



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

ESTABLISHING METHODS FOR ANALYSIS OF DNA METHYLATION IN BREAST CANCER AND CELL-FREE CIRCULATING DNA

Master by Research in Genomics and Experimental Medicine
Institute of Genetics and Experimental Medicine
Centre for Genomics and Experimental Medicine
University of Edinburgh

Clara Domingo Sabugo

Supervisor: Tim Aitman



THE UNIVERSITY
of EDINBURGH



TABLE OF CONTENTS

I. Declaration of Own Work.....	4
II. Acknowledgement.....	5
III. Abstract.....	6
IV. Lay Summary.....	7
1. Introduction	8
1.1. Breast Cancer.....	10
1.2. DNA Methylation in Cancer	13
1.3. Breast Cancer Screening Methods.....	15
1.4. Genotyping Tumor Tissue vs liquid biopsies.....	16
1.5. cfDNA as a potential novel screening method.....	17
1.6. Methodology in CpG Dinucleotide Methylation Analysis.....	20
1.7. Background data for this research project.....	21
2. Hypotheses.....	23
3. Research project aims.....	23
4. Methodology.....	24
4.1. Research samples.....	25
4.2. Fluidigm Access Array Amplification	25
4.2.1. Marker Design and Optimization for Fluidigm Access Array.....	25
4.2.2. PCR Primer Design and Validation.....	27
4.2.3. Preparation of Fully Methylated and Unmethylated DNA.....	27
4.2.4. Bisulfite Conversion of DNA.....	28
4.2.5. Fluidigm Access Array Amplification.....	28
4.2.6. Barcoding and AMPure Clean-up.....	30
4.2.7. Illumina MiSeq Sequencing, Read Alignment and Methylation Calling..	31
4.2.8. Data Analysis.....	32
4.2.9. Statistical Analysis of CpG Dinucleotide Methylation.....	33
4.3. Validation methods.....	34
4.3.1. Preparation of Fully Methylated and Fully Unmethylated Genomic DNA Mixtures and Quantification.....	34
4.3.2. Primer design and validation for Bisulfite Sanger sequencing.....	35
4.3.3. Sanger Sequencing Analysis.....	36
4.3.4. Primer design and validation for Bisulfite Pyrosequencing.....	37
4.3.5. Bisulfite pyrosequencing.....	38

5. Results	39
5.1. Marker Design Optimization.....	39
5.2. Fluidigm Access Array Amplification.....	40
5.1.1. Primer design and validation for Fluidigm Access Array Amplification...40	
5.1.2. Fluidigm Library and Quality Control.....	42
5.1.3. Quality Control of Illumina MiSeq raw data	44
5.1.4. Data Analysis.....	46
A. Analysis of controls coverage.....	46
B. Analysis of samples coverage.....	49
C. Analysis based on complete bisulfite conversion.....	49
D. Analysis of Technical Replicates.....	49
5.1.5. Statistical Analysis.....	50
E. Normality of Methylation Data.....	50
F. CpG Dinucleotide Methylation Analysis of Breast Tumour and Leucocyte Samples.....	51
G. CpG Dinucleotide Methylation Analysis of cfDNA samples.....	52
5.3. Validation of Fluidigm Methylation Results using other Methods.....	54
5.3.1. Bisulfite Sanger Sequencing.....	54
H. Primer Pair Validation.	54
I. Validation of primer pairs in methylated and unmethylated DNA mixtures.....	54
J. Sanger Sequencing Results.	55
K. Quantification of Methylation Level in Sanger Sequencing Traces Using the ab1 Peak Reporter Tool.....	56
5.3.2. Bisulfite Pyrosequencing.....	58
L. Primer Pair Validation.....	58
M. Bisulfite Pyrosequencing Results.....	60
5.4. Correlation of CpG Dinucleotide Methylation.....	64
6. Conclusions	66
7. Discussion	67
8. Future directions	73
9. Appendices	86

Declaration of own work

I hereby declare that all of the work contained within this thesis is the original work of the author and that any work of another person has been appropriately acknowledged.

A handwritten signature in blue ink, appearing to read 'Clara DS', with a large, sweeping flourish underneath.

7th December, 2017

ACKNOWLEDGEMENTS

I would like to give thanks to the following people for their encouragement and support this year:

Professor Tim Aitman for allowing me to work in his group to pursue my Master in Genomics and Experimental Medicine. Prof. Aitman supervised my project with wisdom and patience during this year.

Dr Fiona Semple for her endless patience and wisdom, and for her invaluable advice day to day. I would not have been able to complete it without her encouragement and help. I learnt so much from her. Also to PhD student Danny Laurent who enthusiastically and tirelessly taught and assisted me in learning laboratory and computational skills.

I would also like to thank the bioinformaticians Gil Tomas and Sophie Marion de Proce, for their support and advice during the informatics and statistics of my work. They have listened to all my doubts and have ensured the correct solutions.

During my MRes project I also received support from a lot of people in Prof. Tim Aitman's group. They were David Ross, Dr Holly Black, David Parry, Elaine Ross, Erica Loring and Dominique Balharry. Thank you very much for teaching my techniques and giving me a lot of inputs through discussions.

I would also like to thank my thesis committee, Dr Richard Mehaan and Professor David Cameron for their supervision and useful feedback after the reports.

I am also grateful to the Wellcome Trust Research Facility, particularly William Hawkins for his help with pyrosequencing.

Lastly, recognition must be given the staff at the IGMM, from the front house reception staff, to the HR and the staff of the Nucleus café – they made this a truly fantastic place to work for a year.

ABSTRACT

Breast cancer is the most common cancer in women worldwide. To date, diagnosis and metastasis monitoring are mainly carried out through tissue biopsy, a very invasive procedure limited only to certain locations and not always feasible in clinical practice. Tumour cells release DNA into the blood as circulating cell-free DNA (cfDNA), which can be sampled from circulating blood, an approach known as liquid biopsy. This provides a resource for biomarkers that could allow the use of minimally invasive liquid biopsies for cancer-related research, diagnostics, prognostics, and targeted therapy. The levels of cfDNA have already been shown to be higher in cancer individuals than healthy individuals, and correlate with tumour metastasis, response to therapy and recurrence. Recent technological advances have enabled the identification of both genetic and epigenetic aberrations in cfDNA that reflect changes also found in patients' tumours. The host group performed methylation analysis using the Illumina EPIC Methylation Array, which interrogated CpG dinucleotide methylation at over 850,000 DNA sites. A total of 3172 CpGs showed median methylation differences of more than 40% between tumour and buffy coat of patients with breast cancer. This MRes project aimed to establish methods for detecting these methylation changes between matched tumour samples and leucocytes of breast cancer patients, and in cfDNA by using the Fluidigm 48.48 Access Array microfluidics system and the Illumina MiSeq sequencer. This approach provided quantitative, medium-throughput targeted measurement of DNA methylation at single nucleotide resolution. Finally, bisulfite pyrosequencing was used as a sensitive validation technique for detecting differences in CpG methylation, providing a set of potential biomarkers that could be reliably detected by circulating tumour DNA-based tests. Translating the alterations that are seen in the primary tumour into an assay that is applicable to cfDNA will have important diagnostic implications, such as monitoring tumour progression, drug response and disease recurrence, as well as the early detection of cancer, which could ultimately complement or even avoid the need for tumour tissue biopsies.

Keywords: cell-free DNA, methylation, breast cancer, Fluidigm 48.48 Access Array microfluidics, biomarkers

LAY SUMMARY

Breast cancer is the most common cancer in women worldwide. To date, diagnosis and monitoring is mainly carried out through surgery of tissue biopsy, a very invasive procedure limited only to certain locations and not always feasible in clinical practice. Like many other cells, cancerous cells shed DNA into the blood as cell-free DNA (cfDNA). These molecules can be easily sampled from circulating blood, an approach known as liquid biopsy. A liquid biopsy may be used to help find cancer at an early stage. It may also be used to help plan treatment or to find out how well treatment is working or if cancer has come back. Being able to take multiple samples of blood over time may also help to understand what kind of molecular changes are taking place in a tumour. Tests in development examine these bits of DNA, finding different types of alterations serving as markers for specific cancer types. This project aimed to establish methods for detection of specific changes in primary tumour samples and in cfDNA of breast cancer patients. Establishing methods that could make cancer detection so painless that it becomes part of routine check-ups will have important diagnostic implications, such as monitoring tumour progression, drug response and disease recurrence, as well as the early detection of cancer, which could ultimately complement or even avoid the need for tumour tissue biopsies.

1. INTRODUCTION

Cancer is a very heterogeneous disease in most aspects, including its cellularity, different genetic alterations and several clinical behaviors. The combinatorial origin, the heterogeneity of malignant cells, and the variable host background produce multiple tumour subclasses. Many analytical methods have been used to study human tumours and to classify them into homogeneous groups that can predict clinical behavior. Currently, cancer classifications are mostly based on clinical and histomorphologic features that only partially reflect the molecular heterogeneity of the tumour, reducing the probability of the most appropriate diagnostic, prognostic and therapeutic strategy for each patient¹. Progress in cancer genomics research over the past few decades has revealed that cancer is driven by various genomic alterations. As a result of different international initiatives such as The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC), the use of next-generation sequencing (NGS) has helped define the genomic landscape of early stage cancers². In breast cancer, these studies have reported the high level of tumour heterogeneity between patients that consists of several molecular subsets, which are driven by different molecular alterations³.

Breast cancer emerges by a multistep process which involves the transformation of normal cells via the steps of hyperplasia, premalignant change and *in situ* carcinoma⁴. Like all cancers, breast cancer is considered to result in part from the accumulation of multiple genetic alterations leading to oncogene overexpression and tumour suppressor loss⁵. Recently, the role of epigenetic changes has emerged as a crucial and characteristic mechanism in many cancer types⁵. Epigenetic modifications are believed to occur early in carcinogenesis and precede genetic alterations⁶. Changes in DNA methylation have been recognized as one of the most common molecular alterations in cancer and hypermethylation of gene-promoter regions is a suggested mechanism of loss of gene function. For instance, in primary lung carcinomas, the inactivation of the tumour suppressor gene *p16*⁷, the DNA repair gene *MGMT*⁸, and the detoxifying gene *GSTP1*⁹ by promoter hypermethylation have been well described. At the same time that certain CGIs become

hypermethylated, the degree of genomic DNA hypomethylation increases from the benign proliferations to the invasive cancers¹⁰.

Molecular understanding of breast carcinogenesis has been accumulated during the last decades but has hardly been translated into the clinic as strategies for early detection or prevention of cancer. Screening for breast cancer by mammography is well known to reduce mortality¹¹. However, an alternative DNA-based approach for early detection of breast cancer might be promising since DNA extracted from the patient's plasma, serum or other body fluids could be easily amplified by PCR technology and is potentially more sensitive than current tests.

Epigenetic alterations, including aberrant DNA methylation, have been implicated in abnormal expression patterns of different tumours, including breast cancer. DNA methylation is a reversible chemical alteration of DNA that mostly affects cytosines when located 5' to a guanosine. CpG dinucleotides are not randomly distributed throughout the genome¹². Rather they are frequently clustered into CpG islands, regions that are rich in CpG sites and these areas are frequently located at the promoter regions of the coding genes present in the mammalian genome, where they modulate gene transcription¹³. It has been increasingly recognized that the CpG islands of a large number of genes, which are mostly unmethylated in normal tissues, are methylated in human cancers, including breast cancer⁵. Detection of promoter CpG island hypermethylation offers several advantages compared to other DNA alterations in cancer¹⁴. Methylated DNA can be detected with a very high degree of specificity, even in the presence of a vast excess of unmethylated DNA. MethyLight technology for instance can detect a single hypermethylated allele against a background of 10.000 unmethylated alleles¹⁵. Methylated DNA from patients with manifest breast cancer has been detected in blood¹⁶ as well as in ductal lavage fluid¹⁷. Detection of DNA methylation in blood might prove to be useful as a predictive marker at the moment of primary diagnosis or as a marker for early detection of relapse of disease. Therefore, analysis of epigenetic alteration is a potential approach in early detection of breast cancer¹⁸.

1.1. Breast Cancer

Breast cancer is the most common malignant disease and the leading cause of cancer death among females, expected to account for 29% of all new cancer diagnoses in women¹⁹. Even though breast cancer survival in the UK has doubled in the last 40 years, this improvement is achieved by earlier stage diagnosis²⁰. Detection of localized breast cancer at an early stage, when cancer cells have not yet invaded neighboring normal tissue, has better prognosis and requires less severe treatment with a survival rate of 98%²¹. However, breast cancer is frequently diagnosed at a late stage after tumour metastasis, which lowers the survival rate to 14%²² (Figure 1). Thus, the identification of indicators characterizing the early stages of formation and progression of breast cancer may reduce the incidence of this disease²³.

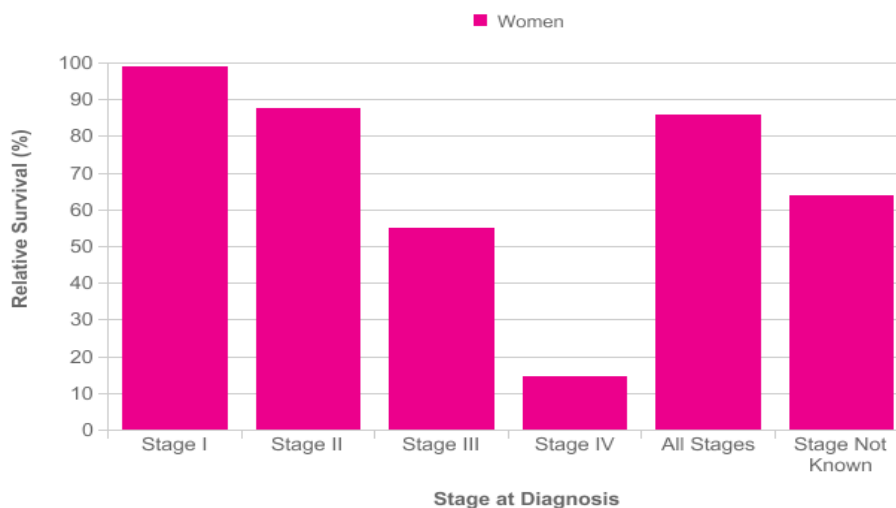


Figure 1: Five-Year Relative Survival (%) by Breast Cancer Stage at diagnosis in adults aged 15-99, Former Anglia Cancer Network. Five-year survival for female breast cancer shows a much more rapid decrease in survival between Stages I and IV. Five-year relative survival in women ranges from 99% at Stage I to 15% at Stage IV for patients diagnosed during 2002-2006 in the former Anglia Cancer Network. From: Cancer research UK²⁴.

There are several different types of breast cancer, which can occur in different parts of the breast. According to the World Health organization (WHO) classification, breast cancer can be classified up to 21 distinct histological types on the basis of cell morphology, growth and architecture patterns²⁵. The most common type is invasive ductal breast carcinoma (IDC) of no special type (NST), which comprises all tumours without the specific differentiating features

that characterize the other categories of breast cancers²⁶. IDC represents approximately 60% to 75 % of all breast cancers, whereas breast cancer special types represent 25%. This minority group includes the classic lobular invasive carcinoma, as well as mucinous, tubular, adenoid cystic and cribriform carcinomas, among others²⁷.

Breast cancer is often divided into non-invasive breast cancer (carcinoma *in situ*) when cancerous cells remain in a specific location of the breast without spreading to surrounding tissue, or invasive breast cancer if cancerous cells break through normal breast tissue and spread to other parts of the body through the bloodstream and lymph nodes. Other less common types of breast cancer include invasive (and pre-invasive) lobular breast cancer, inflammatory breast cancer and Paget's disease of the breast²⁸. Most breast cancers are carcinomas if they start in the epithelial cells that line organs and tissues. In fact, breast cancers are often a type of carcinoma called adenocarcinoma, which starts in cells that make glands (glandular tissue). Breast adenocarcinomas start in the ducts (the milk ducts) or the lobules (milk-producing glands). There are other types of breast cancers, such as sarcomas, which start in the cells of the muscle, fat, or connective tissue, and sometimes a single breast tumour can be a combination of these different types²⁹.

Breast cancer can also be classified based on the receptor status, which is important in deciding treatment options. Breast cancer cells can express three different receptors: estrogen receptor (ER), progesterone receptor (PR), which are both endocrine receptors, or epidermal growth factor receptor (HER2). Cancers are called hormone receptor-positive or hormone receptor-negative based on whether or not they have these receptors, and classified as:

- Endocrine receptor-positive (estrogen or progesterone receptors).
- HER2-positive.
- Triple positive: positive for estrogen receptors, progesterone receptors, and HER2.
- Triple negative: not positive for estrogen receptors, progesterone receptors, and HER2.

Currently, cancers are being classified into molecular subgroups based on the variability in treatment response and prognosis than current histological classifications. For instance, gene expression patterns have allowed classification of breast tumours into five different molecular subtypes, namely basal-like, ErbB²⁺, normal breast like, luminal subtype A and luminal subtype B^{30,31,32,33}. More recently, a new subtype classified as “claudin-low” has also been identified³⁴. Ongoing molecular classification studies will ultimately allow stratification of patients for a more personalized treatment based on molecular alterations. This will provide useful information regarding patient-specific prognosis and risk of relapse probability for complete response, not only in breast cancer, but also in many other cancer types.

It is well known that breast cancer is a genetic disease. Breast tumorigenesis is best described by a multi-step progression model⁴, in which the normal breast epithelium evolves via hyperplasia and carcinoma *in situ* into an invasive cancer, which eventually can disseminate via lymph and vascular systems to form metastases. Each of these steps may be the result of one or more distinct mutations in regulatory genes. Several somatic mutations in these genes have been found in breast cancer cells. Mutational activation of oncogenes, often coupled with non-mutational inactivation of tumour suppressor genes, is probably an early event in sporadic tumours, followed by mutations in multiple additional genes, leading to tumorous cancer development. Oncogenes that have been reported to play an early role in sporadic breast cancer are *MYC*, *CCND1* (Cyclin D1) and *ERBB2* (HER2/neu). Sporadic breast cancer accounts for 70-80% of the cases. Less commonly, germline mutations predispose that individual to develop breast cancer. However, the inheritance of a mutated cancer susceptibility allele of a high penetrance susceptibility gene (such as *BRCA1*, *BRCA2*, *TP53* or *PTEN*) is only the first step in promoting the development of a malignancy and does not guarantee that an individual will go on to develop a particular malignancy³⁵. The development of a heritable cancer, as well as most other cancers, is postulated to be dependent on the occurrence of a second genomic alteration (Knudson’s “two-hit” model)³⁶. The current consensus is that these are ‘caretaker’ genes, which, when inactivated, allow other genetic defects to accumulate³⁷. Moreover, analysis of molecular features

of early stage breast cancer using NGS has confirmed that *TP53* and *PIK3CA* mutations are the most frequent genomic alterations overall in all intrinsic subtypes (28% for both genes)³. Amplifications in *ERB2*, *FGFR1* and *CCND1* follow in frequency, being observed in 10-20% of all breast cancer subtypes. *PTEN* mutations and deletions, among many others genes, including *KRAS*, *APC*, *NF1*, *SKT11*, *MAPK2K4*, *MAP3K1* and *AKT2* can also be altered in breast cancer². Environmental and lifestyle factors also influence whether a person will develop breast cancer³⁸. Recognized risk factors for breast cancer development include hormonal, reproductive, and menstrual history, age, lack of exercise, alcohol, radiation, benign breast disease, and obesity³⁹.

1.2. DNA Methylation in Cancer

Decades of research have led to a substantial understanding of the factors that cause or are associated with development of breast cancer and lead to the appearance of a neoplasm, characterized by a variety of genetic lesions, such as gene amplifications, gene deletions, point mutations, loss of heterozygosity, chromosomal rearrangements, and overall aneuploidy. Even though the genetic origin of cancer is widely accepted, epigenetic alterations are among the most common molecular alterations in human neoplasia⁴⁰. These include DNA methylation, histone modifications, nucleosome positioning and aberrant expression of non-coding RNAs, specifically microRNAs⁴¹. Both genetic and epigenetic alterations interact at all stages of cancer development to promote cancer progression^{42,43,44,45}. For instance, *EZH2*, the writer of the histone H3 lysine 27 trimethylation (H3K27me3) mark associated with Polycomb repressive complex 2 (PRC2), is overexpressed in aggressive breast cancer as a consequence of genetic upstream mutations in *BRCA1*. As a consequence, *EZH2* leads to cancer cell migration and invasion by methylating and silencing important differentiation genes⁴⁶.

Altered DNA methylation patterns is the most studied epigenetic modification in cancer, which has been considered as a hallmark of the disease^{47,48,49}. DNA methylation refers to the addition of a methyl group covalently to the base cytosine. In vertebrates, DNA methylation mainly occurs at cytosines in a CpG dinucleotide context⁵⁰. Most CpG dinucleotides in the human genome are

methyated. However, CpGs are not normally distributed, as they have been severely depleted in the vertebrate genome to about 20% of the predicted frequency. The only exception for this global CpG depletion resides in a specific category of GC- and CpG-rich sequences termed CpG islands that are found at increased density in the promoters of genes⁵¹, where they are generally unmethylated⁵².

This modification is involved in regulating many processes, including embryonic development, transcription, genomic imprinting, among others¹⁸. Consistent with these important roles, DNA methylation is associated with transcriptional silencing and aberrant methylation may play a role in silencing of tumour suppressor genes. For gene transcription to occur, the gene promoter should be accessible to transcription factors (TFs) and other regulatory units⁵³. DNA methylation can directly prevent transcription factor binding and lead to changes in chromatin structure that restrict access of TFs to the gene promoter⁵⁴. Interestingly, some developmentally important human TFs prefer to bind to methylated CpG sites, thus methylation exerts a selective effect on factors that binds to a target sequence⁵⁵. In many human diseases, including cancer, it is acknowledged that aberrant methylation or hypermethylation of promoters alters the expression of a variety of critical genes, such as tumour suppressor genes, affecting different transcriptional pathways and hence facilitating the development of malignant tumours⁴¹.

Overall, DNA methylation alterations play an important role in all stages of multistep tumorigenesis from the early onset of malignant transformation^{56,57,58}. With the advent of NGS technologies we have obtained genome-wide DNA methylation maps at high resolution that reveal new key players in cancer, including breast cancer. Initial clinical data strongly suggests that DNA methylation signatures may be useful tumour biomarkers for cancer detection, diagnosis and prognosis⁵⁹, and especially in early detection of the disease. Moreover, unlike genetic alterations, DNA methylation is reversible, rendering it a potential target for novel therapy approaches⁶⁰. For this reason, epigenetic alterations are leading candidates for the development of specific markers. Although epigenetic alterations are well characterised to be involved in cancer

development, current diagnostic techniques are well behind current biological knowledge.

1.3. Breast Cancer Screening Methods

Mammography is the most widely used screening method, with solid evidence of benefit for women aged 40 to 74 years. Clinical breast examination and breast self-exam have also been evaluated but are of uncertain benefit⁶¹. Overall, the breast screening program finds cancer in about 8 out of every 1,000 women having screening⁶². This benefit is greater for women who are at higher risk for breast cancer based on older age or other risk factors such as family history.

Despite mammography screening may be effective in reducing breast cancer mortality in certain populations, it can pose harm to women who participate. The limitations are best described as false positives (related to the specificity of the test), overdiagnosis (true positives that will not become clinically significant), false negatives (related to the sensitivity of the test), discomfort associated with the test, radiation risk, and anxiety⁶³. About half or more of women who have a mammogram early for 10 years will have a false-positive mammogram, and up to 20% of these women will need a biopsy. For some women undergoing regular screening, the mammogram may find an invasive cancer or noninvasive condition (ie, ductal carcinoma *in situ*) that would never have caused problems. There is about a 19% chance that the cancer is being overdiagnosed, and patients will receive unnecessary treatment. Moreover, 13% of breast cancers are undetectable by mammography due to tumour size and age of patients, thus being more advanced when diagnosed, as they may grow longer before being detected by a screening mammogram⁶³. In addition, currently used biomarkers with low accuracy, such as cancer antigen CA15-3 and carcinoembryonic antigen (CEA), have been recommended against for accurately diagnosing breast cancer⁶⁴.

For decades, there has been strong interest in screening strategies that would be able to detect early cancers before they progress, thereby reducing mortality. Even though mammography screening appears to reduce breast cancer mortality, for some patients, the harms may outweigh the benefits. Therefore,

better breast cancer screening tests are needed. Nevertheless, early breast cancer detection is dependent on sensitive and specific screening methods⁶⁵.

1.4. Genotyping Tumour Tissue vs liquid biopsies

Imaging studies such as mammogram and MRI, often along with physical exams of the breast, can lead doctors to suspect that a person has breast cancer. However, the only way to confirm the existence of a tumour is to take a sample of the tissue from the suspicious area and examine it under a microscope⁶⁶. Tissue biopsy is also the gold standard for clinical and investigational sequencing, but barriers exist in terms of acquisition and utility. Likely, the major limitation of tissue biopsy is heterogeneity, which characterizes most advanced cancers^{67,68}. Cancers are heterogeneous, with different areas of the same tumour showing different genetic profiles (ie, intratumoral heterogeneity); likewise, heterogeneity exists between metastases within the same patient (ie, intermetastatic heterogeneity). A biopsy or tissue section from one part of a tumour will miss the molecular intratumoral as well as intermetastatic heterogeneity⁶⁹. Moreover, biopsies are expensive and invasive procedures for patients that can lead to clinical complications. To overcome the limitations of tissue biopsies, less invasive, cost-effective and highly sensitive and specific techniques capable of capturing tumour heterogeneity and the molecular changes that cancer cells undergo are needed to detect and diagnose breast cancer⁶⁴.

When a comprehensive analysis of the overall disease is required or when tissue specimens are difficult to obtain or are unavailable, liquid biopsies are an attractive alternative option. This is because circulating tumour DNA fragments contain identical genetic defects to those in the tumours themselves and cancer-related molecular alterations can be detected in cfDNA. These include somatic point mutations, loss of heterozygosity (LOH), translocations, gene copy number changes and DNA methylation changes⁷⁰. Thus, cfDNA tests may potentially overcome problems related to tumour heterogeneity and accessibility⁷¹. Also, repeated blood samples can be taken in order to monitor changes in cfDNA in the natural course of the disease or during cancer

treatment⁷². In the detection of metastatic breast cancer, cfDNA shows superior sensitivity to that of other conventional tumour biomarkers and has a greater dynamic range that correlates with changes in tumour burden⁷³. Furthermore, the accuracy of advanced qualitative analysis showed even higher level of discriminatory power in breast cancer detection⁶⁴.

1.5. cfDNA as a potential novel screening method

In 1948, Mandel and Metaís discovered for the first time the presence of DNA in human blood⁷⁴, termed as circulating cell-free DNA (cfDNA). All cells, including tumour cells and non-malignant cells, shed DNA, called cfDNA, into the circulatory system. Circulating tumour DNA (ctDNA) is cfDNA that is shed from tumour cells into the circulatory system. These molecules can be easily sampled from circulating blood, an approach known as liquid biopsy. In healthy individuals, plasma cfDNA is believed to derive primarily from apoptosis of normal cells of the hematopoietic lineage, with minimal contributions from other tissues⁷⁵. Furthermore, the levels of cfDNA have been shown to be higher in cancer individuals than in healthy individuals, and correlates with tumour metastasis, response to therapy and recurrence⁷⁶. As the tumour increases in volume, so too does the cellular turnover and hence the number of apoptotic and necrotic cells^{67,68}. Under normal physiologic circumstances, apoptotic and necrotic remains are cleared by infiltrating phagocytes. This does not happen efficiently within the tumoral mass, leading to the accumulation of cellular debris and its inevitable release into the circulation⁶⁹.

In addition to quantitative changes, qualitative alterations of circulating cfDNA have also been observed, such as microsatellite alterations⁷⁷, oncogenes, tumour suppressor, and other somatic gene mutations⁷⁸, mitochondrial DNA, viral DNA⁷⁹ and tumour-specific methylated DNA.

