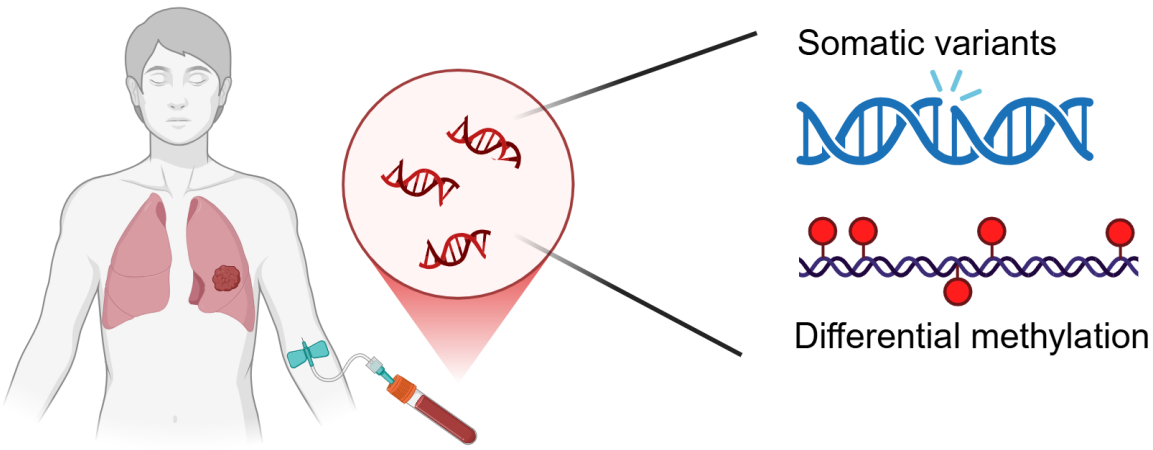
CHALMERS
UNIVERSITY OF TECHNOLOGY



Methylation and mutational analysis of circulating tumour DNA in non-small cell lung cancer using short and long read sequencing technologies

Master's thesis in Biotechnology

Therese Grönqvist

DEPARTMENT OF LIFE SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Methylation and mutational analysis of circulating tumour DNA in non-small cell lung cancer using short and long read sequencing technologies

Therese Grönqvist

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Life Sciences
Division of Chemical Biology
Chalmers University of Technology
Gothenburg, Sweden 2025

Methylation and mutational analysis of circulating tumour DNA in non-small cell lung cancer using short and long read sequencing technologies
Therese Grönqvist

Supervisor: Anna Rohlin, Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital

Examiner: Fredrik Westerlund, Division of Chemical Biology, Department of Life Sciences, Chalmers University of Technology

Master's Thesis 2025
Department of Life Sciences
Division of Chemical Biology
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Illustration of a patient with lung cancer, highlighting somatic variants and differential methylation in ctDNA collected via liquid biopsy. Created with Biorender.com.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Methylation and mutational analysis of circulating tumour DNA in non-small cell lung cancer using short and long read sequencing technologies
Therese Grönqvist
Department of Life Sciences
Chalmers University of Technology

Abstract

To this day, non-small cell lung cancer remains one of the malignancies with the highest mortality rates in the Nordic countries. Since the introduction of targeted molecular therapy and immunotherapy, treatment outcomes have greatly improved, however, issues still remain. Patients are often diagnosed at late stages of the disease and not all patients respond to treatment. Mutational and methylation signatures in specific genes have been introduced as potential biomarkers to predict and monitor treatment response. Further, it has been shown that these biomarkers can be identified in cell-free DNA extracted from blood samples.

In this thesis, a long read sequencing technology, capable of detecting both mutational and methylation information, was set up and used in parallel with two short read sequencing technologies, one utilising enzymatic methyl sequencing to find methylation signatures and one using standard Illumina sequencing to detect somatic variants. This was done in an effort to evaluate the methods and find potential biomarkers in cell-free DNA from non-small cell lung cancer patients, that could be used to predict and monitor treatment response, as well as the spread of metastases.

The results show that, while more optimisation is needed to enable proper analysis of differential methylation, methylation signatures could be visualised using long read sequencing. The methylation patterns found were also consistent with the patterns found using the enzymatic methyl sequencing, and some regions were identified were certain samples showed differential methylation patterns. Somatic variants were also identified but all results require further analysis and evaluation before conclusions can be drawn with regard to potential biomarkers. The conclusion of the study was that, while the long read sequencing method require additional time before it is perfected, both long and short read sequencing can be used to detect differential methylation. The results also suggest the possibility of detecting potential biomarkers.

NSCLC, Methylation, ctDNA, cfDNA, Long read sequencing, Short read sequencing, Liquid biopsies

Acknowledgements

First, I would like to thank my supervisor Anna Rohlin for giving me the opportunity to work with this project. Your help and support has been invaluable and I have been able to gain so much new knowledge about this very fascinating subject. I am thankful for all interesting discussions, both regarding the project but also the research field over all. Your skills and knowledge within the field are truly inspiring!

I would also like to extend a big thank you to Johanna Svensson, who has also taught me so much and has supported me throughout the project. I appreciate the time you have devoted to helping me and discussing with me when I have felt stuck. I am truly grateful to have had your support.

Further, I would like to thank everyone at Clinical Genetics and Genomics at Gothenburg University and Sahlgrenska Academic Hospital for your warm welcome into the department and for the support I have been offered this past year. Special thanks go to Eddie Vuong, for introducing me to Oxford Nanopore™ long read sequencing, Maria Yhr and Josephine Wernersson for teaching me and helping me throughout this project, and Alvar Almstedt for your technical support and assistance with bioinformatic analyses. I would also like to extend a thank you to the patients of the BioLung study, without whom this research would never be possible.

This project has not always been straight-forward and has required a lot of time and effort. Despite this, this past year has been amazing and I am truly happy I got to work on this project. I am deeply grateful for the knowledge and experience I have gained, as well as the amazing people I have had the pleasure of getting to know. Last, but not least, I would like to thank my friends and family for your support and encouragement, not only during this past year, but always, it means a lot to me.

Therese Grönqvist, Gothenburg, May 2025

List of Acronyms

Below all relevant acronyms used in this thesis will be presented. The acronyms are alphabetically ordered.

5-mC	5-methylcytosine
5-gmC	5-(β-glucosyloxymethyl)cytosine
5-hmC	5-hydroxymethylcytosine
bam	Binary alignment map
cfDNA	Cell-free DNA
CHIP	Clonal haematopoiesis of indeterminate potential
CNV	Copy number variation
ctDNA	Circulating tumour DNA
DMR	Differentially methylated region
DNMT	DNA methytransferase
EM-Seq	Enzymatic methylation sequencing
FFPE	Formalin-fixed paraffin-embedded
gDNA	Genomic DNA
GOF	Gain-of-function
HQ	High quality
IGV	Integrative Genomics Viewer
LOF	Loss-of-function
MHB	Methylation haplotype block
NSCLC	Non-small cell lung cancer
PCA	Principal component analysis
PD	Progressive disease
PD-1	Programmed death 1
PD-L1	Programmed death ligand 1
PR	Partial response
QC	Quality control
SBS	Sequencing by synthesis
SD	Stable disease
SNV	Single nucleotide variants
STR	Short tandem repeats
SV	Structural variants
VMR	Variable methylation region
WES	Whole exome sequencing

Contents

List of Figures

List of Tables

1
Introduction

In 2022 there were almost 20 million new cases of cancer globally, as well as approximately 9.7 million cancer-related deaths [1]. In the Nordic countries lung cancer has the highest mortality out of the different types of cancer malignancies, resulting in more than 12 000 deaths annually [2]. Lung cancer can be divided into small cell lung cancer and non-small cell lung cancer (NSCLC), with NSCLC affecting the majority of the patients [2]. New treatment methods using targeted molecular therapy and immunotherapy have improved the outcome of NSCLC patients [3], however, many patients still do not respond to these treatments or experience relapses [2]. One of the problems with treatment of lung cancer is that the majority of patients are diagnosed when they are already in stage III (locally advanced) or IV (distant metastases) [2]. Mutations in specific genes can act as biomarkers, indicating which patients will respond to treatment as well as allowing monitoring of treatment [3]. Recently, it has also been shown that methylation signatures in cell-free DNA (cfDNA) can be used as a similar biomarker [4]. Further, methylation signatures in cfDNA show correlation between the signature and the tissue of origin, enabling "back tracking" of the cell-free DNA [5]. Thus, by utilising mutational and methylation signatures, it can be possible to identify biomarkers that could be used to predict and monitor treatment response, enabling better and more personalised treatment for the patients.

This project is done using patient samples from the BioLung cohort. BioLung is a NSCLC patient study conducted at Sahlgrenska University Hospital and Skaraborg Hospital.

1.1 Aim

The main aim of this thesis was to set up the Oxford Nanopore™ long read sequencing method and compare it to short read sequencing technologies, with regard to detection of mutational and methylation signatures. To this aim, the short read sequencing technologies Hologic® Diagenode's "Human Methylome Service" (Diagenode Cat# G02180000, Belgium) and Eurofins Genomics INVIEW Liquid Biopsy Oncoprofiling (728 genes) service (Eurofins Genomics, Germany) were also used. Samples from circulating tumour DNA (ctDNA) from patients with stage III-IV NSCLC were used. Additionally, data from the sequencing technologies was analysed to identify potential biomarkers indicative of response to treatment and spread of metastases.

1.2 Clarification of task

To specify the aim the following questions will be answered.

- How does the Oxford Nanopore™ sequencing method compare to other methylation analysis methods?
- Can distinct mutational and methylation signatures be observed using the different sequencing techniques?
- Can connections be found between the methylation signatures and response to treatment as well as to the spread of metastases?

1.3 Limitations

One limitation with this project is the number of samples used. Due to limited time and capacity only a selected amount of samples are included in the cohort and not all samples could be used for all methods. This affects the statistical power of the analyses, and also causes individual variation to have a larger impact on the results. The results therefore need to be verified in a larger cohort. It should also be noted that the amount of cfDNA retained from each sample is also subject to individual variation, and will affect what downstream analyses are possible to perform. Due to the limited material, repetition of the experiments using the same sample is not always possible either.

2
Theory

In this chapter the relevant theory will be presented, including sections on non-small cell lung cancer, circulating tumour DNA, somatic mutations, DNA methylation and different next generation sequencing technologies.

2.1 Non-small cell lung cancer

NSCLC is the most common form of lung cancer, accounting for approximately 80-85% of lung cancer diagnoses [2], and it can be divided into adenocarcinoma, squamous cell carcinoma and large cell carcinoma [6]. Approximately 60% of NSCLC patients are already at stage IV when diagnosed, 25% are at stage III, and the remaining 15% are diagnosed at stage I-II [2]. In the USA in 2018, the 5-year overall survival rate of cancer patients with stage IV NSCLC was 1-8% [7]. Common treatment methods for NSCLC include surgery, chemotherapy and radiation therapy [6], however surgery is not an option for patients with stage III-IV NSCLC [2]. In recent years, the standard has instead become that patients diagnosed with stage IV NSCLC are assessed for immunotherapy as part of their first-line treatment [2].

2.2 Immunotherapy

In recent years, immunotherapy has become an essential treatment method for various diseases [8]. It is defined as a treatment method that uses substances that either intensifies or re-establishes the immune system's ability to protect itself from malignancies [8]. The immune system has an inherent ability to protect itself from anything it registers as foreign [9]. It does this by allowing T-cells to detect and bind antigens displayed on antigen-displaying cells [9]. This process is regulated by several immune checkpoint pathways, the programmed death 1 (PD-1) pathway being one of the central ones [9]. When T-cells bind to the programmed death ligand 1 (PD-L1) on an antigen-displaying cell, its activation is reduced [9]. Tumour cells thus display PD-L1 which prevents T-cell activation and allows the tumour cell to survive [9]. Additionally, the binding to PD-L1 can result in the inhibition of the proliferation of both T and B lymphocytes, thereby resulting in an overall weakening of the immune response [9].

For patients with lung cancer, so called immune checkpoint inhibitors have proved themselves beneficial, resulting in both longer survival and increased quality of life [9]. Treatment with PD-1 or PD-L1 inhibitors has shown to significantly increase

overall survival in patients with NSCLC, particularly for patients with high expression (≥50%) of PD-L1 treated with PD-1 inhibitors [2]. Monoclonal antibodies act as the inhibitors and will block either PD-1 or PD-L1, resulting in reactivation of tumour-infiltrating lymphocytes which in turn allows for the detection and elimination of the tumour cells [9]. However, despite the success, not all patients respond to the treatment, and there is also a risk that the patient develops immune-related toxicity, which in some, albeit few cases, can be associated with increased morbidity and mortality [9].

2.3 Clinical response, RECIST

The ability to clinically assess the evolution of a tumour is vital for the evaluation of treatment response [10]. The RECIST guideline was originally published in 2000 and was revised into RECIST (version 1.1) in 2009 [10]. Assessment of the tumour is primarily done by imaging techniques such as CT scans or MRI:s, and the response criteria can be summarised as follows [10]:

- **Complete response (CR):** Total disappearance of all target lesions. Additionally, any pathological lymph nodes must have a reduction of the short axis to less than 10 mm.
- **Partial Response (PR):** A decrease in the sum of the diameters of the lesion of at least 30%, using the baseline diameters as reference.
- **Progressive disease (PD):** An increase of the sum of the diameters of the lesion of at least 20%, using the smallest measured diameters during study as baseline. Additionally, the absolute increase must be at least 5 mm. In case of appearance of one or more new lesions, this is also labelled as progression.
- **Stable disease (SD):** If a lesion, at evaluation, does not qualify as either partial response or progressive disease, it is labelled as stable disease.

2.4 Liquid biopsies and circulating tumour DNA

The ”gold standard” for pathological diagnosis today is considered to be tissue biopsies [11]. It can provide a lot of necessary information needed to make accurate assessments of malignancies, however, it still has some significant issues in terms of acquisition and utility [11]. Tissue biopsies are invasive, require additional time and cost and are associated with clinical complications [11]. Additionally, tissue biopsies are unable to properly show tumour heterogeneity [12] which can affect treatment response of patients [11].

Thanks to recent advancements in medicine and research, liquid biopsies have emerged as a good alternative method to tissue biopsies [13]. Liquid biopsies are minimally invasive and are used to detect biomarkers present in the liquid sample that could have diagnostic or prognostic significance in relation to malignancies, including cancer [14]. Apart from being more easily accessible and less invasive than tissue biopsies, liquid biopsies also allow for better reflection of both spatial and temporal

heterogeneity of the tumour, sequential sampling as well as real-time monitoring [12]. Liquid biopsies can also allow for earlier tumour detection since CT scans usually requires the tumour to be at least 7-10 mm in size, containing one billion malignant cells, to be detectable, whereas a liquid biopsy can detect a tumour containing just 5000 cells [13].

The most common form of liquid biopsy is blood samples as it contains a large array of biological analytes, including biomarkers [12]. "Biomarkers" are biological measurable markers that gives information about the state of the body and the interactions between the biological system and potential hazards, such as diseases [15]. Some relevant analytes that can be found through analysis of blood samples include circulating tumour DNA (ctDNA) and circulating tumour cells [12].

Cell-free DNA (cfDNA) is small fragments of DNA that are released into the blood during cell degradation such as necrosis and apoptosis [11]. The fragment size produced by cell apoptosis is around 180 base pairs, whereas necrosis results in fragments that are over 1000 base pairs [11]. There can be a large variation of cfDNA between different people [16], but also within one individual during the day, with levels of cfDNA typically being higher in the morning and gradually decreasing [17]. Additionally, regular activities such as exercise, but also events such as cancer, inflammation, injury or surgery will cause an increase of cfDNA [17]. Once in the blood stream, the cfDNA is cleared away through renal excretion, but studies have also suggested that immune-mediated clearance as well as endo- and exonuclease activity could contribute to the degradation [17]. Regardless, the mean half-life of cfDNA has been approximated to 15 minutes which makes it a useful tool to show the present state of the body [17].

