotide of a growing DNA chain, leading to the termination of the chain (Sanger, Nicklen & Coulson, 1977) (See Figure 3).

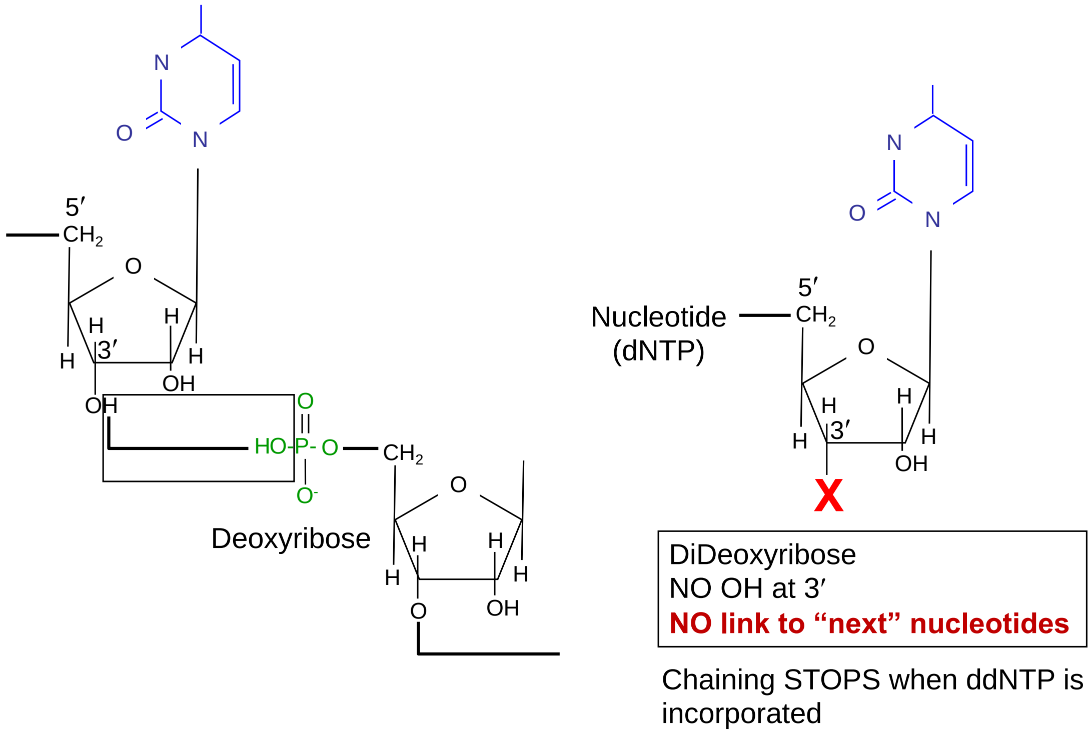


Figure 3: Chain Termination due to ddNTP

With the inception of the Human Genome Project in 1990, faster sequencing technologies were developed, thereby providing significant improvements to DNA sequencing over the years (Church, 2005). These revolutionary advances paved way for the invention and commercial introduction of the first massively parallel sequencing platform in 2004, so-called Next Generation Sequencing (NGS) (Mardis, 2008). This innovation heralded the era of high throughput genomic analysis (Next Generation Sequencing or NGS). The broadest application of NGS may be the resequencing of human genomes to enhance our understanding of how genetic differences affect health and disease (Metzker, 2010).

2.3 Next Generation Sequencing

Next Generation Sequencing (NGS) has empowered genomics by making it possible to sequence genomes at a lower cost and in less time compared to the traditional Sanger method (Sanger, Nicklen & Coulson, 1977). The latter was used in the Human Genome Project (Lander *et al.*, 2001; Venter *et al.*, 2001), which took about three years. Nowadays, with the use of high throughput NGS, a human genome can be sequenced within a week.

The first commercially available (2005) NGS platform was the Roche/454 FLX Pyrosequencer which uses a pyrosequencing sequencing technology (Margulies *et al.*, 2005). Following this, the Illumina sequencing platform was released in 2006. Illumina technology is based on the sequencing by synthesis method (Mardis, 2008). A year later "Sequencing by Oligo Ligation Detection" (SOLiD) was developed by Life Technologies (Valouev *et al.*, 2008). The Ion Semiconductor Sequencing technology was developed by Ion Torrent (now a subsidiary of Life Technologies) in 2010. An important distinction of this technology is its use of semiconductor technology rather than optical detection of nucleotides using fluorescence making it quicker, cheaper and smaller than previously mentioned platforms (van Dijk *et al.*, 2014). Several other NGS platforms have been developed including Helioscope Single Molecule Sequencer by Helicos Biosciences (Pushkarev, Neff & Quake, 2009), Single-molecule real-time sequencing commercialized by Pacific Biosciences in 2011 (van Dijk *et al.*, 2014), Polony sequencing (Porreca, Shendure & Church, 2006), and DNA nanoball sequencing by Complete Genomics (Drmanac *et al.*, 2010).

NGS technologies rely on a complex combination of enzymology, chemistry, optical sensors (excluding Ion semiconductor sequencing technology), hardware and software (Ledergerber & Dessimoz, 2011). Each platform requires raw genomic material to go through a series of stages to produce a DNA sequence. These stages are broadly classified as the library preparation, the imaging and sequencing, and the data analysis phases (Metzker, 2010). The final step in the sequencing process, known as base calling, involves using software to identify individual bases.

The ability to sequence a whole genome offered by these technologies has resulted in an abundance of comparative and evolutionary studies that were previously not possible. NGS has been applied to several areas of biology, which include mutation detection, alternative splicing, microRNA profiling, and mapping of protein DNA interactions (Wang *et al.*, 2012). Despite the revolutionary advancements Next Generation Sequencing technologies have brought to sequencing, they still fall short when compared to the traditional Sanger method, due to reduced accuracy and shorter read lengths (Ledergerber & Dessimoz, 2011).

Given the impact of the conclusions drawn from DNA sequencing. Quality control to check the trustworthiness of the applied sequencing technology is therefore of utmost importance; its lack of may have led to the publication of erroneous results in previous studies that are still accepted and used as the basis for further research.

It is the harmful influence of side-effects generated at the library stage, such as the shearing of the source DNA into fragments of a particular length, that forms the basis of this dissertation. The methodology chosen as the subject of my research is Illumina dye sequencing, which is described in the next section.

2.4 Illumina dye sequencing

The Illumina dye sequencing method was originally developed by Shankar Balasubramanian and David Klenerman of Cambridge University, who originally utilized this technique in their Solexa sequencing platform, which was acquired by Illumina in 2007 (Bharagava *et al.*, 2019). This technique utilizes the sequencing by synthesis (SBS) method wherein a DNA sample is sheared into a large number of fragments. Subsequently, specific short sequences ("adapters") are joined to the fragments that enable them to be attached to a physical substrate (the flow cell). Prior to this, the fragments are (amplified) into a large number of clones by a process called polymerase chain reaction (PCR). Once attached to the flow cell, each fragment will again be "amplified" to form a cluster of identical subsequences (this process is called cluster amplification). A single cluster contains roughly one million copies of the original fragment, which sufficiently reports incorporated bases (nucleotides) at a reliable signal intensity for detection during sequencing (Mardis, 2008). Following cluster amplification, a reaction mixture is added to the flow cell, which contains primers[1], DNA polymerase[2] and four terminator nucleotides each labelled with a fluorescent dye. Next, the terminator nucleotide is identified by its fluorescent dye using a CCD (charge-coupled device) camera (Ansorge, 2009). At the end of the imaging step, the reaction mixture is washed away and the cycle is repeated (Mardis, 2008). The synthesis step is repeated for a specific number of cycles as required by the user. Following the sequencing run a base calling algorithm is used to assign sequences and allocate quality scores to each read (Mardis, 2008). All Illumina sequencing platforms (MiniSeq, MiSeq, HiSeq, NovaSeq, etc.) are based on this method (Bharagava *et al.*, 2019).

[1] A primer is a string of nucleotides that serves as the starting point for DNA replication (Princeton.edu, n.d.).

[2] DNA polymerase is an enzyme which is responsible for the replication of DNA (Nature.com, n.d.).

2.5 Stages of Illumina sequencing

This section outlines the stages required for a sequencing run using the Illumina/Solexa platform.

2.5.1 Library preparation

This stage of Illumina sequencing involves preparing a sample of double-stranded genomic DNA. First, the DNA sample is fragmented into smaller pieces, then an end-repair of the fragments is carried out to remove uneven ends. Following this, A-tailing is performed to allow ligation of adapters to the ends of the fragments. Then a size selection step is performed to select fragments of the required size and to remove un-ligated adapters. Finally, PCR amplification is carried out, to increase the representation of fragments in the library. Each stage is explained in more detail below. See **Figure 4** for a graphical outline of the process.

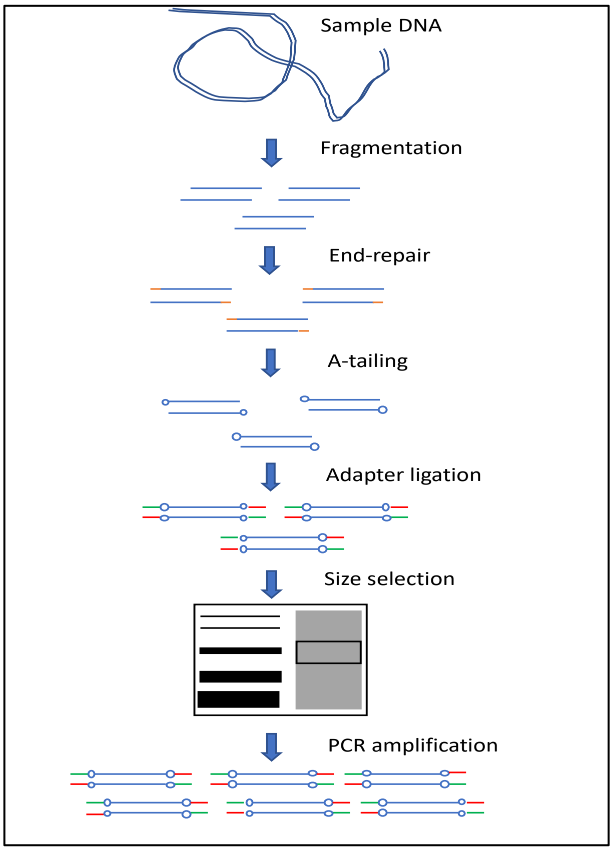


Figure 4: Library preparation workflow.

DNA fragmentation

The first step of library preparation involves the breaking down of sample DNA to fragments of a desired size. This is typically achieved by using mechanical, enzymatic or chemical fragmentation methods (Head *et al.*, 2014). Several techniques can be used to carry out these processes including sonication, acoustic shearing, nebulization and enzymatic shearing.

Sonication involves subjecting DNA samples to ultrasonic waves, the vibrations from the waves produce gaseous cavitations in the medium that contains the DNA sample. The cavitation implodes and the energy that is released by this shears high molecular weight DNA molecules (Knierim *et al.*, 2011). The acoustic shearing method fragments a DNA sample by focusing high frequency, short wavelength energy on the sample. The size of fragments produced using this protocol is controlled by modifying the intensity and duration of the acoustic waves(Apone, Dimalanta & Stewart, 2017). The main difference between sonication and acoustic shearing is the frequency at which they operate. Sonicators operate at a low frequency which leads to long ultrasonic wavelengths. The high frequency used in acoustic shearing produces shorter wavelengths that enable a higher level of precision in the shearing process (Covaris, 2012).

In nebulization breaking up a DNA sample involves forcing it through a tiny hole, using compressed nitrogen or air resulting in random sheared DNA fragments. The size of the resulting fragments is dependent on the preset gas pressure used to force the DNA through the hole (Knierim *et al.*, 2011; New England Biolabs, 2014).

Enzymatic based methods involve the incorporation of two enzymes into a volume of DNA. One of the enzymes generates a partial break in the double-stranded DNA while the other completes the breakage at the opposite end. The end result of this process is a volume of double-stranded DNA fragments (Knierim *et al.*, 2011; New England Biolabs, 2014).

End-repair

DNA fragments from the fragmentation process usually end up with varying 3’ and 5’ overhangs (see **Figure 5**). The end-repair process converts the resulting overhangs to blunt ends using a mixture of enzymes comprising *E. coli* DNA polymerase, T4 DNA polymerase and Klenow enzyme (Illumina, 2008; Son & Taylor, 2011). This enzyme mixture catalyses high 3’ to 5’ exonuclease activity and 5’ to 3’ polymerase activity. The former removes the 3’ overhangs while the latter fills in the 5’ overhangs (Bankier, 2001) (See **Figure 6**). This process prepares the fragments for addition of an A (adenine) base to their 3 ends.

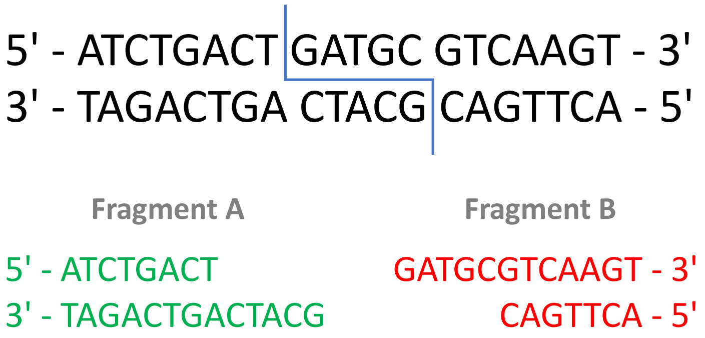


Figure 5: Visual description of overhangs.

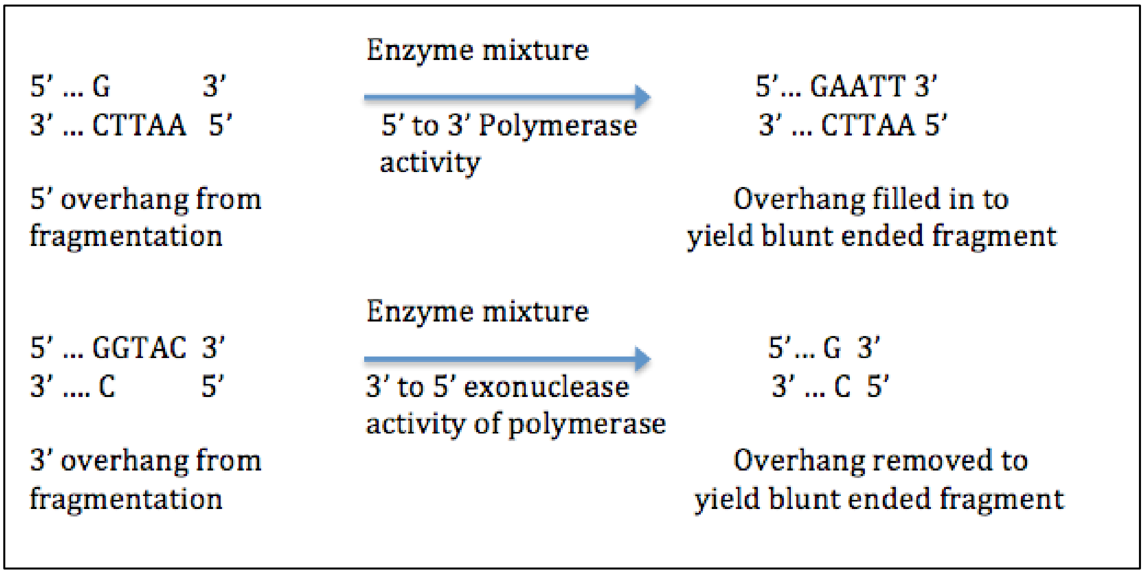

5’ ... G 3’
3’ ... CTTAA 5’
5’ overhang from fragmentation
Enzyme mixture
5’ to 3’ Polymerase activity
5’... GAATT 3’
3’ ... CTTAA 5’
Overhang filled in to yield blunt ended fragment
5’ ... GGTAC 3’
3’ .... C 5’
3’ overhang from fragmentation
Enzyme mixture
3’ to 5’ exonuclease activity of polymerase
5’... G 3’
3’ ... C 5’
Overhang removed to yield blunt ended fragment

Figure 6: Polymerase activity fills in 5 overhangs and exonuclease activity removes 3’ overhangs.

A-tailing

A-tailing involves the addition of an A base to the 3’ ends of a fragment (see **Figure 7**). This is done to enable the ligation of T-tailed adapters and to prevent the formation of concatemers[3]. The modification is carried out by adding the end-repaired DNA fragments to a reaction mix containing Klenow buffer, dATP (deoxyadenosine triphosphate) and Klenow fragment (Son & Taylor, 2011; Illumina, 2008). Polymerase

[3] A concatemer is a DNA molecule that consists of multiple copies of the same DNA attached together sequentially (Kutter, 2001).

activity of the Klenow fragment adds an A base to the 3’ end of the DNA fragments (Illumina, 2008).

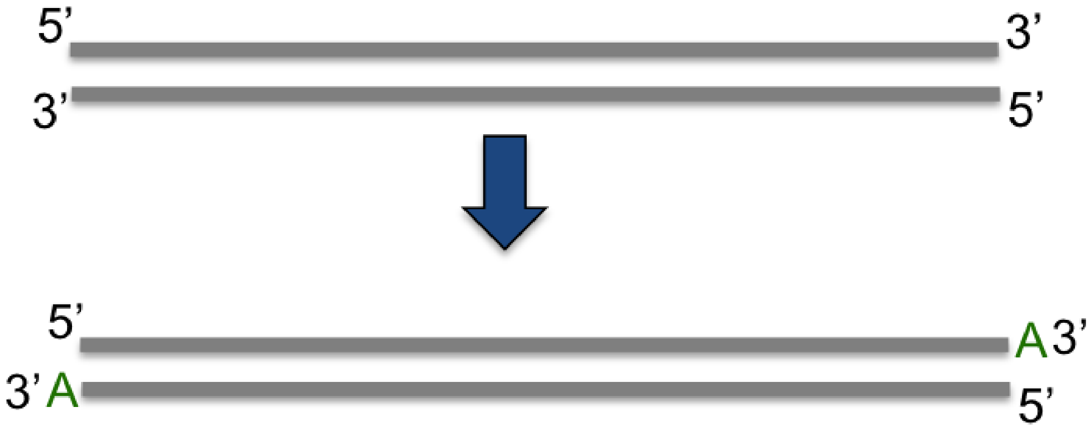


Figure 7: Adding adenine to 3' ends (Labster.com, 2014)

Adapter ligation

To sequence the DNA library, adapters would need to be ligated to the modified fragments. The function of the adapters is to connect the fragments to a flow cell (the substrate on which actual sequencing is performed). In addition, they are required for the cluster amplification stage. Adapter ligation adds distinct (adapter) sequences to DNA fragments by creating a phosphodiester bond between the 3’ end of the fragments and 5’ end of the adapter sequence (see **Figure 8**). The reaction is catalysed using T4 DNA ligase enzyme, which facilitates the ligation of both ends (Gaastra & Hansen, 1984).

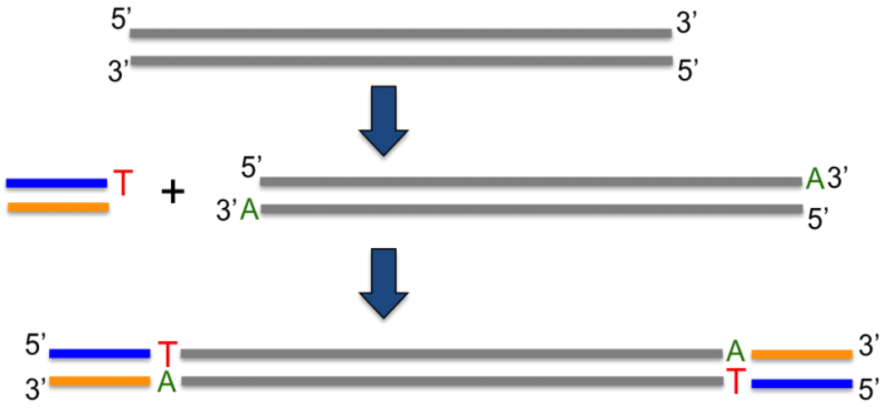


Figure 8: Ligation of adapter to DNA fragment.

Size Selection

After the ligation of adapters to the fragments, a purification process is required to remove un-ligated adapters and adapters that may have ligated to each other, and select a size range of fragments for the library which would be appropriate for the cluster amplification step (Illumina, 2008). The purification can be carried out using gel electrophoresis and band excision or solid-phase reversible immobilization (SPRI) beads (Bronner *et al.*, 2009). Gel electrophoresis is a process whereby DNA fragments are separated by size in an agarose gel. This process is carried out by loading DNA samples into slots made in the gel, and then an electric current is applied to the top (negative end) of the gel, which causes the negatively charged DNA molecules to move towards the bottom (positive end) of the gel. The smaller fragments move faster and end up at the bottom of the gel. A fluorescent dye is also added to the gel, making it easier to visually track the movement of the DNA fragments across the gel using ultraviolet light. At the end of the process, the desired size of DNA is excised from the gel. The desired size may vary depending on the protocol being followed (Carr, 2012; Roberts & Dryden, 2013). **Figure 9** shows an example of gel electrophoresis.

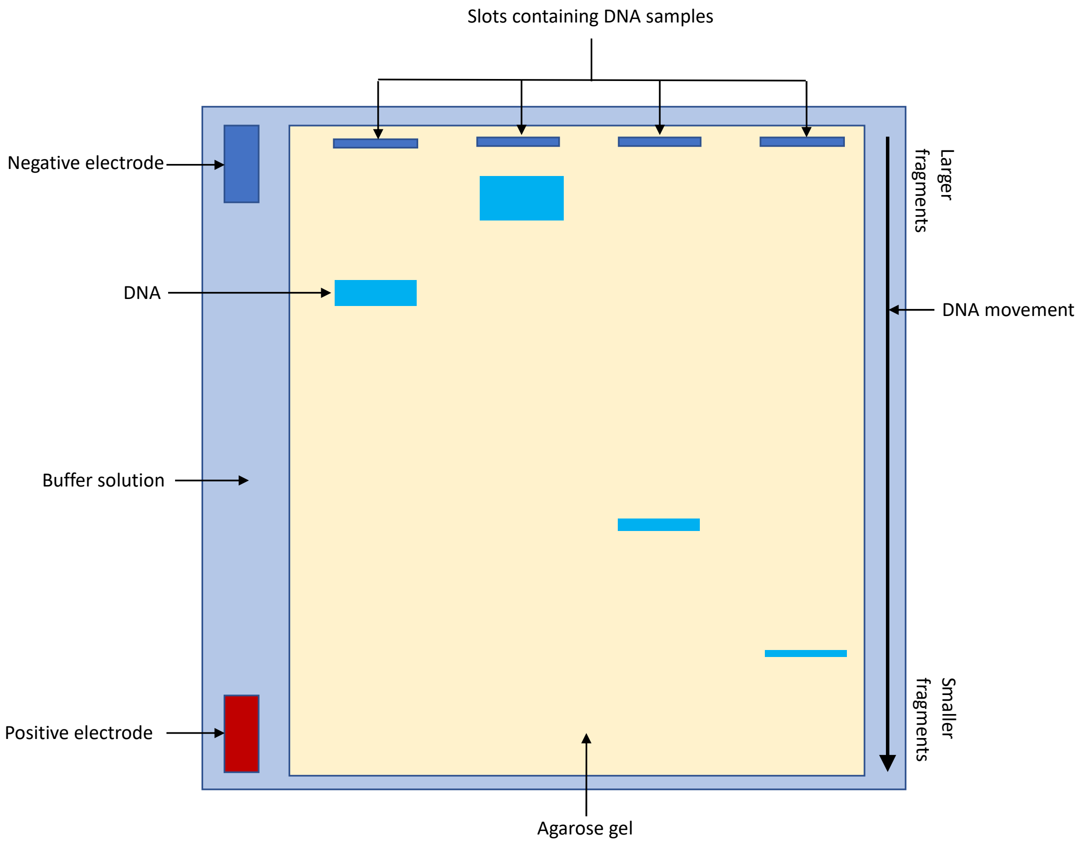


Figure 9: Gel electrophoresis

PCR amplification

The penultimate step in the library preparation process is the amplification of DNA fragments that have adapters ligated to both ends. Polymerase chain reaction (PCR) is a technique to increase the sample size by several orders of magnitude by recurrently cloning the DNA fragments. PCR is carried out by heating and cooling a reaction mixture containing primers, dNTPs (see section 2.2) and DNA polymerase repeatedly. The cycle begins with heating the reaction mixture to about 93°C (temperature could vary depending on the library preparation protocol), which denatures the target DNA into two strands (Chantler, 2004; Illumina, 2011a) (Figure 10a). The temperature is then reduced to allow the primers to attach to the separated DNA strands (Figure 10b). Next, the temperature is increased to enable DNA polymerase to elongate the primers by attaching complementary bases to the strand (Figure 10c). Finally, the PCR product is denatured from the initial DNA strand. The second and third stage are repeated with the primer annealing to the template strands and newly cloned strands enabling elongation by DNA polymerase (Chantler, 2004). Subsequent cycles are carried out until the required sample size is achieved.

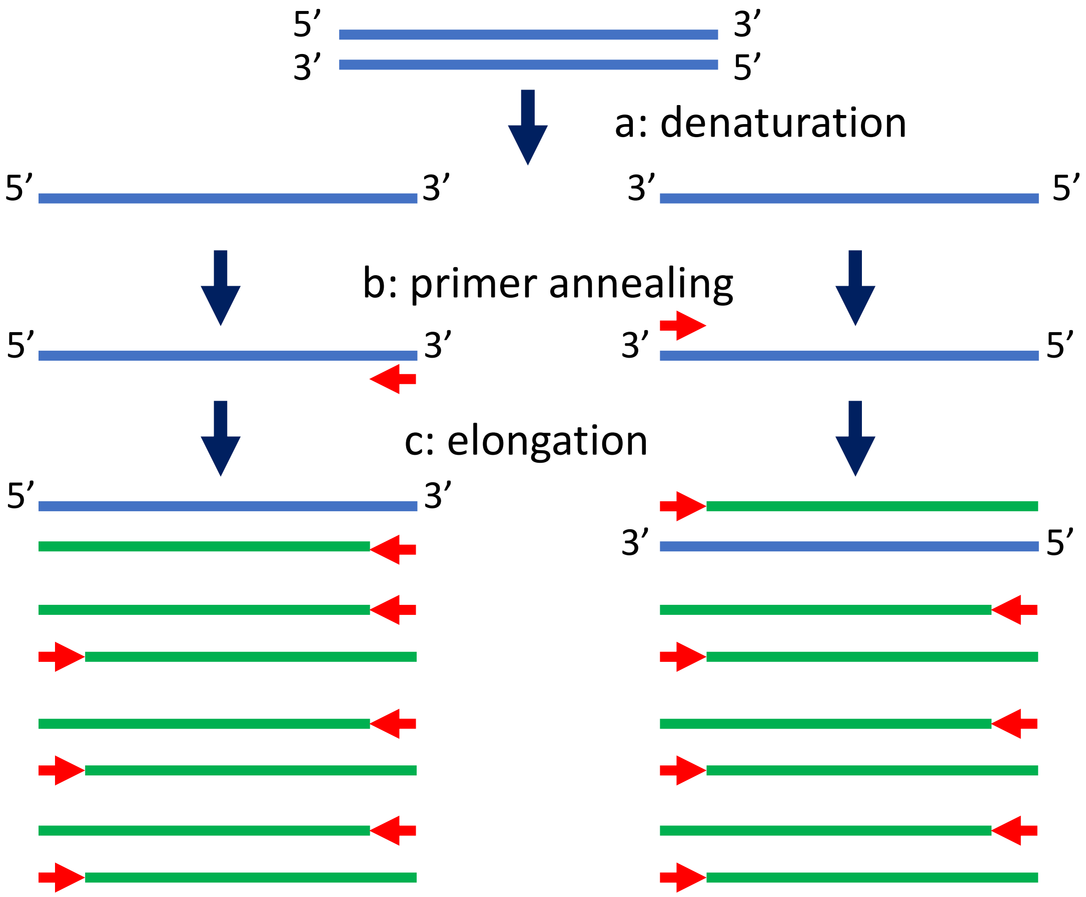


Figure 10: PCR cycle

Library quantification

The final stage of library preparation is a quality control measure, which is recommended by Illumina. It involves verifying the size of the PCR enriched fragments and checking DNA fragment size distribution. To validate the size range of the enriched fragments (which ideally should be the same as it was during the purification stage), gel electrophoresis is carried out on 10% of the volume of the library. Illumina also recommends quantification of the sample library using qPCR (quantitative real-time PCR). This is done to ensure optimum clusters are generated for the lanes on the flow cell (Illumina, 2011a) (see section 2.5.3). If an excessive amount of DNA is loaded on to the flow cells, generated clusters will overlap into adjacent lanes causing a reduction in the quality of sequencing data. If an insufficient amount of DNA is loaded, the generated clusters would have a reduced density, thereby reducing the efficiency of resulting sequencing data (Buehler *et al.*, 2010).

2.5.2 Library denaturation

The product of the library preparation phase is a double-stranded DNA library. To hybridize individual strands of DNA to primers on the flow cell the library is denatured. The denaturation is accomplished by incubating the library in sodium hydroxide (Quail, Swerdlow & Turner, 2009; Illumina, 2011a). Alternatively, the library could be denatured by heating but this could present bias issues with AT-rich fragments and GC rich fragments (Quail, Swerdlow & Turner, 2009). These bias issues are further discussed in section 2.6.

2.5.3 Cluster Amplification

Cluster amplification transforms libraries into clonal clusters on the surface of a flow cell. The Illumina flow cell is a glass slide with microfluidic channels, which dNTPs, polymerases, and buffers flow through (Figure 11). The surface of a flow cell is coated with oligonucleotides, which are complementary to the sequences of the adapters ligated to DNA fragments during library preparation. During cluster amplification, single-stranded fragments are connected by hybridization to the oligonucleotides on the flow cell (Figure 12a). To ensure that only one end of a fragment hybridizes, the sequence of one of the ligated adapters is the reverse complement to the oligonucleotides.

This end is attached to an adjacent oligonucleotide creating a bridge (Figure 12b). To begin the process, the attached fragments are copied using DNA polymerase, creating a reverse strand of the original fragments (Figure 12c). The double-stranded fragments are then denatured, and the ends are freed allowing them to attach to oligonucleotides on the flow cell once again (Figure 12d). This process is repeated several times. Finally, the reverse strands are washed away leaving dense clusters of matching strands derived from the original fragments (Quail, Swerdlow & Turner, 2009) (Figure 12e).