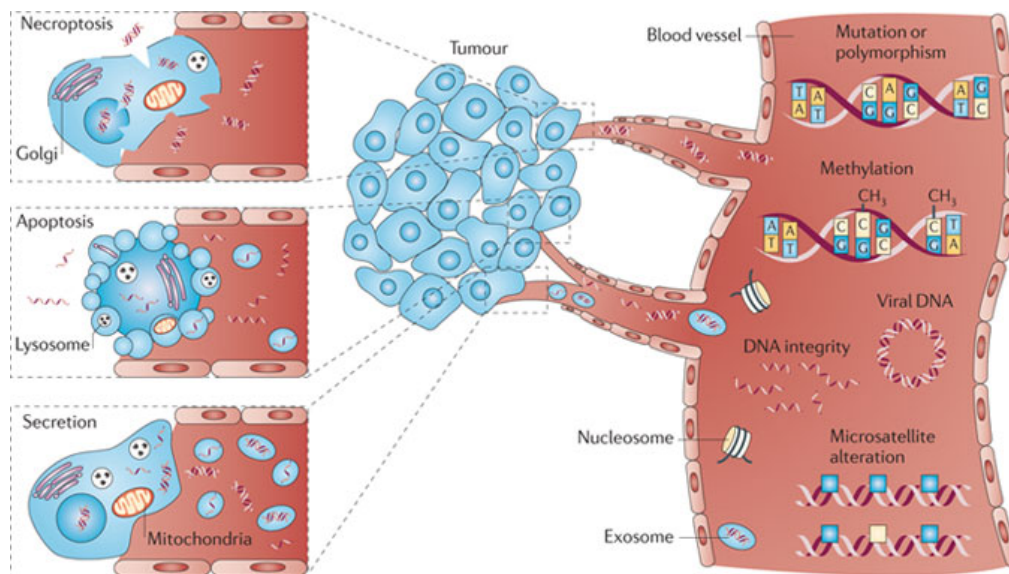


Figure 2: Mutations, methylation, DNA integrity, microsatellite alterations and viral DNA can be detected in cfDNA in blood. Reproduced from Schwarzenbach, Hoon and Pantel, *Nat Rev Cancer*, 426–437 (2011). The release of DNA from tumour cells can be through various cell physiological events such as apoptosis, necrosis and secretion. The physiology and rate of release is still not well understood; tumour burden and tumour cell proliferation rate may have a substantial role in these events⁸⁰.

The release of cell-free nucleic acids (cfNAs) into the bloodstream occurs by different sources, including the primary tumour, tumour cells that circulate in the blood, and micro-metastatic deposits that are present at distant sites, (for example, bone marrow and liver), and normal cell types, such as hematopoietic and stromal cells⁸¹. The physiological events that lead to the increase of cfNAs in the blood during cancer development and progression comprise increased apoptotic and necrotic cell deaths as well as active secretion into the blood circulation^{82,83}. Necrotic and apoptotic cells are usually phagocytosed by macrophages or other scavenger cells⁸⁴. Macrophages then release fragmented cfDNA into the bloodstream. These mechanisms are passive cfDNA release mechanisms. It is also hypothesised that cancer cells can secrete cfDNA into the bloodstream actively. Importantly, DNA can be shed as both single-stranded and double-stranded DNA bound the nucleosome or are found inside exosomes⁸⁰. cfDNA is highly fragmented, with most molecules being approximately 150 bp in length. This matches the length of DNA occupied by a nucleosome, the primary unit for spatial organization of DNA in the nucleus⁸⁵. On average, the size of this DNA varies between small fragments of 70 to 200 base pairs and large fragments of approximately 21 kilobases. Physiologically,

cfDNA is removed from 15 minutes to several hours from the circulation by blood nuclease activity and filtration of kidney and liver⁸⁰. The cfDNA is extracted from the plasma or serum fraction of the blood. However, there are some challenges in working with cfDNA, such as the low concentration of cfDNA in the circulation and high admixture of normal DNA in cfDNA pose major challenges for the development of sensitive and robust detection pipelines^{84,86}.

In addition to cfDNA, other circulating bioamarkers can be analysed using liquid biopsies, including circulating tumour cells (CTCs), cell-free mRNA (cfRNA), microRNAs (miRNAs), exosomes, proteins and metabolites. In terms of detection, only 1 to 10 CTCs per ml of whole blood are found in patients with metastatic disease⁸⁷, therefore the isolation and characterization of CTCs presents a technical challenge. In fact, current CTC technologies only can detect 60 to 80% of patients with known metastases. In contrast, cfDNA analysis is demonstrably more sensitive than CEA measurement (the current standard blood biomarker) to define stage II colon cancer patients at very high risk of recurrence after resection, even after completion of adjuvant chemotherapy⁸⁸. In addition, cfRNA and circulating mRNA are also present in the blood, either in circulating ribonucleoprotein complexes or packaged into exosomes. These molecules have been detected in the plasma of cancer patients, using microarray technologies and quantitative real-time RT-PCR, and led to the detection of expression patterns specific for aggressive prostate cancer⁸⁹. Furthermore, exosomes with cancer mRNA transcripts as well as dsDNA as a result of active secretion by tumour cells, are promising biomarkers⁹⁰. Nevertheless, these are difficult to isolate and are very rare in clinical samples compared to those coming from normal cells that are much more abundant.

Taken together, further studies of the biology of these molecules are required and standardised techniques need to be developed, but it is clear that cfDNA, CTCs and circulating RNA have the potential to be translated into the clinic when acceptable levels of sensitivity and specificity are achieved. Ultimately, combined assessment of different biomarkers will increase the sensitivity required for early cancer detection. For instance, Fackler et al have already

developed cMethDNA, a quantitative multiplexed methylation-specific PCR assay for a panel of genes, and have detected and validated methylation signatures of known breast cancer markers in tumour DNA of metastatic breast cancer patients with a sensitivity and a specificity of 91% and 96%, respectively⁹¹.

1.6. Methodology in CpG Dinucleotide Methylation Analysis

Current epigenetic analysis focus in the investigation of DNA methylation and chromatin modification patterns, and more recently, studies have been directed toward a genome-wide assessment⁹². As DNA methylation effectively down-regulates gene activity, methylation profiling of specific CpG islands appears to be a promising approach for cancer risk assessment for its known role in tumorigenesis, but also because it is involved in many other genetic disorders^{49,80}. Of especial interest is the study of cfDNA methylation patterns, since plasma contains a mixture of DNA from different tissues and organs. As certain methylation patterns are tissue specific, they could serve as an epigenetic signature for the respective cells or tissues that release their DNA into the circulation⁹³.

There are many techniques to analyze changes in CpG dinucleotide methylation, depending on factors such as the availability of the DNA or the number of targets to analyze. CpG dinucleotide methylation analysis is enabled through bisulfite modification of DNA, which was first investigated by Hayatsu et al (1970)⁹⁴. Bisulfite modification converts nonmethylated cytosines to uracils, which are then converted to thymines during DNA amplification by PCR, whereas methylated cytosines are protected from bisulfite modification⁹⁵. After the bisulfite treatment, PCR amplification with specific methylation primers allows to determine the methylation status in the CpG of interest. Sequencing analysis of bisulfite-converted DNA is regarded as a gold-standard technology to qualitative and quantitatively reveals the methylation status at single nucleotide resolution⁹⁶. In order to reach this goal, PCR amplification in this study was carried out with Fluidigm Access Array. It is a microfluidics-based technology that enabled amplification of 48 samples with 48 primer pairs in one run on a single chip. The workflow consists of sample and primer loading, PCR

amplification and sample pooling, followed by sequencing and analysis⁹⁷. Fluidigm enables amplification of 2304 (48 x 48) reactions per chip by separating these reactions in nanoliter microchambers. Approximate amount of starting DNA for Fluidigm can be as low as 0.05µg, which is suitable for the low concentration of cfDNA that can be extracted from plasma⁹⁸.

1.7. Background data for this research project

The Illumina EPIC Methylation Array previously carried out in the host group interrogated over 850,000 methylation sites quantitatively across the genome at single-nucleotide resolution. The Infinium MethylationEPIC BeadChip includes more than 90 % of the CpGs on the HumanMethylation450 (HM450) and an additional 413,743 CpGs at regions identified as potential enhancers. The proportion of DNA methylation at a particular CpG site (also called the methylation beta-value (β)) is then ascertained by taking the ratio of the methylated (C) to unmethylated (T) signal, using the formula:

$$\beta = \frac{\text{Intensity of the methylated signal}}{(\text{Intensity of the unmethylated signal} + \text{Intensity of the methylated signal})}$$

A β of 0 represents a completely unmethylated CpG site and a β -value approaching 1 represents a fully methylated CpG site⁹⁹. This platform allowed profiling of CpG dinucleotide methylation in 18 paired samples consisting of tumour tissue and buffy coat recruited from patients with breast cancer. A total of 3172 probes obtained median methylation differences of more than 40% between tumours and the paired buffy coat in these patients (Fiona Semple and Gil Tomas, IGMM, Personal communication). At this stage of the project, buffy coat is used as a surrogate for plasma cfDNA because most non-tumour cfDNA is anticipated to derive from leucocytes^{100,75}.

Probes having extreme beta-values in leucocyte samples (i.e. methylation levels of less than 5% and more than 95%) were ranked by q-value, median differences and widest beta value separation between tumour and buffy coat samples (Figure 3). Unique probes intersecting these 3 criteria made up a total of 168 probes that were used for Fluidigm Access Array primer design.

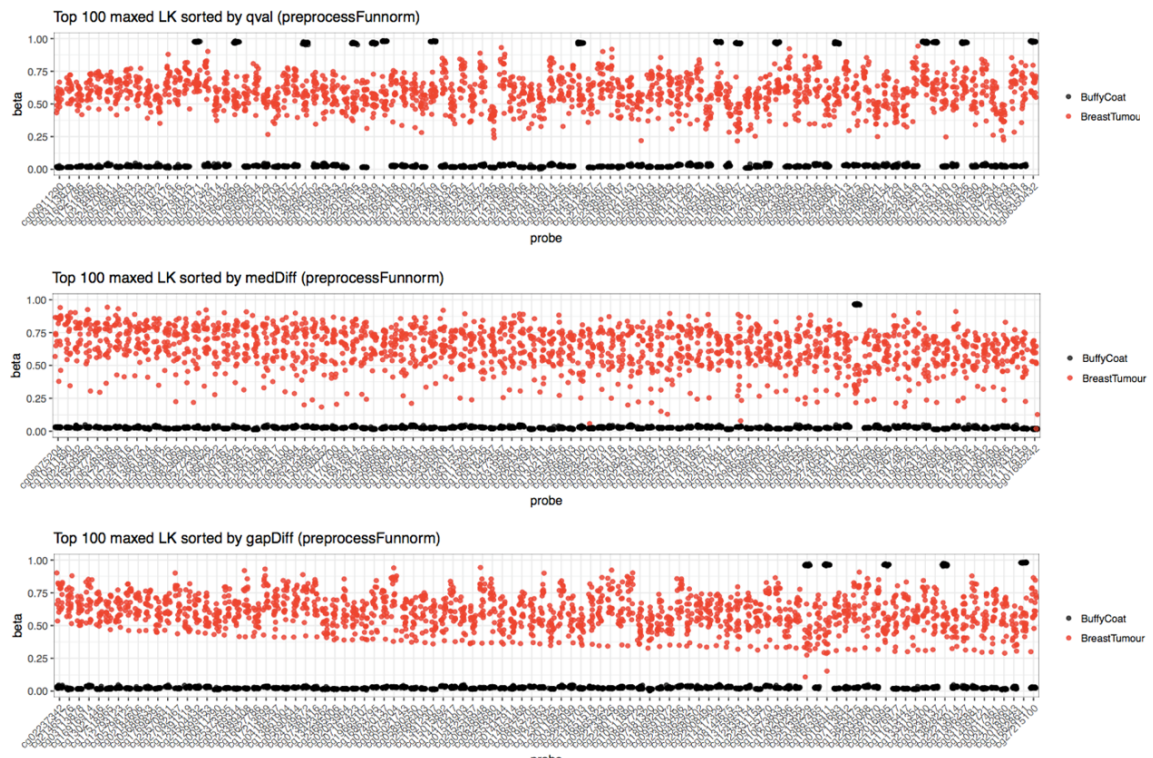


Figure 3: Dot plots showing comparisons between tumour and buffy coat of patients with breast cancer measured by EPIC array. Probes were ranked by qval, median difference and widest beta value separation between tumour and buffy coat samples using Funnorm pre-processing method. (Unpublished data from host group: Fiona Semple and Gil Tomas, IGMM, Personal communication).

2. HYPOTHESES

The working hypothesis of our group is that methylation differences observed between primary tumours and leucocytes can be translated into methylation patterns in cfDNA.

3. RESEARCH PROJECT AIMS

This research aims to establish methods to reliably detect methylation changes to allow development of cancer biomarkers for non-invasive early detection.

In order to achieve this, this research aims to (1) detect methylation differences between matched tumour samples and leucocytes of patients with breast cancer, and in cfDNA using the Fluidigm Access Array, and (2) test the sensitivity of bisulfite Sanger sequencing and bisulfite pyrosequencing as validation techniques for detecting differences in CpG methylation.

4. METHODOLOGY

The methodology of this project (Figure 4) consisted of DNA bisulfite conversion, optimization of marker design, primer pair design and validation for Fluidigm and Sanger Sequencing and Pyrosequencing as validation methods, Fluidigm Access Array Amplification, Illumina MiSeq sequencing, read alignment and methylation calling, statistical analysis and correlation of methylation data.

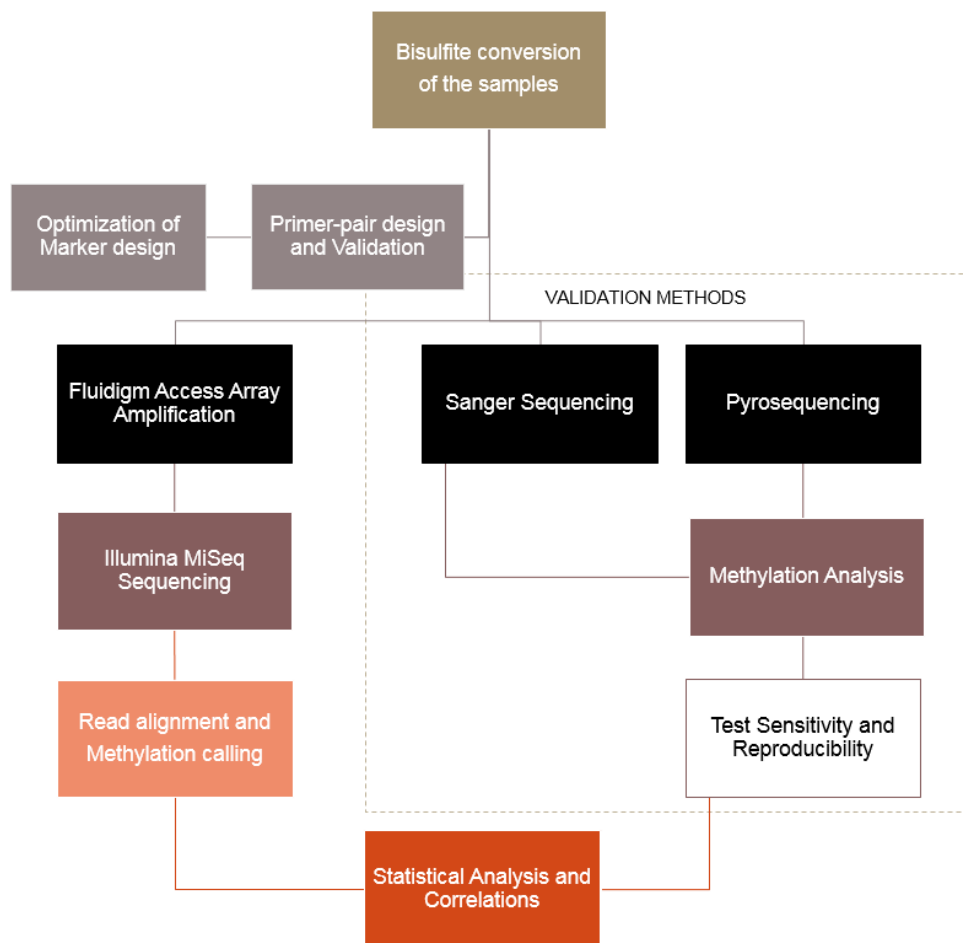


Figure 4: Research project flowchart.

4.1. Research samples

Research samples for this study were provided by Olga Oikonomidou, clinician scientist at the Edinburgh Cancer Research Center. DNA samples consisted of extracted DNA samples of breast tumour tissue and leucocytes from 10 different breast cancer patients before receiving neoadjuvant treatment. Patients ranged ages 36 to 69 years old and they were all classified as ductal carcinoma of no special type (NST) of breast cancer. In addition, 4 cfDNA samples from different healthy individuals were used in this experiment. Plasma was separated in SeraLabs laboratory by centrifugation within 30 minutes after blood draw. This prevented from lysis of leucocytes which would result in a non-cfDNA contribution to the plasma. After that, cfDNA samples were sent to the University of Edinburgh and cfDNA extraction was carried out by Dr. Fiona Semple from Institute of Genomics and Molecular Medicine (IGMM), University of Edinburgh (UoE). All the samples used had been previously isolated by Dr. Fiona Semple.

4.2. Fluidigm Access Array Amplification

4.2.1. Marker Design and Optimization for Fluidigm Access Array

The original raw data from the Illumina EPIC methylation array was re-analysed with the package `minfi`¹⁰¹ using an alternative pipeline (Fig. 3). The normalization of the data from the Illumina EPIC array was optimized using the `Funnorm` pre-processing method¹⁰², which is recommended for datasets where global biological methylation differences exist between samples, such as cancer and normal tissues.

In addition, `minfi` provides a quality control based on the log median intensity coming from CpGs in the EPIC array, where good samples may cluster together, while failed samples tend to separate and have lower median intensities¹⁰¹. After that, questionable probes such as those containing polymorphisms and those with potential to cross-hybridise were also removed.

Probes having extreme beta-values in leucocyte samples (i.e. methylation levels of less than 5% and more than 95%) were ranked by q-value, median

differences and widest beta value separation between tumour and buffy coat samples (Figure 5). Q-value provided statistical robustness for the assessment of methylation differences between tumour and healthy samples by the EPIC array; median differences accounted for the variability on CpG methylation among the different samples; the widest beta value provided the maximum difference in methylation between cancer and healthy samples. Hence, these criteria aimed to identify the probes that could be used for reliably distinguishing methylated cytosines in tumour DNA in a background of non-methylated cytosines from the non-tumour DNA in a tumour sample or vice versa. Thus, unique probes intersecting these three criteria made up a total of 168 probes for primer design and validation.

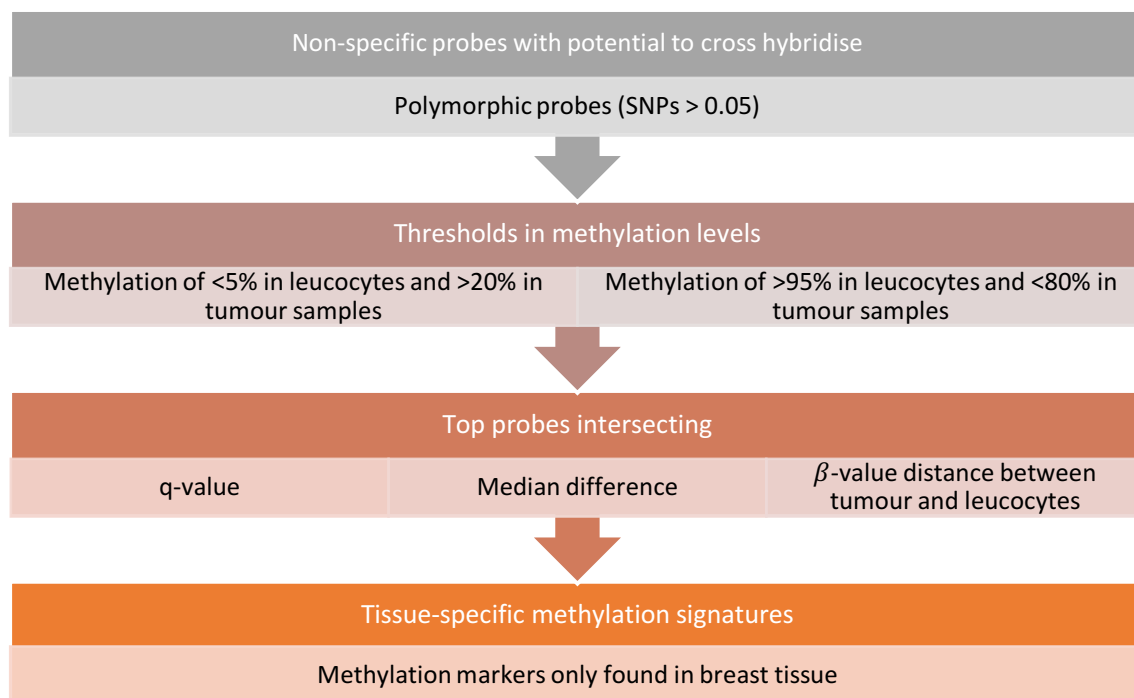


Figure 5: Pipeline for optimal marker design for CpG dinucleotide methylation detection.

4.2.2. PCR Primer Design and Validation

Human Genome data was obtained from University of California, Santa Cruz (UCSC) genome database. Human genome from December 2013 assembly (GRCh38/hg38) was used as donor sequence for primer design. Strand specific primers were designed across the CpGs of interest by using PrimerSuite primer design tool¹⁰³.

Primers were designed to be between 20bp and 30bp, with a melting temperature (T_m) between 60°C and 64°C and to avoid CpGs in the primer sequences. As cfDNA is overall consistently shorter than the fragment length of normal cell-free DNA¹⁰⁴, the amplicon size ranged between 125bp and 140 bp. They were tested by *in silico* PCR against the bisulfite converted human genome (UCSC hg38). The pipeline allows selection of primer pairs that produce one specific amplicon and removal of products larger than 135bp. Designed primer pairs (Appendix 1) were added with common sequence tags for Fluidigm analysis. The common sequence tag for forward and reverse primer were 5'- ACACTGACGACATGGTTCTACA-3' and 5'- TACGGTAGCAGAGACTTGGTCT-3', respectively.

Primer pairs were validated by PCR amplification with bisulfite converted human genomic DNA to ensure primers could successfully amplify the regions of interest. The PCR reaction consisted of 1µl of primer solutions (1µM CS1- TS Forward Primer and 1µM CS2-TS Reverse Primer) and 4µl primer validation mixture (1x FastStart High Fidelity Reaction Buffer without MgCl₂, 4.5mM MgCl₂, 5% DMSO, 200µM PCR Grade Nucleotide Mix, 0.05U/µL FastStart High Fidelity Enzyme Blend, 1x Access Array Loading Reagent, 8.3ng/µl bisulfite converted genomic DNA and PCR grade dH₂O). PCR cycles consisted of a cycle of 70°C 20 minutes, a cycle of pre-denaturation at 95°C for 10 minutes, 40 cycles of denaturation, annealing and extension at 95°C for 15 seconds, 57°C for 30 seconds and 72°C for 60 seconds, respectively, and a final extension cycle at 72°C for 2 minutes. PCR products were run on 1,5% (w/v) agarose gel electrophoresis with 10µl SybrSafe per 100ml of gel. DNA ladder used in the experiment was 100bp quickload DNA ladder (NEB). Gels were visualised under UV light in gel documentation system.

4.2.3. Preparation of Fully Methylated and Unmethylated DNA controls

Fully methylated human genomic DNA was commercially available (Roche). Fully unmethylated DNA was prepared using the REPLI-g whole-genome amplification kit (Qiagen) following the manufacturer's instructions.

4.2.4. Bisulfite Conversion of DNA

DNA bisulfite conversion was carried out to enable quantification of CpG dinucleotide methylation. The bisulfite conversion comprises DNA denaturation, bisulfite deamination, desulfonation and several washing steps. DNA of each sample was bisulfite-converted using MethylCode™ Bisulfite Conversion kit (Life Technologies) according to the manufacturer's instruction with modifications. Modifications involved the use of a total of 250ng diluted in 20 µl water for paired tumour and leucocyte samples, and 20 µl of cfDNA sample as input for DNA bisulfite conversion. Centrifugation steps were carried out at 12000g, and a dry centrifugation step at 15000g was carried out before elution with new collection tube.

4.2.5. Fluidigm Access Array Amplification

Fluidigm Access Array, a 48.48 microfluidics technology, was used to amplify samples of bisulfite converted DNA with validated primer pairs. Primer solutions 20x were prepared from 50µM CS1-TS forward primer, 50µM CS2-TS reverse primer, 20x access array loading reagent and TE buffer with final concentration of 1µM for each primer, 1x access array loading reagent and a total volume of 100µl. Primer solutions 20x were mixed with vortex for 30 seconds and centrifuged briefly to spin down all components.

Fluidigm master mix contained 1x FastStart High Fidelity Reaction Buffer without MgCl₂, 4.5mM MgCl₂, 5% DMSO, 200µM PCR Grade Nucleotide Mix, 0.05U/µL FastStart High Fidelity Enzyme Blend, 1x Access Array Loading Reagent and dH₂O. A total volume of 3.5µl was aliquoted into 96-well plate for each DNA wells. Bisulfite converted DNA samples were added into master mix in each well of 96-well plate. Two technical replicates were used in 8 out of 10 paired tumour and leucocyte samples. Technical replicates involved same bisulfite-converted DNA sample analyzed twice by using two different wells in the Fluidigm plate. CfDNA samples were not replicated.

	B41	B41	T41	T41	B84	T84
	B55	B55	T55	T55	B86	T86
	B66	B66	T66	T66	FM.ca	FUM.ca
	B69	B69	T69	T69	FM	FUM
	B72	B72	T72	T72	Cf20	Cf85
	B77	B77	T77	T77	Cf86	Cf87
#	B82	B82	T82	T82	w	w
	B83	B83	T83	T83	w	w

Figure 6: Primer plate designs of the samples and primer pairs (1B) used in Fluidigm Access Array amplification. “B” and “T” samples correspond to buffy coat and tumour matched samples from breast cancer patients, respectively; FM and FUM are the fully methylated and fully unmethylated DNA controls, respectively, and “cf” corresponds to cfDNA from plasma of healthy individuals.

A 48.48 access array integrated fluidic circuit (IFC) was injected with control line fluid and added with 500µl harvest solution in well H1-H4. IFC was primed in pre-PCR IFC controller AX with Prime (151x) program.

Primed IFC was added with 4µl of sample mix in left inlets and primer mix in right inlets (Figure 7). Samples and primer pairs were mixed by using pre-PCR IFC controller AX with Load Mix (151x) program. Mixed samples and primer pairs were put in thermocycler. PCR cycles consisted of a cycle of 70°C for 20 minutes, pre-denaturation at 95°C for 10 minutes, and 40 cycles of denaturation, annealing and extension at 95°C for 15 seconds, 57°C for 30 seconds and 72°C for 60 seconds respectively.

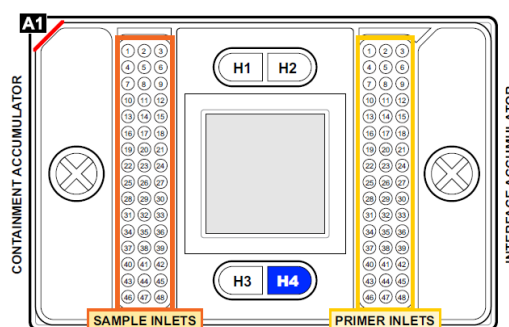


Figure 7: Sample and primer inlets in the Fluidigm Access Array 48.48.

