NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOMKAMPUS

An analysis of the effect of transformation on global- and gene-specific DNA methylation in four cultured cell lines

Jean du Toit, Honns. BSc.

Dissertation submitted for the fulfilment of the requirements for a Masters degree in Biochemistry at the North-West University (December 2010)

Supervisor: Prof. P.J. Pretorius

School for Physical and Chemical Sciences, Division of Biochemistry, North-West University, Potchefstroom Campus, Potchefstroom, South Africa

Optimis parentibus

Know yourself is the whole of science

Only when he has attained a final knowledge of all things will man have come to know himself. For things are only the boundaries of man.

Friederich Nietzsche, "Daybreak" (1881)

Contents

Acknowledgements

Firstly, I want to express my gratitude towards Prof P.J. Pretorius for his amazing guidance during this project. His unwavering dedication and patience helped bring this dissertation to fruition and, furthermore, his moral guidance helped shape this project and my personal life.

I also wish to thank Chrisna Gouws for her unwavering assistance with the practical aspects of my project and for providing me with an excellent role model of a true scientist.

For culturing of experimental cell lines, I wish to thank Dr. Oksana Levanets, Etresia van Dyk, Lizelle Zandberg and Chrisna Gouws. Also Jaco Wentzel and Chrisna Gouws for their assistance with the cytosine-extension assays.

Finally, big thanks to all my friends, both from Biochemistry and from other disciplines. I also want to express gratitude towards my parents for all their help and encouragement during the past year.

Abstract

DNA methylation plays a role in several biological functions, such as gene expression regulation, and several endogenous and exogenous factors affect these DNA methylation patterns in the cell. One such alteration of a cell line's DNA methylation pattern is caused by the insertion of a vector into the cell line. Using the cytosine-extension assay and realtime methylation-specific PCR, alterations of DNA methylation levels on both global and gene-specific levels were investigated. In some cell lines the cellular transformation led to an increase in DNA methylation levels, and in others a decrease in DNA methylation amounts was observed. The same phenomenon was seen in the promoter regions of specific genes, showing that vector-insertion into a cell line caused DNA methylation alterations in many regions of the genome. These alterations in DNA methylation are investigated in this reduced representation study using enrichment of the methylated fraction of fragmented DNA and subsequent GS FLX Titanium sequencing of these methylated fragments. The results of sequence data analysis showed that methylated fragments are distributed over the whole genome, but could be related to only a few specific genes. These results have implications for cell culture work, biotechnological applications and uses in gene therapy.

Key words: Epigenetics, DNA methylation, cell culture, transfection, DNA sequencing, Cytosine-extension assay, real-time methylation-specific PCR

Opsomming

DNS-metilering speel ‘n belangrike rol in verskeie biologiese funksies, soos byvoorbeeld gedurende geen-uitdrukking, en daar is verskeie eksogene en endogene faktore wat ‘n rol speel om DNS-metileringspatrone te beïnvloed. Een so verandering van die metiloom word veroorsaak deur die invoeging van ‘n vektor in ‘n sellyn. Deur gebruik te maak van die sitosien-verlengings toets (Engels: CEA) en intydse metilerings-spesifieke PKR (Engels: Realtime PCR) kan veranderinge in DNS-metileringsvlakke op ‘n globale en geen-spesifieke vlak gemeet word. In hierdie studie is ‘n toename in DNS-metileringsvlakke waargeneem by sommige sellyne, terwyl die DNS-metileringsvlakke van ander afgeneem het. Dieselfde verskynsel is waargeneem in spesifieke geen-promotor gebiede, wat daarop dui dat die invoeging van ‘n vektor binne ‘n sellyn veranderinge in die DNS-metileringspatrone veroorsaak. Hierdie studie benader die verskynsel vanuit ‘n verlaagde voorstelling oogpunt (Engels: Reduced representation view) om die veranderinge wat in vier sellyne se DNS-metileringsvlakke voorkom, as gevolg van die invoeging van ‘n vektor, waar te neem. Hiervoor word verryking van die gemetileerde fraksie van gefragmenteerde DNS, gevolg deur DNS-volgordebepaling (Engels: Sequencing) van die gemetileerde fragmente, gebruik. Die resultate van die volgordebepaling het getoon dat die gemetileerde fragmente met ‘n paar gene gekorreleer kon word en dat gemetileerde fragmente regoor die genoom versprei is. Hierdie resultate is belangrik aangesien dit toepassings het in selkultuur werk, biotegnologie toepassings en in geenterapie.

Key words: Epigenetika, DNA metilering, selkulture, transfeksie, DNS volgorde-bepaling, sitosien-verlengings toets, intydse metilerings-spesifieke PKR

List of figures

List of tables

List of abbreviations

A

A: Adenine

ATCC: American Type Culture Collection

B

bp: Base pair

BLAST: Basic Local Alignment Search Tool

C

C: Cytosine

cat: Catalogue

CEA: Cytosine extension assay

CH_3: Methyl group

CpG: Cytosine coupled to a guanine by means of a phosphate link

C_T: Threshold cycle

D

ddH_2O: Double distilled water

DHF: Dihydrofolate

DNA: Deoxyribonucleic acid

DNMT: DNA methyltransferases

dTMP: Deoxythymidine monophosphate

dUMP: Deoxyuridine monophosphate

E

EDTA: Ethylenediaminetetra acetic acid

Et al: And others (Latin)

EtBr: Ethidium Bromide

EtOH: Ethanol

F

FAA: Fumarylacetoacetate

G

G: Guanine

g: Gram

g: Gravity (9.8 $m.s^{-2}$)

GC: Guanine Cytosine dinucleotide

H

H: Hydrogen

HCC: Hepatocellular carcinogenesis

HT1: Hereditary Tyrosinemia Type I

HTA: Human Tissue Act

I

IEM: Inborn errors of metabolism

M

MAA: Maleylacetoacetate

MBD: Methyl CpG binding domain

Mg: Magnesium

miRNA: MicroRNA

ml: Millilitre

MM: Mastermix

MSP: Methylation-specific PCR

MTHFR: Methylenetetrahydrofolate reductase

N

ng: Nanogram

P

PBS: Phosphate-buffered saline

PCR: Polymerase chain reaction

pH: Potential of Hydrogen

R

RNA: Ribonucleic Acid

rpm: Revolutions per minute

RQ: Relative Quantification

S

SAM: S-adenosylmethionine

T

T: Thymine

TAE: Buffer solution containing a mixture of **T**ris base, **a**cetic acid and **E**DTA

THF: Tetrahydrofolate

TSA: Trichostatin A

U

UV: Ultraviolet

W

WGA: Whole genome amplification

List of symbols

µl: Micro Litre

^{o}C: Degrees Celsius

λ: Lambda

Chapter 1

Introduction

1.1. Project introduction

DNA methylation describes the epigenetic mechanism whereby a methyl-residue is attached to the carbon-5 position of the cytosine ring on a specific DNA sequence (Wilson *et al*., 2007; Prokhortchouk and Defossez, 2008).

DNA methylation alterations in the promoter regions of certain genes can silence these genes (Esteller, 2008). This occurs when CpG islands (CG-rich regions) in the gene promoter areas of the genes are hypermethylated (Das and Singal, 2004) and also takes place in cancer when silencing of tumour suppressor genes occurs (Lee and Lee, 2003). Hypomethylation is associated with cancer development (Esteller, 2008) and global hypomethylation has been identified as a major contributing factor to oncogenesis (Das and Singal, 2004). The effects on cellular function by hypo- and hypermethylation indicate the importance of DNA methylation pattern regulation in cells, and shows that deregulation of DNA methylation can be associated with disease states such as cancer.

Cell culture work often makes use of transformation to study specific effects in cell cultures. However, alteration of DNA methylation patterns and DNA methylation amounts (induced by transformation) may have influences on the cell cultures that a researcher may not be aware of. Transformed cell lines often produce unexpected proteins due to alterations in cellular characteristics, which may be attributed to silencing of some genes and the activation of others (Esteller, 2008).

There are several endogenous and exogenous factors that affect the DNA methylation patterns of cell lines (Das and Singal, 2004). Work in our laboratory has shown that vector-insertion into a cell line has effects on the DNA methylation amounts, causing hypomethylation in some cell lines (Kok, 2009). In this study, the effect of transformation on the DNA methylation levels of several cell lines (143B, fibroblasts, HepG2 and HeLa) was investigated and DNA methylation amount comparisons were made between untransformed and transformed samples of each of the four cell lines.

A reduced representation view approach was taken in this study; this approach can be used in various methylation analyses (Wiedmann *et al*, 2008). Reduced representation reduces the complexity of the genome being studied by many orders of magnitude, allowing identification of several gene candidates that are methylated in the sample DNA (Wiedmann *et al*, 2008).

This study made use of the cytosine extension assay to determine global amounts of DNA methylation in samples. Real-time methylation-specific PCR (MSP) was used to identify the amount of DNA methylation present in the promoter regions of the *MGMT*- and *P16*^{INK4a}-gene. Both these techniques were used to determine whether the amount of DNA methylation was indeed altered with transformation. Results produced showed that there were indeed alterations in the DNA methylation levels after cellular transformation.

After preliminary work, the sample DNA was fragmented using enzymatic methods. Enrichment of methylated fragments was done using a commercially available kit (the MethylMiner kit from Invitrogen).

The methylated fragments were then sent for DNA sequencing at Inqaba Biotec to generate sequence data. This data was analyzed and genes that were methylated in the samples were identified. Other analyses included investigations into the CG-content of the fragments and fragment-length investigations.

1.2. Dissertation outline

In Chapter 2, relevant literature on DNA methylation will be discussed. The literature section will be linked with the final conclusion in chapter 6, which incorporates study results whilst taking into consideration available DNA methylation literature. Chapter 3 will discuss the study aims and experimental approach, while chapter 4 presents the results of a pilot study which was done in order to investigate the effect on DNA methylation of the transformation of cultured cells with a cloning vector. This section includes results from the cytosine-extension assay and real-time MSP analyses. Chapter 5 discusses the selection and partial characterisation of methylated DNA isolated from different cell lines by means of high-

throughput DNA sequencing. Finally, chapter 6 will give a summative conclusion, relating study results to literature and discussing future application and expansion of the project.

Chapter 2

Literature Review

2.1. Classic genetics and epigenetics

The world of science is synonymous with paradigm shifts: discoveries that forever change the way the world and science itself is viewed. Such a revolution happened in the world of molecular biology with the elucidation of the chemical structure of deoxyribonucleic acid (DNA) by Watson and Crick (Watson and Crick, 1953). This monumental proposal of the chemical structure of DNA heralded the age of classic genetics. It allowed scientists to view hereditary, through the use of DNA assorted into chromosomes, with new insight and is the main reason for the remarkable growth of molecular biology and genetics since that time. The growth of knowledge concerning the nucleic acids and proteins, and subsequently of the machinery of the cell itself, also advanced exponentially (Olby, 2003). In short, Watson and Crick's model of DNA provided amazing explanatory power in molecular biology (Olby, 2003; Watson and Crick, 1953).

There are, however, several phenomena that cannot be explained by simply using the theories and techniques of classic genetics. These include the occurrence of a large number of varying phenotypes in a given population, as well as cloned organisms and monozygotic twins which share identical DNA sequences but possess different phenotypes (Esteller, 2008), or various disease states where siblings with the same metabolic disease have different clinical presentations (Mitchell *et al*., 2001). Epigenetics provide a stimulating avenue for investigations into these phenomena.

2.2. DNA methylation

"Epigenetics" refers to heritable changes in the functioning of genes that occur without the modification of the nuclear DNA sequence, i.e. without changing the underlying DNA sequence, the phenotype and gene expression is altered (Wilson *et al*., 2007; Prokhortchouk and Defossez, 2008). The main mechanisms of epigenetics are microRNAs that regulate gene expression, histone modifications and the establishment of DNA methylation patterns (Laird, 2005; Wilson *et al*., 2007). This study focuses on DNA methylation.

DNA methylation is the best-known epigenetic marker in eukaryotes and its role in the nucleus architecture of an eukaryotic cell is essential for gene-activity control (Esteller, 2008; Laird, 2010). It is also a key component in the aging process and plays an important role in the development of several diseases, e.g. cancer, due to the effects it has on gene expression (Wilson *et al*., 2007). Furthermore, epigenetic mechanisms provide stability to the phenotype (Laird, 2010).

DNA methylation describes a covalent chemical modification that results in the addition of a methyl (-CH_3) group at the carbon 5 position of the cytosine ring (Das and Singal, 2004; Laird, 2010). S-adenosylmethionine (SAM) serves as a methyl-donor during this process, transferring a methyl-group to the DNA methyltransferases as illustrated in figure 2.1.

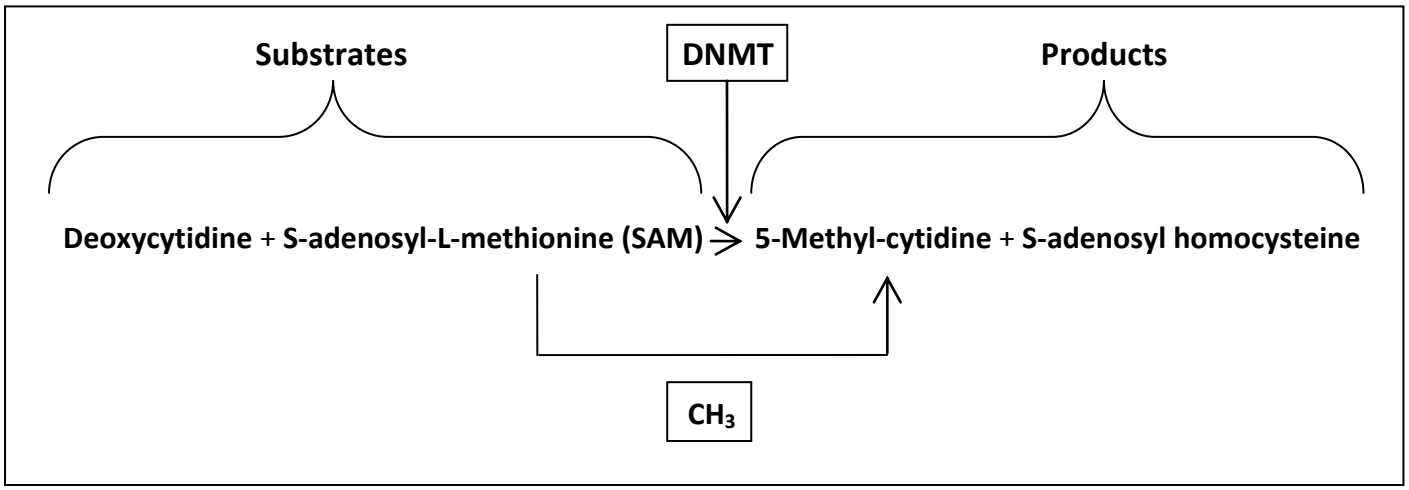


Figure 2.1: DNA methylation catalyzed by DNA methyltransferases. A methyl-group is transferred from S-adenosyl-L-methionine (SAM), the methyl-donor, to the C5 position of the 5-Methyl-cytidine by means of DNA methyltransferase (DNMT) compounds.

DNA methylation does not occur uniformly in the human genome; there are methylated regions in the DNA sequence that are interspersed with unmethylated regions (Das and Singal, 2004). DNA methylation occurs on the cytosine in CG-dinucleotides spread out throughout the entire human genome (Das and Singal, 2004; Laird, 2005) and in CpG islands, which are CG-rich regions found mostly in the gene-promoter regions (Esteller, 2008; Laird, 2010). Gardener-Garden defined a CpG island as a region greater than 200 bp with CG-content greater than

50% (Gardiner-Garden, 1987). CpG islands span the 5' end promoter regions of about 50-60% of genes with tissue-specific patterns of expression and in housekeeping genes (Bird, 2002; Das and Singal, 2004; Laird, 2005; Prokhortchouk and Defossez, 2008) and often overlap with transcription start sites (Laird, 2010). CpG islands tend to be unmethylated (Das and Singal, 2004; Duffy *et al*., 2009), and DNA methylation is usually a repressive mark when located in gene promoters (Laird, 2010).

Investigations into the complex mechanisms of epigenetics, such as DNA methylation, could potentially lead to therapeutic targets for the treatment of several disease states (Das and Singal, 2004).

2.3. DNA methyltransferases

A question now arises of how DNA methylation patterns are established in the DNA sequence, both during cellular replication (DNA methylation is not preserved during DNA replication (Nephew and Huang, 2003; Jones and Liang, 2009)) and in the somatic developmental phases.

DNA methylation is catalyzed by enzymes known as DNA methyltranferases. Three types are active in mammals: DNMT1, DNMT3a and DNMT3b (Bird, 2002; Prokhortchouk and Defossez, 2008). These methyltransferases have highly conserved catalytic motifs (Laird, 2010).

Hemi-methylated DNA is generated during the replication of methylated DNA. This happens because the DNA methylation patterns on the original DNA strands are not transferred to the newly generated DNA (Nephew and Huang, 2003). DNMT1 is a major factor in causing the DNA to become fully methylated again (Prokhortchouk and Defossez, 2008) and in this way methylation patterns are preserved on newly-replicated DNA strands (Das and Singal, 2004). This process is known as "maintenance" methylation (Oakeley and Chiang, 1999).

DNMT3a and DNMT3b, in unison with the DNMT3L co-factor, are *de novo* methylating agents which cause new methylation patterns to be established on DNA sequences which were unmethylated (Oakeley, 1999; Prokhortchouk and Defossez, 2008). Whilst *de novo*

methylation is usually restricted to germ cells or the early embryo, some research suggest that *de novo* methylation could also occur in adult cells (Bird, 2002).

DNA methylation patterns are established early in embryogenesis (Das and Singal, 2004). Blastocyst cells divide without any detectable methylation levels early in embryonic development and DNA methylation becomes detectable around the time of implantation. This is important for the viability of individual cells. During development, around the time of implantation, a portion of all CpG islands become methylated, which causes the associated promoter to be stably silenced. These specific methylation patterns are also associated with X-chromosome inactivation and genomic imprinting (Razin and Shemer, 1995; Bird, 2002).

2.4. The significance of DNA methylation

The functions of DNA methylation are numerous, but the most important functions are regulatory effects on gene expression (Das and Singal, 2004). The following sections briefly discuss the significance of DNA methylation *in vivo*.

2.4.1. Housekeeping functions of DNA methylation

Manel Esteller states that DNA methylation plays a crucial role in control of gene activity (Esteller, 2008). As such, DNA methylation represents a level of control for certain tissue-specific genes, is required for genomic imprinting to ensure monoallelic expression and is involved in X-chromosome inactivation in females (Laird, 2005).

Furthermore, in almost all disease states the epigenetic regulation which is specific for certain cell types becomes deregulated – refer to section 2.5 (Gargiulo and Minucci, 2009). In these cases, the role of DNA methylation can be seen unmistakably, specifically in cells where the DNA methyltransferases are defective or disrupted (Esteller, 2008). Alterations in DNA methylation are common in tumours (Das and Singal, 2004) and in cancerous cells and cells where DNA methyltransferases do not function correctly, nuclear abnormalities are prominent. This is due to the lack of DNA methylation's stabilizing effect (Esteller, 2008).

2.4.2. Gene expression

Gene expression is affected by gene promoter methylation. DNA methylation is therefore important as a regulator for gene transcription and has a leading role in carcinogenesis when aberrant (Das and Singal, 2004). DNA methylation affects gene transcription by means of altering protein-DNA interactions (Razin and Cedar, 1991).

2.4.3. Gene silencing

DNA methylation has differing effects on the expression of genes, depending on where the increased methylation occurs. If methylation is increased in the promoter regions of a gene, the effect will usually be reduced expression of that gene and, possibly, silencing of the gene. Methylation in the transcribed region of a gene will, however, have a variable effect on gene expression (Das and Singal, 2004). As has already been stated, gene transcription is affected by means of the alterations of protein-DNA interactions when DNA is methylated (Razin and Cedar, 1991). One of the mechanisms by which this is achieved involves direct interference with the binding of specific transcription factors to their recognition sites in the promoter regions. Another mode of repression involves direct binding of transcriptional repressors to methylated DNA (Das and Singal, 2004).

2.5. Aberrant DNA methylation

Aberrant DNA methylation describes the state wherein DNA is either hyper- or hypomethylated beyond the norms for regular functioning of the DNA; it is usually associated with disease states (Gargiulo and Minucci, 2009).

Typically, hypermethylation of DNA involves CpG islands, whilst hypomethylation involves repeated DNA sequences (for example long interspersed nuclear elements) that are found spread out across the human genome (Das and Singal, 2004).

2.5.1. Hypomethylation

Global hypomethylation has been identified as a major contributing factor to oncogenesis (Das and Singal, 2004). One of the most obvious examples of DNA hypomethylation in the disease state occurs in tumours. A particularly low level of DNA methylation has been noted in tumour cells as compared to the level of DNA methylation in healthy cells from the same tissues as the specific cancerous cells. This loss of methylation can be ascribed to hypomethylation of repeat DNA sequences and to the demethylation of introns and coding regions. The degree of hypomethylation that occurs in cancerous genomic DNA increases with the development of cancer from benign to invasive (Esteller, 2008).

2.5.2. Hypermethylation

Hypermethylation of promoter CpG islands generally affects tumour-suppressor genes, especially in cancer cells (Laird, 2005). Aberrant hypermethylation represses the transcription of tumour-suppressor regions and promoter regions, which leads to gene-silencing (Das and Singal, 2004). This CpG island hypermethylation forms a part of an integrated series of changes in histone modifications and chromosome structure that occurs during cancer development (Laird, 2005).

Translocations, chromosomal instability and gene disruption due to the reactivation of transposable DNA sequences may possibly be prevented through hypermethylation of repetitive genomic sequences (Esteller, 2008).

2.6. Factors that affect DNA methylation

Several endogenous and exogenous factors can cause DNA methylation levels to be altered (Das and Singal, 2004). The goal of this study is to investigate the effect of vector-insertion into a cell line, but several other factors have an effect on DNA methylation patterns in the *in vitro* and *in vivo* environment. These factors also indirectly affect gene expression regulation, a function of DNA methylation discussed in section 2.4.

Nutrients play an important role in affecting gene expression by means of modulation of DNA methylation and interactions with genetic polymorphisms (Das and Singal, 2004). Disruption of methyl-group metabolism may lead to several diseases (including cancer), especially when folate and cobalamin (vitamin B12) are deficient (Das and Singal, 2004; Okochi-Takada *et al*., 2004; Esteller 2008). Furthermore, diets which are deficient in choline and methionine (or other methyl-donors) affect the levels of the universal methyl-donor S-adenosyl-L-methionine (SAM) negatively, which could lead to DNA hypomethylation (Okochi-Takada *et al*., 2004; Esteller, 2008).

Figure 2.2 shows that the metabolism of methyl groups can be divided into two "branches": the first involving purine and thymidine synthesis and the second that of methionine and s-adenosylmethionine synthesis for the purpose of protein and polyamine generation and DNA methylation reactions (Das and Singal, 2004).

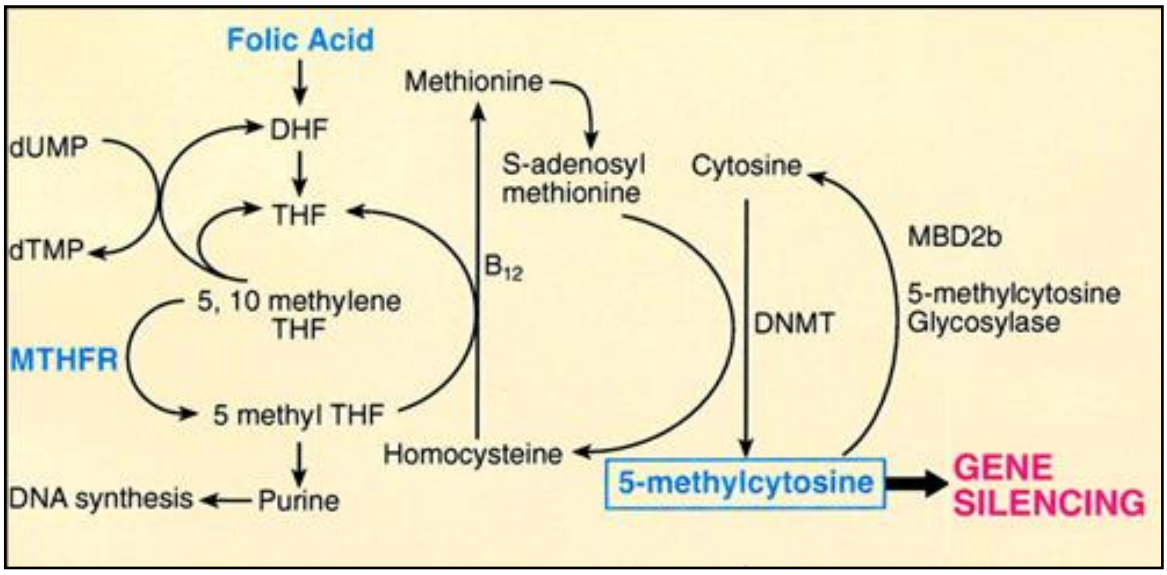


Figure 2.2: Overview of the folic acid pathway illustrating the generation of 5-methylcytosine and how gene silencing is caused via dietary intake (adapted from Das and Singal, 2004).

The importance of the correct functioning of this metabolism is evident from individuals where the MTHFR enzyme (refer to figure 2.2) is deficient (Das and Singal, 2004; Esteller, 2008). The purpose of this enzyme is to shift methyl groups from the first "branch" of this metabolic

pathway to the second. In patients where this metabolism does not function correctly, an increased amount of homocysteine in the blood and urine can be noted and mental retardation and thrombo-occlusive vascular disease are often the result (Das and Singal, 2004).

Diets deficient in vitamin B12, folate or malate (Das and Singal, 2004; Okochi-Takada *et al*., 2004; Esteller, 2008) also have serious implications for the DNA methylation patterns, causing widespread DNA hypomethylation and leading to associated disease states.

The influences of diet and energy metabolism, aging and disease states are endogenous influences on DNA methylation patterns. There are also exogenous chemical agents that affect DNA methylation patterns. These include demethylases (including exogenous demethylating agents such as 5-aza-2'-deoxycytidine and valproic acid epigallocatechin-3-gallate) that cause hypomethylation (Okochi-Takada *et al*., 2004). Several other mechanisms also influence the methylation status of DNA sequences, such as methylation centres that trigger DNA methylation and methylation protection centres (Das and Singal, 2004; Okochi-Takada *et al*., 2004).

2.7. Harnessing the power of DNA methylation

Studies into regional DNA methylation patterns, as well as global methylation profiles, may help researchers understand how epigenetic changes - such as DNA methylation alterations – might enable aberrant gene expression patterns and lead to disease states (Laird, 2010). Techniques such as epigenomic profiling have the eventual aim of scanning the entire epigenome for alterations that may cause or maintain oncogenic alteration. In this way, epigenetics may be integrated with classic genetics (i.e. structural information integrated with gene regulation) (Gargiulo and Minucci, 2009). Particularly, techniques for studying cytosine methylation at specific loci can lead to the elucidation of a methylome for specific disease states. Thus far aims have been to compare the methylation profiles of different cell types with each other, as well as those of tumourous and non-tumourous samples. These studies may have important clinical and diagnostic implications in the future. An interesting possibility is the early prognosis of cancer (Das and Singal, 2004), for example by using methylated genes

as cancer biomarkers (Duffy *et al*., 2009). Determination of the full extent of CpG island methylation, however, will require more detailed genome-wide analyses (Bernstein *et al*., 2007).

2.8. Techniques for studying DNA methylation

Originally, DNA methylation studies were based on gel electrophoresis of methylation-sensitive enzyme digestion of DNA samples (Laird, 2010). Today, several different research approaches exist. These include non-specific DNA methylation analyses (including thin-layer chromatography, use of SssI methyltransferases, the chloracetaldehyde reaction and immunological studies) and gene-directed methylation analyses. Some restriction endonucleases may be used for the latter type of studies, as these restriction endonucleases are usually isoschizomeres of each other with different methylation-sensitivities allowing discernment of methylated and unmethylated genome regions (Oakeley and Chiang, 1999). Restriction enzymes are used in the cytosine-extension assay (see section 4.2). In this study, a technique based on enrichment of methylated DNA fragments is used to investigate the methylome. After enrichment of methylated DNA, direct sequencing of methylated DNA is done.

Chapter 3

Aims and study approach

3.1. Introduction

The title of this study makes reference to investigating the effect of “transformation” in cell cultures. The process of transformation refers to the alteration of a cell due to the uptake, incorporation and expression of exogenous genetic material. Most commonly, this happens when a plasmid is inserted into a competent bacterial cell. Transformation is most commonly used to refer to non-viral DNA transfer in bacteria, non-animal eukaryotic cells and plant cells. “Transfection” refers to the delivery of a vector into a eukaryotic cell by non-viral means to introduce nucleic acids into a cell (Wilson and Walker, 2005). In this study, the terms “transformation” and “transfection” will be used interchangeably to refer to the insertion of a vector into one of several eukaryotic cell lines.

It is hypothesized that the insertion of a vector into a cell causes alterations in DNA methylation beyond those changes that occur naturally in cells. The alterations that have been observed in our laboratory indicate that the insertion of a vector into a cell line may be correlated with a state of hypomethylation (Kok, 2009). This study will expand upon those observations by directly investigating the DNA sequencing data of methylated regions in sample DNA.

3.2. Aims of this study

The aim of this study is to investigate alterations in DNA methylation caused by the insertion of a vector. Study aims have been divided into primary and secondary aims.

The primary aim of this study is to investigate perturbations in global and gene-specific DNA methylation when an expression vector is inserted into eukaryotic cultured cells. As already stated, a previous study in our laboratory showed that there exists DNA methylation differences between cell lines and transfected versions of those same cell lines (Kok, 2009). In this study, further investigation of this phenomenon will be done in an attempt to identify and characterize the DNA species involved in the perturbation of DNA methylation.

The secondary aim of this study is to characterize the DNA sequences that are involved in these perturbations through the use of high-throughput sequencing, with the possibility of identifying genes that are methylated in both the cell line and its transformed counterpart.

3.3. Study approach

Four untransformed cell lines (143B, fibroblasts, HepG2 and HeLa cell lines), as well as their vector-containing counterparts, are the focus of this study. Three of the cell lines are cancer cell lines (143B, HepG2 and HeLa) and the remaining one is a healthy human fibroblast cell line.

DNA was isolated from the eight cell cultures after the cell lines had been cultured successfully. This isolated DNA was used to determine the effect of transformation on global, as well as gene-specific DNA methylation levels. Subsequently, the MethylMiner kit was used to enrich methylated DNA fragments from the samples for GS FLX Titanium sequencing. Analysis of generated data was done after DNA sequencing of all samples was completed. An element of this step was to compare the generated data with available BLAST data for each of the cell lines. Data analysis was done using Microsoft Excel, CLC Bio Genomics Workbench and the NCBI BLAST search engine (URL: *www.ncbi.nlm.nih.gov/blast/*).

A reduced representation view is taken in the study to identify several methylated gene candidates for each of the cell lines (Wiedmann *et al*., 2008).

Chapter 4

A study of the effect on DNA methylation of the transformation of cultured cells with a cloning vector

4.1. Introduction

Human cells will continue to grow in an *in vitro* setting - under the correct environmental conditions and if supplied with sufficient nutrients - once they are removed from a living environment (*in vivo*). This statement forms the main theoretical basis of cell culture work, which simply refers to cells that divide and grow once placed in a suitable *in vitro* environment. Most batch cell lines will continue to grow until limits are imposed by restricting nutritional factors (Chaudry, 2004). Once such a cell culture has successfully been established, research on the specific cells may commence (Wilson and Walker, 2005).

There are several technical aspects that should be considered when working with cell cultures. Cells may be grown as either suspension cultures or as cells attached to a solid surface, but all cells used in this study were adherent cell cultures. “Primary cultures” refer to freshly isolated cells from mammalian tissues, whilst “continuous cell lines” refer to the cells after several sub-cultures onto fresh media (Wilson and Walker, 2005). All the cells used in this study have become continuous cell lines after extended periods of sub-culturing. Continuous cell lines may have different cellular characteristics when compared to primary cell cultures; these include changes in cell morphology and chromosomal variation. Consideration should be given to these alterations when considering the results obtained from cell culture studies (Chaudry, 2004).

It is known from literature that various factors play a role in the methylation of cellular DNA (refer to chapter 2 for a complete discussion). One factor under investigation in this study is the effect of transformation of a cell line by means of a cloning vector (i.e. transfection of a cell line’s genetic material). This study investigates the effect that the insertion of an expression vector into cell cultures has on the DNA methylation patterns of specific cell lines. The human-derived cell lines used in this study are shown in table 4.1.

Table 4.1: List of cell lines investigated in this study

Cell line name	Type of cell line	Origin of cells
143B	Cancer cell line	Bone osteocarcinoma
HepG2	Cancer cell line	Hepatocellular liver carcinoma
HeLa	Cancer cell line	Cervical cancer cells
Fibroblasts	Connective tissue cell line	Epithelial fibroblasts

Cancer cell lines are derived from tumours. These cells can replicate indefinitely under the correct conditions and express at least some of the characteristics of the cells of origin (Alberts *et al.*, 2002). Each of the four basic cell lines listed in table 4.1 was investigated in the study, while another aliquot of each was transformed and studied. This produces a total of eight different samples, which are the samples examined in the study.

Different transfection techniques were used to generate the transformed cells. All samples have been stably transfected and a discussion of each of the transfection techniques follows below.