Tumour cells also release DNA into the bloodstream during apoptosis and necrosis as well as during autophagy and other cell mechanisms [13]. This results in circulating tumour DNA (ctDNA) which becomes a part of cfDNA [13]. CfDNA and ctDNA can be found in blood plasma and just like cfDNA, the levels of ctDNA present in the blood can vary a lot between different individuals [13]. Cancer patients have a much higher level of cfDNA compared to other individuals, and the levels can vary depending on cancer progression, cellular turnover, vascularity and response to therapy [13]. CtDNA can also be characterised by a variety of genetic and epigenetic mutations related to the tumour, such as copy number variations (CNVs), loss of heterozygosity (LOH) and differential DNA methylation [13]. Analysis of DNA variants contained in ctDNA allows for a real-time mutational profile, and makes it possible to track dynamic changes in the tumour as well as tumour heterogeneity [11].

2.5 Somatic variants

Mutations in the genome are divided into somatic mutations and germline mutations [18]. Germline mutations are mutations that occur in so called germ cells, i.e. sperm and eggs [18]. These mutations can therefore be passed on to offspring [18]. Somatic

mutations are mutations that occur in somatic cells, diploid cells that make up the other tissues in the body [18]. Many somatic mutations affect only one or a few cells and do not have significant effect on the health of the individual [19], however, some can still cause disease such as cancer [18]. In many cases, cancers are a result of an accumulation of somatic mutations that affect the genes that control cell division and cell growth [19]. In this thesis the "mutations" in the genome are not assumed to be neither benign nor pathogenic and thus are referred to as variants, as the word "mutation" in some cases implies pathogenicity.

The most common variants are single-nucleotide variants (SNVs) which are caused by a substitution of a single nucleotide in the genome [20]. If sections of the genome have been deleted or inserted the variants are called deletions or insertions [20]. Together, these can be referred to as indels [21]. Indels that are not a multiple of three and thus alters the reading frame of the codon, can be referred to as frameshift mutations [22]. If the insertion or deletion affects a larger area of the genome the variant is referred to as a copy number variation (CNV) [21]. Different variants can also result in different genomic effects. When a SNV occurs in a gene and results in an amino acid change in the protein, the variant is called a missense variant [23]. When a SNV results in a stop codon instead of an amino acid the variant is called a nonsense variant [23]. Indels can result in inversions of a genomic segment, duplications of a genomic segment or translocations, the movement of a genomic segment to a new location [23]. These alterations, CNVs, indels, duplications, inversions and translocations, when affecting a larger area of the genome, are referred to as structural variants (SVs) [24]. A summary of different somatic variants can be seen in figure 2.1. The figure visualises deletions, a few different types of insertions, inversions, translocations, SNVs, CNVs, and frameshift mutations.

Figure 2.1: Figure showing a summary of different somatic variations. These include deletions, tandem and interspersed duplications, inversions, translocations, CNVs, frameshift mutations as well as novel sequence, mobile-element, intra-chromosome and inter-chromosome insertions. Figure created using Biorender.com.

Additionally, loss-of-function (LOF) and gain-of-function (GOF) variants can be defined as variants that result in a partial or complete reduction of, or increase of, activity of a specific gene, respectively [25]. LOF variants are often recessive, while GOF variants are often dominant [25]. Genes in which a GOF variant would cause the cell to become cancerous are called proto-oncogenes, and the corresponding varied genes are called oncogenes [26]. Genes in which a LOF variant would cause the cell to become cancerous are referred to as tumour suppressor genes [26]. Further, genes can be divided into drivers and passengers, where driver variants aid in driving the cancerous development of the cell, and passenger variants are variations that are not connected to the disease but happen to be present in the same cell [27].

Further, all genomes contain so called variable number tandem repeats (VNTRs), which are sequences of DNA that are repeated multiple times, and where the number of repetitions is individual [28]. When this repeating sequence is only two to six nucleotides long, it is referred to as short tandem repeats (STRs), and both STRs and VNTRs can be used for individual analysis [28].

Haematopoiesis, which is the process by which different blood cells are formed [29], generally involve haematopoietic stem cells with equal potential of developing into each different blood cell [30]. However, variants can occur in the stem cell that give

them particular selective fitness advantages, which results in a clonal expansion of these stem cells [30]. An increased risk of these variants occurring can be the result of several different factors, with age being one of them [30]. While these variants can be malignant, they do not guarantee the development of a malignancy, as other additional variants in other genes are often required to cause a malignant transformation [30]. As a result, these mutations are referred to as clonal haematopoiesis of indeterminate potential (CHIP) [30].

2.6 Next generation sequencing technologies

In this section a couple of different next generation sequencing technologies used in this thesis will be explained.

2.6.1 Illumina Sequencing

Illumina sequencing is a short read next generation sequencing technique that utilises "sequencing by synthesis" (SBS) technology [31]. The technology works by clonal amplification of DNA fragments that have been bound to the flow cell, forming clusters, and the sequential addition of fluorescently labelled nucleotides to the template DNA strand [32]. The nucleotide is detected through excitation of the clusters, resulting in characteristic fluorescent signals representing individual nucleotides [32]. This is performed for both the forward and reverse DNA strands [32]. By controlling the number of cycles performed during sequencing, reads of desired lengths can be produced [32].

2.6.2 Oxford Nanopore™ sequencing

Oxford Nanopore™ sequencing (Oxford Nanopore™ Technologies, UK) is a long read sequencing technique that can sequence everything between short and ultra long DNA or RNA fragments in real time [33]. Since it can sequence long reads, PCR is generally not necessary, in contrast to short read sequencing, and thus PCR bias is removed as a source of error [33]. Additionally, another advantage of being able to sequence DNA without PCR or complex library preparation is that the DNA remains undamaged and epigenetic modifications remain intact and can be detected during sequencing [34].

The technique uses nanoscale protein pores called "nanopores" in which a single stranded DNA molecule or RNA molecule passes through starting from the negatively charged *cis* side to the positively charged *trans* side [35]. The pore acts as a biosensor and it is surrounded by a membrane made of electrically resistant polymers [35]. A constant voltage is applied to an electrolyte solution which produces an ionic current which results in the DNA or RNA strand being pushed through the pore [35]. A motor protein pushes the strand through the pore step by step, which determines the translocation speed, and the motor protein also unwinds any double stranded DNA or DNA-RNA duplexes through an inherent helicase activity [35].

The nucleotide bases present inside the pore will cause a change in the ionic current, and this change will be detected by a sensor and analysed using computational algorithms [35]. Modified bases, such as methylated bases, will produce a different change in the ionic current, distinct from the changes produced by unmodified bases, which will be detected by the sensor [36].

2.7 DNA methylation

Epigenetics is defined as a type of heritable modification of the chromatin structure that results in a stable expression of the gene and that is not encoded for in the genetic sequence [37]. Epigenetics include several different modifications, where DNA methylation is one of them [38]. Methylation is used in eukaryotes to regulate gene expression, and the methylation related processes are managed by a few different enzymes [39]. Maintenance methylases methylate new strands of DNA to ensure that they keep the same methylation pattern as the template strand [39]. *De novo* methylases add new methyl groups to DNA strands and demethylases remove methyl groups from methylated DNA [39]. *De novo* methylases are also known as DNA methyltransferases (DNMTs) [38]. The methyl group is usually added to cytosine when it occurs in the sequence CG, a cytosine nucleotide in sequence with guanine [39]. This sequence can also be denoted as CpG, with the p referring to the phosphate group that links the two nucleotides [39]. In DNA, several CpG sites can occur in sequence and form so called CpG islands, and it is common for genes to be located near these islands [39]. If these CpG islands are heavily methylated, the expression of the nearby genes are silenced [39]. This process can be used to silence just one gene, a part of a chromosome or the entire chromosome [39]. The methylation pattern of DNA therefore varies a lot between different types of tissues, depending on what genes need to be expressed for that specific tissue's function [39]. Differential methylation patterns can be divided into hyper- and hypomethylation, with hypermethylation indicating an increased methylation of a certain region and hypomethylation indicating a decrease in methylation [40][41].

The methylation of cytosine is usually denoted as 5-mC (5-methylcytosine), as the methyl group is attached to the 5' position of the nucleotide [4]. The base can also be hydroxymethylated in which case it is denoted as 5-hmC (5-hydroxymethylcytosine) [4]. Since methylation is an important regulatory mechanism for cells, responsible not only for stable gene expression and tissue differentiation, but also genome stability and cell identity maintenance, changes to oncogenes and tumour suppressor genes play an important role in cancer development and tumour phenotype development [4]. The methylation of CpG sites has been used to identify cancers as well as differentiate between different tissues of origin of the tumours [42], and can be used to track the origin of metastases with an unknown primary cancer [4].

CpG islands that are located close to each other have a tendency to have similar methylation patterns, which creates what is called methylation haplotype blocks (MHBs) [42]. These MHBs are easier to distinguish using liquid biopsies than other DNA characteristics when analysing samples from cancer patients [42]. Additionally

it is worth noting that in variable methylation regions (VMRs) it has been observed that cancer derived DNA show a discordant methylation pattern rather than the concordant pattern that would be expected from healthy tissues [43]. Analysis of the MHBs could be used to detect early stage cancer or asymptomatic cancer [42] and could also have great potential use in cancer monitoring, disease prognosis, disease treatment as well as analysis of drug and treatment efficacy and resistance development [4].

2.7.1 Methylation detection methods

There are a few different methods available to detect methylation in DNA. Two of the most relevant methods for this thesis will be described here.

2.7.1.1 The Golden Standard: Bisulfite sequencing

Bisulfite sequencing is a methylation detection technique that utilises sodium bisulfite mediated conversion of cytosines to uracils [44]. Unmethylated cytosines on single-stranded DNA are converted into uracils, while methylated cytosines remain unchanged [44]. The DNA is then amplified using PCR which will result in strands of DNA where the uracils have been converted to thymines and the methylated cytosines remain as cytosines [44]. The DNA can then be sequenced and analysed to find the methylated bases [44]. One issue with bisulite sequencing is that correct reference alignment is crucial for analysis, however sequences never match references perfectly [45]. Additionally the forward and reverse strands need to be considered individually since methylation is not symmetrical and can vary between different cells [45]. As such, regular alignment tools cannot be used, but rather specific bisulfite sequencing analysis tools are required [45]. The coverage that results from bisulfite sequencing is also often subpar, resulting in many reads with low or no coverage [46]. Furthermore, the method requires extreme temperatures and pH values which damages the DNA, and disproportionately damages unmethylated cytosines [46]. This results in degradation and fragmentation and subsequently material loss of DNA, and the disproportionate damage of cytosines causes biased sequencing data with an under-representation of CG dinucleotides [46].

2.7.1.2 Enzymatic methyl sequencing

Enzymatic methyl sequencing (EM-seq) is a methylation detection method that utilises enzymatic conversion of methylated cytosines into products that cannot be deaminated, and then conversion of unmethylated cytosines into uracils, to detect methylation [46]. The enzyme tet mehtylcytosine dioxygenase 2 (TET2) will convert 5-mC to 5-hmC and then the enzyme T4-phage beta-glucosyltransferase (T4-BGT) will convert all 5-hmC to 5-(β-glucosyloxymethyl)cytosine (5-gmC) [46]. 5-gmC is protected from deamination and will thus be unaffected by the next reaction step [46]. Next, an engineered version of the enzyme apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A) is used to deaminate cytosine into uracils [46]. This allows for detection of the methylated cytosines using DNA sequencing [46]. An engineered version of APOBEC3A is used as the human version of the

enzyme has been found to be biased toward TC and CC dinucleotides [46]. EM-Seq has showed advantages over bisulfite sequencing with regards to, amongst other things, coverage, sensitivity, accuracy of methylation calls, CG distributions and number of CpGs identified within genomic features [46]. Additionally, EM-Seq does not result in the same extent of DNA damage [46].

3
Methods

In this chapter the methods used during this thesis will be explained. The study design and meta data for the study will be presented. Then, the methods used for sample preparation will be explained briefly and the techniques used for sequencing and analysis will be presented.

3.1 Study design

The BioLung cohort study is an ongoing observational study which was started in 2019. It includes patients from Sahlgrenska University Hospital as well as Skövde Skaraborg Hospital, and patients are still being recruited. The patients in this cohort study are being treated with immune checkpoint blockade (ICB). Longitudinal sampling is performed at 6 different time points, and blood samples are collected in connection to treatment. There are 3 weeks between the five treatment time points. If progressive disease is observed a progression sample is collected. If progressive disease is not observed, a one year sample is collected. A flowchart over the study design is pictured in figure 3.1.

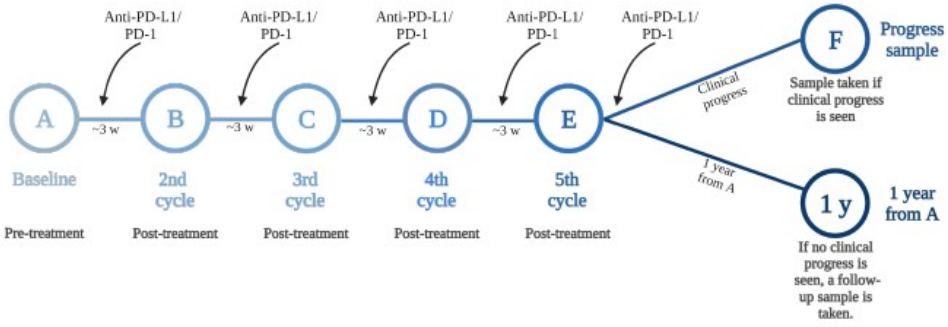


Figure 3.1: Flowchart over study design and sampling for the BioLung cohort study. There are five treatment time points (A-E) and 6 sampling time points (A-F/1y). Source: J. Svensson, "Flowchart over study design and sampling in the BioLung cohort study", *Genetic profiling in non-small cell lung cancer*, 2023, Gothenburg [47].

Note that if progression is observed early during the treatment, a progression sample (F) will be collected immediately, and thus, all time points may not exist for all samples. Blood samples to be used for extraction of ctDNA are collected in Streck™

Cell-free DNA BCT tubes. The Streck™ tube is specialised to reduce release of genomic DNA, and therefore the samples used for this thesis have been collected from these tubes. Another common type of tube used for blood sample collection is the EDTA tube, which is specialised in preventing coagulation. However, EDTA tubes do not stabilise the sample the way Streck™ tubes do and therefore require much more prompt handling after collection.

3.1.1 Meta data

From the larger BioLung cohort, 21 subjects were selected based on cancer stage, metastases and response to treatment. All subjects have stage III - IV NSCLC and have responded poorly to treatment. Out of the 21 samples, 3 were never analysed as other samples were deemed better suited based on concentration of cfDNA, available patient information as well as due to time restraints. 1 cfDNA sample that was sent to Eurofins Genomics for their INVIEW Liquid Biopsy Oncoprofiling panel, but due to very poor quality control (QC) scores of the corresponding normal blood, the sample could not be used for analysis. This left a total of n=17 samples in the final cohort. Table 3.1 presents an overview of these 17 patients used for the study, including the distribution of male and female patients, their age, and histology. Figure 3.2 depicts the patients' clinical response to treatment according to RECIST version 1.1 criteria, which is described in section 2.3.

Table 3.1: Overview of the patients used for this thesis. The table presents the distribution of male and female patients, their age, as well as histology.

Patients	n (%)
Sex	
Female	10 (59%)
Male	7 (41%)
Age (years)	
Mean	67
Median	70
Range	36 - 88
Histology	
Adenocarcinoma	10 (71%)
Squamous cell carcinoma	4 (29%)

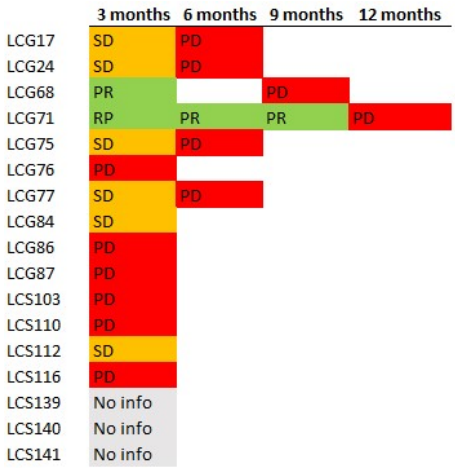


Figure 3.2: A summary of the patients' response to treatment at 3, 6, 9 and 12 months according to RECIST 1.1 criteria. PR stand for "Partial response", SD stands for "Stable disease" and PD stands for "Progressive disease".

Figure 3.3 and figure 3.4 presents the distribution of metastases for the patients. Figure 3.3 depicts all metastases that were found and their frequency in total number of instances. Figure 3.4 presents all the metastases that were found as well as their distribution per patient. The lighter gray areas represents probable metastases that at present have not been properly confirmed. These have been included in the total frequency overview in figure 3.3.