Access array harvest reagents in H1-H4 wells of IFC were replaced with 600µL of fresh reagent. Each sample inlets were added with 1µl of 1x access array harvest reagent. Samples were harvested by putting IFC in post-PCR IFC controller AX with Harvest (151x) program. Samples were transferred into a 96-well plate.

4.1.7. Barcoding and AMPure Clean-up

Bisulfite-converted DNA samples were barcoded. Barcoding PCR master mix contained 1x FastStart High Fidelity reaction buffer without MgCl₂, 4.5mM MgCl₂, 5% (v/v) DMSO, 200mM of each nucleotide from PCR grade nucleotide mix, 0.05U/μl FastStart High Fidelity enzyme blend and water to a total reaction volume of 14μl. A mixture of 14μl mastermix, 4μl of barcode and 2μl of bisulfite-converted DNA samples was prepared in each well for each sample. The mixture was spun down and put in thermocycler. PCR program consisted of a cycle of pre-denaturation at 95°C for 10 minutes, 15 cycles of denaturation, annealing, and extension at 95°C for 15 seconds, 60°C for 30 seconds, and 72°C for 60 seconds respectively, and a cycle of final extension at 72°C for 3 minutes.

Barcoded samples were pooled into a tube by transferring 4μl for each sample. Pooled barcoded sample was purified with AMPure SPRI magnetic beads. A total volume of 12μl of pooled sample was added with 24μL of water and 43.5μl of AMPure beads. Mixture was incubated at room temperature for 15 minutes, applied to magnetic field for 15 minutes and 70μl of the liquid was removed from the tube. DNA bound to magnetic beads was washed with 200μl of 80% (v/v) ethanol twice, air-dried and re-suspended in 40μl of water. Samples were mixed with vortex briefly, incubated at room temperature for 5 minutes, and applied to magnetic field for 5 minutes. Finally, 35μl of purified pooled barcoded DNA was collected.

The pre and post cleaned-up barcoded libraries, ready for sequencing, and random samples pre and post-barcode diluted by adding 40 and 10 μl water, respectively, were analyzed with Agilent Bioanalyser DNA HS chip to determine Fluidigm product size at the Wellcome Trust Clinical Research Facility, UoE.

Qubit HS DNA quantification was carried out at pre- and post-barcode random individual samples and DNA library pools to verify the presence of Fluidigm product.

4.1.8. Illumina MiSeq Sequencing, Read Alignment and Methylation Calling

Fluidigm libraries were sequenced with paired-end Illumina MiSeq Next Generation Sequencing (David Ross at Edinburgh Clinical Genetics). Read alignment, methylation calling and analysis of bisulfite sequencing data were carried out according to the analysis pipeline of Dr. Duncan Sproul, University of Edinburgh (unpublished).

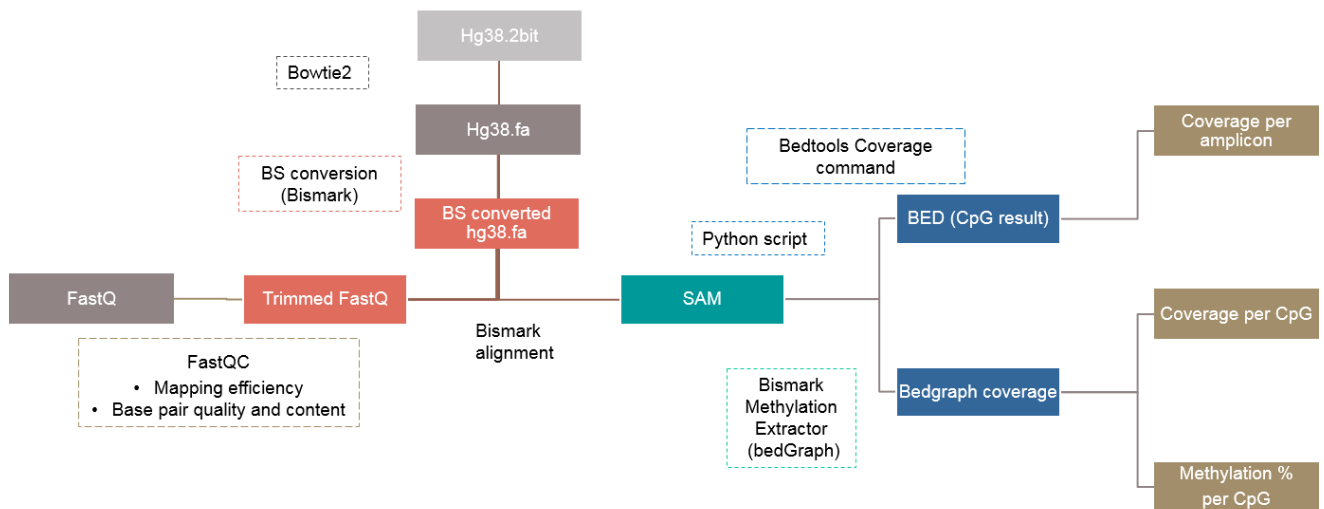


Figure 8: Pipeline used to analyse Fluidigm-MiSeq sequencing data. FastQ files derived from bisulfite converted targeted Illumina MiSeq sequencing were the input files to obtain coverage and methylation percentage for each CpG of interest.

Human genome December 2013 assembly¹⁰⁵ (GRCh38/hg38) was downloaded from University of California Santa Cruz (UCSC) genome database¹⁰⁶. Genome was bisulfite converted and prepared with bismark v0.14.3 and bowtie2 according to bismark user guide manual with default options. FastQC software provided assessment of the quality of the sequenced bases and Phred scores were used to exclude low quality reads. Then, adapter trimming and alignment against the reference genome were performed. Alignment of sequencing results to bisulfite converted genome was also performed with bismark v0.14.3 with default options to obtain SAM files. SAM files of each target amplicon were processed (python script) and BEDtools v2.23.0 with bedtools coverage command and default options to obtain coverage of each amplicon. SAM files were also processed with bismark methylation extractor command from bismark v.0.14.3 according to the user

guide with default options. Output of bismark methylation extractor was processed (perl script) to separate total methylated and unmethylated amplicons from the files. Output of perl script was BED files. BED files were processed with BEDtools v.2.23.0 coverage command to obtain total methylated and unmethylated per target amplicons. Bismark methylation extractor command was used to obtain individual CpG methylation percentages. Data analysis was performed in UNIX programming language. Python and Perl scripts used above were provided by Dr. Duncan Sproul and automation of commands was assisted by Dr. David Parry (IGMM, University of Edinburgh). Methylation percentage was calculated with formula:

$$\text{Methylation \%} = \frac{\text{Total methylated amplicons}}{(\text{Total methylated amplicons} + \text{total unmethylated amplicons})} \times 100$$

4.1.9. Data Analysis

Bisulfite sequencing data was then analyzed based on successful amplification of the samples and complete bisulfite conversion. Successful amplification of the FMD, FUM and water controls were analyzed to ensure little contamination in water used in the experiment and an indicator of successful sequencing. After that, samples and amplicons were filtered based on the number of reads per amplicon to ensure amplification during the Fluidigm experiment. In addition, CpGs were filtered based on successful bisulfite conversion in both the FUD and FMD positive controls.

Moreover, this methylation study aimed to detect differentially methylated CpGs between tumour and leucocyte samples and in cfDNA. Methylation levels for the CpGs of interest are unknown and may be different between the different tissues studied. Moreover, the variation present in the Fluidigm platform poses the challenge of determining whether the differences in methylation are caused by biological differences or by statistical chance. The best way to address this challenge is to use technical replicates of the different samples. Thus, two technical replicates were used for 8 out of 10 paired tumour and leucocyte samples included in the experiment. Pearson tests were performed for correlation analysis by using R studio. Technical replicates allowed averaging

methylation data from CpGs and thus increased the power to detect differentially methylated CpGs.

4.1.10. Statistical Analysis of CpG Dinucleotide Methylation

Statistical analysis was performed using R studio. Normality distribution of leucocyte, tumour and cfDNA data was analyzed with Shapiro Wilk normality test with 95% confidence interval. Paired test was only performed to CpG dinucleotide methylation data that passed all the filtering criteria. Comparisons were carried out to samples with positive strong correlation between technical replicates. P-values were adjusted by using the “BH” method (Benjamini, Hochberg, and Yekutieli) to control the false discovery rate.

4.3. VALIDATION METHODS

Specific and sensitive analytical procedures must be developed and optimized to target circulating molecules to show methylation differences between patients and healthy subjects. Most of these recent efforts rely on the methylation level of individual CpG sites, and they are fundamentally limited by the technical noise and sensitivity in measuring single-CpG methylation⁸⁶. In order to establish a validation method for CpG dinucleotide methylation detection in tumour-cfDNA, Sanger sequencing and Pyrosequencing were performed to test their sensitivity and reproducibility for detecting small methylation differences. Probes that showed clear methylation differences between tumour and leucocytes were selected, and primer pairs were designed independently for each assay to test whether these methods are able to produce results comparable to those produced by the EPIC methylation array and Fluidigm Access Array.

Detailed information regarding the principle of Sanger sequencing and Pyrosequencing is given in Appendix 1.

4.3.1. Preparation of Fully Methylated and Fully Unmethylated Genomic DNA Mixtures and Quantification

Fully methylated human genomic DNA was commercially available (Roche). Fully unmethylated DNA was prepared using the REPLI-g whole-genome amplification kit (Qiagen) following the manufacturer's instructions (as described above) and the purified with simple DNA precipitation. A total of 1/25 volume sodium acetate 3M and 1 volume of 100% ethanol were added into the DNA. The mixture was incubated overnight at -20°C, and then centrifuged at 12000rpm for 20 minutes. The pellet was washed with 70% (v/v) ethanol, centrifuged at 12000rpm for 20 minutes, air-dried and re-suspended in 100µl of dH₂O. Purified fully unmethylated DNA was quantified using Qubit dsDNA broad range assay kit and the concentration of both fully methylated and unmethylated DNA were adjusted to 100ng/µl. When purified fully unmethylated DNA was too diluted, the vacuum system in Human Genetics Unit (HGU) was used to reach 100 ng/µl .

Both methylated and unmethylated DNA were mixed as shown in table 1 to sum

the volume up to 5 µl, and the final concentration of 500 ng/µl, which is the optimal for performing the bisulfite conversion, was achieved. DNA concentration was measured with the Qubit ssDNA broad range assay kit (Life Technologies) according to the manufacturer's instructions.

Table 1: Percentages of fully methylated (FM) and unmethylated (UM) DNA used to make the samples. The asterisk * indicates that these samples have a replicate.

Sample name	FM DNA (%)	UM DNA (%)
1	0	100
2*	10	90
3	25	75
4*	50	50
5	75	25
6*	90	10
7	100	0

Batch effect was avoided by replicating samples 2, 4 and 6, from two independent bisulfite treatments, with separate PCR amplification, as well as Sanger sequencing and Pyrosequencing experiments were performed for each replicate.

4.3.2. Primer design and validation for Bisulfite Sanger sequencing

9 probes that showed well defined differences (Figure 5) in methylation between tumour and leucocytes in the EPIC array were randomly selected to be analysed by Sanger sequencing.

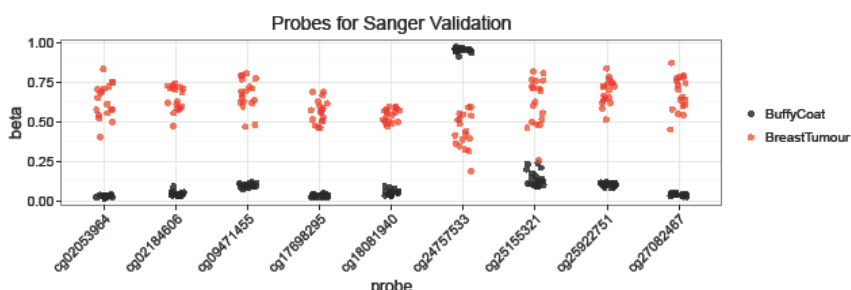


Figure 10: Dot plot showing comparisons between tumour and buffy coat of patients with breast cancer measured by EPIC array. The 9 probes obtained median differences in methylation of more than 40%. Previous data from host group (Fiona Semple and Gil Tomas, IGMM, Personal communication).

Primers were designed as described above except amplicon size 150 bp and 250 bp across the CpGs of interest by using Bisulfite Primer Seeker Tool¹⁰⁷. Then, primer pairs were tested by *in silico* PCR against the reference human genome (Hg38) by using BiSearch primer design and search tool¹⁰⁸. This software performs *in silico* bisulfite conversion prior to primer design. It should be noted that primer design is strand specific. Therefore, if the designed primers are complementary to the DNA reverse strand and the reverse strand sequence serves as an input data, the settings of the software should remain unchanged (“default-forward strand”) otherwise the designed primers will be complementary to the forward strand¹⁰⁸.

After that, primers were selected looking for a single match on the appropriate strand. Primer pairs that passed the *in silico* PCR step were ordered from Sigma. Primer pairs were then validated by PCR amplification to ensure primers could successfully amplify these regions, and the best annealing temperature was chosen for each pair of primers with human genomic DNA. The PCR reaction consisted of 2 µl primer solutions, 19 µl of PCR grad distilled water, 25 µl of Zymo Taq PreMix and 2 µl of the template DNA which was the bisulfite converted DNA of each mixture. The PCR cycles are listed in Table 2.

Finally, the validated primers pairs and the PCR products of the samples were sent for Sanger sequencing at Edinburgh Genomics. The results were analysed using the Sequencher 5.4.6.

4.3.3. Sanger Sequencing Analysis

Sanger DNA sequencing electropherograms allowed the identification of methylated cytosines, which would appear as cytosine in the traces, whereas thymine nucleotides would appear in unmethylated cytosines.

Quantitative information from Sanger sequencing traces was obtained by using the ab1PeakReporter web-based tool¹⁰⁹, which converted Sanger sequencing trace files into comma separated value (.csv) files. These output files contained the peak height and quality values for each nucleotide and peak height ratios for all four bases at any given locus allowing the detection and assessment of small changes in methylation at any given allele.

4.3.4. Primer design and validation for Bisulfite Pyrosequencing

11 probes that showed well-defined differences in methylation between tumour and leucocytes in both the EPIC array and the Fluidigm experiment were selected to be analyzed by Bisulfite Pyrosequencing. None of these probes were analysed by Sanger Sequencing, as new criteria for optimal marker design was developed after Sanger Sequencing was performed during the Master project.

Human Genome data was obtained from University of California, Santa Cruz (UCSC) genome database. Human genome from December 2013 assembly (GRCh38/hg38) was used as donor sequence for primer design. Assays were designed using the PyroMark Assay Design and PyroMark 24 Software (Qiagen), programs provided by Dr. Alex Adams. This software automatically performed the assay design including PCR primers and pyrosequencing primers. The size of the amplification product was restricted to 130 bp or less, as otherwise secondary structures such as loops can be formed in the single-stranded template which could interfere with or inhibit the sequencing reaction or increase the background signal due to the extension of the 3'-end of the terminus. In addition, capture efficiency of the biotinylated amplification product decreases with size.

Primers were designed to be between 15bp and 30bp, with a melting temperature (T_m) between 60°C and 64°C and to avoid CpGs in the primer sequences. In addition, the pyrosequencing software incorporated internal controls for bisulfite treatment.

4.3.5. Bisulfite pyrosequencing

Highly quantitative bisulfite pyrosequencing¹¹⁰ was performed using the PyroMark Q24 system (Qiagen, Valencia, CA, USA). Bisulfite Pyrosequencing was performed by William Hawkins at the Wellcome Trust Clinical Research Facility of UoE.

Each assay was validated by means of a series of standards of 0, 25, 50, 75 and 100%-methylated DNA. The standards were also mixtures created from whole genome amplified DNA using the REPLI-g whole-genome amplification kit (Qiagen), representing 0% methylation, and commercially available fully methylated human genomic DNA (Roche), representing 100% methylation, which were mixed in relative proportions to create the same mixtures used in Sanger Sequencing (Table 1). Only the primer sets that worked well in the Pyrosequencing with the DNA mixtures were then tested in five bisulfite-converted DNA paired samples of tumour and leucocytes, and four cfDNAs from healthy individuals.

5. RESULTS

5.1. Marker Design Optimization

Marker design optimisation was performed to select for CpG sites used in the subsequent analysis. The original raw data from the Illumina EPIC methylation array was re-analysed with the package minfi. The starting point of minfi is reading the .IDAT files, with the built-in function *read.metharray*. Probes containing polymorphisms and those with potential to cross-hybridise were also removed. Minfi provides a quality control based on the log median intensity coming from CpGs in the EPIC array. Good samples clustered together, while failed samples obtained lower median intensities. As can be seen in figure 11, samples 1 and 18 failed the quality control and were excluded for downstream analysis.

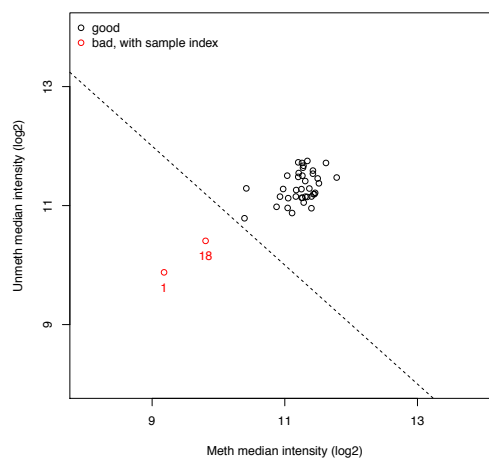


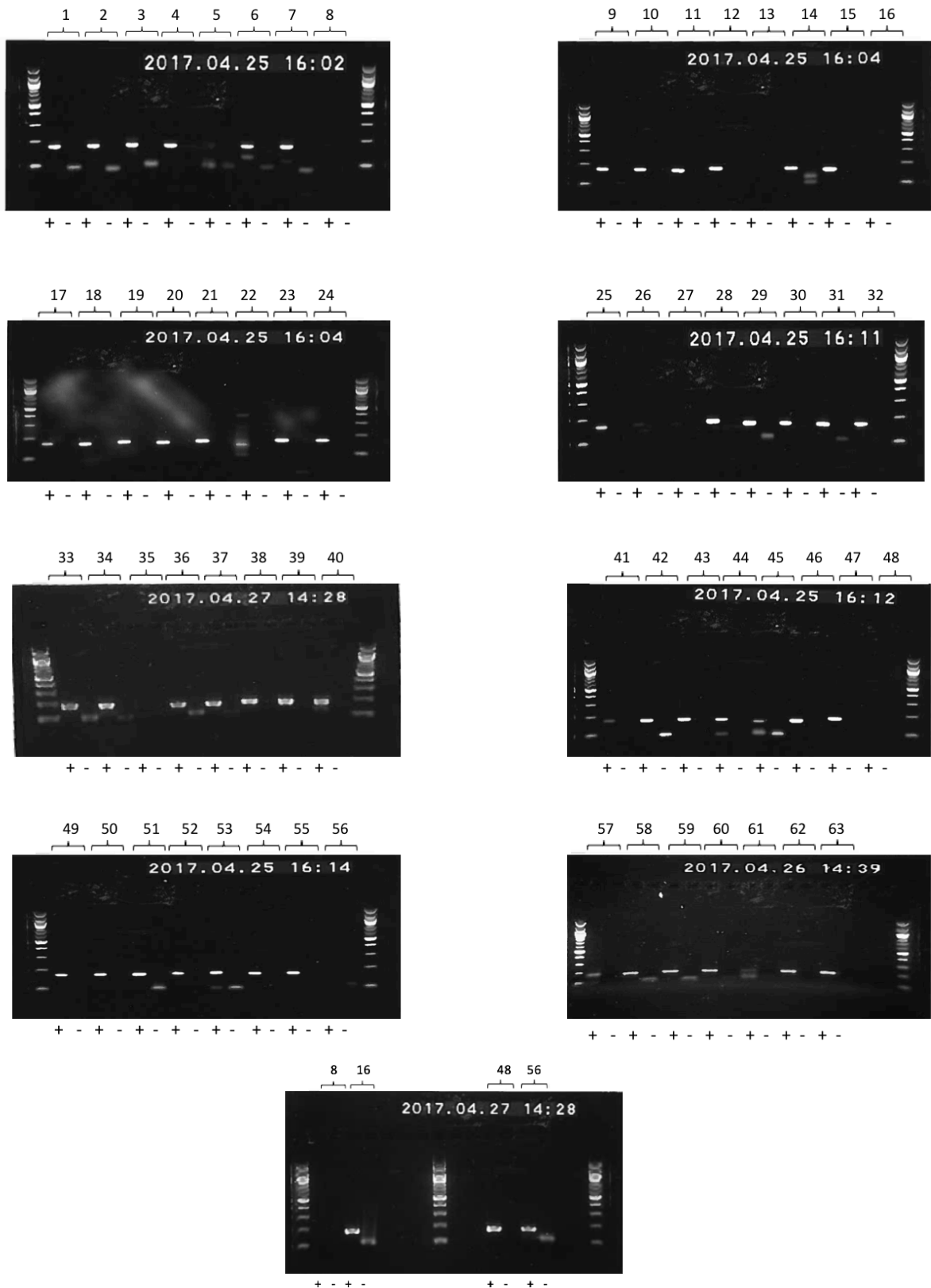
Figure 11: Quality control of the samples used in the 850k EPIC array provided by minfi package. All samples passed the QC based on the median intensities from the CpGs except from samples 1 and 18.

The data was normalized by using Funnorm preprocessing method, which uses internal control probes present on the array to infer between-array technical variation. The input for the function *preprocessFunnorm* is a *RGChannelSet*, which is the initial object for the analysis containing the raw intensities in the green and red channels. The phenotype data, such as sample names, sample wells, array, slides and probes was accessed via the accessor command *pData*.

5.2. Fluidigm Access Array Amplification

5.2.1. Primer design and validation for Fluidigm Access Array Amplification

Primer design was performed using the PrimerSuite tool, which provided a set of 168 primer pairs targeting 23 probes showing differential methylation between breast tumour and buffy coat in the EPIC array. Primer validation was performed to ensure that primer pairs produced one specific amplifon. Primer pairs that passed the *in silico* PCR step were validated by PCR amplification with bisulfite converted human genomic DNA (hgDNA) to ensure primers could successfully amplify the CpGs of interest.



Figures 13A-13I: Agarose gels (1,5% (w/v)) showing amplicons of 130 bp after PCR amplification with 63 different primers pairs. “+” lanes are PCR reactions with human genomic DNA, “-”lanes are water blanks (negative controls). Figure 13I shows repetition of five primers that where in the last rows that may have been evaporated in the other PCR reactions.

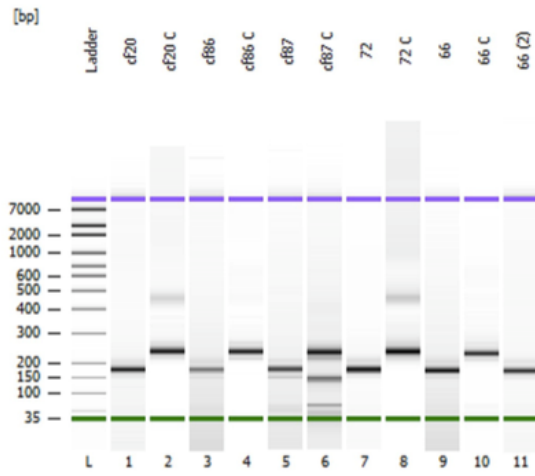
There were 168 potential targets however due to the limitations of bisulfite primer design only 87 primer pairs could be designed to comply with the parameters for Fluidigm amplification. After validation by *in silico* PCR, 63 primer pairs were tested for PCR amplification with hgDNA (Figures 13A-13I). A final primer set of 23 primer pairs (Appendix 2) were chosen in order to amplify 23 CpGs that showed differential dinucleotide CpG methylation in the EPIC array. Together with these, 25 primer pairs that previously amplified CpGs of interest were used as positive controls to show successful amplification in the Fluidigm. These made up a total of 48 targets to test on the Fluidigm.

5.2.2. Fluidigm Library and Quality Control

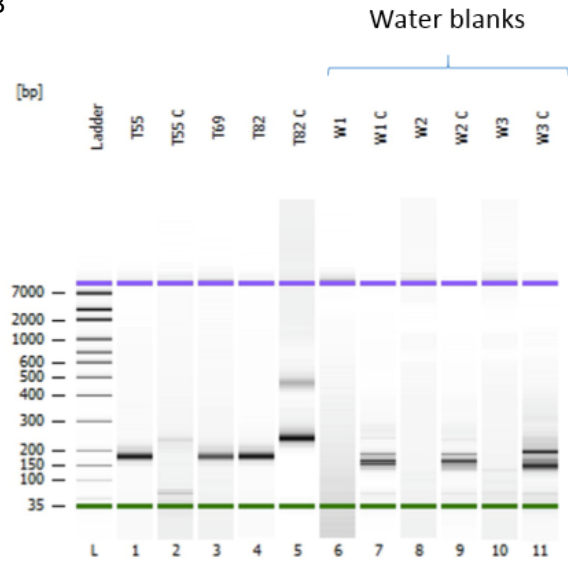
DNA concentrations of randomly selected Fluidigm amplicons were checked with Qubit HS dsDNA assay to ensure successful PCR amplification. After barcoding, another set of randomly selected Fluidigm samples and the pre and post cleaned-up libraries were checked with bioanalyser prior to next generation sequencing to confirm expected product size in each Fluidigm stage.

DNA concentrations of samples increased from post-Fluidigm pre-barcode stage to post-Fluidigm post-barcode stage. Negligible amount of amplicon was generated in water sample during Fluidigm amplification. The Fluidigm library concentration was higher pre-SPRI clean-up than post-SPRI clean-up. This indicated successful removal of non-amplicon DNA such as primer dimers. DNA concentration of the library ranged from 28.4 ng/μl before clean-up to 9.2ng/μl after clean-up.

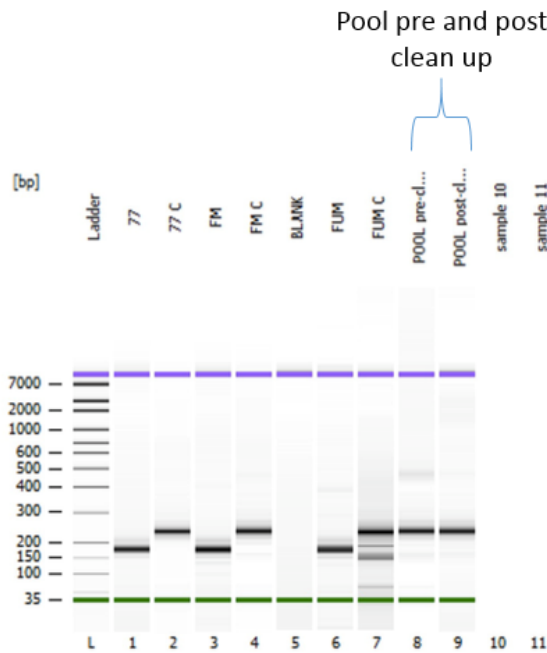
14A



14B



14C

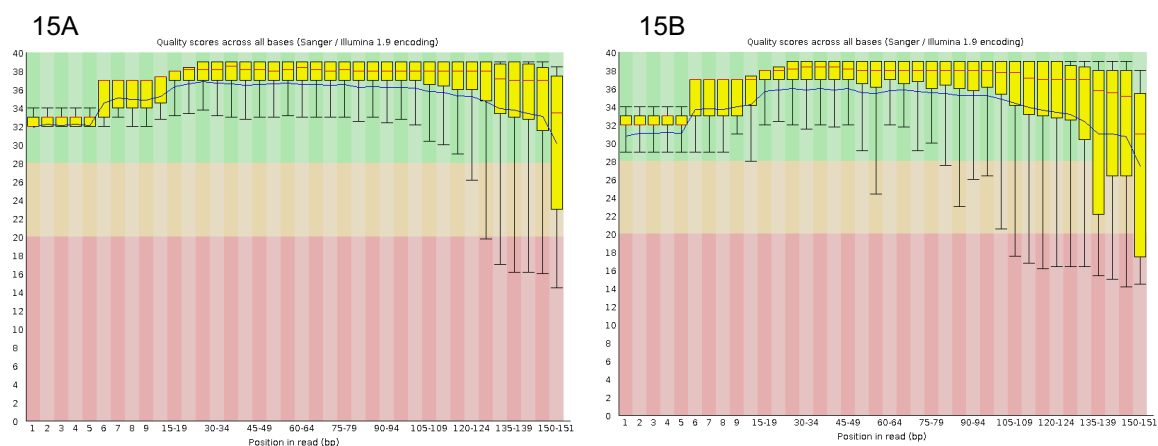


Figures 14A-14C: DNA bioanalyser traces before and after barcoding of the Fluidigm library. Numbers correspond to matched tumour and leucocyte samples from breast cancer patients; “cf” corresponds to cfDNA samples from healthy individuals; “C” samples correspond to barcoded samples; pool pre-cleaned up and pool post-cleaned up correspond to pooled post-barcode post-Fluidigm samples before and after the clean-up step; FM and FUM are the fully methylated and fully unmethylated DNA controls, respectively and “w” correspond to the water controls used in the Fluidigm.