143B cells: Transfection of the 143B cells was achieved using ExGen *in vitro* transfection reagents from Fermentas (cat. # R0511), which is a sterile solution of linear polyethylenimine (PEI) in water. This was followed with antibiotic selection (using Puromycin) before development of individual clones (after several days) through the use of a serial dilution method. Transfection and cell maintenance was performed by Dr Oksana Levanets.

HepG2: A stable HepG2 tTs cell line was generated by means of transfection with a ptTS-Neo vector, followed by selection with antibiotic selection (using the antibiotic G418). Transfected cells may be isolated after several weeks. After isolation and propagation of these cells, a new HepG2 tTS cell line is established. This cell line is expected to provide a high level of tTS expression. Transfection and cell maintenance was performed by Chrisna Gouws.

HeLa: Transformed cells contain the same vector as the transformed HepG2 cells (vector ptTS-Neo vector). HeLa tTS cells were purchased from Clontech (cat. # 630928). Transfection and cell maintenance was performed by Etresia van Dyk.

Fibroblasts: Transfection of the cell line was achieved by means of electroporation using 0.4 cm cuvettes. 25ug of plasmid was used in 30ul of solution. Preparation of the transfected fibroblast cells is based on the work of Litzkas P *et al*, who used a recombinant pRSVneo vector (Litzkas *et al*., 1984). In the preparation of these cells however, a pTracer vector was inserted into cells by means of electroporation and antibiotic selection was done through the use of G418. Transfection and cell maintenance was performed by Lizelle Zandberg.

All the cell cultures used during this study are adherent cells, and therefore a trypsination step is needed to place the cells in suspension for use in experiments. Trypsin is used for this purpose. An accurate measurement of the number of cells suspended in a particular volume is also needed for optimal DNA isolation from the cells. Cells are stained with Trypan blue and counted using a counting chamber to accurately determine the number of cells.

A great deal of ethical consideration and control is required to obtain cells for culturing, as is exemplified by the Human Tissue Act 2004 (HTA) in Great Britain (Anon., 2004), and whilst further cell culture work use does not require ethical approval, the original ethical concerns should not end with the appropriation of the cell cultures from the original donor and should be recognized by subsequent researchers as well. Two main issues concerning cell cultures today is that of privacy of the original donor and ownership of subsequent research and of the cell cultures themselves (Upile *et al*., 2009). However, if these ethical considerations are kept in mind, cell culture research provides an appealing avenue for various types of molecular biology studies.

Major advantages of using cell cultures in molecular biology studies are reproducibility of results and consistency in studies. Furthermore, the relatively small amount of ethical concern (except for the factors mentioned in the previous paragraph) in comparison to animal models

are also advantageous. Some disadvantages, however, become pressing after a long period of cell culture growth. Firstly, cell characteristics may change over time and, secondly, some cell lines may adapt to the culture environment by varying enzyme activity (Chaudry, 2004). These elements have serious implications for the results generated from cell culture work and show that results of cell culture work may be influenced by several external factors.

4.2. Global DNA methylation

Global DNA methylation refers to the overall amount of DNA methylation present in a specific sample. Using global DNA methylation measurement techniques, a researcher is able to assess the total number of methylated cytosines in comparison to the total number of unmethylated cytosines.

4.2.1. Cytosine-extension assay (CEA)

The central hypothesis to be tested during this pilot study is whether there is a difference in global DNA methylation levels between untransfected cell lines and their transfected counterparts (Kok, 2009). The first step to investigating this principle is to measure the total (global) DNA methylation levels. These differences in global DNA methylation levels between samples were quantified by means of the CEA, a technique already standardized in our laboratory (Wentzel *et al*., 2010).

4.2.2. Theoretical basis of CEA

CEA makes use of methylation-sensitive restriction enzymes (*HpaII* and *MspI*) to cleave DNA isolated from cells. These two enzymes are isoschizomers, i.e. they have the same recognition sites, but they are not equally sensitive to DNA methylation at this restriction sites (5'-C^CGG-3'). *HpaII* is sensitive to site-specific DNA methylation of the cytosine in the CG-dinucleotide (it will not cleave the DNA at sites of CG-dinucleotides if methylation is present and will only cleave unmethylated forms of the restriction sites), whilst *MspI* will hypothetically cleave all restriction sites, irrespective of methylation state. Differentiation therefore depends on the

methylation status of the second cytosine in the recognition site. This differential sensitivity allows DNA methylation differences to be determined.

The enzyme digestion of DNA via these enzymes generates a 5' guanine overhang, where [3H]dCTP can be incorporated (this will be described in section 4.2.3). [3H]dCTP causes a single nucleotide extension at the point of DNA cleavage and radioactive cytosine incorporation into the DNA is used, by means of scintillation counting (counting of the number of disintegrations per minute due to the radioactive material integration into the samples), to calculate the absolute DNA methylation percentage present in the samples.

4.2.3. Overview of CEA protocol

DNA was isolated from the cells using FlexiGene DNA kit from Qiagen (cat. # 51204) according to the manufaturer's guidelines and two separate enzyme digestions of the same sample was done, one using *MspI* and the other using *HpaII*.

Next, [3H]-dCTP was integrated at the 5' guanine overhang using GoTaq DNA polymerase from Promega (cat. # M3001) to create an "isotope mixture". This step forms the central part of the CEA. The "isotope mixture" for each sample was then transferred to Whatman DE-81 ion-exchange filter (cat. # 3658-325) and a vacuum was generated. Samples were washed with PBS (phosphate-buffered saline); the isotope-labeled DNA remained fixed on the ion-exchange filters while the rest of the mixture flowed through. Once the filters were completely dry, each was placed in a separate glass counting vial with scintillation fluid. The samples were then placed in the scintillation counter and the DPM (disintegrations per minute) was measured.

4.2.4. Results of global DNA methylation

Results are expressed as relative [3H]-dCTP incorporation per 0.5mg DNA. These results are given as percentages. A total of four counts are generated per sample, because each sample is counted twice by the scintillation counter and the experiment is done in duplicate. The

average is calculated with sigma = 2% and all values are expressed as DPM-values. The main calculation, used to estimate the degree of methylation, is calculated by dividing the DPM-value of the *HpaII*-digested samples with the DPM-value of the *MspI*-digested samples.

Analysis of the data was done in Microsoft Excel based on the DPM-values generated with the scintillation counter. CEA results are presented as a graph showing a comparison of the untransfected cell lines' global methylation levels with the global methylation levels of the transfected cell lines (i.e. the cell lines with an inserted vector). The results of the CEA are given in figure 4.1 below:

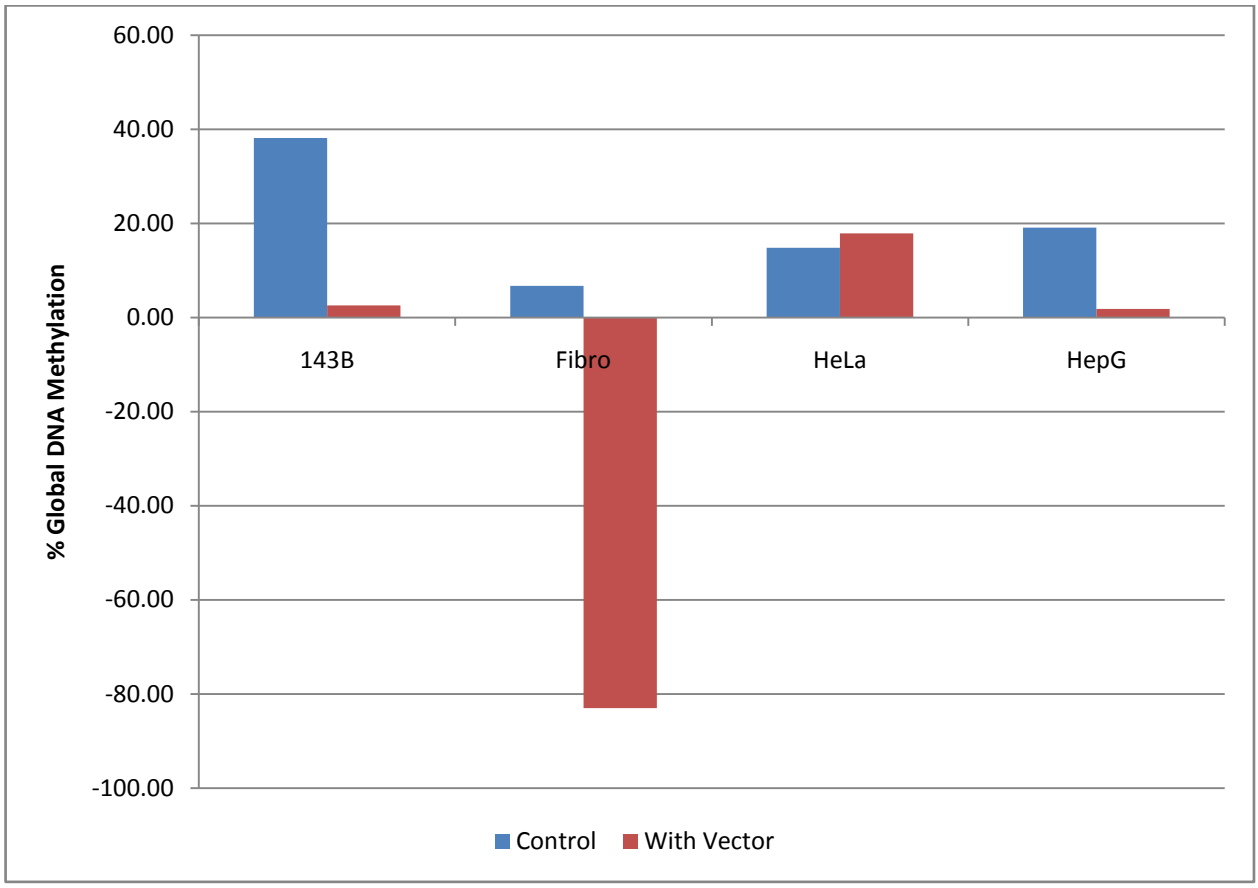


Figure 4.1: A comparison of the effect of transformation on global DNA methylation in different cell lines. No error bars are shown in the figure, due to the fact that calculations are done relative to two enzyme digestions.

Figure 4.1 is summative of three different CEA experiments and shows the averages of the results. This was done to minimize variation between experimental runs and to gain a clearer view of the actual global methylation levels of the various samples. Each of the three

experiments was done twice (duplicate), and each of the vials were counted twice in the scintillation counter.

The results show that the average level of global DNA methylation in the 143B cells is 38.17%, in comparison with 2.6% methylation for the transfected 143B cells. The fibroblasts show a global DNA methylation of 6.75%, whilst the transfected cells show 89.77% less global DNA methylation. This means that the value of the global DNA methylation is negative for the transfected fibroblast cells - this result will be discussed in the next paragraph. The transfected HeLa cell line shows 17.91% global DNA methylation in comparison with 14.84% global methylation for untransfected HeLa cells. The HepG2 cells show a similar global DNA methylation patterns to the 143B cells, in the sense that the untransfected HepG2 cells have a higher degree of global DNA methylation (19.14%) than their transfected counterparts (1.81%).

A strange phenomenon in the CEA results is that the DNA methylation levels measured for the transfected fibroblasts suggest a negative value. However, in this case it should be remembered that the CEA gives only relative levels of DNA methylation and the results for the fibroblasts might simply suggest that the transfected fibroblast cells have a much lower amount of global DNA methylation than the untransformed cells. In this case, the percentage value given for the transformed fibroblasts might not be accurate, but the difference in the amount of DNA methylation observed would allow one to make a qualitative statement, i.e. that the amount of DNA methylation in the transfected fibroblasts is between 6.75% and 0.

Can this be ascribed to the way in which the transfected fibroblasts were generated, which was not chemical as was the case for the other cell lines, but by electroporesis? This question may be investigated in an extension of this study. Another possibility is that ineffective digestion by the restriction enzymes occurred in the transfected fibroblast sample. Manufacturer's guidelines from New England Biolabs for the two enzymes state that when the external cytosine of the sequence 5'-C^CGG-3' is methylated, the enzymes cannot cleave at the position indicated (New England Biolabs, 2010). This occurs irrespective of the methylation

status of the second cytosine. This external cytosine may have interfered with the correct enzyme digestion of one or either of the enzymes used in the fibroblast experiment, although a more in-depth investigation will be required to test this postulation. However, literature supports this idea. E.J. Oakeley states that, if the external cytosine of the sequence is methylated, digestion with *HpaII* may occur, but *MspI* will not digest the sample successfully (Oakeley and Chiang, 1999).

In summary of the CEA experiment, figure 4.1 shows that there are marked differences between the global DNA methylation status of the basic cell lines and those of the corresponding transfected cells. As these conclusions were derived from the global DNA methylation amounts, the next logical step would be to investigate the levels of DNA methylation on gene-level for some of the cell lines. This would give a clear indication as to the extent that DNA methylation is affected on both a global and gene-specific level with the introduction of a vector into a cell line and would serve to verify conclusions made from these CEA experiments.

These CEA results correlate with the results of another study, which shows that genome-wide hypomethylation occurs due to transformation. However, this study also found that global hypomethylation is not necessarily seen in all transformed cell lines, and some cancer cell lines show a change in DNA methylation after transformation (Wild and Flanagan, 2010).

4.3. Gene-specific methylation

Gene-specific DNA methylation studies attempt to investigate the amount of DNA methylation present at a specific gene promoter region, rather than the overall amount of methylation in the sample. In this way, determinations of the methylation density at specific positions of the genome may be made. This was done by means of real-time methylation-specific PCR (MSP).

4.3.1. Real-time MSP

Global methylation experiments indicated that there were differences in the amount of DNA methylation between transformed and untransformed samples of the same cell line. A real-time MSP analysis was done next to determine whether transformation of cultured cells have any effect on the methylation of DNA on the gene-specific level, i.e. whether the DNA methylation levels in specific gene promoters differ between basic cell line samples and transformed samples of the same cell line.

Real-time MSP investigates the DNA methylation levels in specific genes (as chosen by the researcher) and the basic technique has already been standardized in our laboratory (Van Heerden, 2006; Du Toit, 2009).

The real-time MSP assay produces reliable, reproducible results in large sample sets and is quantitative in nature. Other techniques, such as direct genomic sequencing or Southern Blot, are not used at this point due to high cost and decreased sensitivity, respectively, in comparison to real-time MSP. However, direct genomic sequencing will be used later in this study to investigate aspects of the study aims which are beyond the scope of the real-time MSP technique.

Only 143B cell line samples, containing a variety of different inserted vectors (including vectors which have knockdown effects), were investigated in the following real-time MSP experiments. This was done because transfection of 143B cells produced the greatest influence on the DNA methylation amounts in the CEA-experiments. This part of the pilot study therefore serves only as verification that global DNA methylation differences are also present on the gene-specific level. A further development of the pilot study would be to also investigate the other three cell lines (fibroblasts, HeLa and HepG2 cell lines) through the use of real-time MSP.

4.3.1. Theoretical basis of real-time MSP

The 7500HT Real-time PCR System from Applied Biosystems (for real-time PCR) differs from regular PCR, which is usually based on simple end-point analysis of amplicons. In real-time PCR reactions the point during cycling when amplification of the target reaches the exponential phase, rather than the amount of accumulated product at the end of the amplification, is of importance.

Two types of quantitative real-time PCR exist: relative and absolute. In the specific runs done for this study, relative quantification (RQ) is used. By means of threshold cycle (comparative C_T) analysis, the change in expression of a nucleic acid sequence (the target) in a test sample relative to the same sequence in a calibrator sample (usually an untreated control) is determined. Quantification is done relative to an internal methylation insensitive reference gene: *ACTB*. This provides accurate comparison between the initial amounts of template present in each sample. The results are given as an amplification plot (refer to figure 4.2).

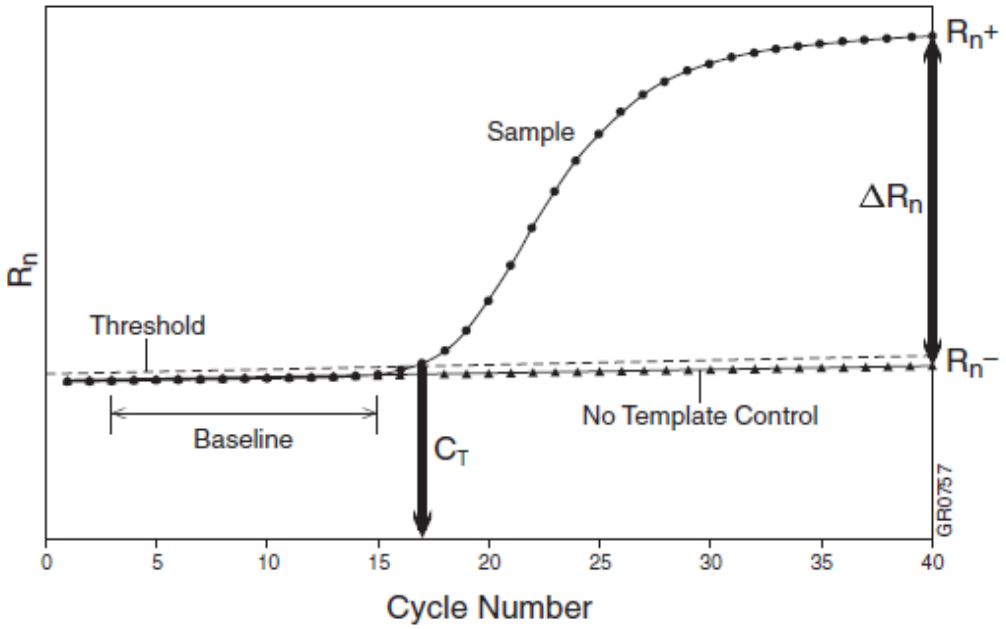


Figure 4.2: Illustration of an amplification plot. The PCR-cycles are shown on the x-axis and the logarithmic indication of reporter intensity is shown on the y-axis (taken from the Applied Biosystems instrument manual) (Anon., 2007).

Results are determined based on the amount of reporter fluorescence measured and are displayed on a graph. C_T refers to the threshold cycle, which describes the fractional cycle number where the fluorescence intersects with the threshold. Rn is the normalized reporter that shows the ratio of fluorescence emission intensity (of passive reference dye) to fluorescence emission intensity (of reporter dye). Delta Rn is the signal generated by the specific PCR conditions (thus, it is Rn with the baseline subtracted)

Real-time PCR runs can be performed in multiplex (multiple primer pairs in one reaction tube) or singleplex (a single primer pair in a single reaction tube). In each experiment, a target (the nucleic acid which is studied), calibrator (sample used as basis for comparative study) and a reference gene or endogenous control (gene present at consistent expression level in all runs) is present. The main purpose of a reference gene is to normalize the quantification of the sample DNA targets. In the experimental run, expression levels of the reference gene are subtracted from the expression levels of the sample to prevent problems which may arise due to differences in the amount of template added in the reaction. Replicate wells (two or more) are used to verify results of individual analyses by running them in duplicate or triplicate.

In this study, a variation of real-time PCR known as real-time methylation-specific PCR (real-time MSP) is used. Real-time MSP is based on sodium-bisulfite induced differences between methylated DNA and unmethylated DNA. DNA is treated with sodium-bisulfite before a real-time methylation-specific PCR is run. In this study, sodium-bisulfite treatment was necessary to determine whether there are DNA methylation differences between the basic cell lines and their transformed counterparts in a subsequent real-time methylation-specific PCR run. Methylation-specific real-time PCR is based on the sequence changes induced by the treatment of DNA with sodium-bisulfite, which converts all unmethylated cytosines to uracils, whilst methylated cytosines remain unaltered. In this way, different DNA sequences are generated for methylated and unmethylated DNA, forming a basis of differentiation between them. This can be seen in figure 4.3 on the next page.

The most important aspect of the sodium-bisulfite treatment of DNA is that complete conversion of unmethylated cytosines should occur. This is critical for the determination of the

sample's methylation levels and if inadequate conversion occurs, subsequent analysis of results may be problematic. Complete sodium-bisulfite conversion is achieved by incubating DNA in high concentrations of bisulfite at low pH and high temperature. These are harsh conditions that cause complete sodium-bisulfite conversion, but lead to a high degree of DNA fragmentation. In the experimental runs, sodium-bisulfite treated DNA is used as a template in real-time MSP runs that use primers specific for the bisulfite-treated methylated DNA sequence (the gene promoter regions of methylated genes). The Qiagen EpiTect Bisulfite kit (cat. # 59104) was used for sodium-bisulfite treatment according to the manufacturer's guidelines. As has already been noted, complete sodium-bisulfite conversion is necessary for use in quantitative experimental approaches and strict adherence to manufacturer guidelines ensures this.

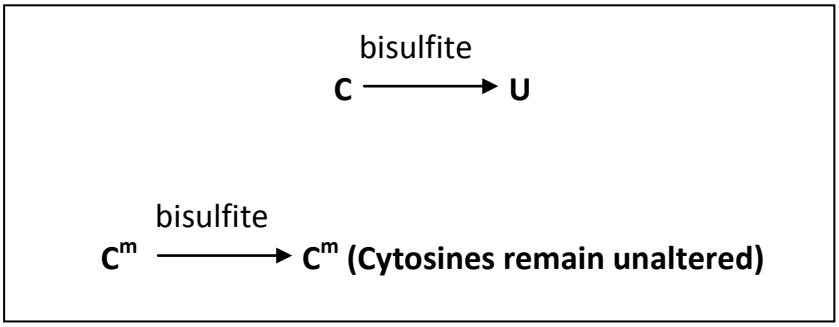


Figure 4.3: Illustration showing the differences induced by sodium-bisulfite treatment in unmethylated and methylated DNA. Unmethylated cytosines are converted to uracils, whilst methylated cytosines remain unaltered. The differences that are induced in the DNA sequences form the basis upon which differentiation between methylated and unmethylated DNA sequences can be based during MSP or Real-time MSP (adapted from Hayatsu, 2008).

Real-time MSP is a quantitative method, used for the detection of aberrant promoter methylation and provides several advantages over methylation detection by means of normal MSP. The most important difference between regular MSP and real-time MSP is that regular MSP is only qualitative in nature whilst real-time MSP analyses provide quantitative data. In regular MSP, numeric determinations cannot be made and thus have to be judged by the researcher, which severely limits the possible applications of the MSP assay. Real-time MSP provides numerical results. Values are given as percentages that are comparable to the

reference sample. Regular MSP relies on end-point analysis of amplified DNA, whilst Real-time MSP analysis occurs in real-time. This provides the opportunity to study the amplification of sample DNA in great depth. The real-time MSP is also more sensitive than other techniques such as Southern analysis and can detect very low concentrations of DNA template. Only genomic sequencing can provide a more direct analysis, which is the next step in this study.

TaqMan® chemistry is the selected detection chemistry for real-time MSP in this study. The fluorescence of samples are registered by the 7500 HT Real-time PCR System from Applied Biosystems and results displayed on a graph. Primers and probes, manufactured by Applied Biosystems and designed according to the specifications of published papers (refer to table 4.2), were used in the samples. These sequences, directed at the gene promoter regions of the specified genes, are shown in table 4.2 below.

Table 4.2: Sequences of primers and probes used in real-time MSP experiments

Gene	Forward Primer	Reverse Primer	TaqMan ® Probe	Source
MGMT	5'-CGA ATA TAC TAA AAC AAC CCG CG-3'	5'-GTA TTT TTT CGG GAG CGA GGC-3'	6FAM5'-AAT CCT CGC GAT ACG CAC CGT TTA CG-3'MGBNFQ	Brabender and Usadel, 2003
P16^{INK4a}	5'-TAG CGG GCG GCG GGG GA-3'	5'-CGC ACC TCC TCT ACC CGA-3'	6FAM5'-ATG GAG TCG GCG GCG G-3'MGBNFQ	Roth, Abnet *et al*., 2006

4.3.3. Overview of real-time MSP analysis

The conditions of the specific run differ depending on the type of run, e.g. whether the experimental run is singleplex or multiplex, number of samples used, etc. The universal thermal cycling conditions used in the real-time MSP experimental runs were those suggested by the TaqMan manufacturers (Applied Biosystems, cat. # 4304437). These thermal cycling conditions are as follows: 95°C for 10 minutes, then 95°C for 15 seconds and 60°C for 1 minute. The last two steps were repeated for 50 cycles as per manufacturer's guidelines for standard master mix (MM) conditions.

4.3.4. Results of multiplex real-time MSP of the *MGMT*-gene promoter region

A multiplex real-time MSP was done to determine the specific methylation levels of the 143B samples' *MGMT*-gene promoter region. This was the first of two real-time MSP runs to determine whether there were DNA methylation differences between cell lines and their transformed counterparts, specifically in the *MGMT*- and *p16*^{INK4a}-genes. The control gene was *ACTB*, which is assumed to have a constant state of DNA methylation (due to the absence or low numbers of CG-dinucleotides). The manufacturer's guidelines for standard master mix (MM) conditions were used (refer to section 4.3.3). Sodium-bisulfite treatment was done in preparation of the real-time MSP run according to the manufacturer's guidelines (Qiagen).

Figure 4.4 shows the raw data of the real-time MSP analysis, *i.e.* the rate of amplificon generated with *MGMT*-gene promoter region directed primers:

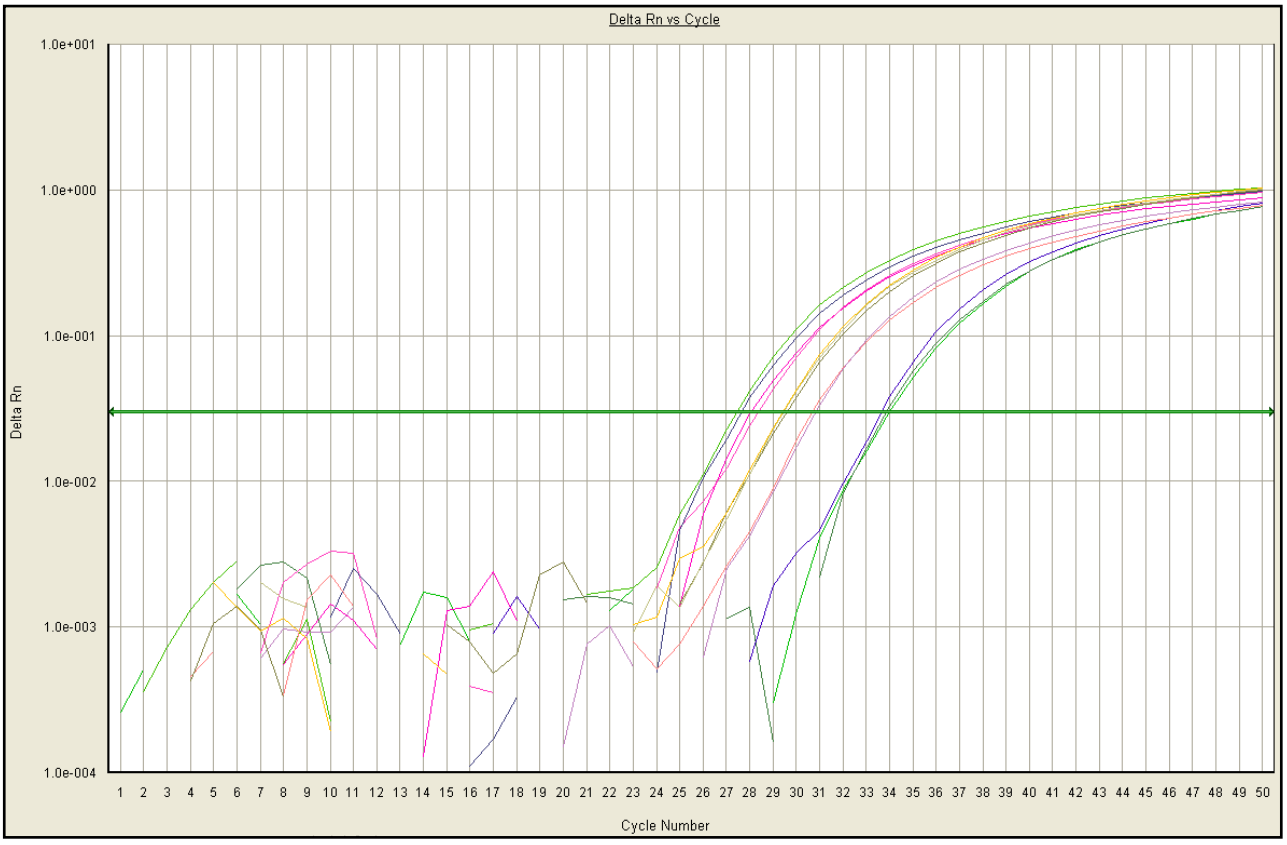


Figure 4.4: Raw data output of a real-time MSP experiment in the *MGMT*-gene promoter region of the 143B cell line and the 143B cell line with inserted vectors. The lines are indicative of the amplification of this region.

Figure 4.4 shows the amplification curve generated in a real-time MSP experiment. The horizontal green line represents the C_T value (refer to figure 4.2), which intersects the amplification curve at the point of linear product amplification. The threshold is set at 0.0200. The lines represent the real-time PCR amplification. The fact that the lines are grouped close together shows that amplification of duplicates was similar, indicating accurate measurement of the amplification of the *MGMT*-gene promoter region.

Figure 4.5 gives the gene-expression results for several inserted vectors.

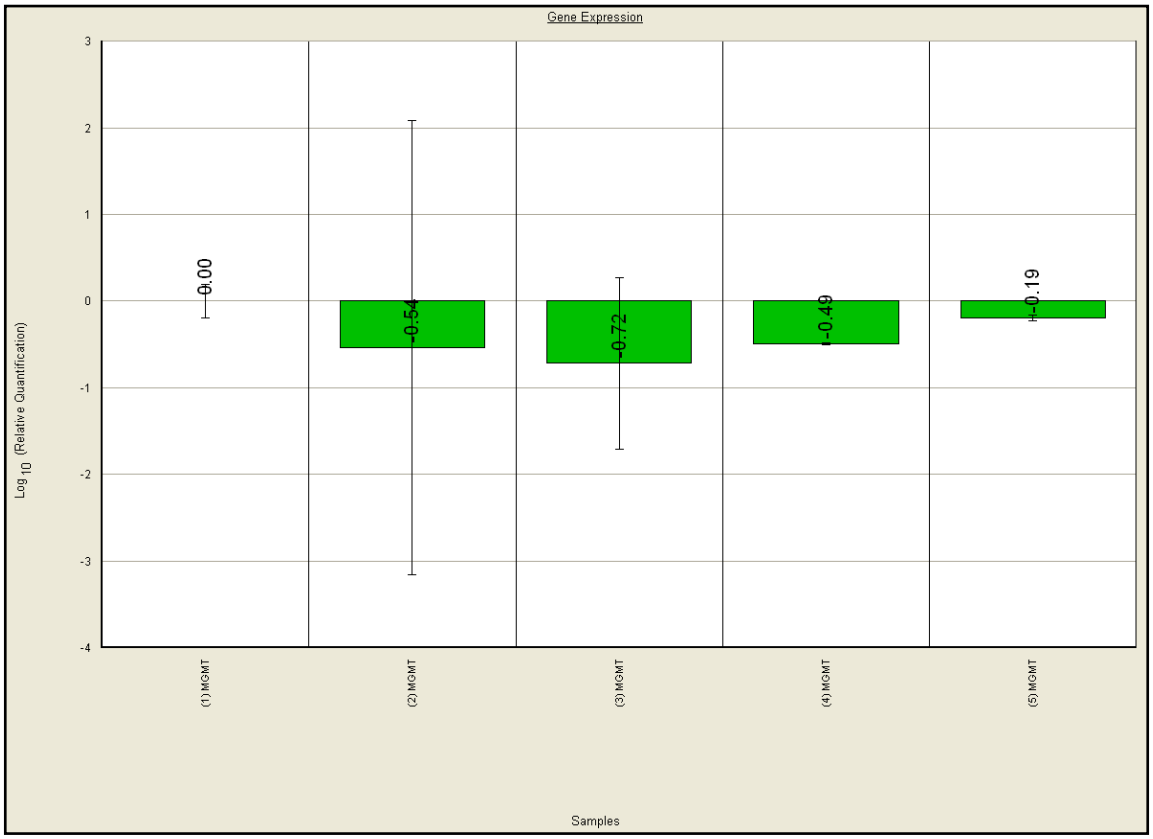


Figure 4.5: Gene expression results of a real-time MSP experiment using untransformed 143B cells and 143B cells with several inserted vectors. The first sample is the untransformed 143B cells. The second sample has an "empty" vector inserted, the third sample has an unrelated GFP knockdown inserted, the fourth sample has a vector inserted that causes a knockdown of complex I of the mitochondrial electron-transfer chain (subunit targeted by RNAi: S.U. NDUFS (Fe-S protein3)) and the fifth sample has a vector that causes the knockdown of complex 3 of the mitochondrial electron-transfer chain (subunit targeted by RNAi: S.U. UQCRFS (Rieske Iron-Sulphur protein)).