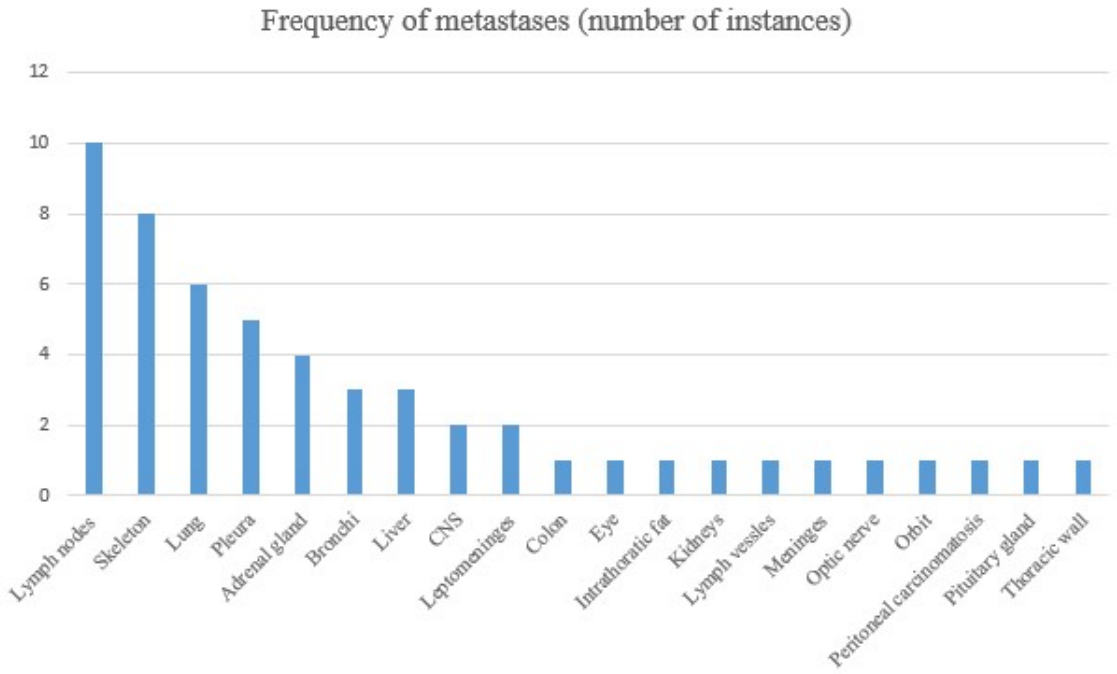


Figure 3.3: An overview of all metastases that were found as well as their total frequency in number of instances.

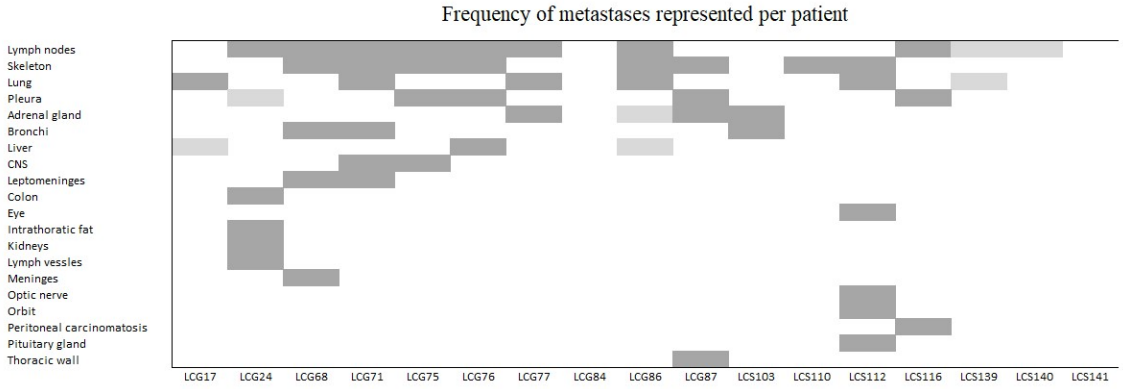


Figure 3.4: An overview of all metastases that were found as well as their distribution per patient. The lighter gray areas represents probable metastases that have not been confirmed.

Three different main methods were used as part of this thesis: EM-Seq performed by Hologic® Diagenode's human methylome service, Eurofins Genomics INVIEW liquid biopsy oncoprofiling panel utilising short read sequencing and Oxford Nanopore™ Technologies long read sequencing. Some methods were used on several samples to enable comparative analysis, but not all methods were used on all samples. Eurofins Genomics INVIEW Oncoprofiling panel for formalin-fixed paraffin-embedded (FFPE) tumour DNA, an in-house GMS560 panel [48] and in-house whole exome sequencing (WES) were used as supplementary methods to Eurofins Genomics INVIEW liquid biopsy oncoprofiling panel. A summary of which samples were run using which methods can be found in table 3.2. The methods will be presented further down in this chapter.

Table 3.2: This table presents a summary of which samples have been run on which methods. The methods include Hologic® Diagenode's human methylome service, Eurofins Genomics INVIEW liquid biopsy oncoprofiling panel and INVIEW oncoprofiling panel for FFPE tumour DNA, Oxford Nanopore™ Technologies long read sequencing, in-house GMS560 gene panel for FFPE tumour DNA and in-house tumour WES.

Sample	Hologic® Diagenode, Human Methylome Service	Eurofins Genomics, INVIEW Liquid Biopsy Oncoprofiling	Oxford Nanopore Technologies, long read sequencing	Eurofins Genomics, INVIEW Oncoprofiling	In-house, GMS560 gene panel	In-house, WES
LCG17	x		x	x		
LCG24	x		x	x		
LCG68	x		x	x		
LCG71	x			x		
LCG75		x			x	
LCG76		x			x	
LCG77	x	x	x			x
LCG84		x			x	
LCG86			x			
LCG87			x			
LCS103		x				
LCS110	x	x	x	x		
LCS112	x					
LCS116	x					
LCS139		x			x	
LCS140		x				
LCS141		x			x	

3.2 Extraction of cfDNA

Blood samples were separated into plasma, buffy coat, and remaining blood. The cfDNA was then extracted from plasma using the automated EZ1 & 2 ccfDNA Kit (Qiagen, Germany), a magnetic bead-based extraction technique. Alternatively, the magnetic bead-based manual extraction kit QIAmp MinElute ccfDNA Mini Kit (50) (Qiagen, Germany), was used. After extraction, the cfDNA samples were stored at -80° C.

3.3 Enzymatic methyl sequencing

Eight cfDNA samples were sent to Hologic® Diagenode for methylation analysis. Their Human Methylome Service (Diagenode Cat# G02180000) was purchased, which utilises EM-Seq technology to sequence the DNA and retain methylation information. The technique includes enzymatic conversion for library preparation, hybrid-capture using a Human Methylome panel, spanning a wide range of targets including over 84% of human CpG island sites [49], and then sequencing using paired-reads Illumina sequencing with an average of 60 M raw reads generated per sample. The service also includes quality control (QC) and a standard bioinformatic analysis where statistics for the whole sample and for the CpG sites are presented.

Apart from the standard analysis report provided by the company, a comparative analysis of different patient groups was requested. These groups were determined based on different response to treatment (at 3 months) as well as different metastases, and were the following:

- Partial response vs Progressive disease (PR vs PD)
- Partial response vs Stable disease (PR vs SD)
- Few metastases (<2) vs Many metastases (>=2)
- Skeleton metastases vs No skeleton metastases
- Lymph node metastases vs No lymph node metastases.

The metastases for comparison were chosen based on frequency in the cohort. A summary of which patients belong to which group is presented in table 3.3. LCS110 and LCS116 appear in both progressive disease and stable disease as the comparative analysis ”Partial response vs Stable disease” was done first and ”Partial response vs Progressive disease” was added later.

Table 3.3: This table provides a summary over which samples were included in which groups for the comparative analysis performed by Hologic® Diagenode.

Partial response	Progressive disease	Stable disease
LCG68	LCS110	LCG17
LCG71	LCS116	LCG24
		LCG 77
		LCS110
		LCS112
		LCS116
Few metastases	**Many metastases**	
LCG17	LCG24	
LCG71	LCG68	
LCS110	LCG77	
	LCS112	
	LCS116	
Skeleton metastases	**No skeleton metastases**	
LCG68	LCG17	
LCS110	LCG24	
LCS112	LCG71	
	LCG77	
	LCS116	
Lymph node metastases	**No lymph node metastases**	
LCG24	LCG17	
LCG68	LCG71	
LCG77	LCS110	
LCS116	LCS112	

For the analysis of the results, the statistics from the comparative analysis reports provided by the company was first regarded. This report included, amongst other things, dendrograms, principal component analysis (PCA), volcano plots and heat maps showing differential methylation between the pre-decided sample groups. The report also included information about where in the CpG and genomic regions the majority of the differentially methylated regions (DMRs) were found, as well as differential methylation percentage per chromosome and percentage of hyper- versus hypomethylated DMRs. The company provided the comparative data files annotated either by DMR or CpG regions as well as between intronic, intergenic, promoter, CpG island, CpG shore, CpG shelf or CpG open sea regions. All files had been filtered to only contain CpGs or DMRs with more than 25% difference in methylation and a q-value (adjusted p-value) of less than 0.01. The files containing CpG sites in the promoter region were chosen for additional filtering and further analysis in the Integrative Genomics Viewer (IGV). Additional filtering was performed in RStudio to only include sites with a differential methylation between 25-50% with a q-value cut-off of less than $1 * 10^{-10}$. These cut-offs were decided based on literature studies and patterns observed when analysing the raw data files.

The code used to filter the files can be found in Appendix A.

From the filtered files, positions with gene ID ”NA” were hidden, as well as all positions with gene ID beginning with ”LOC”, as these were deemed less relevant, and all genes with only 1 remaining differentially methylated position were ignored. All remaining positions were viewed in IGV using the binary alignment map (bam) files containing unique and non-multimapped read pairs. Regions where significant differential methylation between the sample groups were observed were saved in a separate excel document, with positions, p-value, q-value, differential methylation percentage and gene ID annotated. In this document, comments about the associated gene, potential outliers and the visuals in IGV were also added. Additionally, pictures from IGV were included. In this manner, genetic regions of interest were determined. It should be noted that the X and Y chromosomes were not considered during this analysis.

Apart from the analysis of the pre-decided patient groups, differential methylation analysis was also performed between all samples utilising the CpG.methylKit.gz files provided by the company. These files contain all information about the methylation calling for cytosines in the CpG context. This analysis was done using the MethylKit package in RStudio. The samples were filtered to remove regions with less than 10x coverage as well as the 99.9th percentile of coverage and then normalised. Then differential methylation statistics between the samples was determined for the CpG regions. This data could then be used to get the correlation between the samples, as well as dendrograms and PCA plots. The top 13 most significantly differentially methylated CpGs were determined and annotated by region and gene and plotted in a separate heatmap. Additionally, the genes of interest determined by the visual analysis in IGV were plotted in a heatmap. First, the overlap between the differentially methylated CpGs and the genes of interest was determined and an average methylation for the CpGs associated with a certain gene was determined. This data could then be used to create a heatmap showing the differential methylation of the samples for the genes of interest. The code for the methylation analysis using MethylKit can be found in Appendix B.

3.4 Oxford Nanopore™ sequencing

Seven cfDNA samples from the sample cohort were selected to be used for Oxford Nanopore™ sequencing. Additional trial samples were run as well, resulting in a total of twenty one samples sequenced. For two cohort samples and one trial sample, the protocol Ligation sequencing V14 - Human cfDNA singleplex (SQK-LSK114) was used, starting from step 4, ”DNA repair and end-prep”. Sequencing was performed on the PromethION 2 Solo device. For eighteen samples, including five cohort samples, the protocol Ligation sequencing V14 - Human cfDNA multiplex (SQK-NBD114.24), with native barcoding, was used, starting from step 4, ”DNA repair and end-prep”. One multiplex sequencing using five trial samples was performed on the previously mentioned PromethION 2 Solo device. The other samples were sequenced on the PromethION 2 Integrated device. These sam-

ples were divided into three multiplex experiments, one containing five trial samples, one containing five cohort samples and a reference cell line, and one containing two trial samples of cfDNA extracted from blood collected in EDTA tubes, rather than Streck™. For the multiplex experiments, the dorado high accuracy dna_r10.4.1_e8.2_400bps_hac@v4.3.0 basecalling model with 5hmC & 5mC (CG contexts) modifications turned on, was used. For the singleplex experiments, the high accuracy dna_r10.4.1_e8.2_260bps_hac@v4.1.0 basecallig model with 5hmC & 5mC (CG contexts) modifications turned on, was used. Sequencing was performed for 72 hours, according to standard settings.

Before sequencing, the concentration of the samples was measured using Qubit™ Fluorometric Quantification (Thermo Fisher Scientific Inc., US) and the dsDNA HS (High Sensitivity) Assay kit (Thermo Fisher Scientific Inc., US). Quality control was performed using the Agilent 4200 TapeStation system (Agilent Technologies, Inc. US) with the D1000 or High Sensitivity D1000 DNA ScreenTape assay (Agilent Technologies, Inc. US). An additional quality control using Lunatic (Unchained Labs, Inc. US) or DeNovix® DS-11 Spectrophotometer (DeNovis Inc., US) was also performed before sequencing.

After sequencing, the samples were analysed using two different pipelines on the EPI2ME open-analysis platform. One cohort sample and eight trial samples were analysed using the wf-human-variation workflow and all samples were analysed using the wf-somatic-variation workflow. For the human variation workflow, single nucleotide variant (SNV), indel, structural variant (SV) and copy number variation (CNV) calling was turned on, as well as modified basecalling and short tandem repeat (STR) expansion genotyping. For the somatic variation workflow, SNV, indel, SV and modified basecalling was turned on. For both workflows, alignment and QC information was provided.

Compressed and combined versions of the bam files produced for each sample were also analysed in IGV to observe the methylation. For the samples that had been run on both Oxford Nanopore™ long read sequencing and Hologic® Diagenode EM-Seq, the methylation was compared in IGV.

3.5 Variant classification

Nine cfDNA samples, together with corresponding normal blood, were sent to Eurofins Genomics and the service INVIEW Liquid Biopsy Oncoprofiling (728 genes) (Eurofins Genomics, Germany), which utilises a hybridisation-based target capture technology, was purchased. The service purchased provides QC statistics as well as read mapping, read alignment, variant analysis and gene fusion analysis. The variant analysis provides information regarding SNVs and indels, amongst other things. After receiving the data from Eurofins Genomics, the results were analysed using IGV to determine classification.

In order to compare the plasma variants to variants found in the tumour, DNA from FFPE tumour tissue was analysed as well. For one sample this had been done previously using Eurofins Genomics INVIEW oncoprofiling FFPE tumour tissue panel (Eurofins Genomics, Germany) and the existing bam files were used in IGV for comparison with the plasma sample. For five samples, an in-house GMS560 panel [48] was used to analyse FFPE tumour DNA as well as corresponding normal blood. For one sample, in-house whole exome sequencing was performed on the tumour DNA and corresponding normal blood.

The variants that were analysed in IGV were picked from the Variant Call Format (vcf) files based on a few criteria. The variant needed to have an allele frequency equal to or less than 1% in the Genome Aggregation Database and intronic variants were removed. A variant was also only considered if it was present in at least 1% of the reads and/or had at least 2 variant reads in IGV. In general, a variant needed to satisfy both these conditions to be considered, but in cases were a variant had more than 2 variant reads but a frequency just under 1%, it could still be considered.

Based on the results from the analysis of the plasma, tumour and whole blood samples, the variants were classified as somatic, plasma specific or potential clonal haematopoiesis of indeterminate potential (CHIP). The classification was done in the following manner:

- Plasma specific
  - Variant found only in ctDNA in plasma.
- Somatic
  - Variant found in ctDNA in plasma and in tumour but not in blood.
  - Variant found in ctDNA in plasma and in tumour. A small variant frequency found in blood, but less than frequency in tumour and plasma.
- Possible CHIP
  - Variant found in ctDNA in plasma, in tumour and in blood, with a higher frequency in blood than in tumour.
- Germline
  - Variant found in plasma ctDNA or tumour, and in blood, with a frequency between 40-60%.