Bioanalysis results showed the expected amplicon size in all random individual samples (figures 14A-14C): 220-250bp in post-Fluidigm pre-barcode, 270-300bp in post-Fluidigm post-barcode, pooled post-barcode post cleaned-up library (270-300bp). The four water blanks showed no amplicon in post-Fluidigm pre-barcode stage and the expected low background amplification from barcoding of the unused primer pairs in the post barcode samples.

5.2.3. Quality Control of Illumina MiSeq raw data

Amplicon coverage, methylated amplicon counts and CpG methylation counts were obtained from paired-end Illumina MiSeq next generation sequencing (NGS) data analysis. The paired-end Illumina MiSeq dataset consists of the sequencing of both ends of the same fragment. Therefore, the sequences came in two data files, one with the forward paired-end sequences and one with the corresponding reverse paired-end sequences. The FastQC software provided assessment of the quality of the sequenced bases based on the Phred quality score, which is logarithmically related to the probability of an error of the identification of the nucleobases generated by automated DNA sequencing. A score of 10 means a 10% error probability and 30 means a 0.1% chance, etc. A score of 20 is generally accepted as the minimum acceptable score. The FastQC for all the FastQ files obtained a high Phred score over 30 or 40 (Figures 15A and 15B).



Figures 15A, 15B: Phred quality scores of the raw data for the forward and the reverse reads of sample B55. Quality scores across bases obtained >30, except the last bases of reverse reads due to the nature of the Illumina Sequencing platform.

However, reverse reads for all the samples obtained lower Phred scores in the last bases. This may be due to sequencing by synthesis and bridge amplification clusters of the Illumina Sequencing platform. For a variety of reasons, including decay of reagents in the sequencing machine or lose of synchronisation when synthesis of some templates lags behind that on other templates, errors can easily be accumulated in the forming strand and, consequently, the quality of base calls decrease as sequencing progresses. Ultimately, the 5' end of the reads tend to have higher quality compared to the 3' ends, and forward reads have better quality than reverse reads. Sequence bases across the whole content were also checked. Taking into account that the Illumina adaptors added for library preparation were not bisulfite converted, a higher percentage of Cytosine was expected at the end of the graph in the forward sequences. Inversely, a higher percentage of Guanine appeared in the reverse sequences. Moreover, the quality control showed overrepresented sequences. This was expected as the Fluidigm is a targeted experiment, in contrast to for example WGBS.

Following FastQC, Trim Galore was used to perform adaptor trimming in two subsequent steps. A second FastQC was run to ensure that adaptors were trimmed based on the percentage of cytosines and guanines in the forward and reverse sequences, respectively. Trimmed FASTQ sequences showed a higher percentage of thymine in the read ends indicating successful removal of adaptors.

Next step involved bismark alignment of the trimmed FastQ sequences against the bisulfite-converted Human genome from December 2013 assembly (GRCh38/hg38). Genome was bisulfite converted and prepared with bismark v0.14.3 and bowtie2. As primer pairs from PrimerSuit tool were designed to amplify the forward and some others the reverse strand, two alignments were carried out and all the next steps were performed for both lists. Bismark alignment produced SAM files by default, containing all the alignments plus methylation call strings, and text files containing alignment, methylation summaries and mapping efficiency reports. The total mapping efficiency for each sample obtained around 80% and above, meaning that sequences were correctly aligned. Water blanks were expected to be low, and the four obtained

mapping efficiencies ranging from 3% to 9%, indicator of very low contamination.

SAM files were the input files to obtain the coverage and the methylated and unmethylated counts from all the sequences. SAM files of each target amplicon were processed (python script) and BEDtools v2.23.0 with bedtools coverage command and default options to obtain coverage of each amplicon. Finally, in order to obtain the total methylated and unmethylated counts per CpG, SAM files were also processed with bismark methylation extractor command with paired end option. Output of bismark methylation extractor was also processed with perl script, which allowed splitting total methylated and unmethylated amplicons in BED files as output. BED files were processed with BEDtools v.2.23.0 coverage command to obtain total methylated and unmethylated per target amplicons.

Moreover, BED files were used in a bedgraph pipeline to obtain methylation percentage per CpG site. The last step involved parsing all the output files. In total, methylated and unmethylated counts for 568 CpG sites (taking into account that each amplicon could contain more than one CpG dinucleotide) were obtained across all 48 amplicons assayed. Further analysis focused on the new 48 CpG sites that were targeted in the EPIC array.

5.2.4. Data Analysis

A. Analysis of controls coverage

Amplicon coverage of bisulfite sequencing was obtained from Illumina MiSeq next generation sequencing data analysis as described above. Amplicon coverages of water samples were markedly lower than fully methylated DNA (FMD) and fully unmethylated DNA (FUD) controls, though not completely absent for some amplicons. This indicated very little DNA contamination in water samples.

Amplicon coverages of FMD were not significantly different and both obtained more than ten times of amplicon coverage of water samples (Figure 16). Amplicon coverage of the FUM obtained less reads compared to the FMD, but still has more than 1000 reads when comparing with the water controls. As well

as showing little DNA contamination, this indicated successful sequencing and sequence alignment of the test sample amplified DNA. The mean amplicon coverages of other samples used in Fluidigm obtained at least 1000 reads more than water samples.

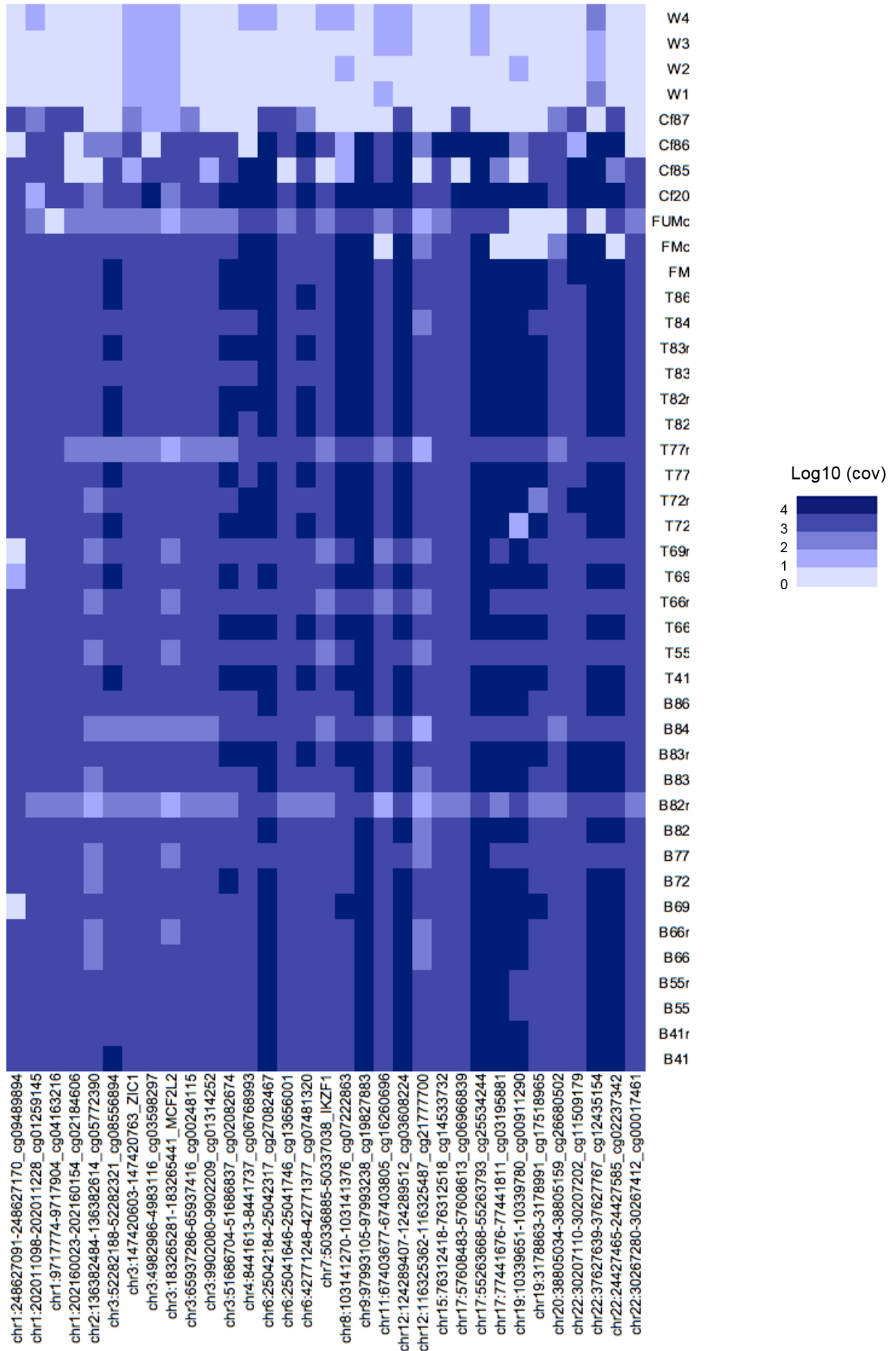


Figure 16: Heat map of the amplicon coverage of the samples. Most of the samples obtained more than 1000 reads and water controls showed less than 1000 reads. “B” samples represent leucocyte samples; “T” samples represent tumour samples; cfDNA20, cfDNA85, cfDNA86 and cfDNA87 represent cfDNA samples from healthy individuals; FM and FUM are the fully methylated and fully unmethylated DNA controls, respectively and “w” correspond to different water controls used in the Fluidigm. Samples and amplicons that did not obtain 1000 reads and did not shown complete bisulfite conversion are not shown.

B. Analysis of samples coverage

Six amplicons that did not obtain 1000 reads, reflecting poor amplification, were excluded, leaving 42 out of the initial 48 amplicons. All samples were applied to the Fluidigm platform in replicate as air bubbles in the machine can prevent sample completely entering the microfluidics chamber. In this current experiment, it appears that this was the case for 5 sample replicates (samples B69r, B72r, B77r, T41r and T55r) which had coverage below the 1000 reads threshold. In addition, one FMD sample demonstrated coverage below this threshold and was also excluded (Figure 16).

C. Analysis based on complete bisulfite conversion

FUD and FMD were used as positive controls to ensure successful bisulfite conversion. CpGs that did not demonstrate <10% methylation in the FUM control and >90% methylation in the FM control were not included in the analysis (8/42 amplicons).

D. Analysis of Technical Replicates

Technical replicates came from the same bisulfite-converted DNA sample which was analyzed twice by using two different wells in the Fluidigm plate. Correlation analysis was performed across technical replicates. Table 2 shows Pearson’s correlation coefficients of methylation percentages. All technical replicates showed a positive correlation (0.83-0.99) indicating that CpG dinucleotide methylation percentage is reproducible using Fluidigm Access Array Amplification. Samples T66 and T72 were excluded from downstream analysis as the experimental replicates did not correlate.

In future Fluidigm experiments it would be useful to include more technical replicates, as their comparison can be used to locate outlier values that may occur due to aberrations within the Fluidigm, the sample, or the experimental procedure.

Table 2: Pearson’s correlation coefficients of the methylated CpGs obtained with Fluidigm. Two technical replicates of 8 paired samples of tumour and leucocytes from breast cancer patients were included in the Fluidigm Access Array. Leucocyte samples showed a very positive strong correlation whereas tumour samples T66 and T72 obtained a poorer positive correlation.

Sample	Correlation	p-value
B41	0.996732	2.2e-16
B55	0.9814165	2.2e-16
B66	0.9906116	2.2e-16
B69	0.9976815	2.2e-16
B72	0.9972151	2.2e-16
B77	0.9975796	2.2e-16
B82	0.9952245	2.2e-16
B83	0.9889942	2.2e-16
T41	0.8280931	2.2e-16
T55	0.7455518	9.413e-16
T66	0.1361715	0.1415
T69	0.6398119	8.137e-15
T72	0.4232273	2.209e-06
T77	0.8311624	2.2e-16
T82	0.8677775	2.2e-16
T83	0.8741456	2.2e-16

5.2.5. Statistical Analysis

E. Normality of Methylation data

Individual Saphiro-Wilk tests were performed to check for normality of leucocyte, tumour and cfDNA samples’ distributions. The null-hypothesis of this test is that the population is normally distributed. Tumour samples came from a normally distributed population (p-value = 0.8948). Leucocyte samples were not normally distributed (p-value of 4.314e-06). This was expected as methylation levels for leucocyte were chosen to be > 90% or < 5% in methylation levels for marker design. CfDNA samples were neither normally distributed (p-value of 0.03474).

F. CpG Dinucleotide Methylation Analysis of Breast Tumour and Leucocyte Samples

CpG dinucleotide methylation status for 30 probes of breast tumour and leucocyte was performed by using Wilcoxon signed rank non-parametric t-test. Only CpGs that were also targeted by the EPIC array were analysed. All CpG sites across the genome were differentially methylated (figure 17) (Appendix 3).

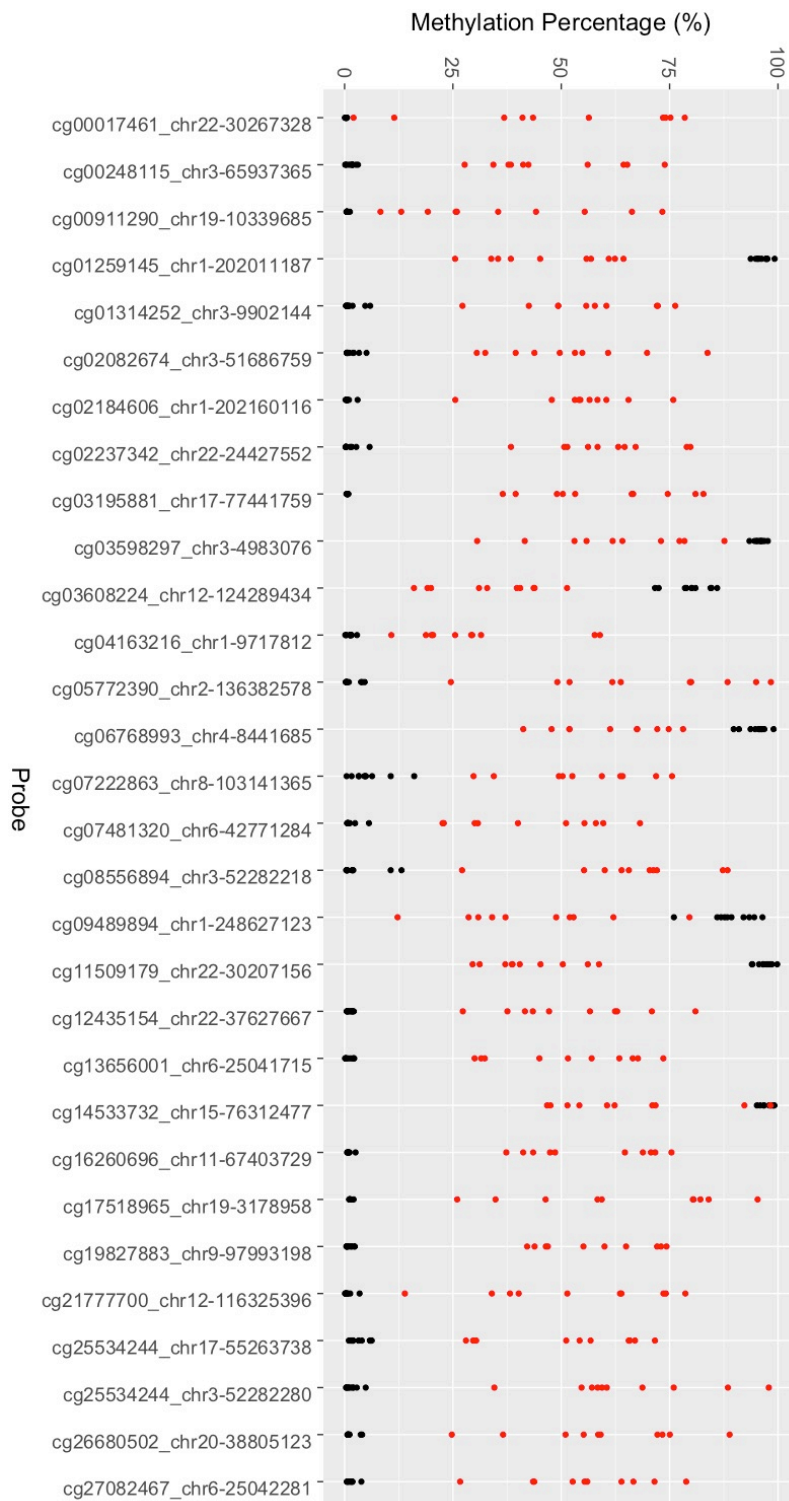


Figure 17: Differential Methylation between paired tumour and leucocyte samples on Fluidigm Access Array. Tumour samples (red dots) showed clear differences in methylation compared to leucocytes (black dots). Probes and genomic positions for each targeted CpG on GRCh38/hg38 Human Genome are provided.

G. CpG Dinucleotide Methylation Analysis of cfDNA samples

From 10ml of blood generally 5-50ng cfDNA is recovered. To test the possibility that less than 5ng cfDNA can be successfully amplified on the Fluidigm platform, amplicon coverage of four cfDNA samples (containing ~5ng cfDNA) was analyzed (Figure 18). CfDNA20, cfDNA85 and cfDNA86 samples produced more than 1000 amplicon coverage. CfDNA87 showed inconsistent amplification for most of the amplicons (only 10 amplicons demonstrated coverage >1000 reads). Failure in amplification of this cfDNA sample may be due to pipetting error in preparation of the primer plate for the Fluidigm assay. Successful amplification of cfDNA20, cfDNA85 and cfDNA86 indicated that even at very low concentrations, the Fluidigm assay can detect CpG dinucleotide methylation.

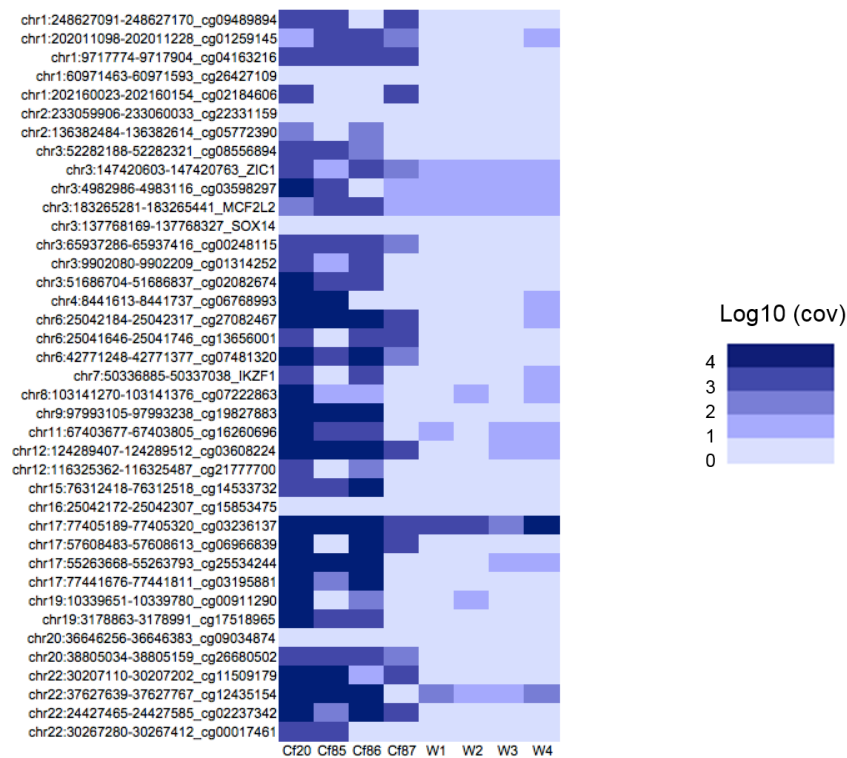


Figure 18: Heat map of the coverage of cfDNA samples and water controls. Most of the amplicons obtained more than 1000 reads compared to water controls. Cf20, Cf85, Cf86 and Cf87 represent cfDNA samples from healthy individuals and “w” correspond to different water controls used in the Fluidigm. The six amplicons excluded based on the analysis of samples coverage are included in this figure.

A Wilcoxon signed rank test was performed to test whether or not CpG dinucleotide methylation was different between cfDNA from healthy plasma and tumour samples. As previously said, only CpGs that were also targeted by the

EPIC array were analysed. P-values (Appendix 4) were significant for 21 CpGs with an alpha level of 0.05. The same test was performed to compare healthy plasma cfDNA with leucocytes. All CpG sites across the genome were not differentially methylated (Appendix 5). These results are shown in figure 19, where some tumour samples have similar methylation levels to those of leucocytes and cfDNA samples. These probes will be excluded from further analysis. The remaining CpGs will be the best candidates identified in this study for useful biomarkers for early detection of breast cancer in an assay of cfDNA obtained from blood.

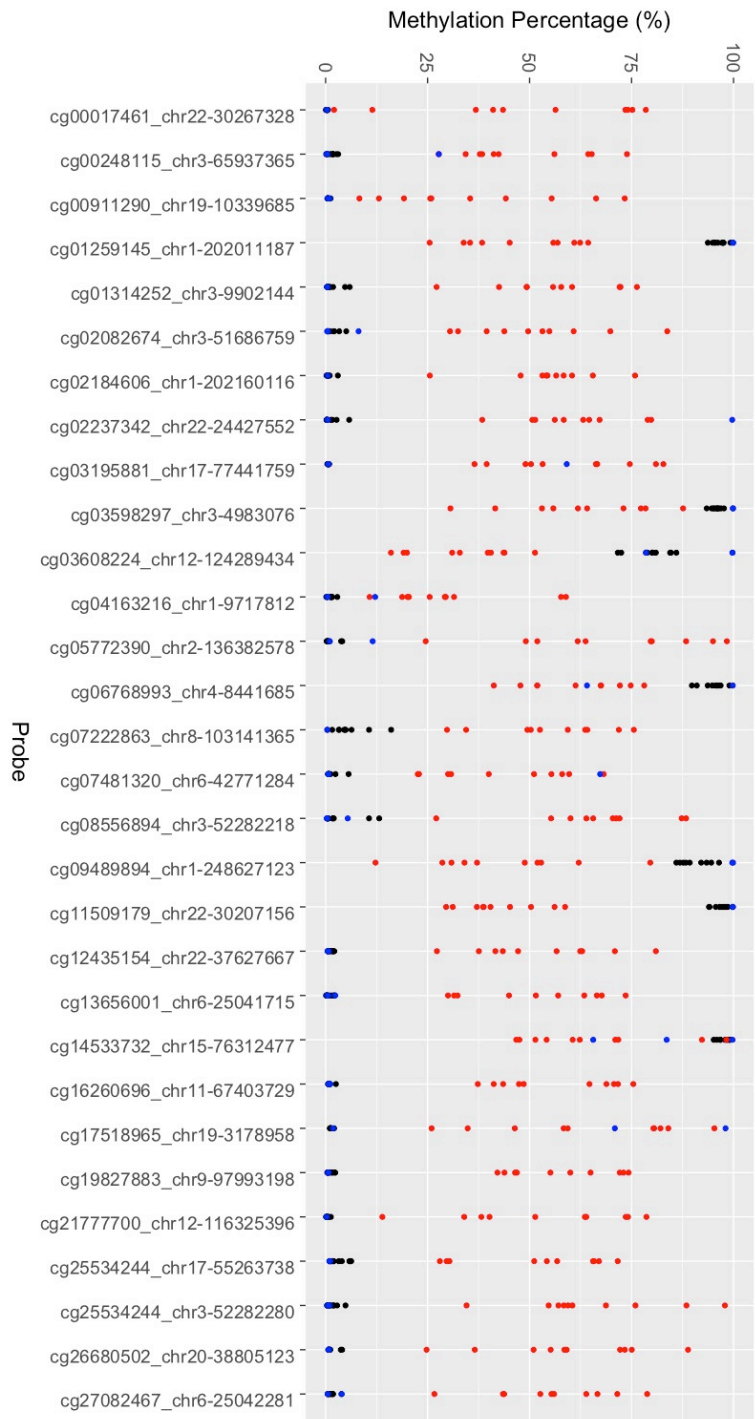


Figure 19: Differential Methylation Detection on Fluidigm Access Array. Leucocyte samples (black dots) showed methylation percentage of less than 5% and more than 95% as expected. cfDNA samples (blue dots) from healthy individuals showed similar methylation percentages compared to leucocytes. In contrast, the majority of tumour samples (red dots) showed clear differences in methylation compared to leucocytes. Probes and genomic positions for each targeted CpG on GRCh38/hg38 Human Genome are provided.

5.3. Validation of Fluidigm Methylation Results using other Methods

5.3.1. Bisulfite Sanger Sequencing

H. Primer Pair Validation

Primer pairs from the nine CpGs shown in Figure 10 were designed for Sanger sequencing (Appendix 6) and were tested in PCR amplification and optimized for six different annealing temperatures: 54°C, 56°C, 58°C, 60°C, 62°C and 64°C. PCR products were visualised under UV light after electrophoresis with 1,5% agarose gel TBE (0.5x) with 10µl SYBRSafe per 100ml of gel. Five out of nine pairs of primers successfully amplified bisulfite converted human genomic DNA and produced amplicons of around 250 bp (Figure 20). PCR amplification was repeated for each primer set to ensure reproducible results.