Figure 4.5 shows that gene expression in the 143B cells was altered with the insertion of an expression vector, due to the fact that variation is seen between the same samples. The only difference in each was the insertion of a vector. Figure 4.6 gives a graphical representation of the same results, but in terms of expected *MGMT*-gene promoter DNA methylation levels. Data analysis for this figure was done with Microsoft Excel. For this analysis, the DNA methylation level of the control cell DNA is set as a baseline (100% gene expression = 0% DNA methylation). This is illustrated in figure 4.6 below:

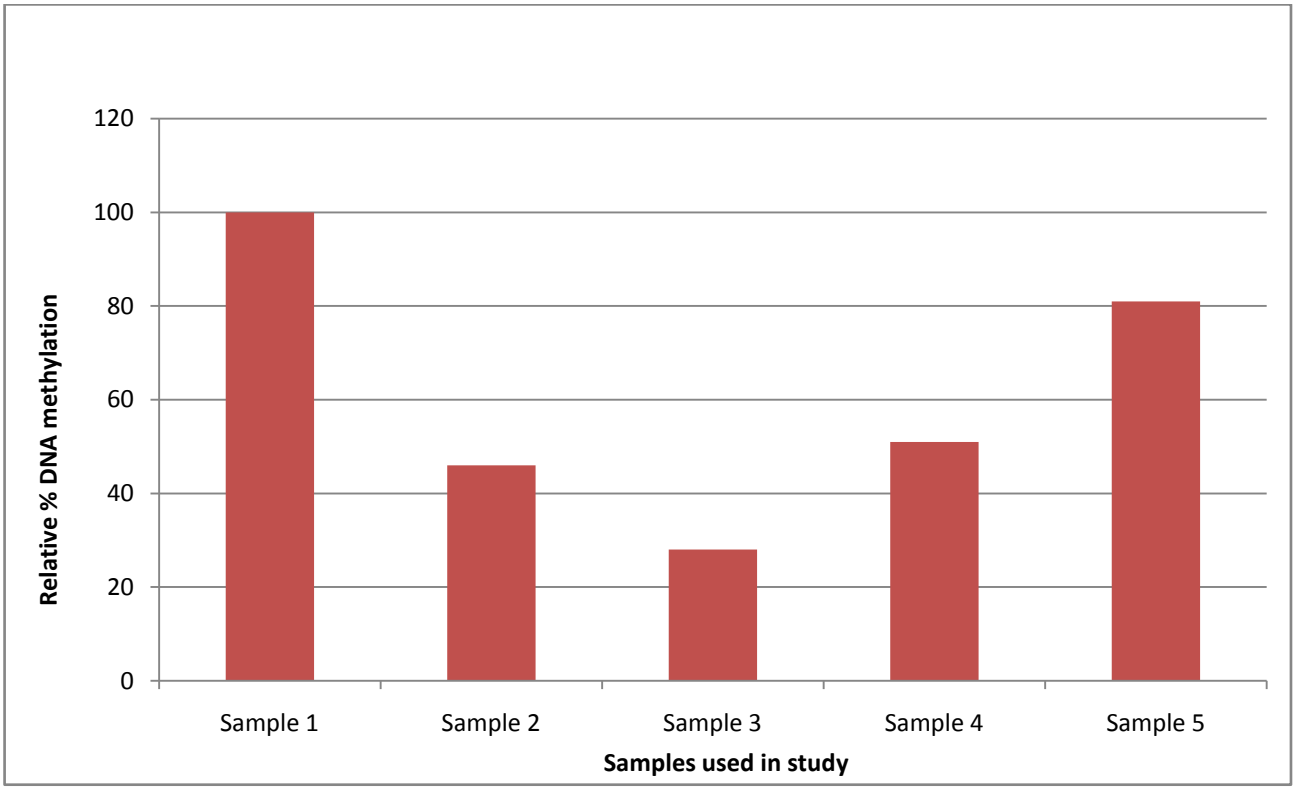


Figure 4.6: Amount of DNA methylation in *MGMT*-gene promoter of various samples of the 143B cell line - in an untransformed 143B sample and in 143B cells with inserted vectors. The first sample is the untransformed 143B cells. The second sample has an "empty" vector inserted, the third sample has an unrelated GFP knockdown inserted, the fourth sample has a vector inserted that causes a knockdown of complex I of the mitochondrial electron-transfer chain (subunit targeted by RNAi: S.U. NDUFS (Fe-S protein3)) and the fifth sample has a vector that causes the knockdown of complex 3 of the mitochondrial electron-transfer chain (subunit targeted by RNAi: S.U. UQCRFS (Rieske Iron-Sulphur protein)).

When the DNA methylation level of *MGMT* of the untransformed cells are taken as a baseline (100% DNA methylation = 0% gene expression), the vector-inserted cells showed less DNA

methylation than the untransfected cells. The second sample shows 46% DNA methylation, the third sample 28% DNA methylation, the fourth a level of 51% and the fifth sample shows a DNA methylation level of 81% for this gene. Figure 4.6 therefore shows that a state of hypomethylation exists in the vector-inserted cells in comparison with that of the control cells (for the *MGMT*-gene specifically).

Figure 4.7 shows a comparison of the *MGMT* gene-expression results of a similar run done by a previous researcher in our laboratory (Kok, 2009), with the results obtained in this study. These results seem to indicate a similar tendency in the same samples for both this study's results and the results generated by another researcher.

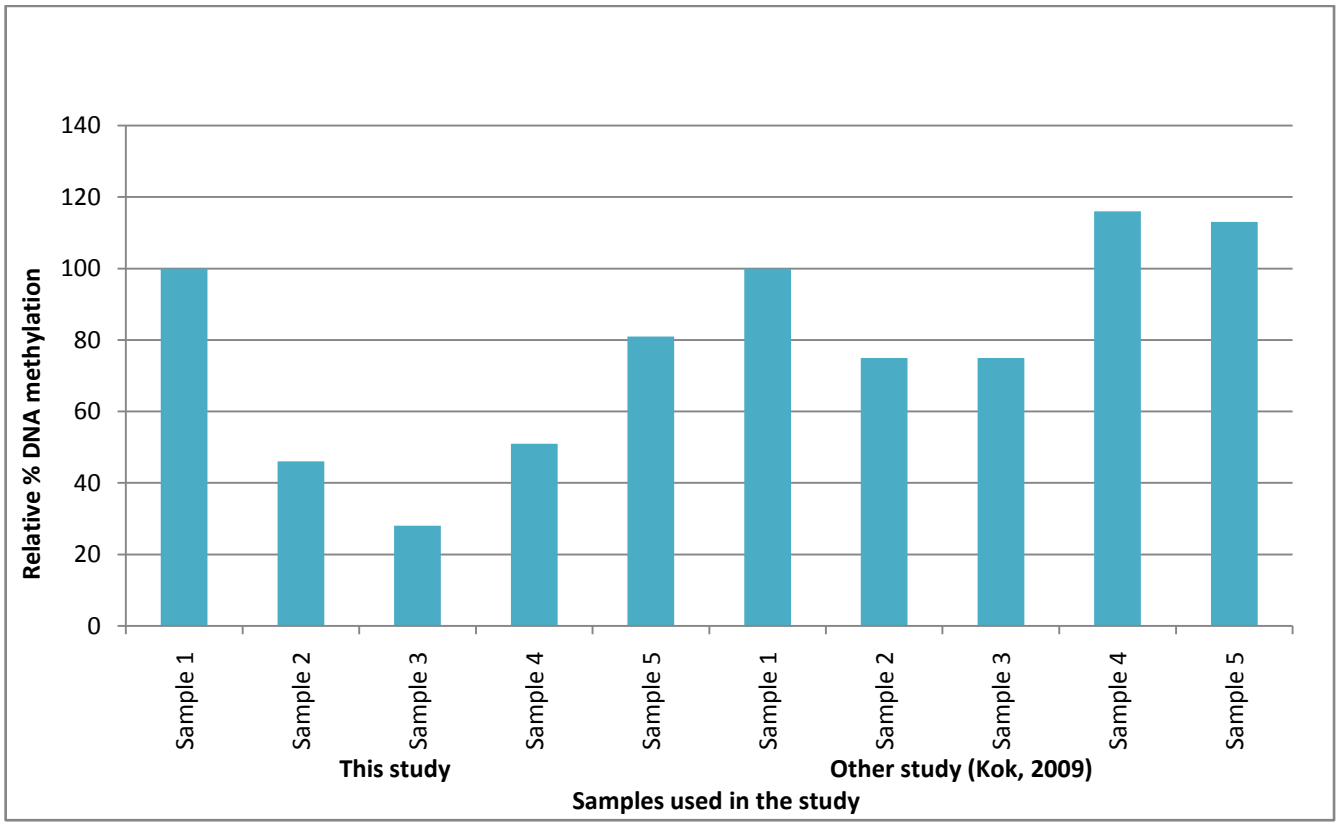


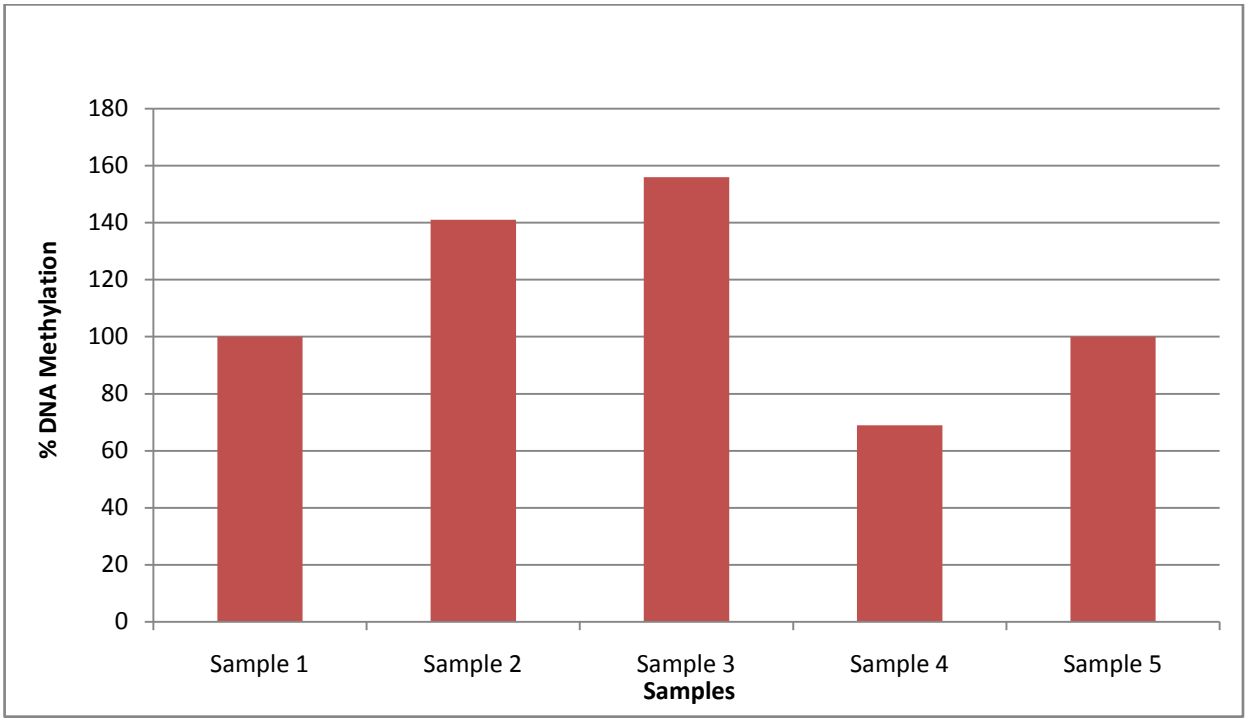
Figure 4.7: Comparison of two real-time MSP experiments, one showing the amount of DNA methylation expected in the MGMT-gene promoter region of an untransformed 143B sample and in 143B cells with inserted vectors as done in this study and the other showing the results of the same experiment as done in another study in our laboratory (Kok, 2009).

Figure 4.7 shows that a similar trend may be seen in both experiments (both experiments showed a state of hypomethylation in the transformed cells, except for sample 4 and 5 of the

other study). If the averages of sample 4 and 5 are calculated for both experiments, the values show a decrease in DNA methylation amounts for both samples. These results seem to indicate that the insertion of a vector into a cell line causes a decrease in the amount of DNA methylation, at least for the *MGMT*-gene in the 143B cell line and its transformed counterparts. A next step would be to test the methylation amounts of the same samples in a different gene.

4.3.5. Results of multiplex real-time MSP of the *p16^{INK4a}*-gene promoter region

The amount of DNA methylation in the *MGMT*-gene promoter of various samples was determined in the previous experiment and, because variation was seen in this region, it was deemed necessary to measure the amount of DNA methylation in another gene. In the next experiment, the amount of DNA methylation in the *p16^{INK4a}*-gene promoter was determined. By doing this, the amount of DNA methylation variation in the *p16^{INK4a}*-gene promoter with the insertion of a vector could be accurately determined. This multiplex real-time MSP run serves as further proof of principle for the study (on a gene-specific level). The *ACTB* gene was used as control. Sodium-bisulfite treated DNA was generated according to manufacturer's guidelines. Raw real-time amplification curve data and gene expression data is not shown for this experiment, but the gene-expression results are shown in figure 4.8. The C_T-line is not shown, but the threshold was set to 0.0200. Although the amplification plot is not shown, this value is indicated because it is important to ensure that the C_T-values are the same when different real-time MSP experiments are compared.



The first sample is the untransformed 143B cells. The second sample has an "empty" vector inserted, the third sample has an unrelated GFP knockdown inserted, the fourth sample has a vector inserted that causes a knockdown of complex I of the mitochondrial electron-transfer chain (subunit targeted by RNAi: S.U. NDUFS (Fe-S protein3)) and the fifth sample has a vector that causes the knockdown of complex 3 of the mitochondrial electron-transfer chain (subunit targeted by RNAi: S.U. UQCRFS (Rieske Iron-Sulphur protein)).

Figure 4.8 gives a graphical representation of the gene-expression results for 143B cells, as well as 143B cells with several inserted vectors, in the $p16^{INK4a}$-gene promoter region. This analysis is done to investigate the DNA methylation levels of this promoter region. The first sample is the untransformed 143B cells. The second sample has an "empty" vector inserted, the third sample has an unrelated, GFP knockdown inserted, the fourth sample has a vector inserted that causes a knockdown of complex I of the mitochondrial electron-transfer chain (refer to section 4.3.4 for more details) and the fifth sample has a vector that causes the knockdown of complex 3 of the mitochondrial electron-transfer chain (refer to section 4.3.4 for more details). From figure 4.8, it was apparent that the gene expression in the transformed cells was altered with the insertion of a vector. The alterations in the amount of DNA methylation in each sample do not show the same trend as seen in the *MGMT*-gene promoter region, however, but there are clear perturbations of the DNA methylation amounts following transfection.

Altered DNA methylation of the *p16*^{INK4a}-gene promoter was observed following insertion of the different vectors into the 143B cells. The second sample, when compared to the first sample's baseline of 100% DNA methylation, shows 41% more DNA methylation, the third sample shows 56% higher levels of DNA methylation, the fourth a level of 69% and the fifth sample shows a DNA methylation level comparable to that of the first sample. This seems to indicate that the insertion of a vector into a cell has implications for the DNA methylation levels of the *p16*^{INK4a}-gene promoter region in the 143B cells.

Literature states that if DNA methylation is increased in the promoter regions of a gene, the effect will normally be reduced expression of that gene and possibly silencing of the gene (Das and Singal, 2004). This is relevant for the *p16*^{INK4a}-gene, where an increase in DNA methylation levels has been noted in the real-time MSP experiment. This provides an interesting avenue for further research. These results, coupled with those of the previous experiment, indicate that transfection does have an effect on gene-specific DNA methylation levels. The effect appears to vary from gene to gene, and appears to be dependent on the type of inserted vector. It may be possible to correlate the effect of vector-insertion on DNA methylation levels with cellular characteristics, although further research will be needed to investigate this avenue of study.

4.4. Chapter Summary

Both the CEA and the real-time MSP showed that there were differences in the amount of DNA methylation on both a global- and gene-specific level. In some cases, the DNA methylation levels were seen to increase, whilst in other cases there was a decrease in the amount of DNA methylation. This seems to support the original premise that transfection of cell cultures has an effect on the DNA methylation patterns and correlates with results generated by other researchers.

The CEA showed that there was a decrease in the amount of methylation in the 143B and HepG2 cells. The HeLa cells showed an increase in the amount of DNA methylation in

transfected cells. Fibroblasts show a large decrease in DNA methylation levels upon transfection in comparison to untransfected fibroblasts. The results of the fibroblast samples, however, negate a quantitative conclusion as to the specific amounts of DNA methylation. From the CEA results, one may conclude that variation of DNA methylation levels are seen in different cell lines upon transfection. This correlates with the results of another study, which showed that changes in DNA methylation levels occur upon cellular transfection (Wild and Flanagan, 2010).

In the real-time MSP experiments, variation in the amount of DNA methylation of a specific gene is seen. These variations depend on the gene in question, as well as the type of inserted vector. Some genes show an increase in DNA methylation, whilst others show a decrease. This means that the effect of transfection is not the same throughout the genome, and varies from gene to gene.

Both of the methods of analysis used in this pilot study are based on indirect analysis of DNA methylation and cannot provide precise information concerning the positions of methylated CG-dinucleotides or the DNA sequences wherein methylated CG-dinucleotides are located. The next step was to analyze the samples by using DNA sequencing.

Chapter 5

Selection and partial characterisation of methylated DNA isolated from different cell lines by means of high-throughput sequencing

5.1. Introduction

In the previous chapter, different techniques affirmed that there is a difference in the levels of DNA methylation between cell lines and their transfected counterparts after transfection. Both the CEA and the real-time MSP are indirect measurements of the amount of DNA methylation and do not provide information at the DNA sequence level. A more direct analysis of could be done by means of high-throughput DNA sequencing.

Before direct DNA sequencing was done, however, it was necessary to decide which selection method was best suited for the isolation of methylated DNA fragments from the complete cellular DNA. The technique chosen for this study was based on the fragmentation of DNA to generate short fragments, followed by enrichment of methylated fragments using a commercially available kit.

5.2. Sample preparation

Enrichment of methylated DNA fragments using a kit formed the basis of the sample preparation steps. The DNA was first fragmented using restriction enzyme digestion, then selection and separation of methylated fragments from unmethylated DNA was done by using the MethylMiner Kit from Invitrogen (cat. # ME10025), and finally, the samples were sequenced using high-throughput GS FLX Titanium sequencing technology from 454 Life Sciences and bought by Roche. The service provider was Inqaba Biotec.

5.2.1. DNA isolation and enzyme digestion

Eight samples were prepared for MethylMiner enrichment; one for the 143B cell line, one for fibroblasts, one for the HeLa cell line and another for the HepG2 cells, as well as samples for the transfected counterparts of each of the cell lines. DNA was isolated from each of the cell lines by means of the Qiagen Flexigene DNA Kit (#51204) using the manufacturer's recommended protocol.

Before the enrichment step the DNA was fragmented using restriction enzyme digestion. Double digestion of DNA with *AluI* and *BSuRI* was used to fragment the DNA. This was done to fragment the DNA at known locations (the restriction site of *AluI* is 5'-AG^CT-3' and the restriction site of *BSuRI* is 5'-GG^CC-3') into fragments of sizes of between 50 bp and 600 bp. This is of suitable size for enrichment via the MethylMiner kit, which can only enrich fragmented DNA of a specific size. The enzymes fragment the DNA sequence at recognition sites found at repeated positions in the genome.

Initially, a test run was done to determine optimal conditions for restriction enzyme digestion. Buffer selection for the restriction digestion was done via the Fermentas website (URL: *www.fermentas.com/doubledigest*) to determine the recommended ratio of enzymes to be used in the double digestion (AluI:BSuRI in ratio 1:2) and to select the appropriate buffer (Tango Buffer). The DNA was electrophoresed on a 1.4% agarose gel at 20V for 5 min and then 50V until adequate separation occurred. The voltage was changed to separate the fragments more clearly. This was done to evaluate the process of DNA digestion. The gel is shown in figure 5.1 below:

Figure 5.1: Gel photo showing enzyme digested DNA. Lane 1 shows the 50 bp ladder, lane 2 shows the clump of undigested DNA and the third lane shows the streak of fragments generated by means of enzyme digestion with AluI and BSuRI (left picture is inverted).

It was decided that a digestion of 3 hours was adequate for complete enzyme digestion of DNA. At this time, a clean-up step was also included - using sodium acetate and ethanol, but due to a large loss of sample DNA during the cleanup step it was decided that use of a cleanup step was undesirable. Subsequent experiments would therefore not make use of the clean-up steps in order to preserve as much of the limited sample as possible.

AluI and *BSuRI* were used to digest the DNA from the 143B cell sample at 37°C for 3 hours and then for 20 minutes at 80°C (to denature enzymes and prevent further enzymatic activity) using the method described in the preceding paragraph. This generated a range of fragments with sizes ranging between 50bp and 600bp (refer to figure 5.1), which could be successfully enriched using the MethylMiner Kit. This indicates that *AluI* and *BSuRI* are suitable enzymes for DNA digestion for downstream use with the MethylMiner kit.

A next step was to upscale the reaction to generate enough DNA for use with the MethylMiner Kit. The MethylMiner kit requires a minimum amount of input DNA of 10-25ug. This test also verified that DNA digestion could be done without the use of a clean-up step while still preserving a high level of purity.

Once these steps were optimized, preliminary work in preparation of enzyme digestion was complete. DNA from 143B cells, fibroblasts, HeLa cells and HepG2 cells were isolated and enzyme digested with *AluI* and *BsuRI*.

5.2.2. Enrichment of methylated DNA

Enrichment of methylated fragments was done using the MethylMiner kit, according to the manufacturer's guidelines. This process led to the separation of the methylated fragments of DNA from the unmethylated fragments of DNA. This process was completed successfully to generate eight methylated samples of each of the basic cell lines and their transformed counterparts.

5.2.2.1. MethylMiner methylated DNA enrichment

The MethylMiner kit is used to enrich fractioned, double-stranded DNA based on the degree of DNA methylation. Separation is done via binding of the methyl-CpG binding domain of human MBD2 protein to methylated CG-dinucleotides in the sample DNA. The MBD2 protein is coupled to magnetic Dynabeads via a biotin linker and a magnetic rack is then used to separate the beads from DNA in solution.

This technique provides increased sensitivity in comparison to methyl-DNA immunoprecipitation, or MeDIP, with 5x more hits and 16x more unique hits observed for the MethylMiner kit from Invitrogen (Jia *et al*., 2010, Prediger (ed.), 2010, Yu *et al*., 2010). The MethylMiner kit is also suited for downstream high-throughput DNA sequencing, as the double-stranded nature of DNA is preserved. This allows direct ligation of double-stranded adaptors for tagging during DNA sequencing applications. Furthermore, the protocol and methods used in the kit provide consistent and repeatable results.

For this step, the MethylMiner™ Methylated DNA Enrichment Kit from Invitrogen (cat # ME10025) and a magnetic rack was used. The methylated fragments can be separated based on degree of DNA methylation, but in this study only a single enriched population was eluted using a high salt buffer. This means that the methylated fragments of the DNA were separated from the unmethylated fragments, regardless of the amount of methylation present on the methylated DNA fragments.

The protocol for enrichment of methylated DNA may be divided into several separate steps. During the first step, the beads were prepared by coupling the MBD-Biotin Protein to magnetic Dynabeads®. The beads were first removed from the storage solution, washed in a wash buffer, and resuspended in a suspension liquid provided with the kit in preparation for experimental use. This prepared the beads for MBD-Biotin Protein coupling before a second washing step. At this point of the protocol, the beads were ready for methylated DNA capture

Validation of the MethylMiner Kit was done using control DNA (methylated DNA and unmethylated DNA) included in the kit, which were added to K-562 DNA. After MethylMiner enrichment, two PCR amplification runs with specific primer sets were done. Validation of the separation of methylated fragments from unmethylated fragments is investigated using these two primer sets. When the methylation-specific primers amplify the methylated fraction of enriched DNA, whilst non-methylation-specific primers amplify the unmethylated fraction, successful separation of the two fractions has occurred.

After each cell line was enriched, validation of adequate separation of methylated DNA from unmethylated DNA was done (appendix A). All the validation studies showed that adequate separation of methylated fragments from unmethylated fragments had occurred for each of the cell line samples.

The control capture reaction was performed first, followed by the sample capture reaction. This involved mixing fragmented DNA (sample and controls) in different reactions with the MBD-biotin protein coupled beads. The methylated DNA fragments were bound with MBD during an incubation step, while the unmethylated fraction was removed in a series of washing steps. Ethanol precipitation was used to remove the methylated fraction from the beads and the DNA was stored at -20°C until further use.

5.2.3. Ligation to PEZseq vector and PCR amplification

Measurements of the amount of DNA generated after the MethylMiner enrichment step showed that the DNA amounts were low (appendix B). This could be attributed to the DNA methylation amounts in a particular cell line; some cell lines will have less DNA methylation amounts than others, and therefore less DNA would be isolated from these cell lines. However, too little sample DNA for high-throughput sequencing using the GS FLX titanium system was isolated. This sequencing technology required 125ng of sample DNA per sample and more DNA was therefore needed before sequencing could begin. Due to the fact that the MethylMiner kit is very expensive and only provides materials to apply to a very limited number of samples, another strategy was needed to generate

more samples for DNA sequencing. Therefore, ligation of the blunt, methylated fragments to a vector and PCR amplification of the samples were done to obtain sufficient amounts of DNA for sequencing. Ligation onto an appropriate vector provides a site for primers to amplify from, as no primers could be designed specifically for the amplification of these fragments as they were derived from various unknown positions in the genome. A pEZseq vector and Clonesmart DNA ligase was used for these steps (part of the pEZSeq Blue/White Cloning Kit from Lucigen, cat. # 89002-518). The pEZseq vector is shown in appendix C.

The thermal reaction conditions for the ligation steps were as follows: Room temperature for 2 hours, 70°C for 15 minutes, room temperature for 15 seconds and 4°C for 15 seconds. A test run was done using fragmented λ *HindIII* DNA. In this first test run, the ligation step was done without PCR amplification. Enough DNA was ligated to the vector to allow direct visualisation on a gel. Several factors were altered to find the optimal conditions for ligation of the DNA sample fragments to a vector. In figure 5.2 below, 450 ng of each ligated sample was loaded onto the 1.5 % agarose gel and electrophoresed.

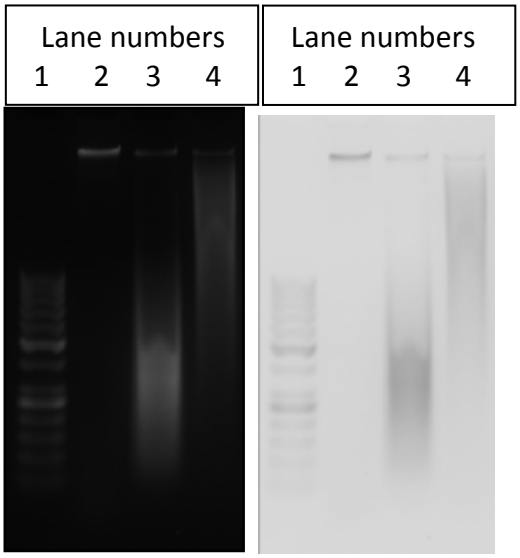


Figure 5.2: Gel photo showing methylated DNA fragments ligated to a vector. The first lane (left lane) shows the 50 bp ladder, the second lane shows the unligated DNA, the third shows the *AluI*- and *BSuRI*-digested DNA and the fourth shows the successful ligation of the fragments to the pEZseq vector (left picture is inverted).

Figure 5.2 shows a gel photo of methylated DNA fragments ligated to the pEZseq vector using Clonesmart DNA ligase. From the gel photo it may be concluded that the fragments were correctly ligated to the pEZseq vector (this is seen as a shift in the density of the fragment smear). A fragment smear is expected due to the variation in fragment sizes between 50 bp and 600 bp, but the increase in the size of the fragments (which are seen to move more slowly through the gel compared to the *AluI* and *BSuRI* digested fragments) shows that ligation to the vector has occurred correctly. It should be noted that several optimization steps and different vectors were used before it was decided that the pEZseq vector was most suitable for use with the MethylMiner-enriched fragments.

A second test run was done to determine whether the PCR amplification of the samples could be done successfully after ligation of methylated fragments to a vector had occurred. Ligation was done as described in the previous paragraph. A series of optimization steps were done before it was decided that Kapa high-fidelity Taq (cat. # KK2101) and M13-forward and –reverse primers (part of the pEZSeq Blue/White Cloning Kit from Lucigen, cat. # 89002-518) were to be used for further PCR amplification runs. The sequences of the M13 primers are shown in appendix C. A gel photo showing the amplified fragments after PCR amplification using these reagents is shown in figure 5.3.

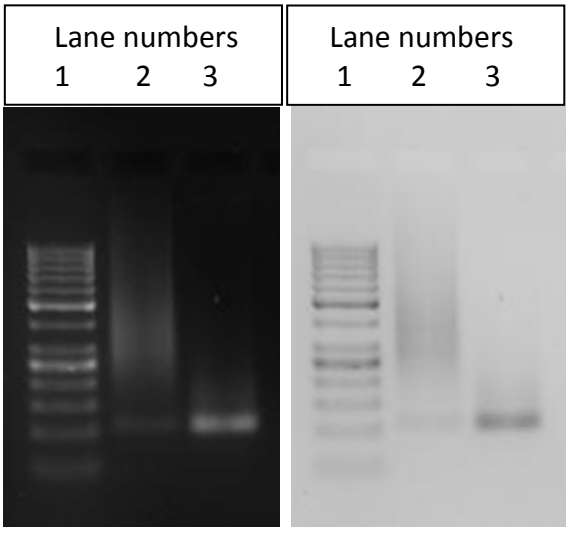


Figure 5.3: Gel photo of amplicons generated via PCR amplification. Kapa high-fidelity Taq was used during amplification and a 2% agarose gel was used. The first lane (left lane) shows a 50bp ladder, and the other two lanes show the amplicon of both samples. The pEZseq vector did not amplify and only the fragments of the samples are visible (left picture is inverted).

The results in figure 5.3 show that amplification of the sample fragments occurred correctly. The thermal cycling parameters for the PCR amplification step were as follows:

Table 5.1: Thermal cycling conditions for PCR amplification of amplicon

4°C	HOLD (prepare on ice)	
94°C	3 minutes	
94°C	30 seconds	x 30 cycles
55°C	30 seconds	
72°C	1 minute	
72°C	1 minute	
4°C	HOLD	

At this juncture it is apt to note that, according to literature, DNA methylation information is erased by some molecular biology techniques, such as PCR amplification (Laird 2010). At first glance this may be seen as problematic for this study, which focuses on DNA methylation. However, the methylated fragments have already been selected by means of the MethylMiner Kit and removed from the unmethylated fraction. As the DNA sequencing will only provide sequence information, the fact that epigenetic marks may have been removed does not pose a problem for further investigation of the methylated sequences, as it is already known that the methylation mark resides in CG-dinucleotides of the sequenced fragments (Das and Singal, 2004; Laird, 2010).

It should be remembered that the ligation and PCR amplification steps were all done to generate a greater sample volume, and at this point the aim was to generate more than 125ng DNA per sample (which was required for GS FLX Titanium sequencing). The ligation of fragments to the Lucigen vector using Clonesmart DNA ligase and PCR amplification of the ligated fragments using Kapa high-fidelity Taq and M13-forward and –reverse primers was completed successfully for all eight samples. Enough sample DNA was generated for sequencing.

5.3. Processing of sequencing results

DNA sequencing was done using the GS FLX Titanium System from Roche / 454 Life Sciences. The system is based on sequencing-by-synthesis technology with an average read length of 400 bases, which falls in the midrange of the fragment lengths prepared for this study. A 10MB lane was used, with eight ¼ library reactions.

Data analysis was done using a range of basic and more sophisticated data analysis tools, mostly Microsoft Excel and CLC Bio Genomics Workbench. CLC Bio is a bioinformatics program which is particularly suited for analysis of the large data sets generated by high-throughput DNA sequencing. In this section of the dissertation, the data-analysis steps will be described and results discussed. Microsoft Excel was used for simpler data analysis requirements, e.g. data sorting.

A note on terminology: A (C) after the name of the cell line indicates that the label refers to the untransfected cell lines, while a (V) after the name refers to the transfected counterparts of the cell line.

5.3.1. Reads and coverage of data

The single read data distribution per sample is as follows (as was supplied by Inqaba Biotech), which gives an indication of DNA sequencing efficiency:

143B (C) sample – 9054 reads
143B (V) sample - 8935 reads
Fibroblasts (C) sample - 287 reads
Fibroblasts (V) sample - 1979 reads
HepG2 (C) sample - 5831 reads
HepG2 (V) sample - 7800 reads
HeLa (C) sample - 10569 reads
HeLa (V) sample - 8136 reads

The first step in high-throughput sequencing data analysis is to use *de novo* assembly to generate contiguous sequences (contigs) of the fragmented DNA. This was done in CLC Bio. Contiguous sequences were created from the large number of short fragment reads generated during sequencing and were used to verify the accuracy of sequencing. By matching similar fragments sequenced from one of many copies of the same fragment (due to PCR amplification or repeats in original DNA sample), one can be certain that the correct DNA sequence for each fragment is assembled by the software. A minimum contig length of 30bp is used; any contig shorter than this would not be useful for further analysis. After preliminary data analysis, the following numbers of contiguous sequences were generated for each of the samples (the list that follows shows the contig count for each sample):

143B (C) sample – 216 contigs (Mean contig length= 212; Total contig length in bases= 45860)
143B (V) sample – 143 contigs (Mean contig length= 204; Total contig Length= 29312)
Fibroblasts (C) sample – 10 contigs (Mean contig length= 185; Total contig Length= 1855)
Fibroblasts (V) sample – 35 contigs (Mean contig length= 203; Total contig Length= 7114)
HepG2 (C) sample – 207 contigs (Mean contig length= 206; Total contig Length= 42767)
HepG2 (V) sample – 143 contigs (Mean contig length= 217; Total contig Length= 31100)
HeLa (C) sample – 16 contigs (Mean contig length= 187; Total contig Length= 3039)
HeLa (V) sample – 213 contigs (Mean contig length= 200; Total contig Length= 46201)

The above data shows the number of contigs generated per sample. A contig is generated using several overlapping fragments to generate a single fragment sequence. In this study, the consensus sequence of each contig is taken to represent a single methylated fragment. This fragment will later be matched to specific positions in the genome. The mean contig length refers to the average length of all assembled contiguous sequences and the total contig length refers to the total length of all contigs (i.e. if assembled end to end).