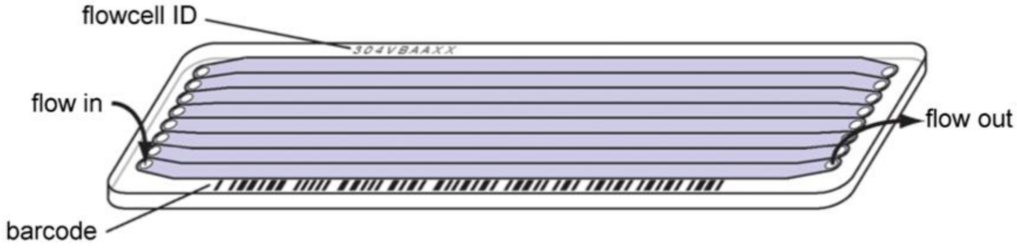


Figure 11: Illumina flow cell (Quail, Swerdlow & Turner, 2009)

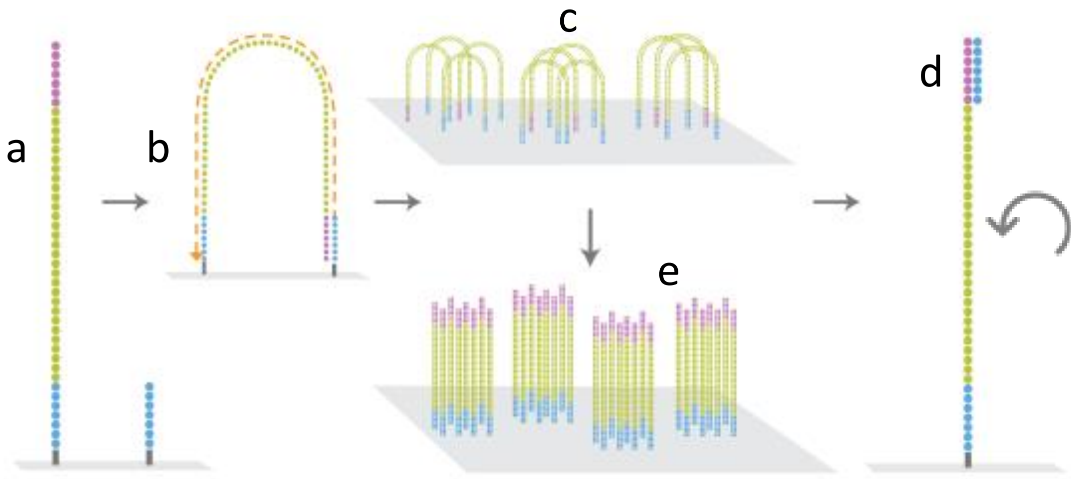


Figure 12: Cluster amplification process. Modified from (CeGaT, 2014).

2.5.4 Sequencing by synthesis

To initiate the sequencing by synthesis process, a flow cell containing millions of clusters is loaded into the sequencer. The first step involves the addition of a polymerase enzyme with the four nucleotides (A, C, G, T). Each nucleotide has a unique fluorescent marker and a "terminator" – to prevent the incorporation of additional nucleotides after the first complementary nucleotide is attached to be read (Mardis, 2008) (Figure 13A). After this addition, the clusters are excited by a light source and an image of the flow cell is captured using fluorescence microscopy (Figure 13B). Following the imaging step, the terminator and fluorescent markers are washed away for the next base to be incorporated (Figure 13C). This cycle is repeated several times until all fragments are read (Figure 13D and E). The number of cycles carried out makes up the length of a "read".

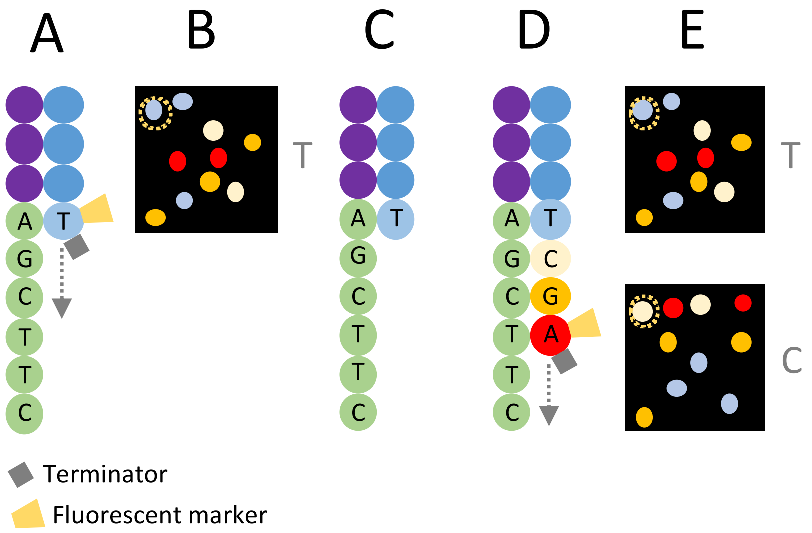


Figure 13: Sequencing by synthesis

There are two types of reads in Illumina sequencing, namely, single-end reads, and paired-end reads. When single-end reads are utilized the DNA fragment is sequenced from one end to the other as seen in Figure 13. With paired-end reads the DNA fragment is sequenced from both ends (Figure 14). Single-end sequencing is the simplest, fastest and most economical way to sequence a DNA sample. Paired-end sequencing due to its ability to read form both directions produces a larger number of reads, thus improving accuracy and enabling enhanced detection of variations in DNA structure (Illumina, 2017; Nakazato, Ohta & Bono, 2013).

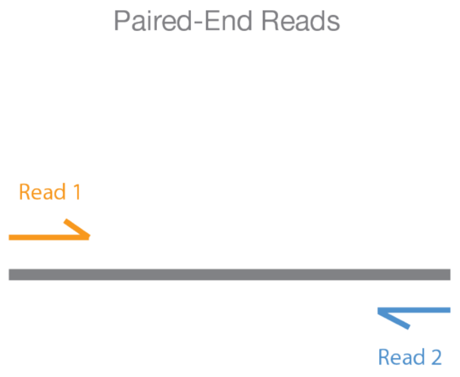


Figure 14: Paired-End Reads

2.5.5 Base calling

"Base calling" is the name of the process, which determines the identity of a base (A, C, G, T) during a sequencing cycle (Illumina, 2011b). To this end specific algorithms (e.g. the "Bustard" base caller) are employed that classify bases in accordance to the fluorescence of the highest intensity.

2.5.6 Quality scoring

A quality score is used to predict the probability of an error in a base call. To assign a quality score a set of quality predictor values are computed. These quality predictor values are observable traits, such as fluorescence intensities of the clusters on the flow cell. The values derived are assigned to a quality table, which relates them to quality scores. This relationship is determined by a calibration process where reads are aligned to a reference genome[4] to confirm the identity of a called base.

2.5.7 Sequence assembly

Sequences generated at the end of the sequencing process are formed of many short reads, which need to be put back together to represent a whole genome. This process is carried out using dedicated algorithms called assemblers. Assemblers work by finding overlapping fragments and from these reconstructs a whole genome. Although this process may sound simple, it comes with its own challenges that may lead to erroneous sequences (Naumenko *et al.*, 2018). For example, DNA segments with repeated nucleotide sequences (so-called "repeats") produce similar or even identical

[4] A reference genome is an already known genome, which is sequenced from several individuals using different sequencing platforms to ensure accuracy.

fragments, which may be from different parts of a genome and are therefore difficult to pinpoint to a location (Nagarajan & Pop, 2013). A follow-on process to assembly is finishing. In this process assembled data are checked and edited to correct any errors if found (Baxevanis & Ouellette, 2004).

2.6 Artefacts in sequencing

Next generation sequencing methods are not completely accurate. They are prone to errors which could lead to miscalled bases causing misaligned reads and mistakes in sequence assembly (Robasky, Lewis & Church, 2014). Errors can arise at various stages of the sequencing process leading to poor quality sequencing output. The table below lists errors derived from literature. Each of the errors are discussed in more detail below.

Table 1: Sequencing artefacts

Sequencing stage	Artefact	Source
Fragmentation	Sequence-dependent cleavage bias	(Grokhovsky, 2006)
	Oxidative DNA damage	(Costello *et al.*, 2013)
End repair		
A-Tailing		
Ligation	Ligation bias	(Seguin-Orlando *et al.*, 2013)
Size Selection	None	
PCR	Amplification Bias	(van Dijk, Jaszczyszyn & Thermes, 2014)
	Slipped strand mispairing	(Fazekas, Steeves & Newmaster, 2010)
	Chimera formation	(Sharifian, 2010)
Cluster Amplification		
Sequencing-by-synthesis	Phasing and Pre-phasing	(Kircher, Heyn & Kelso, 2011)
	Crosstalk	(Li & Speed, 1999)

2.6.1 Sequence-dependent cleavage bias

The fragmentation stage of library preparation produces short DNA fragments by cleavage of a DNA sequence at (supposedly) random positions. However, a study by Grokhovsky et al. (2006), revealed that fragmentation by sonication resulted in a biased cleavage rate between cytosine and guanine in 5’ - CpG - 3’[5] dinucleotides (**Figure 15**). Furthermore, this bias depends on the flanking sequences; it is stronger when both strands contain a mix of purines (A, G) and pyrimidines (C, T) but weaker if the flanking sequences consist of just purines in one of its strands (Grokhovsky, 2006).

Also, a subsequent study by Grokhovsky *et al.* (2008) found that cleavage commonly occurred at the 3’ side of cytosine. The cleavage intensity increased in the order CG > CA = CT > CC. The CA, CT, CC steps have a higher cleavage rate than their complementary steps (TG, AG, GG). The unequal cleavage rate of bases at opposite strands results in overhangs. This is not the case with CG pairs, as its complements have the same identity, which may be the reason for increased cleavage at this position (Grokhovsky *et al.*, 2013).

The mechanism that leads to this bias is the sequence-dependent variation in the serial structure of the nucleotide chain which in turn is controlled by the carbon structures that hold the nucleotides together (Grokhovsky *et al.*, 2011; Poptsova *et al.*, 2014). The increased level of reads in GC-rich areas of the genome is generally attributed to PCR (Benjamini & Speed, 2012). This may not always be the case, as the sequence dependence of cleavage points especially between C and G would lead to the majority of fragments coming from GC-rich areas of a genome. The effects of the splitting bias should be considered when sequencing as it may lead to biased outcomes. Modifications to experimental procedures and the addition of specific reagents may produce a solution to this bias (Poptsova *et al.*, 2014).

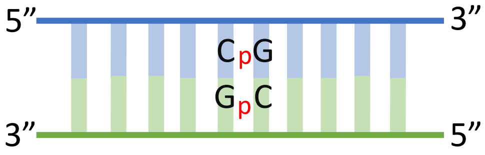


Figure 15: CpG dinucleotide

[5] p is a phosphate which links a pair of nucleotides together.

2.6.2 Oxidative DNA damage

Costello *et al.* (2013) discovered an unexpected high number of otherwise uncommon variants (C → A and G → T mutations) in certain cancerous tissues. These variants appeared to be specifically flanked by C and G (CCG → CAG). Following further inspection, the authors hypothesised that these variants were induced by artefacts in the library preparation or the sequencing process.

Surprisingly, the rate of occurrence of the C → A and G → T mutations varied between sequencing projects run at different laboratories. This instigated Costello et al. to analyse the sequencing projects carried out in their own lab. A comparison of the different sequencing chemistries used (Illumina HiSeq, MiSeq and Ion Torrent) showed no difference in the occurrence of the variants, suggesting the effect was induced before the sequencing stage. Going through the library preparation protocols, they found that DNA fragmented using high powered 150bp sonication showed a significant increase in the occurrence of the variants. However, the effect was only found in less than half of 150bp sonicated libraries implying that the fragmentation method on its own was not enough to explain the artefact. After comparing incoming DNA samples from other collaborating institutions, it was found that the aberration varied between collection sites and could be attributed to heat from high sonication energy in addition to contaminants in the DNA samples. The combination creates a highly oxidative environment leading to the conversion of guanine to 8-Oxoguanine (denoted as G*). 8-oxoG pairs with adenine, hence G(C) becomes G*(A) leading to C → A and (complementary) G → T substitutions (Cheng *et al.*, 1992).

2.6.3 Ligation bias

Another step of the library preparation process that can introduce bias is ligation. Seguin-Orlando *et al.* (2013) discovered failed ligation of adapters to fragments with a specific nucleotide on their 5’ and 3’ ends.

This effect suggests that ligation probability depends on the identity of the nucleotide at the beginning of a fragment and is specifically averse to fragments with a T on their 5’ end. Increasing the concentration of adapters during ligation reduced the loss of fragments. Unfortunately, the converse holds as well; reduction of the adapter concentration amplifies the loss of fragments, this is undesirable, because lowering the concentration of adapters is an effective way of diminishing the presence of adapter dimers[6].

Ligation bias can lead to an under-representation of AT-rich areas and over-representation of GC-rich areas of a genome, resulting in uneven coverage and hence a low sequencing quality.

2.6.4 Amplification bias

The PCR amplification process introduces a bias in sequencing coverage because not all fragments are amplified with the same efficiency (van Dijk, Jaszczyszyn & Thermes, 2014). Especially fragments with extreme base compositions (GC-rich or AT-rich) can be underrepresented or completely lost during library preparation (Aird *et al.*, 2011). This effect can cause difficulties when sequencing important organisms with unbalanced genomic base composition such as *Plasmodium falciparum* (AT-rich: 80% AT) and *Mycobacterium tuberculosis* (GC-rich: 65.6% GC).

Temperatures used during the denaturation and elongation steps of PCR have been shown to be responsible for this. Dutton et al. (1993) found that denaturing GC-rich fragments at 94°C led to a loss of such fragments due to incomplete denaturation. Su *et al.* (1996) reported that elongation of AT-rich fragments at 72°C after denaturation led to a loss of AT-rich fragments. This is most likely brought about by the strength of bonds between GC and AT pairs; GC pairs are held together with three hydrogen bonds, while AT pairs are bound by two. Because the number of bonds holding the pairs together and their neighbouring nucleotides determine the stability of a DNA double helix (Yakovchuk, 2006), the higher number of bonds in the GC pair requires a higher temperature to dissociate it from its template strand, whereas AT pairs can be separated at lower temperatures.

Several solutions have been suggested to lessen these effects: Dutton, et al. (1993) proposed a PCR protocol where the denaturation temperature was set at 98°C to improve the representation of GC rich fragments. The addition of betaine to the PCR reaction mix was also found to improve GC-rich fragment representation (Aird *et al.*,

[6] Adapter dimers are created when adapters ligate to themselves. These dimers can go through the sequencing process and take space on flow cells, thereby leading to reduced sequencing efficiency (Head *et al.*, 2014).

2011). However, although the addition of betaine reduces the melting temperature[7] and thus favours denaturation of GC-rich fragments, it negatively affects the elongation of AT-rich fragments.

In an empirical study, Su *et al.* (1996) found that a reduction of elongation temperatures to 60°C from the routinely used 72°C led to an improvement in coverage of AT-rich DNA. By reducing this temperature, AT-rich fragments were less likely to get denatured during elongation.

Kozarewa *et al.* (2009) proposed an amplification-free library preparation protocol that skips the PCR stage. By doing this they were able to achieve higher coverage for GC-rich sequences. This method uses the attached adapters from the ligation stage to directly adhere the fragments to Illumina flow cells for bridge amplification, hence eliminating the need for a PCR step. As this method is mainly dependent on the presence of adapters on each fragment, extra steps need to be taken to quantify the amount of fully ligated fragments to determine the portion of the DNA library that will be successfully sequenced. Due to the lack of PCR, which increases the representation for each fragment, this PCR-free method generally requires a larger volume of DNA to improve the representation of each fragment on the flow cells (van Dijk, Jaszczyszyn & Thermes, 2014; Oyola *et al.*, 2012; Kozarewa & Turner, 2011).

In a comparison of the enzymes used for PCR, Kapa Hifi (Kapa Biosystems) was found to provide better coverage across a genome than the routinely used Phusion polymerase (Quail *et al.*, 2012). The use of Kapa reduced the amplification bias and resulted in an improved coverage for both AT-rich and GC-rich fragments. Its improvements were close to those of the amplification-free library preparation protocol without the need for increased volumes of DNA.

Oyola *et al.* (2012) proposed an alternative PCR protocol which used Kapa in combination with tetramethylammonium chloride (TMAC). This reaction mix led to vast improvements in the coverage of extremely AT-rich genomes such as *Plasmodium falciparum*. The addition of TMAC improved the stability of AT base pairs (Chevet, Lemaitre & Katinka, 1995).

[7] The melting temperature (Tm) of double stranded DNA is the temperature at which half of the template strand is disassociated from its complementary strand.

2.6.5 Slipped strand mispairing

Slipped strand mispairing, also called Simple Sequence Repeats (SSRs), is an artefact of PCR amplification caused by repetitive nucleotide sequences. It usually occurs when a polymerase stalls elongation of a template strand due to nucleotide repeats. The polymerase dissociates from the strand and disrupts the base pairing process. This causes the template strand to form a loop in the repeat region, which results in the deletion of nucleotides in the loop when elongation is reinitiated. These repeats are mostly found in AT-rich genomes (Fazekas, Steeves & Newmaster, 2010).

2.6.6 Chimera formation

Chimera formation is another artefact produced during the PCR stage of sequencing. A chimera is a sequence composed of DNA from two or more sources (Zhang & Min, 2005). This artefact is caused by incomplete primer extension (Figure 15a). The partial elongation product can attach to a template strand as a primer in the next PCR cycle (Figure 15b). This will synthesise a new strand formed of the 2 template strands (Figure 15c), therefore creating chimeric DNA (Sharifian, 2010).

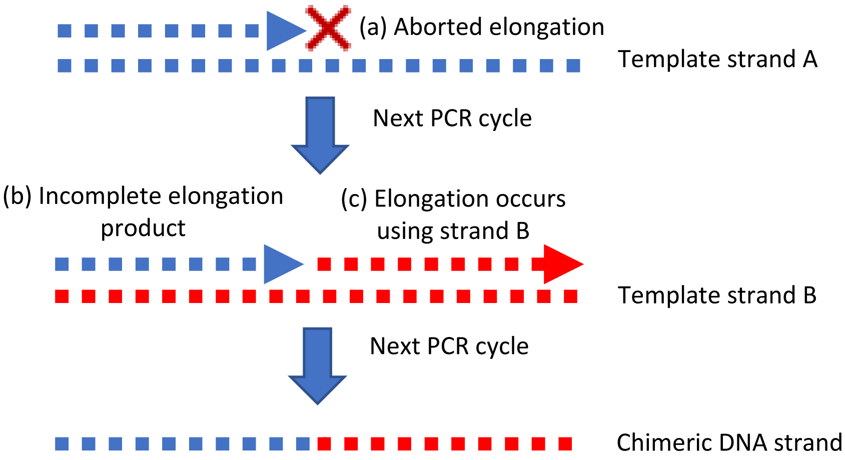


Figure 16: Chimera formation. Modified from (EzBioCloud, 2019)

Sections 2.6.7 and 2.6.8 discuss artefacts from the sequencing by synthesis stage of sequencing which is out of the scope of this thesis.

2.6.7 Phasing and Pre-phasing

Phasing and pre-phasing are errors, which are caused by inefficiencies of the chemistry during the sequencing by synthesis stage. Pre-phasing occurs when nucleotides without effective 3’ end terminators are incorporated in a cycle; this results in the continuous attachment of nucleotides and therefore skipping a base during base calling. Conversely, phasing occurs when 3’ end terminators and fluorescent markers are not washed out at the end of a cycle resulting in a failed incorporation during the next cycle. The failed incorporation causes the base call for that cycle to fall behind (Kircher, Heyn & Kelso, 2011).

2.6.8 Crosstalk

In sequencing by synthesis, the sequencer uses two lasers and four filters to excite and detect the dyes attached to each nucleotide. Frequency crosstalk occurs when the fluorescent dyes of the nucleotides overlap creating non-independent images for each base during a sequencing cycle. It is measured using a 4x4 matrix called a colour matrix. The matrix shows how the 4 nucleotides (A, C, G, T) crosstalk into the 4 spectral channels used for exciting the fluorescent dyes (Li & Speed, 1999).

2.7 NGS Read Simulation

Simulating the NGS process can aid researchers in planning sequencing experiments and testing hypothesis at a lower cost. With a simulator, several parameters of the sequencing process can be tested to find their outcomes without wasting resources on actual sequencing runs. Several computational tools have been developed that are able to generate NGS data. Here, I outline three of such tools and their functionality.

ART (Huang *et al.*, 2012) is a sequencing read simulator that supports the generation of Roche/454, Illumina and SOLiD reads. It utilises platform-specific and user-generated profiles to generate sequencing data. The user-customised profiles are able to generate sequencing data with custom read length and base call error characteristics. Specifically, it is able to model two types of sequencing errors: indels[8] and base substitutions. The characteristics for the errors are derived from empirical models built for each platform. The main error mode for its Illumina read simulation is base substitution. For Roche/454 read simulation, indels are the principal error type used. However, for SOLiD read simulation, the developer failed to state a dominant error

[8] An indel (insertion or deletion) is a genetic variation where a specific sequence of nucleotides is either present(insertion) or absent(deleted). (Rodriguez-Murillo & Salem, 2013)

type. ART is also able to emulate PCR amplification bias by specifying the number of reads for each copied fragment (Escalona, Rocha & Posada, 2016). The simulated reads for ART are returned as SAM and BED files.

The pIRS (Profile-based Illumina pair-end reads simulator) is an Illumina read simulator (Hu *et al.*, 2012). It generates Illumina reads using empirical base calling and GC%-depth (relationship between GC content and coverage depth) profiles. The GC%-depth profile enables the simulation of reads that have sequence-dependent coverage bias. Its empirical base calling profiles are derived from the analysis of sequence alignment results of known genomes. The tool also provides error profiles that are based on empirical models or can be user-generated. Errors modelled include indels, base substitution and single nucleotide polymorphisms (SNPs). A completed run of the tool generates results in the FASTQ file format.

GemSIM is an NGS read simulator supporting both Ilumina and Roche/454 reads (McElroy, Luciani & Thomas, 2012). It utilises empirical sequence-context based error models, fragment length and quality score distributions to simulate sequencing data. The tool consists of four modules: GemErr, GemHaps, GemReads and GemStats. GemErr is used to generate error models from real sequencing data using SAM format alignment data as input. GemStats is an optional module that generates statistics for the generated error models when simulating paired-end reads. It reports error rates for base positions and each nucleotide within a read. The GemHaps module accepts a DNA sequence, haplotype[9] frequency and the number of SNPs in the haplotypes. This input data is used to randomly generate SNP positions which can optionally be used for read generation. Finally, the GemReads module takes a FASTA file, error model generated by GemErr, a haplotype file generated by GemHaps and a species-abundance file when the GemSIM metagenomic mode is utilised. This data is processed and used to generate reads which are returned as FASTQ files.

The tools outlined here provide a lot of features and are efficient at their specific task of generating NGS reads. In my search for simulators, the functionality of most of the tools I found was solely focused on the sequencing-by-synthesis stage of NGS and its related errors. Three tools (ART, Flowsim, Grinder) did offer a simulation of PCR amplification, but I was unable to find any tools that mainly focused on the library

[9] A haplotype is a set of DNA variations on a chromosome that are usually inherited from a single parent (Silverman, 2007). These variants can be SNPs and alleles.

preparation stage of NGS, which introduces its own fair share of errors into sequencing data. In my work, I propose a tool, LpSIM, that simulates the library preparation stage of NGS and integrates some of the artefacts and biases that can occur at this stage. LpSIM and its implemented features are discussed in the next chapter.

2.8 Third Generation Sequencing

Third generation sequencing (TGS) is a newer iteration of sequencing technologies. Oxford Nanopore Technologies and Pacific Biosciences (PacBio) introduced platforms based on this technology in 2011.

The Oxford Nanopore platform utilizes nanopores immersed in an electrically resistant membrane to sequence a DNA sample. When an electrical charge is applied to the membrane the current only flows through the nanopore. The flow of current is then observed to determine the composition of DNA in a molecule (Figure 17).

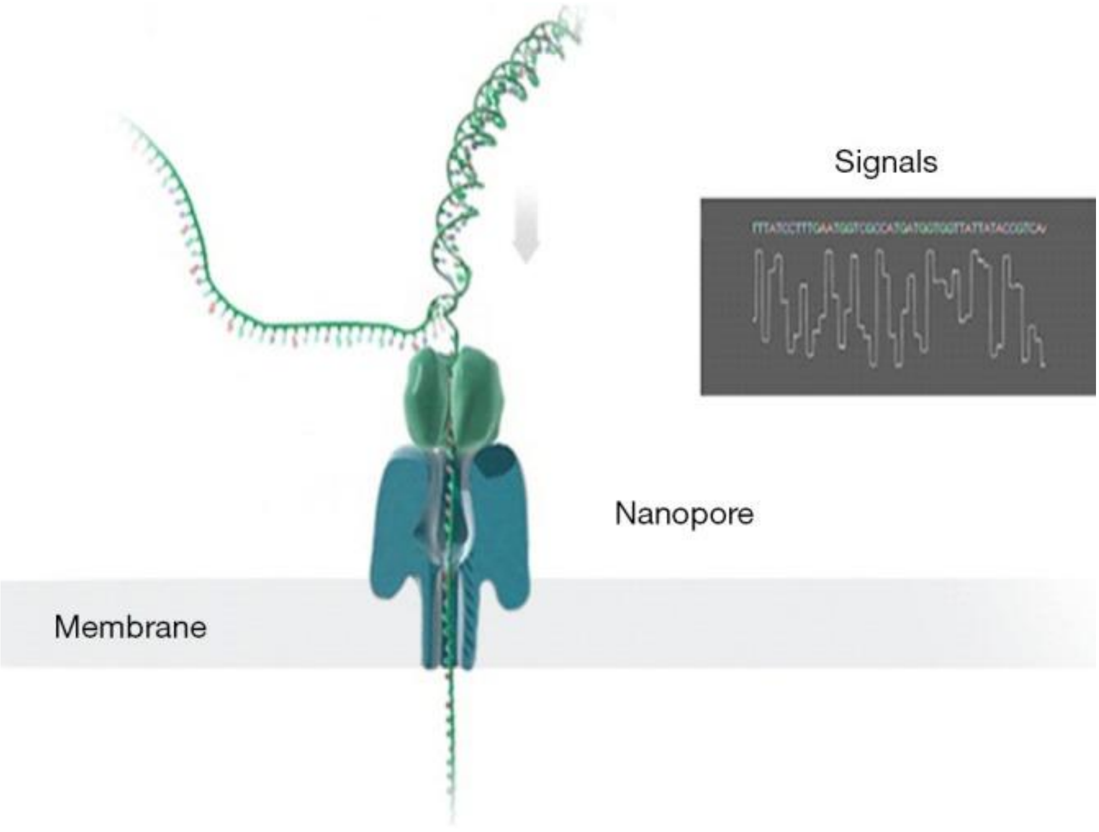


Figure 17: Nanopore Sequencing. (Xiao & Zhou, 2020)

The Pacific Biosciences platform utilizes single molecular real time (SMRT) technology. This technology relies on a SMRT cell containing millions of tiny wells called zero-mode waveguides (ZMWs). Each molecule from a volume of DNA is immobilized in the ZMWs and polymerase is used to incorporate fluorescent labelled nucleotides that are used to identify the nucleotide composition of the molecules. A camera system records the colour of the emitted fluorescence in real time to identify each nucleotide (Figure 18).

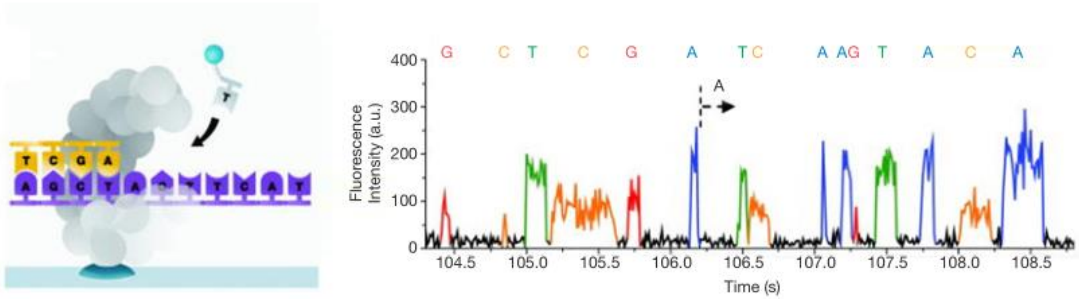


Figure 18: SMRT Sequencing (Xiao & Zhou, 2020)

This newer generation of sequencing technology brings with it advantages such as the ability to produce much longer reads when compared to NGS technologies, this provision tackles issues in genome assembly caused by shorter reads (Bleidorn, 2016). It also provides faster sequencing speeds which are a great advantage in clinical settings where quick analysis is usually required. Despite these improvements the issue of accuracy still remains a major issue when using TGS as error rates are much higher compared to NGS technologies (Bleidorn, 2016).

2.9 Chapter Summary

In this chapter, the theoretical background needed to understand the research carried out for this thesis is outlined. An overview of the structure and function of DNA is provided, followed by a discussion of DNA sequencing from the initial procedures introduced in 1977 to those introduced with the inception of NGS.

The focus of this thesis is the analysis of artefacts in the library preparation stage of NGS using a simulation. Therefore, NGS and the Illumina sequencing platform are described in more detail. The main content of this chapter is found in Sections 2.5 and 2.6. These sections delivered a systematic description of the different stages of sequencing using the Illumina platform followed by descriptions of the artefacts that can occur, along with indications of why they may occur and measures that have been taken to reduce their presence in sequencing outcomes.

Finally, a brief description of existing simulators is laid out in Section 2.7 along with their limitations and an introduction to Third generation sequencing technologies is provided in Section 2.8. The next chapter gives the details of the library preparation simulator that I have developed in order to carryout research into the effects of artefacts that occur in this stage of sequencing.

Chapter 3 Modelling the Library Preparation of NGS

This chapter describes the library preparation model used in my work and the methods used to analyse its output. The main objective of the model is to provide a virtual platform to simulate the different stages of the Illumina sequencing library preparation workflow and the possible artefacts that may be introduced at each stage. The simulator aims to uncover the effects of flaws (variations and biases) at each stage and their concurrent influence on subsequent stages. The table below outlines the stages that have been developed and their implemented parameters and biases.

Table 2: LpSIM parameters

Stage	Parameters	Bias
DNA Fragmentation	Fragment size distribution parameters (μ, σ)	Probability of cleavage at a CpG site
Adapter Ligation		Probability of ligation
PCR Amplification	Denaturation temperature Elongation temperature Number of PCR cycles	Temperature dependent amplification

The model simulates three stages of DNA library preparation: DNA fragmentation, adapter ligation and PCR amplification (**Figure 19**). The first stage involves creating fragments, the size of which is derived from a lognormal distribution and cleavage points are drawn randomly from a uniform distribution. In the second stage, adapters are attached to the DNA fragments based on a user defined probability of ligation and the identity of the nucleotides at their terminus. Finally, the fragments are amplified depending on their melting temperature. The result of the simulation is the dispersal of fragments of different sizes over an input DNA sequence. From this, the number of fragments at each nucleotide (coverage) and the uniformity of coverage across the sequence (evenness) can be computed.

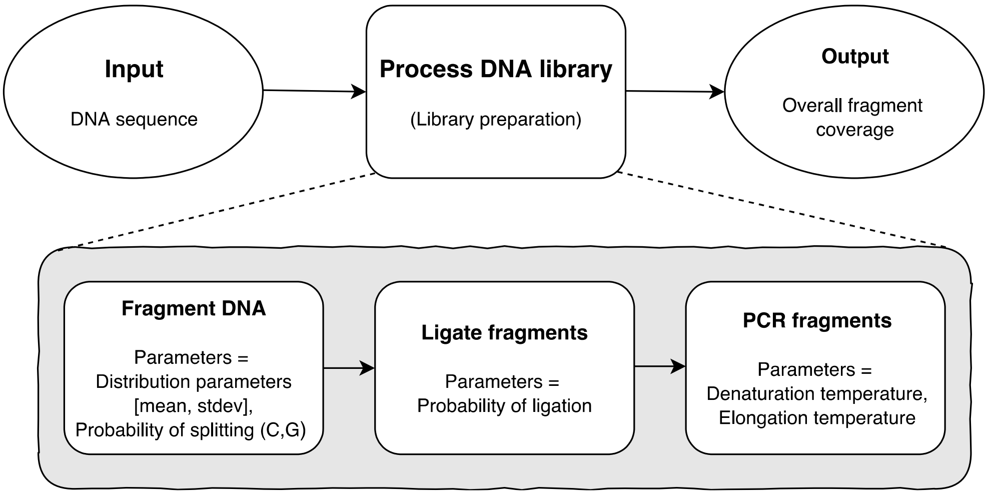


Figure 19: Library preparation workflow in LpSIM.