4
Results

In this section the results will be presented, including the results from the long and short read sequencing technologies, as well as from the different comparative analyses that have been performed.

4.1 Enzymatic methyl sequencing

Eight samples were sent to Hologic® Diagenode for sequencing and analysis using their Human Methylome Panel, which utilises enzymatic methyl sequencing and focuses on methylation analysis. Some analysis reports were provided by the company and additional analysis was then performed, both visually in IGV and by analysis in RStudio. The results are presented below and are divided into QC and methylation analysis.

4.1.1 Quality control

Hologic® Diagenode provided a report outlining the QC results from the sequencing. This report included information about the quality scores of the samples, number of mapped read pairs, mapping efficiency as well as information about multimapping and unique read pairs. Additionally, the report provided information about the coverage of CpG sites and targeted statistics of the CpG sites. A summary of the amount of starting material as well as sequencing statistics can be found in table 4.1 and a summary of the CpG statistics can be found in table 4.2. The Q30 value is the percentage of bases with a phred quality score $\geq$30 and the mean quality is the average quality score of all bases sequenced for each sample. The phred quality score is a logarithmic quality score scale, where a score of 30 represents a base calling accuracy of 99.9%. The number of read pairs kept after trimming refers to the number of read pairs that remain after low quality reads have been removed. Mapping efficiency refers to the percentage of mapped read pairs calculated as mapped read pairs divided by read pairs kept after trimming. The number of unique and no multimapped read pairs is the number of read pairs that remain after PCR duplicates and multimapped read pairs are removed. The average coverage of CpGs with coverage $\geq$ 10, and the average coverage of targeted CpGs with coverage $\geq$ 10, shows the average number of reads mapped to all CpGs detected and all targeted CpGs detected, respectively, with a minimum requirement of 10 reads mapped. Targeted CpGs detected refers to the number of detected CpG sites that also corresponds to sites targeted by the Human Methylome Panel. The proportion of targeted CpGs

detected is the percentage of targeted CpG sites detected with respect to the total number of sites that are targeted by the Human Methylome Panel.

Table 4.1: Summary of QC statistics as well as general statistics from Hologic® Diagenode's standard analysis report. The table includes information about the percentage of bases with a quality score greater than or equal to 30 (Q30), the mean quality of bases, number of read pairs kept after trimming, the mapping efficiency and the number or unique and non-multimapped read pairs for all samples.

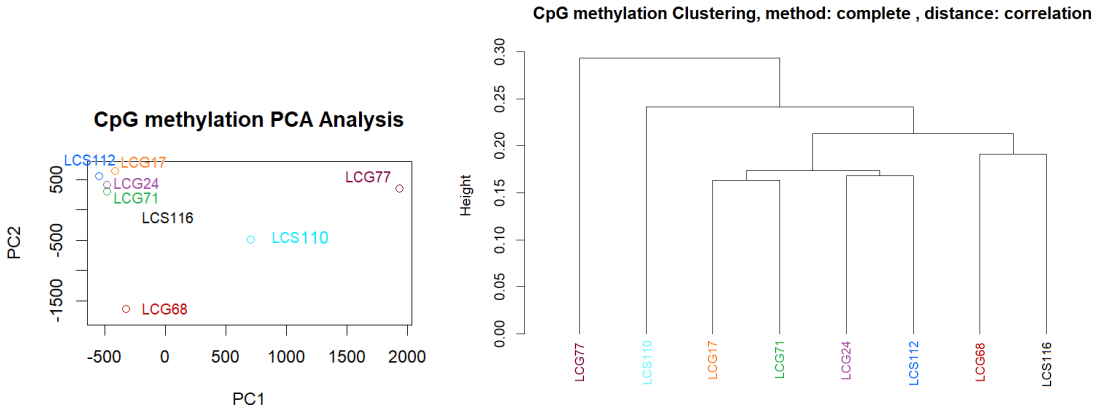
Sample	Starting amount of cfDNA (ng)	Q30 (%)	Mean quality	Number of read pairs kept after trimming	Mapping efficiency (%)	Unique and no multimapped read pairs
LCG17	12.90	91.82	35.57	74,461,930	91.6210	44,839,216
LCG24	15.35	91.56	35.52	74,464,103	90.9420	43,882,707
LCG68	11.78	89.42	35.11	87,120,643	73.3897	36,492,132
LCG71	11.94	91.47	35.51	79,540,387	91.2834	46,448,490
LCG77	13.33	91.38	35.49	70,087,500	88.7497	39,802,417
LCS110	16.27	91.66	35.54	74,631,672	90.6327	43,770,114
LCS112	13.86	91.29	35.47	81,623,446	87.3022	45,885,027
LCS116	15.28	91.51	35.51	84,104,160	89.4370	48,331,409

Table 4.2: Summary of CpG statistics and targeted CpG statistics from Hologic® Diagenode's standard analysis report. The table includes information about average coverage of all detected CpGs as well as targeted CpGs with a minimum coverage of 10x, the number of targeted CpGs detected as well as the proportion of targeted CpGs detected compared to all targeted CpGs included in the panel.

Sample	Average coverage of CpGs, cov $\geq$ 10	Targeted CpGs detected	Proportion of targeted CpGs detected (%)	Average coverage of targeted CpGs, cov $\geq$ 10
LCG17	34.63	3,921,555	98.5086	38.02
LCG24	31.92	3,924,637	98.5861	34.88
LCG68	22.70	3,912,891	98.2910	24.32
LCG71	35.37	3,919,260	98.4510	39.07
LCG77	27.84	3,916,920	98.3922	30.01
LCS110	31.15	3,915,440	98.3550	33.86
LCS112	30.85	3,925,829	98.6160	33.95
LCS116	34.52	3,924,929	98.5934	38.18

4.1.2 Differential methylation analysis

The methylKit browser extensible data (bed) files produced by Hologic® Diagenode's comparative analysis, which contain the methylation information for all samples, was run in RStudio using the MethylKit package. In this analysis, the methylation data of CpG sites was extracted and analysed. A PCA plot as well as a dendrogram was created to visualise the correlation between the different samples, as can be seen in figure 4.1. In the PCA plot, the first principal component (PC1) reveals most of the variance between the samples, while the second principal component (PC2) reveals the second most variance between the samples. The dendrogram shows unsupervised hierarchical clustering between the samples, using correlation distance. The height of the branches shows the degree of difference between the samples, with longer lines indicating larger difference.



(a) PCA plot showing similarity between samples.

(b) Dendrogram showing correlation between samples. The complete clustering method and correlation distance was used.

Figure 4.1: This figure shows similarity and correlation, with respect to the methylation of CpGs, of the samples analysed using the Human Methylome Panel.

In figure 4.1a several samples can be seen clustering in the top left, suggesting that these samples are all similar in regards to the methylation of their CpGs. LCG68 falls in the lower left corner, indicating that this sample is similar to the cluster with regard to PC1, but is dissimilar to the cluster on the PC2 axis. LCG77 falls in the top right corner, indicating that this sample shows a large difference from the other samples with regard to the first principal component, but less difference with regard to the second principal component. LCS110 falls in the middle of the plot indicating that is has some dissimilarities from the other samples in both the first and second principal components. In figure 4.1b as well, it can be seen that LCG77 clusters on its own and therefore shows clear outlier tendencies in relation to the other samples. LCS110 also clusters on its own but shows more correlation the the other samples than LCG77 does.

To visualise the differential methylation of the CpG sites of the samples, the top 13 most significantly methylated CpGs, detected using MethylKit in RStudio, were plotted in a heatmap. The heatmap was done using unsupervised clustering. The CpGs were then annotated with associated region and gene. The resulting heatmap can be seen in figure 4.2. The positive red scale represents hypermethylation, with darker reds representing more hypermethylation. The negative blue scale represents hypomethylation, with darker blues representing more hypomethylation.

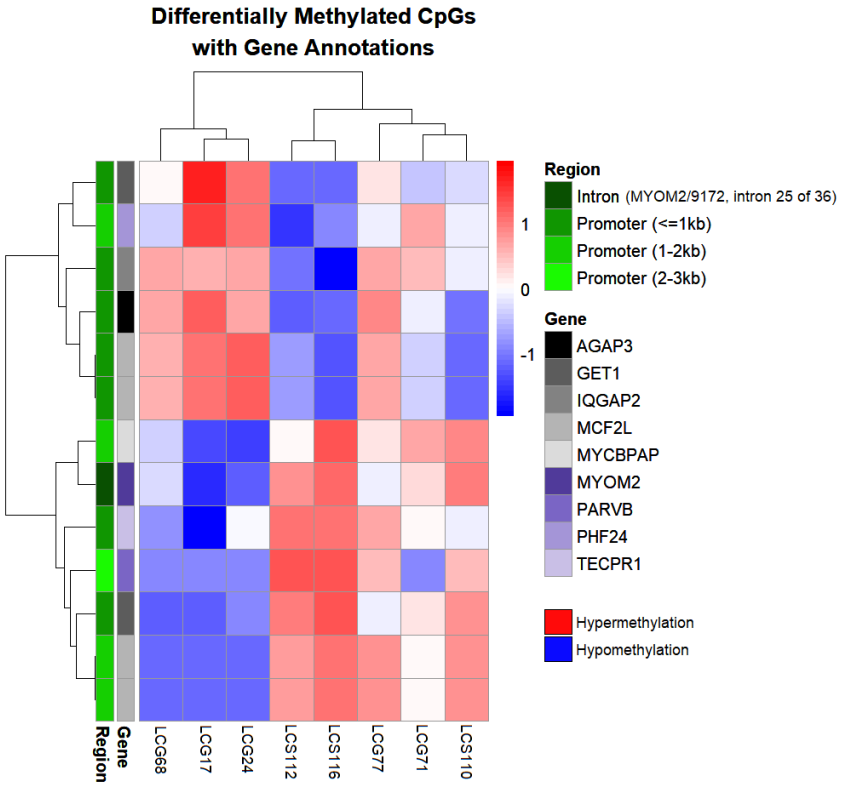


Figure 4.2: Heatmap or the top 13 most significantly differentially methylated CpGs annotated by associated gene and region.

In this heatmap LCG68, LCG17 and LCG24 show similarity to each other, forming a cluster, and LCS112 and LCS116 form another cluster. LCG77, LCG71 and LCS110 cluster slightly, but do not contain as distinct differentially methylated patterns as the other two clusters. The analysis focused on CpG islands in the promoter region, but it can be observed that one intronic region, associated with the gene *MYOM2*, was also found.

To analyse the results of the comparative analysis between pre-decided patient groups, the samples were visualised in IGV. This resulted in a list of 56 genes that were deemed of interest. These 56 genes include both areas where differential methylation was observed between the pre-decided patient groups, as well as areas where one or a few samples showed differential methylation compared to the other samples. The genes that showed differential methylation between the pre-decided patient groups are presented in table 4.3. Note that the genes that showed an in-

crease in methylation compared to the other group have been denoted ”hyper” and that the genes that showed a decrease in methylation compared to the other group have been denoted ”hypo”. Note also that hyper and hypo, in this case, only refer to differential methylation between the groups, and does not denote differential methylation in reference to cfDNA from healthy tissue.

Table 4.3: This table presents the genes that showed differential methylation between the patient groups after comparative analysis visualisation in IGV. The table also show which genes showed an increase in methylation (hyper) and which genes showed a decrease in methylation (hypo) compared to the other group.

Partial response vs Progressive disease	Partial response vs Stable disease	Few metastases vs More metastases	Skelton metastases vs No skeleton metastases	Lymph node metastases vs no lymph node metastases
Hyper for PR:	**Hyper for PR:**	**Hyper for samples with few metastases:**	**Hyper for samples with skeleton metastases**	**Hyper for samples with lymph node metastases:**
S100A1	*EPHB2*	*RNF222*	*TRIM6*	*LMNB2*
S100A14	*S100A1*	*MYO15B*	*NLRC5*	*NINL*
KCNN2	*S100A14*	*NCL*	*NXPH2*	*DNAAF5*
TECPR1	*PARVB*	*PTCSC2*	*GET1*	*UBAC1*
MTUS1	*IQGAP2*	*NALT1*	*FBLN2*	
PHF24			*PLXNB1*	
URAD			*GINS4*	
Hypo for PR:	**Hypo for PR:**	**Hypo for samples with few metastases:**	**Hypo for samples with skeleton metastases:**	**Hypo for samples with lymph node metastases:**
exosc10	*S100A16*	*DPYSL4*	*SMAD3*	*SMYD2*
CYMP	*PM20D1*	*PSMD13*	*MIR4520-2*	*RASSF4*
NGF-AS1	*MYCBPAP*	*RASA3-IT1*	*RNF213*	*PSMD1*
CAPN2	*EPHA1*	*TOGARAM2*	*MIR4458HG*	*EIF4G1*
SNAP47	*AGAP3*	*NINL*	*MYOM2*	*TACC3*
MKLN1				
EPHA1				
MCF2L				
			Showed both hyper and hypo patterns in samples with skeleton metastases:	
			MCF2L	

Of note in table 4.3 is the fact that the gene *MCF2L* showed both increased and decreased methylation in comparison to the group with no skeleton metastases.

These 56 genes of interest were investigated in RStudio as well. A heatmap was created showing the differential methylation of all samples for the genes of interest, as can be seen in figure 4.3. In this heatmap, brighter red colours represents more hypermethylation while darker blue colours represents more hypomethylation.

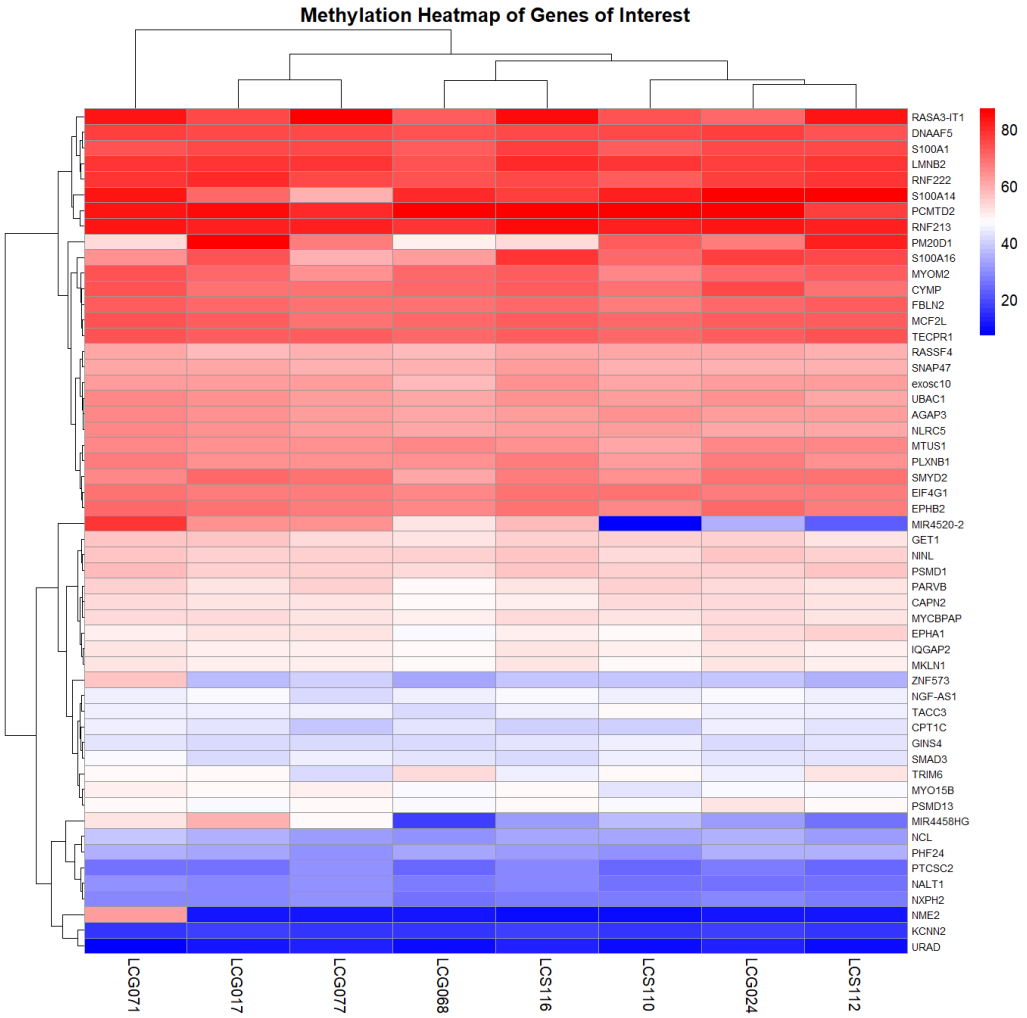


Figure 4.3: Heatmap showing the differential methylation of all main genes of interest found during the visual analysis in IGV.