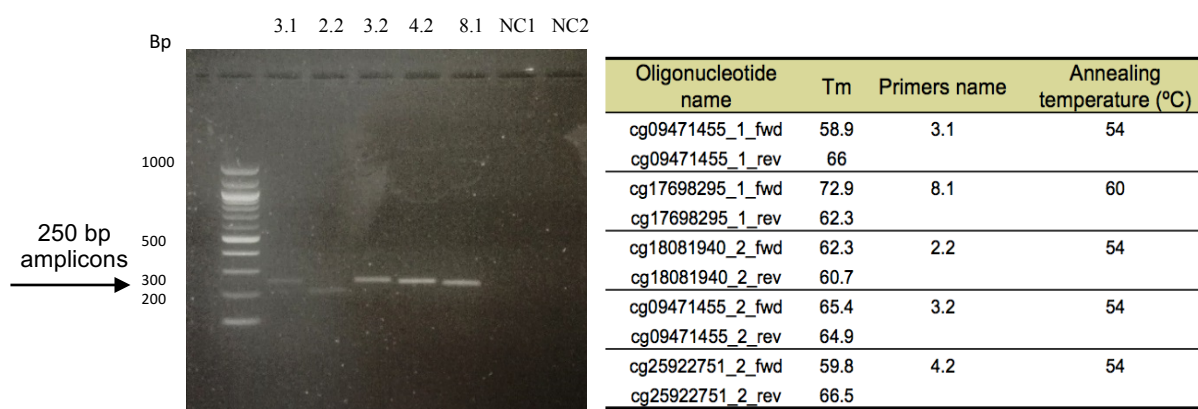
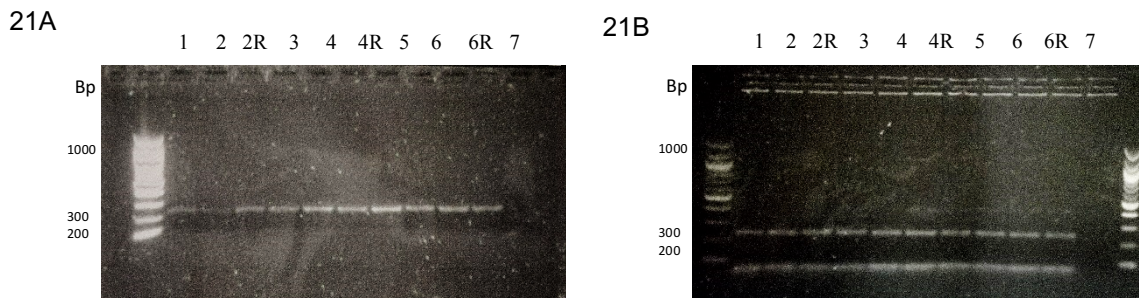


Figure 20: On the left, agarose gel showing 250 bp amplicons corresponding to the 5 validated pairs of primers. Lanes 1-5: primer pairs 3.1, 2.2, 3.2, 4.2 and 8.1. Line 6: negative control without DNA (NC1). Line 7: negative control without primers (NC2). On the right, table summarizing the melting temperature (T_m) of the primers and the best annealing temperature for each pairs of primers obtained in PCR reactions.

I. Validation of primer pairs in methylated and unmethylated DNA mixtures

Amplification of 10 samples, which were mixtures of fully methylated and unmethylated DNA, with two primer pairs (3.1 and 8.1) was tested in PCR reactions. Clear bands of 250bp were observed for all the samples (figures 21A and 21B). Amplicons were sent for Sanger sequencing at Edinburgh Genomics.

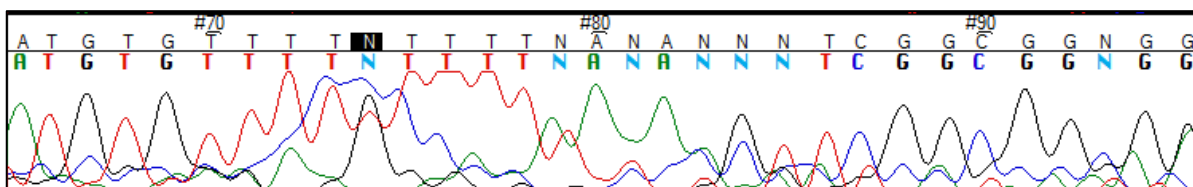


Figures 21A-21B: Agarose gels (1,5%) showing amplicons of 250 bp after PCR amplification of the 10 samples with the pairs of primers 3.1 (3A) and 8.1 (3B). R: replicate; lanes correspond to samples with increasing percentages of fully methylated and fully unmethylated DNA from 0-100%.

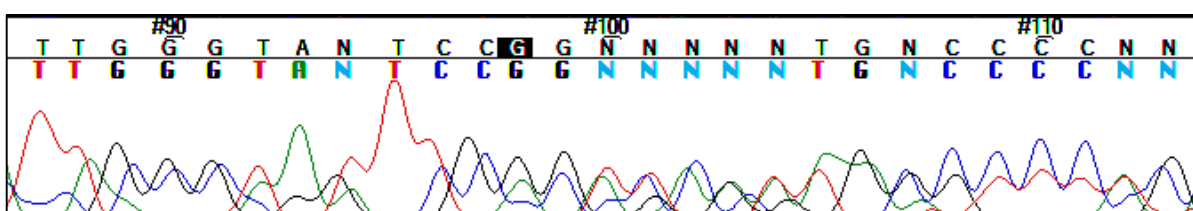
J. Sanger Sequencing Results

The initial Sanger sequences showed suboptimal results and did not distinguish 50% fully methylated from 50% unmethylated DNA (Figures 20A and 20B, show typical electropherograms). The low quality of the sequences might be caused by DNA contamination with mixed templates from the prepared fully unmethylated DNA, as some sequences could have not been amplified, or due to technical problems during the sequencing.

20A



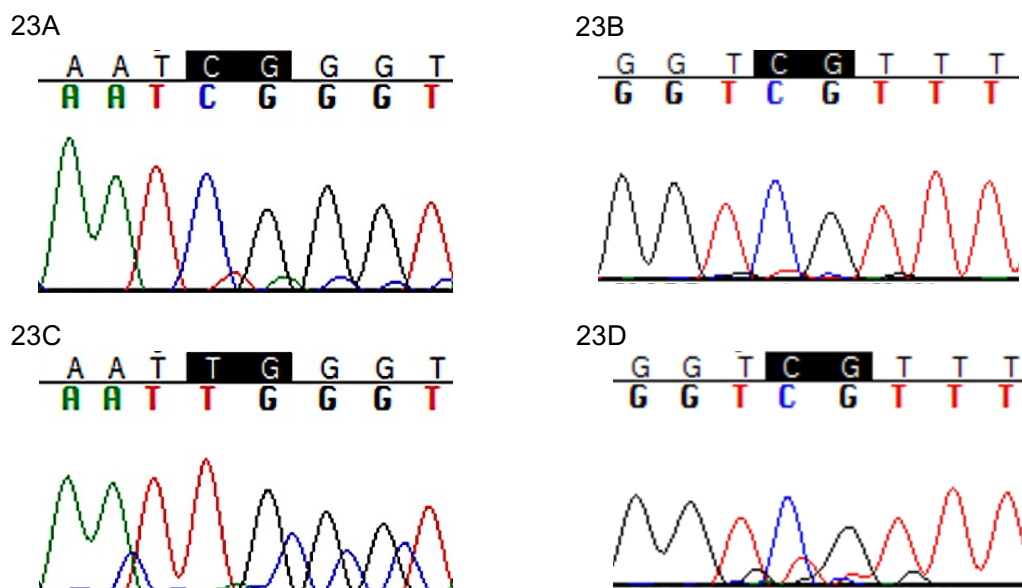
20B



Figures 22A-22B: Sanger sequencing electropherograms of sample 4 containing 50% of fully methylated and unmethylated DNA tested with the primer pairs 8.1 (4A) and 3.1 (4B) for the probes cg17698295 and cg09471455 which contain a CpG in the position 74 and 98, respectively, on Human genome from December 2013 assembly (GRCh38/hg38). "Noisy" data can be identified by the presence of multiple peaks and numerous "N"s within the sequence.

K. Quantification of Methylation Level in Sanger Sequencing Traces Using the ab1 Peak Reporter Tool

The samples with the best Sanger sequencing traces (Figures 23A and 23B) were used to quantify methylation levels. These samples contained 100%, 90%, 75% and 50% of fully methylated DNA and were sequenced with primer pairs 3 and 8.



Figures 23A, 23B, 23C, 23D : Representative Sanger sequencing electropherograms of samples containing 100% (A) 90% (B), 75%(C), 50%(D) of fully methylated DNA.

Table 3 shows peak height ratios for all four bases obtained using ab1 Peak Reporter tool. Sample S1 being 100% methylated showed a peak height ratio of 1 for cytosine and the smallest ratio for thymine, 0.019. Samples with decreasing levels in CpG dinucleotide methylation ranging from 100% methylated until 50% methylated (samples S1-S7), showed a higher peak ratio for cytosine at the same time that the ratio for thymine increased (figure 24).

Results showed that Sanger sequencing can detect fully methylated and fully unmethylated DNA, however when samples contained small methylated DNA amounts, Sanger sequencing was not sensitive enough. In addition, technical replicates did not indicate reproducible methylation measurements by bisulfite Sanger Sequencing. In conclusion, methylation analysis using Sanger sequencing may not be sensitive enough to detect small differences in methylation levels. Analysis of cfDNA challenging and requires highly sensitive

techniques because of the small fraction of tumour specific DNA present within background levels of normal cfDNA. Therefore, Sanger sequencing was demonstrated to not be an appropriate validation method for Fluidigm amplification, and neither appropriate for cfDNA-based tests.

Table 3: Quantification of methylation of samples S1, S2, S3 and S4 obtained with ab1 Peak Reporter tool. Signal strength, signal strength ratio and peak height ratios of CpG sites. Ratio for thymine gradually increased when samples had lower amounts of methylated DNA.

Sample	Base Call	Signal Strength				Ratio				Quality Value
		G	A	T	C	G	A	T	C	
S1	C	0	1	5	256	0	0.004	0.019	1	59
S2	C	25	3	14	497	0.050	0.006	0.028	1	62
S3	C	0	13	62	647	0	0.020	0.0958	1	43
S4	T	0	2	133	3	0	0.015	1	0.023	27

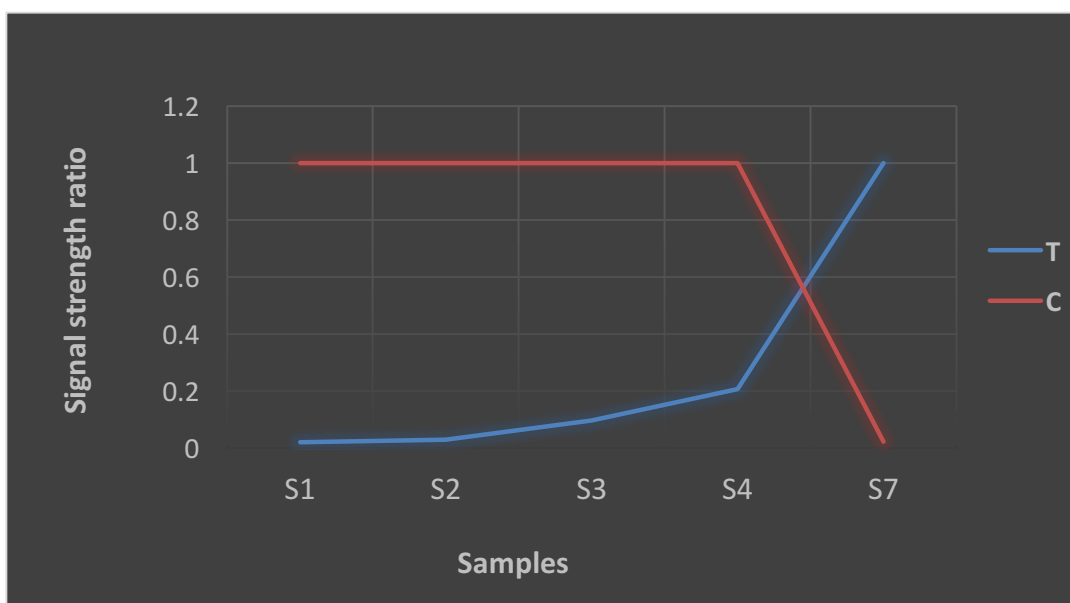


Figure 24: Signal strength ratio of Thymine and Cytosine of the samples S1, S2, S3, S4 and S7 that contained decreasing percentages of methylation from 100-0%.

5.3.2. Bisulfite Pyrosequencing

L. Primer Pair Validation

Primer pairs from 11 CpGs (Table 4) were designed and validated by PCR amplification with bisulfite-converted human genomic DNA (hgDNA) and visualized in agarose gels as previously described for Sanger sequencing.

Table 4: Forward and reverse PCR primers and sequencing primers designed for Pyrosequencing. Biotinylated PCR primers are marked with “bio”. The sequencing primer was designed to be opposite to the biotinylated strand, which is isolated and used in sequencing (the non-biotinylated strand is lost at the denaturation step).

Name	Probe	Genomic location on GRhg38	Primer fwd	Primer rev
18	cg03608224	chr12:124289407-124289512	AGTAGTTTTATTGTGGAGTGG	CCAAATTCCTCAACTTTCAATAAATAATA-bio
27	cg06768993	chr4:8441613-8441737	GGTGTTTGAAGGTTTTAAAGAG	ATCCCCAACACTTTTACACCTA-bio
40	cg11509179	chr22:30603070-30603219	GGTTGTGGTATTAGGAGTTGTTAG	TATCATTACCTTCCAACCTCTCT-bio
ZIC1a	ZIC1a	chr3: 147420614-147420750	AGGTTTTTGTGGGTTTAGTA	AACTAAAAAACCTCTACTCCATATCTCT-bio
ZIC1b	ZIC1b	chr3: 147420614-147420750	GTTGAGTTAGGTAAGAGATATGGAGTA	CCTTTTTTCTACCCAAAC-bio
*1v2	cg00248115	ch3:65937286-65937416	TTGGTGAGTTAGTTGGGGAAGGA	CCTAACACTTCTCCCTTACCCTTTTAC-bio
*2v2	cg07222863	chr8:103141270-103141376	GAGGAGGTGGGTTGTTTTTATT-bio	AACTCCAACAACCCAATACT
3v2	cg04163216	chr1:9717774-9717904	GAAAATAGGAAGTGGGGAGGG	ACCCATAACCTCCACCAAA-bio
1v3	cg19827883	chr9:97993105-97993238	GTTTGGTTTTGAAGAGGAAGTAGATA	CAATCAACAATACCCACAACAT-bio
2v3	cg27082467	chr6:25042184-25042317	GAAGATGATGGGGAGGTAATTTATTTAAGT-bio	TACCCTCCCCTACCATTACA
*3v3	cg16260696	chr11:67403677-67403805	AGAAATAAATAAAGAATTGGAGGTGG-bio	CCAATAACTACAAACTTAAATTCCTATACTAATA

All validation assays included four negative controls:

- PCR without template DNA, to test for non-specific signal from primer in the pyrosequencing reactions.
- Sequencing primer without PCR product (no template control, NTC).
- Biotinylated primer without PCR product (NTC).
- Sequencing primer and biotinylated primer together without PCR product (NTC).

The three latter NTC were included to ensure no secondary structures, such as hairpins or duplexes, were contributing to background signal in pyrosequencing reactions. All primer pairs successfully produced amplicons of less than 130bp, as shown in figure 25. Non-specific background bands were not a problem as

only one primer pair (18) showed a band in the negative control with the sequencing and biotinylated primer together.

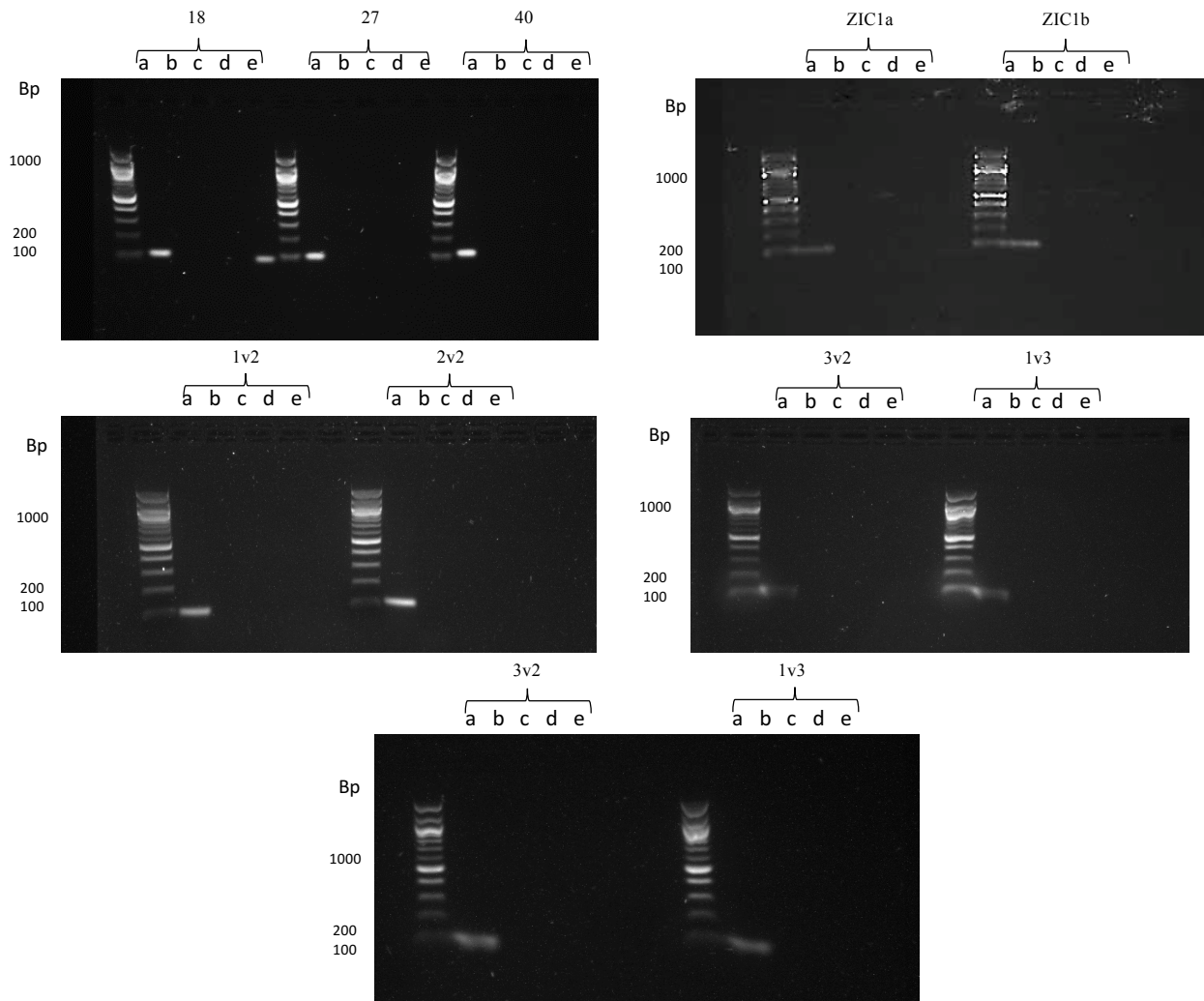


Figure 25: Agarose gels (1,5%) showing 100 bp amplicons corresponding to the pyrosequencing validated pairs of primers: Primers 18, 27, 40 ZIC1a and ZIC1b. “a” lanes are PCR reactions with human genomic DNA; “b” lanes are negative controls without hgDNA; “c” lanes correspond to biotinylated primer without hgDNA. “d” lanes correspond to sequencing primers without hgDNA. “e” biotinylated and sequencing primers without hgDNA.

M. Bisulfite Pyrosequencing Results

Using pyrosequencing, the percentage of methylation was individually determined at multiple CpG sites across each locus. Technical replicates of three of the mixtures showed very low variability (tables 5A, 5B and 5C), indicating reliable reproducible methylation measurements by bisulfite pyrosequencing.

Tables 5A, 5B and 5C: Methylation percentages for two technical replicates for samples 2, 4 and 6, containing 90%, 50% and 10% of fully methylated DNA obtained by separate pyrosequencing assays.

		Sample replicates					
		90%		50%		10%	
Primer set	CpG position	2	2R	4	4R	6	6R
ZIC1a	Pos1	84	80	41	36	7	8
	Pos2	84	79	40	36	7	8
	Pos3	86	77	40	37	7	9
ZIC1b	Pos1	88	84	50	49	11	11
	Pos2	76	74	45	41	11	11
	Pos3	90	86	51	47	12	12
	Pos4	87	85	50	47	11	12

		Sample replicates					
		90%		50%		10%	
Primer set	CpG position	2	2R	4	4R	6	6R
18	Pos1	83	82	39	35	8	8
	Pos2	84	85	40	36	8	9
27	Pos1	76	76	31	30	7	7
40	Pos1	87	82	44	43	10	10
	Pos2	82	79	42	40	10	12
	Pos3	83	78	42	40	9	11
	Pos4	86	82	44	43	11	12

		Sample replicates					
		90%		50%		10%	
Primer set	CpG position	2	2R	4	4R	6	6R
1V2	Pos1	78	79	24	24	4	4
	Pos2	70	72	23	23	3	4
	Pos3	68	69	23	23	6	7
2V2	Pos1	77	79	37	38	7	10
	Pos2	79	79	37	38	9	10
	Pos3	74	74	32	33	7	8
3V3	Pos1	62	63	19	20	6	6
	Pos2	64	65	22	23	6	7

Internal controls ensured complete bisulfite conversion. Cytosines that are not followed by guanines in template sequences are not methylated, and should therefore have been converted to thymine by bisulfite treatment and PCR. Full bisulfite conversion of the samples was confirmed in all templates, as built-in quality control sites showed thymine and no cytosine in these positions (Figure 26).

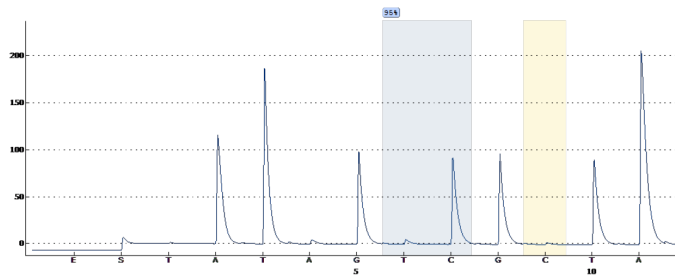


Figure 26: Representative pyrogram showing complete bisulfite conversion of the internal control in a fully methylated sample (S1). While the methylated cytosine in the CpG (highlighted in blue) showed a peak for cytosine, no peak is shown for the built-in quality control (highlighted in yellow) ensuring successful bisulfite conversion.

Methylation percentages for the mixtures provided series of standards (figures 27A-27D) for validation of assays in tumour and leucocyte from breast cancer patients and cfDNA samples from healthy individuals.

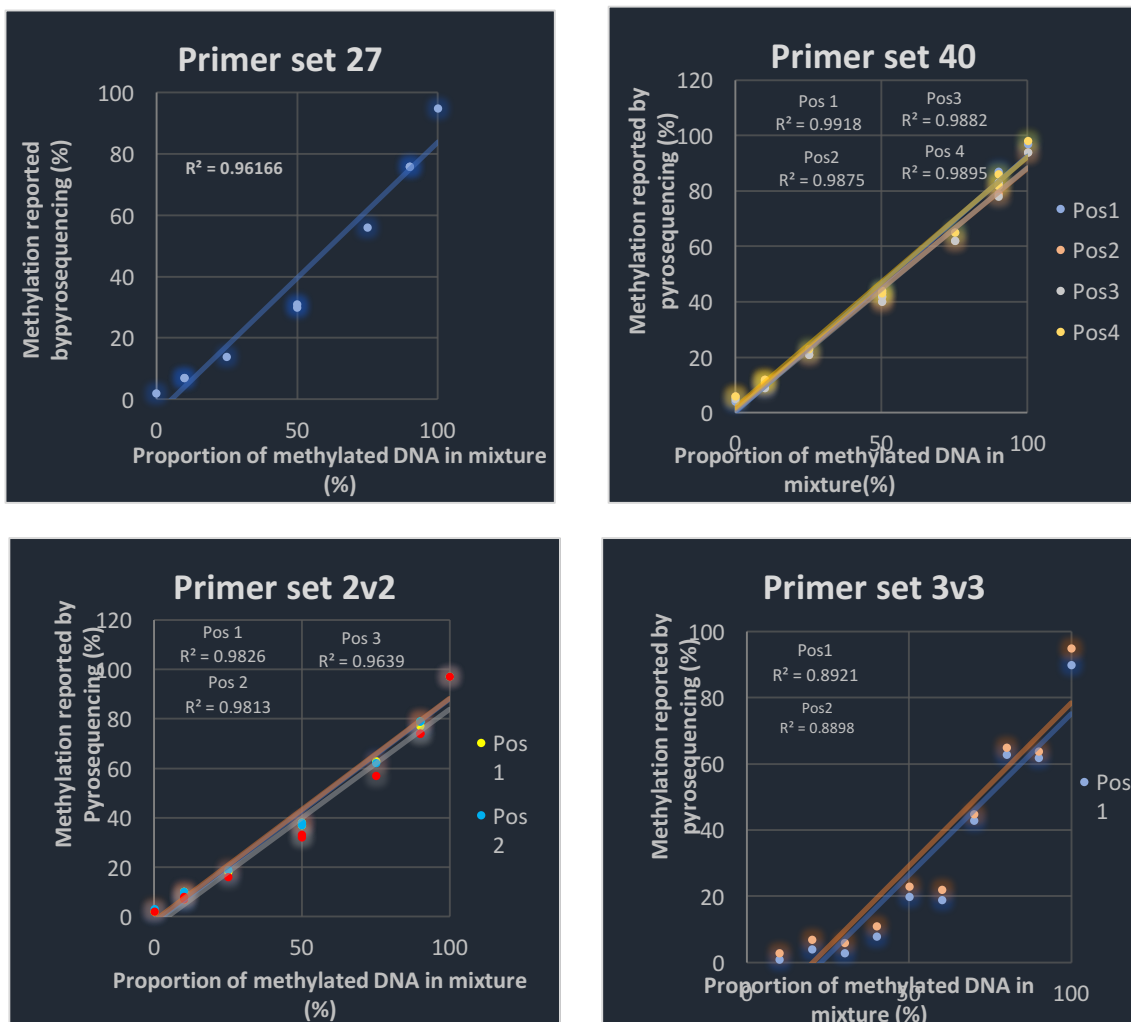


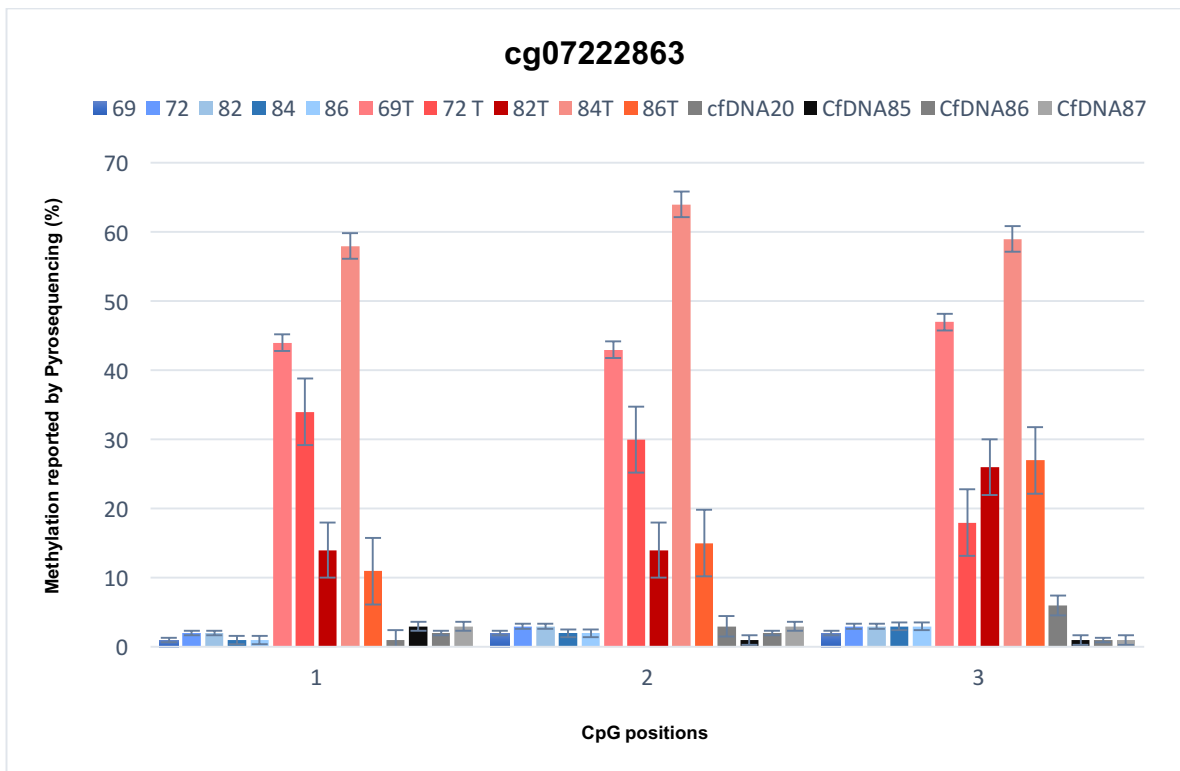
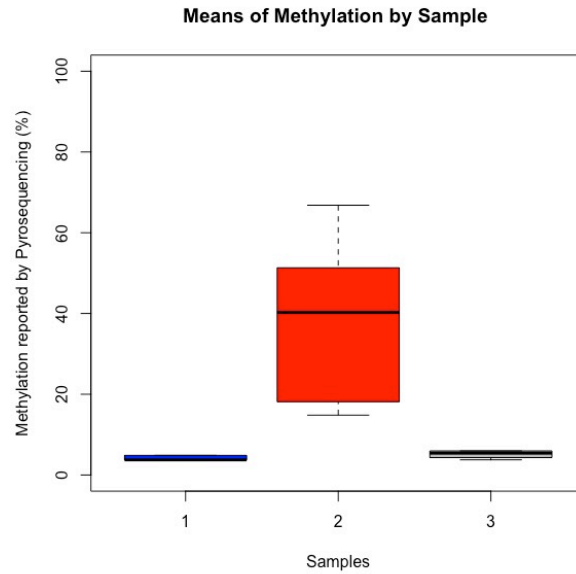
Figure 27A-27D: Linearity of methylation quantification by pyrosequencing. Methylation scales for primer pairs 27 and 40 for different CpG sites obtained from mixtures of whole genome amplified (WGA) DNA (expected 0%) and fully methylated DNA commercially available (expected 100%). The line is the average linear regression with its coefficient of determination (R²). Both primer sets showed a strong positive correlation.