3.1 DNA Sequence Generator

In order to develop and test the model, artificial DNA sequences with modifiable characteristics were needed. Having such a sequence allows for the testing of the effects of such things as different levels of GC content. A pseudo-random DNA generator is used to produce a series of four types of nucleotides such that the identity of a nucleotide at a given position N_i does not depend on the identity of the preceding nucleotides (**Algorithm 1**). Creating DNA sequences with regions of biased nucleotide content (high GC or AT) involves specifying the required regions and the percentage of GC or AT content for that region of the sequence. So, for example, setting p(C) = p(G) = 0.4 and p(A) = p(T) = 0.1 would create a GC-rich sequence.

Algorithm 1: Generate artificial DNA

```
Select DNA sample size $D$
Set proportions of nucleotides $p(A), p(C), p(G), p(T)$
for $i := 1$ to $D$ do
    if random number (0,1) ≤ p(A)
        $N_i$ = "A"
    else
        if random number (0,1) ≤ $p(A) + p(C)$
            $N_i$ = "C"
        else
            if random number (0,1) ≤ p(A) + p(C) + p(G)
                $N_i$ = "G"
            else $N_i$ = T
return $N$
```

A pre-existing sequence generator could have been used for generating basic sequences, but I chose to implement this solution to enable customization of sequences for my use case. For example, creating sequences with varying nucleotide content in different regions to enable visualisation of the effects of biases on nucleotide content.

3.2 DNA Fragmentation

This stage of the model involves splitting of an input DNA sequence into fragments. The distribution of fragment sizes appears to depend on the method of shearing. Fragments resulting from acoustic shearing, sonication and enzymatic fragmentation are typically distributed with a positive skew, whereas nebulization may result in negatively skewed distributions (**Figure 20**).

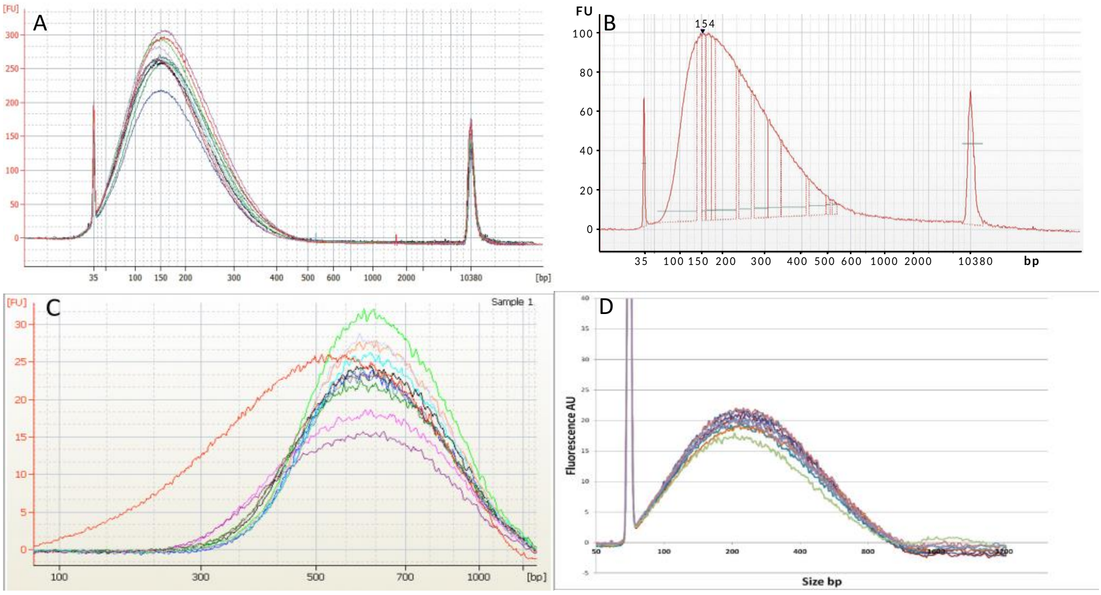


Figure 20: Size distributions of DNA fragments sheared using A: Bioruptor Pico sonicator (Diagenode, 2013), B: KAPA enzymatic fragmentation kit (Kapa Biosystems, 2016), C: Nebulization (Lundin *et al.*, 2010), D: Covaris acoustic shearing (Covaris, 2016).

On theoretical grounds, Kolmogorov (1941) concluded that particle sizes from sequential breakage processes tend to be lognormally distributed. This pattern mostly occurs when solid materials are subjected to mechanical forces (Neĭkov, Naboychenko & Yefimov, 2018). Thus, the lognormal distribution is extensively used to model breakage processes, for instance in geological research to model grain size distributions derived from explosive rock fragmentation (Fowler & Scheu, 2016), and mathematical modelling of brain parcellation in neural circuit research (Ferrante, Wei & Koulakov, 2014).

Using the lognormal distribution here allows for the regulation of the skew of the fragment size distribution to better match what is found in reality (**Figure 20**) and to also find the effects of this skewness. Its use in modelling breakage processes as discussed above makes it the preferred choice for modelling the resultant sizes of the DNA fragmentation process in this simulator.

The lognormal distribution is a continuous probability distribution where the logarithm of a random variable $Ln(x)$ is normally distributed (Maymon, 2018). It is a positively, semi-bounded (i.e. only considers positive values) skewed distribution that is characterised of positive values. A shape and location parameter (σ and μ) are used to specify the distribution and can be set to obtain different degrees of skewness (**Figure 21**). The mean and variance of lognormal random variables can be derived from these parameters.

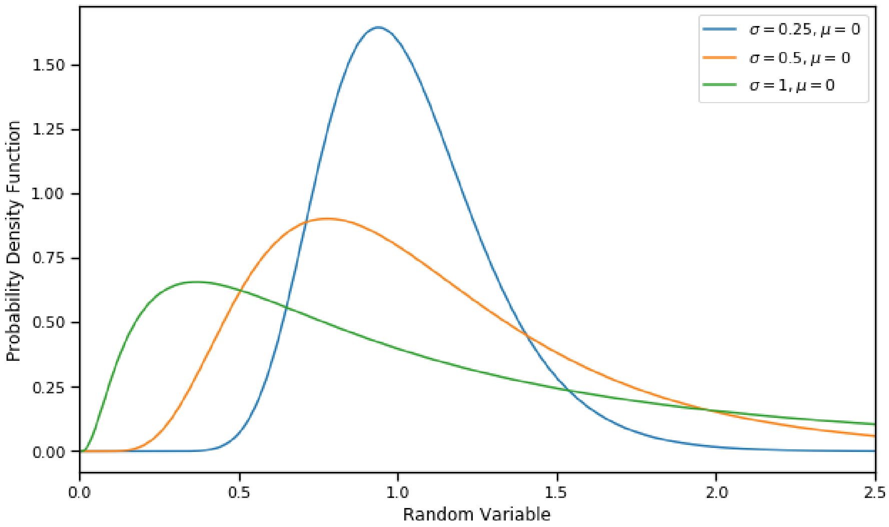


Figure 21: Lognormal density function with different shape parameters.

To generate lognormally distributed fragments from input DNA a chosen number n of fragment sizes f are drawn from a lognormal distribution with a mean of $\boldsymbol{m} = \mathbf{exp}\left(\boldsymbol{\mu} + \frac{\boldsymbol{\sigma}^2}{2}\right)$ and variance of $\boldsymbol{v} = \mathbf{exp}(2\boldsymbol{\mu} + \boldsymbol{\sigma}^2)\,(\mathbf{exp}(\boldsymbol{\sigma}^2) - \mathbf{1})$. For each fragment size f_i, a random cleavage start point C_{start} is drawn from a uniform distribution $[1, L]$ where L denotes the length of the input DNA sequence. Given C_{start}, the cleavage end point C_{end} is computed as $f_i + C_{start}$. In this way the cleavage start and end points are defined for each of the n fragments. A sequence-dependent (CpG) cleavage bias (Section 2.6.1) is integrated in the fragmentation model by implementing a probability

$0 \leq$ B.SPLIT ≤ 1, where B.SPLIT = 0 means no bias and B.SPLIT = 1 implies that a split between C and G always occurs. The model returns an array of DNA fragments ready for the ligation stage (**Algorithm 2**).

Algorithm 2: Fragment DNA

```
1. Given a DNA sample $(D)$ generate a list of fragment sizes $(f)$ from a lognormal distribution
2. Set splitting probability $p(S)$
3. Create fragment list $(F)$
4. for $i$ in $f$
5.      $C_{start}$ = random number (0, length($D$))
6.      $C_{end} = f_i + C_{start}$
7.      if $D[C_{end}]$ = “G” and $D[C_{start}] - 1$ = “C” and random number (0,1) $<= p(S)$
8.           $n = D[C_{start}: C_{end}]$
9.      else if $C_{end}$ ! = “G” and $C_{start} - 1$ ! = “C” and random number (0,1) $>= p(S)$
10.          $n = D[C_{start}: C_{end}]$
11.     else go to step 4
12.     $F \leftarrow n$
13. return $F$
```

3.3 Adapter Ligation

After the DNA has been fragmented following the procedure described above, a preset string representing an adapter is appended to each fragment from a list of fragments depending on the probability parameter B.LIGATE. When the value of B.LIGATE is zero all fragments are ligated. As the parameter goes up, the likelihood of adapters not ligating to fragments increases. Also, the identity of the nucleotide at the ends of the fragments is taken into account: adapters probabilistically bind to fragments with a T on their 5’ end or an A on their 3’ end based on the value of B.LIGATE. Consequently, there will be fragments with adapters on only one end. These fragments, like the ones without any adapters, will not go through to the next (PCR) stage (**Algorithm 3**).

Algorithm 3: Ligate DNA

```
Given a list of fragments ($F$) ligate adapters $A$ and $B$ to each fragment ($F_i$)
Set ligation probability $p(L)$
Create list of ligated fragments ($LF$)
for $i$ in $F$
    if $F_i[0]$ = "T" and random number (0,1) $>= p(L)$
        $F_i = A + F_i$
    else if $F_i[0] != $ "T"
        $F_i = A + F_i$
    if $F_i[-1]$ = "A" and random number (0,1) $>= p(L)$
        $F_i = F_i + B$
    else if $F_i[-1] != $ "A"
        $F_i = F_i + B$
    $LF \leftarrow F_i$
return $LF$
```

3.4 PCR Amplification

The polymerase chain reaction (PCR) amplifies a volume of DNA exponentially by duplicating the number of fragments in a series of cycles. Heating up the DNA sample, splits the double-stranded fragments into single strings (denaturation), then followed by ramping down to a lower temperature to attach PCR primers to the single-stranded fragments (annealing). Subsequently, the temperature is increased once again to the optimal temperature for an enzyme (a polymerase) to synthesize a new complementary strand (elongation). These thermal cycles are run several times until the desired number of clones is attained. One issue that plagues the PCR process is amplification bias (Section 2.6.4): due to the stronger bonds between complementary GC pairs (three hydrogen bonds) it requires a much higher temperature to denature GC-rich fragments. AT pairs are connected with just two hydrogen bonds and denature at a lower temperature. However, the higher temperature needed for elongation could therefore lead to dissociation of such fragments which in turn would lead to a lower yield of cloned AT-rich fragments.

To model this amplification process the melting point of a double stranded DNA fragment has to be determined. The melting point (T_m) of DNA refers to the temperature at which 50% of the nucleotide pairs dissociate. Several procedures exist for establishing the melting point of short DNA sequences (e.g. fragments and primers). For my simulation, I used a nearest-neighbour model formulated by Breslauer *et al.* (1986).

This method predicts the stability and melting behaviour of nucleotide pairs by using the temperature-dependent behaviour ($\Delta G°$) and relative stability ($\Delta H°$) of bonds between neighbouring nucleotide pairs (Breslauer *et al.*, 1986). The predicted relative stabilities of all possible nearest-neighbour combinations are used to calculate the overall thermal stability of a given fragment (**Equation 1**).

$$\mathrm{T}_m = \left\{ \frac{\Delta H° \times 1000}{\Delta S° + R \ln\left(\frac{C_t}{4}\right)} \right\} - 273.15$$

Equation 1: Nearest-neighbour model equation (Le Novere, 2001)

In this equation, $\Delta H°$ and $\Delta S°$ respectively represent the sum of nearest-neighbour enthalpy and entropy changes for a given DNA fragment, of which the values can be looked up from a table as shown (**Table 3**). R is the gas constant (1.987 cal deg^{-1} mol^{-1}(calorie per degree per mole)) and C_t represents the total molar ratio of strands (Le Novere, 2001; Sigma-Aldrich, 2015). The melting temperature calculations for my model were computed using the Biopython MeltingTemp module (Cock *et al.*, 2009).

Table 3: Nearest-Neighbour (NN) thermodynamic values for *ΔH°* and *ΔS°* (Allawi & SantaLucia, 1997). These values were experimentally derived from optical melting studies.

NN interactions[1]	*ΔH°* (kcal/mol)	*ΔS°* (kcal/mol)
AA/TT	-7.9	-22.2
AT/TA	-7.2	-20.4
TA/AT	-7.2	-21.3
CA/GT	-8.5	-22.7
GT/CA	-8.4	-22.4
CT/GA	-7.8	-21.0
GA/CT	-8.2	-22.2
CG/GC	-10.6	-27.2
GC/CG	-9.8	-24.4
GG/CC	-8.0	-19.9
Terminal G/C base pair[2]	0.1	-2.8
Terminal A/T base pair[2]	2.3	4.1

[1] The sum of interaction values is taken for the subject sequence (fragment).

[2] These are duplex initiation parameters which account for stability changes when a sequence is terminated by a G/C or A/T base pair.

The PCR step of my model accepts: an array of ligated fragments, a pre-set denaturation and elongation temperature, and the required number of PCR cycles as input. First, partially ligated, and non-ligated fragments are filtered out because they are missing the adapters required for primer attachment during the annealing stage of PCR (Quail *et al.*, 2008). Next, the melting temperature is computed for each fragment after which fragments with a T_m higher than the set denaturation temperature[10] are filtered out (because their strands will not dissociate). Fragments are also discarded during the elongation step, namely those with a T_m lower than the set elongation temperature (because their strands would disassociate at higher elongation temperatures). This rejection is controlled by a function that probabilistically allows

[10] The required temperature values are set by the user

fragments with a T_m that is in close proximity to the set temperatures to go through the process. The module output is an array of duplicated fragments spread out over the input DNA sequence (**Algorithm 4**).

Algorithm 4: PCR processing

```
Given a list of ligated fragments ($LF$) process each fragment ($LF_i$) with a denaturation temperature
  (d) and elongation temperature (e) based on its melting temperature ($T_m$)
Set a denaturation probability $p(D)$ based on denaturation temperature difference ($dd$)
Set an elongation probability $p(E)$ based on elongation temperature difference ($ed$)
Create a list of PCR processed fragments ($PF$)
for $i$ in $LF$
    $dd = T_m[LF_i] - d$
    $ed = e - T_m[LF_i]$
    if $dd < 0$
        if $ed < 0$
            $PF_i = LF_i$
        else if random number (0,1) <= $p(E)$
            $PF_i = LF_i$
    else if random number (0,1) <= $p(D)$
        if $ed < 0$
            $PF_i = LF_i$
        else if random number (0,1) <= $p(E)$
            $PF_i = LF_i$
return $PF$
```

3.5 Fragment Coverage Metrics

In sequencing, coverage is an important metric that shows how many reads cover a nucleotide position using a reference genome to confirm the nucleotide identity of each position. A reference genome is an already known genome, which is sequenced from several individuals using different sequencing platforms to ensure accuracy. These reference sequences may not be completely accurate, but they are updated frequently to improve their accuracy. Due to this coverage results may not be completely accurate in some cases and will need to be revised when a reference sequence is updated.

A high coverage value signifies the reliability of a read for that particular nucleotide position. As the main output of this simulation is DNA fragments, coverage is characterised as the number of fragments covering a nucleotide position ("fragment coverage").

Here I utilise this metric to observe how implemented parameters and biases might affect fragment coverage across a given sequence. To compute the fragment coverage value, the number of fragments covering each nucleotide position of a sequence is calculated and an array of coverage values is returned. A visualisation of this output is shown below (**Figure 22**).

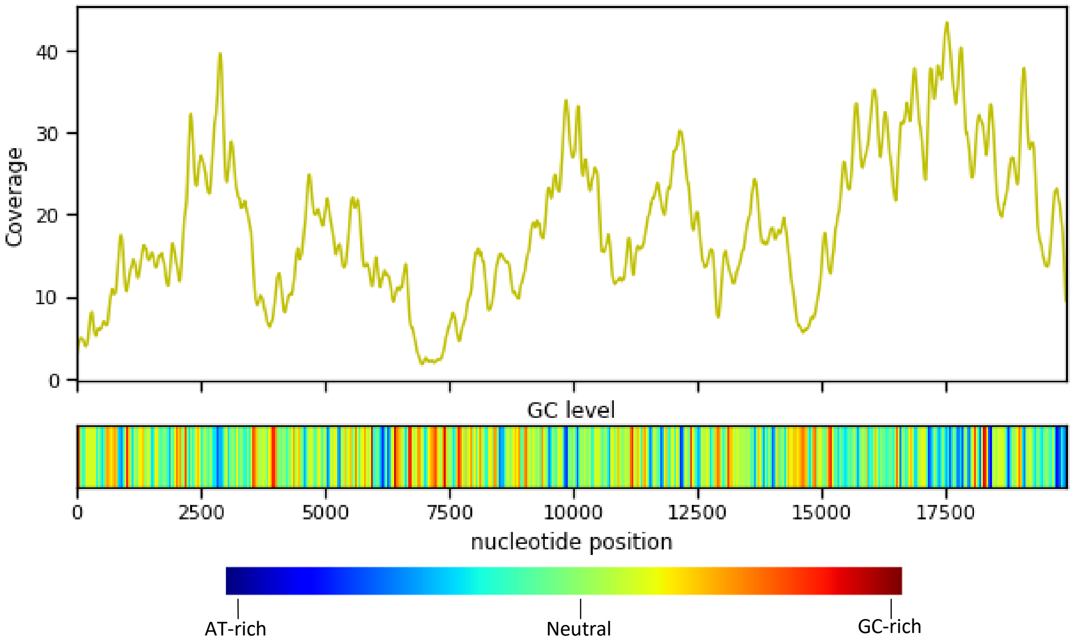


Figure 22: Coverage plot of an artificial GC-rich sequence.

The colour bar in **Figure 22** visualizes the GC-content across a DNA sequence and allows for inspecting possible associations between coverage and DNA structure. The colour scale is blue when GC-content is low (AT-rich), then transitions to green when GC-content is neutral and finally to red when GC-content is high. The GC-content is calculated as a moving average for a window size of 80bp. This size was chosen as it was the best of a range of values (0 – mean fragment size) at revealing the intensity of GC-content across a sequence, as found by a quick experiment that was undertaken. To emulate the reads from the real coverage data used in comparisons (Chapter 4) made in this thesis, paired end reads (Section 2.5.4), which take in to account the identity of nucleotides on both ends of a fragment are used. A read length of 80bp for each end of a fragment was used to ensure consistency between the real and simulated coverage results and across all tests carried out in Chapter 5.

Another important metric in determining the quality of sequencing output is the homogeneity of coverage across a given sequence (Gnirke *et al.*, 2009). This metric determines the variation in coverage across a sequence. Higher levels of homogeneity indicate that a sequence is evenly covered by reads (fragments in this case), thus providing reliable read quality (Sims *et al.*, 2014). A low level of homogeneity would signify uneven coverage leading to poor quality sequencing output. This metric is most often calculated by three methods: listing regions in a DNA sequence with coverage above a specific value (Horton, 2016), using the coefficient of variation to measure variability of the coverage values, and using an evenness score (Mokry *et al.*, 2010) which is the chosen method for this simulation. The evenness score (E) is defined as the portion of coverage values that are evenly distributed across a sequence (Mokry *et al.*, 2010) and is stated as:

$$E = \left\{ \sum_{i=1}^{C_{ave}} \frac{P_i}{C_{ave} * N_{TP}} \right\} * 100\%$$

Equation 2: Evenness score

Where C_{ave} is the average coverage depth of the whole sequence (sum of coverage values/length of sequence), P_i is the number of nucleotide positions having a coverage value of at least C_i, and N_{TP} is the length of the sequence (Mokry *et al.*, 2010). A high value for E means that the coverage of a given sequence is homogeneous, while a lower value signifies poor homogeneity.

I have chosen the approach by Mokry et al. because it allows for comparing the quality of coverage from different experiments with different average coverage depths (C_{ave}), as the value of E is independent of C_{ave}. This relative independence made it the best choice for my work as I will need to compare different sets of coverage values for different levels of the parameters implemented in the simulation. So using this metric and analysis of variance (ANOVA) tests, a comparison of the effects of varying parameters at the different stages of library preparation on coverage can be made (Chapter 5).

3.6 Analysis of variance

Analysis of variance (ANOVA) is a statistical procedure for testing the significance of possible differences of a dependent variable between two or more samples (Searle, 1997), where a sample represents a set of values of that dependent variable measured under a particular (experimental) condition. In other words, ANOVA tests for possible significant effects of the conditions (the independent variable) on the dependent variable.

Formally this is done by checking if the variance between samples (due to factors like experimental conditions) outweighs the variance within them (due to error). If the variance between conditions is larger (as expected by chance) than the variance within conditions, then the null hypothesis of no effect of (experimental) conditions has to be rejected. However, if this is not the case, the error within conditions overrules the effect of the experimental conditions and hence there is no significant difference between the samples. An ANOVA table with the formulas needed for its calculation is presented below **(Table 4)**.

Table 4: ANOVA Table

Source of variation	Sum of squares (SS)	Degrees of freedom (DF)	Mean Sum of Squares (MS)	F	P
Between conditions (B)	$SSB = n\sum(x_j - x_t)^2$	$DFB = c - 1$	$MSB = \frac{SSB}{DFB}$	$\frac{MSB}{MSW}$	p
Within Conditions (W)	$SSW = \sum(x_{ij} - x_j)^2$	$DFW = c(n-1)$	$MSW = \frac{SSW}{DFW}$		
Total (T)	$SST = \sum(x_{ij} - x_t)^2$	$DFT = N - 1$	$MST = \frac{SST}{DFT}$		

[a] n = number of repeated outcomes for each condition
[a] x_j = group means
[b] x_t = grand mean
[c] x_{ij} = individual observation
[d] c is the number of conditions.
[e] $N = c * n$
[f] p is the probability of obtaining an F value under the null hypothesis

The table is structured as follows:

- The variance between conditions is calculated by dividing SSB, the sum of the squared differences between each sample mean and the grand mean multiplied by the number of repeated outcomes for each condition (n), by the corrected number of conditions (DFB).
- The variance within conditions is computed by dividing the sum of the squared differences between each individual observation and its sample mean (SSW) by the appropriate degrees of freedom (DFW)

- The total sum of squares (SST) is calculated by summing up the squared differences between each observation and the grand mean. This boils down to summing the sums of squares between (SSB) and within (SSW) samples, $SSB + SSW$.
- The degrees of freedom for the between conditions is $DFB = c - 1$ and for the within conditions $DFW = c(n - 1)$. The total degrees of freedom can also be represented as $DFT = DFB + DFW$.
- The mean squares are derived by dividing the sum of squares by the degrees of freedom for each row and thus yield the between and within sample variances
- Finally, the difference between the within and between variances is measured by the variance ratio F as $F = \frac{MSB}{MSW}$. The distribution of F under the null hypothesis is a mathematically defined and known probability density function.
- The null hypothesis assumes that all samples are drawn from a population whose means are equal, $\mathrm{H}_0\!: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$. If this hypothesis is true, the computed variance ratio follows the F-distribution for given degrees of freedom (DFB and DFW respectively). This allows for the probability (p) of obtaining an F value under the null hypothesis to be calculated. A significance level of $\alpha = 0.05$ is typically set. This means that if a computed variance ratio exceeds a critical F value (i.e. an F value which, by pure chance, would occur with a probability lower than the set significance level) the null hypothesis has to be rejected. In that case, at least one of the sample means is said to differ significantly from any one of the others and a significant effect of the conditions is found.

The main requirements of an ANOVA are:

- **Equality of sample variances (Heteroscedasticity):** The variance of the tested samples should not differ significantly.
- **Normality:** The error (deviations from the sample means) should be normally distributed.

In the data analysis of this thesis, the Levene, Kolmogorov-Smirnov and Shapiro-Wilk tests have been used to check these assumptions. The Levene test is used to assess the homogeneity of variances between groups (Levene, 1960). A p-value greater than the set alpha of 0.05 for this test indicates there is no significant difference between the groups tested thereby retaining the null hypothesis for the assumption of homogeneity. While a p-value lower than 0.05 signifies a failure to meet this assumption. When faced with a failure, a log transformation of the data can be effective in restoring the equality of variance in most cases. This transformation modifies the skewness of the data and can restore the required symmetry. The Kolmogorov-Smirnov (KS) and Shapiro-Wilk (SW) tests are both used in testing the assumption of normality. The KS test quantifies the distance between a population's observed cumulative distribution function (CDF) and its hypothesized CDF (normal distribution), the percentage of values deviating from the hypothesised distribution are used as the test statistic (Massey, 1951). If this percentage is low the null hypothesis of normality is accepted as a lower percentage would result in a p-value larger than 0.05. When a higher percentage of deviation is observed, and the p-value is less than 0.05 the null hypothesis is rejected. The SW test measures a W statistic that quantifies if a random sample is from a normally distributed population (Shapiro & Wilk, 1965). A small W value signifies a departure from normality while higher W values would signify the samples have been drawn from a normal distribution. If the assumption of normality is not met, a log or square root transformation can be used to modify the skew of the data to make it normally distributed.

Higher order ANOVAs

In my analysis, two- and three-way ANOVAs are used to compare the effects of the six different library parameters on the coverage of specific DNA sequences (Sections 5.1 and 5.2). Whereas in a single-factor ANOVA the conditions are the levels of a single independent variable or "factor", higher order ANOVAs (also called multi-factor ANOVAs) deal with designs containing more than one factor. These multi-factor ANOVAs assess the proportion of the overall variance that is due to the effect of treatments (conditions). The variance explained by each factor is computed as a main effect and the more factors that are included, the less unexplained variance (i.e. error or within variance) remains. Not only does this reduction of within variance lead to a larger variance ratio than would result from separate single factor ANOVAs, it also circumvents repeated testing of the same data. The latter is an ill-advised strategy because it increases the number of outcomes that are statistically significant by chance alone. Furthermore, multi-factor ANOVAs have the added advantage of allowing to test for a possible effect of combinations of treatment levels, so-called interaction effects. An interaction between two main factors A and B (denoted as A*B), implies that the effect of one or more levels of A depends on that of one or more levels of B (**Figure 23**). A two-way ANOVA with formulas for each of its elements is provided in **Table 5**.

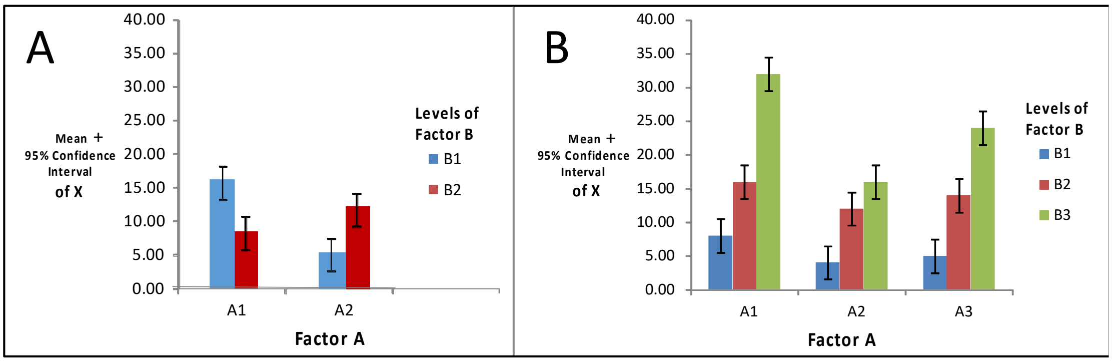


Figure 23: Examples of interactions in ANOVAs. **A**: The mean of dependent variable X measured under condition B1 is larger than condition B2 of factor B, but only for condition A1 of factor A. **B:** For all levels of A, the lowest values of X are found in condition B1 of factor B. However, **significant** differences in X due to conditions B2 and B3 are found under conditions A1 and A3, but not under condition A2 of factor A.

Table 5: Two-way ANOVA table

Source of variation	Sum of squares (SS)	Degrees of freedom (DF)	Mean Sum of Squares (MS)	F	P
Factor X Between conditions (B)	SSBX	$DFBX = x - 1$	$MSBX = \frac{SSBX}{DFBX}$	$\frac{MSBX}{MSW}$	p
Factor Y Between conditions (B)	SSBY	$DFBX = y - 1$	$MSBY = \frac{SSBY}{DFBY}$	$\frac{MSBY}{MSW}$	p
Interaction X * Y	SSXY	$DFXY = (x - 1)(y - 1)$	$MSXY = \frac{SSXY}{DFXY}$	$\frac{MSXY}{MSW}$	p
Within Conditions (W)	SSW	$DFW = xy(n - 1)$	$MSW = \frac{SSW}{DFW}$		
Total (T)	SST	$DFT = N - 1$	$MST = \frac{SST}{DFT}$		

[a] x = number of repeated outcomes for each condition of factor X
[b] y = number of repeated outcomes for each condition of factor Y
[c] c is the number of conditions
[d] $N = xy * n$
[e] p is the probability of obtaining an F value under the null hypothesis

3.7 LpSIM

LpSIM is a library preparation simulator that produces estimated coverage of a DNA sequence. Its main function is to test the effects of the different stages of library preparation on sequence coverage.

It is developed in Python and utilises several Python libraries to model the library preparation process. The fragmentation, ligation and PCR stages of library preparation are implemented in the tool.

3.7.1 Installation

LpSIM is available to download at https://github.com/ebewo/LpSIM

Requirements:

1. Unix based operating system
2. Python 3
3. Git

Installation instructions:

1. Clone the git repository:

```
$ git clone https://github.com/ebewo/LpSIM.git
```

2. Install the required python libraries:

```
$ cd LpSIM/
$ pip3 install -r requirements.txt
```

3.7.2 Usage

LpSIM has two command scripts: "seqgen.py" for generating an in-silico DNA sequence and "run.py" for running the simulator on an input DNA sequence.

Generating an in-silico DNA sequence

The sequence generation command depends on the configuration file "generator.yaml" which contains the characteristics of the required DNA sequence:

```
GC : 80 # GC level of sequence
seq_size : 20000 # DNA sequence size
pclump : 0 # clumping probability
shuffle_opt : 0 # 0 = [no shuffle], 1 = [1 GC-rich area,
1 neutral area, 1 AT-rich area], 2 = [1 GC-rich area, 2
neutral areas,  1 AT-rich area]
```

Running the command returns the input parameters and a confirmation of completion of the task. This generates a sequence using the input parameters which is then saved to a sequence directory.

```
$ python3 seqgen.py

Parameters:

GC: 80
pclump: 0
seq_size: 20000
shuffle_opt: 0

Sequence generated!
```

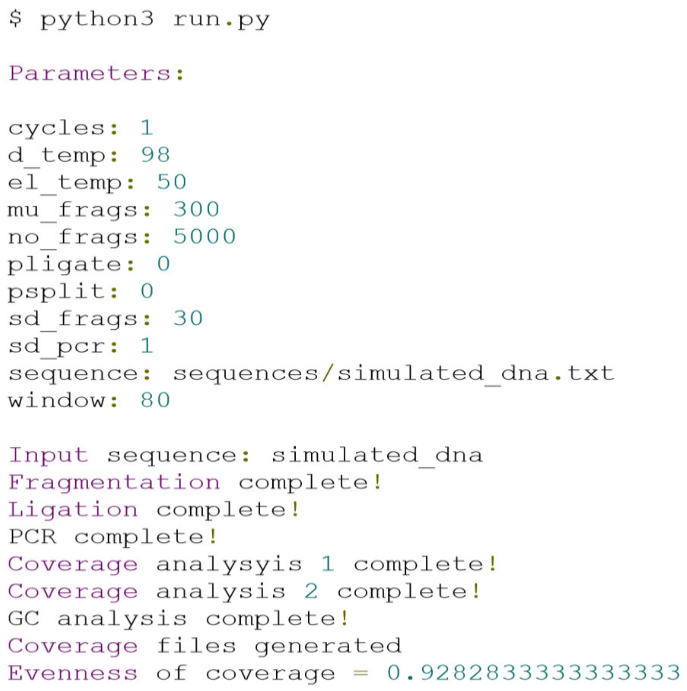
The created DNA sequences can be used in LpSIM to test the effects of library preparation on different types of DNA compositions (e.g. GC-rich and AT- Rich).