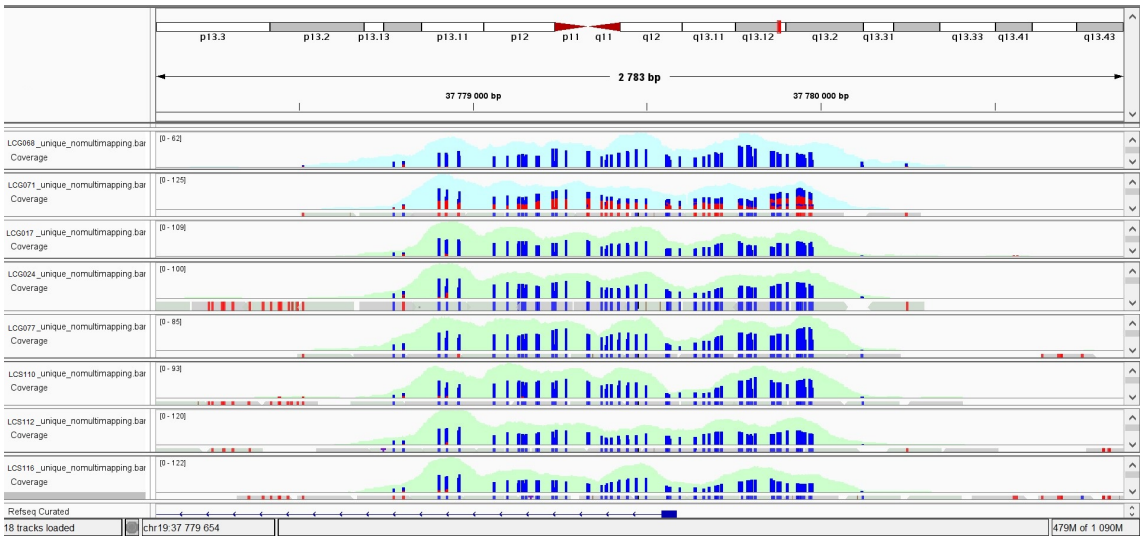
It is clear that most samples appear similar for most genes of interest, however some difference can be observed, particularly for the genes *PM20D1*, *MIR4520-2*, *ZNF573*, *MIR4458HG* and *NME2*.

Both the visual analysis in IGV and the heatmap visualising the genes or interest revealed individual differential methylation of some samples in certain regions. Firstly, some outliers were observed when the patient groups were compared visually. Ad-

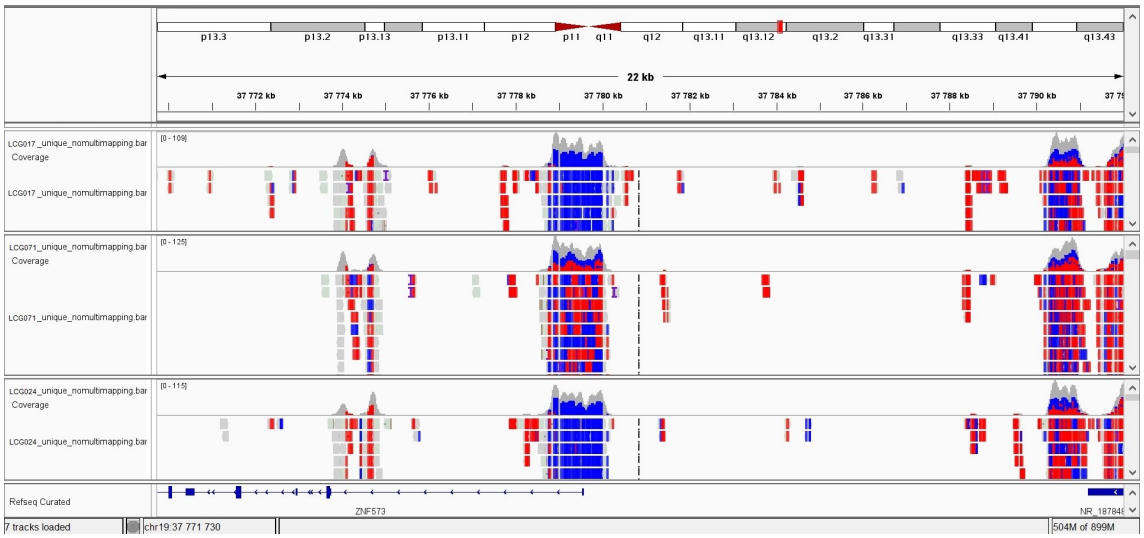
ditionally, some samples had very clear individual patterns in certain regions, and these will be presented in more detail in the following section.

4.1.2.1 Case studies

During the visual analysis in IGV it was observed that LCG71 showed differential methylation over a region of the *ZNF573* gene, in contrast to all other samples. This is shown in figure 4.4, where methylated bases can be seen as red and unmethylated bases are shown as blue. In the coverage bar, the proportion of methylated and unmethylated bases are shown by the colouring. Figure 4.4a shows a slightly smaller region of this gene together with all the samples from the comparative analysis. The samples with a light blue coloured background belong to the "Partial response" (PR) group, while the samples with a light green coloured background belong to the "Stable disease" (SD) group. Figure 4.4b shows the entire region but with fewer samples to more clearly show the differential methylation observed.



(a) A picture of IGV showing all samples from the PR vs SD comparative analysis and a smaller region of the gene *ZNF573*. LCG71 can be seen on the second row from the top.



(b) A picture of IGV showing the entire region of the gene *ZNF573*. LCG71 can be seen in the middle row.

Figure 4.4: This figure shows the differential methylation of LCG71 over the *ZNF573* region. Figure (a) shows a 2783 bp region and all samples from the PR vs SD comparative analysis. Figure (b) shows a 22 kb region with only three samples: LCG17, LCG71 and LCG24, visualised in that order from top to bottom.

LCG71 also showed differential methylation over a region of the *NME2* gene as can be observed in figure 4.5. In this figure all samples from the PR vs SD comparative analysis are included, with the samples belonging to the PR group having a light blue background and the samples belonging to the SD group having a light green background.

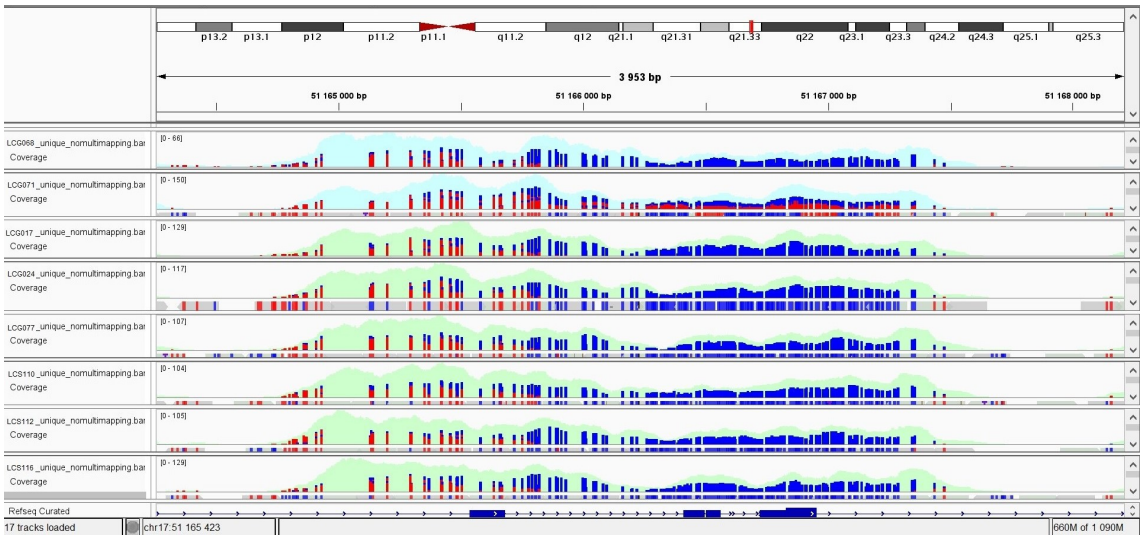
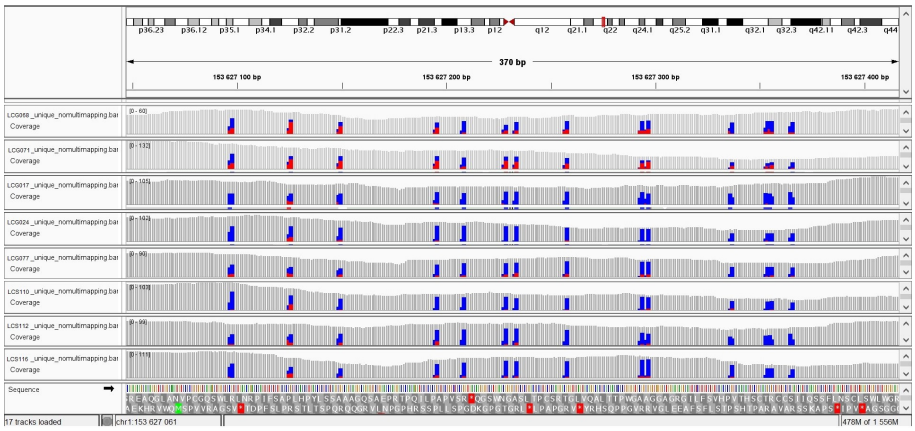
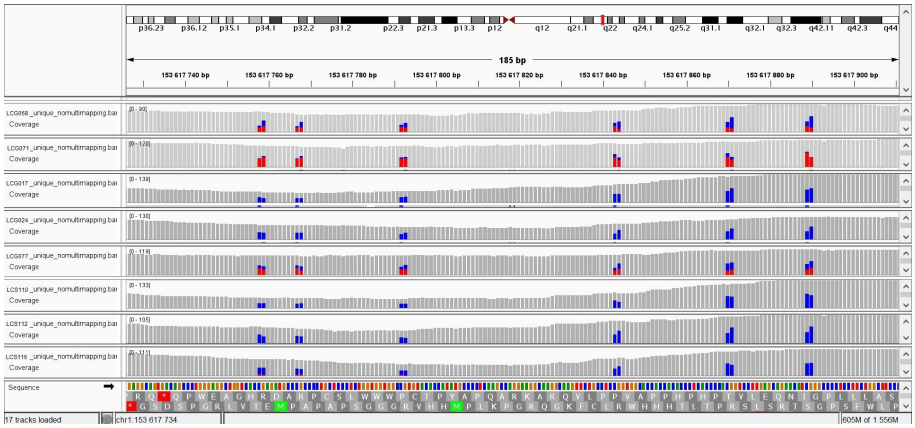


Figure 4.5: Differential methylation of LCG71 over a region of the *NME2* gene. LCG71 can be seen in the second row from the top. All samples from the PR vs SD comparative analysis are included and a region of 3953 bp can be seen.

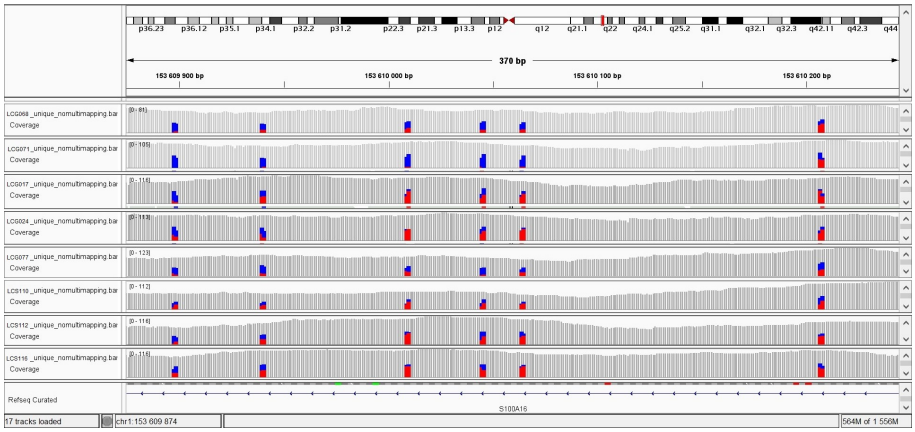
LCG77 was found to be an apparent outlier for some genes included in the *S100A* family, specifically for regions over the genes *S100A1*, *S100A14* and *S100A16*. Despite being classified as stable disease, the methylation pattern was more similar to the samples with partial response. For *S100A14* and *S100A16* the differential methylation was most apparent. Figure 4.6 depicts three pictures of the three different regions from IGV. The samples belonging to the PR group are the top two rows and have a slightly lighter grey background than the other samples. The other samples belong to the SD group. LCG77 can be seen on the fifth row from the top. Figure 4.6a depicts a region over the gene *S100A1*, figure 4.6b depicts a region over the gene *S100A14* and figure 4.6c depicts a region over the gene *S100A16*.



(a) This figure pictures a 370 bp region over the *S100A1* gene.



(b) This figure pictures a 185 bp region over the *S100A14* gene.



(c) This figure pictures a 370 bp region over the *S100A16* gene.

Figure 4.6: This figure pictures LCG77 as an outlier for three different regions, *S100A1*, *S100A14* and *S100A16*, belonging to the *S100A* gene family. The top two rows belong to the PR group and have a lighter grey coloured background, the other samples belong to the SD group. LCG77 can be seen on the fifth row from the top in all pictures.

LCG77 and LCS116 showed differential methylation compared to other samples over a region associated with the *CPT1C* gene, as can be seen in figure 4.7. In this figure

the samples with a light blue coloured background are samples with lymph node metastases and the samples with a light green coloured background are samples with no known lymph node metastases. LCG77 and LCS116 can be seen on the third and forth row, respectively, and show lesser degree of methylation compared to the other samples.

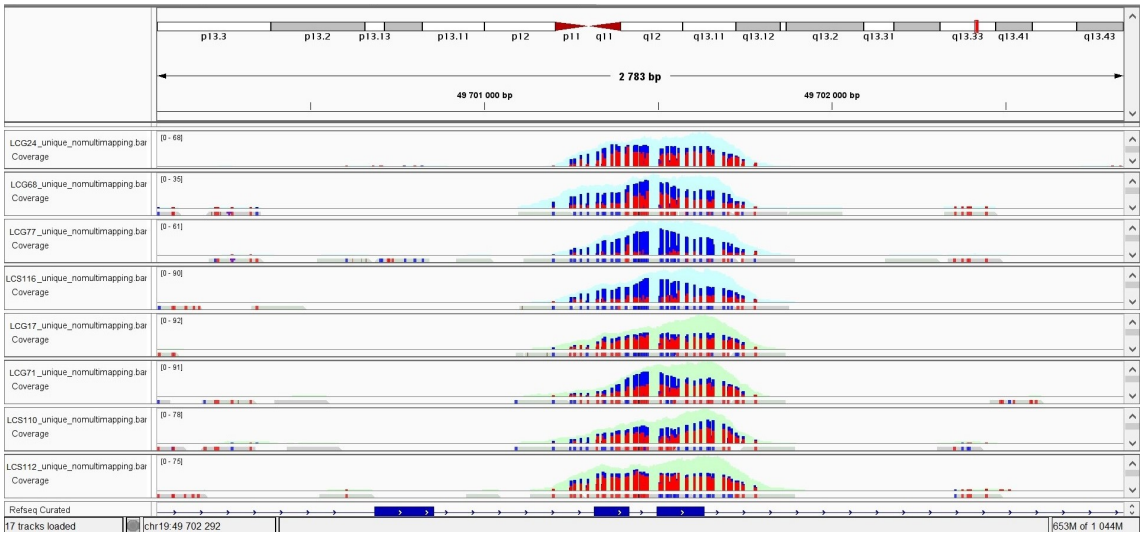


Figure 4.7: This figure shows differential methylation of LCG77 and LCS116, seen on the third and forth row respectively, over a 2783 bp region in the *CPT1C* gene.

LCS112 shows an increase in methylated bases over a region associated with the *PCMTD2* gene not observed in other samples. In figure 4.8, LCS112 can be observed on the third row. Samples with a light blue coloured background are samples with skeletal metastases, while the other samples have no known skeletal metastases.