The single read data distribution of sample contigs are as follows (analysis by means of CLC Bio), which gives an indication of DNA sequencing efficiency:

143B (C) sample – 3207 reads (Mean read length = 168.69; Total read length = 540991)
143B (V) sample - 3961 reads (Mean read length = 142.76; Total read length = 565457)
Fibroblasts (C) sample - 70 reads (Mean read length = 184.96; Total read length = 12947)
Fibroblasts (V) sample - 616 reads (Mean read length = 138.21; Total read length = 85136)
HepG2 (C) sample - 4174 reads (Mean read length = 160.42; Total read length = 669601)
HepG2 (V) sample - 6238 reads (Mean read length = 135.42; Total read length =844731)
HeLa (C) sample - 8640 reads (Mean read length = 124.23; Total read length = 1073339)
HeLa (V) sample - 5744 reads (Mean read length = 148.28; Total read length = 851717)

The amount of reads refers to the total read count, or the total number of actual reads generated for the specific sample after *de novo* assembly has been done. Mean read length refers to the average read length per sample and total read length indicates the sum of the length of all the reads.

Because each of these contigs consists of at least two or more overlapping sequences, only contiguous sequences were used for further data analysis and, indeed, the creation of these contigs forms a central part of data preparation for high-throughput DNA sequencing data analysis. The accuracy of the DNA sequencing is given by means of the coverage, which indicates the number of overlapping of fragments. The average, minimum and maximum coverage of the samples is given in the section below.

5.3.1.1. Coverage data of the 143B sample

Table 5.2 shows the minimum, maximum and average coverage of the 143B cell line samples. The coverage is an indication of the fragment overlap when contigs are assembled using *de novo* methods. This gives an indication of how well the DNA sequencing was done, with very low coverage indicating problems with the sample sequencing. Of course, the PCR steps done before sending the samples for sequencing may also have had an influence on DNA sequencing efficiency.

Table 5.2: Coverage data of the 143B cell line

Samples	143B (C)	143B (V)
Average Coverage	8.48	11.12
Minimum Coverage	1.53	1.51
Maximum Coverage	276.25	675.37

For the untransfected 143B sample, 45490 positions (bases) have coverage of between 1 and 72 and 370 positions (bases) have coverage of above 72. For the transfected sample, 29210 positions (bases) have coverage of between 1 and 92 and 102 positions (bases) have coverage of above 72. This shows that the average coverage is suitable for further data analysis steps, as averages lower than 5 would show that the sequence data might not be accurate.

5.3.1.2. Coverage data of the fibroblast samples

Table 5.3 below shows the minimum, maximum and average coverage of the fibroblast samples.

Table 5.3: Coverage data of the fibroblasts

Samples	Fibroblasts (C)	Fibroblasts (V)
Average Coverage	5.04	8.27
Minimum Coverage	2.00	1.42
Maximum Coverage	16.45	136.72

For the fibroblast cell line, 1785 positions (bases) have coverage of between 1 and 22 and 70 positions (bases) have coverage of above 22. The transformed fibroblasts have 7023 positions (bases) with coverage of between 1 and 108 and 91 positions (bases) with coverage of above 108.

The coverage data shows that sequencing of the transfected fibroblast sample was adequate, but that the untransfected sample showed low coverage. This was kept in mind throughout further data analysis steps, as low coverage might indicate ineffective sequencing of the specific sample or problems in sample preparation beforehand.

5.3.1.3. Coverage data of the HepG2 samples

Table 5.4 below shows the minimum, maximum and average coverage of the HepG2 cell line samples.

Table 5.4: Coverage data of the HepG2 samples

Samples	HepG2 (C)	HepG2 (V)
Average Coverage	10.02	14.36
Minimum Coverage	1.73	1
Maximum Coverage	812.87	730.89

For the HepG2 cell line, 42698 positions (bases) have coverage of between 1 and 193 and 69 positions (bases) have coverage of above 193. For the HepG2 (V) sample, 30924 positions (bases) have coverage of between 1 and 71 and 176 positions (bases) have coverage of above 71. The average coverage of both samples is high and indicates trustworthy sequencing.

5.3.1.4. Coverage data of the HeLa samples

Table 5.5 below shows the minimum, maximum and average coverage of the HeLa cell line samples.

Table 5.5: Coverage data of the HeLa samples

Samples	HeLa (C)	HeLa (V)
Average Coverage	245.83	12.35
Minimum Coverage	5.6	1.4
Maximum Coverage	3738.14	1358.15

For HeLa (C), 2952 positions (bases) have coverage of between 1 and 154 and 87 positions (bases) have coverage of above 154. For HeLa (V), 16132 positions (bases) have coverage of between 1 and 158 and 69 positions (bases) have coverage of above 158. The coverage for the untransfected HeLa cells was very high, whilst the average coverage of the transfected HeLa sample was similar to the coverage of most of the other samples. The high coverage of the untransfected HeLa cells is attributable to the low amount of unique contig sequences generated. Many copies of the same fragments of the untransfected HeLa cells were therefore sequenced – many similar fragments were generated, but only 16 unique fragments could be assembled. Coverage of transfected HeLa cells were suitable for further data analysis steps.

5.3.1.5. Summary

There were some problems with the sequencing data. The overall average coverage of the samples are lower than expected. According to Inqaba Biotec technicians, a single lane produces around 50000 reads per lane. The samples showed lower amounts of reads. A recommendation for future analysis would be to use a lane larger than the 10MB lane, or to not split the lane for sequencing of the different samples. This does not fully explain the low coverage seen in some samples, however, nor the fact that some fragments had very high coverages but only a small number of contigs.

A possible reason for sequencing difficulties could be the low amount of DNA sample enriched by the MethylMiner kit, which necessitated the use of PCR amplification. PCR bias could have

occurred, which means that some fragments may have been amplified more effectively than others. This could explain the results seen with the untransfected HeLa cells. In future expansions of this study, it would be preferable that no PCR amplification be used.

5.3.2. Fragment length of fragments (data analysis)

The process of *de novo* assembly generates consensus sequences for each of the contigs. These consensus sequences are useful for further data analysis, such as determining the positions of the DNA fragments in the genome. These sequences allow greater accuracy for data analysis than the contig sequences. In the next data analysis steps, the consensus sequences for each fragment will be regarded as representative of each fragment in a sample. The list below shows the number of useful fragments which were assembled using *de novo* assembly in CLC Bio:

143B (C) sample – 216 useful consensus sequences of fragments generated
143B (V) sample – 143 useful consensus sequences of fragments generated
Fibroblasts (C) sample – 10 useful consensus sequences of fragments generated
Fibroblasts (V) sample – 35 useful consensus sequences of fragments generated
HepG2 (C) sample – 207 useful consensus sequences of fragments generated
HepG2 (V) sample – 143 useful consensus sequences of fragments generated
HeLa (C) sample – 16 useful consensus sequences of fragments generated
HeLa (V) sample – 213 useful consensus sequences of fragments generated

The following paragraphs show the data of fragment length distribution for all eight samples investigated in this study. Tables are also given which show the minimum, maximum and average fragment lengths. It should be stated again that each fragment is in fact the consensus sequence of a single contig, taken to represent the specific fragment from which the sequence is derived.

5.3.2.1. Fragment lengths of 143B samples

Table 5.6 below shows the minimum, maximum and average fragment lengths of the 143B cell line samples.

Table 5.6: Fragment length data of 143B cell line samples

Sample	143B (C)	143B (V)
Fragment Length Average	212	205
Fragment Length Minimum	98	51
Fragment Length Maximum	451	408

Whilst the minimum and maximum fragment lengths of the untransformed 143B sample are higher than the minimum and maximum values of the transformed 143B sample, the average fragment lengths for both samples appear similar. This suggests that the fragment size distribution is comparable for both samples.

Graphical representations of the size distribution of the fragments for each of the 143B cell lines are illustrated in figures 5.4 and 5.5 on the next page. The fragments have been sorted according to identification numbers assigned to the fragments by CLC Bio.

In the untransformed 143B cells, 25 fragments are shorter than 150 bp, 157 fragments have a length between 150 bp and 274 bp and 34 fragments are longer than 274 bp.

In the transfected 143B sample, 20 fragments are shorter than 140 bp, 103 fragments have a length between 140 bp and 272 bp and 20 fragments are longer than 272 bp.

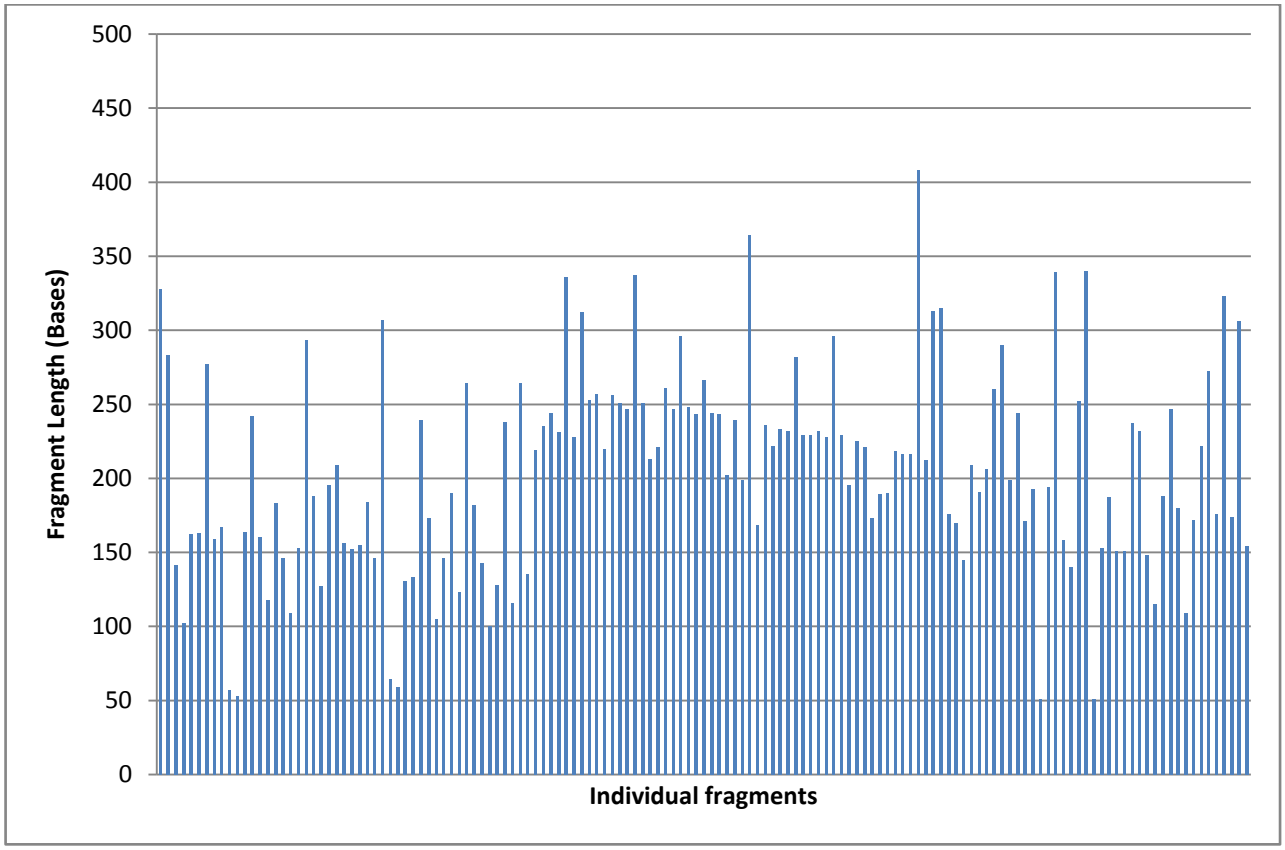


Figure 5.4: Distribution of fragment sizes in the untransformed 143B sample

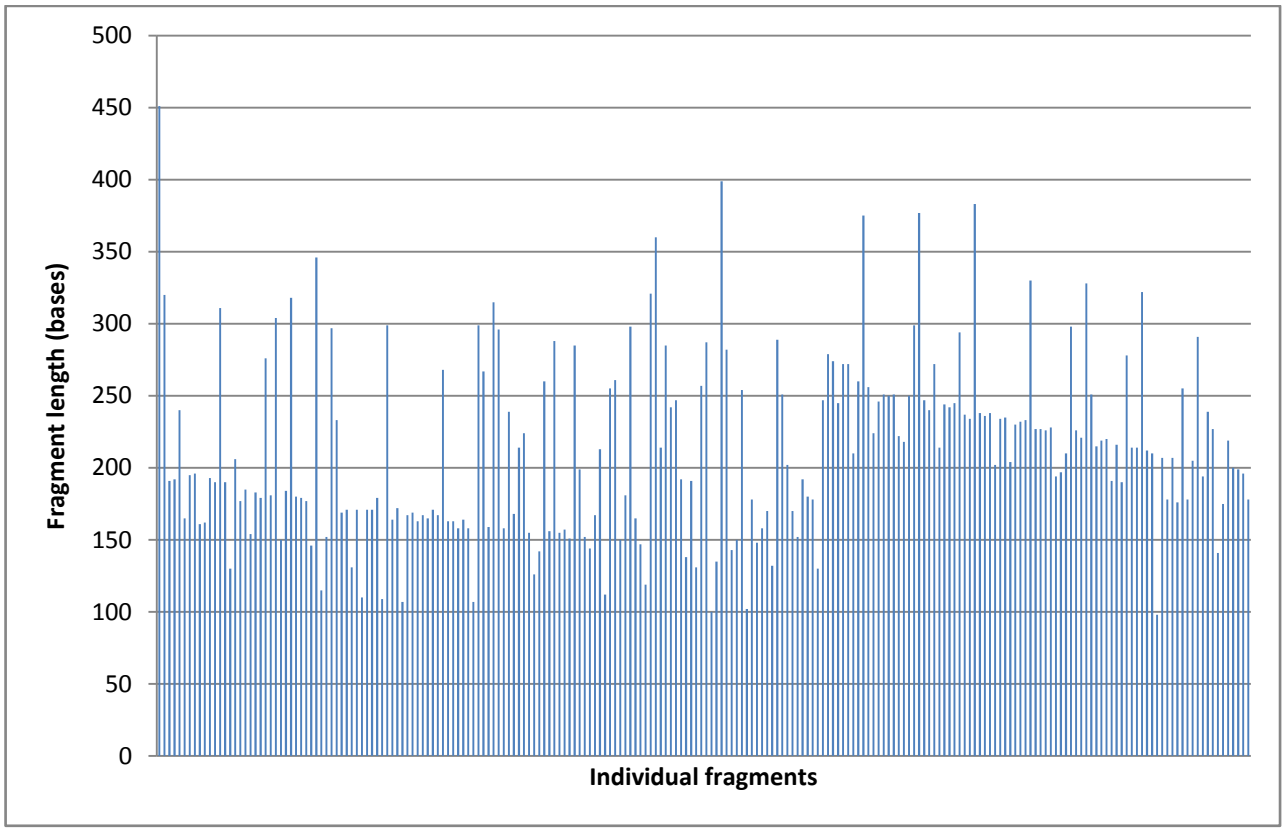


Figure 5.5: Distribution of fragment sizes in the transformed 143B sample

5.3.2.2. Fragment length of fibroblast sample

Table 5.7 below shows the minimum, maximum and average fragment lengths of the fibroblast samples.

Table 5.7: Fragment length data of fibroblast cell line samples

Samples	Fibroblasts (C)	Fibroblasts (V)
Fragment Length Average	185	203
Fragment Length Minimum	115	69
Fragment Length Maximum	253	339

The minimum and maximum fragment lengths of the fibroblast cells show a greater degree of variation in comparison to the 143B cell line samples. The untransformed fibroblast cells show a greater minimum fragment length in comparison to the transformed fibroblasts, whilst the maximum fragment lengths for the untransformed fibroblast cells are lower than those of the transformed fibroblasts. This means that there was less variation in the fibroblast fragment lengths than in the fragments of the 143B samples, which showed a similar fragment size distribution for both the transfected and untransfected 143B cells. These values differ from the fibroblast values. The reason for this is evident from figure 5.6, where it can be seen that the number of sequenced fragments are lower for the untransfected fibroblasts than for the transfected fibroblasts and 143B samples. The transfected fibroblast sample has a higher number of fragments, which means that statistically it could be correlated with the relatively high number of fragments seen in the 143B cell lines. The average fragment length values of the two fibroblast samples correlate with this idea, as the untransfected cells have a much lower average fragment length than those of the transfected fibroblasts. Overall, the transfected fibroblasts correlate more closely with the data of the 143B cells.

Graphical representations of the size distribution of the fragments for each of the fibroblast cell groups are shown in figures 5.6 and 5.7. The fragments have been sorted according to identification numbers assigned to the fragments by CLC Bio.

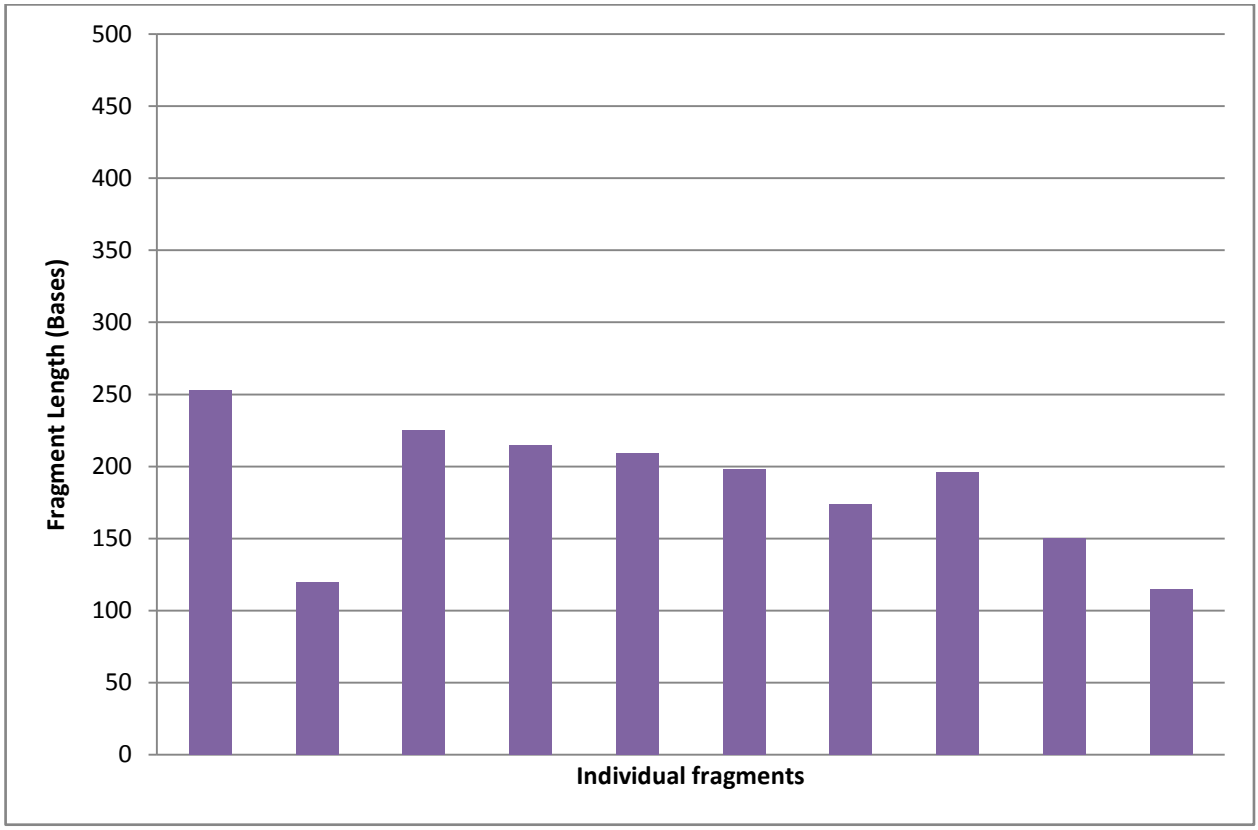


Figure 5.6: Distribution of fragment sizes in the untransformed fibroblast sample

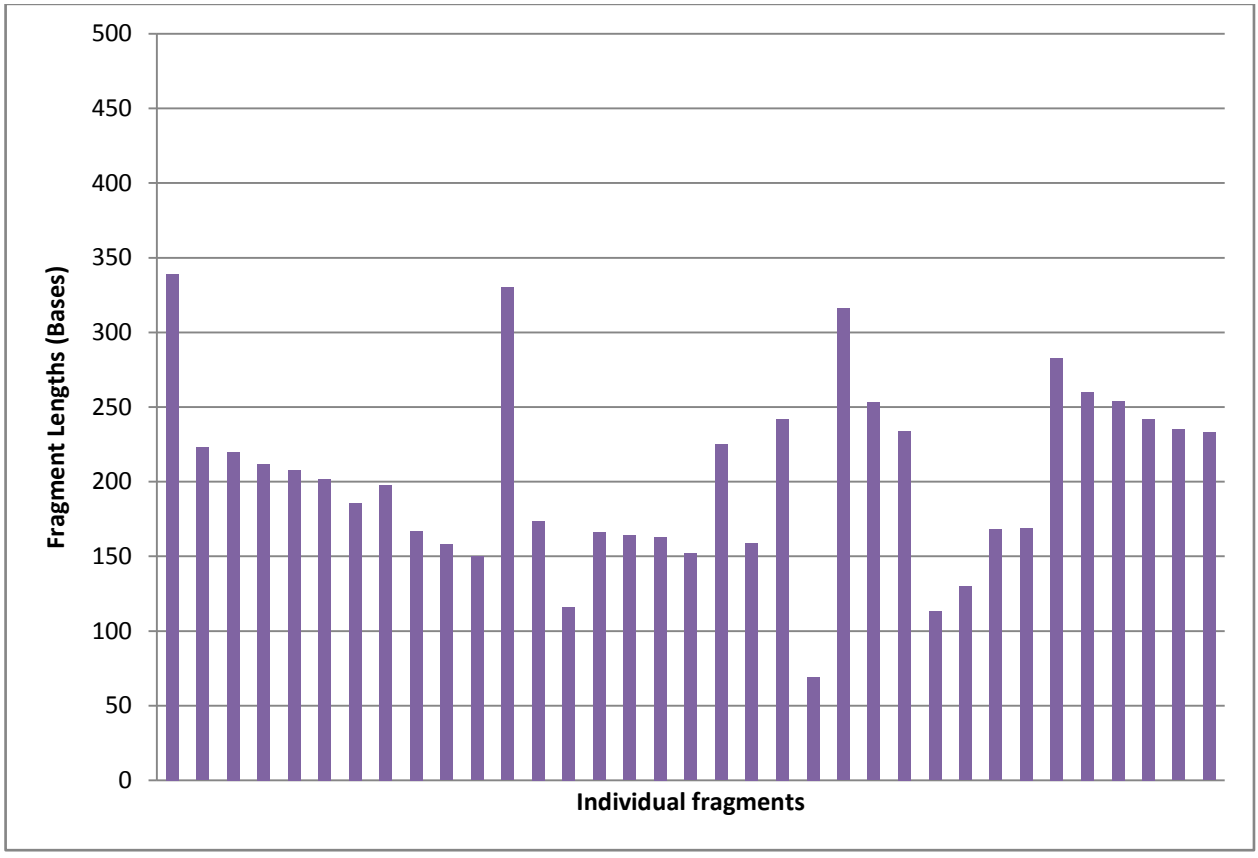


Figure 5.7: Distribution of fragment sizes in the transformed fibroblast sample

For the fibroblasts cell line, 2 fragments are shorter than 150 bp, 7 fragments have a length between 150 bp and 225 bp and 1 fragment is longer than 225 bp.

For the transformed fibroblasts, 4 fragments are shorter than 150 bp, 27 fragments have a length between 150 bp and 260 bp and 4 fragments are longer than 260 bp.

5.3.2.3. Fragment lengths of HepG2 samples

Table 5.8 below shows the minimum, maximum and average fragment lengths of the HepG2 cell line samples.

Table 5.8: Fragment length data of HepG2 cell line samples

Samples	HepG2 (C)	HepG2 (V)
Fragment Length Average	207.81	217.48
Fragment Length Minimum	84	36
Fragment Length Maximum	445	517

The number of fragments assembled for the HepG2 samples are much higher than those of the fibroblast samples. This allows more accurate determinations of the average sizes of the HepG2 fragments to be made, as shown in table 5.8. The average fragment lengths of the HepG2 samples correlate closely with the 143B sample averages, and with each other, whilst the same values of the fibroblast cells differ much more. This is expected, due to the shortcomings of the fibroblast data in comparison to the high number of sequenced fragments in the HepG2 samples.

The minimum and maximum values in table 5.8 show the same pattern of variation as seen in the corresponding values of the fibroblast samples. The values of maximum and minimum fragment length for the untransfected HepG2 sample are closer together, indicating shorter fragments than those seen in the transformed HepG2 sample. This is similar to observations

made in the fibroblast samples, where the untransfected sample shows overall shorter fragments. Further investigation could be potentially relevant for the study. This aspect will be discussed further in the final paragraph of section 5.3.2.5, where the summative results of all samples' fragment data will be investigated.

It should be noted that a large number of fragments had been assembled for the HepG2 cell line samples. This indicated that there were adequate amounts of samples enriched with the MethylMiner kit, all with suitable purity for sequencing. Graphical representations of the size distribution of the fragments for each of the HepG2 cell lines are shown in figures 5.8 and 5.9 on the next page. The fragments have been sorted according to identification numbers assigned to the fragments by CLC Bio.

For the HepG2 cell line, 18 fragments are shorter than 145 bp, 166 fragments have a length between 145 bp and 265 bp and 23 fragments are longer than 265 bp. In the transfected HepG2 cells, 23 fragments are shorter than 148 bp, 101 fragments have a length between 148 bp and 288 bp and 19 fragments are longer than 288 bp.

Refer to next page (page 62) for figures 5.8 and 5.9 of HepG2 samples.

5.3.2.4. Fragment lengths of HeLa samples

Table 5.9 below shows the minimum, maximum and average fragment lengths of the HeLa cell line samples.

Table 5.9: Fragment length data of HeLa cell line samples

Samples	HeLa (C)	HeLa (V)
Fragment Length Average	189.93	200.00
Fragment Length Minimum	148	46
Fragment Length Maximum	256	377

<continued on page 63>

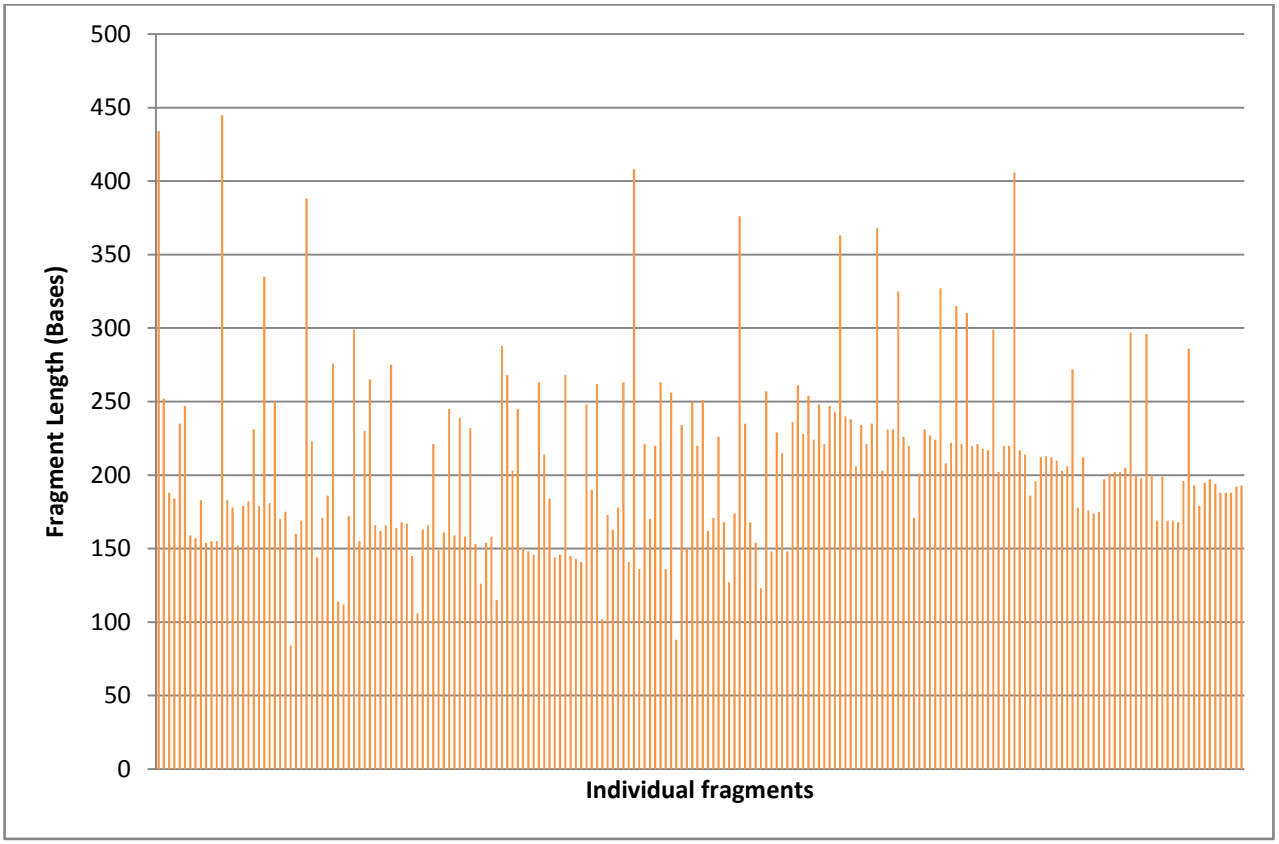


Figure 5.8: Distribution of fragment sizes in the untransformed HepG2 sample

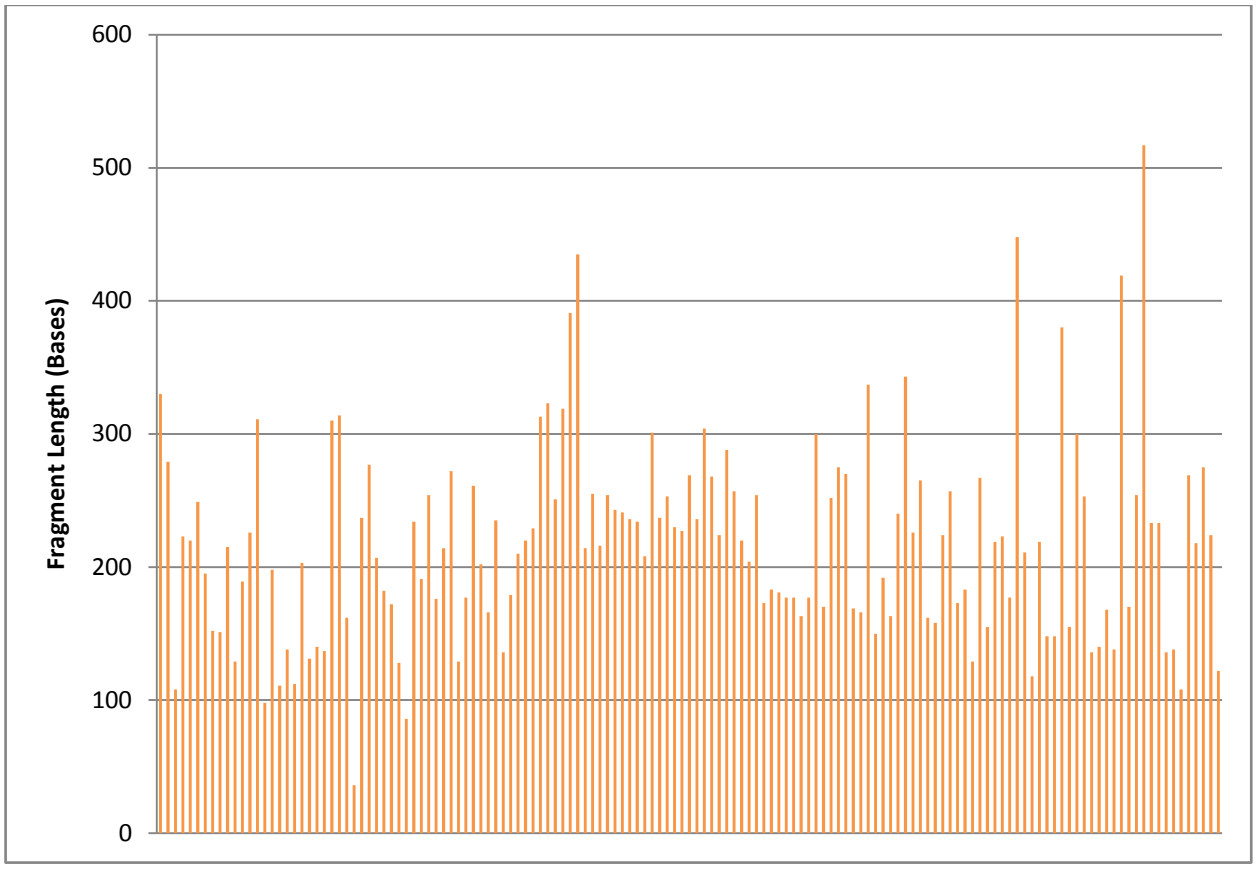


Figure 5.9: Distribution of fragment sizes in the transformed HepG2 sample

<continued from page 61>

There is large variation between the minimum fragment length of the untransfected HeLa cells and the minimum fragment length of the transfected HeLa cells, which have a fragment length of 46 bp in comparison to the 148 bp minimum of its counterpart. This variation might again be attributable to the relatively low amount of untransformed HeLa sample fragments that were assembled from the sequencing data. Shorter fragments may not have been sequenced. The variation in the maximum fragment lengths may be attributable to the low number of fragments as well. However, the average fragment length of both samples does not vary a great deal, which shows that the larger and smaller fragments of the untransfected HeLa cell samples were not sequenced successfully. Otherwise, these fragments were "deselected" during nebulization or emPCR prior to the sequencing process.