Generate coverage values for a given DNA sequence

To run the simulator on a given sequence a configuration file “parameters.yaml” is populated with the required parameters for the implemented stages of the library preparation process:

```
sequence: "sequences/simulated_dna.txt" # input sequence
mu_frags: 300 # mean of fragment size distributiom
sd_frags: 30 # standard deviation of fragment size
distribution
no_frags: 5000 # number of fragments
psplit: 0 # splitting probabilty
pligate: 0 # ligation probability
d_temp: 98 # denaturation temperature
el_temp: 50 # elongation temperature
sd_pcr: 1  # standard deviation of PCR probability
distribution
cycles: 1 # number of PCR cycles
window: 80 # window size for moving average of GC levels
and coverage
```

Executing the “run.py” script returns a confirmation of the input parameters followed by a confirmation of completion for each library preparation stage. After the simulation is completed the resultant coverage is calculated and saved to a csv file containing the nucleotide identity for each base position and its coverage value (**Table 6**). The coverage values and GC content across the sequence are then used to generate a plot that will help in identifying regions lacking coverage (**Figure 24**). Finally, the evenness score for the given coverage values is returned.



```
$ python3 run.py

Parameters:

cycles: 1
d_temp: 98
el_temp: 50
mu_frags: 300
no_frags: 5000
pligate: 0
psplit: 0
sd_frags: 30
sd_pcr: 1
sequence: sequences/simulated_dna.txt
window: 80

Input sequence: simulated_dna
Fragmentation complete!
Ligation complete!
PCR complete!
Coverage analysyis 1 complete!
Coverage analysis 2 complete!
GC analysis complete!
Coverage files generated
Evenness of coverage = 0.9282833333333333
```

Table 6: Sample coverage results

base position	nucleotide id	coverage
1	C	0
2	C	1
3	G	1
4	G	1
5	G	1
6	C	1

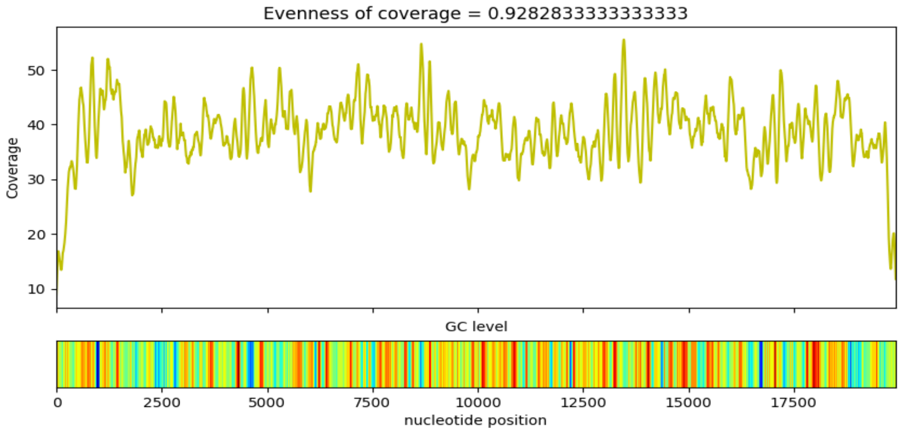


Figure 24: Coverage plot

In Chapter 5, LpSIM is used to test the effects of different levels of parameters for each library preparation stage on the coverage of different types of DNA structures than can be found in natural genomes. It is expected that such experiments where the influence of library preparation parameters are studied will be the main use case for this tool.

3.7.3 Performance Metrics

The specifications of the computer used for the performance tests in this section is given below:

Processor	AMD Ryzen 3700X @ 4.3 GHz
Installed RAM	32GB
Operating System	Ubuntu 18.04 LTS
Python Version	3.6.9

Testing the performance of the sequence generator involved generating sequences of varying sizes and measuring computational time and memory usage. The collected metrics are presented in **Table 7** and **Figure 25**. Generating a 10000bp sequence takes 0.103 seconds and the run time scales linearly with sequence size. Utilised memory slightly increases with larger sequence sizes.

Table 7: Sequence generator test results

DNA sequence size	Time (seconds)	Memory (kilobytes)
10000	0.103	56
20000	0.103	56.2
30000	0.113	56.2
40000	0.12	56.6
50000	0.123	56.5
60000	0.131	57.2
70000	0.14	57.2
80000	0.143	57.2
90000	0.153	57.2
100000	0.158	57.4

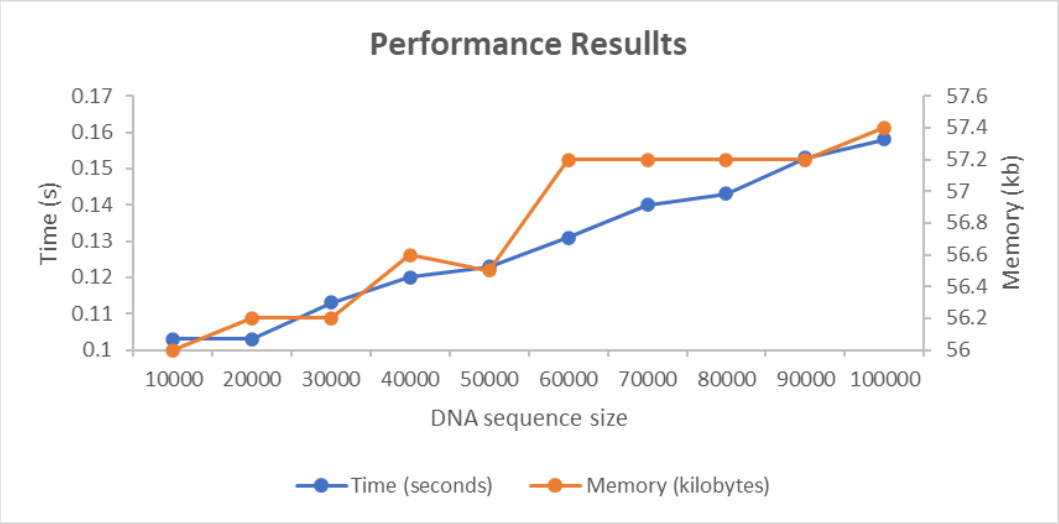


Figure 25: Performance results for the sequence generator

The same approach as above was used to test the simulator. Each sequence generated in the previous test was used to run the simulator with a static set of parameters (**Table 8**). In **Figure 26** the same trend as above can be seen where the computational time scales linearly with increasing fragment sizes. Memory utilisation is mildly erratic but generally sits between 160kb – 195kb for the sequence sizes tested. These results will be used to determine the computational time and memory requirements for experiments to be carried out in Chapters 4 and 5.

Table 8: Simulator Parameters

Library Preparation Step	Parameter	Value
Fragmentation	Mean[1]	300
	Standard Deviation [1]	30
	Splitting Bias	0
Ligation	Ligation Bias	0
Amplification	Denaturation Temperature	98
	Elongation Temperature	50

[1] *Parameter of the fragment size distribution (lognormal).*

Table 9: Simulator test results

DNA sequence size	Time (seconds)	Memory (kilobytes)
10000	12.8	172.5
20000	21.7	166.3
30000	31	171.7
40000	40	162.2
50000	48.3	172.2
60000	57.6	166.9
70000	67.1	165
80000	76	165.8
90000	83.8	167.4
100000	94.6	192.6

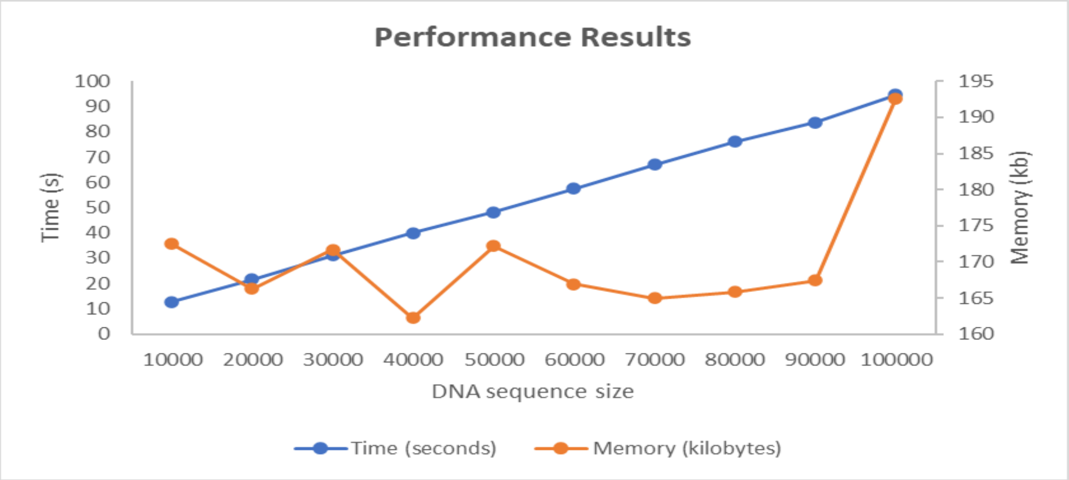


Figure 26: Performance metrics for the simulator

3.8 Chapter Summary

This chapter described the development of the library preparation model and metrics used to measure the quality of the simulator's output. The functionality of methods and modules that simulate different stages of the library preparation process and quantify its output were outlined. The functionality of ANOVAs used to check the effects of the parameters of each implemented module was also discussed. Finally, the installation, usage and performance of the developed tool is presented. The next chapter discusses the validation of LpSIM using a genetic algorithm.

Chapter 4 Matching Model Outcomes with Results of Real Sequencing using a Genetic Algorithm

Following the development of the simulator, it became necessary to ensure the coverage results returned were comparable to those found in real-world sequencing. To obtain such results, a search for optimal parameters needs to be carried out. Genetic Algorithms have been found to provide a robust parameter search solution for optimisation problems (Selig & Coverstone-Carroll, 1996).

Genetic algorithms are a type of optimisation algorithm inspired by evolution which was introduced by Holland (1975). They are used to implement optimisation strategies by imitating the natural processes of reproduction and natural selection to provide very good solutions to a computational problem (Goldberg, 1989).

These natural processes are simulated by first creating a random population of individuals, then the fittest individuals are selected for a crossover step where they produce offspring which are further diversified by a mutation step (Mitchell, 1996).

The individuals take the form of values representing a solution to a given problem. Each individual is assessed by a fitness function that assigns a score to it based on its ability to solve an assigned problem. Following the allocation of scores to each individual, a selection process picks the highest scoring individuals. The crossover operator mimics the sexual reproduction process found in nature where the genes from a pair of parent chromosomes combine to form offspring. Mutations can occur during the reproduction process which can lead to errors in copying the genes of the parents to the offspring. This randomly changes the solution offered by an individual. The probability of this occurring is typically low (Mitchell, 1996).

The stages of the GA are run for several iterations until the fittest individual stays consistent for many generations. At the end, this individual is picked as the best solution for the problem presented.

The following sections outline how the genetic algorithm was setup for finding the best possible parameters for my simulator and the coverage results obtained with the parameters found.

The Distributed Evolutionary Algorithms in Python (DEAP) framework (Fortin *et al.*, 2012) was used to implement the genetic algorithm used in this thesis.

4.1 Initial Population

The initial population consists of 50 randomly generated individuals. Each individual is encoded as an array of values representing the simulator's six parameters: fragment distribution parameters (mean and standard deviation), splitting bias level, ligation bias level, denaturation temperature and elongation temperature. A randomly chosen real number from the domain of each parameter is allocated to each individual with a uniform probability. With this, the values of each encoded individual are used to run the simulator and its fitness is assessed. In using random numbers, it is assumed that the parameters are independent of each other, allowing for a large variety of solutions and thereby increasing the search space.

4.2 Fitness Function

After the initial population is generated, parameters for each individual are used to run the simulator on a preselected DNA sequence. The result of this is a set of coverage values for these individuals. For the comparison, real-world coverage results for the preselected DNA sequence are obtained from sequencing experiments stored on the NCBI (National Center for Biotechnology Information) SRA (Sequence Read Archive) database (Leinonen *et al.*, 2011). The fitness function checks the correlation between the coverage generated by the simulator and coverage from the real-world sequencing experiment. This similarity is measured using the Pearson correlation efficient (r) which assigns numerical values between -1 and 1, where -1 represents a negative linear relation, 0 represents no linear relation and 1 represents a positive linear relation. The r value assigned to each individual is used as its fitness score.

4.3 Selection

With fitness scores assigned to each individual, the best individuals are chosen using tournament selection. This is a commonly used strategy in GAs, that starts out with randomly choosing a group of individuals from the population with equal probability. The individual with the highest fitness score from this group is inserted into a secondary population ("mating pool"). Several tournaments are run until the secondary population list is the same size as the original population list (50 tournaments). Increasing the size of tournament groups improves the chances of getting an individual with a much higher fitness score. This modification of group sizes is known as "selection pressure" (Xie & Zhang, 2013). For this work a tournament size of three is used as the selection pressure at this size was found to produce the best fitness scores in my experiments.

An elitist approach was not used here as some parameters (the fragment distribution parameters and resultant cleavage points) lead to variable outcomes and are better served by new solutions (parameter values) from each GA run.

The mating pool is processed by the crossover and mutation operators which are explained in the succeeding sections.

4.4 Crossover

The crossover stage combines two individuals (parents) from the mating pool to create a possibly improved set of solutions (Sastry, Goldberg & Kendall, 2005). Here the blend crossover (BLX-α) operator (Eshelman & Schaffer, 1993) is used to create offspring.

This operator accepts parents y^1 and y^2 formed of real numbers. For each parameter y_i^c (i-th parameter) in an offspring y^c , it randomly selects a value between Y_i^1 and Y_i^2 based on a uniform probability, where

$$Y_i^1 = min(y_i^1, y_i^2) - \alpha d_i$$

$$Y_i^2 = max(y_i^1, y_i^2) + \alpha d_i$$

$$d_i = |y_i^1 - y_i^2|$$

Modifying the α of this operator has an effect on the range y_i^c is picked from. The default α of 0 leaves the original range between y_i^1 and y_i^2. An α greater than 0 increases the range which could lead to a value outside the interval, while a negative α reduces the range (Takahashi & Kita, 2001). Thus, this crossover stage produces one offspring for each pair of parents. The value of each parameter in the child solution is based on its parents' values but is markedly different from either. An α of 0 is used in my experiments to retain offspring parameter values within the range of both parent solutions.

4.5 Mutation

After offspring are created in the crossover stage, they are then subjected to the mutation operator. I have chosen to use the polynomial bounded mutation operator (Deb & Agrawal, 1999) as it is better suited to real numbers.

This method uses a polynomial probability distribution to change a current parameter value y_i to a mutated value z_i. The distribution has its mean at the current parameter value and its variance is a function of the distribution index j. To mutate y_i a perturbance factor δ is defined as:

$$\delta = \frac{z_i - y_i}{\beta_{max}}$$

Where β_{max} is the pre-set maximum perturbation value allowed between y_i and z_i. The polynomial probability distribution used to calculate the mutated value depends on the perturbance factor δ and is defined as:

$$\mathcal{P}(\delta) = 0.5(j+1)(1-|\delta|)^j$$

The valid range of the distribution is between -1 and 1. Next a random number w between 0 and 1 is generated and used in the equation below to calculate the perturbance factor δ corresponding to it using the probability distribution:

$$\bar{\delta} = \begin{cases} (2w)^{\frac{1}{j+1}} - 1, & if\ w < 0.5 \\ 1 - [2(1-w)]^{\frac{1}{j+1}}, & if\ w \geq 0.5 \end{cases}$$

Finally, a mutated value is calculated as follows:

$$z_i = y_i + \bar{\delta}\beta_{max}$$

Setting a high distribution index value results in a mutant akin to the original parameter value, while a smaller index produces a less similar value (Deb & Goyal, 1996; Deb & Agrawal, 1999; Zeng *et al.*, 2016). The probability of a parameter being mutated is $1/s$, where s is the size of an individual. The distribution index value is set at 20 for my experiments in order to obtain mutated values near the original value.

4.6 Results

The GA was run using sections of three genomes to derive parameters that result in coverage comparable to what was found in actual sequencing experiments. The first was a 50kbp (kilo base pair) section of the *Mycobacterium tuberculosis* strain H37Rv genome. The second a 50kbp section of the *Plasmodium falciparum* strain 3D7 genome. And finally, *TP53* (*Tumor Protein P53*), a gene from chromosome 17 of the human genome, which is 19,148bp long. These sequences were respectively selected to represent the different types of nucleotide composition bias (GC-rich and AT-rich) and a sequence with neutral base content (equal levels of A, C, G and T). This allows me to ensure the simulator can handle such sequences.

For each DNA sequence, the GA was run for a number of generations. I found the solutions to have converged at the 20th generation or earlier in all cases (**Figure 27**).

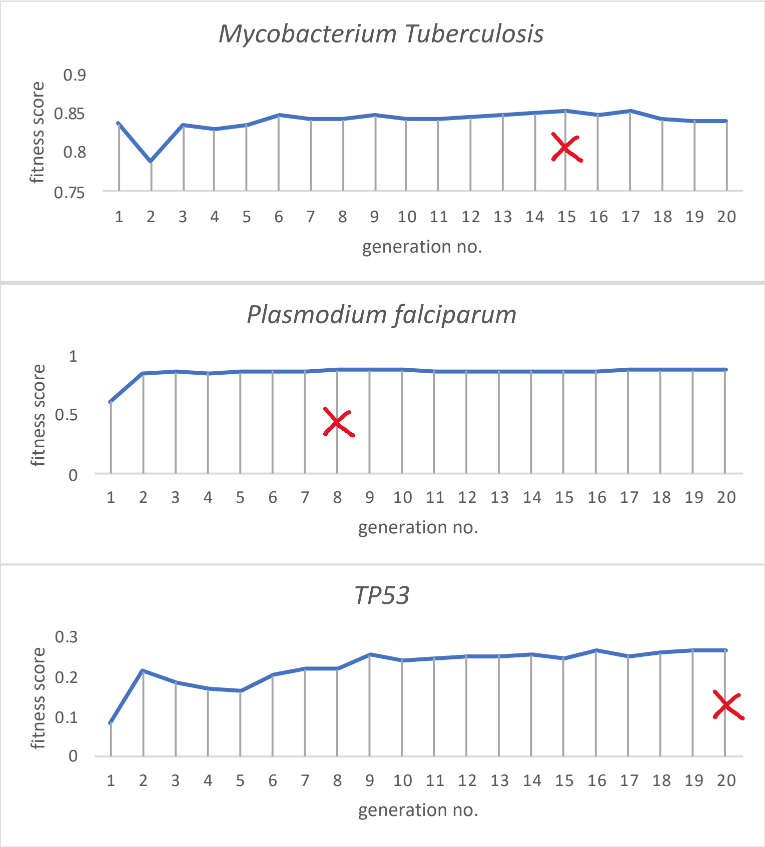


Figure 27: GA runs for each of the sequences tested (the red marks signify the generation where the highest fitness was achieved).

Coverage results from different actual sequencing runs for each genome were tested. Those results with which the GA was able to produce the highest fitness score (similarity of coverage) are presented here. In the case of *M. tuberculosis* the GA successfully produced a set of parameters (**Table 10**) which, when applied to my simulator was able to produce coverage similar to what was found in an actual sequencing run SRR6257109[11] (**Figure 28**). A key trend seen here is the ability of the simulator to better mimic coverage in areas of homogeneous GC-content. The same trend is also seen after running my simulator with the best parameters (**Table 11**) found for the *Plasmodium falciparum* sequence (SRR5161262)[12] (**Figure 29**). Here the mimicking capability is better in areas of homogeneous AT-content. In the coverage comparison for *M. tuberculosis* (**Figure 28**) there is a clear overestimation of coverage in the region between nucleotide position 30000 and the end of the sequence. This is most likely due to the lack of homogeneous nucleotide content (AT- or GC-rich) in this region. A possible reason for this is that the simulator preferentially captures the coverage of homogenous regions while failing to properly capture the features of areas with a higher variability in nucleotide content.

Interestingly, the denaturation temperature chosen by the GA for the GC-rich *M. tuberculosis* sequence was low (85C). Possibly, but this is a speculation, the denaturation temperature in the original experiment was low, leading to lower coverage in GC-rich areas (which are notoriously difficult to denature). If this was the case, then the simulator was particularly able to capture that effect. Similarly, the elongation parameter value chosen for *P. falciparum* was high (75C). This would affect coverage in an AT-rich sequence as such a high temperature would lead to a loss of fragments during PCR.

Table 10: Best parameters for *M. tuberculosis*

Parameter	Value
Fragment distribution mean	224.602
Fragment distribution standard deviation	43.585
Splitting bias probability	0.731
Ligation bias probability	0.360
Denaturation temperature	85.289
Elongation temperature	85.289

[11] https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=run_browser&run=SRR6257109

[12] https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=run_browser&run=SRR6257109

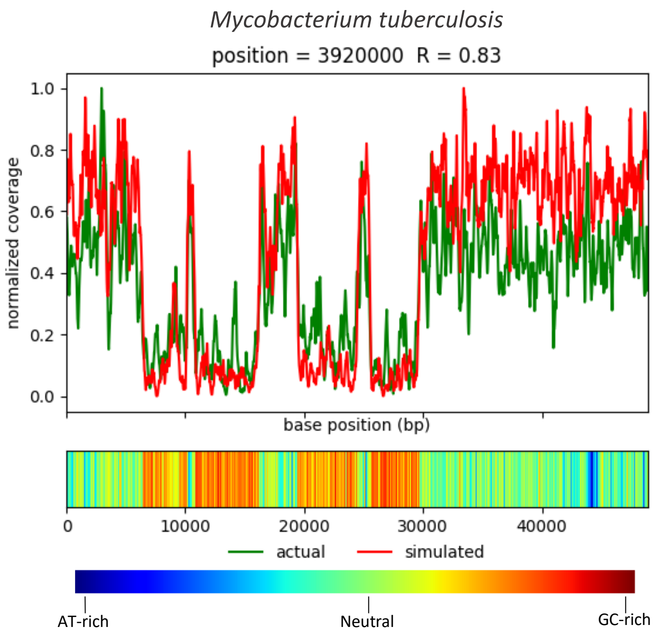


Figure 28: *Mycobacterium tuberculosis* coverage comparison. The colour bar shows levels of base composition bias (blue → red = increasing GC content)

Table 11: Best parameters for *P. falciparum*

Parameter	Value
Fragment distribution mean	464.894
Fragment distribution standard deviation	64.085
Splitting bias probability	0.567
Ligation bias probability	0.455
Denaturation temperature	89.259
Elongation temperature	75.844

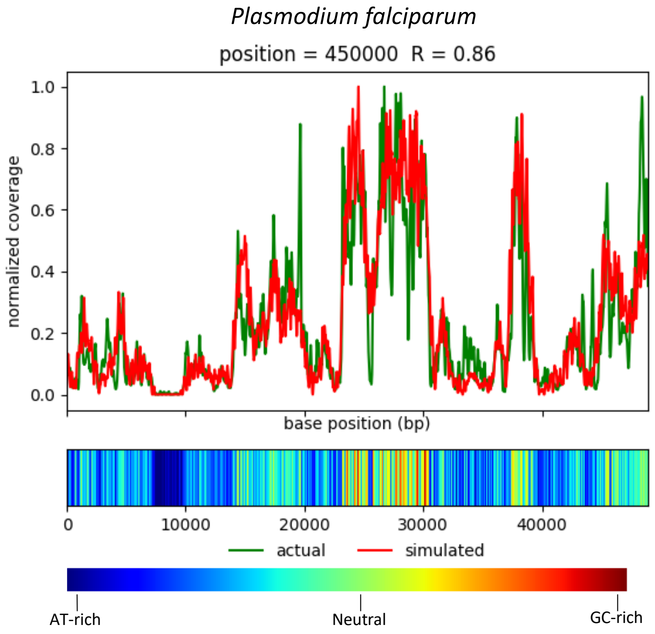


Figure 29: *Plasmodium falciparum* coverage comparison

There were difficulties in obtaining parameters (**Table 12**) to mimic the coverage found in sequencing results for *TP53*[13] (Auton *et al.*, 2015). This is evident from the lower fitness scores attained in the GA run (**Figure 27**) and the poor overlap of the two plots in **Figure 30**. Here there were no large areas of homogeneous biased nucleotide content as the sequence is made of a neutral base composition. This led me to believe the simulator is better able to mimic real world coverage values when there are homogeneous areas of biased nucleotide content.

Table 12: Best parameters for *TP53*

Parameter	**Value**
Fragment distribution mean	490.172
Fragment distribution standard deviation	50.077
Splitting bias probability	0.398
Ligation bias probability	0.329
Denaturation temperature	75.524
Elongation temperature	77.343

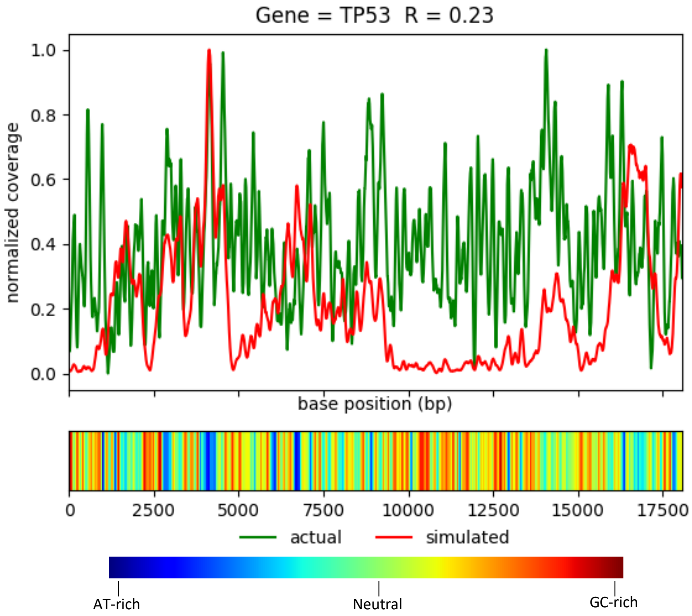


Figure 30: *TP53* coverage comparison

To ensure the failure in obtaining parameters was not due to the genetic structure of *TP53*, a GA run was carried out on an entire contig (AC012627.4[14]) of the human genome containing both coding and non-coding regions. Such regions are usually

[13] HG00154: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/HG00154/

[14] HG00154: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/HG00154/

harder to sequence due to the presence of nucleotide repeats. Once again, the simulator failed to capture coverage from this new region (**Figure 31**). Although the chosen region does include repeats, they do not cover large areas as in the case of *P. falciparum* and *M. tuberculosis*. This further reiterates my previous view that the simulator works better when an input sequence includes large areas of homogeneous nucleotide content. The reason for this occurrence is further discussed in the next section.

Table 13: Best parameters for AC012627.4

Parameter	Value
Fragment distribution mean	525.016
Fragment distribution standard deviation	52.368
Splitting bias probability	0.593
Ligation bias probability	0.446
Denaturation temperature	73.259
Elongation temperature	53.884

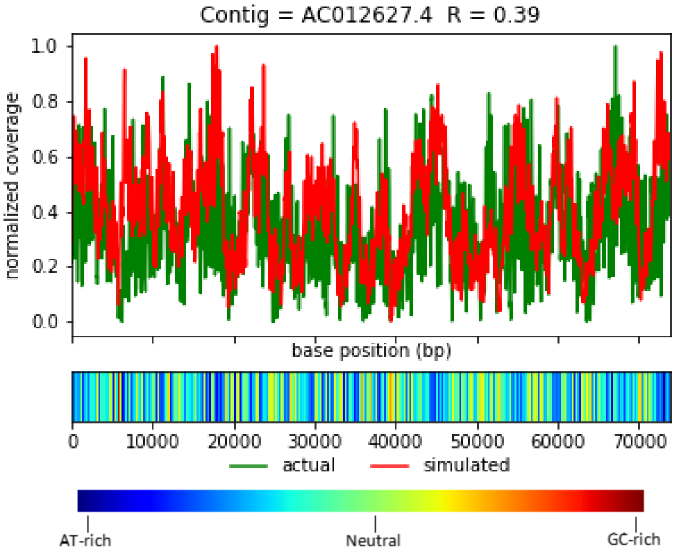


Figure 31: AC012627.4 coverage comparison

4.7 Chapter Summary and Discussion

In this chapter, I set out to find if my simulator is able to produce coverage results that bear a resemblance to what is found in real-word sequencing. The implemented GA was able to derive parameters that led to a good fit between the actual and simulated coverages for the selected regions of DNA samples from *P. falciparum* and *M. tuberculosis* but was less successful for the human samples. The parameters derived for the tested region of *P. falciparum* were also able to provide similar coverage when tested on other parts of the genome (**Figure 32**). This was not the case for *M. tuberculosis* as the parameters derived for the tested region did not always lead to similar coverage for other parts of the genome (**Figure 33**).

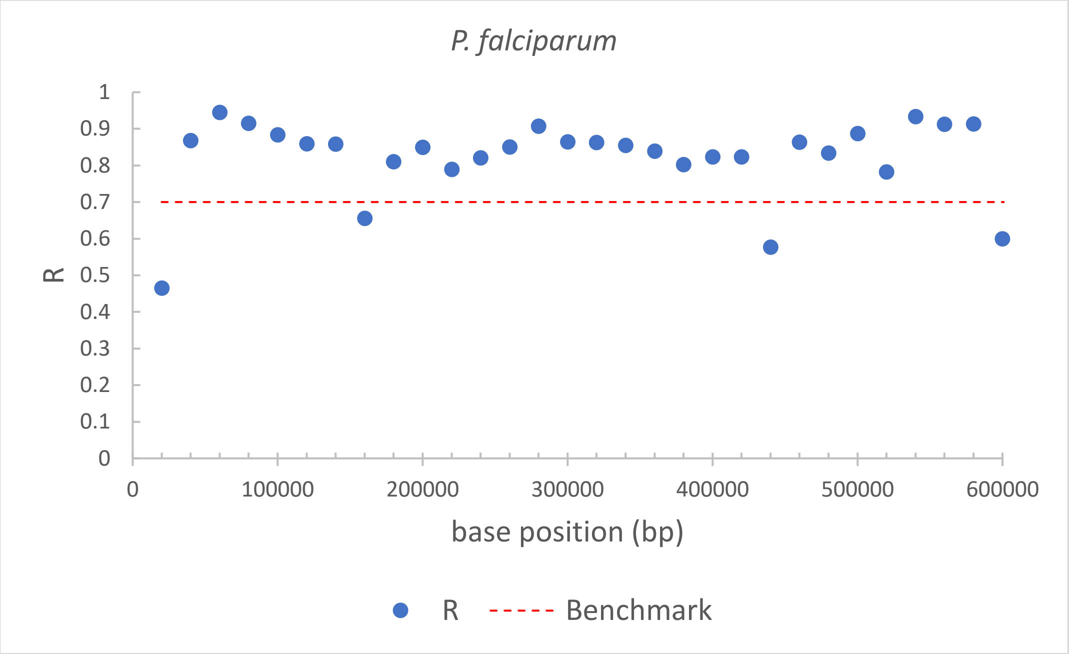


Figure 32: Parameters with the highest fitness score taken from a section of the *P. falciparum* genome are tested on other parts of the genome. Each point is the fitness score (R) for the tested region. The benchmark line is the level at which below it the coverage of the given point lacks similarity to the original coverage values for that region

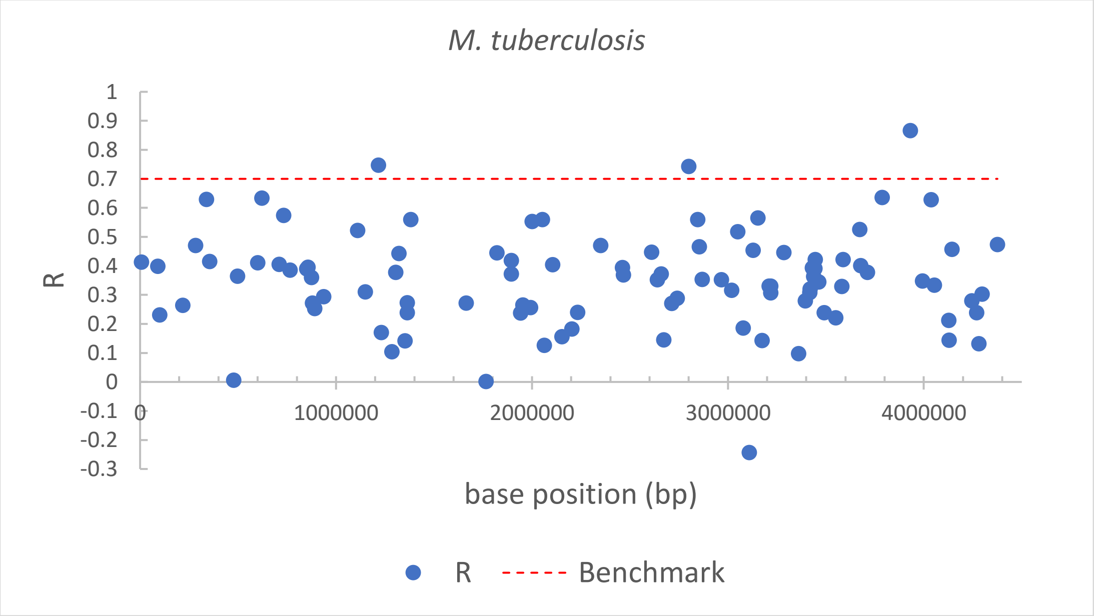


Figure 33: Parameters with the highest fitness score taken from a section of the *M. tuberculosis* genome are tested on other parts of the genome.