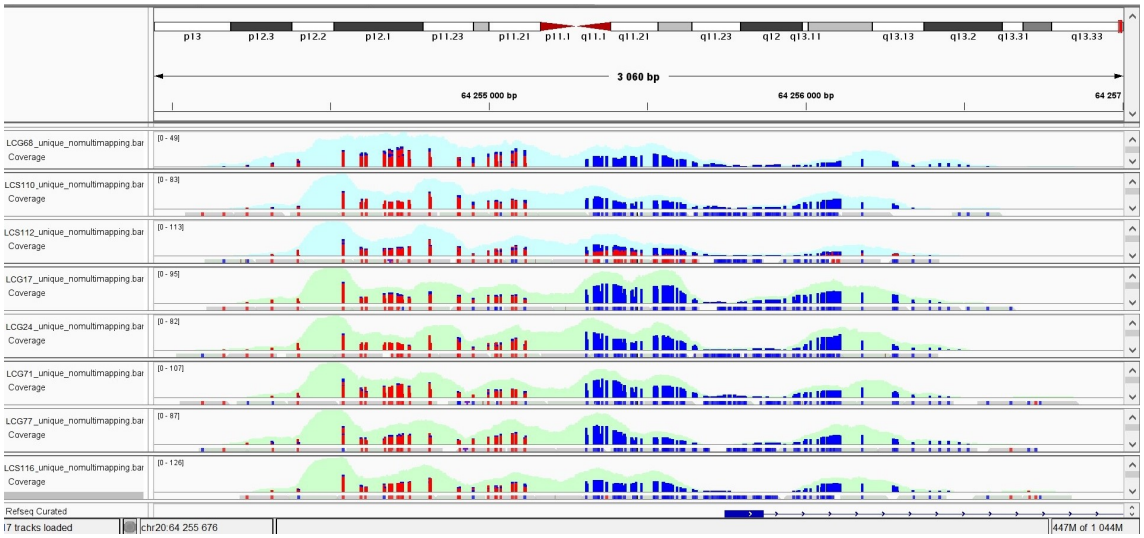


Figure 4.8: This figure shows the differential methylation of LCS112, visible on row three. The picture shows a 3060 bp region over the *PCMTD2* gene.

Over the gene *PM20D1* clear individual differences were observed during the analy-

sis between patients with and without lymph node metastases. LCG17 and LCS112 were significantly methylated, LCG24, LCG77 and LCS110 were noticeably methylated and LCG68, LCG71 and LCS116 were notably non-methylated. These individual patterns can be observed in figure 4.9. The first four rows are coloured light blue and show samples with lymph node metastases, while the bottom four have no known lymph node metastases.

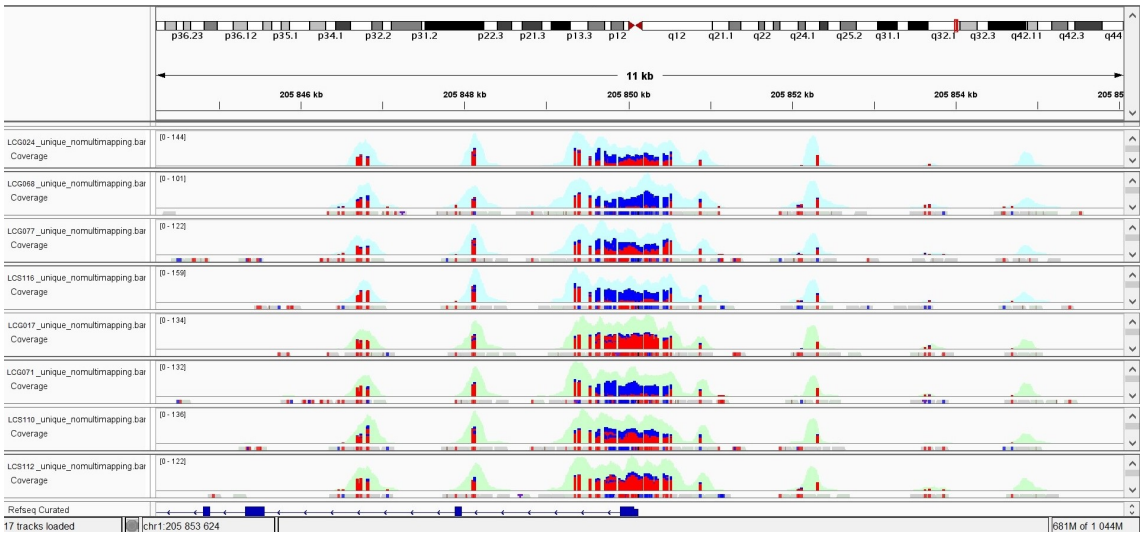


Figure 4.9: This figure shows individual methylation patterns for the different samples over a 11 kb region associated with the *PM20D1* gene.

4.2 Oxford Nanopore™ sequencing

Seven samples from the study cohort were analysed using Oxford Nanopore™ long read sequencing. Apart from these seven samples, several trial samples were run, as well as two samples from blood collected with EDTA tubes, and a reference sample from a cell line. Some of the test samples, as well as one of the study cohort samples were analysed using the wf-human variation pipeline on the EPI2ME platform. All samples were analysed using the wf-somatic variation pipeline on the EPI2ME platform. Focus in this section will be given to the 7 samples from the study cohort.

The SQK-LSK114 singleplex protocol recommends an input amount of 30 ng of cfDNA, while the SQK-NBD114.24 multiplex protocol recommends a minimum of 6 ng cfDNA per sample as input material, preferably adding up to a minimum of 30 ng of cfDNA in total being loaded onto the flow cell. As the flow cells are sensitive to underloading, all cfDNA available was used as input for each experiment ranging between 21-112 ng for the singleplex experiments, and between 17-84 ng in total for the multiplex experiments.

The quality of the samples was checked before library preparation for the sequencing began. The results from Qubit™ (Thermo Fisher Scientific Inc., USA) and TapeStation (Agilent Technologies, Inc. US) suggested that while some samples appeared to

contain some contamination, they were generally of acceptable quality. The results from the Lunatic (Unchained Labs, Inc. US) or DS-11 Spectrophotometer (DeNovis Inc., US) quality check however, showed unexpected 260/280 and 260/230 ratios compared to genomic DNA.

The output from the EPI2ME bioinformatic pipelines generate QC reports as well as variant reports and modified base reports. A summary of the QC reports for the different samples can be seen in table 4.4, including the amount of generated reads per sample and the coverage from the wf-somatic variation pipeline, as well as the coverage from the wf-human variation pipeline, in the case when both were used.

Table 4.4: QC summary of the study cohort samples run with Oxford Nanopore™ long read sequencing. The table includes the number of generated reads, the protocol used, and the coverage from the wf-somatic and the coverage from wf-human variation pipeline.

Sample	Protocol used	Generated reads (*wf-somatic variation*)	Coverage (*wf-somatic variation*)	Coverage (*wf-human variation*)
LCG17	SQK-NBD114.24	10997971	1.36x	
LCG24	SQK-NBD114.24	3322899	0.328x	
LCG68	SQK-NBD114.24	3064520	0.342x	
LCG77	SQK-NBD114.24	5272794	0.514x	
LCG86	SQK-LSK114	9230429	1.26x	1.25x
LCG87	SQK-LSK114	3943073	0.637x	
LCS110	SQK-NBD114.24	11021832	1.35x	

Only one sample from the study cohort was run on both EPI2ME bioinformatic pipelines, but the coverages are comparable. It should be noted that the coverages are very low, with a majority of the samples not reaching 1x coverage.

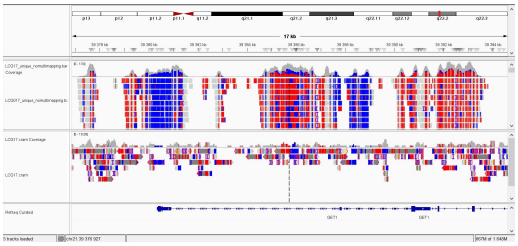
Apart from the QC report, the EPI2ME pipelines also produce reports with CNV, SNV, indels, STR and SV information. Due to low coverage and also, on occasion, poor output quality, not all variants could be determined. The SV calling failed for all samples when run on the wf-somatic variation pipeline. In table 4.5 the amounts of SNVs and indels found are listed for the cohort samples.

Table 4.5: A summary of the amount of SNVs and indels found during variant calling using the wf-somatic variation pipeline and wf-human variation pipeline.

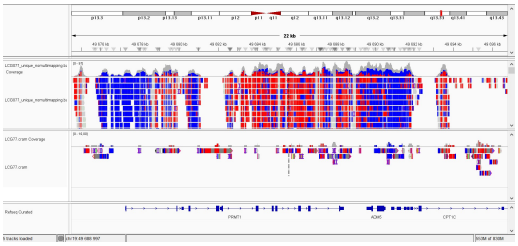
Sample	SNV (*wf-somatic variation*)	Indels (*wf-somatic variation*)	SNV (*wf-human variation*)	Indels (*wf-human variation*)
LCG17	2894	45		
LCG24	550	0		
LCG68	993	0		
LCG77	969	1		
LCG86	1952	19	651,106	35,006
LCG87	996	0		
LCS110	3737	70		

The number of SNVs and indels found in the wf-human variation pipeline is significantly larger than for the wf-somatic variation, which is to be expected since the two pipelines uses two different variant callers. It should be noted that table 4.5 lists all SNVs and indels found, they have not been filtered or classified based on pathogenicity.

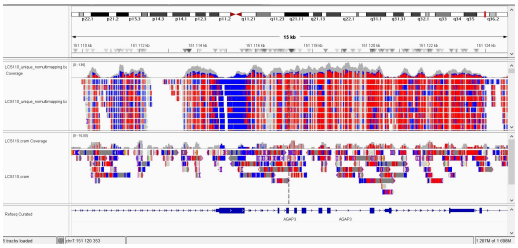
The cohort samples were also viewed in IGV and compared to the results from the short read EM-Seq provided by Hologic® Diagenode. Pictures from a few selected regions are shown in figure 4.10. Figure 4.10a shows LCG17 over a region of the *GET1* gene, with the results from the EM-Seq of the top and the long read sequencing on the bottom. Figure 4.10b shows LCG77 over a region of the *CPT1C* gene with the results from the EM-Seq on the top and the long read sequencing on the bottom. Figure 4.10c and figure 4.10d show LCS110 over a region of the *AGAP3* gene and the *MCF2L* gene, respectively, with the results from the EM-Seq on the top and the long read sequencing on the bottom.



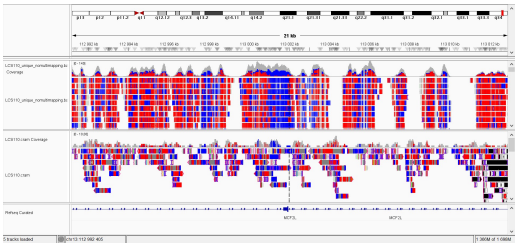
(a) LCG17 over a 17 kb region of the *GET1* gene. EM-Seq results are shown on the top row and long read sequencing on the bottom.



(b) LCG77 over a 22 kb region of the *CPT1C* gene, EM-Seq results are shown on the top row and long read sequencing on the bottom.



(c) LCS110 over a 15 kb region of the *AGAP3* gene. EM-Seq results are shown on the top row and long read sequencing on the bottom.



(d) LCS110 over a 21 kb region of the *MCF2L* gene. EM-Seq results are shown on the top row and long read sequencing on the bottom.

Figure 4.10: This figure shows a selection of pictures from IGV, visualising the methylation results from the EM-Seq and long read sequencing. Red areas represent methylated bases while blue areas represent unmethylated bases.

It can be observed that the coverage from the long read sequencing is significantly lower than for the enzymatic methyl sequencing, causing the comparative analysis to become a bit more complicated. However, in general, the methylation patterns observed for the EM-Seq and the long read sequencing appear to correspond to each other.

4.3 Variant classification

Nine samples, including both cfDNA from plasma (P) and from normal blood (B) for each sample, was sent to Eurofins Genomics for mutational analysis. The sequencing is performed using short read Illumina sequencing. The company provided reports with QC statistics as well as variant information. Additional analysis was then performed by visualisation and classification in IGV, as well as comparison to variants found in FFPE tumour DNA. The amount of input material for each sample was 14 ng of cfDNA. The results can be found below and are divided into QC and the classification of variants.

4.3.1 Quality control

The variant analysis report provided by Eurofins Genomics contained information about, amongst other things, the quality of the samples as well as the number of reads, number of high quality reads and bases, the percentage of bases with a phred quality score above 30, number of unique reads, as well as mean coverage with and without duplicates. A summary of the QC statistics can be found in table 4.6. The letter "P" after the sample name indicates that the cfDNA is from a plasma sample and the letter "B" after the sample name indicates that the cfDNA comes from the normal blood. The total number of high quality (HQ) reads refers to the number of high quality reads that remain after sequence cleaning and filtering. The HQ bases (Q30) is the percentage of high quality bases with a phred quality score of at least 30. Unique reads refer to the number of reads that are mapped exactly on the reference, and the mean coverage (without duplicates) refers to the average coverage of the reference after duplicates have been removed.

Table 4.6: Summary of QC and alignment statistics from Eurofins Genomcs Variant Analysis Report. The table includes information about the total amount of HQ reads, the percentage of HQ bases with a phred score of 30, the number of unique reads as well as the mean coverage after removal of duplicates. "P" indicates plasma samples, while "B" indicates normal blood samples.

Sample	Total HQ reads	HQ bases, Q30 (%)	Unique reads	Mean coverage, without duplicates
LCG75-P	98.29 M	95.75	95.14M	780.58x
LCG75-B	21.61 M	93.27	21.00M	318.18x
LCG76-P	113.71 M	96.18	110.04M	1278.52x
LCG76-B	24 M	92.18	23.37M	317.16x
LCG77-P	100.75 M	95.98	97.67M	1099.98x
LCG77-B	21.37 M	93.77	20.80M	337.25x
LCG84-P	99.73 M	95.59	96.68M	961.41x
LCG84-B	14.67 M	88.63	14.27M	264.47x
LCS103-P	110.95 M	96.10	106.08M	1235.25x
LCS103-B	23.25 M	93.01	22.26M	309.95x
LCS110-P	104.17 M	96.07	99.61M	1040.97x
LCS110-B	23.46 M	92.85	22.43M	319.31x
LCS139-P	107.12 M	95.75	102.17M	939.26x
LCS139-B	12.64 M	94.95	12.04M	208.94x
LCS140-P	116.18 M	96.12	110.81M	1103.84x
LCS140-B	18.59 M	95.79	17.86M	311.18x
LCS141-P	113.51 M	96.05	108.28M	1030.26x
LCS141-B	25.29 M	95.62	24.26M	367.97x

The blood samples have significantly lower numbers of reads and coverage, however, this is in accordance with the service provided, which focuses on the plasma samples

and uses the blood samples as a normal.

4.3.2 Classification of variants

The variants reported were visualised in IGV and the corresponding plasma and normal blood samples were compared against each other. For samples with FFPE tumour DNA data from Eurofins Genomics INVIEW oncoprofiling panel, or from the in-house GMS560 panel or whole exome sequencing service, the variants were also compared against the corresponding tumour DNA and tumour normal blood samples. A summary of the results is presented in table 4.7. For samples with no available tumour material, it was not possible to distinguish between somatic, plasma specific and potential CHIP variants, and thus only the total number of variants found are reported. For these samples, the remaining cells have been filled with "-", to show that analysis was not possible. The table includes the total number of variants found, the number of somatic variants, the number of plasma specific variants, the number of germline variants and the number of potential CHIP variants. The reason most samples have no detected germline variants is because an oncoprofiling panel was used which is not intended to report germline variants.

Table 4.7: A summary of the classified variants found during analysis of Eurofins Genomics liquid biopsy oncoprofiling panel.

Sample	Total number of variants found	Somatic variants	Plasma specific variants	Germline	Potential CHIP
LCG75	15	3	3	1	1
LCG76	11	6	3	0	0
LCG77	34	26	1	0	0
LCG84	7	2	1	0	0
LCS103	11	-	-	0	-
LCS110	5	1	0	0	1
LCS139	5	2	1	0	0
LCS140	8	-	-	0	-
LCS141	1	0	0	0	0

Only one sample had a variant that was classified as germline, and only two samples had variants that could be classified as potential CHIP. The individual variation between the samples was large and LCS110 was noisy in both plasma and tumour making analysis more difficult. In LCS141 only 1 variant, classified as not somatic, was found.