These results, along with those seen in the fibroblast samples, show that the number of fragments assembled should be kept in mind throughout the other data analysis steps. If only a small number of fragments are analyzed, the researcher should take care not to make rash global conclusions. The researcher can make conclusions with much more confidence when more fragments are included in a sample. Therefore, it is important that the quality of the generated data should be kept in mind throughout the data analysis steps.

Graphical representations of the size distribution of the fragments for each of the HeLa cell lines are illustrated in figures 5.10 and 5.11 on the next page. The fragments have been sorted according to identification numbers assigned to the fragments by CLC Bio.

For the untransfected HeLa cells, 3 fragments are shorter than 158 bp, 10 fragments have a length between 158 bp and 220 bp and 3 fragments are longer than 220 bp.

For the transfected HeLa samples, 23 fragments are shorter than 142 bp, 175 fragments have a length between 142 bp and 257 bp and 33 fragments are longer than 257 bp.

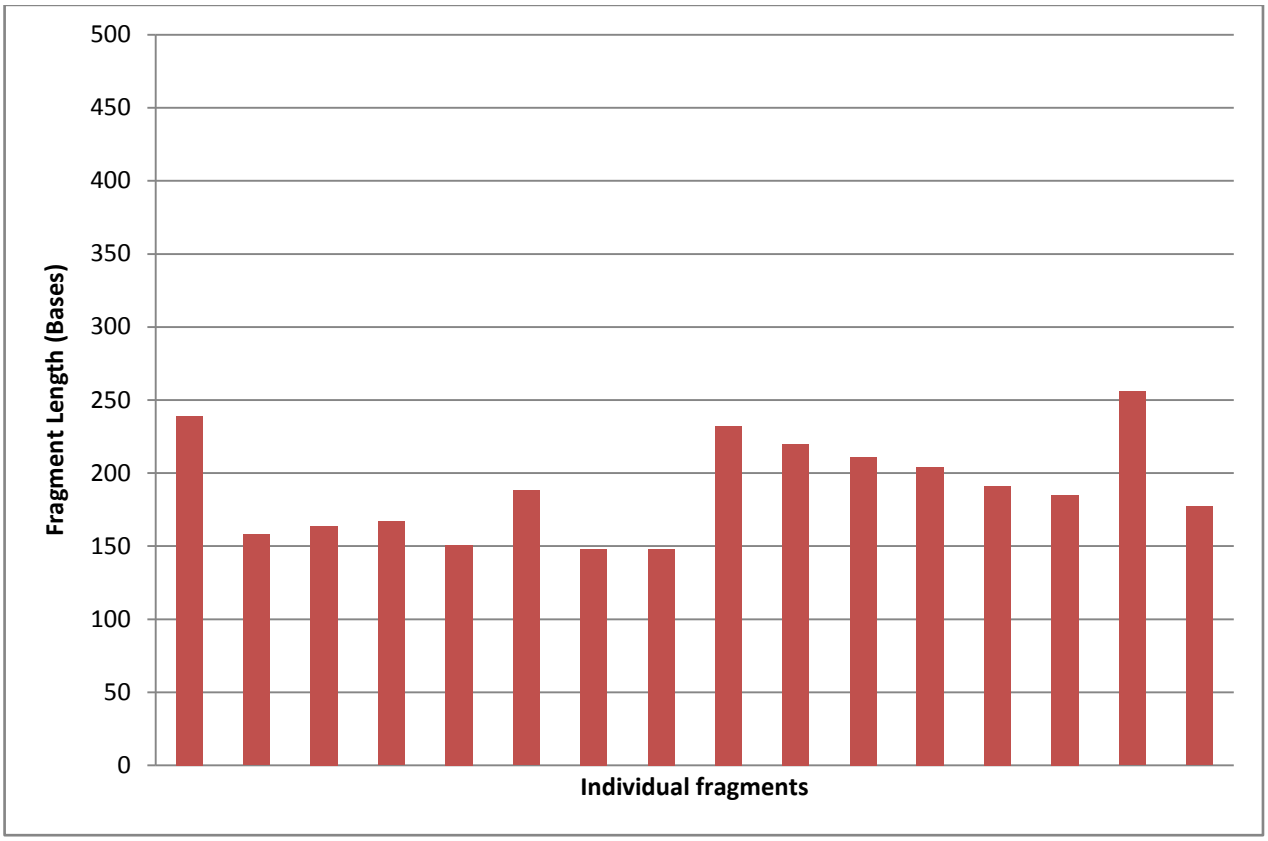


Figure 5.10: Distribution of fragment sizes in the HeLa sample

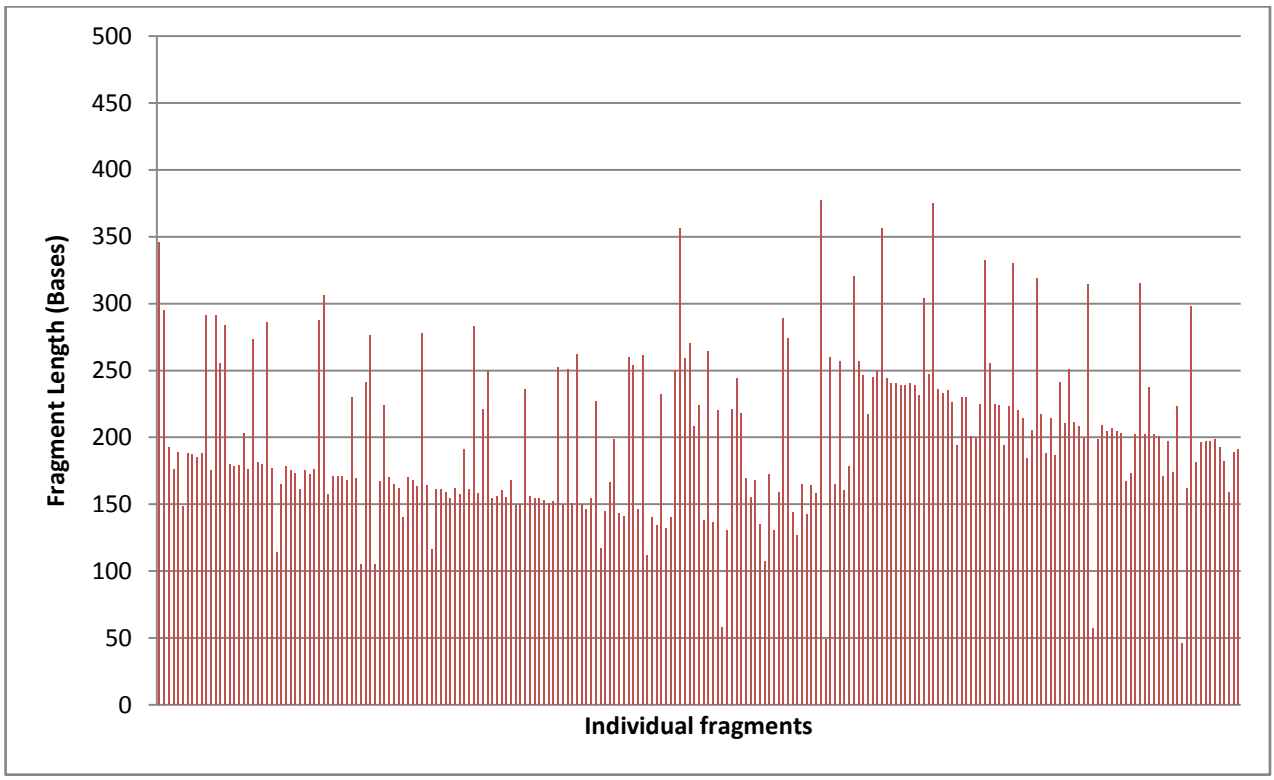


Figure 5.11: Distribution of fragment sizes in the transformed HeLa sample

5.3.2.5. Summary

In the previous paragraphs, all the results of the individual cell lines were compared, with only minor reference made to other cell lines. Table 5.10 gives the minimum, maximum and average fragment lengths of all cell lines in a single table.

Table 5.10: Summative fragment length data of all cell line samples used in the study

Cell line	143B (C)	143B (V)	Fibro (C)	Fibro (V)	HepG (C)	HepG (V)	HeLa (C)	HeLa (V)	Averages
Fragment Length Average	212.31	205.3	185.5	203.23	207.82	217.48	189.94	200	202.70
Fragment Length Minimum	98	51	115	69	84	36	148	46	80.875
Fragment Length Maximum	451	408	253	339	445	517	256	377	380.75

Included in table 5.10 are averages for the fragment length values of all samples used in the study. Most of the average values of the fragment lengths do not vary far from 202.7 bases per fragment. The highest average number of basepairs per sample is 217.48 bp for the transfected HepG2 sample, whilst the lowest average number of basepairs per sample is 185.5 bp for the untransfected fibroblast cells. The low fibroblast values might be attributable to the lower efficiency of fragment sequencing for the specific sample. However, the high average length of the transfected HepG2 sample is accurate in terms of the quality of the sample sequencing.

The minimum and maximum fragment sizes vary much more. This is attributable to the quality of fragment sequencing in some cases, where very short or very long fragments may not have been sequenced effectively. This ineffective sequencing may be due to low levels of sample DNA sent for sequencing, or may be due to problems with the sample preperation. However, the relatively small variation in average fragment length suggests that the main fragments in all samples are centralized around the same sizes, and that these were sequenced correctly.

Figure 5.12 gives the longest and shortest fragments in every sample of the study:

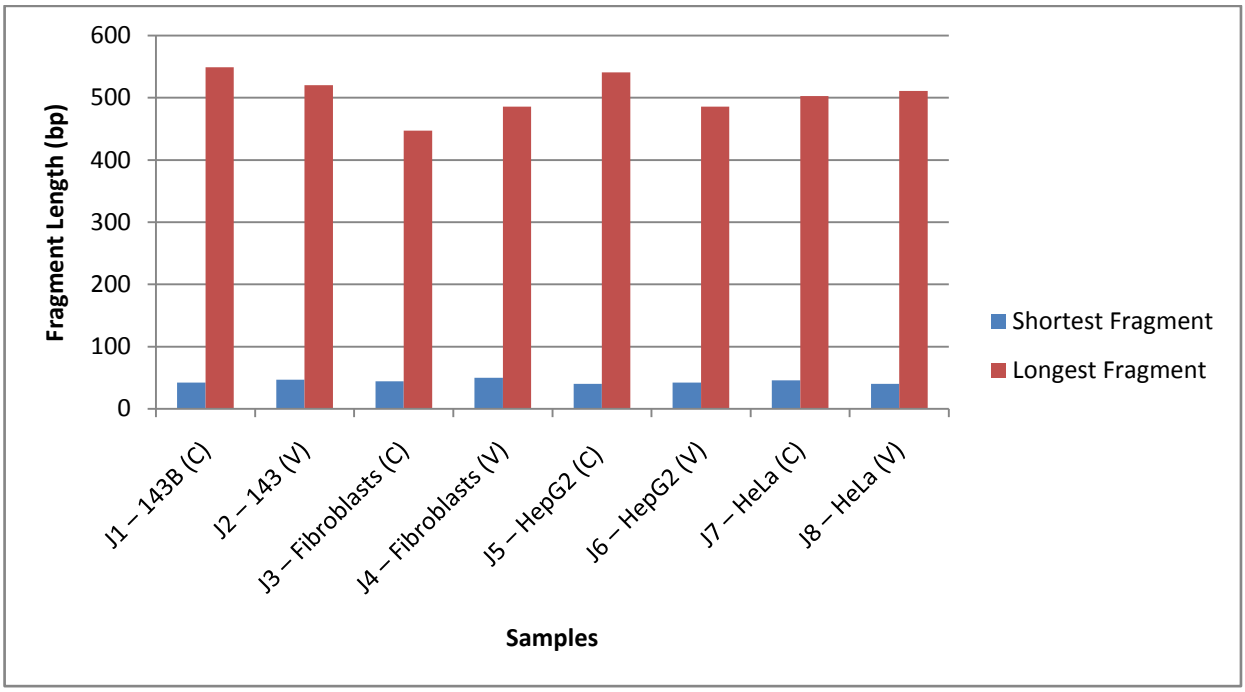


Figure 5.12: Maximum and minimum fragment sizes in the cell line samples of the study

The range of fragment sizes of the sequenced fragments are similar to those seen in a previous gel photo (figure 5.1) illustrating fragment size after restriction enzyme digestion, indicating that the selected fragments ranged from 50 bp to 600 bp in length. The small amount of variation in fragment length shows that the enzyme digestion effectively fractioned the sample before MethylMiner kit use and that enrichment of fragments of the varying sizes was done correctly.

5.3.3. CG-content of fragments (data analysis)

In the following section, the number of CG-dinucleotides in the sample fragments are investigated for each of the cell lines. DNA methylation occurs in CG-dinucleotides and CpG islands - refer to chapter 2 for a full discussion of literature concerning CG-dinucleotides.

5.3.3.1. CG-content of 143B fragments

Table 5.11 gives the CG-content data of the fragments of the untransfected and transfected 143B cell line samples.

Table 5.11: CG-content data of 143B cell line samples

Samples	143B (C)	143B (V)
CG Average (per fragment)	10.61	9.87
CG Minimum (per fragment)	2	1
CG Maximum (per fragment)	23	24
CG-content in % (total)	52.21	51.90

The minimum number of CG-dinucleotides is of importance to verify the correct selection of the samples using the MethylMiner kit. A minimum number of a single CG group is expected per fragment, as selection of methylated DNA fragments is based on methylation (which occurs in the CG-dinucleotide). This also shows that enrichment with the MethylMiner kit is sensitive even if only a single methylated cytosine is present in a sample. As has been noted, the maximum and minimum numbers of CG-groups of both the transfected and untransfected samples are roughly the same, showing the repeatability of MethylMiner kit enrichment.

The total percentage of CG-content is roughly 50%, but it should be noted that not all CG-dinucleotides are necessarily methylated. At least one of the CG-groups must have been methylated for enrichment with the MethylMiner kit to have occurred, but this does not negate the methylation of more than one CG-group. It is likely that more than one CG-group was methylated in each fragment, as literature states that 70-90% of CpG islands are methylated in mammals (Ehrlich *et al*., 1982; Tucker, KL. 2001). This means that roughly 50% of each fragment consists of CG-dinucleotides, and according to literature 70-90% of the CG-dinucleotides are methylated. Approximately 40% of each fragment's length is expected to consist of methylated CG-groups. Figure 5.13 and 5.14 show the distribution of CG-dinucleotides in sample fragments:

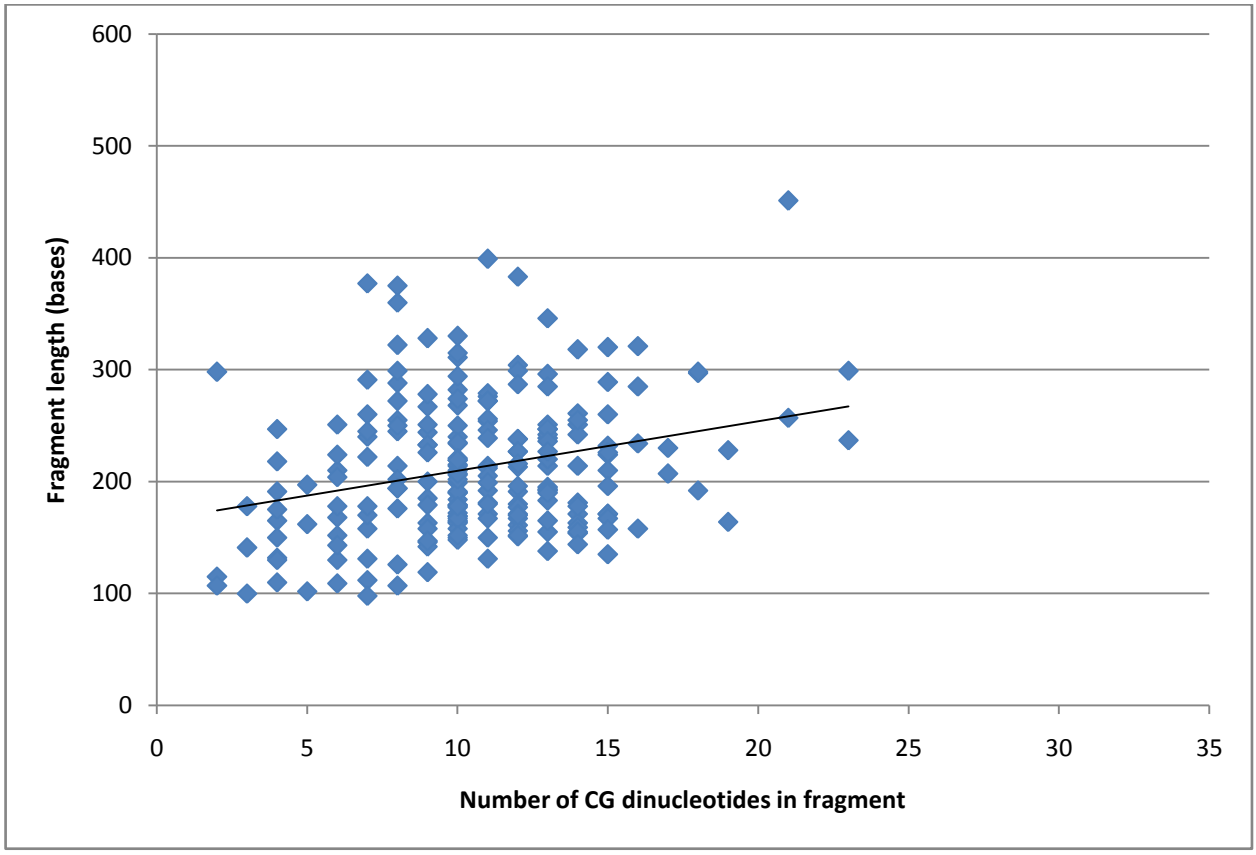


Figure 5.13: Distribution of CG-dinucleotides in the 143B cell sample with trend line shown

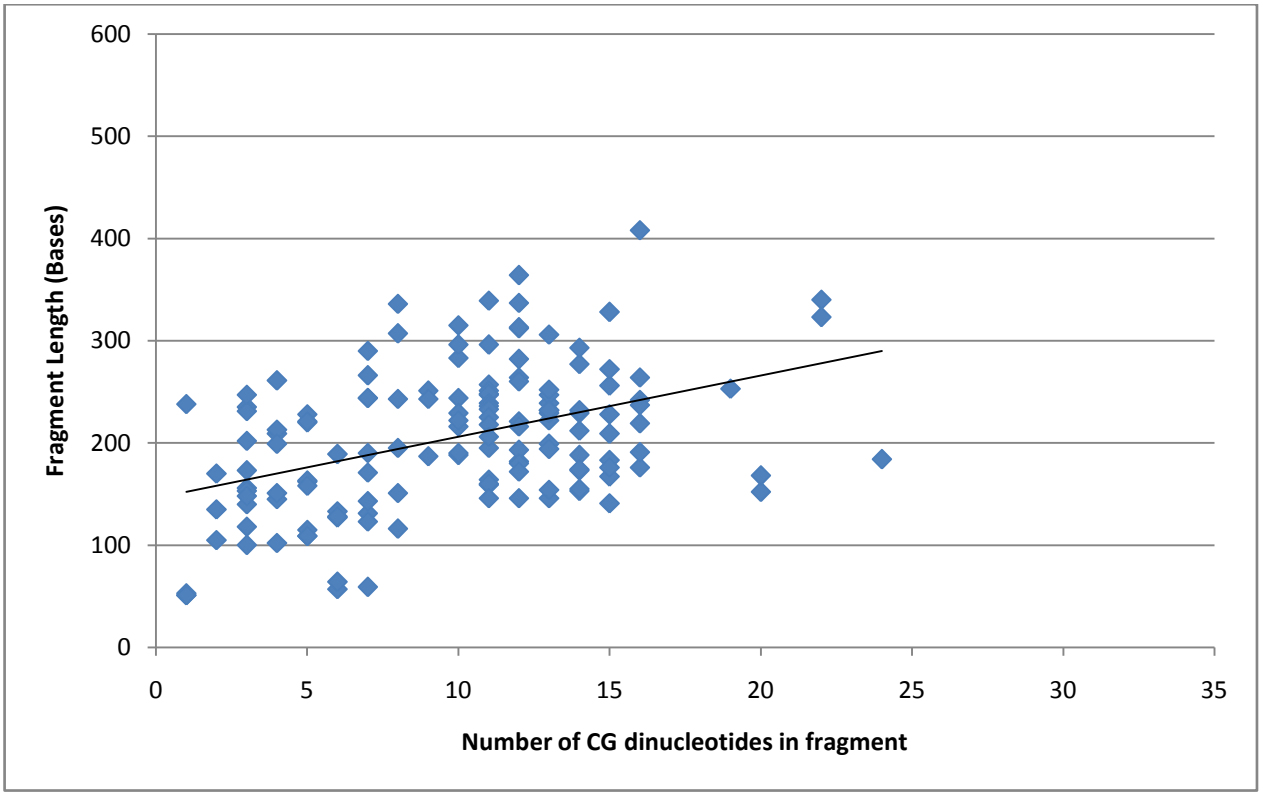


Figure 5.14: Distribution of CG-dinucleotides in the transfected 143B cell sample with trend line shown

Figures 5.13 and 5.14 give a similar distribution of CG-dinucleotides in the sequenced, methylated fragments. The charts both show trend lines. These trend lines are used only to indicate the rough distribution of the data points in a linear fashion. These trend lines suggest that, the longer the fragment length, the higher the number of CG-dinucleotides per fragment. However, the fact that some data points fall outside the area of the trend line suggests that selection by means of the MethylMiner kit is not dependent on higher numbers of CG-dinucleotides to enrich larger methylated fragments.

5.3.3.2. CG-content of fibroblast fragments

Table 5.12 gives the CG-content data of the fragments of the untransfected and transfected fibroblast cell line samples.

Table 5.12: CG-content data of fibroblast samples

Samples	Fibroblasts (C)	Fibroblasts (V)
CG Average (per fragment)	10.90	12.11
CG Minimum (per fragment)	5	1
CG Maximum (per fragment)	16	25
CG-content in % (total)	57.57	58.42

Table 5.12 shows that there are larger variations between the minimum and maximum values of CG-dinucleotides of the fibroblast samples than in the 143B samples. The clearest difference is in the maximum values, but the minimum values provide interesting information on the sequencing of the untransfected fibroblasts.

The fact that the fragment with the lowest number of CG-dinucleotides in the untransfected sample had five CG-dinucleotides suggests that the smaller fragments of the sample were either not sequenced or are unrepresented due to PCR bias. The same may be said for the larger fragments, as the maximum number of CG-dinucleotides in these fragments is 16 CG-dinucleotides. This seems to indicate that larger fragments are also not represented.

The average percentages of CG-content are similar, but higher than the same percentages for the 143B fragments. This aspect will be discussed more fully in the concluding paragraphs (5.3.3.5) of this section, where all results of the samples in terms of CG-dinucleotide content are discussed.

Figures 5.15 and 5.16 on the next page give the distribution of CG-dinucleotides in the methylated DNA fragments. A trend line is included in both charts, though the data distribution of the untransfected fibroblasts is inadequate for accurate comparison between the two samples. The trend line for the transfected fibroblast sample is similar to the trend lines seen in the charts of the 143B cell lines (figures 5.13 and 5.14). This is expected, as longer fragments would contain a higher number of CG-dinucleotides.

5.3.3.3. CG-content of HepG2 fragments

Table 5.13 gives the CG-content data of the fragments of the untransfected and transfected HepG2 cell line samples.

Table 5.13: CG-content data of HepG2 cell line samples

Samples	HepG2 (C)	HepG2 (V)
CG Average (per fragment)	12.06	12.05
CG Minimum (per fragment)	2	2
CG Maximum (per fragment)	27	32
CG-content in % (total)	54.90	52.76

The average number of CG-dinucleotides in the cell lines are approximately the same, as is the minimum number of CG-dinucleotides per fragment. The maximum number of CG-dinucleotides per sample shows variation within five CG groups of each other.

< continues on page 72 >

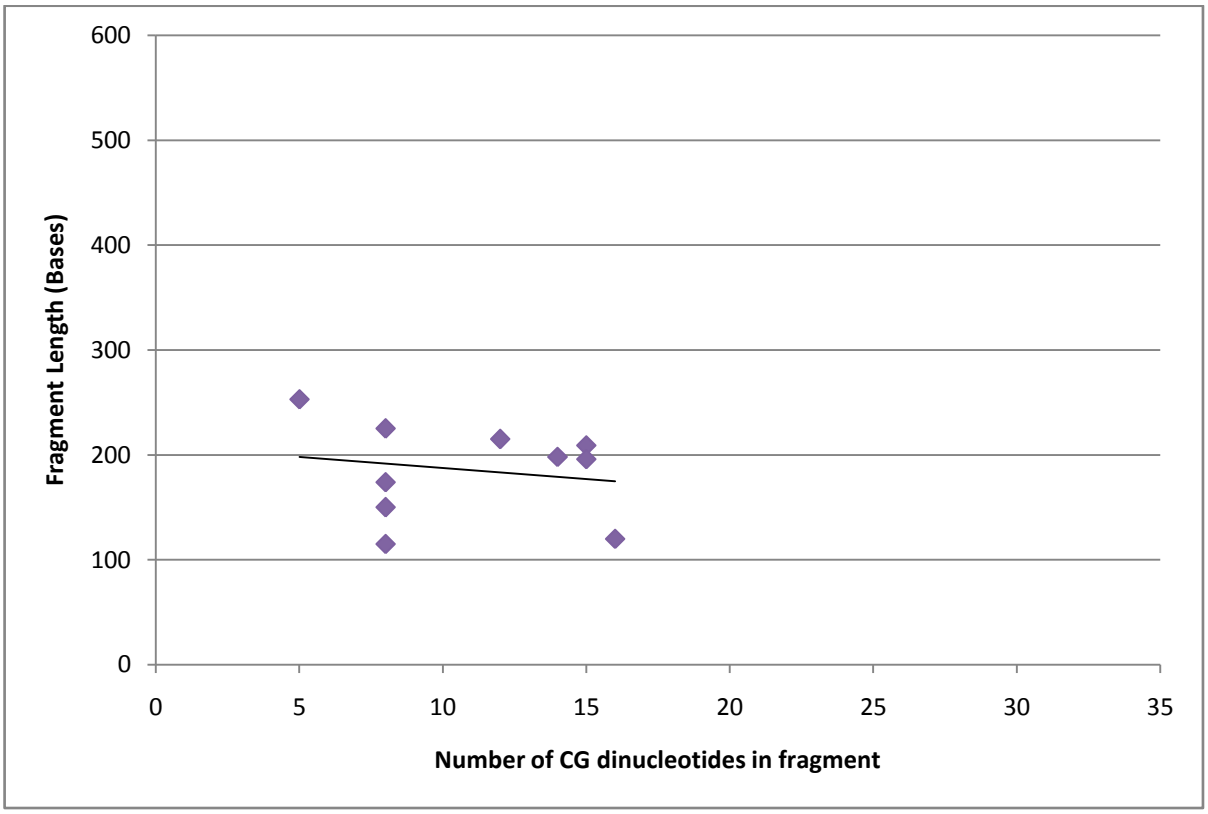


Figure 5.15: Distribution of CG-dinucleotides in the fibroblast sample with trend line shown

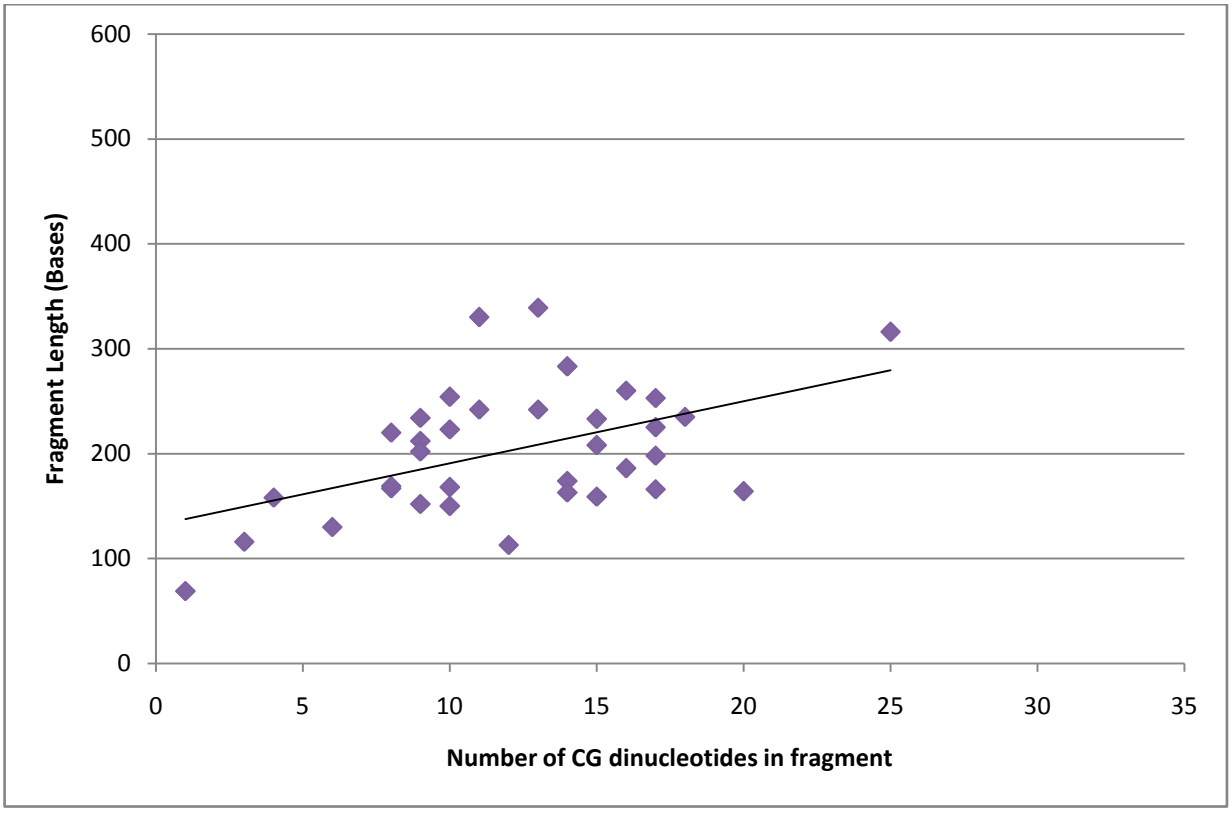


Figure 5.16: Distribution of CG-dinucleotides in the transfected fibroblast cell sample with trend line shown

< continued from page 70 >

This is likely attributable to chance, although it might suggest that the transfected HepG2 sample have some fragments with a higher number of CG-dinucleotides present. The fact that the CG-content is similar suggests that these fragments in the transfected HepG2 cells with a high number of CG-dinucleotides are more numerous than the same size fragments in the untransfected sample. However, these larger fragments are still relatively few in the transfected samples.

The total percentage of CG-content is similar. These values are lower than those observed in the fibroblast samples, but correlate with those of the 143B cells.

Figures 5.17 and 5.18 give the distribution of CG-dinucleotides in sample fragments and are illustrated on the next page (page 73). These figures show a similar distribution of CG-dinucleotides in the methylated fragments of both HepG2 samples. The charts show trend lines which are similar and suggest that, the longer the fragment length, the higher the number of CG-dinucleotides per fragment. This is expected and correlates with the trend lines observed in the 143B cell line samples.

Again, the fact that some data points fall outside the surrounding area of the trend line suggests that longer fragments are enriched with the MethylMiner kit regardless of the number of CG-dinucleotides present in the fragments.