The explanation for these observations is that the GA is best able to optimise parameters for sequences that are characterised by well-defined homogeneous areas (i.e. areas that are dominated by two nucleotide types such as in GC-rich or AT-rich regions). The human genome is relatively free of such regions compared to *P. falciparum* and *M. tuberculosis*. In the latter, homogeneous regions (GC-rich) are distributed less regularly over the genome than the homogeneous regions of *P. falciparum* (**Figure 34**).

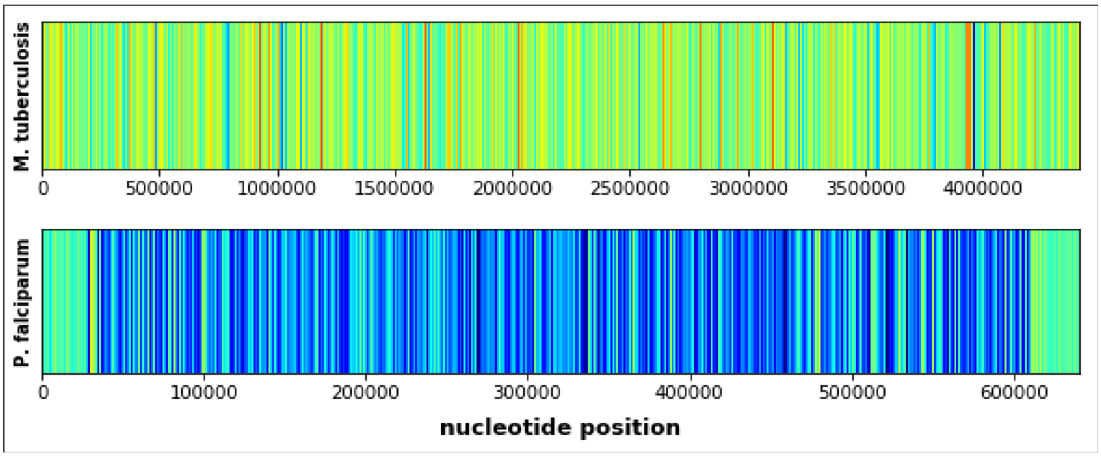


Figure 34: Nucleotide identity across the *P. falciparum* and *M. tuberculosis* genomes.

As the sequencing results used here are not without their own deficiencies, it appears that the GA is particularly able to mimic low coverage areas. However, the good performance of the GA is not caused by low coverage, but because it is better able to make predictions in problem areas that are characterised by homogeneous nucleotide content which would already have poor coverage. This lower coverage could be a result of inappropriate choices of melting and elongation temperatures during PCR; an insufficient melting temperature would fail to denature GC-rich regions of the genome, leading to underrepresentation of these regions. While an elongation temperature that is too high affects the cloning of AT-rich fragments. The implication of this is that using parameters from the GA, the simulator is better able to reconstruct the shortcomings of the original sequencing procedures. In the next chapter, the individual and combined effects of each simulated library preparation parameter is tested on different types of DNA sequences.

Chapter 5 Effects of Library Preparation

This chapter examines the individual and combined effects of the three implemented steps of library preparation: fragmentation, ligation and amplification. The aim is to evaluate the extent to which these steps lead to a deviation from optimal (uniform) coverage. The uniformity of coverage is measured using the evenness score (E) as described in section 3.5.

The effects of the library preparation steps on evenness of coverage were analysed using multiway analysis of variance (ANOVA) with replication in two stages. In the first stage, two independent variables were tested, the first being the DNA structure (S.DNA) and the second representing one of the three library preparation steps. In all cases, evenness of coverage is the dependent variable.

As an example, the library preparation step of attaching adaptors to the fragments is represented by chosen values of the parameter ligation bias. These values are the levels of the independent variable representing ligation bias (B.LIGATION) and the test is a two-factor ANOVA (S.DNA and B.LIGATION) with replication where S.DNA represents different types of DNA sequence structures (The tested sequences are outlined below). Besides assessing the proportion of variance accounted for by the main factors S.DNA and B.LIGATION, the ANOVA also evaluates their interaction, i.e., in how far the effect of one independent variable depends on the levels of the other. For an overview of the other independent variables and their levels, see **Table 14**.

Table 14: Overview of independent variables.

Library Preparation Step	Parameter	Independent Variable	Levels (Increment)	Default Parameter Value [3]
Fragmentation	Mean[1]	M.SIZE	100 – 1000 (100)	300
	Skewness [1,2]	SKEW	10 – 100 (10)	30
	Splitting Bias	B.SPLIT	0.0 – 1.0 (0.1)	0
Ligation	Ligation Bias	B.LIGATION	0.0 – 1.0 (0.1)	0
Amplification	Denaturation Temperature	T.DENAT	90 – 100 (1)	120
	Elongation Temperature	T.ELON	60 – 74 (2)	60

[1] Parameter of the fragment size distribution (lognormal).

[2] Measured as standard deviation. The relationship between skewness and standard deviation is described in the text.

[3] These are tested baseline values at which there is no effect on coverage output.

In the second stage, three independent variables were analysed; DNA structure (S.DNA) was tested together with a combination of two of the three library preparation steps.

The main assumptions of the ANOVAs were checked by means of Kolmogorov - Smirnov and Shapiro - Wilk tests (for normality) and Levene’s test (for heteroscedasticity) (For results of the verification of assumptions, see appendix A). Violations of the assumptions was one of the reasons to carry out several ANOVAs instead of one overall test including all six library preparation variables. Also, such an ANOVA would be difficult to interpret because of the large number and complexity of interactions. All ANOVA results were generated using IBM SPSS Statistics for Windows, Version 25.0 (IBM Corp., 2017).

Concerning DNA structure, natural genomes differ by their nucleotide content and sequential dependency. Some are characterised by having high GC-content or having high AT-content or some other deviations from equal proportions of nucleotide bases. In addition, the bases may not be distributed independently of each other thus forming a heterogeneous (“clumped”) sequence. To study how far these structural features affect the evenness of coverage, the following four types of artificial DNA sequences, each 20,000 bp long, representing these characteristics were generated for this part of my study:

- Sequence 1 (GC80) is GC-rich: 40% of its nucleotides are G and 40% are C (i.e. a GC composition of 80 %).
- Sequence 2 (AT80) has an AT composition of 80% (“AT-rich”).
- Sequence 3 (GCAT80) consists of two regions of 5000bp situated at each side of the central base position, the first has an 80% GC content and the second an 80% AT content. The remaining regions of 5000bp at the start and end of the sequence have a neutral base composition (A:25%, C:25%, G:25%, T:25%).
- Sequence 4 (GC50) has neutral base composition (A:25%, C:25%, G:25%, T:25%).

These sequences respectively represent genomes with high GC-content, high AT-content, with clumped areas of biased nucleotide content (AT-rich and GC-rich) and with equal quantities of all bases.

5.1 Single Effects

The effects of the separate library preparation steps on evenness of coverage were analysed using a two-way ANOVA. Two independent variables were tested, the first being the DNA structure (S.DNA) and the second being one of the six parameters associated with the library preparation steps (See Chapter 3).

5.1.1 Fragmentation

This section is devoted to the effects of the distribution of fragment sizes on coverage uniformity. Also, attention will be paid to a possible fragmentation bias, where splitting preferentially occurs between CpG dinucleotides (Poptsova *et al.*, 2014). This bias is discussed in Chapter 2.

In my model, fragmentation was modelled by drawing values, representing fragment sizes, from a lognormal distribution (Section 3.2). This distribution is characterised by a shape and location parameter. The first is related to the standard deviation of the distribution and the second to the mean of lognormally distributed fragment sizes. The values of the mean were varied between 100 and 1000 with increments of 100 and the standard deviation between 10 and 100 with increments of 10. These values were transformed to the location and shape parameter of the lognormal distribution respectively and used to generate fragment sizes. Modifying the standard deviation here affects the skewness of the lognormal distribution (see Section 0 for a formal description of this relationship).

Mean Fragment Size

For each of the artificial DNA sequences, the value of E was plotted for varying mean fragment sizes. The same trend can be seen in all four artificial sequences (Figure 35): E is high for mean fragment sizes between 200bp and 500bp and declines as mean fragment size increases from 500bp to 1000bp.

The results of the ANOVA show that the main factors, mean fragment size (M.SIZE, p = 0.000) and DNA structure (S.DNA, p = 0.000), both have a statistically significant effect on coverage uniformity (**Table 15**). However, the interaction between S.DNA and M.SIZE is not significant (p = 0.430), implying that the effects of a main factor do not depend on the level of the other.

Figure 35 confirms this as the line plots retain the same shape for each type of DNA structure: the only difference found is the height of the curves, with the lowest values found in GCAT80 (see also **Figure 37**). It demonstrates the main effect of DNA structure as stated above, the significantly lower evenness of coverage of DNA with sequentially dependent ("clustered") nucleotides (GCAT80), and the absence of an interaction effect.

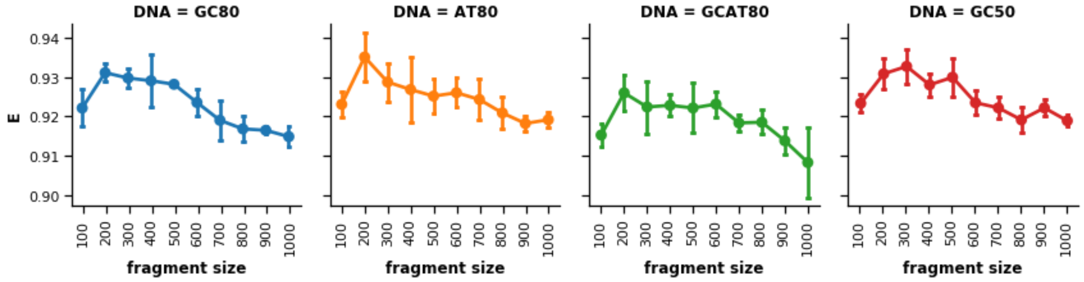


Figure 35: Effects of mean fragment size on the uniformity of coverage for all four generated sequences. For all figures in this chapter, the error bars are standard deviations.

Table 15: ANOVA of S.DNA and M.SIZE.

Dependent Variable: E

Source	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	F – Value (F)	P – Value (p)
S.DNA	0.001	3	0.000	17.202	0.000
M.SIZE	0.005	9	0.001	23.007	0.000
S.DNA * M.SIZE	0.001	27	2.283E-5	1.032	0.430
Error	0.004	160	2.212E-5		
Total	0.010	199			

Skewness

Skewness was altered by varying the standard deviation of the fragment size distribution; for the lognormal distribution, skewness is (almost) linearly dependent on standard deviation, especially for smaller mean values (**Figure 36**).

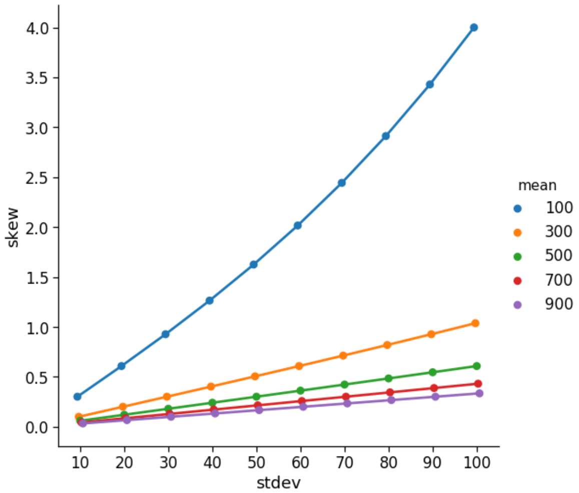


Figure 36: Skew dependence on standard deviation for different mean values.

The effects of skewness (SKEW) for each of the four types of DNA sequences are shown in **Figure 37**. As expected, the effects of DNA structure are significant (p = 0.000). The factor S.DNA was kept in the ANOVA design because the second factor (SKEW) and the possible interaction S.DNA * SKEW, lead to a different error term (and hence result in different F and p values) than when left out.

The ANOVA reveals that the effect of SKEW is not statistically significant (p = 0.150) (

Table 16) and that there is no indication of an interaction effect between S.DNA and SKEW (p = 0.562). As before, the effect of DNA type shows up as lower values seen in the heterogeneous series GCAT80 (**Figure 37**).

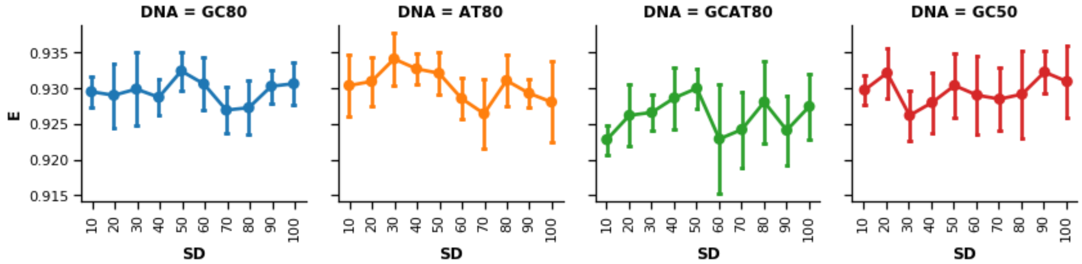


Figure 37: Effects of the skewness of the fragment distribution on uniformity of coverage.

Table 16: ANOVA of S.DNA and SKEW.

Dependent Variable: E

Source	SS	df	MS	F	p
S.DNA	0.001	3	0.000	8.793	0.000
SKEW	0.000	9	3.117E-5	1.505	0.150
S.DNA * SKEW	0.001	27	1.936E-5	0.935	0.562
Error	0.003	160	2.072E-5		
Total	0.005	199			

Splitting Bias

In the model, non-random DNA fragmentation is simulated by a parameter governing the probability of a split between a C and a G (see Section 0). The values of this parameter are the levels of the independent variable (B.SPLIT).

The main factors B.SPLIT (p = 0.000) and S.DNA (p = 0.000) both have a statistically significant effect on coverage uniformity. There is also a statistically significant interaction (S.DNA * B.SPLIT, p = 0.000) (**Table 17**).

Coverage uniformity appears to be somewhat lower for the highest levels of B.SPLIT. This trend is especially noticeable in the clumped sequence (GCAT80), with a sharp decline in E for B.SPLIT in the range from 0.7 to 0.9 (**Figure 38**). This difference is the cause of the significant interaction (S.DNA * B.SPLIT) in the ANOVA. Note that of all three variables representing aspects of fragmentation (M.SIZE, SKEW and B.SPLIT), B.SPLIT most strongly elevates the effects of DNA structure.

Table 17: ANOVA of S.DNA and B.SPLIT

Dependent Variable: E

Source	SS	df	MS	F	p
S.DNA	0.005	3	0.002	76.711	0.000
B.SPLIT	0.013	8	0.002	71.563	0.000
S.DNA * B.SPLIT	0.019	24	0.001	35.013	0.000
Error	0.003	144	2.279E-5		
Total	0.041	179			

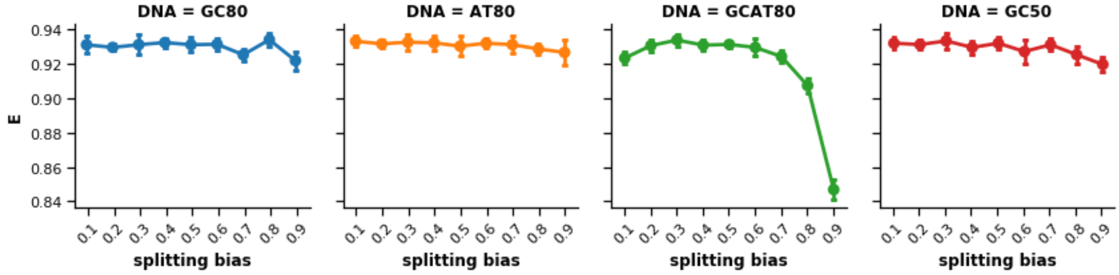


Figure 38: Effects of non-random fragmentation bias on the uniformity of coverage.

5.1.2 Ligation

The ligation of adapters to fragments during library preparation could influence coverage, as it determines which fragments will be cloned during PCR. As explained in Chapter 3, this process is modelled in my simulation by a ligation bias parameter, which reflects the likelihood of fragments to be ligated given the identity of their terminal base. The ligation bias parameter is the probability with which a fragment with a T at the 5' end will be attached to an adaptor. Because adapters are biased against fragments with a T on their 5' end, a high value of this parameter corresponds to a low binding affinity. To investigate the possible effect of ligation bias, the coverage uniformity at different values of the bias are compared. These values are the levels of the independent variable B.LIGATION

Both of the main factors (B.LIGATION and S.DNA) have a significant effect on coverage uniformity (p = 0.000 and p = 0.000 respectively) (**Table 18**). The significant interaction term (p = 0.000) means the effect of B.LIGATION depends on the DNA structure of the sequences tested. Whereas the AT-rich (AT80) and clumped sequence (GCAT80) show a decline in E as B.LIGATION increases (**Figure 39**), the GC-rich sequence does not show any trend, while the neutral base composition sequence (GC50) shows a less steep reduction in E compared to the sequences with high AT content.

Table 18: ANOVA of S.DNA and B.LIGATION.

Dependent Variable: E

Source	SS	df	MS	F	p
S.DNA	0.008	3	0.003	81.782	0.000
B.LIGATION	0.015	8	0.002	58.798	0.000
S.DNA * B.LIGATION	0.008	24	0.000	10.776	0.000
Error	0.005	144	3.152E-5		
Total	0.035	179			

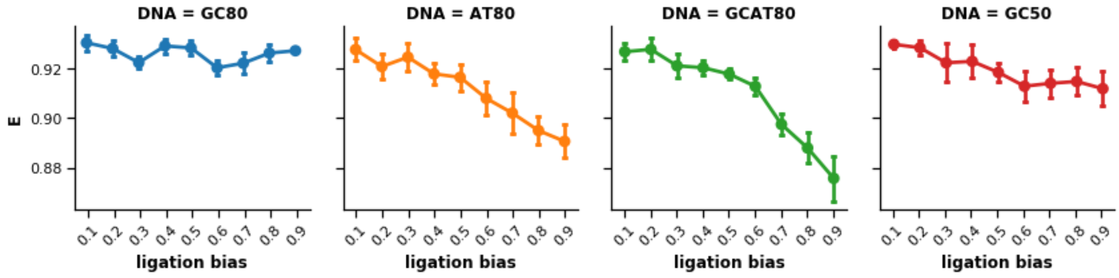


Figure 39: Effects of the ligation bias on coverage uniformity.

5.1.3 Amplification

The thermodynamics at play during the PCR amplification process can pose a challenge in situations where inappropriate (low) temperatures are selected. In my simulation, the denaturation and elongation stages of PCR are modelled by parameters expressing the temperature at which a template DNA strand disassociates from its complementary strand (see Chapter 3). A range of temperatures is applied to a sequence during the PCR denaturation and elongation phases to determine their effects on coverage.

For denaturation, temperatures from 90°C to 100°C with increments of 1°C were chosen. This covers the 94°C to 98°C range used in standard PCR protocols (Lorenz, 2012). A temperature range from 60°C to 74°C with increments of 2°C was chosen to test the effects of elongation. This range includes the conventionally used temperature of 72°C (Innis & Gelfand, 1999). In both cases, the ranges were chosen in order to check if the routinely used values were the only appropriate temperatures. The selected values of the denaturation and elongation temperatures are the levels of the independent variables T.DENAT and T.ELON respectively.

Denaturation

The ANOVA indicates the main factors (T.DENAT and DNA) as both having a statistically significant effect on coverage uniformity (p = 0.000 and p = 0.000 respectively) (**Table 19**). Also, the interaction between DNA and T.DENAT is significant (p = 0.000).

The sequences with high GC-content (GC80 and GCAT80) perform worst (have lower values of E) between 90°C and 93°C (**Figure 40**). At higher temperatures, E converges to its maximum. For the sequences with lower levels and average levels of GC (AT80 and GC50), E attains maximal uniformity and does not vary across the range of temperatures. These differences in coverage uniformity between the sequences with high and low GC-content explain the significant effect of DNA base composition and the interaction term. The coverage plot in **Figure 41** shows how a low denaturation temperature (90°C) affects coverage in the GC-rich region (red area in the colour bar) of GCAT80.

Table 19: ANOVA of S.DNA and T.DENAT.

Dependent Variable: E

Source	SS	df	MS	F	p
S.DNA	0.305	3	0.102	3432.120	0.000
T.DENAT	0.660	10	0.066	2229.626	0.000
S.DNA * T.DENAT	0.930	30	0.031	1047.855	0.000
Error	0.005	176	2.958E-5		
Total	1.900	219			

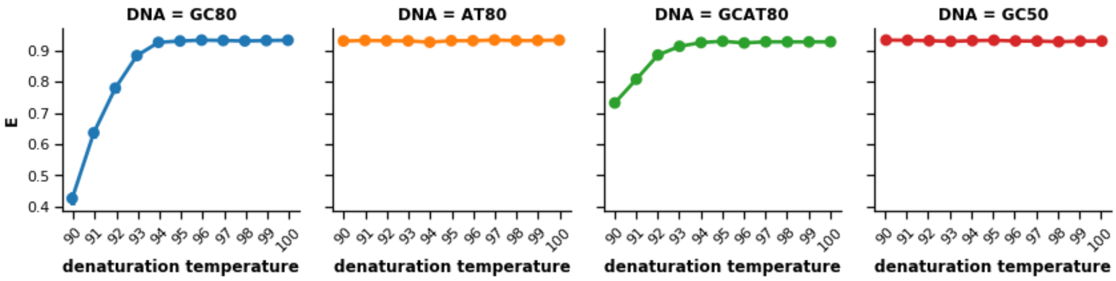


Figure 40: Effects of denaturation temperature on coverage uniformity.

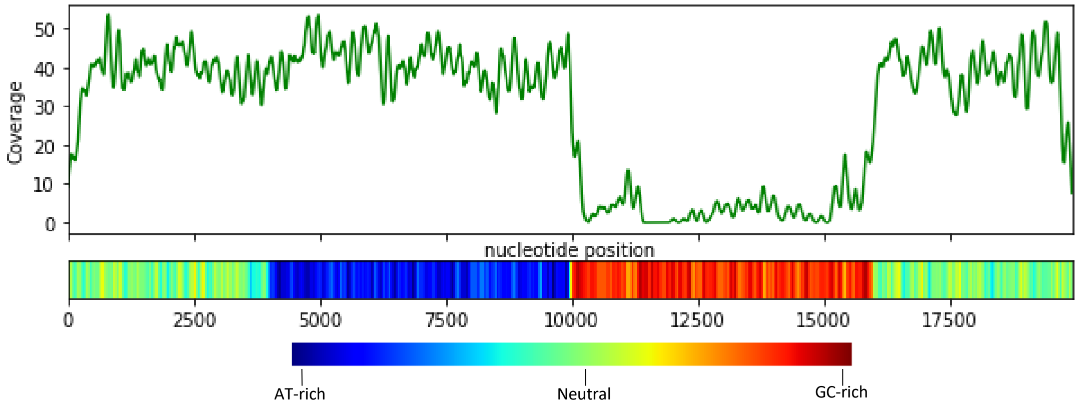


Figure 41: Coverage plot of GCAT80 sequence. The colour bar shows levels of base composition bias (blue - red = increasing GC content).

Elongation

Because the data did not meet the assumptions of normality (Kolmogorov-Smirnov test: p = 0.022), I applied a one-way ANOVA for each separate DNA sequence type (**Table 20**). The effect of elongation temperature (T.ELON) was significant in the AT-rich sequence (AT80) (p = 0.000) and clumped sequence (GCAT80) (p = 0.000), but not in the other two sequences (GC80 and GC50) (p = 0.457 and p = 0.366 respectively).

The results suggest that sequences with high AT content (AT80 and GCAT80) suffer from reduced levels of E at higher elongation temperatures (72°C to 74°C) (**Figure 42**). At lower temperatures (60°C to 70°C) E remains stable through the range. In sequences with lower and average levels of AT content (GC80 and GC50), the range of temperatures has no effect on E. The effect of a high elongation temperature (74°C) on the AT-rich region (blue area in colour bar) of GCAT80 can be seen in **Figure 43**.

Table 20: ANOVA of S.DNA and T.ELON.

Dependent Variable: E

DNA	Source	SS	df	MS	F	p
AT80	T.ELON	0.618	7	0.088	1582.283	0.000
	Error	0.002	28	5.580E-5		
	Total	0.620	35			
GC50	T.ELON	0.000	7	3.532E-5	1.135	0.366
	Error	0.001	32	3.110E-5		
	Total	0.001	39			
GC80	T. ELON	0.000	7	1.900E-5	0.988	0.457
	Error	0.001	32	1.923E-5		
	Total	0.001	39			
GCAT80	T.ELON	0.229	7	0.033	1848.080	0.000
	Error	0.001	32	1.773E-5		
	Total	0.230	39			

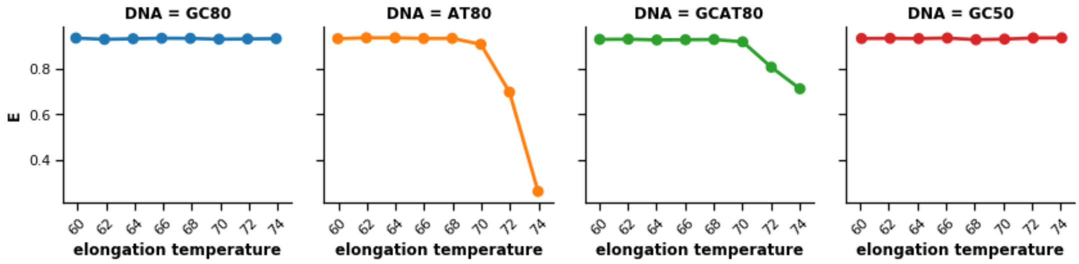


Figure 42: Effects of elongation temperature on coverage uniformity.

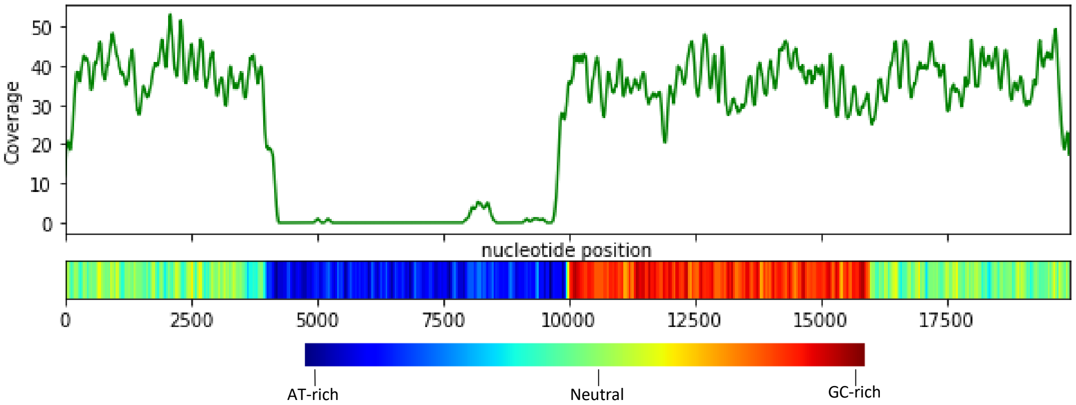


Figure 43: Coverage plot of GCAT80 sequence. The colour bar shows levels of base composition bias (blue - red = increasing GC content).

5.1.4 Amplification-Free

In Section 2.6.4, Kozarewa and colleagues' amplification-free library preparation method, which skips the PCR stage was explored (Kozarewa *et al.*, 2009). To assess the benefits of this approach, all parameters were retested without the PCR module. The results of this test (**Figures Figure 44,Figure 45,Figure 46 and Figure 47**) match the previous results found when static PCR parameters emulating optimal denaturation and elongation temperatures (**Table 14**) were used. This shows that excluding the PCR stage is indeed beneficial in avoiding the coverage deficiencies caused by it, but on the other hand this exclusion has no real influence on the individual effects of its preceding stages (fragmentation and ligation).