Additionally, for the samples with known driver mutations, it was determined whether these mutations could be found in the plasma cfDNA. The known driver mutations in question include:

- **LCG76:** *KRAS* G12V

- **LCG77:** *ERBB2* amplification without treatment predictive significance

Both these driver mutations were observed during the classification analysis. Additionally, a variant was found in *ERBB2* for LCG77 that was classified as somatic.

5
Discussion

In this section the previously presented results will be discussed. The section will include discussions regarding the study design, as well as the differential methylation analysis and variant classification.

5.1 Study design

It is important to note that as per the study design, several samples are collected from each patient at different time points. Thus, a few different time point samples have been used for the patients, which could affect the analysis. It can also affect the levels of ctDNA detected, as levels of cfDNA present in the plasma is expected to decrease during treatment. The majority of the samples used were collected at time point A or F, see section 3.1, as these samples are expected to have the highest levels of ctDNA. One sample was a time point C sample, which could affect the ctDNA content. For all samples run using both enzymatic methyl sequencing and Oxford Nanopore™ long read sequencing, samples from the same time point were used for both methods, which should eliminate cfDNA levels and tumour evolution as a potential error source for the comparison between the two methods.

For Eurofins Genomics INVIEW liquid biopsy oncoprofiling panel, all samples except one were taken from time point A. The exception was one sample taken at time point C. Additionally, the blood used as a normal was not always collected at the same time point as the plasma sample. While this, in theory, should not affect the analysis, since the blood should always look the same, in practice it could have an affect due to contamination of ctDNA in the normal blood.

5.2 Enzymatic methyl sequencing

First of all, it should be noted that no normal samples were used during this analysis. Further, comparison between the differential methylation of the samples and reference methylation data from healthy tissue has not been performed at present either. Thus, it is not possible to draw conclusions about the biological implications, or potential pathogenicity, of the differential methylation observed. The analysis done for this report, and the results presented, are therefore strictly a comparison between the patient groups and individual samples.

The QC reports provided suggests that the quality of the samples was good and the coverage of the CpG sites, as well as the targeted CpG sites, was as expected. It was noted when presenting the method that LCS110 and LCS116 were grouped as both "Progressive disease" and "Stable disease". This was due to the fact that the PR vs SD analysis was performed first, together with the analysis of the patient groups with different metastases. The "Partial response vs Progressive disease" (PR vs PD) analysis was requested later. At the time when the original analysis was requested, not all samples that were in the final study cohort had been added or had the necessary material available. For this reason "Stable disease" was originally chosen as the criteria, and the samples with progressive disease were added to that group as well. Therefore, the PR vs PD comparative analysis can be assumed to provide more robust results in comparison to the PR vs SD comparison, as the SD group contains more in-group variability.

Regarding the patient groups with different metastases it should be noted that new information regarding the metastases was found after the comparative analysis had been requested. To be able to draw concrete conclusions from the analysis, it is vital that the metastases analysed were the ones present, or not present, at the time point when the sample used for analysis was taken. It was, however, noted that the available information concerning the metastases was not entirely correct. This does produce a level of uncertainty in regard to the comparative analysis that should be taken into account.

For the analysis using MethylKit in RStudio it should first be noted that LCG77 is an outlier in both the PCA analysis and dendrogram in figure 4.1. This is consistent with the fact that LCG77 was indicated as an outlier in several instances during the visual analysis of the differential methylation. In the heatmaps in figure 4.2 and 4.3, LCG77 clusters differently and is not as clearly an outlier compared to the other samples. However, when analysing the methylation observed in the heatmaps, it can still be observed, especially in figure 4.2, which shows the top 13 most significantly methylated CpGs, that LCG77 has different methylation patterns compared to the other samples. Additionally, in the PCA and dendrogram plots, LCS110 and LCG68 appear to be slight outliers. These have not been indicated as obvious outliers during further analysis, but LCS110 does have distinct methylation patterns in figure 4.2. In this heatmap it can also be observed that LCG71 shows some individual methylation patterns, which was not necessarily suggested by the correlation analysis.

In figure 4.2, the top 13 most significantly differentially methylated CpG sites have been annotated by region and gene. Amongst the annotated genes, some have known associations to cancer while others appear to have no known links at all. Despite this, some clear clusterings can be observed in the heatmap, which warrants further research.

In figure 4.3, showing the differential methylation of all genes of interest, as determined by the visual analysis, it becomes clear that there are no major groupings to be

inferred. However, the hypermethylation of the genes *ZNF573* and *NME2* for sample LCG71 can be seen. Additionally, distinct methylation pattens can be observed for *MIR4520-2* and *MIR4458HG*. *PM20D1*, which showed individual methylation patterns during the visual analysis is also indicated here. However, *PCMTD2* show no clear differential methylation pattern for any sample, despite having been observed as differentially methylated for LCS112 during the visual analysis. This is true for *CPT1C* as well, which shows no clear patterns despite having been identified as differentially methylated for LCG77 and LCS116. It should be noted that *MIR4520-2* and, to a lesser extent, *MIR4458HG*, did show some individual differential methylation patterns during the visual analysis of patients with and without skeletal metastases. They were not included in the case studies as the patterns appeared ambiguous, but the results from this heatmap suggests that further analysis might be warranted.

The visual analysis in IGV produced a variety of genes of interest with varying biological contexts and associations to malignancies, as backed up by the heatmap showing the differential methylation of the genes in figure 4.3. It should be noted again, that ”hyper” and ”hypo” were used to denote an increase or decrease, respectively, of methylation in comparison to the other patient group, and does not necessarily imply a differential methylation pattern compared to cfDNA from healthy tissue. It was also observed that the gene *MCF2L* showed both some increased methylation and some decreased methylation at different positions. The majority of the positions that showed differential methylation did, however, show an increase in methylation, with only one position showing a decrease.

During the visual analysis, some interesting individual patterns were observed as well. It was noted that LCG71 has a large hypermethylated region over the gene *ZNF573* which was not observed in any other sample. ZNF573 is a zinc-finger protein, expected to enable DNA-binding transcription factor activity as well as RNA polymerase II-specific and cis-regulatory region sequence specific activity [50]. While this particular gene is not well known, other members in the *ZNF* gene family have been identified as both oncogenes and tumour suppressor genes [51]. Similar hypermethylation patterns were found for LCG71 over the *NME2* gene. This gene is associated with nucleotide metabolism as well as transcription and transcription regulation, and has been implicated as a prognostic marker in, amongst other things, lung adenocarcinoma [52]. Apart from having these differential methylation patterns, it should also be noted that this patient was very young, 36 years old, and originally responded to treatment, which suggest additional reason to study the samples further.

LCG77 was shown to be an outlier during analysis of genes belonging to the *S100A* family, showing methylation patterns more similar to responders than the other non-responders. This was particularly evident for the *S100A1*, *S100A14* and *S100A16* genes. The *S100A* family plays a role in several biological processes and these genes have also been implicated in tumour development, progression, differentiation and invasion [53][54], making these genes and methylation patterns interesting for fur-

ther analysis. Additionally, proteins from the S100 family, particularly, S100B, are used as biomarkers for malignant melanoma [55] showing additional ties between the gene family and malignancies.

Three other cases were identified where larger areas around specific genes showed particular differential methylation patterns for certain samples. These include, as previously mentioned, LCG77 and LCS116 over a region of the *CPT1C* gene, LCS112 over a region of the *PCMTD2* gene and somewhat individual methylation patterns over the gene *PM20D1*. The genes have varying degrees of evidence suggesting involvement in cancer but the analysis suggests these results should be analysed further. To note additionally is the fact that for *PM20D1* the groups created by the differential methylation patterns closely resemble the groups in the PR vs SD analysis, with LCG68 and LCG71 belonging to the "Partial response" group, and the other patients to the "Stable disease" group. Viewing the results this way, LCS116 would be the outlier, as it belongs to the group with progressive disease but here resembles the partial response samples. In fact, as can be seen in table 4.3, the gene was identified as a gene of interest during the PR vs SD comparative analysis, and here as well LCS116 has been noted as an outlier.

5.3 Oxford Nanopore™ sequencing

As mentioned in the result, several trial samples were run on the Oxford Nanopore™ long read sequencing device, apart from the samples from the study cohort. The reason trial samples were used is because, as the method was very new, it took time to get satisfactory output, and the samples are limited. Thus, cohort samples could not be "wasted" on runs that might not work. Further, it should be noted that for this method no normal samples were run, and thus the methylation data acquired can not currently be used to draw conclusions about the biological implications of the signatures observed. The differential methylation was mainly used to compare the method to enzymatic methyl sequencing.

While the original goal was to get a coverage of 30x, it was determined early on that that might not be possible for cfDNA samples due to the limited input, and a new goal of about 10x was set. However, as can be seen in figure 4.4, the coverage was very low. Discussions with and presentations from international research groups, does, however, suggest that a coverage under 1x might still be enough for accurate calling of differential methylation. This suggests that with further analysis, it might be possible to get accurate analytical information from these samples.

In table 4.5 a summary of the SNVs and indels found using the two different EPI2ME pipelines are presented. It should be noted, that these are unfiltered data that have been taken directly from the calling reports generated by the pipeline. Therefore, no classification analysis has been done to determine pathogenicity of the variants in question. Additionally, as mentioned briefly in the results, the wf-human variation pipeline generate significantly more variants than the wf-somatic variation pipeline.

This is caused by the different caller models used. The wf-human variation pipeline uses the Clair3 caller which is a germline variant caller, while the wf-somatic variation pipeline uses the ClairS model which focuses on somatic variants. Originally, wf-human variation was used for analysis as this pipeline did not require a normal sample, in contrast to wf-somatic variation. However, later it was found that the somatic variation pipeline could be used without a normal sample as well, but with limited calling output as a results, and the analysis was therefore done on that pipeline instead, as it was better suited for this thesis' purpose.

The comparative analysis of differential methylation patterns done between Oxford Nanopore™ sequencing and Hologic® Diagenode's Human methylation panel, suggests that the methylation patterns observed are accurate and can be visualised using both long and short read sequencing. While the low coverage for the long read sequencing samples causes definite analysis to be difficult, in general the same methylation patterns could be observed in the output from both technologies. This suggest that the modified basecalling of the two different methods is comparable.

5.3.1 Troubleshooting

While attempting to set up this method some issues were encountered. For the first few trial samples, performed using singleplex sequencing the flow cells appeared underloaded, resulting in poor sequencing output and low coverage. In an attempt to fix this, the amount of input cfDNA was increased and all available material from the sample was loaded onto the flow cell. Despite this, poor pore occupancy, and thus low outputs, as well as poor output quality, remained an issue. Quality control of all samples was then introduced including analysis using Agilent 4200 TapeStation system (Agilent Technologies, Inc. US), and spectrophotometric analysis using Lunatic (Unchained Labs, Inc. US) or DeNovix®DS-11 Spectrophotometer (DeNovis Inc., US).

The quality control, as mentioned in the results, suggested that some samples might contain contamination, possibly from genomic DNA (gDNA), RNA fragments or some chemicals. This was investigated thoroughly in order to find what might be causing the issues. The possible contamination shown on the TapeStation system (Agilent Technologies, Inc. US) could be explained by either gDNA contamination, or contamination of longer fragments of cfDNA. The abnormal ratios shown by the Lunatic (Unchained Labs, Inc. US) and the DeNovix® DS-11 Spectrophotometer (DeNovis Inc., US) proved more difficult to explain. It was suggested that the method of extraction of cfDNA might be causing gDNA or RNA contamination. At the time, an automatic bead based extraction technology was used. Therefore, the extraction method was switched to the recommended manual method, but this did not improve the QC results. Additional purification steps were also added to the protocol to ensure that contamination of RNA was not causing these issues, but the ratios did not improve. The additional purification steps also presented the issue of further material loss. While most material was generally retained after purification, some loss is unavoidable. It is also important to note, that the expected ratio scores,

as per the library preparation protocols (SQK-LSK114 and SQK-NBD114.24), are based on gDNA and thus it is possible that the spectrophotometric ratios for cfDNA should be different. However, no reference data for cfDNA appears to be available and other research groups appear to not have performed equivalent spectrophotometric QC checks. Thus, the cause of the abnormal ratios remains uncertain.

To avoid underloading of the flow cell, the decision was made to run multiplex experiments instead. This did fix the issue with pore occupancy caused by the previously underloaded flow cells, however, the issues with quality still remained. One indication of this, was that basecalling of the samples did not reach 100% and several failed bases and reads could be observed. One suggestion for the reason for this was issues with computing power. The PromethIOn 2 Solo device does not have an integrated computer, but rather is connected to an external device. While the computer used was strong, it was suggested that this could be causing an issue. Another suggested issue was the tubes used for blood collection. While Streck™ tubes should be the most suitable tubes for the purpose of extracting cfDNA, other research groups had used EDTA tubes instead. A research group with access to a PromethION 2 Integrated device, essentially the same device but with an integrated computer instead, was contacted and three multiplex experiments were run on this device instead, one of which contained trial samples from EDTA tubes. The basecalling worked better, but still some quality issues remained. It was noted that the EDTA tube samples did not perform better than the Streck™ tube samples.

At this point it was noted by the Oxford Nanopore Technologies support, that a large amount of data and good quality reads was, in fact, acquired from the sequencing, but the bases showed poor quality. Additionally, the number of passed and failed bases did not add up to the given number of total bases. By looking at the underlying reports from the device it was also noted that the quality of the samples appeared satisfactory, and therefore should generate higher coverage than they do. However, the question remains as to whether this is an issue of some failed bases dispersed over the entire genome, resulting in passed reads despite a high number of failed bases, or if there are concentrated regions of failed bases which also cause failed reads. While results from the sequencing have been acquired that shows some analytical promise, as least in terms of differential methylation detection, more work needs to be done.

5.4 Variant classification

The QC results from Eurofins Genomics INVIEW liquid biopsy oncoprofiling panel shows passable results. In the original QC report, warnings were given for all samples concerning the ”On-target rate (%)”, the percentage of reads mapped to the target region with a +/- 100 bp tolerance. Additionally, coverage warnings were shown for samples LCG75-B, LCS139-B and LCS140-B. However, despite this, the quality was deemed satisfactory.

For the variant classification, the analysis was not always straight forward as some ctDNA can contaminate the normal blood samples. Thus, the possibility of some

percentage of the variant being found in the blood, while still being true somatic, needed to be considered. Due to the lack of available FFPE tumour DNA, it also was not possible to definitely classify all variants for all samples as somatic, plasma specific or possible CHIP. It should also be noted that due to the average age of the patients in this cohort, far more CHIP variants are expected than was found. It is possible that not all CHIP variants were detected since an oncoprofiling panel was used, and CHIP variants were not specifically studied. Regardless, more research should be conducted to validate the results.

For the samples with known driver mutations, it was determined that the drivers could be found in the plasma sample during the analysis. This is positive as it suggests that a blood sample could be enough to identify driver mutations in a patient. Additionally, apart from the *ERBB2* amplification driver mutation found in LCG77, a somatic variant was also found in the *ERBB2* gene during classification. This could be expected, since the *ERBB2* gene has been amplified there is a statistically higher chance of a variant in the region, and it does not inherently imply any connection between the variant and the malignancy.

6
Conclusion

In this study a patient cohort consisting of 17 patients with metastatic NSCLC and limited response to treatment, was evaluated using three different short and long read sequencing technologies in order to try to determine methylation and mutational signatures that could be used as biomarkers to predict response to treatment. Mutational variants were found using short read sequencing and classified to determine possible somatic and plasma specific variants. Methylation analysis was done using both short read and long read sequencing. The results from the methylation analysis using short read sequencing suggests the presence of differential methylation patterns that should be analysed further to determine biological and clinical impact. The long read methylation analysis needs more time to perfect, but the results so far suggests that similar differential methylation signatures can be found, as were found in the short read sequencing. At present it is not possible to determine with certainty, the correlation between the differential methylation signatures or the somatic variants found, and response to treatment or the spread of metastases. However, with further analysis, correlations might be found. In conclusion, more time is needed to perfect the long read sequencing protocol and analyse all the results more in depth. When relevant, all results should also be verified using a larger cohort. However, the acquired results do show potential, and suggest it could be possible to find methylation and mutational signatures that could be used as biomarkers.