Standard curves allowed to adjust methylation levels in real samples for each assay. Methylation levels were successfully assessed in five paired tumour and leucocyte samples, and in four cfDNA samples. Statistical analysis was performed to assess differences in methylation. Bartlett test was significant for homogeneity of variances therefore individual t-tests with an alpha of 0.05 were performed. Tumour samples were significantly different from leucocytes (table 6A) and cfDNAs samples (table 6C), whereas leucocytes and cfDNAs were not significantly different with an alpha level of 0.05 (table 6B).

Tables 6A, 6B and 6C: Statistical analysis for pyrosequencing results. Significant p-values were obtained for tumour-leucocyte comparisons and for tumour-cfDNA comparisons for 5 CpG sites assayed with two different primer sets with pyrosequencing.

<i>Tumour-Leucocytes comparisons</i>		<i>Leucocytes-cfDNAs comparisons</i>		<i>Tumour-cfDNAs comparisons</i>	
<i>CpG</i>	<i>P-value</i>	<i>CpGs</i>	<i>P-value</i>	<i>CpGs</i>	<i>P-value</i>
<i>2v2_pos1</i>	0.026798	<i>2v2_pos1</i>	0.180597	<i>2v2_pos1</i>	0.028176105
<i>2v2_pos2</i>	0.031498	<i>2v2_pos2</i>	0.792534	<i>2v2_pos2</i>	0.029706425
<i>2v2_pos3</i>	0.013166	<i>2v2_pos3</i>	0.691964	<i>2v2_pos3</i>	0.011206465
<i>3v3_pos1</i>	0.000253	<i>3v3_pos1</i>	0.110162	<i>3v3_pos1</i>	8.25E-05
<i>3v3_pos2</i>	0.000182	<i>3v3_pos2</i>	0.094493	<i>3v3_pos2</i>	3.62E-05

Means of methylation percentages were calculated. Overall, methylation levels in tumour samples ranged from 18% to 80%, from 1% to 5% in leucocyte samples, and from 1% to 10% in cfDNA samples from healthy individuals. Representative pyrosequencing results are shown in figures 28A and 28B.



Figures 28A and 28B. Representative pyrosequencing results for probe cg07222863. **28A**, Box plot of methylation means in paired leucocyte (1) and tumour (2) samples, and in cfDNA samples (3) reported for probe cg07222863 by pyrosequencing. Tumour samples obtained a mean of methylation percentage of 38.27168, leucocytes of 4.200885 and cfDNAs of 5.14115. **28B**, Comparisons of dinucleotide methylation percentages for three CpG sites contained in probe cg07222863 between paired tumour (blue hue) and leucocyte samples (red hue), and cfDNAs (grey hue) from healthy people reported by pyrosequencing. Numbers 69, 72, 82, 84 and 86 represent different leucocyte samples; same numbers ended with a “T” represent their paired tumour samples; and cfDNA20, cfDNA85, cfDNA86 and cfDNA87 represent cfDNA samples from healthy individuals.

5.4. Correlation of CpG Dinucleotide Methylation

Correlations between assays were assessed by linear regression and by Pearson's correlation. Correlation plots were prepared to investigate the agreement between the 2 assays. Statistical analysis was performed using Tukey(HSD) and $P < 0.05$ was considered statistically significant.

This variability in tumour samples is shown in the correlation plot. Tumour samples show a weak correlation of methylation measurements between the EPIC array and Fluidigm-MiSeq ($R^2 = 0.649948$) whereas leucocyte samples correlated significantly better ($R^2 = 0.993984$) between the two methods. Overall, significant linear associations were observed for methylation measurements between both methods (Table 7) (Appendix 7).

Sample	Correlation	Adjusted p-value
B41	0.9971536	2.2e-16
B55	0.9943155	2.2e-16
B66	0.9979749	2.2e-16
B69	0.9815363	2.2e-16
B72	0.9959881	2.2e-16
B77	0.9935976	2.2e-16
B82	0.9966643	2.2e-16
B83	0.9952834	2.2e-16
B84	0.9978038	2.2e-16
T41	0.8497082	2.86 e-09
T55	0.7671869	7.586e-07
T66	0.1698525	0.3695
T69	0.5950708	0.0005233
T72	0.3833564	0.03651
T77	0.1982793	2.2e-16
T82	0.4351883	0.01624
T83	0.6837339	3.111e-05
T84	0.8092882	6.158e-08

Table 7: Pearson's correlation coefficients of the methylation levels reported by EPIC and Fluidigm-MiSeq. Leucocyte samples showed a very positive strong correlation between both methods whereas tumour samples obtained a poorer positive correlation.

An analysis of Variance (ANOVA) was used to compare methylation means reported by Fluidigm-MiSeq, EPIC array and pyrosequencing. Homoscedasticity was verified performing Bartlett test (P value = 0.9593). Then, a p -value >0.05 was obtained for analysis of variance. This indicated that the three different methods were statistically equal in detecting methylation levels (figures 31A and 31B). A TukeyHSD *posthoc* test was performed to obtain detailed information about mean differences in all possible simple contrasts. Any of the three different methods appeared to be better or worst in detecting methylation levels.

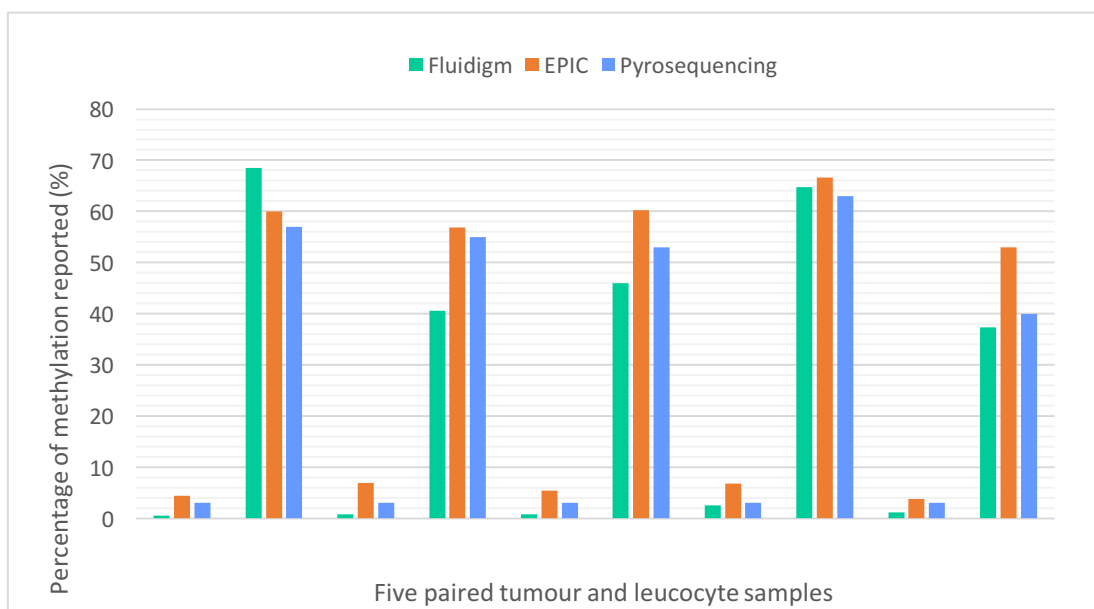
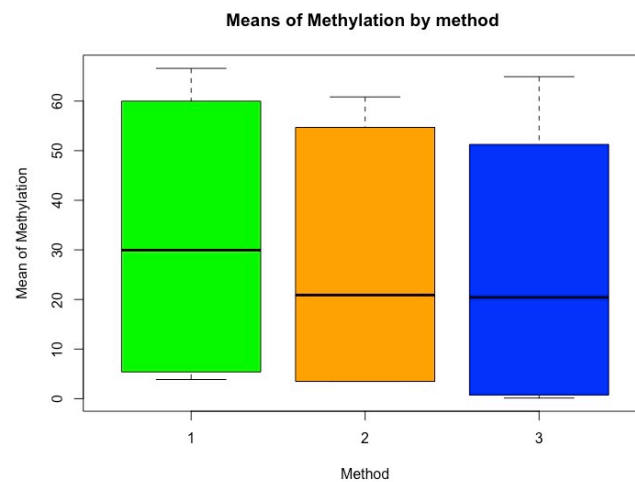


Figure 31A and 31B. **31A**, Box plot showing means of methylation levels measured by Fluidigm (green), EPIC array (orange) and pyrosequencing (blue). **31B**, Comparison of methylation percentage reported by the three methods for probe cg16260696 which contains a differentially methylated CpG in chr11:67403729 position on GRCh38/hg38 Human Genome between tumour and buffy coat of breast cancer patients.

6. CONCLUSIONS

Tumour, leucocyte and cfDNA samples had been successfully analyzed using the Fluidigm Access Array Amplification and Illumina MiSeq Next Generation Sequencing. Twenty-three different primer pairs were designed and validated to specifically amplify CpG sites that showed extreme methylation distribution for leucocytes with a methylation difference of $> 20\%$ in the tumour samples in the EPIC array. Sixteen of the twenty-three CpGs passed all the filtering criteria after the MiSeq Sequencing analysis, and showed significant differences between tumour and buffy coat. Methylation levels between cfDNA from healthy individuals and tumour samples were statistically different.

EPIC and Fluidigm-MiSeq showed moderate positive correlation overall, but showed strong positive correlation for leucocyte samples, the latter being selected to have extreme methylation levels for marker design.

Regarding the validation methods, pyrosequencing showed to be more accurate than Sanger sequencing, which may not be sensitive enough to detect small differences in methylation levels. Pyrosequencing showed high sensitivity reliably detecting 10%, 25%, 50%, 75% and 90% of methylation. Pyrosequencing also showed reproducible correlation across technical replicates with different ratios of methylated DNA. In contrast, Sanger Sequencing while successfully detecting 100% methylated and 100% unmethylated DNA, could not detect small levels in methylation, in addition to showing detection bias towards methylated DNA.

In conclusion, target-specific marker design optimization allowed detection of methylation differences between tumour and leucocyte samples, and in addition showed potential for detection of methylation in cfDNA with the 48.48 Access Array. Pyrosequencing supports these results, and is therefore a useful method for validation of Fluidigm-MiSeq methylation data. These techniques used together may have the potential to ultimately translate methylation patterns from tissue tumours to cfDNA, which could avoid tissue biopsies in the future.

7. DISCUSSION

Liquid biopsies are non-invasive methods for detection and monitoring of cancer-associated alleles and likely to substitute traditional tissue biopsies in the future. Most studies focus on the detection of somatic mutations in cfDNA and have already obtained high levels of sensitivity for cfDNA biomarkers in different metastatic cancer types¹¹¹. In addition, the feasibility of using tumour-specific mutations in cfDNA to monitor the response to therapy has been demonstrated in colorectal¹¹², breast⁷³, ovarian¹¹³ and lung⁷³ cancers. While effective, this approach was found to be time consuming⁷⁸, and difficult to apply to an extended patient population given the highly individual profile of cancer mutations⁶⁷.

Methylated cfDNA has also been reported as a marker of surgery, for instance Liggett and colleagues found methylated sequences in plasma of breast cancer patients that decreased following surgery and tamoxifen treatment¹¹⁴. In addition, gene methylation patterns of both individual genes and gene panels have been correlated with patient survival¹¹⁵. The present study aimed to find a set of CpG sites from a preliminar methylation study using the Illumina EPIC Methylation Array, which interrogated CpG dinucleotide methylation at over 850,000 DNA sites. A total of 3172 CpGs showed clear methylation differences between tumour and leucocytes of patients with breast cancer. 48 CpG sites showing marked methylation differences were used in this study to establish accurate and reliable methods for CpG dinucleotide methylation detection, to ultimately be translated into an assay applicable to circulating tumour DNA. All tumour, leucocyte and cfDNA samples used for this purpose were successfully amplified with the Fluidigm Access Array followed by Illumina MiSeq sequencing. This approach has provided accurate methylation percentages, which demonstrate high correlation with the EPIC array methylation values.

Several other approaches have recently been proposed for non-invasive cancer detection, and potential epigenetic aberrations have been shown to differentiate healthy plasma from tumour in breast cancer patients^{116,117,118,119}. However, these techniques mostly used the candidate gene approach and did not rely on

detecting specific biomarkers to certain tumour types. Importantly, Kang and colleagues proposed the CancerLocator method to infer the proportion and tissue of origin of cfDNA in a blood sample using genome-wide DNA methylation data. Whole-genome bisulfite sequencing data is an increasingly used technique for methylation studies that is allowing the construction of genomic-maps at single-base resolution of nearly every CpG site, including low CpG-density regions. Other experimental approaches for DNA methylation include enzyme digestion, affinity enrichment-based methods, methylated DNA immunoprecipitation (MeDIP)¹²⁰, methylation arrays and commercial DNA methylation kits. In addition, emerging third-generation sequencing technologies, including single-molecule real-time sequencing (SMRT) and Oxford Nanopore technology, are being adopted in epigenetics research. Furthermore, bisulfite conversion sequencing can be done with targeted methods such as amplicon methyl-seq, target enrichment, or with whole-genome bisulfite sequencing (WGBS). Additionally, OxBS and TAB-Seq can be used with NGS for identification of hydroxymethylation (5-hMc) in conjunction with methylation (5-mc) analysis. Research questions, cost, amount of input DNA and the expected degree of methylation changes are the main factors when selecting a particular technique. Validation methods used in this study were the most cost-effective and practical for medium-throughput analysis. Methylation analysis using Sanger sequencing has not been sufficiently sensitive to accurately detect small differences in methylation levels. While Sanger sequencing did reliably detect 100% fully methylated and 100% fully unmethylated DNA, it was not possible to distinguish 50% fully methylated DNA. Low quality of the sequences may be improved by additional cleaning steps in the FUD preparation. However, Sanger sequencing may not be sensitive enough when aiming to establish a reliable method for detection of small changes in methylation.

In contrast, pyrosequencing has provided both quantitative and qualitative methylation data and the results have demonstrated that small changes are detectable in both the mixtures and breast tumour, leucocyte and cfDNA samples. In addition, methylation levels in paired tumour and leucocyte samples reported by this technology correlate well with both the Fluidigm and EPIC array

data. Therefore, pyrosequencing is reproducible, and also works with very low concentrations of DNA. In conclusion, Fluidigm amplification and pyrosequencing together can be used to accurately detect methylation in an assay of cfDNA obtained from blood.

The results described above were obtained by first selecting a methylation pattern in breast tumour samples which was distinct from the patterns shown in leucocytes in the EPIC array, then determining whether that same methylation level was obtained using Fluidigm amplification. In agreement with previous research, methylation levels between cancer samples and paired normal tissue have shown different patterns of methylation with our methodology. Specifically, breast tumour samples demonstrated different methylation levels compared to those in their leucocyte paired samples and cfDNA samples from healthy individuals. On the other hand, tumour samples showed a lower positive correlation in methylation levels than leucocyte samples between Fluidigm-MiSeq and EPIC. The EPIC array bisulfite conversion was done using a different bisulfite conversion kit (EpiTech), which could have led to different fragmentation and different levels of incomplete bisulfite conversion.

Methylation of CpG sites at selected DNA sequences provides a level of regulation over gene expression. Genome methylation undergoes coordinated changes at defined stages of development and in response to environmental stimuli such as diet, chemical toxins and pollutants, and temperature stresses¹²¹. Compared with other cancer-specific biochemical modifications, methylation of DNA is a cancer-specific stable modification that can be obtained from diverse body fluids, including whole blood¹²², plasma¹²³, circulating tumour cells¹²⁴, serum¹²⁵, saliva¹²⁶, urine¹²⁷, bronchoalveolar lavage¹²⁸, sputum⁷, stool¹²⁹, and fine needle aspirate¹³⁰. These features make CpG dinucleotide methylation a suitable target for alternative non-invasive diagnostic strategies. Additionally, the number of aberrantly methylated CpG sites is significantly larger than the number of genetic mutations. This allows for a more flexible design with a larger number of diagnostic targets. Methylated cfDNA is a challenging substrate to work with, largely because cfDNA is highly diluted, with concentrations as low as < 10 ng per ml of plasma in healthy subjects¹³¹. What is more, the methylated component is an even smaller subfraction of this

amount. Thus, many of the technical issues for this study relate simply to limited quantities of starting material. Generation of PCR duplicates could have had an impact on the results. PCR errors in early stages of amplification, especially if occur at CpG of interest could have affected methylation results. In addition, PCR duplicates might have affected methylation estimation. In this case, it might be useful to estimate the number of starting material as input for the Fluidigm. This issue can also be addressed by the incorporation of Unique Molecular Identifiers (UMIs). Moreover, another of the major concerns is the lack of standardised techniques. Collection and processing of clinical samples needs to be even more carefully considered for discovery and validation of circulating biomarkers. Blood in EDTA tubes can only be stored for a limited amount of time before leucocytes start to lyse, and genomic leucocyte DNA will therefore be contributing to the plasma DNA fraction. In addition, bisulfite conversion, used as the gold standard method for assessing DNA methylation, fragments the DNA resulting in lower input for Fluidigm amplification. In addition, a high sequencing depth is required to differentiate real modifications from background sequencing errors.

Another challenge in cfDNA studies is that many aspects of the biological characteristics are still not clear. For instance, the size of cfDNA has been shown to be variable (70bp-200bp)⁷⁰, though Lo and colleagues previously showed a characteristic size pattern at 166bp¹³². In this study, an amplicon length between 125 and 135 bp was successfully amplified in cfDNA samples, obtaining more than 30,000 reads, and methylation levels have been detected. The size of cfDNA reflects the apoptotic origin of the DNA⁸³. A technical consequence of this is that DNA purification methods are not as efficient at extraction of cfDNA, leading to possible DNA losses. Furthermore, cfDNA is known to be found in a background of non-tumour cfDNA, derived primarily from apoptosis of normal cells of the hematopoietic lineage^{133,100,75,134}. Our working hypothesis was based on this assumption. Nevertheless, the characterization of aberrant DNA methylation patterns depends on a normal reference DNA methylome. Since each cell type presents a specific DNA methylation pattern, defining a reference DNA methylome is still a major challenge. However, this study anticipates that methylation levels in leucocytes from breast cancer

patients can be compared to those obtained in healthy cfDNAs. Hence, methylation levels in non-tumour circulating DNA assumed to come mainly from leucocytes could serve as a reference methylome to detect aberrant epigenetic events when comparing with cfDNA in breast cancer patients.

Regarding marker design for this study, probes having extreme beta-values in leucocyte samples in the EPIC that intersected best q-value, median differences and widest beta-value separation between tumour and leucocyte samples were chosen to be tested with the Fluidigm-MiSeq assay. Determining the specificity of a biomarker requires large numbers of healthy controls, ideally age-matched and collected within the same setting and in cancer patients as well as from healthy control population. As noted above, a normal reference DNA methylome is required, but healthy samples are also likely to show changes in methylation due simply to the presence of inflammation or after exercise¹³⁵. Consequently, the resulting biomarker may not be able to differentiate between cancer and other physiological conditions with an inflammation component. Presumably, the establishment of an extreme threshold in methylation levels (less than 5% or more than 95%) of leucocyte samples may have included this variability in methylation as well as modifications coming from other physiological circumstances.

A limitation of this study is the small number of paired samples studied (n=10) as well as cfDNA samples (n=4). In order to consistently translate methylation patterns from tumour to cfDNA, future research should include greater numbers of samples. This could enable full evaluation of the potential and limitations of this approach. Expanding the breadth and quality of the healthy reference datasets to explore the distribution of methylation patterns in leucocytes from healthy people is likely to robustly identify more specific markers. For instance, the identification of CpGs where methylation levels vary depending on other physiological conditions in healthy leucocytes will avoid incorrect comparisons that could lead to false-positives.

In summary, low concentrations, small size of cfDNA, and technical issues can be problematic in methylation studies and need to be overcome for methylation

of cfDNA to be used in the clinical setting as a cancer biomarker. Despite these difficulties, this study demonstrates that cfDNA from healthy samples can be amplified and investigated. These initial results are encouraging and further optimisation will allow liquid biopsies to be routinely applied into the clinic for monitoring, diagnostics, stratification and prognosis of different cancer types.

8. FUTURE DIRECTIONS

Even though sequence events play a critical role in aberrant gene expression in cancer, studying them alone remains insufficient to understand how genomes are translated into transcriptional patterns in normal and cancer cells. Therefore, epigenetics has emerged as a critical process in cancer progression, together with genetic events. The increasing interest in this field has allowed the development of sensitive platforms to detect events that are now known to be involved in many diseases. Aberrant epigenetic events are frequent in early stages of cancer⁴¹. Thus, epigenetic alterations may have the potential to become biomarkers for diagnostics, stratification and prognosis.

Tissue-tumour profiles are subject to sampling bias and provide only a snapshot of tumour heterogeneity, hence missing the clonal composition of each tumour as a whole, and cannot be obtained repeatedly. In contrast, genomic profiles of circulating cell-free tumour DNA overcome tumour heterogeneity and can detect changes that come from both primary and metastatic tumours. CfDNA offers the possibility to detect specifically what kind of molecular changes are happening in the tumour in real time.

CfDNA has an enormous potential as a “liquid biopsy” and the results obtained in this project suggest several translational implications and important avenues for future research when bigger datasets are explored. A limiting step in the analysis of cfDNA is the minute amounts of cfDNA available from the plasma of cancer patients and/or healthy volunteers. However, Fluidigm MiSeq-sequencing and pyrosequencing have both detected methylation in cfDNA from healthy plasma and have shown a very positive strong correlation when using paired tumour and leucocyte samples.

In conclusion, these 2 methodologies used and developed in this Masters project show evidence of their potential utility and that of epigenetic analysis of cfDNA for detection of tumour-derived cfDNA in patients that can, with further work, be useful in clinical practice.

REFERENCES

1. He, M., Rosen, J., Mangiameli, D. & Libutti, S. K. Cancer Development and Progression. in *Microarray Technology and Cancer Gene Profiling* (ed. Mocellin, S.) 117–133 (Springer New York, 2007). doi:10.1007/978-0-387-39978-2_12
2. OSBREAC, T. O. B. C. C. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 1–7 (2012).
3. Arnedos, M. *et al.* Precision medicine for metastatic breast cancer—limitations and solutions. *Nat. Rev. Clin. Oncol.* **12**, 693–704 (2015).
4. Beckmann, M. W., Niederacher, D., Schnüren, H. G., Gusterson, B. A. & Bender, H. G. Multistep carcinogenesis of breast cancer and tumour heterogeneity. *J. Mol. Med.* **75**, 429–439 (1997).
5. Yang, X., Yan, L. & Davidson, N. E. DNA methylation in breast cancer. *Endocr. Relat. Cancer* **8**, 115–127 (2001).
6. Tang, X. *et al.* Hypermethylation of the Death-Associated Protein (DAP) Kinase Promoter and Aggressiveness in Stage I Non-Small-Cell Lung Cancer. **92**, (2017).
7. Sciences, M. Aberrant methylation of p16 INK4a is an early event in lung cancer and a potential biomarker for early diagnosis. **95**, 11891–11896 (1998).
8. Esteller, M., Hamilton, S. R., Burger, P. C., Baylin, S. B. & Herman, J. G. Inactivation of the DNA Repair Gene O₆-Methylguanine-DNA Methyltransferase by Promoter Hypermethylation is a Common Event in Primary Human Neoplasia Advances in Brief Inactivation of the DNA Repair Gene O₆-Methylguanine-DNA Methyltransferase by Prom. *Cancer Res.* 793–797 (1999).
9. Esteller, M. *et al.* Inactivation of Glutathione S-Transferase in Human Neoplasia1 PI Gene by Promoter Hypermethylation. 4515–4518 (1998).
10. Fraga, M. F. A mouse skin multistage carcinogenesis model reflects the aberrant DNA methylation patterns of human tumors. *Cancer Res.* **64**, 5527–5534 (2004).

11. Weedon-Fekjær, H., Romundstad, P. R. & Vatten, L. J. Modern mammography screening and breast cancer mortality: population study. *BMJ* **348**, g3701 (2014).
12. Jones, P. A. & Takai, D. The Role of DNA Methylation in Mammalian Epigenetics. **293**, 1068–1071 (2001).
13. Widschwendter, M. & Jones, P. A. DNA methylation and breast carcinogenesis. *Oncogene* **21**, 5462 – 5482 (2002).
14. Res, C. C., Widschwendter, M. & Jones, P. A. The Potential Prognostic , Predictive , and Therapeutic Values of DNA Methylation in Cancer 1. **8**, 17–21 (2002).
15. Eads, C. A. *et al.* MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Res.* **28**, e32-0 (2000).
16. Silva, J. M. *et al.* Aberrant DNA methylation of the p16 INK4a gene in plasma DNA of breast cancer patients. *Br. J. Cancer* **80**, 1262–1264 (1999).
17. Evron, E. *et al.* Detection of breast cancer cells in ductal lavage fluid by methylation-specific. *Lancet* **357**, 1335–1336 (2001).
18. Robertson, K. D. DNA methylation and human disease. *Nat.Rev.Genet.* **6**, 597–610 (2005).
19. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics. *CA Cancer J Clin* **66**, 7–30 (2016).
20. Bannister, N. & Broggio. Cancer survival by stage at diagnosis for England (experimental statistics): Adults diagnosed 2012 , 2013 and 2014 and followed up to 2015. 1–23 (2016).
21. Etzioni, R. *et al.* The case for early detection. *Nat. Rev. Cancer* **3**, 243–252 (2003).
22. National Cancer Institute. *September 12* (2016). Available at: <https://seer.cancer.gov/statfacts/html/breast.html>. (Accessed: 16th November 2016)
23. Heyn, H., Méndez-González, J. & Esteller, M. Epigenetic profiling joins personalized cancer medicine. *Expert Rev. Mol. Diagn.* **13**, 473–479

(2013).