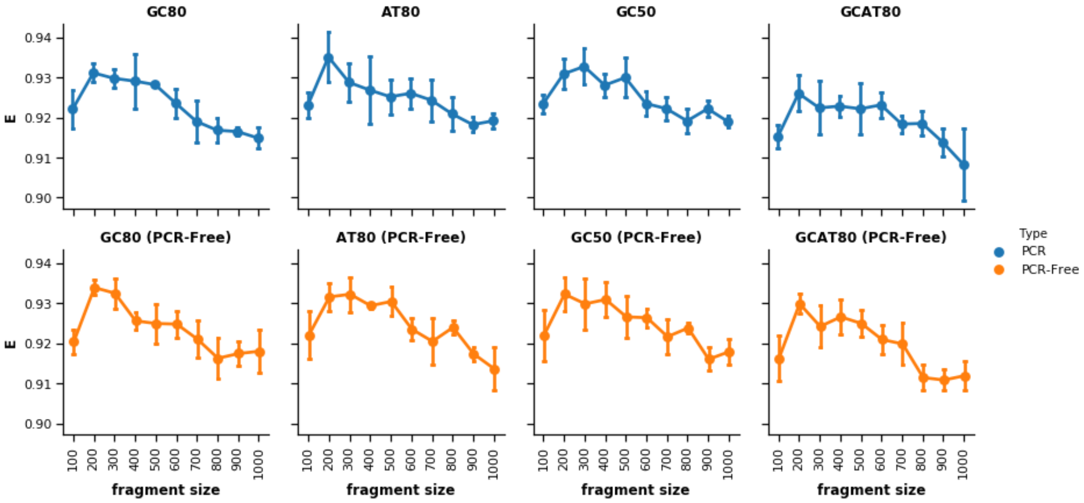


Figure 44: PCR-free mean fragment size test

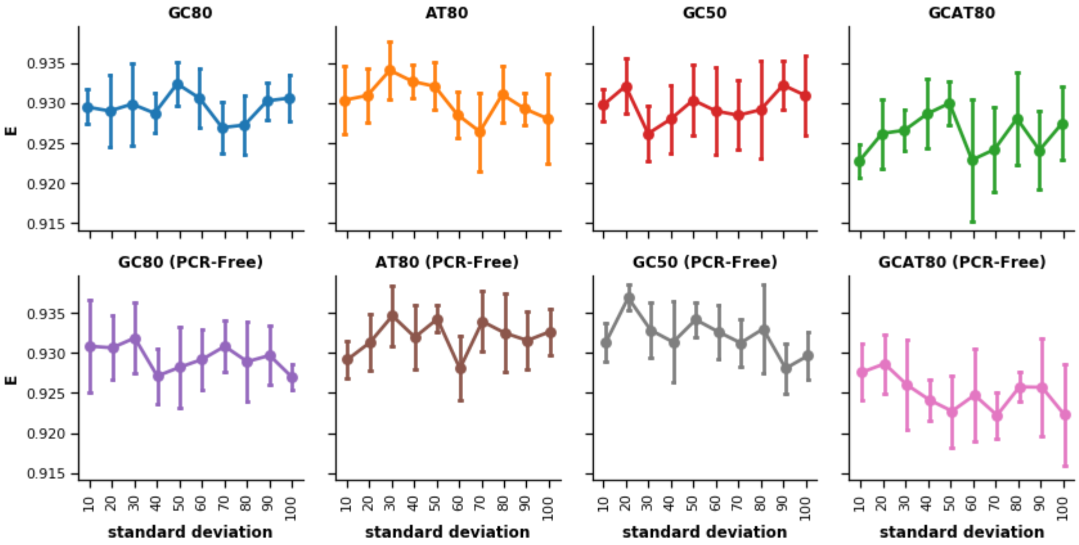


Figure 45: PCR-free skewness test.

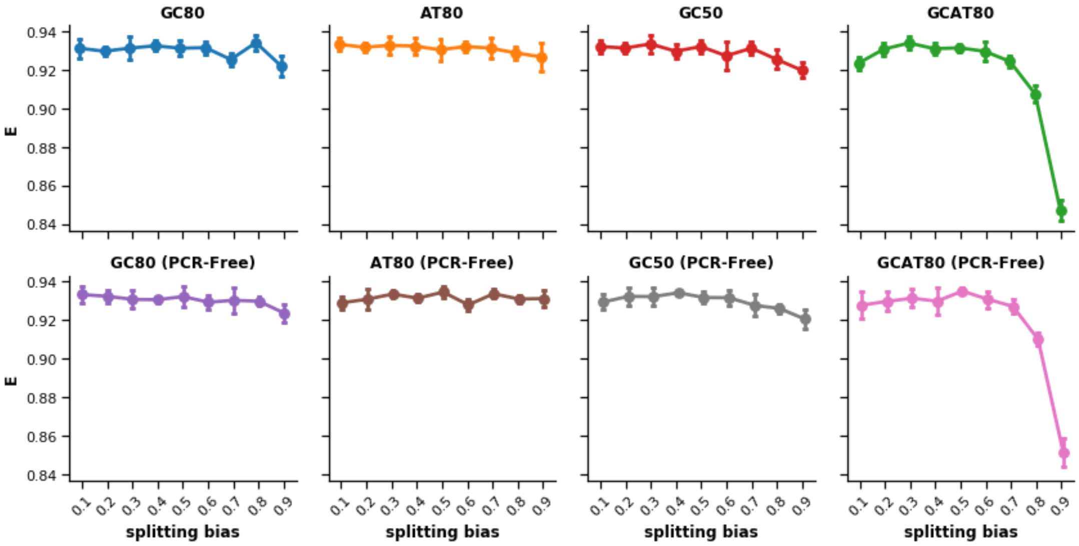


Figure 46: PCR-free splitting bias test.

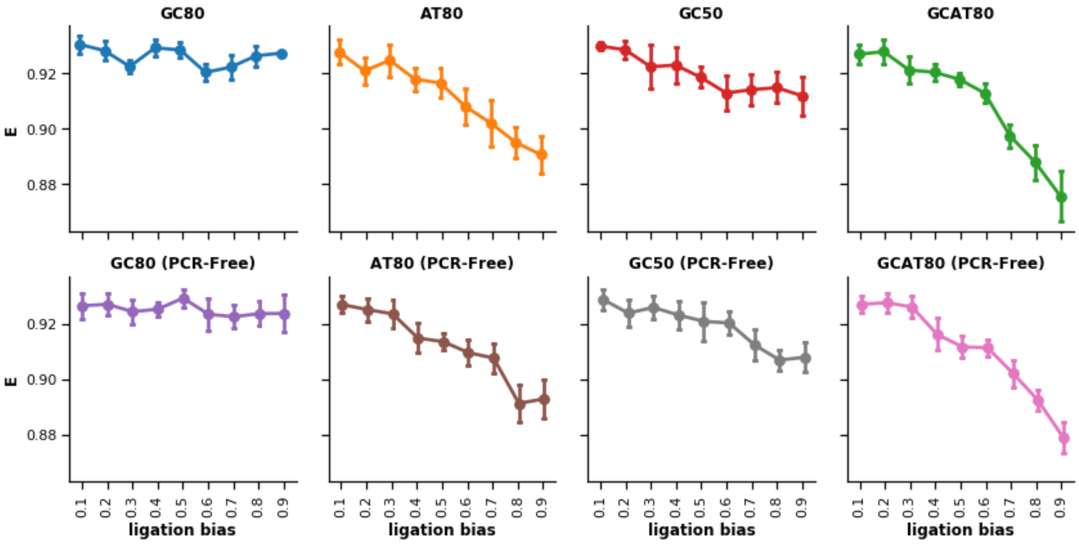


Figure 47: PCR-free ligation bias test.

5.2 Combined Effects

This section examines the combined effects of the three implemented steps of library preparation. This is done using a multi-way ANOVA of three independent variables the first being the DNA structure (S.DNA) the other two being combinations of the six parameters associated with the library preparation steps. Only significant effects (listed in **Table 21**) are discussed. The library parameter SKEW is not considered here, because, as shown in Section 5.1.1, it has no effect on uniformity of coverage.

This analysis is carried out to examine the possible knock-on effects of preceding library steps, that is how far does a preceding step combine with the next step to affect coverage uniformity. For example, in the fragmentation step, a strong splitting bias creates a larger number of fragments that begin with a C and end with a G (Poptsova *et al.*, 2014). This effect was also observed in my analysis of the splitting bias (Section 5.1.1). The effect may interact with the subsequent step, ligation, where higher levels of the ligation bias lead to a loss of fragments starting with a T and ending with an A (see Section 2.6.3). Therefore, in a sequence with areas of biased nucleotide content (i.e. GC-rich and AT-rich), the splitting bias would cause the majority of splits to occur in the GC-rich regions, leading to a lower representation of AT-rich regions. In turn, this may be exacerbated by the ligation bias, which will further reduce region representation due to the loss of AT-rich fragments. The question then becomes, to what extent would this affect uniformity of coverage, which will be explored in detail in the next section.

Table 21: Results of combined effects analysis

	M.SIZE	B.SPLIT	B.LIGATE	T.DENAT	T.ELON
M.SIZE		0	0	1	1
B.SPLIT			1	0	0
B.LIGATE				0	0
T.DENAT					0
T.ELON					

1 = Significant, 0 = Not significant

5.2.1 Splitting Bias and Ligation Bias

An analysis of the combined effects of varying levels of the splitting bias (B.SPLIT) and ligation bias (B.LIGATION) was carried out. The three-way interaction (S.DNA, B.SPLIT and B.LIGATE) and two-way interaction (B.SPLIT and B.LIGATE) were significant (p = 0.000 and p = 0.001 respectively) (**Table 22**). The reason for the statistical significance of the interaction between the three main factors is due to the steep decline of coverage uniformity in the clumped sequence (GCAT80) (**Figure 48**). This decline is especially noticeable for maximum splitting bias in combination with the coverage uniformity reducing effects of increasing ligation bias. In other words, coverage uniformity reduces with increasing levels of both biases in the clumped sequence (GCAT80). Thus, the combination of both biases leads to a stronger effect on coverage uniformity than each would have on its own.

Table 22: ANOVA of S.DNA, B.SPLIT and B.LIGATE.

Dependent Variable: E

Source	SS	df	MS	F	p
S.DNA	0.103	3	0.034	1066.022	0.000
B.LIGATE	0.062	3	0.021	638.810	0.000
B.SPLIT	0.051	3	0.017	531.294	0.000
S.DNA * B.LIGATE	0.031	9	0.003	107.249	0.000
S.DNA * B.SPLIT	0.106	9	0.012	364.754	0.000
B.LIGATE * B.SPLIT	0.001	9	0.000	3.430	0.001
S.DNA * B.LIGATE * B.SPLIT	0.006	27	0.000	6.336	0.000
Error	0.008	256	0.000		
Total	0.368	319			

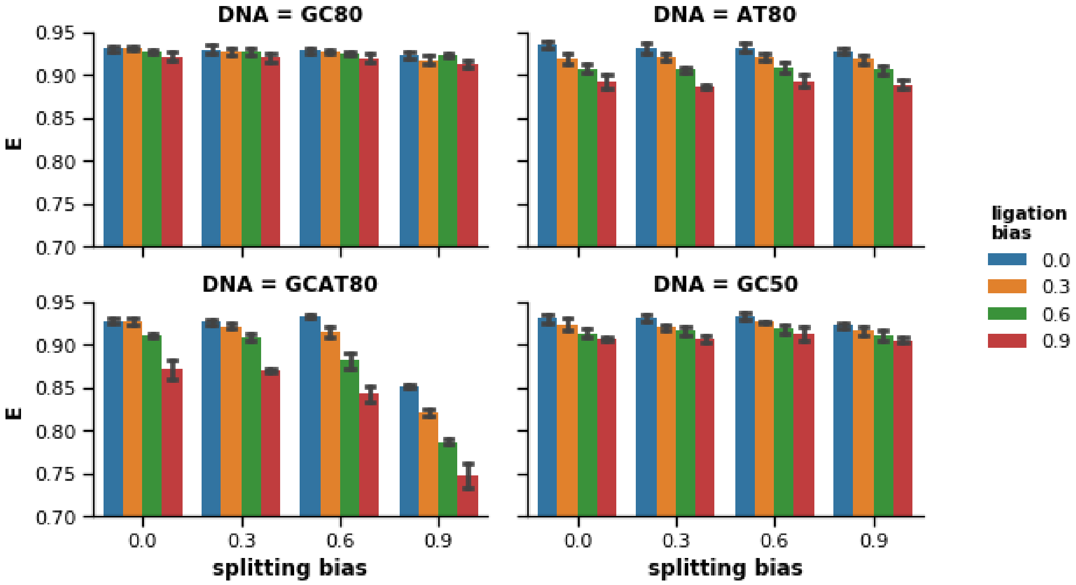


Figure 48: Effects of B.SPLIT and B.LIGATE on E.

5.2.2 Fragment Size and Denaturation

This section deals with the effects of DNA structure (S.DNA) along with those of mean fragment size (M.SIZE) and denaturation temperature (T.DENAT). The ANOVA shows that all main effects and interactions are significant (Table 23). The effect of the interactions can be seen in **Figure 49**: for the GC-rich sequence (GC80) coverage is less uniform for larger fragment sizes when the denaturation temperature is low (94°C). To a lesser extent, the same holds for the clumped sequence GCAT80.

Table 23: ANOVA of S.DNA, M.SIZE and T.DENAT

Dependent Variable: E

Source	SS	df	MS	F	p
S.DNA	0.011	3	0.004	122.723	0.000
T.DENAT	0.015	3	0.005	172.008	0.000
M.SIZE	0.014	3	0.005	156.824	0.000
S.DNA * T.DENAT	0.033	9	0.004	127.659	0.000
S.DNA * M.SIZE	0.012	9	0.001	46.719	0.000
T.DENAT * M.SIZE	0.020	9	0.002	75.035	0.000
DNA * T.DENAT * M.SIZE	0.033	27	0.001	41.539	0.000
Error	0.007	256	0.000		
Total	0.145	319			

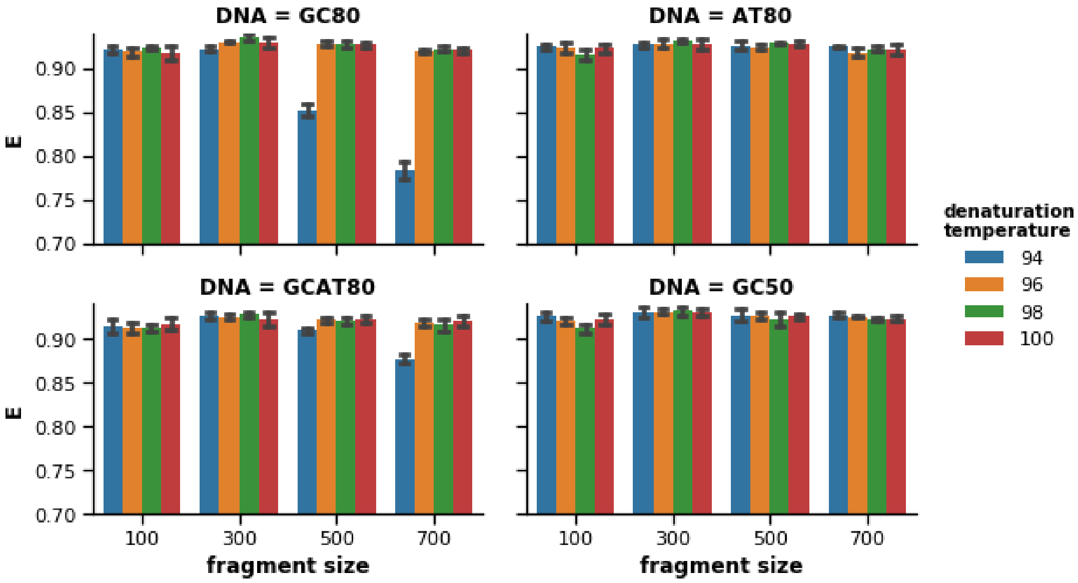


Figure 49: Effects of M.SIZE and T.DENAT on E. Note the low values of E for the low denaturation temperature (94°C) in GC-rich sequences (GC80 and GCAT80), especially for larger fragment sizes.

5.2.3 Fragment size and Elongation

With respect to the combined effects of mean fragment size and elongation temperature, my results show that both the three-way interaction (S.DNA, M.SIZE and T.ELON) and two-way interaction (M.SIZE and T.ELON) are significant (p = 0.000 and p = 0.000 respectively) (Table 24). A similar trend can be seen for the AT-rich sequence (AT80) and clumped sequence (GCAT80); larger fragments sizes in combination with higher elongation temperatures bring about a reduction in coverage uniformity (E) (**Figure 50**).

Table 24: ANOVA of S.DNA, M.SIZE and T.ELON

Dependent Variable: E

Source	SS	df	MS	F	p
DNA	0.003	3	0.001	50.127	0.000
T.ELON	0.004	3	0.001	66.150	0.000
M.SIZE	0.004	3	0.001	68.825	0.000
DNA * T.ELON	0.007	9	0.001	41.380	0.000
DNA * M.SIZE	0.001	9	0.000	7.068	0.000
T.ELON * M.SIZE	0.002	9	0.000	9.665	0.000
DNA * T.ELON * M.SIZE	0.003	27	0.000	5.880	0.000
Error	0.005	256	0.000		
Total	0.030	319			

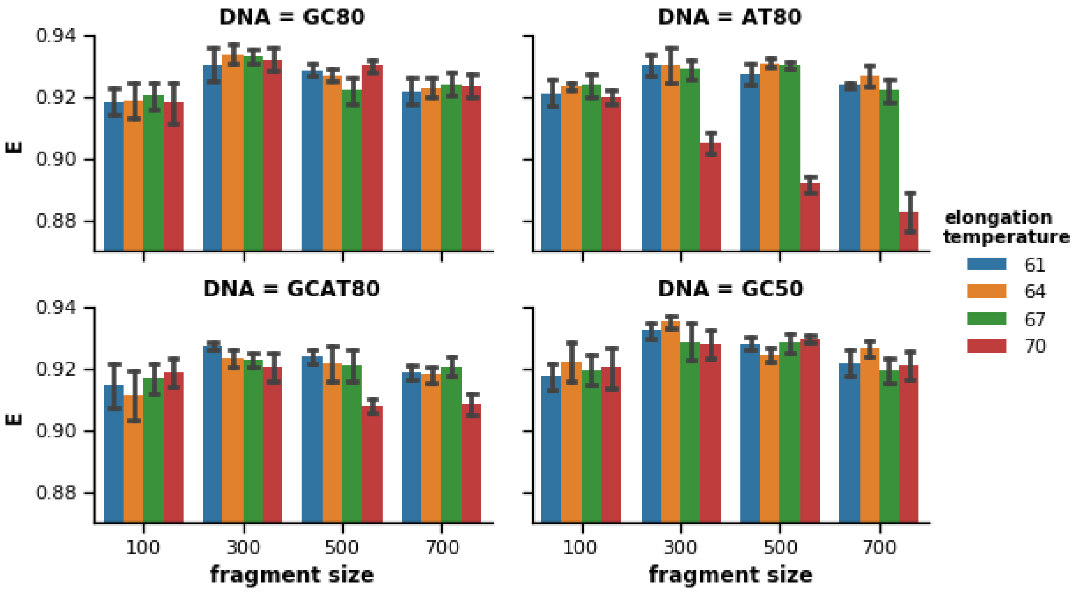


Figure 50: Effects of M.SIZE and T.ELON on E. Note the low values of E for the high elongation temperatures in AT-rich sequences (AT80 and GCAT80), especially for larger fragment sizes.

5.3 Validation of tests with actual DNA

To ensure the above results have a resemblance to what occurs in reality a set of real DNA sequences with features similar to those of the previously used artificial sequences were tested. The effects of each stage of library preparation are tested here and a comparison of results from real and artificial DNA are made. The chosen DNA sequences and their matching artificial sequences are presented below (**Table 25**). A common trend that will be found in the comparisons below is the lower evenness of coverage for *Tuberculosis* and *Plasmodium.* This is due the varying sequential dependencies in different genomes leading to difficulties in delivering similar levels of coverage uniformity for each genome.

Table 25: Matched real and artificial DNA sequences.

Real DNA sequence	Artificial DNA sequence
Mycobacterium tuberculosis[1]	GC80
Plasmodium Falciparum[2]	AT80
Human (TP53 gene)[3]	GC50

[1] GC-content: 70.46%, Region: 3,920,000bp – 3,970,000bp

[2] GC-content: 19.26%, Region: 450,000bp – 500,000bp

[3] GC-content: 47.77%, Region: 7,668,401bp – 7,687,550bp

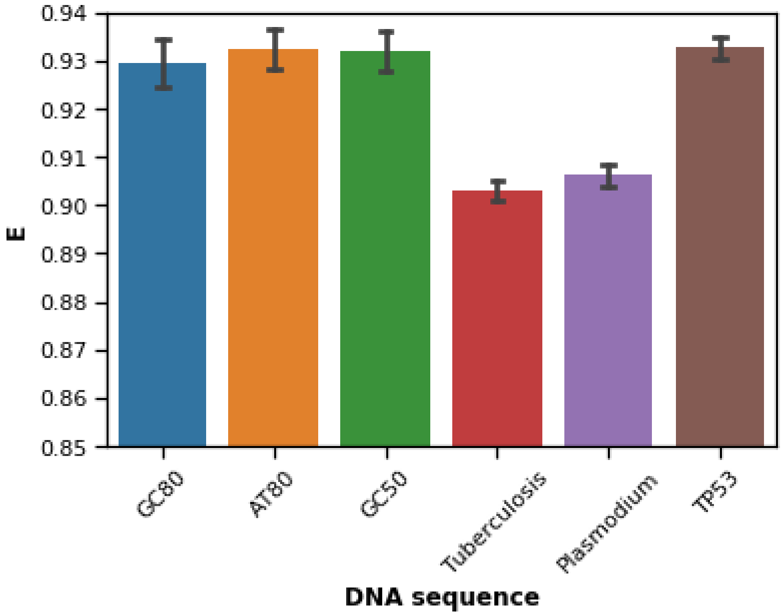


Figure 51: Comparison of evenness of coverage for real and artificial DNA sequences.

Mean Fragment Size

The effects of applying different mean fragment sizes to the selected actual sequences differ to an extent when compared to the artificial sequences (**Figure 52**). When comparing the GC- and AT-rich sequences the lower E observed when an average fragment size below 200bp is used can be observed in both real and artificial sequences. However, a difference can be seen when larger fragment sizes are used (>200bp). In the real sequences the level of E does not reduce by as much when average fragment size increases as it does in the artificial AT- and GC-rich sequences. The trend in both neutral sequences are very similar, showing only minor differences.

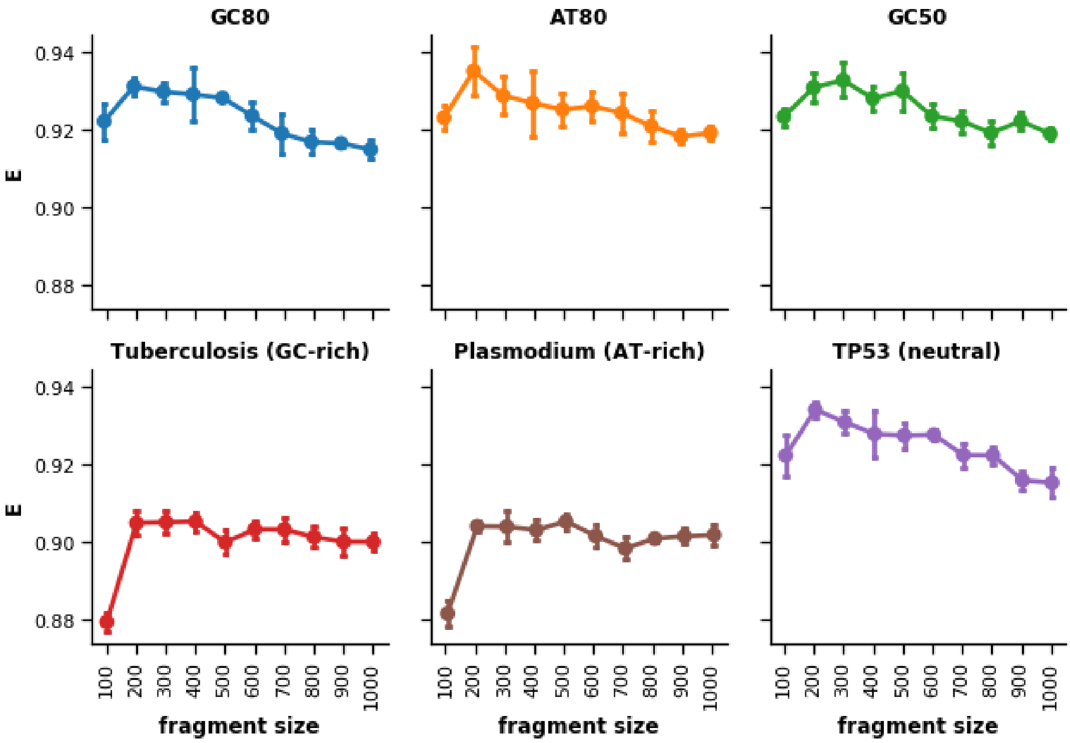


Figure 52: Comparison of effects from varying mean fragment sizes.

Skewness

In the previous analysis of the effects of skewness on coverage uniformity for the artificial sequences, there was no clear trend in its impact on coverage. This same result is found with the real sequences tested (**Figure 53**). This confirms that modifying the skewness of the fragment distribution has no effect on the evenness of coverage.

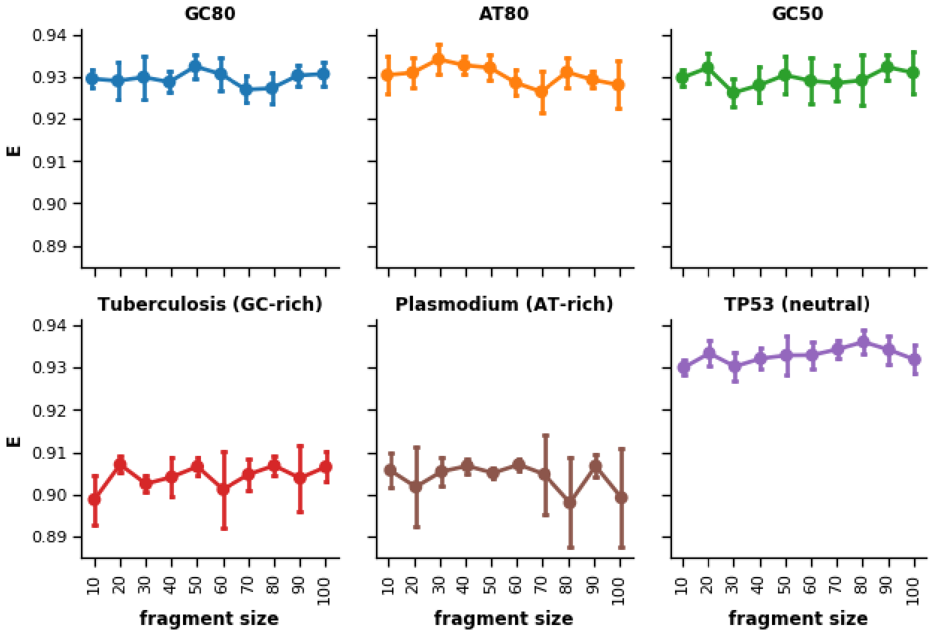


Figure 53: Comparison of effects from varying the standard deviation.

Splitting Bias

In this comparison we see a minute reduction in E at higher levels (0.8 – 0.9) of the splitting bias for the different nucleotide compositions for both the real and artificial sequences (**Figure 54**). The splitting bias does not adversely affect coverage uniformity for these sequences. However in section 5.1.1 this bias was seen to be more effective on the artificial sequence with large areas of homogenous nucleotide content (GCAT80).

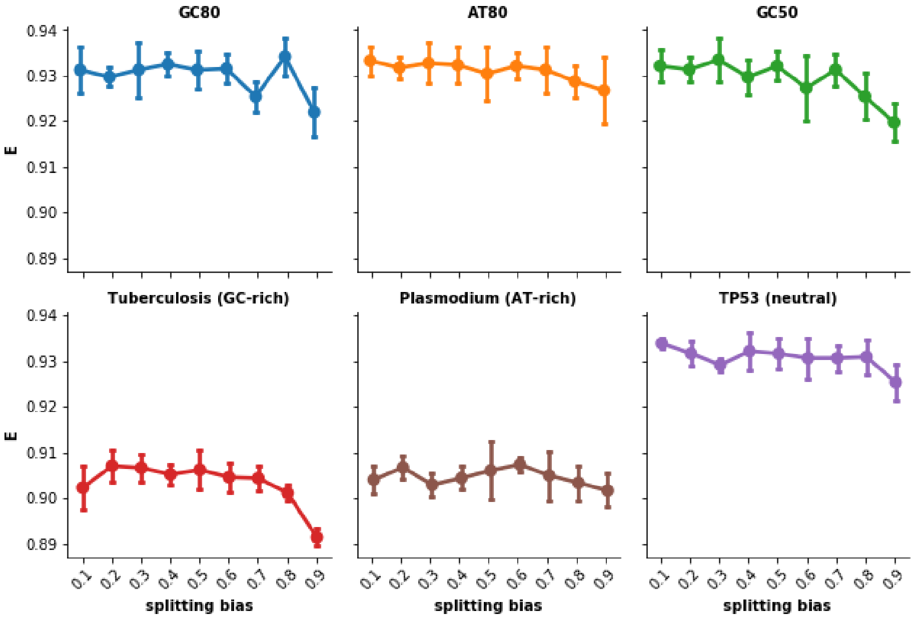


Figure 54: Comparison of the effects from varying splitting bias levels.

Ligation Bias

The ligation bias parameter mostly affects sequences with an AT proportion of 50% or higher and also sequences with areas of homogeneous AT content. This effect can be seen in the real sequences (**Figure 55**). *Plasmodium* and *TP53* share a similar trend with their artificial counterparts, where as the level of the ligation bias increases the evenness of coverage reduces. The trend in the *Tuberculosis* plot is slightly different from GC80, showing a slight reduction in E at the highest levels of the bias (0.8 – 0.9). Due to the minimal difference in E this change can be ignored. The results seen here confirm the effects of the ligation bias on coverage.

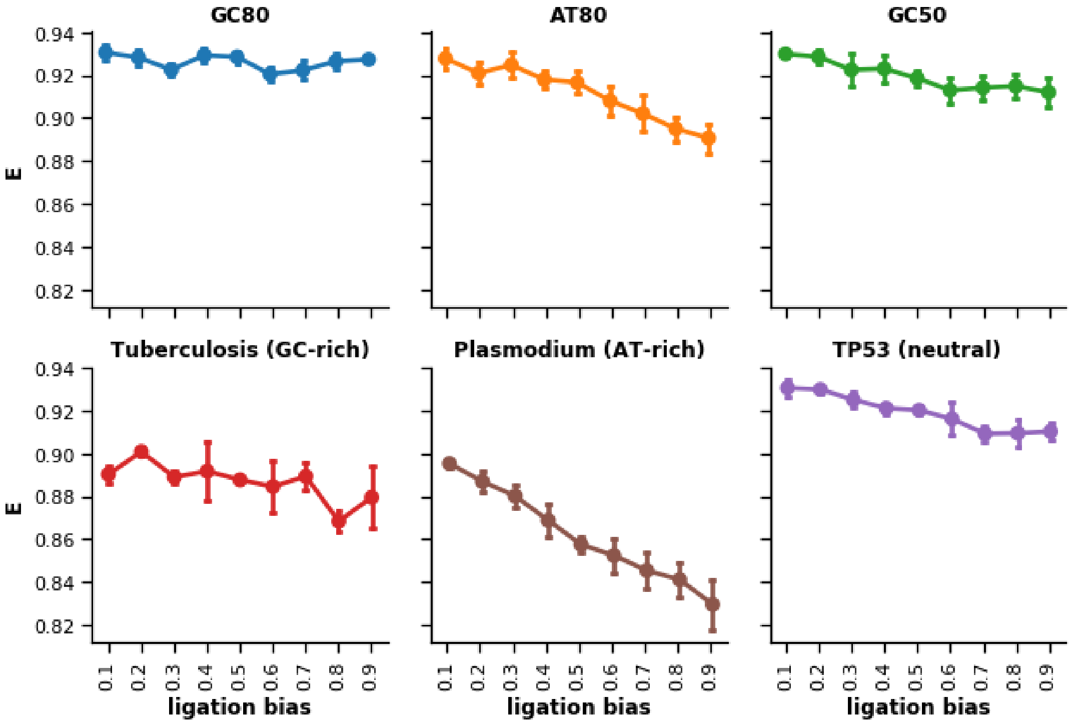


Figure 55: Comparison of the effects from varying ligation bias levels.

Denaturation and Elongation Temperatures

The temperatures set and structure of the sequences provided have a key influence on coverage levels after PCR. When applying different denaturation temperatures to both GC-rich sequences (real and artificial), a similar effect is seen (**Figure 56**). Lower temperatures (< 94C) lead to a reduction in E, especially in the artificial sequence, where its GC-level is higher (80%) than the *Tuberculosis* sequence (70.46%). The levels of E are the same when comparing both types of AT-rich and neutral sequences. Varying the Elongation temperature resulted in closely matching trends once again

(**Figure 57**). AT80 and *Plasmodium* show a steep decline in E with elongation temperatures above 68C. In the case of the GC-rich and neutral sequences there is no effect.