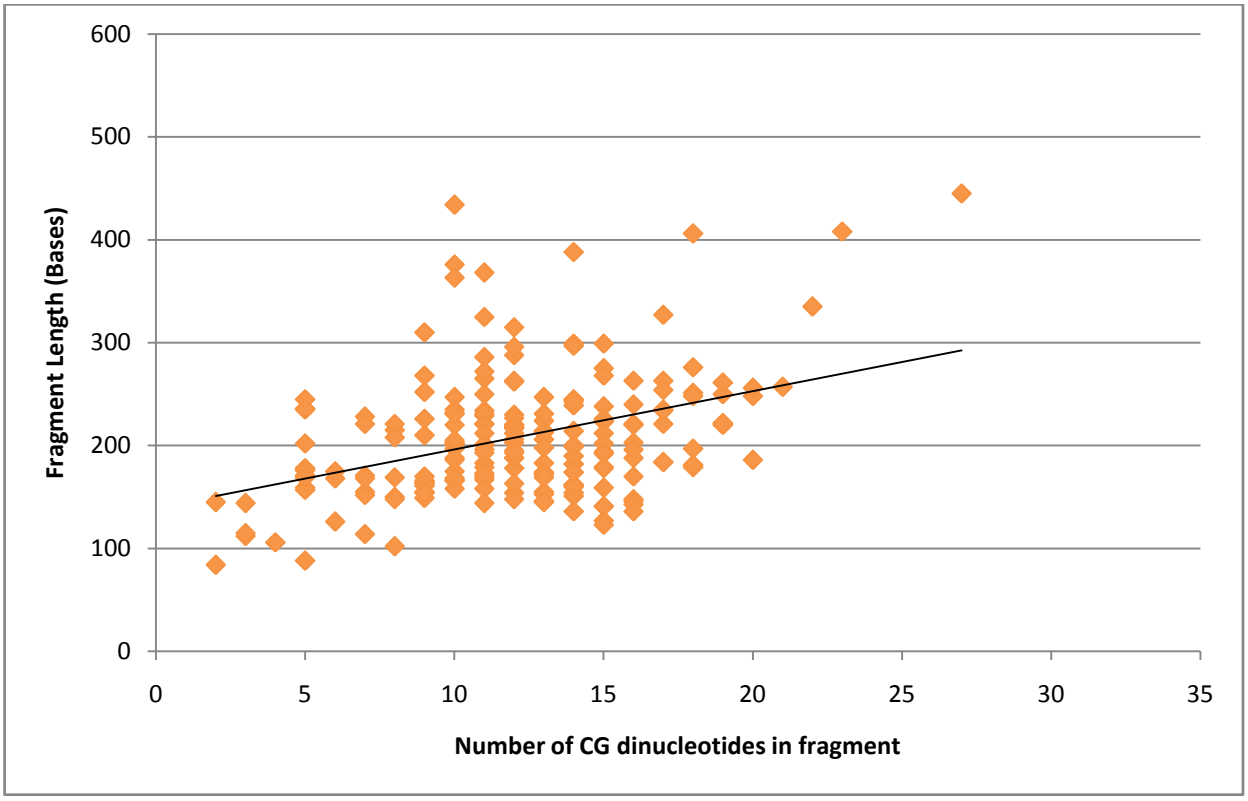


Figure 5.17: Distribution of CG-dinucleotides in the HepG2 cell sample with trend line

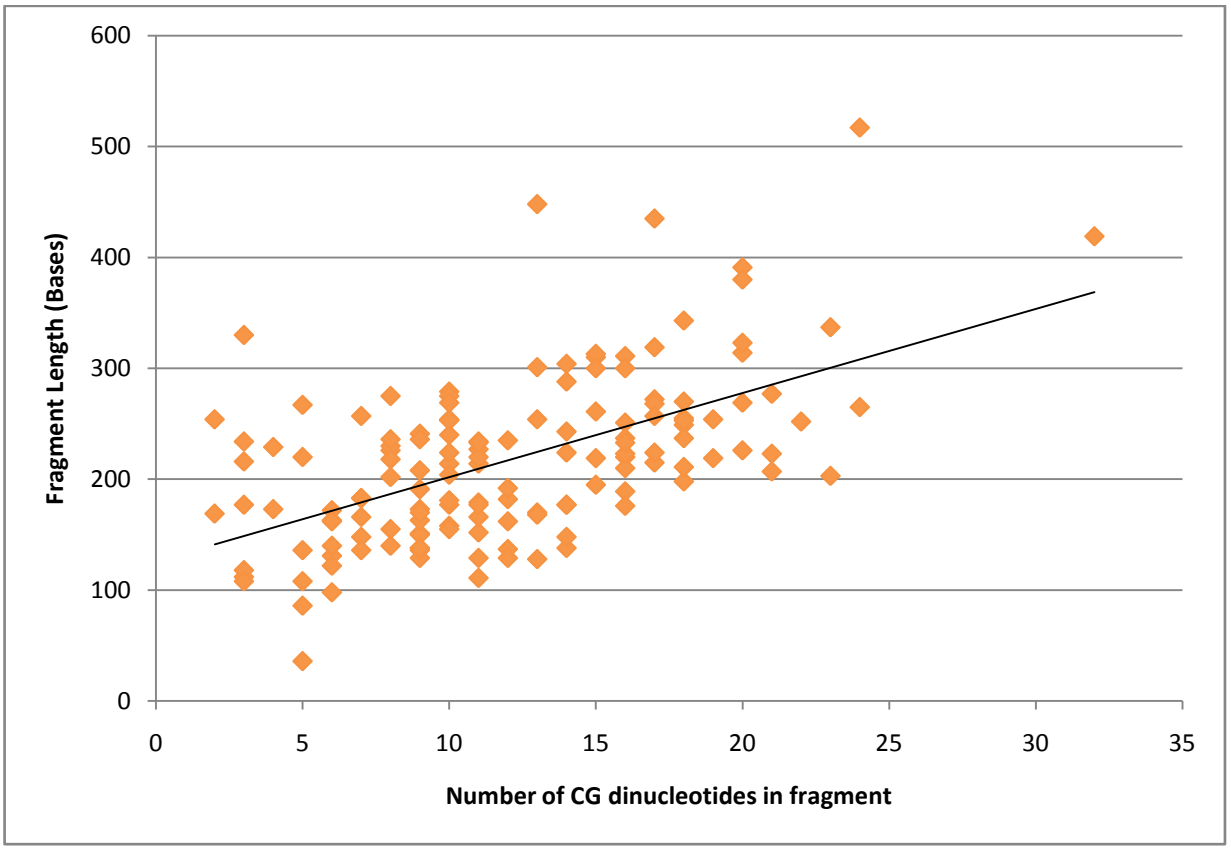


Figure 5.18: Distribution of CG-dinucleotides in the transfected HepG2 cell sample with trend line shown

5.3.3.4. CG-content of HeLa fragments

Table 5.14 gives the CG-content data of the fragments of the untransfected and transfected HeLa cell line samples.

Table 5.14: CG-content data of HeLa cell line samples

Samples	HeLa (C)	HeLa (V)
CG Average (per fragment)	13.00	12.45
CG Minimum (per fragment)	9	1
CG Maximum (per fragment)	18	31
CG-content in % (total)	55.54	54.65

Table 5.14 gives values for the HeLa cell line fragments. These values are similar to those observed in the fibroblasts. The table shows differences between the average, minimum and maximum number of CG-dinucleotides of the HeLa fragments, where the minimum and maximum number of CG-dinucleotides per fragment for each of the HepG2 samples differ. The fact that the fragment with the lowest number of CG-dinucleotides in the untransfected sample had nine CG-dinucleotides instead of the minimum expected number of one suggests that the smaller fragments of the sample were either not sequenced or are unrepresented due to PCR bias; this is similar to the inadequate fragment generation results of the untransfected fibroblast sample. The same may be said for the larger fragments, which also appear to be underrepresented.

The average percentages of CG-content are similar for both HepG2 samples and fall in the midrange between the lower 143B and HeLa average percentages and the higher fibroblast percentages.

Figure 5.19 and 5.20 on the next page gives the distribution of CG-dinucleotides in the methylated DNA fragments of HeLa cell lines.

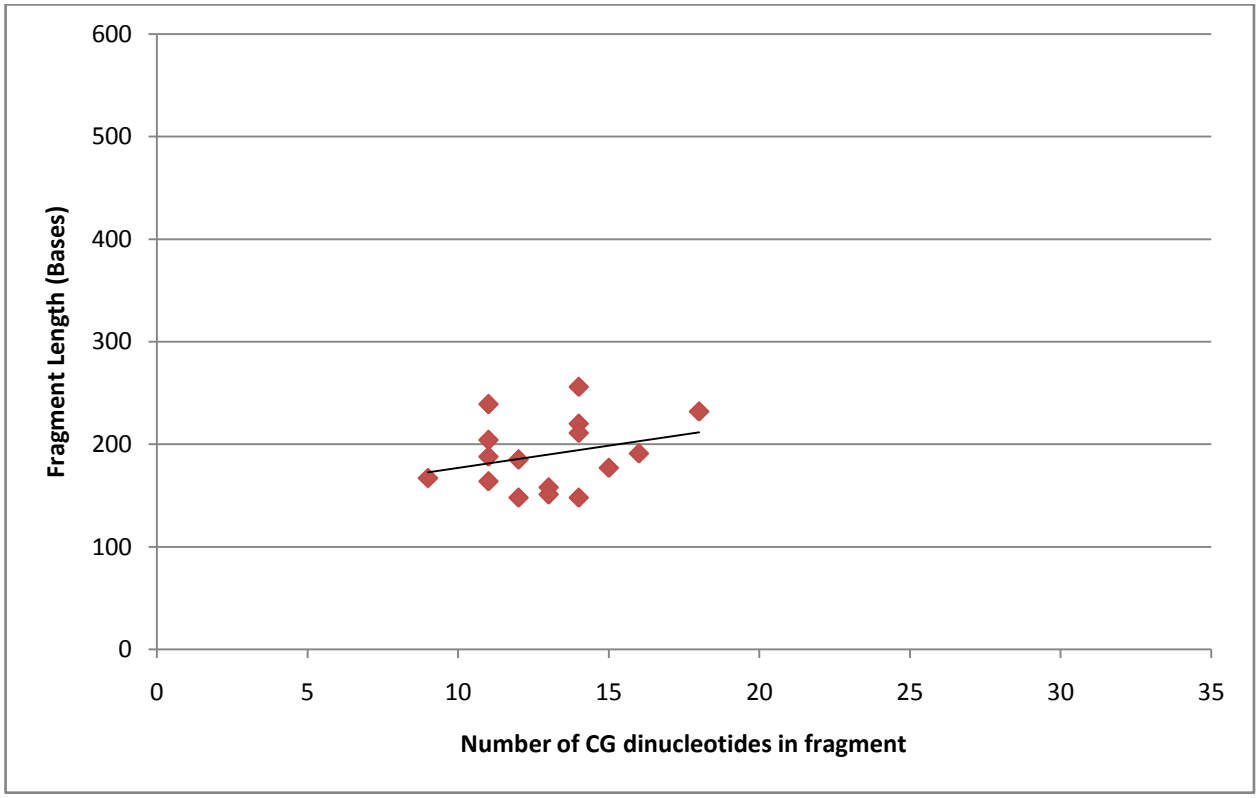


Figure 5.19: Distribution of CG-dinucleotides in the HeLa cell sample with trend line shown

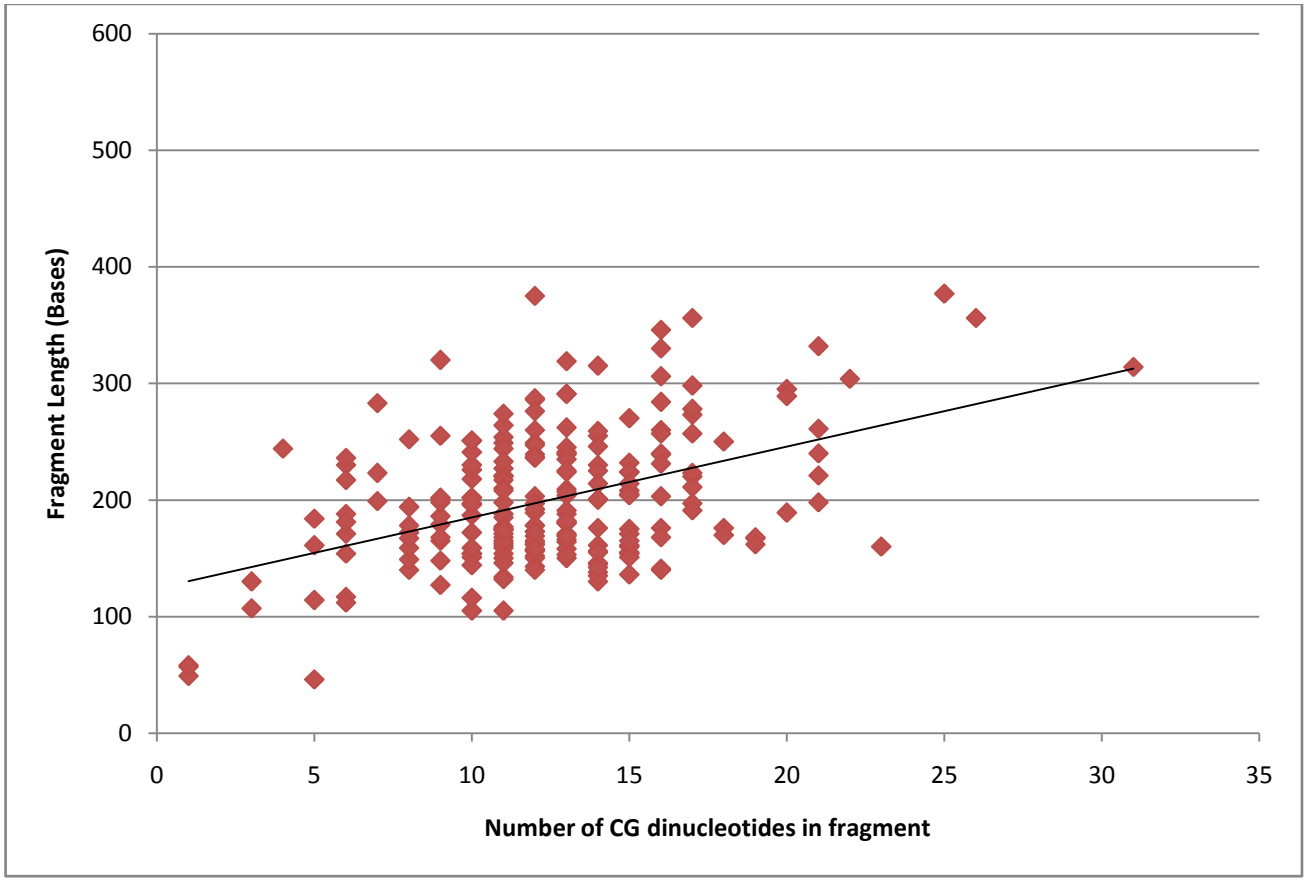


Figure 5.20: Distribution of CG-dinucleotides in the transfected HeLa cell sample with trend line shown

A trend line is included in both charts (figure 5.19 and 5.20), though the data distribution of the untransfected HeLa sample is insufficient for accurate comparison between the two samples. However, both trend lines are similar even though the larger and smaller fragments may be underrepresented in the untransfected HeLa samples. This is possibly simply due to chance. Both trend lines for the HeLa samples are similar to the trend lines seen in the 143B and HepG2 cell lines' charts (figures 5.13, 5.14, 5.17, 5.18).

5.3.3.5. Summary

In the previous paragraphs, all the results of the individual cell lines were compared separately with each other and only minor reference was made to other cell lines. Table 5.15 gives the average, minimum, maximum and percentage CG-dinucleotide content in cell lines in a single table.

Table 5.15: Summative CG-dinucleotide data of all cell line samples used in the study

Samples	143B (C)	143B (V)	Fibroblasts (C)	Fibroblasts (V)	HepG2 (C)	HepG2 (V)	HeLa (C)	HeLa (V)
CG Average (per fragment)	10.61	9.87	10.90	12.11	12.09	12.05	13	12.45
CG Minimum (per fragment)	2	1	5	1	2	2	9	1
CG Maximum (per fragment)	23	24	16	25	27	32	18	31
CG-content in % (total)	52.21	51.90	57.57	58.42	54.90	52.76	55.54	54.65

The possible problems with the data generation of the untransfected fibroblasts and the untransfected HeLa cells have already been noted.

However, these aspects may be attributable to the fact that both these cell lines may be hypomethylated in comparison to the other samples. Except for these two samples, most fragments show similar distribution of CG-regions (maximum, minimum and average).

The CG-content percentages differ between different cell lines, but are invariably the same for the untransfected and transfected samples of the same cell line. This suggests that there is a constant percentage of CG-groups in specific cell lines that remain constant even with the insertion of a vector into the original cell line – the DNA sequence itself remains unaltered. The fact that these variations exist between cell lines suggests that CG-percentages of all samples will remain roughly the same for the same cell lines. However, not all CG-groups are methylated and therefore DNA methylation may be different.

5.3.4. BLAST results of fragments (data analysis)

Basic **L**ocal **A**lignment **S**earch **T**ool (BLAST) is a search algorithm used in bioinformatics to compare DNA sequence information (URL: *www.ncbi.nlm.nih.gov/blast/*). The following results were generated using BLAST analyses of the human sequences (both reference only and using all available assemblies) to align the fragments to specific positions on the human genome.

5.3.4.1. BLAST results of untransfected 143B fragments

A total of 216 methylated fragment sequences were analyzed using BLAST. An analysis was done using all available assemblies (reference and alternate) for the human genome.

In total, 882 BLAST hits were generated for the untransfected 143B cells (these include BLAST hits which are similar to each other, results on repeat areas of the reference genome sequence or multiple similar hits on different assemblies of the human genome).

Eight of the 882 hits could be correlated with specific genes.

These genes were derived from all assemblies of the genome and are given in table 5.16 (only those fragments with a BLAST score of 75 and above are included, as any BLAST values lower than this may indicate that the identified gene is not relevant for the specific sample).

Table 5.16: List of BLAST hits associated with genes for the untransfected 143B fragments

Total Score	Official Symbol	Official Full Name	Gene Type	Location	Additional
470	GTF2I	general transcription factor IIi	protein coding	7q11.23	This gene encodes a multifunctional phosphoprotein with roles in transcription and signal transduction. It is deleted in Williams-Beuren syndrome
487	LOC100131754	hypothetical LOC100131754	miscRNA	1p36.33	
215	LOC100132144	putative protein FAM27E1-like	protein coding	9p11.2	
358	LOC727805	putative POM121-like protein 1-like	pseudo	6p11.2	
150	LOC643650	hypothetical LOC643650	miscRNA	10q11.22	
95.3	LOC100133900	double homeobox protein 4-like	pseudo	Chromosome Unknown	
89.8	LOC100510456	double homeobox protein 4-like	pseudo	Chromosome Unknown	
89.8	LOC100510245	protein FAM27D1-like	protein coding	Chromosome 9	

Table 5.16 shows the genes which can be associated with methylated fragments from the sample. This means that there is a degree of DNA methylation in these genes, because a specific methylated DNA fragment was derived from this genomic position in the original DNA sample (before enrichment through the MethylMiner kit).

Two hits were found in the double homeobox protein 4-like pseudogene, from the DUX family of genes (Homeobox Database, 2010). The DUX family is protein coding. Another identified protein coding gene is the GTF2I (general transcription factor IIi), which encodes a

multifunctional phosphoprotein. Two hypothetical genes were also identified. These are genes which encode for a hypothetical protein – a protein that is predicted but for which no experimental results exist (Hernández *et al*., 2009). Furthermore, three pseudogenes were identified – genes which have lost their protein-coding ability or are no longer expressed in the cell (Mighell *et al*., 2000). The role of the identified genes in DNA methylation and in epigenetic events are not obvious, as they appear on different chromosomes and do not appear to have a functional relationship.

Another BLAST analysis was done using only the reference sequence of the human genome. In this analysis, 135 unique BLAST hits could be associated with positions in the genome sequence. These hits are illustrated in figure 5.21, which shows the positions of the BLAST hits on a chromosomal view.

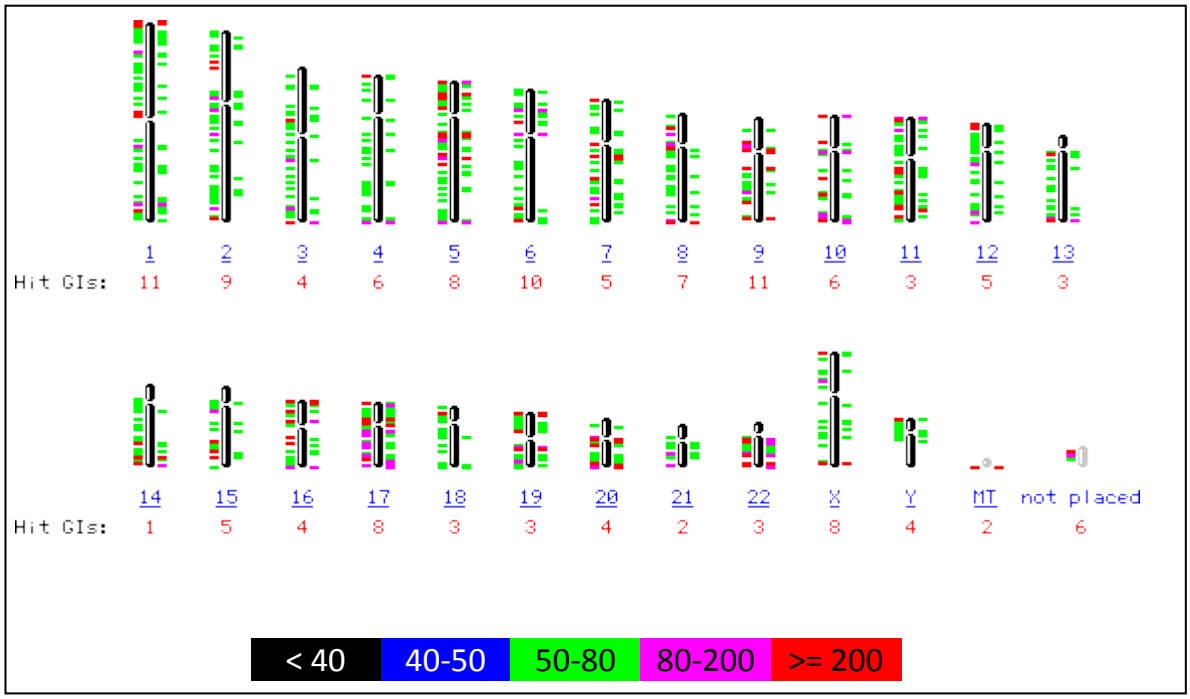


Figure 5.21: BLAST results for the 143B fragments shown on a map of human chromosomes. The colour table at the bottom of the figure is used to score the number of BLAST hits which are correlated with specific genomic positions, with higher numbers indicating greater accuracy.

Figure 5.21 shows a map of human chromosomes 1 through 22, X and Y. Included are BLAST hits in the mitochondrial DNA as well as sequence data which has not yet been placed at a

position on the NCBI genome database (Homeobox Database, 2010). The Hit GI tab refers to unique hits on the chromosome.

Colour markers show the position of the BLAST hit on the chromosome and the relative BLAST hit score. A colour key (given underneath the genomic map) is used to score the number of BLAST hits which are correlated with specific genomic positions. This score gives an indication of how well the specific BLAST hits were aligned to the reference sequence for each of the samples. Higher scores indicate better alignments, while lower scores indicate sample alignments which are less precise. Because these scores are normalized, one can compare the alignment scores between different experiments and different data sets (Madden, 2003)

Figure 5.21 shows that the fragments were derived from many positions in the genome, which means that methylation is present on many different positions in the chromosomes of the untransfected 143B cells.

5.3.4.2. BLAST results of transfected 143B fragments

A total of 143 methylated fragment sequences were analyzed using BLAST. An analysis was done using all available assemblies (reference and alternate) for the human genome. 339 BLAST hits were generated for the transfected 143B cell line sample (these include BLAST hits which are similar to each other, results on repeat areas of the reference genome sequence or multiple similar hits on different assemblies of the human genome).

Three of the BLAST results could be correlated with specific genes. These genes are derived from all assemblies of the genome and are given in table 5.17 (only those fragments with a BLAST score of 75 and above are included).

Table 5.17: List of BLAST hits associated with genes for the transfected 143B fragments

Total Score	Official Symbol	Official Full Name	Gene Type	Location	Additional
634	LOC100131754	hypothetical LOC100131754	miscRNA	1p36.33	
195	SLC6A10P	solute carrier family 6 (neurotransmitter transporter, creatine), member 10 (pseudogene)	pseudo	16p11.2	Similar to sodium- and chloride-dependent creatine transporter
195	LOC653562	sodium- and chloride-dependent creatine transporter 1-like	pseudo	16p11.2	

Table 5.17 shows the genes which can be associated with methylated fragments from the sample. Two of the identified genes are pseudogenes, and the other is a hypothetical gene.

Another BLAST analysis was done using only the reference sequence of the human genome. 70 unique BLAST hits could be associated with positions in the reference human genome sequence. These hits are illustrated in the figure 5.22, which shows the positions of the BLAST hits on a chromosomal view.

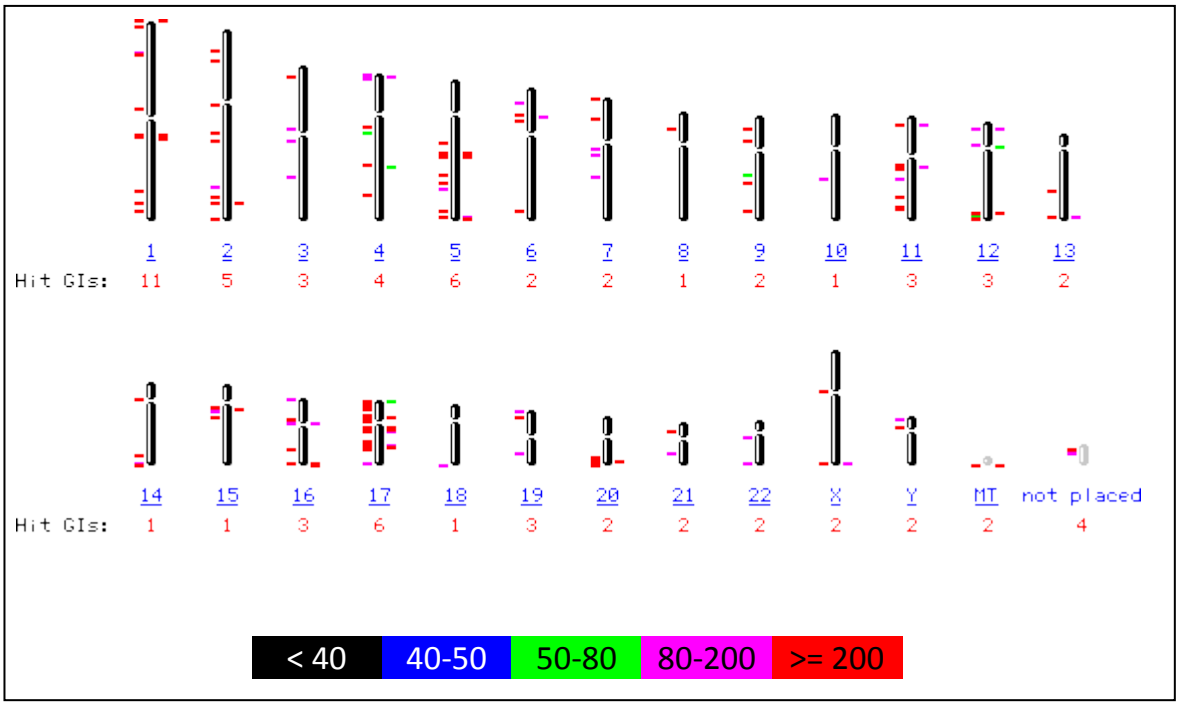


Figure 5.22: BLAST results for the 143B fragments shown on a map of human chromosomes. The colour table at the bottom of the figure is used to score the number of BLAST hits which are correlated with specific genomic positions, with higher numbers indicating greater accuracy.

Figure 5.22 shows that the fragments were derived from various positions in the genome. This indicates that methylation is present on many different positions in the chromosomes of the transfected 143B cells.

5.3.4.3. BLAST results of untransfected fibroblast fragments

A total of 10 methylated fragment sequences were analyzed using BLAST. An analysis was done using all available assemblies (reference and alternate) for the human genome. 121 BLAST hits were generated for the sample (these include BLAST hits which are similar or results on repeat areas of the reference genome sequence).

Twenty fragments could be correlated with specific genes. These genes are derived from all assemblies of the genome and were found in the double homeobox protein 4-like pseudogene. The high number of BLAST hits in this genomic position may be due to bias introduced by the PCR amplification steps.

Another BLAST analysis was done using only the reference sequence of the human genome. 14 unique BLAST hits could be associated with positions in reference human genome sequence. These hits are illustrated in the figure 5.23, which shows the positions of the BLAST hits on a chromosomal view.

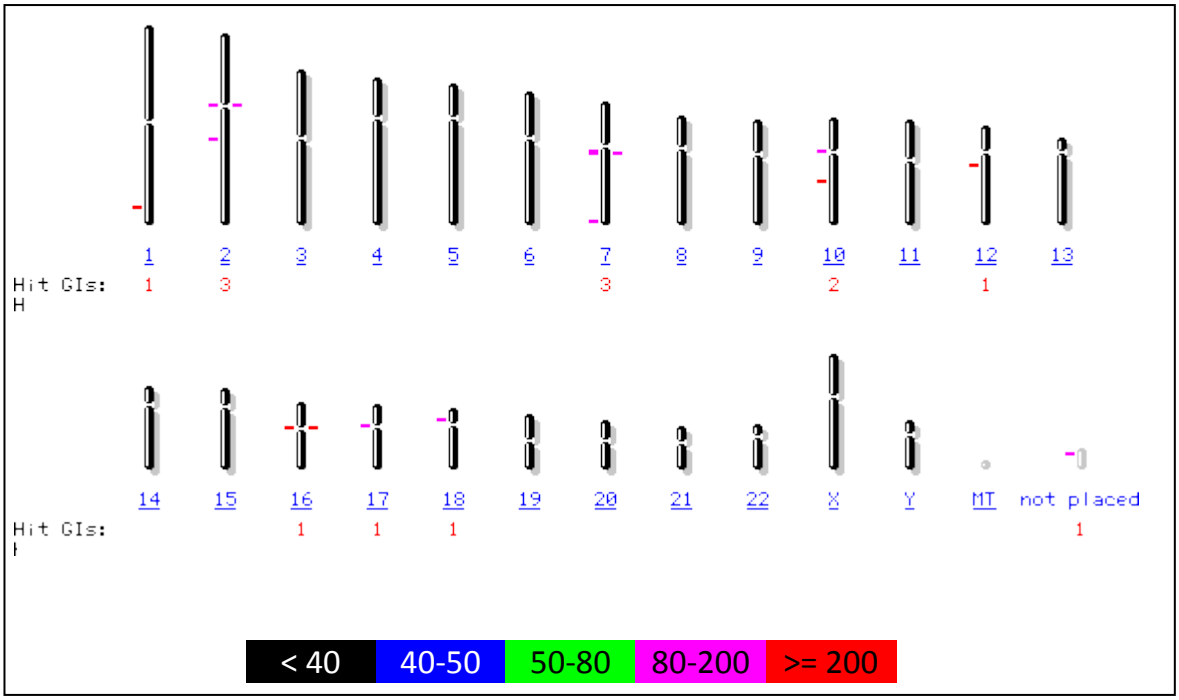


Figure 5.23: BLAST results for the fibroblast fragments shown on map of human genome. The colour table at the bottom of the figure is used to score the number of BLAST hits which are correlated with specific genomic positions, with higher numbers indicating greater accuracy.

Figure 5.23 indicates that the fragments are located on the first, second, seventh, tenth, twelfth, sixteenth, seventeenth and eighteenth chromosomes. Methylation is present on these positions in the genome, which may be correlated with centromeres and telomeres of specific chromosomes. This distribution of BLAST hits is visible due to the small amount of positioned fragments. Investigations into the reason for this may prove interesting for further development of this study. One hit could not be placed.

5.3.4.4. BLAST results of transfected fibroblast fragments

A total of 35 methylated fragment sequences were analyzed using BLAST. An analysis was done using all available assemblies (reference and alternate) for the human genome. 255 BLAST hits were generated for the sample (these include BLAST hits which are similar to each other, results on repeat areas of the reference genome sequence or multiple similar hits on different assemblies of the human genome).

Table 5.18: List of BLAST hits associated with genes for the transfected fibroblast fragments

Total Score	Official Symbol	Official Full Name	Gene Type	Location	Summary
303	DNM1P41	DNM1 pseudogene 41	pseudo	15q25.2	
209	LOC100134208	hypothetical LOC100134208	pseudo	Chromosome Unknown	
200	DUX3	double homeobox 3	protein coding	Unknown	Repeats of the 3.3-kb family in regions associated with heterochromatin. The DUX gene family, including DUX3, resides within these 3.3-kb repeated elements (Beckers *et al*., 2001 [PubMed 11245978]).
189	LOC100508734	ankyrin repeat domain-containing protein 57-like	pseudo	Chromosome Unknown	
178	LOC389834	ankyrin repeat domain 57 pseudogene	pseudo	Chromosome 4	
178	LOC642191	hypothetical protein LOC642191	pseudo	Chromosome 12	
93.5	LOC652802	hypothetical LOC652802	pseudo	Chromosome Unknown	

Seven fragments could be correlated with specific genes. These genes are derived from all assemblies of the genome and are given in table 5.18 (only those fragments with a BLAST score of 75 and above are included).