24. Cancer Research UK. 2005
25. Tavassoli F.A., D. P. (Eds. . *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Breast and Female Genital Organs*. (2003).
26. Sinn, H. P. & Kreipe, H. A brief overview of the WHO classification of breast tumors, 4th edition, focusing on issues and updates from the 3rd edition. *Breast Care* **8**, 149–154 (2013).
27. Dieci, M. V., Orvieto, E., Dominici, M., Conte, P. & Guarneri, V. Rare Breast Cancer Subtypes: Histological, Molecular, and Clinical Peculiarities. *Oncologist* **19**, 805–813 (2014).
28. NHS. 26/09/2016 Available at: <http://www.nhs.uk/Conditions/Cancer-of-the-breast-female/Pages/Introduction.aspx>. (Accessed: 5th July 2017)
29. Team, T. A. C. S. medical and editorial content. American Cancer Society. August 18, 2016 Available at: <https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/types-of-breast-cancer.html>. (Accessed: 1st August 2017)
30. Perou, C. M. *+ et al. Molecular portraits of human breast tumours. [Letter]. **533**, 747–752 (2000).
31. Sorlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* **98**, 10869–10874 (2001).
32. Sorlie, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci.* **100**, 8418–8423 (2003).
33. Sotiriou, C. et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci.* **100**, 10393–10398 (2003).
34. Prat, A. et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **12**, R68

(2010).

35. Hansen, N. M. Management of the Patient at High Risk for Breast Cancer. (2013). doi:10.1007/978-1-4614-5891-3
36. Knudson, a G. Two genetic hits (more or less) to cancer. *Nat. Rev. Cancer* **1**, 157–162 (2001).
37. Cipollini, G. *et al.* Genetic alterations in hereditary breast cancer. *Ann. Oncol.* **15**, 7–13 (2004).
38. Apostolou, P. & Fostira, F. Hereditary breast cancer: the era of new susceptibility genes. *Biomed Res Int* **2013**, 747318 (2013).
39. Yang, X. R. *et al.* Associations of breast cancer risk factors with tumor subtypes: A pooled analysis from the breast cancer association consortium studies. *J. Natl. Cancer Inst.* **103**, 250–263 (2011).
40. Herman, J. G. DNA hypermethylation in tumorigenesis epigenetics joins genetics expression in cancer. **9525**, (2000).
41. Dong, Y., Zhao, H., Li, H., Li, X. & Yang, S. DNA methylation as an early diagnostic marker of cancer (Review). *Biomed. Reports* 326–330 (2014). doi:10.3892/br.2014.237
42. Feinberg, A. P. & Vogelstein, B. Hypomethylation of ras oncogenes in primary human cancers. *Biochem. Biophys. Res. Commun.* **111**, 47–54 (1983).
43. Andrew P. Feinberg & Bert Vogelstein. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**,
44. Unnasch, T. R. & Wirth, D. F. The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res.* **1**, 6883–6894 (1983).
45. Shen, H. & Laird, P. W. Interplay between the cancer genome and epigenome. *Cell* **153**, 38–55 (2013).
46. Wang, L. *et al.* BRCA1 is a negative modulator of the PRC2 complex. *EMBO J.* **32**, 1584–1597 (2013).
47. Cooper, D. N. & Youssoufian, H. The CpG dinucleotide and human genetic disease. *Hum. Genet.* **78**, 151–155 (1988).

48. Costello, J. F. *et al.* Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat. Genet.* **24**, 132–138 (2000).
49. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.* **8**, 286–298 (2007).
50. Ziller, M. J. *et al.* Genomic distribution and Inter-Sample variation of Non-CpG methylation across human cell types. *PLoS Genet.* **7**, (2011).
51. Sproul, D. & Meehan, R. R. Genomic insights into cancer-associated aberrant CpG island hypermethylation. *Brief. Funct. Genomics* **12**, 174–190 (2013).
52. Craig, J. M. & Bickmore, W. a. The distribution of CpG islands in mammalian chromosomes. *Nat. Genet.* **7**, 376–382 (1994).
53. Watt, F. & Mouoy, P. L. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. 1136–1143 (1988).
54. Derek, A. & Eamonn, H. K. L. DNA methylation: a form of epigenetic control of gene expression. 37–42 (2010).
55. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (80-.)*. **356**, eaaj2239 (2017).
56. Leng, S. *et al.* Defining a Gene Promoter Methylation Signature in Sputum for Lung Cancer Risk Assessment. *Clin. Cancer Res.* **18**, 3387–3395 (2012).
57. Hascher, A. *et al.* DNA methyltransferase inhibition reverses epigenetically embedded phenotypes in lung cancer preferentially affecting polycomb target genes. *Clin. Cancer Res.* **20**, 814–826 (2014).
58. Javier Carmona, F. & Esteller, M. Dna methylation in early neoplasia. *Transl. Pathol. Early Cancer* **9**, 101–111 (2012).
59. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome — biological and translational implications. *Nat. Rev. Cancer* **11**, 726-734 (2011).
60. Verigos, J. & Magklara, A. Revealing the complexity of breast cancer by next generation sequencing. *Cancers (Basel)*. **7**, 2183–2200 (2015).


61. Bethesda, M. N. C. I. PDQ® Screening and Prevention Editorial Board. 2017 Jun 19 Available at: <https://www.ncbi.nlm.nih.gov/books/NBK65906/>. (Accessed: 3rd July 2017)
62. Ministry of Health, the E. and C. C. M. Breast Screening Programme. (2012).
63. Pace, L. E. & Keating, N. L. A systematic assessment of benefits and risks to guide breast cancer screening decisions. *Jama* **311**, 1327–35 (2014).
64. Lin, Z. *et al.* Value of circulating cell-free DNA analysis as a diagnostic tool for breast cancer: a meta-analysis. *Oncotarget* **8**, 26625–26636 (2017).
65. Radpour, R. *et al.* Hypermethylation of tumor suppressor genes involved in critical regulatory pathways for developing a blood- based test in breast cancer. *PLoS One* **6**, (2011).
66. Breastcancer.org. April 14, 2016 Available at: <http://www.breastcancer.org/symptoms/testing/types/breast-cancer-index-test>. (Accessed: 25th July 2017)
67. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* (80-.). **339**, 1546–1558 (2013).
68. Spannuth, W. A. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.*
69. Diaz, L. A. & Bardelli, A. Liquid biopsies: Genotyping circulating tumor DNA. *Journal of Clinical Oncology* **32**, (2014).
70. Siravegna, G. & Bardelli, A. Genotyping cell-free tumor DNA in the blood to detect residual disease and drug resistance. *Genome Biol* **15**, 449 (2014).
71. Ignatiadis, M. & Dawson, S. J. Circulating tumor cells and circulating tumor DNA for precision medicine: dream or reality? *Ann Oncol* **25**, 2304–2313 (2014).
72. Swaminathan, R. & Butt, A. N. Circulating nucleic acids in plasma and

- serum: Recent developments. *Ann. N. Y. Acad. Sci.* **1075**, 1–9 (2006).
73. Dawson, S.-J. *et al.* Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* **368**, 1199–209 (2013).
 74. Mandel P, M. P. *Les acides nucléiques du plasma sanguin chez l'homme (Nucleic acids in human blood plasma) C R Acad Sci Paris.e.*
 75. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57–68 (2016).
 76. Zhong, X. Y. *et al.* Elevated level of cell-free plasma DNA is associated with breast cancer. *Arch. Gynecol. Obstet.* **276**, 327–331 (2007).
 77. Nawroz-Danish, H. *et al.* Microsatellite analysis of serum DNA in patients with head and neck cancer. *Int. J. Cancer* **111**, 96–100 (2004).
 78. Diehl, F. *et al.* Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc Natl Acad Sci U S A* **102**, 16368–16373 (2005).
 79. Leung, S. fai *et al.* Pretherapy quantitative measurement of circulating Epstein - Barr virus DNA is predictive of posttherapy distant failure in patients with early-stage nasopharyngeal carcinoma of undifferentiated type. *Cancer* **98**, 288–291 (2003).
 80. Schwarzenbach, H., Hoon, D. S. B. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer* **11**, 426–437 (2011).
 81. Ellinger, J. *et al.* The role of cell-free circulating DNA in the diagnosis and prognosis of prostate cancer. *Urol. Oncol. Semin. Orig. Investig.* **29**, 124–129 (2011).
 82. Stroun, M., Lyautey, J., Lederrey, C., Olson-Sand, A. & Anker, P. About the possible origin and mechanism of circulating DNA. *Clin. Chim. Acta* **313**, 139–142 (2001).
 83. Jahr, S. *et al.* DNA Fragments in the Blood Plasma of Cancer Patients : Quantitations and Evidence for Their Origin from Apoptotic and Necrotic Cells DNA Fragments in the Blood Plasma of Cancer Patients : Quantitations and Evidence for Their Origin from Apoptotic and Necr.

1659–1665 (2001).

84. Schwarzenbach, H., Hoon, D. S. B. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat. Rev. Cancer* **11**, 426–37 (2011).
85. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryotic chromosome. *Cell* **98**, 285–294 (1999).
86. Volik, S., Alcaide, M., Morin, R. D. & Collins, C. C. Cell-free DNA (cfDNA): clinical significance and utility in cancer shaped by emerging technologies. *Mol. Cancer Res.* **14**, molcanres.0044.2016 (2016).
87. Miller, M. C., Doyle, G. V. & Terstappen, L. W. M. M. Significance of Circulating Tumor Cells Detected by the CellSearch System in Patients with Metastatic Breast Colorectal and Prostate Cancer. *J. Oncol.* **2010**, 1–8 (2010).
88. Dorbeau, M., Bazille, C. & Bibeau, F. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Côlon & Rectum* **11**, 117–118 (2017).
89. Olmos, D. *et al.* Prognostic value of blood mRNA expression signatures in castration-resistant prostate cancer: a prospective, two-stage study. **13**, 1114–1124 (2012).
90. Thakur, B. K. *et al.* Double-stranded DNA in exosomes: A novel biomarker in cancer detection. *Cell Res.* **24**, 766–769 (2014).
91. Fackler, M. J. *et al.* Novel Methylated Biomarkers and a Robust Assay to Detect Circulating Tumor DNA in Metastatic Breast Cancer. *Cancer Res.* **74**, 2160–2170 (2014).
92. DeAngelis JT, Farrington WJ, T. T. A. O. of E. A. M. biotechnology. 2008;38(2):179-183. doi:10. 1007/s1203.-007-9010-y. An Overview of Epigenetic Assays. *Mol Biotechnol* **38(2)**, 179–183
93. Perakis, S. & Speicher, M. R. Emerging concepts in liquid biopsies. *BMC Med.* **15**, 75 (2017).
94. Hayatsu, H., Wataya, Y., Kai, K. & Iida, S. Reaction of Sodium Bisulfite with Uracil, Cytosine, and their Derivatives. *Biochemistry* **9**, 2858–2865

(1970).

95. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci.* **89**, 1827–1831 (1992).
96. Tollefsbol, T. O. DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods Mol Biol* **287**, 316 (2011).
97. Lange, V. *et al.* Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics* **15**, 63–63 (2014).
98. Baker, M. Clever PCR: more genotyping, smaller volumes. *Nat. Methods* **7**, 351–356 (2010).
99. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).
100. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci.* **112**, 201508736 (2015).
101. Aryee, M. J. *et al.* Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
102. Liu, J. & Siegmund, K. D. An evaluation of processing methods for HumanMethylation450 BeadChip data. *BMC Genomics* **17**, 469 (2016).
103. Lu, J. *et al.* PrimerSuite: A High-Throughput Web-Based Primer Design Program for Multiplex Bisulfite PCR. *Sci. Rep.* **7**, 41328 (2017).
104. Underhill, H. R. *et al.* Fragment Length of Circulating Tumor DNA. *PLoS Genet.* **12**, 1–24 (2016).
105.  International Human Genome Sequencing Consortium, L. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
106. Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681 (2015).

107. The, C. *et al.* A Comprehensive Guide To Bisulfite Converted Dna.
108. Tusnády, G. E., Simon, I., Váradi, A. & Arányi, T. BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. *Nucleic Acids Res.* **33**, e9 (2005).
109. Roy, S. & E. Schreiber. Detecting and Quantifying Low Level Gene Variants in Sanger Sequencing Traces Using the ab1 Peak Reporter Tool. **25**, (2014).
110. Cairns, P. Identification and Quantification of Differentially Methylated Loci. *Methods* **507**, 165–174 (2009).
111. Bettegowda, C. *et al.* Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. **6**, (2014).
112. Diehl, F. *et al.* Circulating mutant DNA to assess tumor dynamics. **14**, 985–990 (2008).
113. Forshew, T. *et al.* Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA. **4**, (2012).
114. Liggett, T. E., Melnikov, A. A., Marks, J. R. & Levenson, V. V. Methylation patterns in cell-free plasma DNA reflect removal of the primary tumor and drug treatment of breast. *Int. J. Cancer* 492–499 (2011). doi:10.1002/ijc.25363
115. Laytragoon-lewin, N., Chen, F. U., Castro, J. & Elmberger, G. DNA Content and Methylation of p16 , DAPK and RASSF1A Gene in Tumour and Distant , Normal Mucosal Tissue of Head and Neck Squamous Cell Carcinoma Patients. **4648**, 4643–4648 (2010).
116. Hoque, M. O. *et al.* Detection of Aberrant Methylation of Four Genes in Plasma DNA for the Detection of Breast Cancer. *J. Clin. Oncol.* **24**, 4262–4269 (2006).
117. Skvortsova, T. E. *et al.* Cell-free and cell-bound circulating DNA in breast tumours: DNA quantification and analysis of tumour-related gene methylation. 1492–1495 (2006). doi:10.1038/sj.bjc.6603117
118. Ng, E. K. O. *et al.* Quantitative Analysis and Diagnostic Significance of Methylated SLC19A3 DNA in the Plasma of Breast and Gastric Cancer

Patients. **6**, (2011).

119. Guerrero-preston, R. *et al.* Differential promoter methylation of kinesin family member 1a in plasma is associated with breast cancer and DNA repair capacity. 505–512 (2014). doi:10.3892/or.2014.3262
120. Thomson, J. P. *et al.* DNA immunoprecipitation semiconductor sequencing (DIP-SC-seq) as a rapid method to generate genome wide epigenetic signatures. 1–9 (2015). doi:10.1038/srep09778
121. Feil, R. & Fraga, M. F. Epigenetics and the environment: emerging patterns and implications. **13**, 2012 (2012).
122. Hsiung, D. T. *et al.* Global DNA Methylation Level in Whole Blood as a Biomarker in Head and Neck Squamous Cell Carcinoma. **16**, 108–115 (2007).
123. Lofton-day, C. *et al.* DNA Methylation Biomarkers for Blood-Based Colorectal Cancer Screening. **423**, (2008).
124. Chimonidou, M. *et al.* DNA Methylation of Tumor Suppressor and Metastasis Suppressor Genes in Circulating Tumor Cells. **1177**, 1169–1177 (2011).
125. Mori, T. *et al.* Predictive Utility of Circulating Methylated DNA in Serum of Melanoma Patients Receiving Biochemotherapy. **23**, 9351–9358 (2017).
126. Righini, C. A. *et al.* Tumor-Specific Methylation in Saliva: A Promising Biomarker for Early Detection of Head and Neck Cancer Recurrence. **13**, 1179–1185 (2007).
127. Zhu, T. *et al.* Human Cancer Biology A Novel Set of DNA Methylation Markers in Urine Sediments for Sensitive / Specific Detection of Bladder Cancer. **13**, 7296–7305 (2007).
128. Ahrendt, S. A. *et al.* Molecular Detection of Tumor Cells in Bronchoalveolar Lavage Fluid From Patients With Early Stage Lung Cancer. **91**, (1999).
129. Glöckner, S. C. *et al.* Methylation of TFPI2 in Stool DNA: A Potential Novel Biomarker for the Detection of Colorectal Cancer. **69**, 4691–4699 (2011).

130. Jero, C. *et al.* Detection of Gene Promoter Hypermethylation in Fine Needle Washings from Breast Lesions 1. **9**, 3413–3417 (2003).
131. Bronkhorst, A. J., Aucamp, J. & Pretorius, P. J. Cell-free DNA: Preanalytical variables. *Clin. Chim. Acta* **450**, 243–253 (2015).
132. Jiang, P. & Lo, Y. M. D. The Long and Short of Circulating Cell-Free DNA and the Ins and Outs of Molecular Diagnostics. *Trends Genet.* **32**, 360–371 (2016).
133. Lui, Y. Y. N. *et al.* Predominant hematopoietic origin of cell-free dna in plasma and serum after sex-mismatched bone marrow transplantation. *Clin. Chem.* **48**, 421–427 (2002).
134. Guo, S. *et al.* Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat. Genet.* **49**, 635–642 (2017).
135. Breitbach, S., Tug, S. & Simon, P. Circulating cell-free DNA: an upcoming molecular marker in exercise physiology. *Sports Med.* **42**, 565–86 (2012).

9. APPENDICES

Appendix 1: Principle of Sanger Sequencing and Pyrosequencing

Pyrosequencing uses a high-throughput platform that can interrogate many CpG sites within an amplicon in real time. The pyrosequencing platform is designed to detect single-nucleotide polymorphisms, or SNPs, which can be artificially created at CpG sites through bisulfite modification, as the two strands of genomic DNA are no longer complementary as unmethylated cytosines are deaminated to uracils, which do no longer base pair with the unmodified guanine in the formerly complementary strands. Assuming that methylation occurs consistently in both strands around the CpG analyzed, both strands can be used for pyrosequencing amplification assuming that methylation occurs consistently in both strands of a CpG site.

This technology is distinct from Sanger sequencing, in which normal deoxynucleotides and labeled dideoxynucleotides (ddNTPs) are incorporated randomly in the reaction. The latter terminate DNA strand elongation. These chain-terminating nucleotides lack a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, causing DNA polymerase to cease extension of DNA when a modified ddNTP is incorporated. The ddNTPs may be radioactively or fluorescently labeled for detection in automated sequencing machines. As a result, extension of strands is representative of each nucleotide position; rather, pyrosequencing uses a sequencing-by-synthesis system in which nucleotides are dispensed one at a time, incorporated into the extending strand and degraded prior to the next nucleotide dispensation.

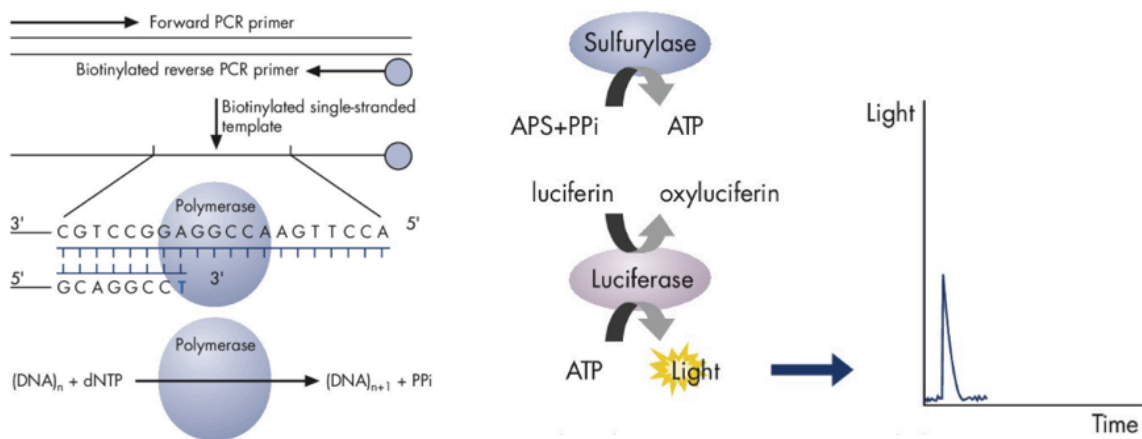


Figure 9: Enzyme cascade system in pyrosequencing (Modified from Qiagen). First, a bisulfite-converted DNA segment is amplified via PCR with one of the primers being biotinylated. After denaturation, the biotinylated single-stranded PCR amplicon is isolated and incubated with a sequencing primer and the enzymes DNA polymerase, ATP sulfurylase, luciferase, and apyrase, as well as the substrates adenosine 5' phosphosulfate (APS) and luciferin. After successful incorporation of a nucleotide by DNA polymerase into the growing DNA strand, PPI is released and reacts with APS in the presence of ATP sulfurylase giving rise to ATP. ATP in the presence of substrate luciferin and enzyme luciferase produces oxyluciferin that generates visible light, which can be detected by a charge coupled device (CCD) and seen as a peak in the raw data output (Pyrogram). Any unincorporated nucleotides and ATP are degraded into its building blocks by enzyme apyrase prior to the next nucleotide dispensation. Addition of dNTPs is performed sequentially, allowing the complementary DNA strand to be built up and the nucleotide sequence is determined from the signal peaks in the Pyrogram trace. ATP (adenosine triphosphate), APS (Adenosine 5' phosphosulfate), PPI (pyrophosphate).

Appendix 2: List of designed 48.48 Fluidigm Access Array Primer pairs.

Green background primer pairs are targeting 23 unique probes and orange background primer pairs were used as positive controls.

<i>Probe</i>	<i>Forward sequence</i>	<i>Reverse sequence</i>
<i>cg00017461</i>	AATACCCTTCCTACCCTCT	GGGGGATGTTGGTTTTGG
<i>cg00911290</i>	GGTAGGTAGGGGAAAAGG	CTAACCCCAAAAACCACA
<i>cg02184606</i>	AAAATATTCCTCCTCCTCC	ATGTTTTTGTGTTTTGTTTTAGTTGT
<i>cg07481320</i>	AAAAAAAAACCCTCCCCC	AAATAAAAAGAAATTAGAGTAGTTAAAA
<i>cg12435154</i>	CTACAACACCACCTCCAC	TTTGTGTTTTATGTTTTGTTTTGA
<i>cg17518965</i>	CATTATTCTACTACAACCAC	TTAGTAAGTTTTTAGTATTATTAGG
<i>cg22331159</i>	TGGAGGTTTAGGGTGGGT	ATAAATAAACCAAAAAATAATTCTTCT
<i>cg26680502</i>	ATCTCAAAAAACAACCCTATC	TATTATTTTTTTTTTTGGAAAGAAGG
<i>cg02082674</i>	AAAAAAAATACCAATACCCTAACCTACTT	GGTGTTTTTTTAGATGGGTTTATGAGG
<i>cg03236137</i>	GATATAGTTTAGTTTTAAATTGTGTTAGAGA	CTAAACACTCCTTATATCCCAACCA
<i>cg08556894</i>	GTGTGTGAGTTAATTGTAAAGAGGA	CCTATACTATTATCAATAAATTTTTTACCAAC
<i>cg09034874</i>	AACCACCTTCCCCAAATTCCTATTT	TTATTTTTTTTAGTTTTGTGTTTTTTGGATAAGA
<i>cg16260696</i>	ATTTTGTTTTAAAAGAAATAAATAAAGAATTGGAG	TCCCACCTTCCCATAACCC
<i>cg19827883</i>	AAGAAAAGTGAAGGAAGGAAGAAG	TTCAAATCAACAATACCCACAACATAATAA
<i>cg21777700</i>	TAGAGTTTTTTATGTAAGATGTTATTATGAGA	TATTTTTAAAATATAATAAAAAAAACTCACCACC
<i>cg25534244</i>	AAACTCAAACTCAAAACAAAAAAACTT	AGGAGAGGGTTGTAAGAGAGGA
<i>cg26427109</i>	TAACTCTCCACTTCCCTTCTTCTT	GTTTGTGTTGGGAGTTTATTTATTAGG
<i>cg03195881</i>	CCAATAAACTAAAATTTTTCTTTC	GGTTAGGTTTGGTAGGAGG
<i>cg15853475</i>	TTTTGTTGAATTTAGTGTTATTTTATAG	ACCCACCTAACCTCCCA
<i>cg27082467</i>	TGTTTTAGTTTTTTGTGTTATGTTTT	CACCTACCCTCCCCCT
<i>cg07222863</i>	YGGATTAGGGGTGGGTGTTGAG	CCCCAACTCCAACAACCCAATACTC
<i>cg09489894</i>	TTAGGAGGGTGAAGTTAGAGGATGATGTG	AACRTATTCTCTTACCTCCTTCTAATCC
<i>cg11509179</i>	TTAGGAGGYGGAATGTAGAGAATGTGG	ACACCTATCRCCCAAACCTTAACAAAC
<i>cg13656001</i>	AGGATAGGGTTGGAAGGAAATAGGAAG	CCCTCTACAAATCTTCTCCAATCTTCCCC
<i>cg14533732</i>	ATTAGTAGAGGAGTAAAGGGGTGTTGGAG	ATTCCCAATTCAACAACAACCTACCTTTCC
<i>MCF2L2_rc2</i>	GTGTAGGGGAGGGATTTAGGAAGAAGTTTATAAG	TACAAACACACAAAAAAACTACACCCACAAC
<i>IKZF1_rc2</i>	GGTTAGGTTAGGGAAGAGTTAGGGATAGGTTG	AAAAATATTCCACCTCCCCTCTTCAACATTAC
<i>ZIC1_c</i>	TTGTTGGGTTTTTGGGATGAGAATTTGGG	CAATCCTTCCCCTCCCTCCAACCTC
<i>SOX14_a</i>	AATTTGAAGGGTTAGGTTGGAAGGGG	TTGGTCTRCACCCTAATTTCTAATTTCCCATCTAAACC
<i>cg00248115</i>	TTGGTGAGTTAGTTGGGGAAGGAGAGG	ACCCACCTAACACTTCTCCCTTACC
<i>cg01259145</i>	GGGGGAATAGAGGTTAGGGTGGAG	CAACCTCATCCTCTCTCCCCCTACC
<i>cg01314252</i>	AATGTTTTATGAAGGGGAGGAAGTTGTGGG	AAAAAAAACRAACCCACACACTAACC
<i>cg03598297</i>	GGTTYGGAGTTGTTGTAGGGTGG	TAACACRCAATTTTTCCAAAACCTTCC
<i>cg03608224</i>	GGGGTTGYGGGAAAGTTAAGTTAGG	TACRTATCCAAAACAACCTCCAAATTCCC
<i>cg04163216</i>	GGTTTAAAATAGGAAGTGGGGAGGGGG	CCCTAAACTAAACCTAACCCATAACCTCC
<i>cg06768993</i>	AAGATGGGATAATTGATTGGGGTGGTG	TCCCAACACTTTTACACCTACAACCTTAC
<i>cg06966839</i>	YGTGGGGGGTTTTGAGGATAGGG	TCCACCCCTCTCCTTATTACATCATAACC

Appendix 3: Adjusted p-values obtained from a Wilcoxon signed rank test between paired tumour and leucocyte samples. Paired t-tests are a form of blocking, and have a greater power than unpaired tests when the paired units are similar with respect to “noise factors” that are independent in the different groups being compared. That way the correct rejection of the null hypothesis can become much more likely, with statistical power increasing simply because the random between-patient variation is eliminated.

<i>Probe and position</i>	<i>Adjusted p-value</i>
<i>cg09489894_chr1-248627123</i>	4.51E-05
<i>cg01259145_chr1-202011187</i>	1.16E-06
<i>cg02184606_chr1-202160116</i>	3.80E-07
<i>cg05772390_chr2-136382578</i>	8.48E-06
<i>cg08556894_chr3-52282218</i>	2.22E-06
<i>cg03598297_chr3-4983076</i>	0.000269799
<i>cg00248115_chr3-65937365</i>	4.36E-06
<i>cg01314252_chr3-9902144</i>	9.81E-07
<i>cg02082674_chr3-51686759</i>	3.27E-06
<i>cg06768993_chr4-8441685</i>	3.04E-05
<i>cg07481320_chr6-42771284</i>	1.48E-05
<i>cg07222863_chr8-103141365</i>	2.60E-06
<i>cg19827883_chr9-97993198</i>	1.84E-07
<i>cg16260696_chr11-67403729</i>	7.15E-07
<i>cg21777700_chr12-116325396</i>	2.49E-05
<i>cg14533732_chr15-76312477</i>	0.000314986
<i>cg25534244_chr17-55263738</i>	7.59E-06
<i>cg03195881_chr17-77441759</i>	1.29E-06
<i>cg00911290_chr19-10339685</i>	0.000673586
<i>cg17518965_chr19-3178958</i>	1.28E-05
<i>cg26680502_chr20-38805123</i>	6.52E-06
<i>cg11509179_chr22-30207156</i>	6.50E-08
<i>cg12435154_chr22-37627667</i>	4.31E-06
<i>cg02237342_chr22-24427552</i>	1.64E-07
<i>cg00017461_chr22-30267328</i>	0.000313157
<i>cg04163216_chr1-9717812</i>	0.000224196
<i>cg25534244_chr3-52282218</i>	2.22E-06
<i>cg27082467_chr6-25042281</i>	2.18E-06
<i>cg13656001_chr6-25041715</i>	3.35E-06
<i>cg03608224_chr12-124289434</i>	1.32E-07

Appendix 4: Adjusted p-values obtained from a Wilcoxon signed rank test between cfDNA from healthy plasma and tumour samples. 21 CpG sites were significant with an alpha level of 0.05. Not significant p-values are highlighted in red.