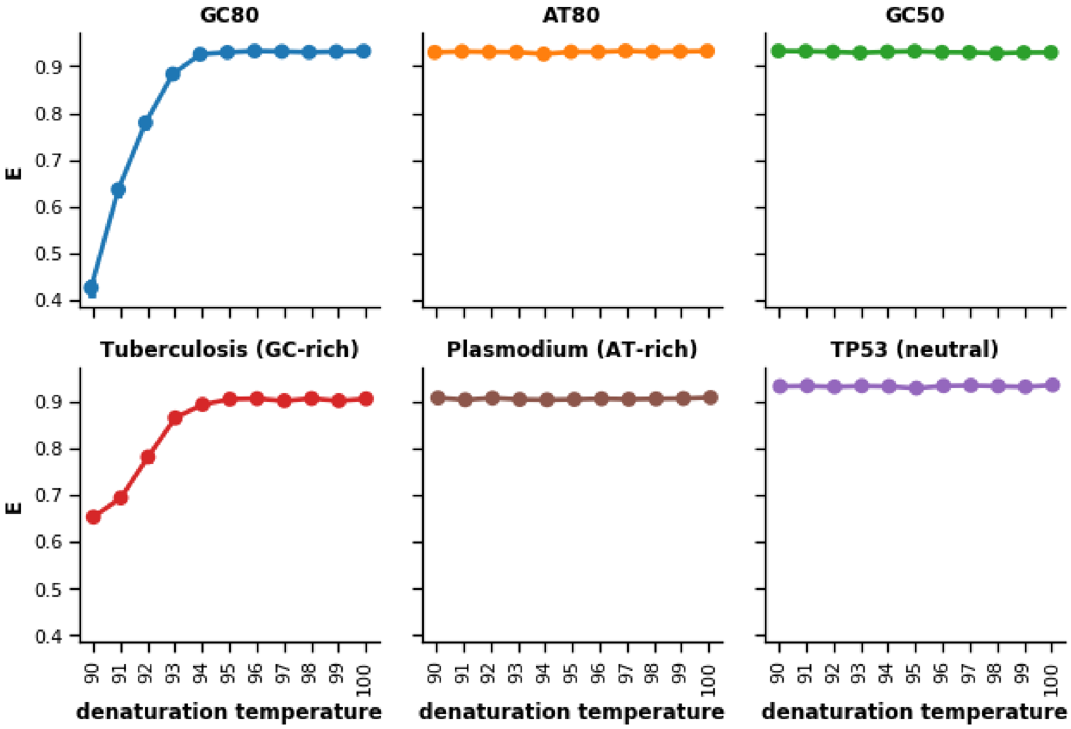


Figure 56: Comparison of the effects from varying denaturation temperatures.

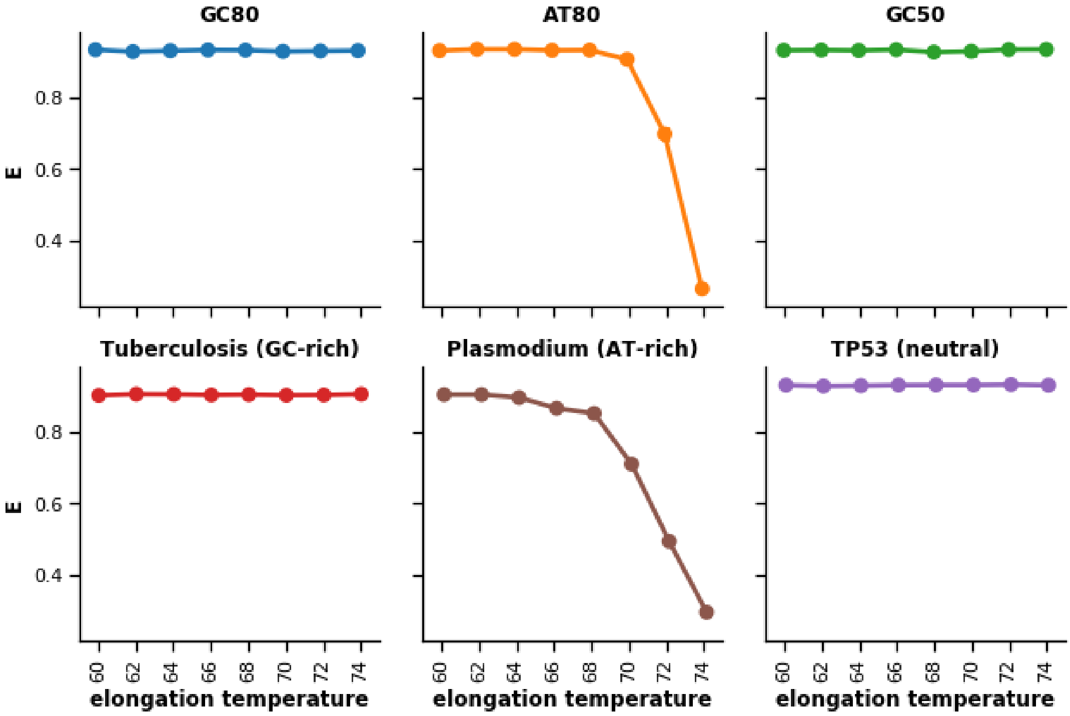


Figure 57: Comparison of the effects from varying elongation temperatures.

In conclusion, the results of these comparisons show that the effects of the tested parameters are not only effective on simulated DNA sequences but also on actual DNA sequences.

5.4 Chapter Summary and Discussion

Fragmentation, ligation, and amplification are important steps of the NGS library preparation process. Irrespective of the particular library preparation steps, the structure of the genome in terms of composition and serial dependence of the nucleotides, has a clear impact on the uniformity of coverage. The results show that a clumped sequence leads to lower coverage uniformity (**Error! Reference source not found.**). This suggests that sequencing results of DNA with areas of biased nucleotide content (AT-rich and GC-rich) are less reliable because of poorer coverage.

My study shows that these steps individually affect the uniformity of coverage at distinct levels of their parameters (mean fragment size, skewness, splitting bias, ligation bias, denaturation, and elongation).

With regards to **fragmentation**, the skewness of the underlying fragment distribution does not have any effects on coverage uniformity, but fragment size does. When fragments are large, there is a decline in coverage uniformity and the highest evenness of coverage was found for fragments between 200 and 400 bp (Figure 35). Interestingly, this is indeed the range of fragment sizes routinely employed by the Illumina platform (Bronner *et al.*, 2009). My study indicates that these values should be adhered to.

Moreover, and in correspondence with the outcomes of the model, Bronner *et al.* (2009) found that larger fragment sizes reduce the efficiency and yield of sequencing experiments. Tan *et al.* (2019) also found that using fragment sizes above 500bp result in lower base call quality and higher error rates when compared with shorter fragments in paired-end sequencing. Read quality only improved when libraries were prepared following Illumina's specifications with fragment sizes of 350bp.

Splitting bias affects the evenness in coverage of the clumped sequence (GCAT80) stronger than the other types of simulated DNA, with a sharp decline in coverage uniformity at higher levels of the splitting bias (**Figure 38**).

This effect is due to the increased presence of CpG dinucleotides in the GC-rich area of the clumped sequence. The preferential splitting between C and G nucleotides leads to an over-representation of fragments from GC-rich areas and of fragments that start with a C or are terminated at a G. The same was reported by Poptsova *et al.* (2014) in an empirical study. Furthermore, this effect is more pronounced in heterogeneous sequences with clumped areas of GC and AT dinucleotides (e.g. GCAT80). The preferential splitting would place a majority of fragments in the GC-rich area, therefore reducing representation of the AT-rich area.

Ligation bias is the tendency for adaptors to connect to fragments with a T at their 5' end (Seguin-Orlando *et al.*, 2013). Because adapters are biased against fragments with a T on their 5' end, a high value of this parameter corresponds to a low binding affinity. A consequence of this bias, a reduced coverage of AT-rich regions of a genome, became apparent in my simulation: it negatively affects the coverage uniformity especially in AT-rich sequences (AT80 and GCAT80) but not in AT-poor DNA (**Figure 39**).

The previously mentioned effect of a strong splitting bias was observed to interact with ligation in sequences with areas of biased nucleotide content (GC-rich and AT-rich). In such regions the splitting bias would cause the majority of fragmentation to occur in the GC-rich regions, leading to a lower representation of AT-rich regions. In turn, this may be exacerbated by the ligation bias, which will further reduce region representation due to the loss of AT-rich fragments.

With respect to **amplification**, the analysis of the effects of PCR **denaturation** temperatures indicates a lower coverage uniformity at lower denaturation temperatures for GC-rich sequences (GC80 and GCAT80) (**Figure 40**). The higher melting temperature of GC-rich double-stranded fragments from such sequences are responsible for this. Incomplete denaturation is one of the main causes of PCR failure (Innis & Gelfand, 1999). At lower temperatures, double-stranded fragments with high GC-content do not completely separate and therefore do not go through the PCR cycle, hence yielding a lower coverage and a lower uniformity of coverage for GC-rich sequences.

During PCR **elongation**, higher temperatures cause a reduction in coverage uniformity for AT-rich sequences (GCAT80 and AT80) (Figure 42) most likely because of the lower melting temperature of AT-rich double-stranded fragments.

AT-rich fragments are denatured usually at an elongation temperature of 72°C (López-Barragán *et al.*, 2011), because AT bonds can be disrupted easily due to their lower melting temperature (Yakovchuk, 2006). Consequently, if the PCR elongation temperature is too high there will be a loss of AT-rich fragments and concurrent coverage loss in AT-rich areas of a sequence, thus leading to uneven coverage. In a previous study, Su *et al.* (1996) found that reducing the PCR elongation temperature from the typical 72°C to 60°C, improves amplification of AT-rich fragments. This reduced temperature can lead to increased coverage yield in AT-rich areas of a sequence. The effects of reduced elongation temperature can be seen in my results where coverage uniformity is higher at lower temperatures (**Figure 42**).

Fragmentation and amplification appear to be interacting library preparation steps in the sense that a lower coverage uniformity of larger fragments is particularly noticeable at low denaturation and high elongation temperatures for respectively GC-rich and AT-rich DNA (Figures **Figure 49** and **Figure 50**). These combined effects of fragmentation, amplification and genome structure may be due to a sampling effect, as I found in the output of my simulation, larger fragment size is associated with a lower standard deviation of the mean GC/AT content of those fragments (**Figure 58**).

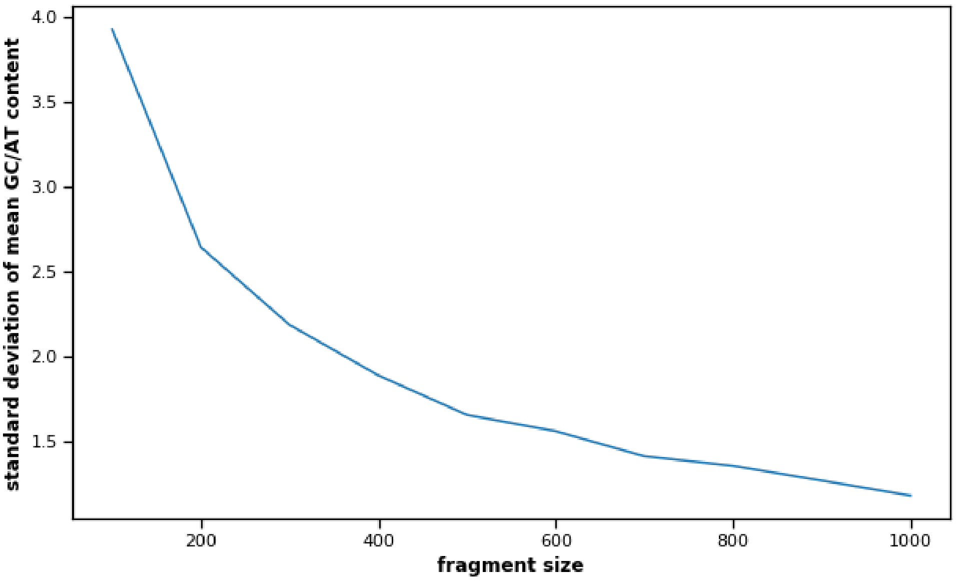


Figure 58: Standard deviation of mean GC/AT content vs fragment size.

This might be a consequence of the central limit theorem, which states that the sample mean and population mean converge and the variance of the distribution of sample means reduces as sample size increases (**Figure 59**)

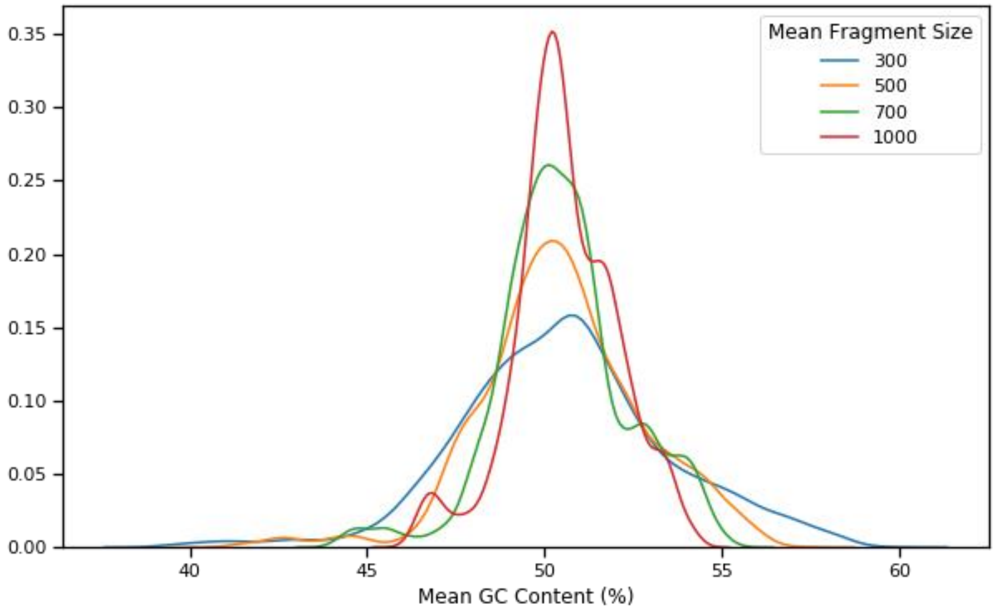


Figure 59: In line with the central limit theorem, the distribution of mean GC content approaches a normal distribution with a smaller variance as sample size increases. The data in this plot was generated from my simulation.

To further explain this, if the number of strong (G or C) or weak binding (A or T) bases in a string of $\boldsymbol{n}$ nucleotides is represented by the binary variable $\boldsymbol{Y}$ (i.e. $\boldsymbol{Y}$ takes on the values {G or C} = 1, {A or T} = 0), then the proportion of G's or C's in that string is the mean ($\boldsymbol{m}$) of $\boldsymbol{Y}$ for a sample size of $\boldsymbol{n}$. According to the central limit theorem, given a set of $\boldsymbol{N}$ such samples (indexed as $\boldsymbol{i}$ = 1, 2, …, $\boldsymbol{N}$), the corresponding sample means $\boldsymbol{m_i}$ are normally distributed with an overall mean of $\boldsymbol{\mu_m = \overline{m}_i}$, a variance of $\boldsymbol{\sigma_m^2 = \frac{\sigma^2}{n}}$, and a standard deviation of $\boldsymbol{\sigma_m = \frac{\sigma}{\sqrt{n}}}$. The latter can be estimated from a sample as the standard error, $\boldsymbol{\frac{s}{\sqrt{n}}}$. From this, it follows that larger samples have on average a smaller variation of $\boldsymbol{Y}$ than smaller samples. The implication is that smaller DNA fragments on average have a more diverse base composition than larger samples.

Thus, larger fragments tend to be either more AT-rich or GC-rich than smaller ones although the average base composition of large and small fragments is the same. A similar suggestion has been put forward by Elhaik *et al.* (2010). Logically, the sampling effect found here should intrinsically lead to the same occurrence in real sequencing.

The consequence for sequencing is that, as fragment size increases there will be a higher number of fragments with increased GC/AT content. During the denaturation stage of PCR, an abundance of such GC-rich fragments leads to a loss of coverage if temperatures are inadequate, as they are less likely to denaturise at lower temperatures. AT-rich fragments are similarly affected during elongation if the elongation temperature is set too high because this results in a loss of such fragments due to their lower melting temperature. Thus, the increased GC/AT content of larger fragments could explain the difficulties seen in the simulated PCR amplification step. These effects will ultimately lead to uneven coverage across the genome.

Chapter 6 Conclusions

The main aim of this body of work was to analyse how artefacts that can occur during the library preparation stage of sequencing affect coverage. To do this, I implemented a model, LpSim, that simulates the fragmentation, ligation and PCR stages. These stages are represented by designated parameters. By varying the parameters, the outcomes of the simulation showed which alterations in library preparation influenced coverage and to what extent.

I used a genetic algorithm to find parameter settings that produced coverage values similar to those from actual sequencing experiments. LpSim simulated real-world coverage well for DNA sequences characterised by a serial dependency due to the presence of distinct stretches of homogeneous nucleotide content (especially AT and GC rich regions). The simulation was less successful for sequences that lacked such regions.

These findings corresponded with results from applying the model to computationally generated DNA sequences. The artificial sequences were especially designed to reflect sequential dependency and the presence or lack of regions with homogeneous nucleotide content.

The model was applied to four types of in silico DNA. Three of these were generated as zero-order Markov chains (i.e. lacked sequential dependency) and consisted of respectively 80% AT (“AT-rich”), 80% GC (“GC-rich”) and equal proportions of all four nucleotides (“neutral”). The fourth type was made to contain blocks of neutral as well as AT/GC rich composition (“clumped”).

After running the simulator on these sequences, I found that the parameter settings modify the evenness of coverage in the following ways:

1. The size of fragments affects coverage in all tested sequences. This is in line with the suggestions by Bronner *et al.* (2009) and Tan *et al.* (2019) to limit fragment sizes to between 200 – 500 as larger sizes may negatively affect coverage uniformity.

2. The splitting bias alters the evenness of coverage of a "clumped" sequence because fragmentation occurs mostly in GC-rich regions thus lowering the proportion of fragments from other regions.
3. Increased ligation bias influences coverage for sequences with high AT content. This is because the adapters are less likely to attach to fragments that terminate with a T.
4. Denaturation and elongation temperature respectively impact GC-rich and AT-rich sequences due to the well-known temperature-related effect of PCR on such sequences.

The effects of some parameters were found to interact with each other, leading to additional reduction in coverage uniformity:

5. The splitting bias reinforces the effect of ligation bias. Reduced coverage of AT-rich regions caused by the splitting bias is decreased even further by the ligation bias because of the lower binding affinity of adapters to fragments from such regions. The resulting low number of AT-rich fragments brings about a less even coverage overall.
6. Fragment size interacts with the impact of PCR related temperature settings. Due to sampling effects, larger fragments have a less diverse base composition than smaller fragments. This leads to difficulties in denaturing these large fragments at lower temperatures if a sequence is GC-rich and a loss of fragments at higher elongation temperatures when a sequence is AT-rich.

Because all the tests were carried out using *in silico* DNA, these results have to be viewed as suggestions for further research on real sequences.

To gauge the usefulness of these suggestions, a validation with real DNA was necessary to ensure the effects found bear relevance to data from the real world. To do this, each step of the model was applied to sequences that match the features of those that were previously tested. In all steps the effects of coverage were quite similar but being that the sequential dependencies of the sequences are quite different an exact match was not expected.

My research provides an insight from a generic perspective on how library preparation methods can affect the reliability of sequencing output. This sets the scene for investigating a lot of shortcoming that can occur during this sample preparation phase.

By providing an in-silico method to do this it is now easier to test different combinations of parameters which would be rather unrealistic to test in the lab.

6.1 Future Work

Further extensions can be applied to this body of work:

1. To simulate fragmentation, fragment sizes were derived from a lognormal distribution because this distribution is commonly used in the modelling of breakage processes. This could be complemented by bottom-up oriented models that consider the finer details of fragmentation such as the physics underlying the breaking up of DNA molecules for the different fragmentation techniques.
2. The model may be further extended by adding library artefacts that were not implemented in my model. These may include artefacts such as slipped strand mispairing and chimera formation which can occur during the PCR step, the former causes a deletion of nucleotides in AT-rich fragments characterised by nucleotide repeats, while the latter leads to the formation of chimeric DNA due to incomplete primer extension during PCR.
3. A model implementing the sequencing stages following library preparation can be used in conjunction with LpSim to find the interactive effects of the properties of these stages and their inherent biases on sequencing output. For example, in the sequencing by synthesis stage, phasing and pre-phasing can lead to the omission of nucleotides in base calling. It would be of interest to see how errors from the follow-on steps interact with library preparation parameters.
4. A genetic algorithm can be used in conjunction with my model to create a tool which searches for parameters that would lead to uniform coverage of a given sequence. The fitness assessment here will utilise the evenness score to rate the parameters. This kind of tool can further assist wet-lab researchers in deciding what set of actions can be taken to improve coverage in a sequencing experiment.

6.2 Publications and Conferences

- Optimizing parameters for a DNA library preparation model

  Nathan Beka, Rene te Boerkhorst, Rod Adams, and Neil Davey

  Workshop on Mathematical and Statistical Aspects of Molecular Biology 2019

  EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK

  April 25-26, 2019

- Validation of a DNA library preparation model using a genetic algorithm

  Nathan Beka, Rene te Boerkhorst, Rod Adams, and Neil Davey

  Engineering and Computer Science Research Conference 2019

  University of Hertfordshire, Hatfield, UK

  April 17, 2019

- Naumenko, F.M., Abnizova, I.I., Beka, N., Genaev, M.A. and Orlov, Y.L. (2018) Novel read density distribution score shows possible aligner artefacts, when mapping a single chromosome. BMC Genomics. 19 (S3), 92. Available from: doi:10.1186/s12864-018-4475-6.

References

Aguilar, W., Santamaría-Bonfil, G., Froese, T. & Gershenson, C. (2014) The Past, Present, and Future of Artificial Life. *Frontiers in Robotics and AI*. [Online] 1. Available from: doi:10.3389/frobt.2014.00008 [Accessed: 23 May 2020].

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology*. [Online] 12 (2), R18. Available from: doi:10.1186/gb-2011-12-2-r18.

Allawi, H.T. & SantaLucia, J. (1997) Thermodynamics and NMR of Internal G·T Mismatches in DNA. *Biochemistry*. [Online] 36 (34), 10581–10594. Available from: doi:10.1021/bi962590c.

Ansorge, W.J. (2009) Next-generation DNA sequencing techniques. *New biotechnology*. [Online] 25 (4), 195–203. Available from: doi:10.1016/j.nbt.2008.12.009.

Apone, L., Dimalanta, E. & Stewart, F. (2017) *Improving Enzymatic DNA Fragmentation for Next Generation Sequencing Library Construction*. [Online]. 2017. Available from: https://www.neb.com/-/media/nebus/files/feature-articles/nebexpressions_feature_nebnext_fs_issueiii_2017.pdf?la=en [Accessed: 9 April 2019].

Arifin, S.N., Zhou, Y., Davis, G.J., Gentile, J.E., et al. (2014) An agent-based model of the population dynamics of Anopheles gambiae. *Malaria Journal*. [Online] 13 (1), 424. Available from: doi:10.1186/1475-2875-13-424.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., et al. (2015) A global reference for human genetic variation. *Nature*. [Online] 526 (7571), 68–74. Available from: doi:10.1038/nature15393.

Bankier, A.T. (2001) Shotgun DNA sequencing. *Methods in molecular biology (Clifton, N.J.)*. [Online] 167, 89–100. Available from: doi:10.1385/1-59259-113-2:089.

Baxevanis, A.D. & Ouellette, B.F.F. (2004) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. [Online]. John Wiley & Sons. Available from: http://books.google.com/books?hl=en&lr=&id=i0W9NBmxewQC&pgis=1.

Benjamini, Y. & Speed, T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*. [Online] 40 (10), e72. Available from: doi:10.1093/nar/gks001.

Bharagava, R.N., Purchase, D., Saxena, G. & Mulla, S.I. (2019) Chapter 26 - Applications of Metagenomics in Microbial Bioremediation of Pollutants: From Genomics to Environmental Cleanup. In: Surajit Das & Hirak Ranjan Dash (eds.). *Microbial Diversity in the Genomic Era*. [Online]. Academic

Press. pp. 459–477. Available from: doi:10.1016/B978-0-12-814849-5.00026-5 [Accessed: 5 October 2020].

Bleidorn, C. (2016) Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*. [Online] 14 (1), 1–8. Available from: doi:10.1080/14772000.2015.1099575.

Boden, M.A. (1996) *The philosophy of artificial life*. Oxford readings in philosophy. Oxford ; New York, Oxford University Press.

Breslauer, K.J., Frank, R., Blöcker, H. & Marky, L. a (1986) Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 83 (11), 3746–3750. Available from: doi:10.1073/pnas.83.11.3746.

Bronner, I.F., Quail, M.A., Turner, D.J. & Swerdlow, H. (2009) Improved Protocols for Illumina Sequencing. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*. [Online] 0 18. Available from: doi:10.1002/0471142905.hg1802s62 [Accessed: 28 March 2019].

Buehler, B., Hogrefe, H.H., Scott, G., Ravi, H., et al. (2010) Rapid quantification of DNA libraries for next-generation sequencing. *Methods (San Diego, Calif.)*. [Online] 50 (4), S15–8. Available from: doi:10.1016/j.ymeth.2010.01.004.

Carr, S.M. (2012) *DNA Electrophoresis*. [Online]. 2012. Available from: https://www.mun.ca/biology/scarr/Gel_Electrophoresis.html [Accessed: 24 August 2014].

CeGaT (2014) *Next-Generation Sequencing Services*. [Online]. 2014. CeGaT GmbH. Available from: https://www.cegat.de/en/services/next-generation-sequencing/ [Accessed: 20 January 2020].

Chantler, P.D. (2004) *rDNA: Polymerase Chain Reaction (PCR)*. [Online]. Available from: http://www.rvc.ac.uk/review/DNA_1/3_PCR.cfm [Accessed: 29 August 2014].

Cheng, K.C., Cahill, D.S., Kasai, H., Nishimura, S., et al. (1992) 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G----T and A----C substitutions. *The Journal of Biological Chemistry*. 267 (1), 166–172.

Chevet, E., Lemaitre, G. & Katinka, M.D. (1995) Low concentrations of tetramethylammonium chloride increase yield and specificity of PCR. *Nucleic Acids Research*. [Online] 23 (16), 3343–3344. Available from: doi:10.1093/nar/23.16.3343.

Church, G.M. (2005) The personal genome project. *Molecular systems biology*. [Online] 1, 2005.0030. Available from: doi:10.1038/msb4100040.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. [Online] 25 (11), 1422–1423. Available from: doi:10.1093/bioinformatics/btp163.

Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., et al. (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research*. [Online] 41 (6), e67. Available from: doi:10.1093/nar/gks1443.

Covaris (2012) *AFA vs. Sonicator Baths & Probes*. [Online]. 28 May 2012. Covaris. Available from: https://covaris.com/pre-analytical/afa-vs-sonicators/ [Accessed: 7 January 2020].

Covaris (2016) *DNA Shearing for Next Generation Sequencing (NGS) with the M220 Focused-ultrasonicator*. [Online]. Available from: https://covaris.com/wp-content/uploads/M020013.pdf [Accessed: 30 October 2019].

Deb, K. & Agrawal, S. (1999) A Niched-Penalty Approach for Constraint Handling in Genetic Algorithms. In: Andrej Dobnikar, Nigel C. Steele, David W. Pearson, & Rudolf F. Albrecht (eds.). *Artificial Neural Nets and Genetic Algorithms*. [Online]. 1999 Vienna, Springer. pp. 235–243. Available from: doi:10.1007/978-3-7091-6384-9_40.

Deb, K. & Goyal, M. (1996) A Combined Genetic Adaptive Search (GeneAS) for Engineering Design. *Computer Science and Informatics*. 26, 30–45.

Diagenode (2013) *Standard protocols DNA shearing for Bioruptor® Pico*. [Online]. Available from: https://www.diagenode.com/files/protocols/Standard_protocols_for_DNAShearing.pdf [Accessed: 30 October 2019].

van Dijk, E.L., Auger, H., Jaszczyszyn, Y. & Thermes, C. (2014) Ten years of next-generation sequencing technology. *Trends in Genetics*. [Online] 30 (9), 418–426. Available from: doi:10.1016/j.tig.2014.07.001.

van Dijk, E.L., Jaszczyszyn, Y. & Thermes, C. (2014) Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research*. [Online] 322 (1), 12–20. Available from: doi:10.1016/j.yexcr.2014.01.008.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., et al. (2010) Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science*. [Online] 327 (5961), 78–81. Available from: doi:10.1126/science.1181498.

Dutton, C.M., Christine, P. & Sommer, S.S. (1993) General method for amplifying regions of very high G + C content. *Nucleic Acids Research*. [Online] 21 (12), 2953–2954. Available from: doi:10.1093/nar/21.12.2953.

Elhaik, E., Graur, D., Josić, K. & Landan, G. (2010) Identifying compositionally homogeneous and nonhomogeneous domains within the human genome using a novel segmentation algorithm. *Nucleic Acids Research*. [Online] 38 (15), e158–e158. Available from: doi:10.1093/nar/gkq532.

Escalona, M., Rocha, S. & Posada, D. (2016) A comparison of tools for the simulation of genomic next-generation sequencing data. *Nature Reviews Genetics*. [Online] 17 (8), 459–469. Available from: doi:10.1038/nrg.2016.57.

Eshelman, L.J. & Schaffer, J.D. (1993) Real-Coded Genetic Algorithms and Interval-Schemata. In: L. DARRELL Whitley (ed.). *Foundations of Genetic Algorithms*. Foundations of Genetic Algorithms. [Online]. Elsevier. pp. 187–202. Available from: doi:10.1016/B978-0-08-094832-4.50018-0 [Accessed: 2 April 2019].

EzBioCloud (2019) *Chimera detection | EzBioCloud Help center*. [Online]. 2019. EzBioCloud. Available from: //help.ezbiocloud.net/user-guide/mtp-pipeline/chimera-detection/ [Accessed: 28 January 2020].

Fazekas, A., Steeves, R. & Newmaster, S. (2010) Improving sequencing quality from PCR products containing long mononucleotide repeats. *BioTechniques*. [Online] 48 (4), 277–85. Available from: doi:10.2144/000113369.

Ferrante, D.D., Wei, Y. & Koulakov, A.A. (2014) Statistical model of evolution of brain parcellation. *arXiv:1412.0603 [cond-mat, q-bio]*. [Online] Available from: http://arxiv.org/abs/1412.0603 [Accessed: 1 November 2019].

Fortin, F.-A., Rainville, F.-M.D., Gardner, M.-A., Parizeau, M., et al. (2012) DEAP: Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research*. 13 (70), 2171–2175.

Fowler, A.C. & Scheu, B. (2016) A theoretical explanation of grain size distributions in explosive rock fragmentation. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*. [Online] 472 (2190), 20150843. Available from: doi:10.1098/rspa.2015.0843.

Gaastra, W. & Hansen, K. (1984) Ligation of DNA with T4 DNA Ligase. In: John M. Walker (ed.). *Nucleic Acids*. Methods in Molecular Biology. [Online]. Totowa, NJ, Humana Press. pp. 225–230. Available from: doi:10.1385/0-89603-064-4:225 [Accessed: 15 April 2019].

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*. [Online] 27 (2), 182–189. Available from: doi:10.1038/nbt.1523.

Goldberg, D.E. (1989) *Genetic algorithms in search, optimization, and machine learning*. Reading, Mass, Addison-Wesley Pub. Co.

Grokhovsky, S., Il'icheva, I.A., Nechipurenko, Y., Golovkin, M., et al. (2013) Mechanochemical cleavage of DNA by ultrasound. *Ultrasonics: Theory, Techniques and Practical Applications*. 1–24.

Grokhovsky, S.L. (2006) Specificity of DNA cleavage by ultrasound. *Molecular Biology*. [Online] 40 (2), 276–283. Available from: doi:10.1134/S0026893306020142.