Table 5.18 shows the genes which can be associated with methylated fragments from the sample. Six of the identified genes were pseudogenes, and three of these pseudogenes were hypothetical. A protein coding gene, double homeobox 3, was also identified in the fragment samples. Furthermore, 58 other BLAST hits could be correlated to the double homeobox protein 4-like pseudogene. The high number of hits in the pseudogene may again be due to bias introduced by the PCR amplification steps, or it may indicate a state of hypermethylation in this genomic region. Another BLAST analysis was done using only the reference sequence of the human genome. 88 unique BLAST hits could be associated with positions in the reference human genome sequence. These hits are illustrated in the figure 5.24, which shows the positions of the BLAST hits on a chromosomal view.

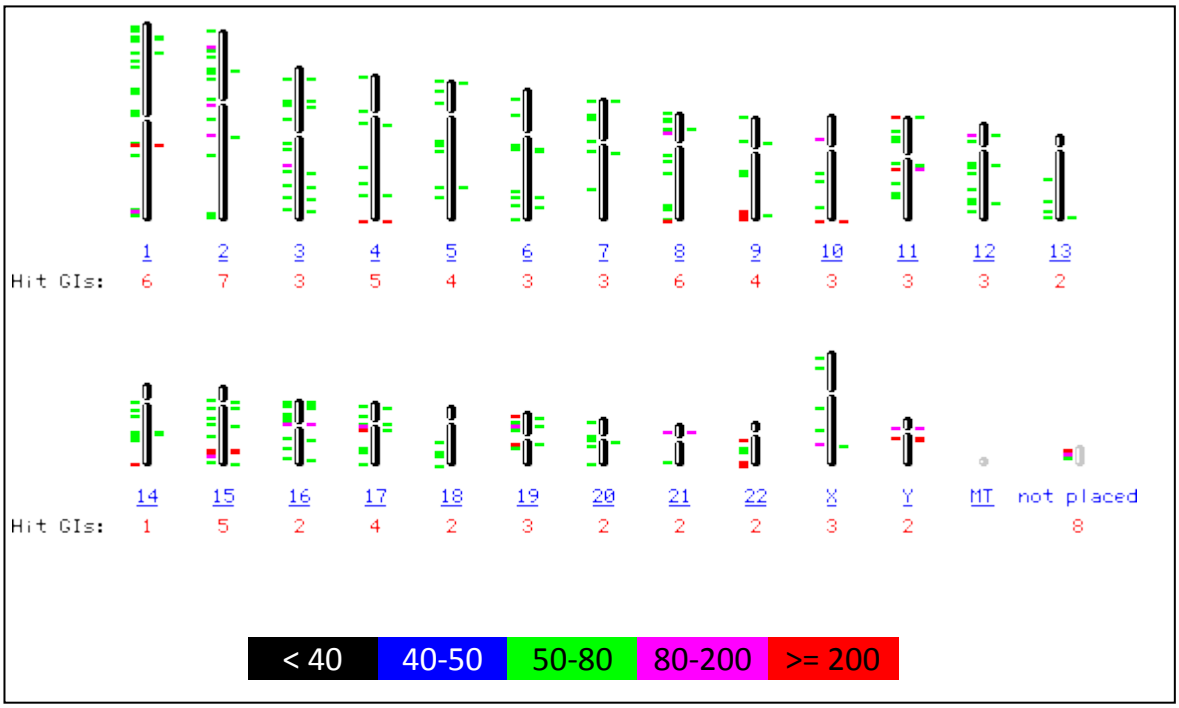


Figure 5.24: BLAST results for the transfected fibroblast fragments shown on a map of human chromosomes. The colour table at the bottom of the figure is used to score the number of BLAST hits which are correlated with specific genomic positions, with higher numbers indicating greater accuracy.

Figure 5.24 illustrates an even spread of the fragments across the genome, which means that methylation is present on many different positions in the chromosomes of the transfected fibroblast cells.

5.3.4.5. BLAST results of untransfected HepG2 fragments

A total of 207 methylated fragment sequences were analyzed using BLAST. An analysis was done using all available assemblies (reference and alternate) for the human genome. 396 BLAST hits were generated for the sample (these include BLAST hits which are similar to each other, results on repeat areas of the reference genome sequence or multiple similar hits on different assemblies of the human genome).

Ten fragments could be correlated with specific genes. These genes are derived from all assemblies of the genome and are given in table 5.19 on the next page (only those fragments with a BLAST score of 75 and above are included).

Table 5.19: List of BLAST hits associated with genes for the HepG2 fragments

Total Score	Official Symbol	Official Full Name	Gene Type	Location	Summary
339	NAIP	NLR family, apoptosis inhibitory protein	protein coding	5q13.1	The protein encoded by this gene is able to suppress apoptosis induced by various signals
211	FAM138D	family with sequence similarity 138, member D	miscRNA	12p13.33	
206	LOC100509396	tektin-4-like protein LOC389833-like	protein coding	Chromosome Unknown	
206	LOC100291646	tektin-4-like	pseudo	Chromosome Unknown	
202	PMS2P11	postmeiotic segregation increased 2 pseudogene 11	pseudo	7q11.23	
202 x 2	DTX2	deltex homolog 2 (Drosophila)	protein coding	7q11.23	DTX2 functions as an E3 ubiquitin ligase
167	LOC643650	hypothetical LOC643650	miscRNA	10q11.22	
145	LOC727805	putative POM121-like protein 1-like	pseudo	6p11.2	
128	LOC440820	similar to breakpoint cluster region isoform 1	protein coding	22q11.21	

Table 5.19 shows the genes which can be associated with methylated fragments from the sample. The results include three pseudogenes, one hypothetical gene and three protein coding genes. 27 additional hits were found in the double homeobox protein 4-like pseudogene. The high number of hits in the pseudogene may be due to bias introduced by the PCR amplification steps.

Another BLAST analysis was done using only the reference sequence of the human genome. 100 unique BLAST hits could be associated with positions in the reference human genome sequence. These hits are illustrated in the figure 5.25, which shows the positions of the BLAST hits on a chromosomal view.

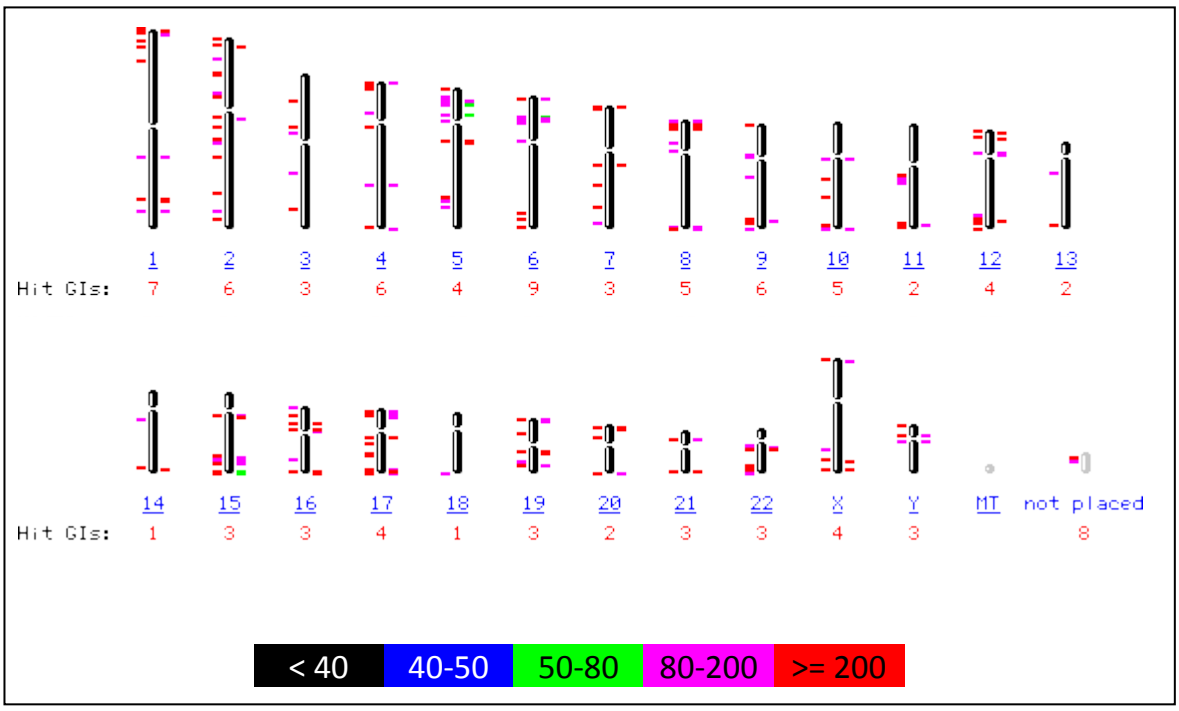


Figure 5.25: BLAST results for the HepG2 fragments shown on a map of human chromosomes. The colour table at the bottom of the figure is used to score the number of BLAST hits which are correlated with specific genomic positions, with higher numbers indicating greater accuracy.

Figure 5.25 again shows that the fragments were derived from many positions in the genome, which means that methylation is present on many different positions in the chromosomes of the untransfected HepG2 cells.

5.3.4.6. BLAST results of transfected HepG2 fragments

A total of 143 methylated fragment sequences were analyzed using BLAST. An analysis was done using all available assemblies (reference and alternate) for the human genome. 431 BLAST hits were generated for the sample (these include BLAST hits which are similar to each other, results on repeat areas of the reference genome sequence or multiple similar hits on different assemblies of the human genome).

Two fragments could be correlated with specific genes. These genes are derived from all assemblies of the genome and are given in table 5.20 (only those fragments with a BLAST score of 75 and above are included).

Table 5.20: List of BLAST hits associated with genes for the transfected HepG2 fragments

Total Score	Official Symbol	Official Full Name	Gene Type	Location	Summary
456	USP17L2	ubiquitin specific peptidase 17-like 2	protein coding	8p23.1	DUB3 is a member of the ubiquitin processing protease (UBP) subfamily of deubiquitinating enzymes.
219	LOC392196	ubiquitin specific peptidase 17-like 2 pseudogene	pseudo	8p23.1	

Table 5.20 shows the genes which can be associated with methylated fragments from the sample. One of the genes is a pseudo gene and the other is an ubiquitin specific peptidase 17-like 2 protein coding gene.

Another BLAST analysis was done using only the reference sequence of the human genome. 91 unique BLAST hits could be associated with positions in the reference human genome

sequence. These hits are illustrated in the figure 5.26, which shows the positions of the BLAST hits on a chromosomal view.

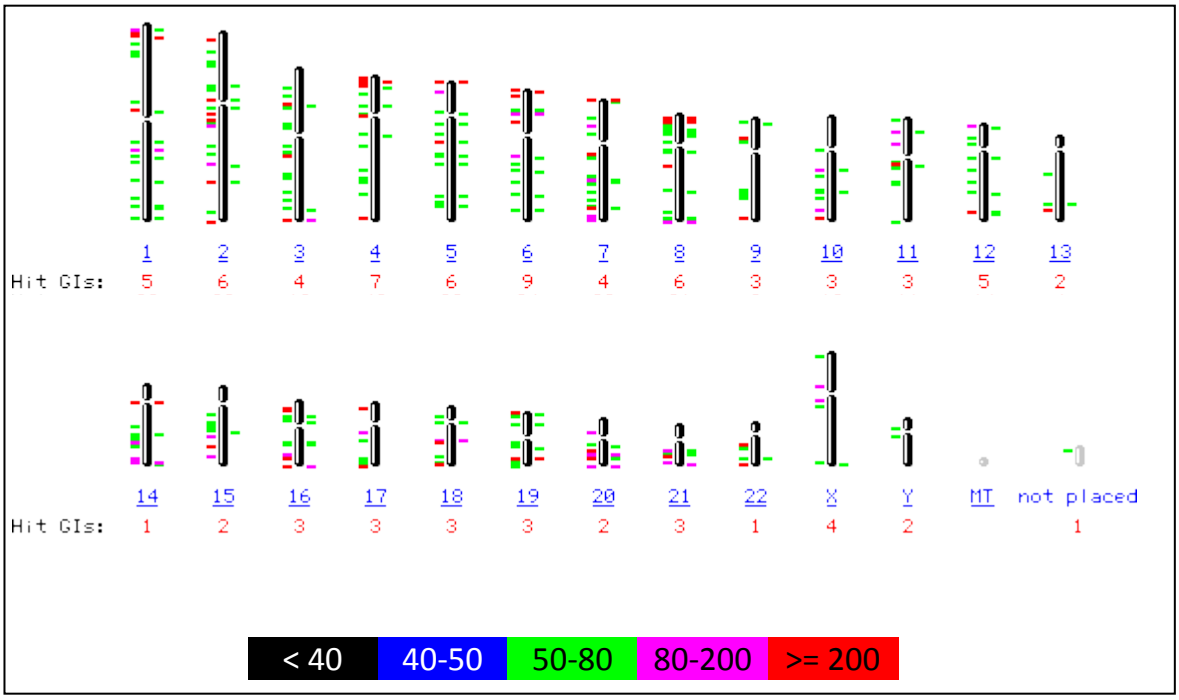


Figure 5.26: BLAST results for the transfected HepG2 fragments shown on a map of human chromosomes. The colour table at the bottom of the figure is used to score the number of BLAST hits which are correlated with specific genomic positions, with higher numbers indicating greater accuracy.

Figure 5.26 shows that the fragments were derived from various positions in the genome; DNA methylation is present on many different positions for the transfected HepG2 sample.

5.3.4.7. BLAST results of untransfected HeLa fragments

A total of 16 methylated fragment sequences were analyzed using BLAST. An analysis was done using all available assemblies (reference and alternate) for the human genome. 47 BLAST hits were generated for the sample (these include BLAST hits which are similar to each other, results on repeat areas of the reference genome sequence or multiple similar hits on different assemblies of the human genome).

Another BLAST analysis was done using only the reference sequence of the human genome. No fragments could be correlated with specific genes. This is probably due to the low quality of fragment assembly and low number of BLAST hits derived for matching with known genes.

Another BLAST analysis was done using only the reference sequence of the human genome. 16 unique BLAST hits could be associated with positions in reference human genome sequence. These hits are illustrated in the figure 5.27, which shows the positions of the BLAST hits on a chromosomal view.

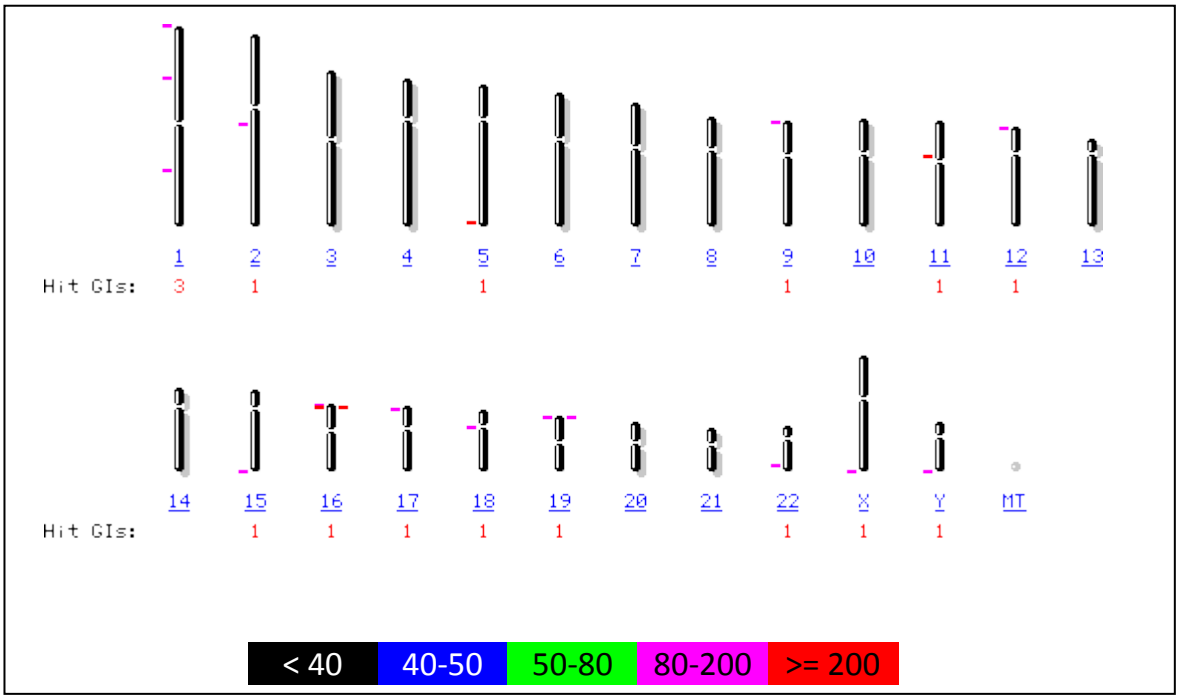


Figure 5.27: BLAST results for the HeLa fragments shown on a map of human chromosomes. The colour table at the bottom of the figure is used to score the number of BLAST hits which are correlated with specific genomic positions, with higher numbers indicating greater accuracy.

Figure 5.27 shows fragments in the first, second, fifth, ninth, eleventh, twelfth, fifteenth, sixteenth, seventeenth, eighteenth, nineteenth, twenty-second and X and Y chromosomes. It is difficult to draw any conclusions from the chromosome map results for the untransfected HeLa cell line samples, as a low number of BLAST hits could be derived from analyses. However, methylated fragments are predominantly localized in centromeres and telomeres of

specific chromosomes. This distribution of BLAST hits is again visible due to the small amount of fragments which could be positioned on specific chromosomes.

5.3.4.8. BLAST results of transfected HeLa fragments

A total of 231 methylated fragment sequences were analyzed using BLAST. An analysis was done using all available assemblies (reference and alternate) for the human genome. 667 BLAST hits were generated for the sample (these include BLAST hits which are similar to each other, results on repeat areas of the reference genome sequence or multiple similar hits on different assemblies of the human genome).

Three fragments could be correlated with specific genes. These genes are derived from all assemblies of the genome and are given in table 5.21 (only those fragments with a BLAST score of 75 and above are included).

Table 5.21: List of BLAST hits associated with genes for the transfected HeLa fragments

Total Score	Official Symbol	Official Full Name	Gene Type	Location	Summary
294	OPLAH	5-oxoprolinase (ATP-hydrolysing)	protein coding	8q24.3	
97.1 + 95.3	DUX3	double homeobox 3	protein coding	Unknown	The human genome contains hundreds of repeats of the 3.3-kb family in regions associated with heterochromatin. The DUX gene family, including DUX3, resides within these 3.3-kb repeated elements

Table 5.21 shows that fragments could be correlated with three protein coding genes. These genes are the 5-oxoprolinase (ATP-hydrolysing) protein coding gene and two hits on the

double homeobox 3 protein coding gene. 53 hits further hits were found in the double homeobox protein 4-like pseudogene. The high number of hits in the pseudogene may again be due to bias introduced by the PCR amplification steps, or it may indicate a state of hypermethylation in this genomic region.

Another BLAST analysis was done using only the reference sequence of the human genome. 87 unique BLAST hits could be associated with positions in reference human genome sequence. These hits are illustrated in the figure 5.28, which shows the positions of the BLAST hits on a chromosomal view.

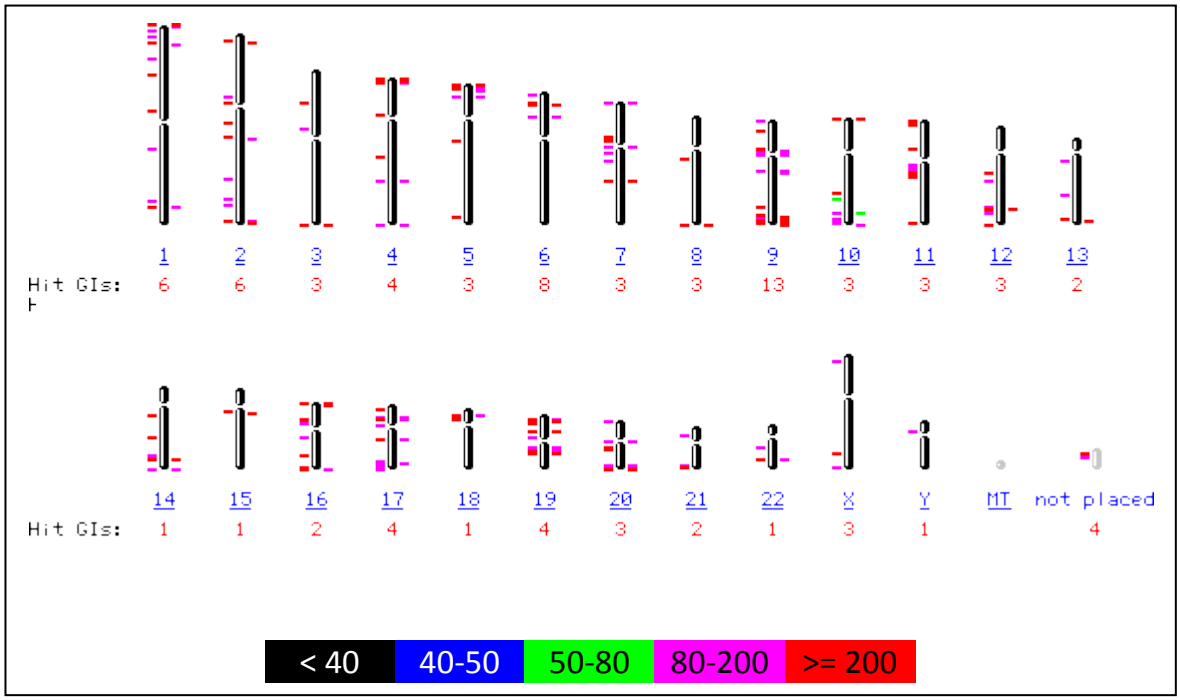


Figure 5.28: BLAST results for the transfected HeLa fragments shown on a map of human chromosomes. The colour table at the bottom of the figure is used to score the number of BLAST hits which are correlated with specific genomic positions, with higher numbers indicating greater accuracy.

Figure 5.28 shows the positions of the fragments on various chromosomes. As with previous examples, a spread of fragments across all chromosomes was seen.

5.3.4.9. Summary

Only a small number of genes could be identified by means of BLAST analysis. The fragment data may not be representative of the entire methylome and several genes may therefore have been missed due to PCR amplification bias or the small amount of sample DNA generated with the MethylMiner kit. There may also have been flaws in the DNA sequencing due to these factors.

Correlation of fragments to positions on the chromosomes was successful, however, and the genomic positions of a large number of fragments could be mapped. These positions are spread out across the entire genome. Distribution of fragments around centromeres and telomeres were seen in the untransfected fibroblast and untransfected HeLa cells. Though this observation is based on only a small number of generated fragments, it would be interesting to investigate this phenomenon further.

5.4. Chapter Summary

Before high-throughput GS FLX Titanium sequencing was done, a series of sample preparation steps were completed. These involved the fragmentation of DNA by means of restriction enzymes, enrichment with the MethylMiner kit and ligation of the samples to the pEZseq vector for PCR amplification.

There were several differences between the four cell lines and their transformed counterparts in terms of the quality of sequencing results generated. In some cases, this may have been due to the amount of DNA provided for the sequencing. These shortcomings were kept in mind throughout the data analysis of the samples.

Preliminary data analysis steps were begun for each sample after DNA sequencing was completed. *De novo* assembly of each sample was done to generate contiguous sequences before further data analysis. These contiguous sequences, as well as generated consensus

sequences for the fragments of each sample, were used for analysis of fragment length, CG-content analysis and NCBI BLAST analysis.

Fragment length analysis showed that the fragments varied in sizes of roughly 50bp to 600bp. These results matched gel photos taken of the samples before they were sent for sequencing.

CG-content analyses showed that the percentage of CG-dinucleotides differed between different cell lines, but was the same for untransfected and transfected samples of the same cell line. The insertion of a vector does not alter the underlying DNA sequence where DNA methylation may be located. The fact that these variations exist between cell lines suggests that CG percentages of all samples will remain roughly the same for the same cell lines.

BLAST results showed that several genes could be identified for each of the samples. Although little correlation could be made between the different identified genes of the various cell lines, the main goal of the study (to identify possibly methylated genes) was achieved. Reduced representation was therefore used to identify several genes which could be investigated in further studies. The double homeobox protein 4-like pseudogene was particularly overrepresented in several samples, which makes this an interesting avenue for further study.

BLAST hits for the fragments were distributed across the entire genome and all chromosomes for most samples (all excluding untransfected fibroblast and untransfected HeLa cells). This correlates with the results of Michael Weber *et al*., who used DNA microarray analysis to determine sites of differential methylation in normal and transformed human cells. Results of that study showed that differentially methylated regions were spread across the entire genome of the investigated cells (Weber *et al*., 2005). These results are similar to those seen in this study when mapping fragments to specific chromosome-positions in the genome.

Chapter 6

Summary and conclusion

The elucidation of the chemical structure of DNA heralded a major paradigm shift in the world of molecular biology (Watson and Crick, 1953; Olby, 2003). Epigenetics, particularly DNA methylation studies, now have the potential to do the same. DNA methylation plays a role in several biological functions, most notably that of gene expression regulation (Esteller, 2008; Laird, 2010). There are several endogenous and exogenous factors that affect DNA methylation (see section 2.6) and various methods to measure the alterations in DNA methylation caused by these factors, with new technologies allowing more thorough investigations into the methylome.

Whilst there are many techniques available to investigate DNA methylation, each with various positive and negative aspects, the technique chosen in this study was based on the enrichment of the methylated fraction of fragmented DNA ("mechanical" separation via magnetic beads) and subsequent high-throughput DNA sequencing. By using this technique, new insights into the DNA methylation amounts and distribution within specific samples could be achieved.

It has been noted that the insertion of a vector into a cell line causes alterations on many levels, such as producing unexpected proteins (Chaudry, 2004). These alterations include changes in the amounts and patterns of DNA methylation in specific transformed cell lines (Kok, 2009). In this study, the effect of transformation by means of a vector inserted into several mammalian cell lines was investigated. These cell lines were the 143B cells, fibroblasts, HepG2 cells and HeLa cells.

The first step of this investigation was to verify the observations that there were alterations in the DNA methylation levels of cell cultures once a vector was inserted. This was done in a pilot study using the cytosine-extension assay (CEA) and real-time methylation-specific PCR (real-time MSP) to determine DNA methylation amounts in cells on a global and gene-specific level respectively.

The CEA results (section 4.2) showed that there were indeed alterations in the amount of DNA methylation between untransformed cells and transformed cells of the same cell

lines. There were also DNA methylation differences seen when comparing different cell lines. Previous studies in our laboratory indicated that a state of hypomethylation was observed with the insertion of a vector into the 143B cell line (Kok, 2009). In the current study, four different cell lines were investigated. This revealed that the insertion of a vector caused a state of hypomethylation in some cells (such as the 143B and HepG2 cell lines) and a state of hypermethylation in others (the HeLa cell line). A study done by Laurence Wild and James Flanagan showed that transformation of a cell line may cause genome-wide hypomethylation in some cells, but that this state of hypomethylation was not consistently seen in all cell lines (such as cancer cells) (Wild and Flanagan, 2010). Their conclusions thus support results observed in this study.

The real-time MSP results (section 4.3) showed that the insertion of a vector into the 143B cell line caused DNA methylation alterations in the promoter region of the *MGMT*-gene and the $p16^{INK4a}$-gene. The *MGMT*-gene showed a decrease in DNA methylation with the insertion of a vector, whilst the $p16^{INK4a}$-gene showed various alterations in the DNA methylation amounts with the insertion of a vector. The results of the real-time MSP experiments, coupled with the CEA data, showed that the insertion of a vector into a cell line caused alterations in the DNA methylation levels on both a gene-specific and global level.

With this preliminary investigative work done, the main examination of the effect of vector-insertion into a cell line on DNA methylation could begin using MethylMiner enrichment and high-throughput DNA sequencing. Preparation of samples for DNA sequencing involved several optimization steps prior to sending the samples for high-throughput GS FLX Titanium sequencing (section 5.2).

The data was analyzed by various means once DNA sequencing was completed (section 5.3). The biological data analysis computer program, CLC Bio, was used for *de novo* assembly of generated fragments and for basic analysis of generated sequencing data. The assembly showed that a relatively low number of different fragments was collected by the MethylMiner enrichment for each sample. A rather disappointing outcome of this

project was the low number of specific genes amongst the sequenced fragments. Even though this study was planned as a reduced representation approach, the aim was to identify and characterize specific sequences (genes) involved in the epigenetic changes induced by transformation of eukaryotic cells.

Further data analysis was done using Excel programming and chart design. This produced data on sequence fragment length and CG-content of fragments. Fragment sizes varied between 50 bp and 600 bp, with an average fragment length of 202.7 bp. This corresponds with size preparation of the genomic DNA with the specific restriction enzymes. CG-content analyses showed that the percentages of CG-dinucleotides in fragments were similar for the transformed and untransformed counterparts of the same cell line, but differed for different cell lines. It was also seen that, in longer fragments, the amount of CG-dinucleotides were more than in shorter fragments. This suggests that enrichment with the MethylMiner kit could enrich sample fragments irrespective of high or low amounts of CG-dinucleotides present, which shows the suitability of this type of selection for studies similar to this one.

Comparisons are often made between the methyl-DNA immunoprecipitation, or MeDIP, and the selection technique used in this study (MethylMiner enrichment). The MethylMiner kit provides increased sensitivity in comparison to MeDIP, with 5x more hits and 16x more unique hits observed for the MethylMiner kit (Jia *et al*., 2010, Prediger (ed.), 2010, Yu *et al*., 2010). Furthermore, the use of MeDIP is dependent on the density of methylated CpG sites, which vary within different genomic regions. This factor makes it impossible to estimate absolute methylation levels using MeDIP, due to the confounding influence of differing CpG density in the genome (Down *et al*., 2008). However, the results of fragment length analysis done in this study showed that MethylMiner kit enrichment could be used to enrich methylated DNA fragments, irrespective of high or low amounts of methylated CpG sites. This makes the approach used in this study much more viable for the accurate investigation of DNA methylation patterns than MeDIP approaches.

Final data analysis steps were done using NCBI BLAST to correlate fragments for each sample with reference genome data. These analyses led to the correlation of methylated regions with specific genomic positions and to the identification of some genes of the cell line which were methylated, as already mentioned. However, the role of the identified genes in DNA methylation and in epigenetic events were not obvious.

The number of identified genes was fewer than expected. It might be deduced that only a few genes of a particular cell line are methylated. Considering the large size of the human genome and the relatively small number of genes, it is possible that DNA methylation in this particular experimental setup is only present in non-gene regions. This is unlikely, however, as DNA methylation is known to be present in a wide range of gene promoter regions in a number of different genes (Esteller, 2008).

Due to the observed global hypomethylation in some transformed cell lines, the small number of identified methylated genes may be due to a decrease of DNA methylation amounts in gene-regions. However, the low number of identified genes also seen in the untransformed samples contradicts the possibility of this explanation, as well as the fact that some cell lines showed an increase in DNA methylation following vector-insertion. A more likely possibility is that the low number of assembled fragments did not give a broad view of the methylome. This may be due to PCR amplification bias of methylated fragments or due to the low amounts of samples generated via MethylMiner enrichment. It was seen in some samples that the methylated fragments were derived mostly from the centromeric and telomeric regions of the chromosome. A further investigation into this phenomenon is suggested, with the belief that more thorough information of a cell line's methylome will yield interesting insights into the reason for this distribution of methylated fragments. The overall low number of identified genes, however, suggests that the generated fragments did not give an accurate view of the entire methylome and was representative of only a small portion thereof.

Fragments could be mapped to genomic positions using BLAST analysis. These results showed that fragments were distributed across the entire genome (section 5.3.4), which

correlates with results generated in a similar study done by another group using a DNA microarray approach (Weber *et al*., 2005).

The investigation of the cell line transformation is particularly important in cell culture studies and this study investigated this technical aspect of laboratory work. The insertion of a vector has several influences on the functioning in a particular cell culture, and cell cultures may even grow for longer or shorter periods after transformation (Chaudry, 2004). This may be associated with the fact that DNA methylation plays a role in cellular aging (Wilson *et al*., 2007) and alterations in the DNA methylation patterns may have an effect on how long cell cultures continue to grow.

The importance of this study is also underlined by the fact that transformed cell lines often produce unexpected proteins or metabolites due to alterations in cellular characteristics, such as the methylome (Chaudry, 2004). This might be attributable to DNA methylation alterations in the promoter regions of certain genes, thereby turning these genes “on” or “off” or otherwise controlling gene activity (Esteller, 2008). It should also be noted that DNA methylation represents a level of control for certain tissue-specific genes (Laird 2005), which is of importance in cell line studies which focus on tissue-specific gene effects, and gene transcription is also affected by means of the alterations of protein-DNA interactions when DNA is methylated (Razin and Cedar, 1991). Because epigenetic mechanisms provide stability to the phenotype (Laird, 2010), any alterations in these DNA methylation patterns due to vector-insertion may cause instability of chromosomal regions.

Biotechnological applications and potential gene therapy may also require careful investigation of the DNA methylation alterations caused by vector-insertion into cells. Especially the safety of gene therapy for use in clinical applications may be brought into question if unknown cellular characteristics are induced when cells are transformed by the insertion of an expression vector into such a cell.

Furthermore, it has been observed that DNA methylation plays a role in disease-development, such as cancer, and that epigenetic deregulation is involved in almost all disease states (Wilson *et al*., 2007; Gargiulo and Minucci, 2009). Alterations in DNA methylation are common in tumours (Das and Singal, 2004) due to the lack of DNA methylation's stabilizing effect (Esteller, 2008). An extension of this study may entail the examination of whether these same effects are present *in vivo* in transformed cells. With more in-depth studies, it is conceivable that alterations in the methylome of the cells may cause disease-like effects (either in the cells themselves or *in vivo*).