<i>Probe</i>	<i>Adjusted p-values</i>
<i>cg09489894_chr1-248627123</i>	0.04329
<i>cg01259145_chr1-202011187</i>	0.04329
<i>cg02184606_chr1-202160116</i>	0.20979
<i>cg05772390_chr2-136382578</i>	0.04329
<i>cg08556894_chr3-52282218</i>	0.020979
<i>cg03598297_chr3-4983076</i>	0.04329
<i>cg00248115_chr3-65937365</i>	0.034965
<i>cg01314252_chr3-9902144</i>	0.04329
<i>cg02082674_chr3-51686759</i>	0.020979
<i>cg06768993_chr4-8441685</i>	0.397103
<i>cg07481320_chr6-42771284</i>	0.397103
<i>cg07222863_chr8-103141365</i>	0.20979
<i>cg19827883_chr9-97993198</i>	0.020979
<i>cg16260696_chr11-67403729</i>	0.020979
<i>cg21777700_chr12-116325396</i>	0.20979
<i>cg14533732_chr15-76312477</i>	0.20979
<i>cg25534244_chr17-55263738</i>	0.020979
<i>cg03195881_chr17-77441759</i>	0.152575
<i>cg00911290_chr19-10339685</i>	0.04329
<i>cg17518965_chr19-3178958</i>	1
<i>cg26680502_chr20-38805123</i>	0.020979
<i>cg11509179_chr22-30207156</i>	0.04329
<i>cg12435154_chr22-37627667</i>	0.020979
<i>cg02237342_chr22-24427552</i>	0.484688
<i>cg00017461_chr22-30267328</i>	0.04329
<i>cg04163216_chr1-9717812</i>	0.034965
<i>cg25534244_chr3-52282280</i>	0.04329
<i>cg27082467_chr6-25042281</i>	0.020979
<i>cg13656001_chr6-25041715</i>	0.020979
<i>cg03608224_chr12-124289434</i>	0.020979

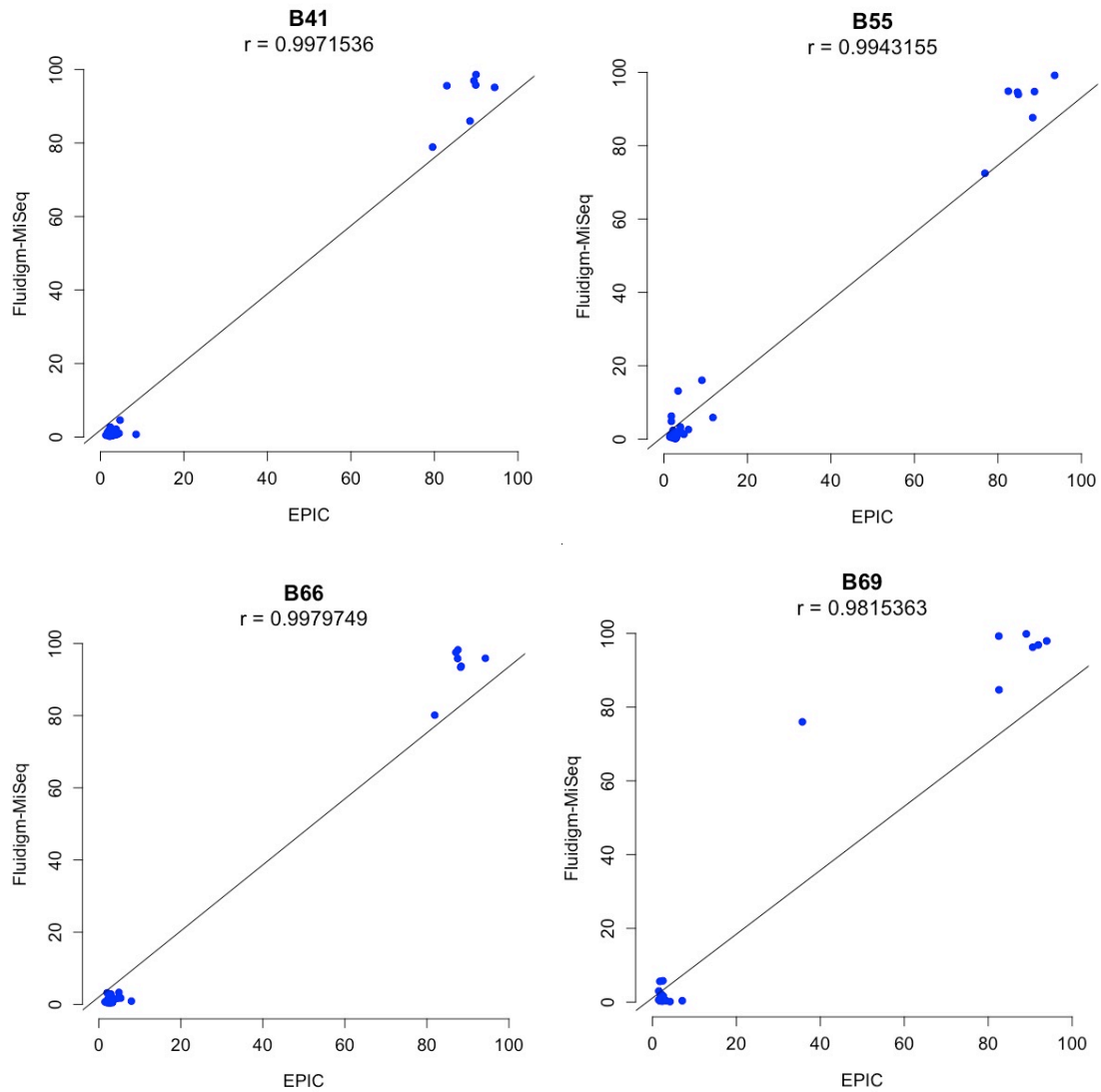
Appendix 5: Adjusted p-values obtained for a Wilcoxon signed rank test between cfDNA from healthy plasma and leucocyte samples. Any of the 30 CpG sites across the genome were differentially methylated.

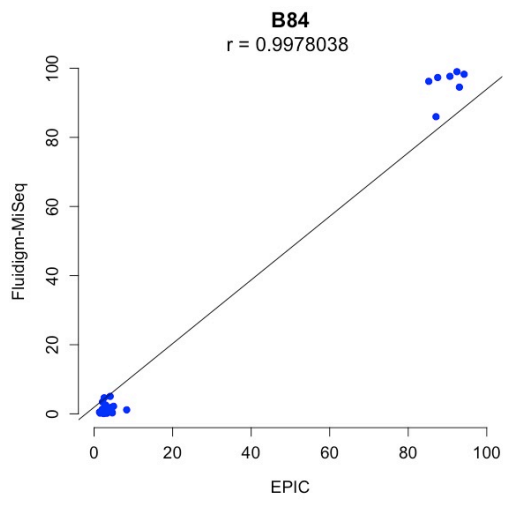
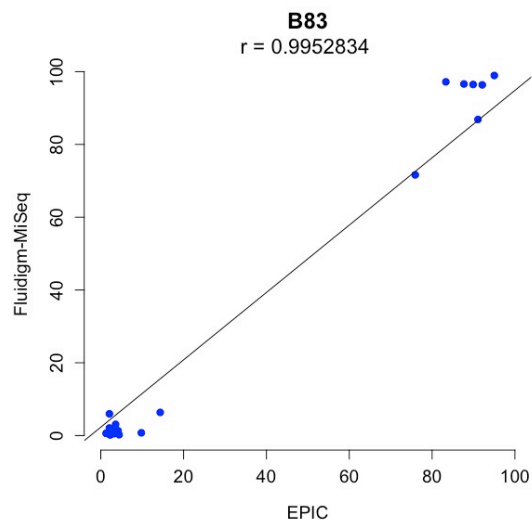
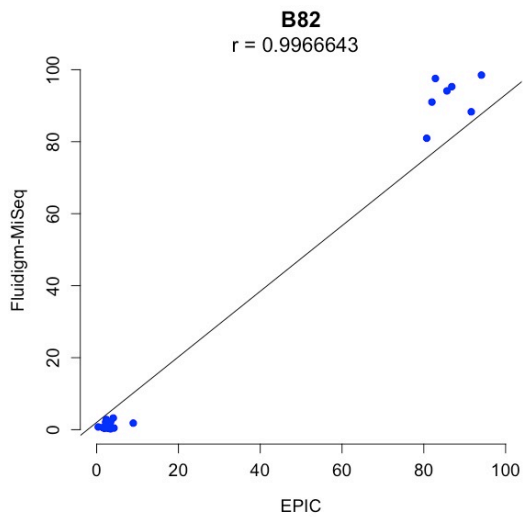
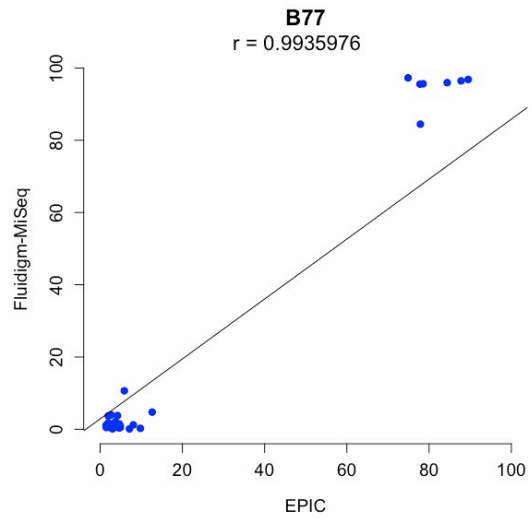
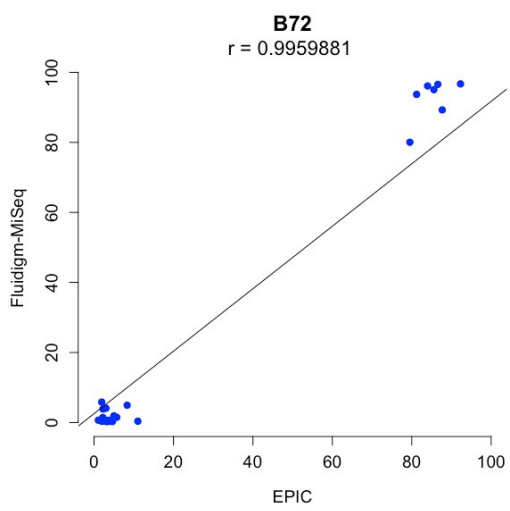
<i>Probe</i>	<i>Adjusted p-values</i>
<i>cg09489894_chr1-248627123</i>	0.210909091
<i>cg01259145_chr1-202011187</i>	0.210909091
<i>cg02184606_chr1-202160116</i>	1
<i>cg05772390_chr2-136382578</i>	0.527272727
<i>cg08556894_chr3-52282218</i>	0.365875447
<i>cg03598297_chr3-4983076</i>	0.210909091
<i>cg00248115_chr3-65937365</i>	0.98018648
<i>cg01314252_chr3-9902144</i>	0.878787879
<i>cg02082674_chr3-51686759</i>	0.98018648
<i>cg06768993_chr4-8441685</i>	1
<i>cg07481320_chr6-42771284</i>	1
<i>cg07222863_chr8-103141365</i>	0.585858586
<i>cg19827883_chr9-97993198</i>	0.98018648
<i>cg16260696_chr11-67403729</i>	0.98018648
<i>cg21777700_chr12-116325396</i>	0.98018648
<i>cg14533732_chr15-76312477</i>	0.878787879
<i>cg25534244_chr17-55263738</i>	0.210909091
<i>cg03195881_chr17-77441759</i>	0.692890443
<i>cg00911290_chr19-10339685</i>	0.98018648
<i>cg17518965_chr19-3178958</i>	0.210909091
<i>cg26680502_chr20-38805123</i>	0.878787879
<i>cg11509179_chr22-30207156</i>	0.502164502
<i>cg12435154_chr22-37627667</i>	0.502164502
<i>cg02237342_chr22-24427552</i>	0.692890443
<i>cg00017461_chr22-30267328</i>	1
<i>cg04163216_chr1-9717812</i>	1
<i>cg25534244_chr3-52282280</i>	0.878787879
<i>cg27082467_chr6-25042281</i>	0.98018648
<i>cg13656001_chr6-25041715</i>	0.978198272
<i>cg03608224_chr12-124289434</i>	0.628671329

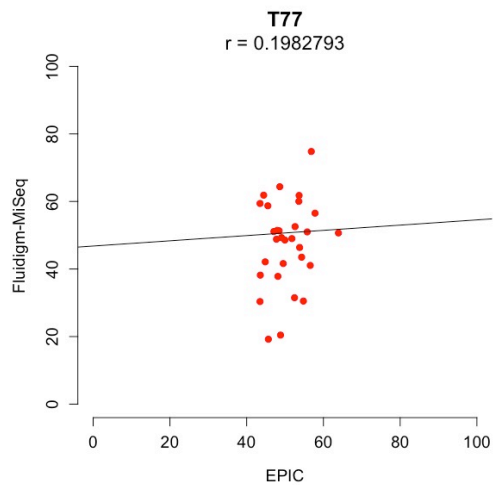
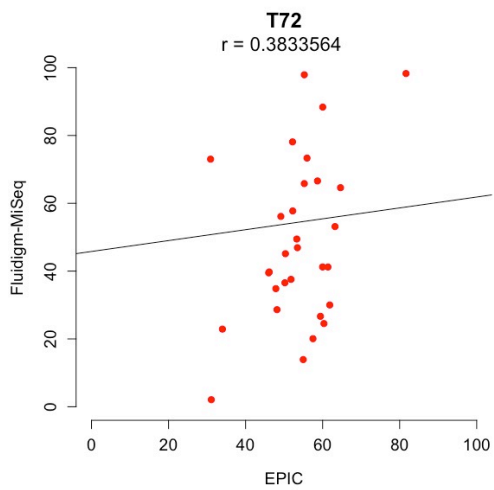
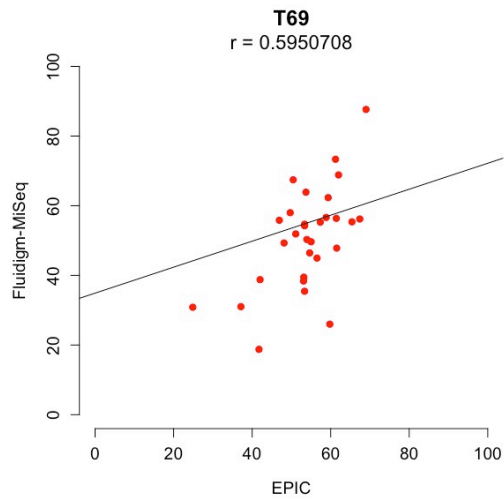
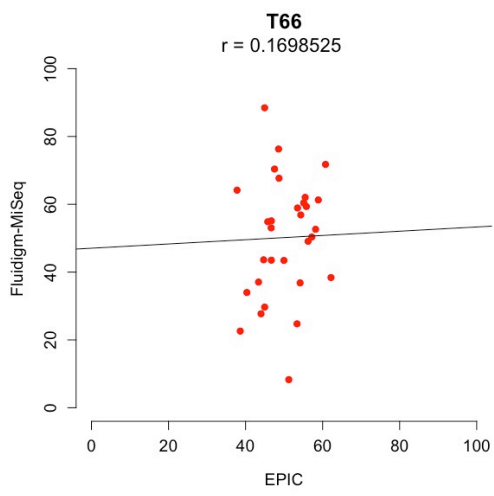
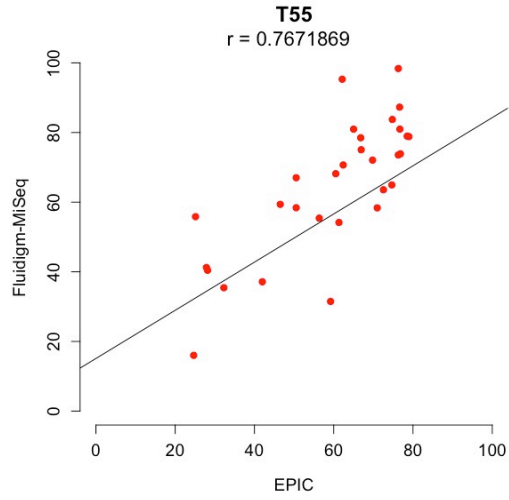
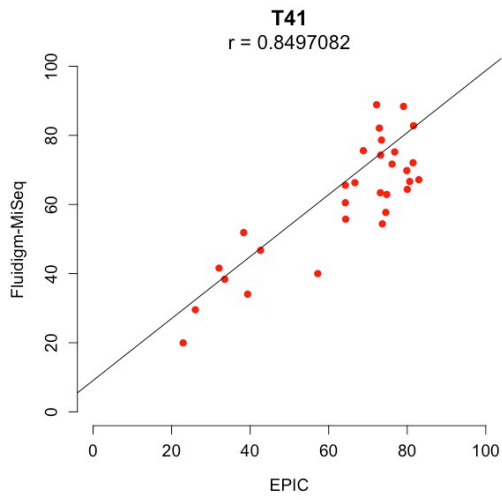
Appendix 6: Table summarizing primers names, primer identifiers (IDs) and the sequences of the primers forward and reverse designed for Sanger Sequencing.

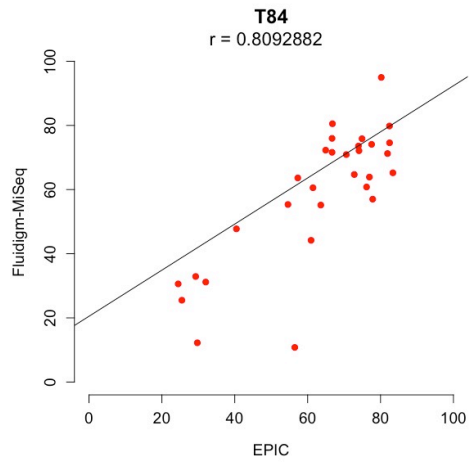
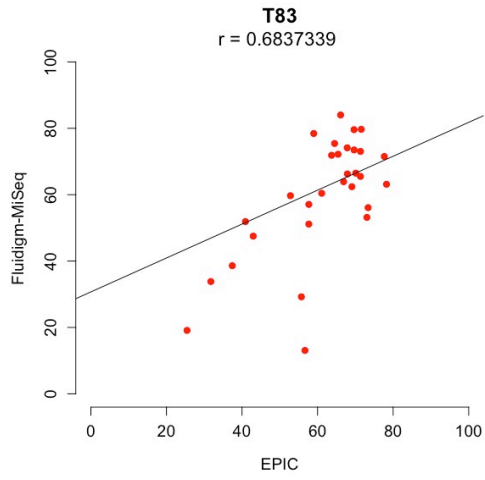
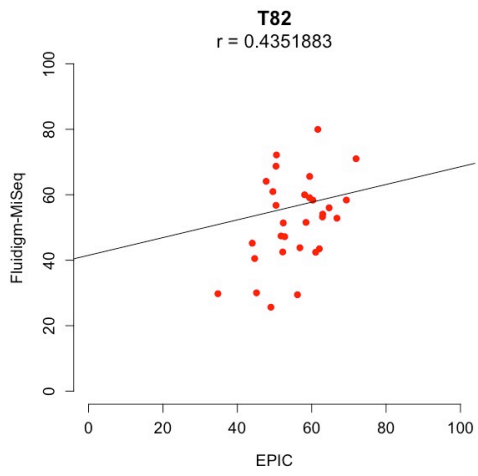
<i>Primer name</i>	Primer ID	Primer forward	Primer reverse
3.1	cg09471455	AATTTGAGGTATTTTAGGTTTGG	CTCACAATTCCTATTCTACCTTAAAACC
8.1	cg17698295	GTTGTTTGGTGGGGAGGGGATTG	CTCACACAACCTCCTAAACCC
2.2	cg18081940	GTGTTTTATTGAGTAGAGAGTTTTTTTTTG	AACCACAACCTCTACCCTACC
3.2	cg09471455	TGGGTTTTAAAGATGGTTTGGG	AATAAACTATCAACCCCTCAATCCTC
4.2	cg25922751	GTTTTTTTTATAGTAATAGTTGGAGGAG	AAATAACAAAATCCCCTCCCCAC

Appendix 7: Linear regressions showing the relationships in methylation measurements (%) between EPIC and Fluidigm-MiSeq. 30 CpGs in both tumour and leucocyte paired samples profiled in both platforms passed all the filtering criteria. Pearson's correlation coefficients (r) for leucocyte samples (blue dots) showed a greater positive correlation between both platforms compared to tumour samples (red dots).









Appendix 9: Adjusted p-values obtained from a Wilcoxon signed rank test between paired tumour and leucocyte samples. All CpG sites were significant with an alpha level of 0.05.

CpG ID	Adjusted p-value ("BH" method)
cg04163216_chr1-9717812	0.000116555516362394
cg01259145_chr1-202011187	5,62E+08
cg02184606_chr1-202160058	5,62E+08
cg02184606_chr1-202160065	5,62E+08
cg02184606_chr1-202160077	8,00E+08
cg02184606_chr1-202160110	4,42E+08
cg02184606_chr1-202160116	2,90E+08
cg02184606_chr1-202160128	1,87E+08
cg09489894_chr1-248627123	0.000228783309183215
cg09489894_chr1-248627141	0.000300654390491659
cg09489894_chr1-248627145	0.000529316798742494
cg09489894_chr1-248627158	0.000751176270436417
cg05772390_chr2-136382523	1,27E+09
cg05772390_chr2-136382547	2,15E+09
cg05772390_chr2-136382578	5,09E+09
cg03598297_chr3-4983076	0.000329154335905657
cg01314252_chr3-9902144	5,40E+08
cg02082674_chr3-51686759	9,29E+07
cg02082674_chr3-51686797	0.000101802634397014
cg25534244_chr3-52282218	7,52E+08
cg25534244_chr3-52282280	6,34E+08

cg00248115_chr3-65937335	1,86E+09
cg00248115_chr3-65937365	1,11E+09
ZIC1_chr3-147420637	0.000430914099506005
ZIC1_chr3-147420645	2,74E+09
ZIC1_chr3-147420657	2,63E+09
ZIC1_chr3-147420676	3,16E+08
ZIC1_chr3-147420698	0.000398905184218684
ZIC1_chr3-147420707	0.0027920717612231
ZIC1_chr3-147420727	0.0012650664537736
MCF2L2_chr3-183265342	0.000137764400011713
MCF2L2_chr3-183265348	0.000167442900503826
MCF2L2_chr3-183265402	0.00437091678868553
cg06768993_chr4-8441685	4,46E+09
cg13656001_chr6-25041715	9,29E+07
cg08578703_chr6-25042218	5,87E+08
cg08578703_chr6-25042240	6,37E+08
cg08578703_chr6-25042247	9,29E+07
cg08578703_chr6-25042267	5,00E+08
cg08578703_chr6-25042269	5,40E+08
cg27082467_chr6-25042281	7,52E+08
cg07481320_chr6-42771271	1,61E+09
cg07481320_chr6-42771276	1,33E+09
cg07481320_chr6-42771279	1,86E+09
cg07481320_chr6-42771284	2,54E+09
cg07481320_chr6-42771288	1,54E+09
cg07481320_chr6-42771296	2,63E+09

cg07481320_chr6-42771312	2,63E+09
cg07481320_chr6-42771317	2,54E+09
cg07481320_chr6-42771326	3,25E+09
cg07481320_chr6-42771331	2,63E+09
cg07481320_chr6-42771342	2,63E+09
IKZF1_chr7-50337000	3,66E+09
cg03629107_chr8-100666181	0.0293513612679617
cg03629107_chr8-100666211	0.0134632866719105
cg03629107_chr8-100666228	0.00803421266593726
cg07222863_chr8-103141321	0.000102282315100579
cg07222863_chr8-103141328	0.000584196458342506
cg07222863_chr8-103141336	7,52E+08
cg07222863_chr8-103141365	8,27E+08
cg26427109_chr11-60971500	1,65E+08
cg26427109_chr11-60971501	0.00798526951336363
cg26427109_chr11-60971524	4,68E+07
cg26427109_chr11-60971534	1,03E+07
cg26427109_chr11-60971548	1,10E+07
cg16260696_chr11-67403720	4,10E+08
cg16260696_chr11-67403729	4,42E+08
cg21777700_chr12-116325396	0.000136059324822913
cg03608224_chr12-124289434	1,87E+08
cg14533732_chr15-76312477	0.000376747488990707
cg15853475_chr16-29746244	0.000271565419970362
cg15853475_chr16-29746264	0.000215609800468024
cg15853475_chr16-29746267	0.000186996131840244

cg25534244_chr17-55263702	6,95E+08
cg25534244_chr17-55263733	9,11E+08
cg25534244_chr17-55263738	1,54E+09
cg25534244_chr17-55263741	1,38E+09
cg25534244_chr17-55263744	1,27E+09
cg25534244_chr17-55263769	1,27E+09
cg03236137_chr17-77405256	1,65E+08
cg03236137_chr17-77405280	4,10E+08
cg03236137_chr17-77405285	9,62E+08
cg03195881_chr17-77441704	8,27E+08
cg03195881_chr17-77441759	5,62E+08
cg03195881_chr17-77441772	0.000242822978379575
cg03195881_chr17-77441785	9,11E+08
cg17518965_chr19-3178888	1,54E+09
cg17518965_chr19-3178893	1,86E+09
cg17518965_chr19-3178901	1,61E+09
cg17518965_chr19-3178905	2,63E+09
cg17518965_chr19-3178907	2,41E+09
cg17518965_chr19-3178915	5,83E+08
cg17518965_chr19-3178925	1,83E+09
cg17518965_chr19-3178941	1,54E+09
cg17518965_chr19-3178951	6,50E+09
cg17518965_chr19-3178958	2,26E+09
cg00911290_chr19-10339670	5,09E+09
cg00911290_chr19-10339685	0.000740337361460006
cg00911290_chr19-10339688	0.000431198846854671

cg00911290_chr19-10339699	8,74E+09
cg09034874_chr20-36646312	0.0012513425335745
cg26680502_chr20-38805082	1,34E+08
cg26680502_chr20-38805089	2,06E+09
cg26680502_chr20-38805103	9,76E+08
cg26680502_chr20-38805123	1,42E+09
cg02237342_chr22-24427488	2,90E+08
cg02237342_chr22-24427507	2,76E+08
cg02237342_chr22-24427542	1,65E+08
cg02237342_chr22-24427547	1,87E+08
cg02237342_chr22-24427552	1,87E+08
cg11509179_chr22-30207152	2,56E+07
cg11509179_chr22-30207156	1,65E+08
cg00017461_chr22-30267301	0.000421607029616094
cg00017461_chr22-30267304	0.000496789543845604
cg00017461_chr22-30267328	0.000376747488990707
cg00017461_chr22-30267338	0.00318520223052771
cg00017461_chr22-30267361	0.000593193935958368
cg00017461_chr22-30267377	0.00375151359264498
cg12435154_chr22-37627667	1,11E+09
cg12435154_chr22-37627678	5,62E+08
cg12435154_chr22-37627693	1,18E+09