Grokhovsky, S.L., Il'icheva, I.A., Nechipurenko, D.Yu., Golovkin, M.V., et al. (2011) Sequence-Specific Ultrasonic Cleavage of DNA. *Biophysical Journal*. [Online] 100 (1), 117–125. Available from: doi:10.1016/j.bpj.2010.10.052.

Grokhovsky, S.L., Il'icheva, I.A., Nechipurenko, D.Yu., Panchenko, L.A., et al. (2008) Ultrasonic cleavage of DNA: Quantitative analysis of sequence specificity. *Biophysics*. [Online] 53 (3), 250. Available from: doi:10.1134/S0006350908030159.

Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., et al. (2014) Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*. [Online] 56 (2), 61–77. Available from: doi:10.2144/000114133.

Holland, J.H. (1975) *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor, University of Michigan Press.

Horton, P. (2016) A commentary on evaluation of the evenness score in next-generation sequencing. *Journal of Human Genetics*. [Online] 61, 575. Available from: doi:10.1038/jhg.2016.29.

Hu, X., Yuan, J., Shi, Y., Lu, J., et al. (2012) pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics (Oxford, England)*. [Online] 28 (11), 1533–5. Available from: doi:10.1093/bioinformatics/bts187.

Huang, W., Li, L., Myers, J.R. & Marth, G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*. [Online] 28 (4), 593–4. Available from: doi:10.1093/bioinformatics/btr708.

Huiru, W., Jinhui, S., Jianying, F., Huiru, F., et al. (2018) An agent-based modeling and simulation of consumers' purchase behavior for wine consumption. *IFAC-PapersOnLine*. [Online] 51 (17), 843–848. Available from: doi:10.1016/j.ifacol.2018.08.089.

IBM Corp. (2017) *IBM SPSS Statistics for Windows*. Armonk, NY, IBM Corp.

Illumina (2017) *An Introduction to Next-Generation Sequencing Technology*. [Online]. p.16. Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf [Accessed: 20 October 2020].

Illumina (2011a) *Paired-End Sample Preparation Guide*. [Online]. Available from: http://supportres.illumina.com/documents/myillumina/e5af4eb5-6742-40c8-bcb1-d8b350bcb964/paired-end_sampleprep_guide_1005063_e.pdf.

Illumina (2008) *Preparing Samples for Sequencing Genomic DNA*. [Online]. Illumina, Inc. Available from: http://support.illumina.com/downloads/genomic_dna_sample_prep_guide_1003806.html.

Illumina (2011b) *RTA 1.13, HCS 1.5, and SCS 2.10 Theory of Operation*. [Online]. Available from: http://res.illumina.com/documents/products/technotes/technote_rta_theory_operations.pdf.

Innis, M. & Gelfand, D. (1999) 1 - Optimization of PCR: Conversations between Michael and David. In: Michael A. Innis, David H. Gelfand, & John J. Sninsky (eds.). *PCR Applications*. [Online]. San Diego, Academic Press. pp. 3–22.

Available from: doi:10.1016/B978-012372185-3/50002-X [Accessed: 18 June 2019].

jbstatistics (2012) *Finding the P-value in One-Way ANOVA*. [Online]. Available from: https://www.youtube.com/watch?v=XdZ7BRqznSA [Accessed: 31 March 2020].

Kapa Biosystems (2016) *KAPA Frag Kit for Enzymatic Fragmentation*. [Online]. Available from: http://netdocs.roche.com/DDM/Effective/0000000000001200000190097_000_01_005_Native.pdf [Accessed: 30 October 2019].

Kircher, M., Heyn, P. & Kelso, J. (2011) Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics*. [Online] 12 (1), 382. Available from: doi:10.1186/1471-2164-12-382.

Knierim, E., Lucke, B., Schwarz, J.M., Schuelke, M., et al. (2011) Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. M. Thomas P. Gilbert (ed.). *PloS one*. [Online] 6 (11), e28240. Available from: doi:10.1371/journal.pone.0028240.

Kolmogorov, A.N. (1941) On the logarithmically normal law of distribution of the size of particles under pulverisation. *Doklady Akademii Nauk SSSR*. 31, 99–101.

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., et al. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. *Nature methods*. [Online] 6 (4), 291–295. Available from: doi:10.1038/nmeth.1311.

Kozarewa, I. & Turner, D.J. (2011) Amplification-Free Library Preparation for Paired-End Illumina Sequencing. In: Young Min Kwon & Steven C. Ricke (eds.). *High-Throughput Next Generation Sequencing: Methods and Applications*. Methods in Molecular Biology. [Online]. Totowa, NJ, Humana Press. pp. 257–266. Available from: doi:10.1007/978-1-61779-089-8_18 [Accessed: 22 October 2020].

Kutter, E. (2001) Concatemer (Genomes). In: Sydney Brenner & Jefferey H. Miller (eds.). *Encyclopedia of Genetics*. [Online]. New York, Academic Press. pp. 435–436. Available from: doi:10.1006/rwgn.2001.0258 [Accessed: 22 October 2019].

Labster.com (2014) *A-tailing - Life Science Learning Wiki*. [Online]. 2014. Available from: http://learn.labster.com/index.php/A-tailing.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., et al. (2001) Initial sequencing and analysis of the human genome. *Nature*. [Online] 409 (6822), 860–921. Available from: doi:10.1038/35057062.

Langton, C.G. (1997) Google-Books-ID: qErpoKjc1h4C. *Artificial Life: An Overview*. MIT Press.

Le Novere, N. (2001) MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics*. [Online] 17 (12), 1226–1227. Available from: doi:10.1093/bioinformatics/17.12.1226.

Ledergerber, C. & Dessimoz, C. (2011) Base-calling for next-generation sequencing platforms. *Briefings in bioinformatics*. [Online] 12 (5), 489–97. Available from: doi:10.1093/bib/bbq077.

Leinonen, R., Sugawara, H., Shumway, M. & on behalf of the International Nucleotide Sequence Database Collaboration (2011) The Sequence Read Archive. *Nucleic Acids Research*. [Online] 39 (Database), D19–D21. Available from: doi:10.1093/nar/gkq1019.

Levene, H. (1960) Robust tests for equality of variances. In: *Contributions to probability and statistics; essays in honor of Harold Hotelling*. Stanford, Calif., Stanford University Press. pp. 278–292.

Li, L. & Speed, T.P. (1999) An estimate of the crosstalk matrix in four-dye fluorescence-based DNA sequencing. *Electrophoresis*. [Online] 20 (7), 1433–42. Available from: doi:10.1002/(SICI)1522-2683(19990601)20:7<1433::AID-ELPS1433>3.0.CO;2-0.

López-Barragán, M.J., Quiñones, M., Cui, K., Lemieux, J., et al. (2011) Effect of PCR extension temperature on high-throughput sequencing. *Molecular and biochemical parasitology*. [Online] 176 (1), 64–67. Available from: doi:10.1016/j.molbiopara.2010.11.013.

Lorenz, T.C. (2012) Polymerase Chain Reaction: Basic Protocol Plus Troubleshooting and Optimization Strategies. *Journal of Visualized Experiments : JoVE*. [Online] (63). Available from: doi:10.3791/3998 [Accessed: 18 June 2019].

Lundin, S., Stranneheim, H., Pettersson, E., Klevebring, D., et al. (2010) Increased Throughput by Parallelization of Library Preparation for Massive Sequencing. *PLOS ONE*. [Online] 5 (4), e10029. Available from: doi:10.1371/journal.pone.0010029.

Mardis, E.R. (2008) Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*. [Online] 9, 387–402. Available from: doi:10.1146/annurev.genom.9.081307.164359.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. [Online] 437 (7057), 376–380. Available from: doi:10.1038/nature03959.

Massey, F.J. (1951) The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*. [Online] 46 (253), 68–78. Available from: doi:10.2307/2280095.

Maxam, A.M. & Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 74, 560–564. Available from: doi:10.1073/pnas.74.2.560.

Maymon, G. (2018) *Stochastic crack propagation: essential practical aspects*. London, United Kingdom ; San Diego, Academic Press is an imprint of Elsevier.

McElroy, K.E., Luciani, F. & Thomas, T. (2012) GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC genomics*. [Online] 13 (1), 74. Available from: doi:10.1186/1471-2164-13-74.

Metzker, M.L. (2010) Sequencing technologies — the next generation. *Nature Reviews Genetics*. [Online] 11 (1), 31–46. Available from: doi:10.1038/nrg2626.

Mitchell, M. (1996) *An introduction to genetic algorithms*. Complex adaptive systems. Cambridge, Mass, MIT Press.

Mokry, M., Feitsma, H., Nijman, I.J., de Bruijn, E., et al. (2010) Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. *Nucleic Acids Research*. [Online] 38 (10), e116. Available from: doi:10.1093/nar/gkq072.

Nagarajan, N. & Pop, M. (2013) Sequence assembly demystified. *Nature reviews. Genetics*. [Online] 14 (3), 157–67. Available from: doi:10.1038/nrg3367.

Nakazato, T., Ohta, T. & Bono, H. (2013) Experimental Design-Based Functional Mining and Characterization of High-Throughput Sequencing Data in the Sequence Read Archive. *PLOS ONE*. [Online] 8 (10), e77910. Available from: doi:10.1371/journal.pone.0077910.

Nature.com (n.d.) *DNA polymerase / DNAP*. [Online]. Available from: http://www.nature.com/scitable/definition/dna-polymerase-dnap-1 [Accessed: 1 October 2014].

Naumenko, F.M., Abnizova, I.I., Beka, N., Genaev, M.A., et al. (2018) Novel read density distribution score shows possible aligner artefacts, when mapping a single chromosome. *BMC Genomics*. [Online] 19 (S3), 92. Available from: doi:10.1186/s12864-018-4475-6.

Neĭkov, O.D., Naboychenko, S. & Yefimov, N.V. (2018) *Handbook of non-ferrous metals powders: technologies and applications*.

New England Biolabs (2014) *DNA Fragmentation | New England Biolabs*. [Online]. 2014. Available from: https://www.neb.com/applications/library-preparation-for-next-generation-sequencing/dna-fragmentation.

Obenrader, S. (2003) *The Sanger Method*. [Online]. 2003. bio.davidson.edu. Available from: http://www.bio.davidson.edu/courses/molbio/molstudents/spring2003/obenrader/sanger_method_page.htm [Accessed: 14 July 2014].

Oyola, S.O., Otto, T.D., Gu, Y., Maslen, G., et al. (2012) Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC genomics*. [Online] 13 (1), 1. Available from: doi:10.1186/1471-2164-13-1.

Poptsova, M.S., Il'icheva, I.A., Nechipurenko, D.Y., Panchenko, L.A., et al. (2014) Non-random DNA fragmentation in next-generation sequencing. *Scientific reports*. [Online] 4, 4532. Available from: doi:10.1038/srep04532.

Porreca, G.J., Shendure, J. & Church, G.M. (2006) Polony DNA sequencing. *Current Protocols in Molecular Biology*. [Online] Chapter 7, Unit 7.8. Available from: doi:10.1002/0471142727.mb0708s76.

Princeton.edu (n.d.) *Primer (molecular biology)*. [Online]. Available from: https://www.princeton.edu/~achaney/tmve/wiki100k/docs/Primer_(molecular _biology).html [Accessed: 1 October 2014].

Pushkarev, D., Neff, N.F. & Quake, S.R. (2009) Single-molecule sequencing of an individual human genome. *Nature Biotechnology*. [Online] 27 (9), 847–850. Available from: doi:10.1038/nbt.1561.

Quail, M.A., Kozarewa, I., Smith, F., Scally, A., et al. (2008) A large genome center's improvements to the Illumina sequencing system. *Nature methods*. [Online] 5 (12), 1005–10. Available from: doi:10.1038/nmeth.1270.

Quail, M.A., Otto, T.D., Gu, Y., Harris, S.R., et al. (2012) Optimal enzymes for amplifying sequencing libraries. *Nature Methods*. [Online] 9 (1), 10–11. Available from: doi:10.1038/nmeth.1814.

Quail, M.A., Swerdlow, H. & Turner, D.J. (2009) Improved protocols for the illumina genome analyzer sequencing system. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*. [Online] Chapter 18, Unit 18.2. Available from: doi:10.1002/0471142905.hg1802s62.

Research By Design (2017) *Foundations of ANOVA – Assumptions and Hypotheses for One-Way ANOVA (12-3)*. [Online]. Available from: https://www.youtube.com/watch?v=Y1pY86G74_c [Accessed: 31 March 2020].

Ripley, J. (2018) *Factorial ANOVA main effects and interactions*. [Online]. Available from: https://www.youtube.com/watch?v=f3WrB11uPcU [Accessed: 31 March 2020].

Robasky, K., Lewis, N.E. & Church, G.M. (2014) The role of replicates for error mitigation in next-generation sequencing. *Nature reviews. Genetics*. [Online] 15 (1), 56–62. Available from: doi:10.1038/nrg3655.

Roberts, G.A. & Dryden, D.T.F. (2013) DNA Electrophoresis: Historical and Theoretical Perspectives. In: Svetlana Makovets (ed.). *DNA Electrophoresis: Methods and Protocols*. Methods in Molecular Biology. [Online]. Totowa, NJ, Humana Press. pp. 1–9. Available from: doi:10.1007/978-1-62703-565-1_1 [Accessed: 15 April 2019].

Rodriguez-Murillo, L. & Salem, R.M. (2013) Insertion/Deletion Polymorphism. In: Marc D. Gellman & J. Rick Turner (eds.). *Encyclopedia of Behavioral Medicine*. [Online]. New York, NY, Springer New York. pp. 1076–1076. Available from: doi:10.1007/978-1-4419-1005-9_706.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 74 (12), 5463–5467. Available from: doi:10.1073/pnas.74.12.5463.

Sastry, K., Goldberg, D. & Kendall, G. (2005) Genetic Algorithms. In: Edmund K. Burke & Graham Kendall (eds.). *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*. [Online]. Boston, MA, Springer US. pp. 97–125. Available from: doi:10.1007/0-387-28356-0_4 [Accessed: 10 January 2020].

Searle, S.R. (1997) *Linear Models: Searle/Linear*. [Online]. Hoboken, NJ, USA, John Wiley & Sons, Inc. Available from: doi:10.1002/9781118491782 [Accessed: 3 December 2019].

Seguin-Orlando, A., Schubert, M., Clary, J., Stagegaard, J., et al. (2013) Ligation Bias in Illumina Next-Generation DNA Libraries: Implications for Sequencing Ancient Genomes. *PLOS ONE*. [Online] 8 (10), e78575. Available from: doi:10.1371/journal.pone.0078575.

Selig, M.S. & Coverstone-Carroll, V.L. (1996) Application of a Genetic Algorithm to Wind Turbine Design. *Journal of Energy Resources Technology*. [Online] 118 (1), 22–28. Available from: doi:10.1115/1.2792688.

Shapiro, S.S. & Wilk, M.B. (1965) An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*. [Online] 52 (3/4), 591–611. Available from: doi:10.2307/2333709.

Sharifian, H. (2010) *Errors induced during PCR amplification*. PhD Thesis. [Online]. Master Thesis ETH Zurich, 2010. Available from: http://e-collection.library.ethz.ch/eserv/eth:1397/eth-1397-01.pdf.

Sigma-Aldrich (2015) *Melting Temperatures of Oligonucleotides*. [Online]. Available from: http://www.sigmaaldrich.com/technical-documents/articles/biology/oligos-melting-temp.html#ref.

Silverman, E.K. (2007) Haplotype Thinking in Lung Disease. *Proceedings of the American Thoracic Society*. [Online] 4 (1), 4–8. Available from: doi:10.1513/pats.200607-145JG.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., et al. (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics*. [Online] 15 (2), 121–32. Available from: doi:10.1038/nrg3642.

Son, M.S. & Taylor, R.K. (2011) Preparing DNA libraries for multiplexed paired-end deep sequencing for Illumina GA sequencers. *Current protocols in microbiology*. [Online] Chapter 1, Unit 1E.4. Available from: doi:10.1002/9780471729259.mc01e04s20.

SRR5161262 (2011) *National Center for Biotechnology Information*. [Online]. Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR5161262 [Accessed: 17 January 2020].

SRR6257109 (2011) *National Center for Biotechnology Information*. [Online]. Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=run_browser&run=SRR6257109 [Accessed: 17 January 2020].

Su, X., Wu, Y., Sifri, C.D. & Wellems, T.E. (1996) Reduced Extension Temperatures Required for PCR Amplification of Extremely A+T-rich DNA. *Nucleic Acids Research*. [Online] 24 (8), 1574–1575. Available from: doi:10.1093/nar/24.8.1574.

Takahashi, M. & Kita, H. (2001) A crossover operator using independent component analysis for real-coded genetic algorithms. In: *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546)*. [Online]. May 2001 pp. 643–649 vol. 1. Available from: doi:10.1109/CEC.2001.934452.

Tan, G., Opitz, L., Schlapbach, R. & Rehrauer, H. (2019) Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*. [Online] 9 (1), 2856. Available from: doi:10.1038/s41598-019-39076-7.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., et al. (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research*. [Online] 18 (7), 1051–1063. Available from: doi:10.1101/gr.076463.108.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., et al. (2001) The sequence of the human genome. *Science (New York, N.Y.)*. [Online] 291 (5507), 1304–1351. Available from: doi:10.1126/science.1058040.

Wang, T.-Y., Chen, K.-C., Hsu, D.F. & Kao, C.-Y. (2009) Combining Agent-Based Models with Stochastic Differential Equations for Gene Regulatory Networks. In: *2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering*. [Online]. June 2009 pp. 405–409. Available from: doi:10.1109/BIBE.2009.47.

Wang, X.V., Blades, N., Ding, J., Sultana, R., et al. (2012) Estimation of sequencing error rates in short reads. *BMC Bioinformatics*. [Online] 13, 185. Available from: doi:10.1186/1471-2105-13-185.

Xiao, T. & Zhou, W. (2020) The third generation sequencing: the advanced approach to genetic diseases. *Translational Pediatrics*. [Online] 9 (2), 163–173. Available from: doi:10.21037/tp.2020.03.06.

Xie, H. & Zhang, M. (2013) Parent Selection Pressure Auto-Tuning for Tournament Selection in Genetic Programming. *IEEE Transactions on Evolutionary Computation*. [Online] 17 (1), 1–19. Available from: doi:10.1109/TEVC.2011.2182652.

Yakovchuk, P. (2006) Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*. [Online] 34 (2), 564–574. Available from: doi:10.1093/nar/gkj454.

Zeng, G.-Q., Chen, J., Li, L.-M., Chen, M.-R., et al. (2016) An improved multi-objective population-based extremal optimization algorithm with polynomial mutation. *Information Sciences*. [Online] 330, 49–73. Available from: doi:10.1016/j.ins.2015.10.010.

Zhang, J., Chiodini, R., Badr, A. & Zhang, G. (2011) The impact of next-generation sequencing on genomics. *Journal of genetics and genomics = Yi chuan xue bao*. [Online] 38 (3), 95–109. Available from: doi:10.1016/j.jgg.2011.02.003.

Zhang, P. & Min, X.J. (2005) EST Data Mining and Applications in Fungal Genomics. In: *Applied Mycology and Biotechnology*. [Online]. Elsevier. pp. 33–70. Available from: doi:10.1016/S1874-5334(05)80004-8 [Accessed: 25 September 2019].

Appendix A – Assumption Checks

A.1 Single Effects

A.1.1 Fragmentation

A.1.1.1 Mean Fragment Size

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
E	Based on Mean	2.814	39	160	0.000
	Based on Median	1.138	39	160	0.285
	Based on Median and with adjusted df	1.138	39	61.915	0.320
	Based on trimmed mean	2.749	39	160	0.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: E

b. Design: Intercept + S.DNA + M.SIZE + DNA * M.SIZE

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for E	0.059	200	0.085	0.987	200	0.059

a. Lilliefors Significance Correction

A.1.1.2 Skewness

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
E	Based on Mean	1.128	39	160	0.297
	Based on Median	0.523	39	160	0.990
	Based on Median and with adjusted df	0.523	39	92.839	0.987
	Based on trimmed mean	1.072	39	160	0.372

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: E

b. Design: Intercept + DNA + SKEW + DNA * SKEW

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for E	0.053	200	0.200*	0.995	200	0.693

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

A.1.1.3 Splitting Bias

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
E	Based on Mean	1.507	35	144	0.049
	Based on Median	0.571	35	144	0.973
	Based on Median and with adjusted df	0.571	35	84.299	0.967
	Based on trimmed mean	1.444	35	144	0.070

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: E

b. Design: Intercept + DNA + B.SPLIT + DNA * B.SPLIT

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for E	0.053	180	0.200*	0.989	180	0.156

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

A.1.2 Ligation

A.1.2.1 Ligation Bias

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
E	Based on Mean	2.829	35	144	0.000
	Based on Median	0.829	35	144	0.737
	Based on Median and with adjusted df	0.829	35	77.584	0.728
	Based on trimmed mean	2.734	35	144	0.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: E

b. Design: Intercept + DNA + B.LIGATE + DNA * B.LIGATE

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for E	0.051	180	0.200*	0.992	180	0.388

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

A.1.3 Amplification

A.1.3.1 Denaturation

Levene's Test of Equality of Error Variances^{a,b}

		Levene Statistic	df1	df2	Sig.
E	Based on Mean	3.918	43	176	0.000
	Based on Median	1.322	43	176	0.108
	Based on Median and with adjusted df	1.322	43	38.051	0.192
	Based on trimmed mean	3.753	43	176	0.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: E

b. Design: DNA + T.DENAT + DNA * T.DENAT

Tests of Normality

	Kolmogorov-Smirnov^{a}			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for E	0.046	220	0.200*	0.977	220	0.001

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

A.1.3.2 Elongation (failed two-way ANOVA)

Tests of Between-Subjects Effects

Dependent Variable: E

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
DNA	0.273	3	0.091	3013.578	0.000
T.ELON	0.527	7	0.075	2496.467	0.000
DNA * T.ELON	0.566	21	0.027	893.229	0.000
Error	0.004	124	3.016E-5		
Total	128.961	156			
Corrected Total	0.950	155			

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
E	Based on Mean	3.281	30	124	0.000
	Based on Median	0.966	30	124	0.525
	Based on Median and with adjusted df	0.966	30	25.211	0.540
	Based on trimmed mean	3.078	30	124	0.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: E

b. Design: Intercept + DNA + T.ELON + DNA * T.ELON

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for E	0.078	156	0.022	0.965	156	0.001

a. Lilliefors Significance Correction

A.1.3.3 Elongation (assumption tests for one-way ANOVA)

Levene's Test of Equality of Error Variances[a,b]

DNA			Levene Statistic	df1	df2	Sig.
AT80	E	Based on Mean	5.237	6	28	0.001
		Based on Median	1.027	6	28	0.428
		Based on Median and with adjusted df	1.027	6	7.251	0.477
		Based on trimmed mean	4.760	6	28	0.002
GC50	E	Based on Mean	1.901	7	32	0.102
		Based on Median	1.193	7	32	0.335
		Based on Median and with adjusted df	1.193	7	26.556	0.340
		Based on trimmed mean	1.876	7	32	0.107
GC80	E	Based on Mean	1.213	7	32	0.324
		Based on Median	0.281	7	32	0.957
		Based on Median and with adjusted df	0.281	7	25.423	0.956
		Based on trimmed mean	1.170	7	32	0.347
GCAT80	E	Based on Mean	1.828	7	32	0.116
		Based on Median	0.774	7	32	0.613
		Based on Median and with adjusted df	0.774	7	20.721	0.615
		Based on trimmed mean	1.744	7	32	0.134

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: E

b. Design: Intercept + T.ELON

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
DNA		Statistic	df	Sig.	Statistic	df	Sig.
AT80	Residual for E	0.101	36	0.200[*]	0.932	36	0.028
GC50	Residual for E	0.106	40	0.200[*]	0.981	40	0.723
GC80	Residual for E	0.190	40	0.001	0.916	40	0.006
GCAT80	Residual for E	0.135	40	0.064	0.962	40	0.194

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

A.2 Combined Effects

A.2.1 Splitting Bias and Ligation Bias

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
E	Based on Mean	1.703	63	256	0.002
	Based on Median	0.916	63	256	0.654
	Based on Median and with adjusted df	0.916	63	97.442	0.643
	Based on trimmed mean	1.583	63	256	0.007

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: E

b. Design: DNA + B.LIGATE + B.SPLIT + DNA * B.LIGATE + DNA * B.SPLIT + B.LIGATE * B.SPLIT + DNA * B.LIGATE * B.SPLIT

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for E	0.038	320	0.200*	0.978	320	0.000

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

A.2.2 Fragment Size and Denaturation

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
E	Based on Mean	2.127	63	256	0.000
	Based on Median	0.984	63	256	0.516
	Based on Median and with adjusted df	0.984	63	140.342	0.519
	Based on trimmed mean	2.062	63	256	0.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: E

b. Design: DNA + T.DENAT + M.SIZE + DNA * T.DENAT + DNA * M.SIZE + T.DENAT * M.SIZE + DNA * T.DENAT * M.SIZE

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for E	0.027	320	0.200*	0.997	320	0.777

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

A.2.3 Fragment Size vs Elongation

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
E	Based on Mean	1.842	63	256	0.001
	Based on Median	0.781	63	256	0.878
	Based on Median and with adjusted df	0.781	63	133.167	0.863
	Based on trimmed mean	1.762	63	256	0.001

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: E

b. Design: DNA + T.ELON + M.SIZE + DNA * T.ELON + DNA * M.SIZE + T.ELON * M.SIZE + DNA * T.ELON * M.SIZE

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for E	0.047	320	0.081	0.989	320	0.016

a. Lilliefors Significance Correction

Appendix B– Conference Abstracts

School of Engineering
and Computer Science

Validation of a DNA library preparation model using a genetic algorithm

Nathan Beka[1*], Rene te Boerkhorst[1], and Rod Adams[1], and Neil Davey[1]
[1]*University of Hertfordshire*
*n.beka@herts.ac.uk

Our research is a study of artefacts associated with the library preparation stage of DNA sequencing and how they may be overcome to improve final sequencing outcomes. To investigate these issues a library preparation model was developed, and its associated issues were implemented in the model. To validate our model a genetic algorithm (GA) is used to find optimal parameters for our library preparation model. Our final results show that using parameters selected by the GA we were able to acceptably mimic real-world coverage.

Keywords: Next Generation Sequencing; Library Preparation; Genetic Algorithm; DNA; Coverage

Introduction

Next-generation sequencing has empowered genomics by making it possible to sequence genomes at a lower cost and less time compared to the traditional Sanger method [1]. However, these improvements suffer from reduced accuracy when compared with the Sanger method. During the library preparation stage of sequencing, artefacts can be introduced that affect the reliability of a read [2]. These artefacts can arise from biases due to the structure of the genome, such as preferential splitting of DNA between specific nucleotides [3], bias of adapter ligation towards certain base pair identities [4], and temperature dependent denaturation due to nucleotide composition [5].

Experimental

To investigate this a library preparation model was developed to simulate the occurrences and effects of such artefacts. Our model simulates the following steps of the library preparation process: i) DNA fragmentation, ii) adapter ligation and iii) PCR amplification. To do this a set of parameters characterizing these three steps and a DNA sequence are fed as input to the model and the expected output is coverage scores across the genome. In order to find optimal parameters that would lead to coverage values comparable to those found in real-world sequencing a Genetic Algorithm (GA) was applied. As a fitness function we used the correlation between an actually sequenced genome and the coverage from subjecting that genome to the model.

Results and discussion

After running the GA, we were able to acquire parameters which delivered coverage results that matched the actual coverage for 2 genomes. The first was a 50kbp (kilo base pairs) section of the *Mycobacterium tuberculosis* strain H37Rv genome where the fitness score was 0.83 (Figure 1a). In the second a 50kbp section of the *Plasmodium falciparum* strain 3D7 genome where the fitness score was 0.86 (Figure 1b). In both cases the acquired parameters were able to acceptably mimic coverage. Following these results, we decided to test the acquired parameters on contiguous sections of the tested genomes. In the case of the *tuberculosis* genome it was not possible to mimic coverage across the genome (Figure 2a), but with plasmodium the parameters were able to mimic coverage (Figure 2b). This led us to believe that mimicking coverage across a genome was dependent on its structure.

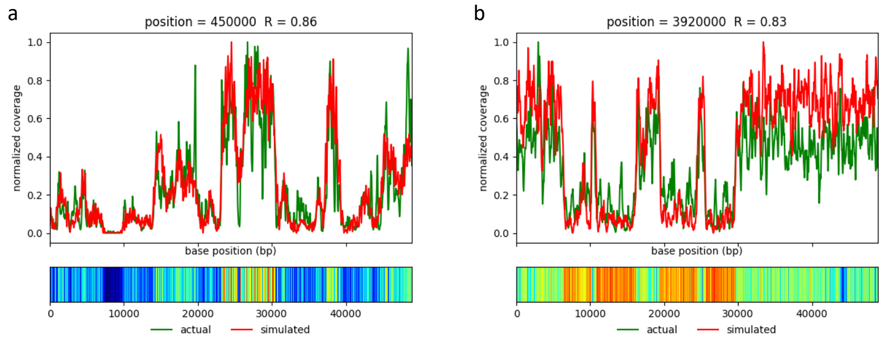


Figure 1. Comparison of results for simulated sequencing and actual sequencing run after evolving model parameters. The colour bar shows levels of base composition bias (blue - red = increasing GC content). (A) Results for section of Mycobacterium tuberculosis genome. (B) Results for section of Plasmodium falciparum genome.

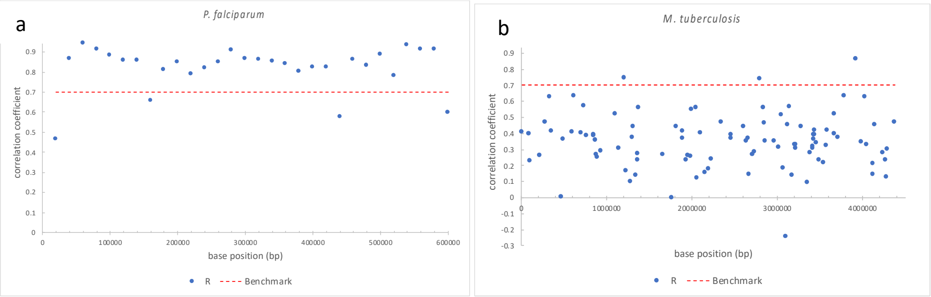


Figure 2. Parameters with the highest fitness score taken from a section of the genome are tested on other parts of the genome. (A) The parameters could not reliably mimic coverage across the Mycobacterium tuberculosis genome. (B) For the Plasmodium falciparum genome, the parameters were able to mimic coverage across the genome.

Conclusion

These results confirm that a GA can be used to optimize our model to obtain coverage values similar to those obtained in real-world sequencing runs. However, in how far the parameters acquired by the GA are representative across a genome depends on the species-specific structure of that genome Our next objective is to analyze the effect of combined and possible knock-on effects of chosen parameter values on coverage given the nucleotide composition of an input genome.

References

[1] E. L. van Dijk, Y. Jaszczyszyn, and C. Thermes, “Library preparation methods for next-generation sequencing: Tone down the bias,” Experimental Cell Research, vol. 322, no. 1, pp. 12–20, Mar. 2014.
[2] M. G. Ross et al., “Characterizing and measuring bias in sequence data,” Genome Biology, vol. 14, no. 5, p. R51, May 2013.
[3] M. S. Poptsova et al., “Non-random DNA fragmentation in next-generation sequencing.,” Scientific reports, vol. 4, p. 4532, Jan. 2014.
[4] A. Seguin-Orlando et al., “Ligation bias in illumina next-generation DNA libraries: implications for sequencing ancient genomes.,” PloS one, vol. 8, no. 10, p. e78575, Jan. 2013.
[5] D. Aird et al., “Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.,” Genome biology, vol. 12, no. 2, p. R18, Jan. 2011.