Das and Singal suggest that investigations into the complex mechanisms of epigenetics, such as DNA methylation, could potentially lead to therapeutic targets for the treatment of several disease states (Das and Singal, 2004). This study forms part of an array of studies to better understand this epigenetic phenomenon with the possible result of clinical application and, with further investigation and data analysis, lends itself to publication in article form. Comparisons of the fragment sequences of the untransformed and transformed cell lines may yield interesting results.

It is suggested that sodium-bisulfite treatment and subsequent DNA sequencing analysis should be done to provide a different set of results in follow-up studies. This analysis may be compared with the results of the current study to produce a more accurate view of methylome alterations with the insertion of a vector into a cell line.

Some other aspects of the study could also be modified to generate more results in a follow-up study. It is expected that methylated fragments that are more representative of the entire methylome could be generated if the vector ligation and PCR amplification of the enriched methylated samples is not done. This would negate fears of PCR bias and would produce more accurate sequencing results for data analysis, but would mean that several MethylMiner kits are required for sample generation. This has financial implications, which may be problematic depending on the funding of the project.

A broader study of the alterations in DNA methylation patterns induced by cellular transformation is also suggested, particularly using a variety of different methods and in-depth investigation of the methylome and metabolism of both transformed and untransformed cells. It is also expected that the use of fewer cell lines, more MethylMiner enrichment steps and no PCR amplification of fragments in a follow-up study may produce interesting results that could give new insights into DNA methylation of transformed cell lines.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002) Molecular Biology of the Cell (4th ed.). Garland Science, p. 205-207.

Anon. (2004) Human Tissue Act. Human Tissue Authority. http://www.hta.gov.uk/legislationpoliciesandcodesofpractice/legislation/humantissueact.cfm. Date accessed: 12 August 2010.

Anon. (2007) Relative Quantitation Using Comparative C_T Getting Started Guide. *Applied Biosystems*, 1-100.

Beckers, M., Gabriels, J., Van der Maarel, S. (2001). "Active genes in junk DNA? Characterization of DUX genes embedded within 3.3 kb repeated elements." Cell (264): 51 - 7

Bernstein, B. E., A. Meissner, Lander, E.S. (2007). "The Mammalian Epigenome." Cell (128): 669–681.

Bird, A. (2002). "DNA methylation patterns and epigenetic memory." Genes & Development 16: 6-21.

Brabender, J., Usadel, H. (2003) "Quantitative *O6-Methylguanine DNA Methyltransferase* Methylation Analysis in Curatively Resected Non-Small Cell Lung Cancer: Associations with Clinical Outcome" Clinical Cancer Research *Vol. 9, 223–227,*

Chaudry, A. (2004) Cell Culture. The Science Creative Quarterley. http://www.scq.ubc.ca/cell-culture/. Date accessed: 24 June 2010

Das, P. M. and R. Singal (2004). "DNA Methylation and Cancer." Journal of Clinical Oncology 22(22): 4632-4642.

Down, T.A., Rakyan, V.K., , Turner D.J., Flicek P., Li H., Kulesha E., Gräf S., Johnson N., Herrero J., Tomazou E.M., Thorne N.P., Bäckdahl L., Herberth M., Howe K.L., Jackson D.K., Miretti

M.M., Marioni J.C., Birney E., Hubbard T.J., Durbin R., Tavaré S., Beck S. (2008) "A Bayesuan deconvolution strategy for immunoprecipitation-based DNA methylome analysis." Nature Biotechnology 26(7): 779-785

Duffy M.J., Napieralski R., Martens J.W., Span P.N., Spyratos F., Sweep F.C., Brunner N., Foekens J.A., Schmitt M. (2009). "Methylated genes as new cancer biomarkers". European Journal of Cancer 45(3):335-46.

Du Toit, J. (2009). Honours report: Standardisation of a Real-time methylation-specific PCR method for quantitative analysis of DNA methylation of the *p16*^{INK4a}-gene promoter. School for physical and chemical sciences, Division of Biochemistry, North-West University, Potchefstroom campus, Potchefstroom, South Africa

Ehrlich, M., Gama Sosa, M.A., Huang L.H., Midgett R.M., Kuo K.C., McCune R.A., Gehrke C. (1982). "Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells". *Nucleic Acids Research* 10 (8): 2709–2721.

Esteller, M. (2008). "Molecular origins of cancer: Epigenetics in cancer." The New England Journal of Medicine 358(11): 1148-1159.

Gardiner-Garden M. and Frommer M. (1987). "CpG islands in vertebrate genomes." Journal of Molecular Biology 196(2):261-82.

Gargiulo, G. and S. Minucci (2009). "Epigenomic profiling of cancer cells." The International Journal of Biochemistry & Cell Biology(41): 127–135.

Hayatsu, H. (2008) "The bisulfite genomic sequencing used in the analysis of epigenetic states, a technique in the emerging environmental genotoxicology research" Mutation Research/Reviews in Mutation Research Volume 659, Issues 1-2, Pages 77-82

Hernández, S., Gómez, A., Cedano, J., Querol, E. (2009). "Bioinformatics annotation of the hypothetical proteins found by omics techniques can help to disclose additional virulence factors". *Current Microbiology* 59 (4): 451–6.

Homeobox Database (2010) Locus information on *LOC100506764.* http://homeodb.zoo.ox.ac.uk/gene_info.get?spf=f&spfm=PRD&sbfm=Dux&id=9531&og=Human. Accessed on: 10 December 2010.

Jia, J., Pekowska, A., Jaeger S., Benoukraf T., Ferrier P., Spicuglia S. (2010) "Assessing the efficiency and significance of Methylated DNA Immunoprecipitation (MeDIP) assays in using *in vitro* methylated genomic DNA." BioMed Central Research Notes 3: 240

Jones, P.A. and Liang, G. (2009). "Rethinking how DNA methylation patterns are maintained. Nature Reviews Genetics – Advanced Online Publication.

Kok, D. (2009). Masters Dissertation: Measuring the influence of ETC complex I and III knockdown on global and gene-specific DNA methylation. School for physical and chemical sciences, Division of Biochemistry, North-West University, Potchefstroom campus, Potchefstroom, South Africa

Laird, P. W. (2005). "Cancer epigenetics." Human Molecular Genetics 14(Review Issue 1): R56-R76.

Laird, P. W. (2010). "Principles and challengs of genome-wide DNA methylation analysis." Nature Reviews/Genetics 11: 191-203.

Lee, S., Lee, H. J., Kim, J., Lee, H., Jang, J.J., and Kang, G. H. (2003). "Aberrant CpG Island Hypermethylation Along Multistep Hepatocarcinogenesis." American Journal of Pathology 163(4): 1371-1378.

Litzkas, P., Jha, K.K., Ozer, H.L. (1984) “Efficient transfer of cloned DNA into human diploid cells: protoplast fusion in suspension.” *Molecular Cellular Biology* pp. 2549–2552.

Madden, T. (2003) The BLAST Sequence Analysis Tool (Chapter 16) from The NCBI Handbook [Internet] (McEntyre, J. and Ostell, J. editors). http://www.ncbi.nlm.nih.gov/books/NBK21097/. Accessed on 2 December 2010.

Mighell, A.J., Smith, N.R., Robinson, P.A., Markham, A.F. (2000). "Vertebrate pseudogenes". *FEBS Lett.* 468 (2-3): 109–14.

Mitchell, G. A., Grompe, M. (2001). "Hypertyrosinemia (In Scriver, C.R., Beaudet, A.L., Sly, W.S., Valle, D., eds. The Metabolic and Molecular Basis of Inherited Disease. 8th ed. New York: McGraw-Hill. p. 1777)."

Nephew, K.P. and Huang, T.H. (2003). “Epigenetic gene silencing in cancer initiation and progression”. Cancer Letters 190: 125-133

New England Biolabs (2010) *MspI* restriction enzyme details. http://www.neb.com/nebecomm/products/productR0106.asp. Accessed on: 17 November 2010.

Oakeley, E.J., Chiang , P.K. (ed) (1999) “DNA methylation analysis: a review of current methodologies.” Pharmacology & Therapeutics 84: 389-400.

Okochi-Takada, E., Ichimura, S., Kaneda, A., Sugimura, T., Ushijima, T. (2004). "Establisment of a detection system for demethylating agents using an endogenous promoter CpG island." Mutation Research 568: 187-194.

Olby, R. (2003). "Why celebrate the golden jubilee of the double helix?" Endeavour 27(2).

Prediger (ed.) (2010) “Methylated DNA Enrichment Simplifies Study of Methylomes. Applied Biosystems Innovations, March 2010. Issue 14.

Prokhortchouk, E. and P.-A. Defossez (2008). "The cell biology of DNA methylation in mammals." Biochimica et Biophysica Acta 1783: 297-2173.

Razin, C. and H. Cedar (1991). "DNA Methylation and Gene Expression." Microbiolical Reviews 55(3): 451-458.

Razin A and Shemer R. (1995) “DNA methylation in early development.” Human Molecular Genetics 4 Spec No:1751-5.

Roth, M. J., C. C. Abnet. (2006). "*p16, MGMT, RARB2, CLDN3, CRBP* and *MT1G* gene methylation in esophageal squamous cell carcinoma and its precursor lesions." Oncology Reports **15**: 1591-1597.

Tucker, K.L. (2001). "Methylated cytosine and the brain: a new base for neuroscience". *Neuron* 30 (3): 649–652.

Upile, T., W. Jerjes, Kafas, P., Singh, S.U., Mahil, J., Sandison, A., Hopper, C., Sudhoff, H. (2009). "Ethical and technical considerations for the creation of cell lines in the head & neck and tissue harvesting for research and drug development (Part II): Ethical aspects of obtaining tissue specimens." International Archives of Medicine **2**(9): 5.

Van Heerden, C. (2006) Masters Dissertation: Establishing a method for measuring the DNA methylation status of specific human genes. School for physical and chemical sciences, Division of Biochemistry, North-West University, Potchefstroom campus, Potchefstroom, South Africa

Watson, J. D. and F. H. C. Crick (1953). "Molecular Structure of Nucleic Acids (A Structure for Deoxyribose Nucleic Acid)." Nature(4356): 737-738.

Wentzel, J.F. Gouws, C., Huysamen, C., Dyk, E., Koekemoer, G., Pretorius, P.J. (2010) “Assessing the DNA methylation status of single cells with comet assay.” Analytical Biochemistry 400: 190-194

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., Schübeler, D. (2005) "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed cells." Nature Genetics 37: 8

Wiedmann, R.T., Smith, P.L.T., Nonneman, D.J. (2008). “SNP discovery in swine by reduced representation and high throughput pyrosequencing.” BioMed Central Genetics. 9:81

Wild, L. and Flanagan, J.M. (2010) "Genome-wide hypomethylation in cancer may be a passive consequence of transformation" Biochim. Biophys. Acta, doi:10.1016?j.bbcan.2010.03.003.

Wilson, A.S., Power, B.E., Molley, P.L. (2007). "DNA hypomethylation and human diseases." Biochimica et Biophysica Acta **1775**(1): 138-162.

Wilson, K. And Walker, J. (2005) Principles and Techiques of Biochemistry and Molecular Biology. Cambridge University Press: New York

Yu, Y., Blair, S., David, G., Randy, J., David, M., Ahmed, B.H., Indraneel, G., Alexander, C. (2010) “Direct DNA Methylation Profiling Using Methyl Binding Domain Proteins.” Analytical Chemistry 82: 5012-5019

Appendices

Appendix A: Verification of MethylMiner enrichment

After each cell line was enriched, a validation of the separation of methylated DNA from unmethylated DNA was done. This involved PCR amplification of a representative portion of both the methylated and unmethylated fraction (controls). The figure below (figure A.1) shows the results of the 143B MethylMiner verification (using controls included in the kit):

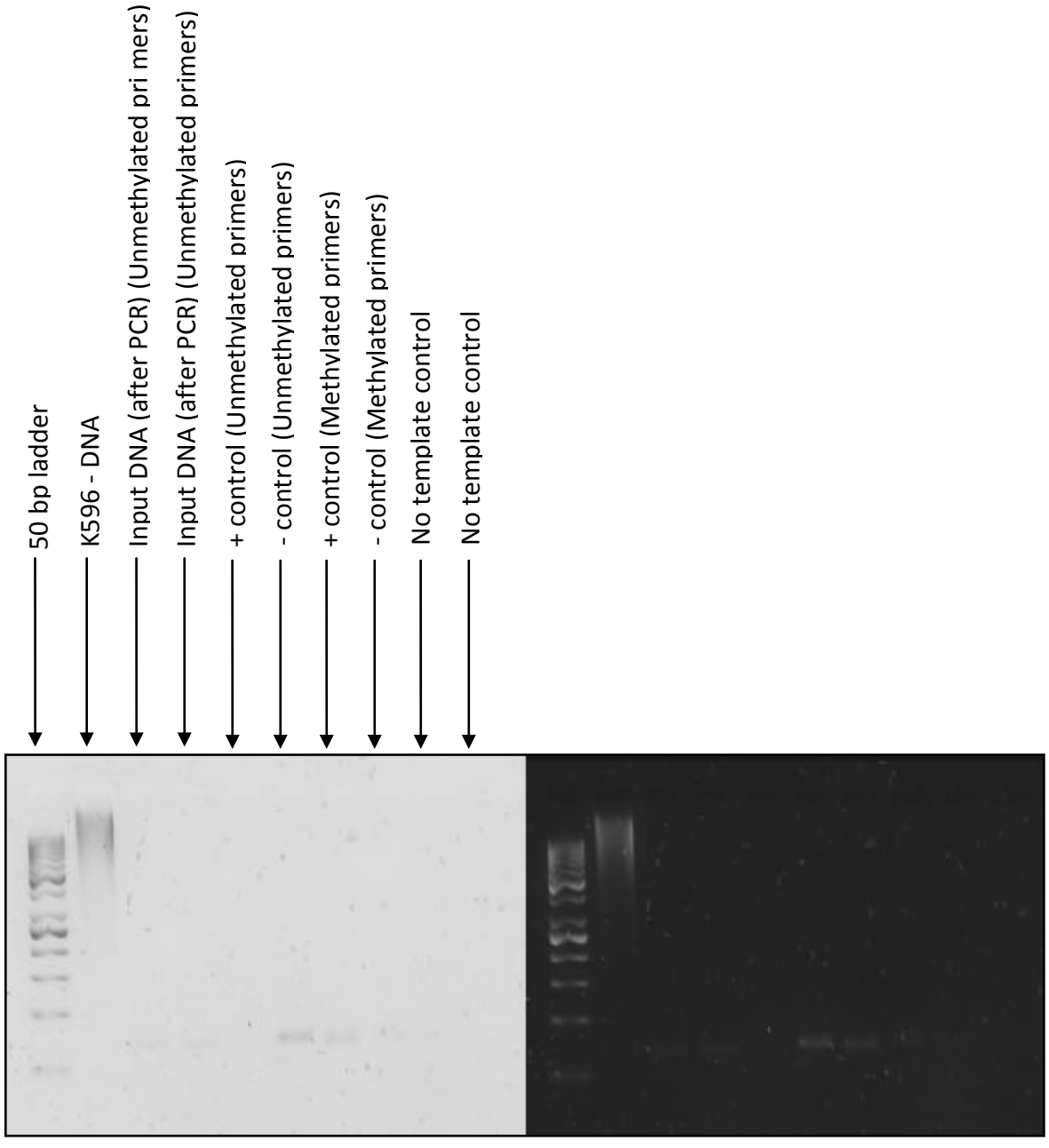


Figure A.1: Results of control validation of the 143B sample enrichment via MethylMiner kit.

The next figure (figure A.2) shows the results of the fibroblast MethylMiner verification (using controls included in the kit):

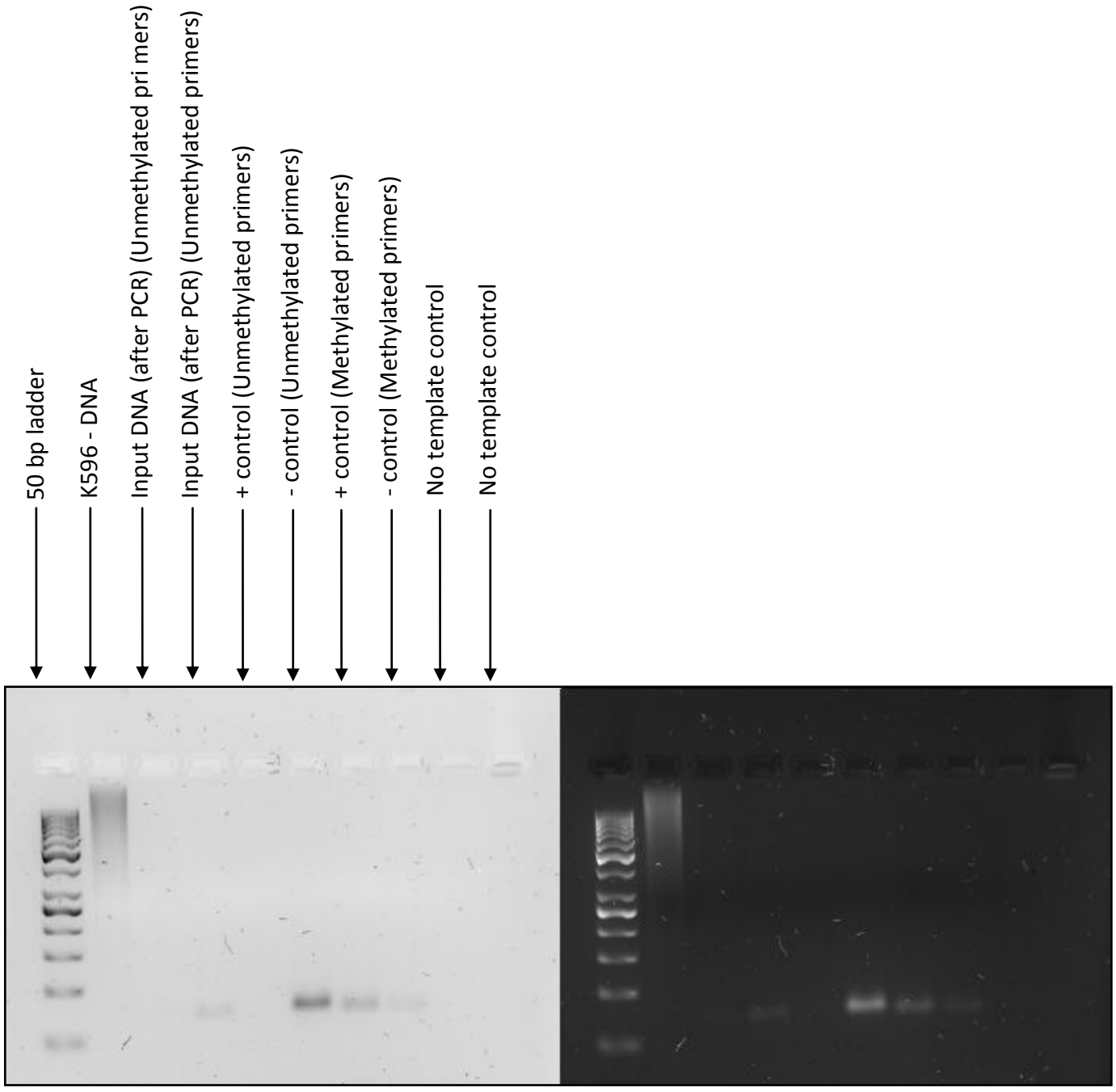


Figure A.2: Results of control validation of the fibroblasts enrichment via MethylMiner kit.

The figure below shows the results of the HeLa control verification (figure A.3):

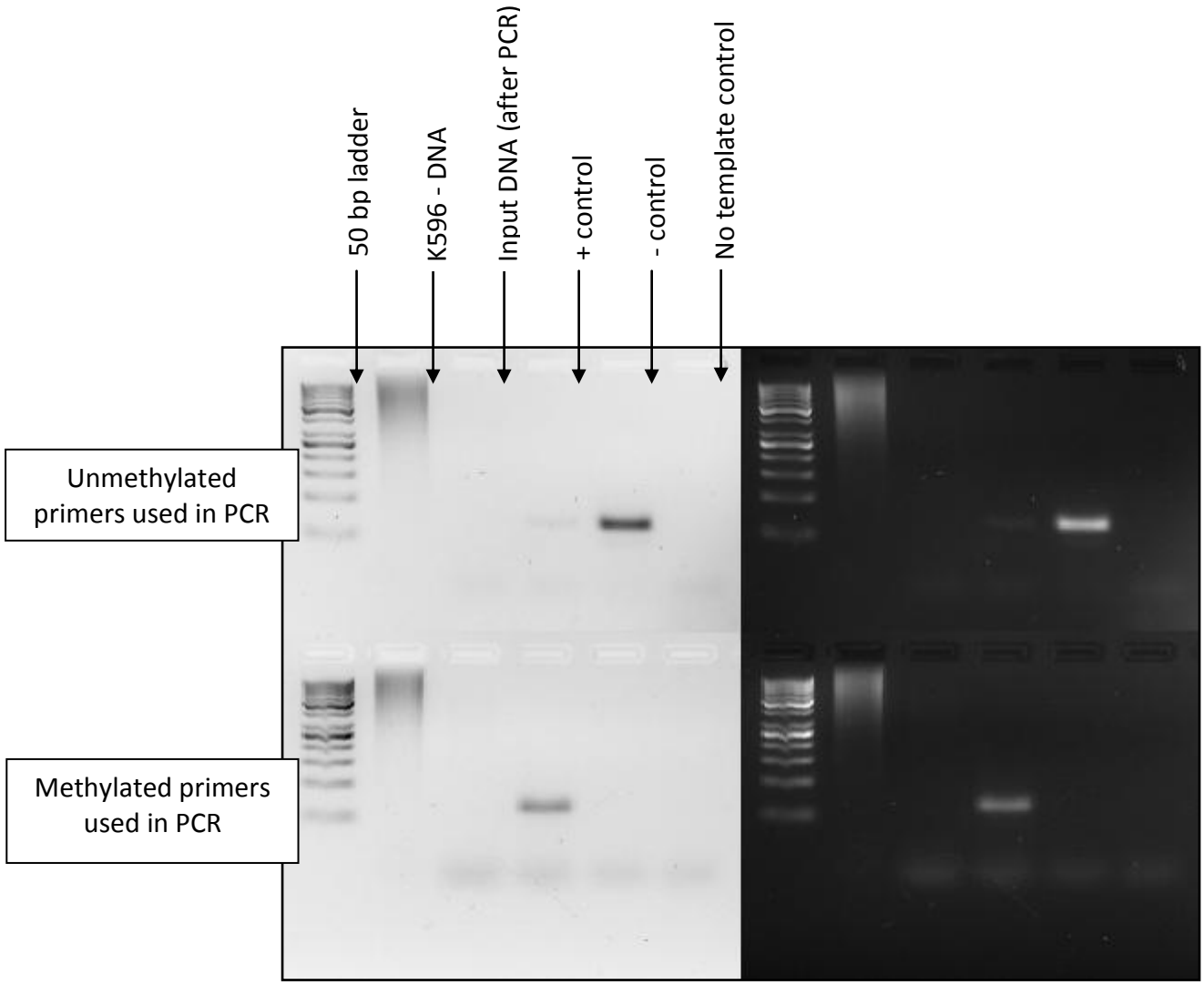


Figure A.3: Results of control validation of HeLa sample enrichment via MethylMiner kit.

The final figure below (figure A.4) shows the results of the HepG2 MethylMiner verification (using controls included in the kit):

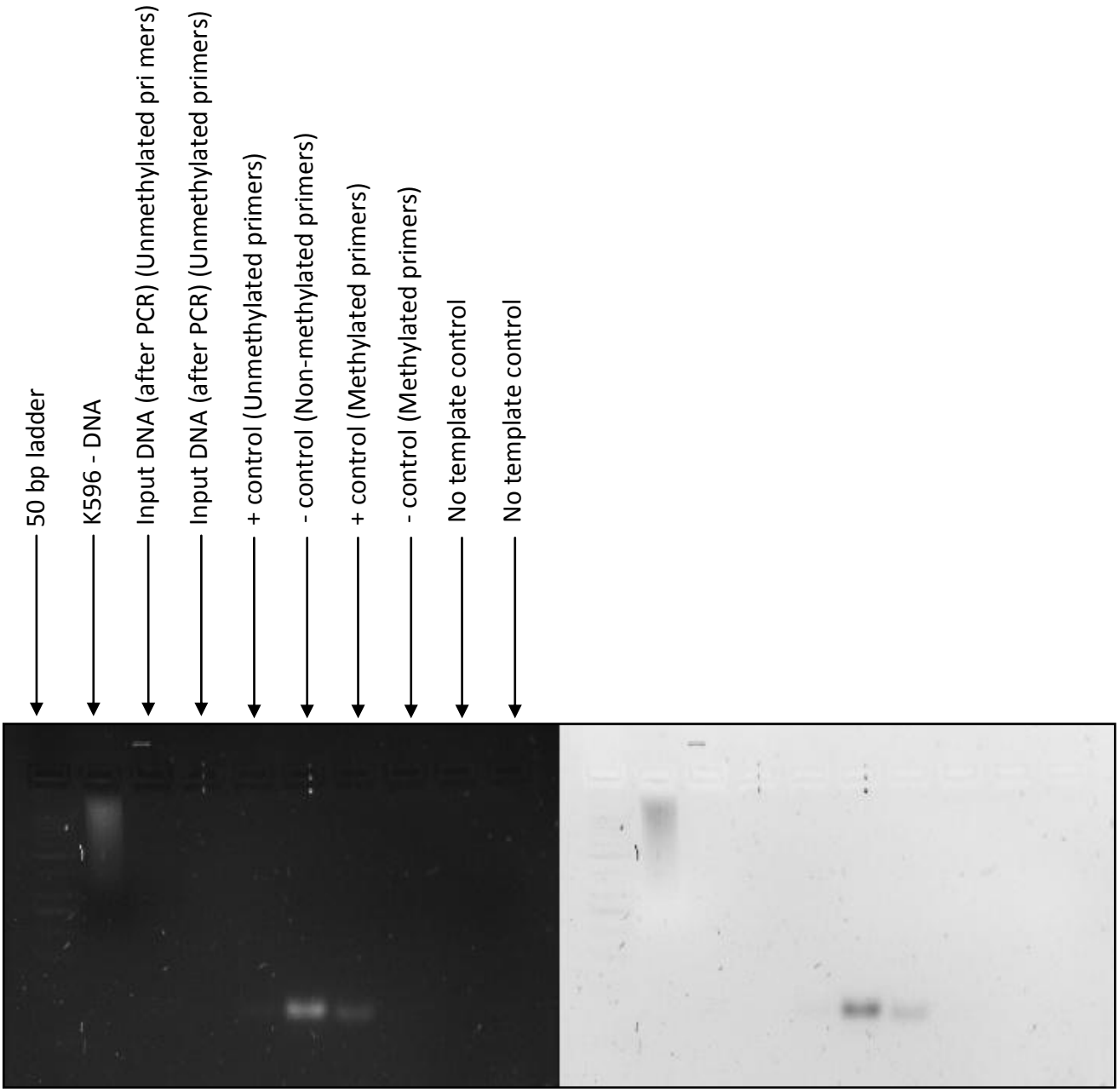


Figure A.4: Results of control validation of HepG2 sample enrichment via MethylMiner kit.

Appendix B: Nanodrop data of samples before DNA sequencing

Methylated DNA enrichment for each of the eight samples was done according to the manufacturer's guidelines of the MethylMiner kit. Nanodrop (ND 1000) results for the samples after enrichment are given below:

[143B cells]

Sample Name: CM

Full Name: 143 B control cells methylated fraction

Nanodrop results: 260/280 = 1,53 260/230 = 0,90 13,4 ng/ul x 55ul = 737 ng

Sample Name: VM

Full Name: 143 B vector-inserted cells methylated fraction

Nanodrop Results: 260 / 280 = 1,43 260/230 = 0,69 20,1 ng/ul x 55ul = 1105,5 ng

[Fibroblasts]

Sample Name: FCF

Full Name: Fibroblast control cells methylated fraction

Nanodrop Results: 260/280 = 1,73 260/230 = 0,98 24,7 ng/ul x 55ul = 1358,5 ng

Sample Name: FVF

Full Name: Fibroblast vector-inserted cells methylated fraction

Nanodrop Results: 260/280 = 1,40 260/230 = 1,67 6.4 ng/ul x 55 ul = 352 ng

[HepG2 Cells]

Sample Name: HepCM

Full Name: HepG2 control cells methylated fraction

Nanodrop Results: 260/280 = 1,63 260/230 = 0,60 13,1 ng/ul x 55 ul = 720 ng

Sample Name: HepVM

Full Name: HepG2 vector-inserted cells methylated fraction

Nanodrop Results: 260/280 = 1,86 260/230 = 0,80 41,9 ng/ul x 55 ul = 2304,5 ng

[HeLa Cells]

Sample Name: HeLaCM

Full Name: HeLa control cells methylated fraction

Nanodrop Results: 260/280 = 2,45 260/230 = 0,39 2 ng/ul x 110 ul = 220 ng

Sample Name: HeLaVM

Full Name: HeLa vector-inserted cells methylated fraction

Nanodrop Results: 260/280 = 1,53 260/230 = 0,65 17,7 ng/ul x 55 ul = 973.5 ng

Note: Many samples have high A260:280 ratios, below the accepted >1.8 value. This may be a contributing factor in the suspected PCR bias.

Appendix C: pEZseq vector and M13 primers

Vector Map and Sequencing Primers (taken from the Blue/White Cloning Kit from Lucigen, cat. # 89002-518)

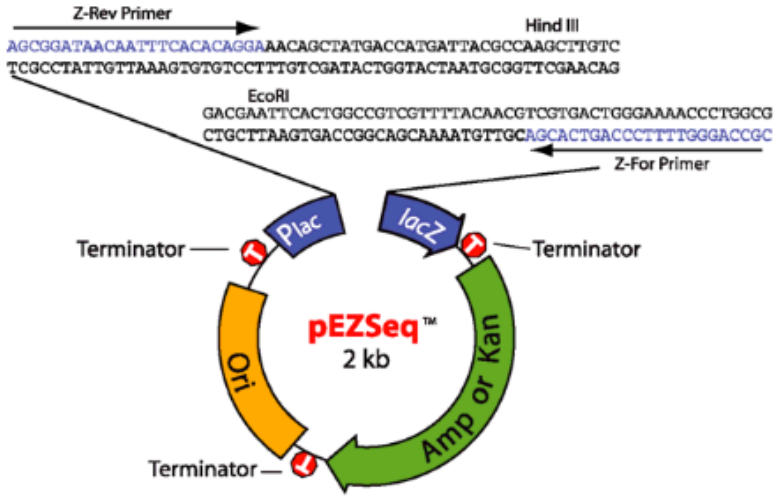
The pEZSeq vector is supplied predigested, with blunt, dephosphorylated ends. Transcriptional terminators border the cloning site to prevent transcription from the insert into the vector.

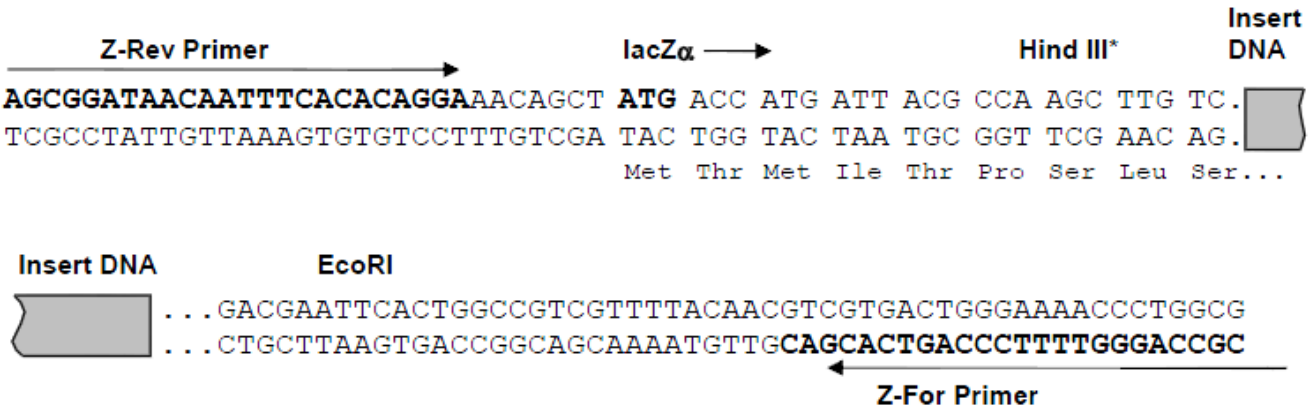
Another terminator at the 3' end of the ampicillin resistance gene prevents this transcript from reading into the insert DNA.

The GenBank accession number of pEZSeq-Amp is AF532109.

The sequences of the Z-Rev and Z-For primers are the same as the M13 Reverse and Forward primers of pUC19:

Z-Rev (M13 Reverse (-48)): 5'–AGCGGATAACAATTTCACACAGGA–3'
Z-For (M13 Forward (-41)): 5'–CGCCAGGGTTTTCCCAGTCACGAC–3'





***The Hind III site is NOT unique in the pEZSeq-Kan vector. Another Hind III site is present in the kanamycin resistance gene.**