7

References

[1] F. Bray, M. Laversanne, H. Sung, *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, May 2024, ISSN: 0007-9235. DOI: 10.3322/caac.21834.

[2] A. Hallqvist, A. Rohlin, and S. Raghavan, "Immune checkpoint blockade and biomarkers of clinical response in non–small cell lung cancer," *Scandinavian Journal of Immunology*, vol. 92, no. 6, Dec. 2020, ISSN: 13653083. DOI: 10.1111/sji.12980.

[3] G. Rossi, A. Russo, M. Tagliamento, *et al.*, "Precision medicine for NSCLC in the era of immunotherapy: New biomarkers to select the most suitable treatment or the most suitable patient," *Cancers*, vol. 12, no. 5, May 2020, ISSN: 20726694. DOI: 10.3390/cancers12051125.

[4] T. Draškovič, N. Zidar, and N. Hauptman, "Circulating Tumor DNA Methylation Biomarkers for Characterization and Determination of the Cancer Origin in Malignant Liver Tumors," *Cancers*, vol. 15, no. 3, Feb. 2023, ISSN: 20726694. DOI: 10.3390/cancers15030859.

[5] K. Sun, P. Jiang, K. C. Chan, *et al.*, "Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 40, E5503–E5512, Oct. 2015, ISSN: 10916490. DOI: 10.1073/pnas.1508736112.

[6] T. Cascone, J. Fradette, M. Pradhan, and D. L. Gibbons, "Tumor Immunology and Immunotherapy of Non-Small-Cell Lung Cancer.," *Cold Spring Harbor perspectives in medicine*, vol. 12, no. 5, May 2022, ISSN: 2157-1422. DOI: 10.1101/cshperspect.a037895.

[7] L. Wang, Y. Hu, S. Wang, J. Shen, and X. Wang, "Biomarkers of immunotherapy in non-small cell lung cancer.," *Oncology letters*, vol. 20, no. 5, p. 139, Nov. 2020, ISSN: 1792-1074. DOI: 10.3892/ol.2020.11999.

[8] M. Dahri, N. Beheshtizadeh, N. Seyedpour, *et al.*, "Biomaterial-based delivery platforms for transdermal immunotherapy," *Biomedicine & Pharmacotherapy*, vol. 165, p. 115 048, Sep. 2023, ISSN: 07533322. DOI: 10.1016/j.biopha.2023.115048.

[9] R. Stanley, S. Flanagan, D. O. Reilly, E. Kearney, J. Naidoo, and C. M. Dowling, "Immunotherapy through the Lens of Non-Small Cell Lung Cancer," *Cancers*, vol. 15, no. 11, p. 2996, May 2023, ISSN: 2072-6694. DOI: 10.3390/cancers15112996.

[10] E. Eisenhauer, P. Therasse, J. Bogaerts, *et al.*, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *European Journal of Cancer*, vol. 45, no. 2, pp. 228–247, Jan. 2009, ISSN: 09598049. DOI: 10.1016/j.ejca.2008.10.026.

[11] M. P. Leers, "Circulating tumor DNA and their added value in molecular oncology," *Clinical Chemistry and Laboratory Medicine*, vol. 58, no. 2, pp. 152–161, Feb. 2020, ISSN: 14374331. DOI: 10.1515/cclm-2019-0436.

[12] B. Tomasik, M. Skrzypski, M. Bieńkowski, R. Dziadziuszko, and J. Jassem, "Current and future applications of liquid biopsy in non-small-cell lung cancer-A narrative review," *Translational Lung Cancer Research*, vol. 12, no. 3, pp. 594–614, 2023, ISSN: 22264477. DOI: 10.21037/tlcr-22-742.

[13] S. Sumbal, A. Javed, B. Afroze, *et al.*, "Circulating tumor DNA in blood: Future genomic biomarkers for cancer detection," *Experimental Hematology*, vol. 65, pp. 17–28, Sep. 2018, ISSN: 18732399. DOI: 10.1016/j.exphem.2018.06.003.

[14] A. Kustanovich, R. Schwartz, T. Peretz, and A. Grinshpun, "Life and death of circulating cell-free DNA," *Cancer Biology & Therapy*, vol. 20, no. 8, pp. 1057–1067, Aug. 2019, ISSN: 1538-4047. DOI: 10.1080/15384047.2019.1598759.

[15] K. Strimbu and J. A. Tavel, "What are biomarkers?" *Current Opinion in HIV and AIDS*, vol. 5, no. 6, pp. 463–466, Nov. 2010, ISSN: 1746-630X. DOI: 10.1097/COH.0b013e32833ed177. [Online]. Available: http://journals.lww.com/01222929-201011000-00003.

[16] A. T. Madsen, J. A. Hojbjerg, B. S. Sorensen, and A. Winther-Larsen, "Day-to-day and within-day biological variation of cell-free DNA," *eBioMedicine*, vol. 49, pp. 284–290, Nov. 2019, ISSN: 23523964. DOI: 10.1016/j.ebiom.2019.10.008.

[17] E. Sanz-Garcia, E. Zhao, S. V. Bratman, and L. L. Siu, "H E A L T H A N D M E D I C I N E Monitoring and adapting cancer treatment using circulating tumor DNA kinetics: Current research, opportunities, and challenges," *Sci. Adv*, vol. 8, no. 4, p. 8618, Jan. 2022. DOI: DOI:10.1126/sciadv.abi8618.

[18] D. P. Clark and N. J. Pazdernik, "Basics of Biotechnology," in *Biotechnology*, Second Edition, Elsevier Inc., 2016, ch. 1, p. 12.

[19] D. P. Clark and N. J. Pazdernik, "Cancer," in *Biotechnology*, Second Edition, Elsevier Inc., 2016, ch. 19, pp. 594–595.

[20] B. Alberts, A. Johnson, J. Lewis, *et al.*, "Analyzing Cells, Molecules, and Systems," in *Molecular Biology of the cell*, S. G. Lewis and E. Zayatz, Eds., Sixth Edition, New York: Garland Science, 2015, ch. 8, pp. 492–494.

[21] B. Alberts, A. Johnson, J. Lewis, *et al.*, "Control of Gene Expression," in *Molecular Biology of the cell*, S. G. Lewis and E. Zayatz, Eds., Sixth Edition, New York: Garland Science, 2015, ch. 7, pp. 404–411.

[22] X. Wang, Q. Dong, G. Chen, J. Zhang, Y. Liu, and Y. Cai, "Frameshift and wild-type proteins are often highly similar because the genetic code and genomes were optimized for frameshift tolerance," *BMC Genomics*, vol. 23, no. 1, p. 416, Dec. 2022, ISSN: 1471-2164. DOI: 10.1186/s12864-022-08435-6.

[23] D. P. Clark and N. J. Pazdernik, "Genomics and Gene Expression," in *Biotechnology*, Second Edition, London: Elsevier Inc., 2016, ch. 8, pp. 266–267.

[24] M. Mahmoud, N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, and F. J. Sedlazeck, "Structural variant calling: the long and the short of it," *Genome Biology*, vol. 20, no. 1, p. 246, Dec. 2019, ISSN: 1474-760X. DOI: 10.1186/s13059-019-1828-7.

[25] B. Alberts, A. Johnson, J. Lewis, *et al.*, "Panel 8-2: Review of Classical Genetics," in *Molecular Biology of the Cell*, S. G. Lewis and E. Zayatz, Eds., Sixth Edition, New York: Garland Science, 2015, ch. 8, p. 487.

[26] B. Alberts, A. Johnson, J. Lewis, *et al.*, "Cancer-Critical Genes: How they are found and how they work," in *Molecular Biology of the Cell*, S. G. Lewis and E. Zayatz, Eds., Sixth Edition, New York: Garland Science, 2015, ch. 20, pp. 1104–1105.

[27] B. Alberts, A. Johnson, J. Lewis, *et al.*, "Cancer-Critical Genes: How they are found and what they do," in *Molecular Biology of the Cell*, S. G. Lewis and E. Zayatz, Eds., Sixth Edition, New York: Garland Science, 2015, ch. 20, pp. 1111–1112.

[28] D. P. Clark and N. J. Pazdernik, "Forensic Molecular Biology," in *Biotechnology*, Second Edition, Elsevier Inc., 2016, ch. 23, pp. 729–731.

[29] M. Jagannathan-Bogdan and L. I. Zon, "Hematopoiesis.," *Development (Cambridge, England)*, vol. 140, no. 12, pp. 2463–7, Jun. 2013, ISSN: 1477-9129. DOI: 10.1242/dev.083147.

[30] C. S. Marnell, A. Bick, and P. Natarajan, "Clonal hematopoiesis of indeterminate potential (CHIP): Linking somatic mutations, hematopoiesis, chronic inflammation and cardiovascular disease.," *Journal of molecular and cellular cardiology*, vol. 161, pp. 98–105, Dec. 2021, ISSN: 1095-8584. DOI: 10.1016/j.yjmcc.2021.07.004.

[31] illumina, *Introduction to NGS*. [Online]. Available: https://www.illumina.com/science/technology/next-generation-sequencing.html.

[32] Illumina Inc., *Intro to SBS technology*, 2017. [Online]. Available: https://www.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-workflow.html.

[33] Oxford Nanopore Technologies, *How nanopore sequencing works*. [Online]. Available: https://nanoporetech.com/platform/technology.

[34] Oxford Nanopore Technologies, *Epigenetics and methylation analysis*. [Online]. Available: https://nanoporetech.com/applications/investigations/epigenetics-and-methylation-analysis.

[35] Y. Wang, Y. Zhao, A. Bollas, Y. Wang, and K. F. Au, "Nanopore sequencing technology, bioinformatics and applications," *Nature Biotechnology*, vol. 39, no. 11, pp. 1348–1365, Nov. 2021, ISSN: 1087-0156. DOI: 10.1038/s41587-021-01108-x.

[36] Y. Liu, W. Rosikiewicz, Z. Pan, *et al.*, "DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation," *Genome Biology*, vol. 22, no. 1, p. 295, Dec. 2021, ISSN: 1474-760X. DOI: 10.1186/s13059-021-02510-z.

[37] B. Alberts, A. Johnson, J. Lewis, *et al.*, "DNA, Chromosomes, and Genomes," in *Molecular Biology of the cell*, S. G. Lewis and E. Zayatz, Eds., Sixth Edition, New York: Garland Science, 2015, ch. 4, p. 194.

[38] B. Jin, Y. Li, and K. D. Robertson, "DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?" *Genes & Cancer*, vol. 2, no. 6, pp. 607–617, Jun. 2011, ISSN: 1947-6019. DOI: 10.1177/1947601910393957.

[39] D. P. Clark and N. J. Pazdernik, "DNA, RNA and Protein," in *Biotechnology*, Second Edition, London: Elsevier Inc., 2016, ch. 2, pp. 45–49.

[40] J. Frigola, X. Solé, M. F. Paz, *et al.*, "Differential DNA hypermethylation and hypomethylation signatures in colorectal cancer," *Human Molecular Genetics*, vol. 14, no. 2, pp. 319–326, Jan. 2005, ISSN: 1460-2083. DOI: 10.1093/hmg/ddi028.

[41] M. Ehrlich, "DNA hypomethylation in cancer cells.," *Epigenomics*, vol. 1, no. 2, pp. 239–59, Dec. 2009, ISSN: 1750-192X. DOI: 10.2217/epi.09.33.

[42] Y. M. Lo, D. S. Han, P. Jiang, and R. W. Chiu, "Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies," *Science*, vol. 372, no. 6538, Apr. 2021, ISSN: 10959203. DOI: 10.1126/science.aaw3616.

[43] S. Guo, D. Diep, N. Plongthongkum, H. L. Fung, K. Zhang, and K. Zhang, "Identification of methylation haplotype blocks AIDS in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA," *Nature Genetics*, vol. 49, no. 4, pp. 635–642, Mar. 2017, ISSN: 15461718. DOI: 10.1038/ng.3805.

[44] M. Frommer, L. E. Mcdonald, D. S. Millar, *et al.*, "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands (genomic sequencing/DNA methylation/bisulfite modification/PCR/kininogen gene)," *Genetics*, vol. 89, pp. 1827–1831, Mar. 1992. DOI: https://doi.org/10.1073/pnas.89.5.1827. [Online]. Available: https://www.pnas.org.

[45] F. Krueger, B. Kreck, A. Franke, and S. R. Andrews, "DNA methylome analysis using short bisulfite sequencing data," *Nature Methods*, vol. 9, no. 2, pp. 145–151, Feb. 2012, ISSN: 1548-7091. DOI: 10.1038/nmeth.1828.

[46] R. Vaisvila, V. K. Ponnaluri, Z. Sun, *et al.*, "Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA," *Genome Research*, vol. 31, no. 7, pp. 1280–1289, Jul. 2021, ISSN: 15495469. DOI: 10.1101/gr.266551.120.

[47] J. Svensson, "Genetic profiling in non-small cell lung cancer," Gothenburg, 2023.

[48] Genomic Medicine Sweden, "Genomic Medicine Sweden GMS560: a broad targeted NGS gene panel for comprehensive genomic profiling of solid tumors," Tech. Rep., 2023. [Online]. Available: www.genomicmedicine.se.

[49] Hologic® Diagenode, *Human Methylome Service*. [Online]. Available: https://www.diagenode.com/en/p/human-methylome-service.

[50] National Library of Medicine, *ZNF573 zinc finger protein 573 [Homo sapiens (human)]*, May 2025.

[51] J. Jen and Y.-C. Wang, "Zinc finger proteins in cancer progression," *Journal of Biomedical Science*, vol. 23, no. 1, p. 53, Dec. 2016, ISSN: 1423-0127. DOI: 10.1186/s12929-016-0269-9.

[52] The Human Protein Atlas, *NME2*. [Online]. Available: https://www.proteinatlas.org/ENSG00000243678-NME2.

[53] M. Zhu, H. Wang, J. Cui, *et al.*, "Calcium-binding protein S100A14 induces differentiation and suppresses metastasis in gastric cancer," *Cell Death & Disease*, vol. 8, no. 7, e2938–e2938, Jul. 2017, ISSN: 2041-4889. DOI: 10.1038/cddis.2017.297.

[54] Y. Zhang, X. Yang, X.-L. Zhu, *et al.*, "S100A gene family: immune-related prognostic biomarkers and therapeutic targets for low-grade glioma," *Aging*, vol. 13, no. 11, pp. 15 459–15 478, Jun. 2021, ISSN: 1945-4589. DOI: 10.18632/aging.203103.

[55] R. Harpio and R. Einarsson, "S100 proteins as cancer biomarkers with focus on S100B in malignant melanoma," *Clinical Biochemistry*, vol. 37, no. 7, pp. 512–518, Jul. 2004, ISSN: 00099120. DOI: 10.1016/j.clinbiochem.2004.05.012.

A

Filtering of Hologic® Diagenode data

```
library(dplyr)

data <- read.csv("Z:/Place/name_of_file.csv")

data_no_duplicates <- unique(data)

data_no_duplicates2 <- data[!duplicated(data), ]

data_no_na <- na.omit(data)

filtered_data <- data %>%
  distinct(start, end, .keep_all = TRUE)

filtered_data_add <- filtered_data[filtered_data$meth.diff <
    50 & filtered_data$meth.diff > -50 & filtered_data$
   qvalue < 1e-10, ]

write.csv(filtered_data_add, "C:/Place/filtered_data_name.
   csv", row.names = FALSE)
```

B

Differential Methylation analysis using MethylKit in RStudio

```
library("pheatmap")

library("methylKit")

library("DMRcate")

library("GenomicRanges")

library("dendextend")

file.list=list("LCG017_CpG.methylKit.gz",
               "LCG024_CpG.methylKit.gz",
               "LCG068_CpG.methylKit.gz",
               "LCG071_CpG.methylKit.gz",
               "LCG077_CpG.methylKit.gz",
               "LCS110_CpG.methylKit.gz",
               "LCS112_CpG.methylKit.gz",
               "LCS116_CpG.methylKit.gz")

myobj=methRead(file.list,
               sample.id=list("LCG017", "lCG024", "LCG068",
                  "LCG071", "LCG077", "LCS110", "LCS112", "
                  LCS116"),
               assembly="hg38",
               treatment=c(1,2,3,4,5,6,7,8),
               context="CpG",
               mincov = 10
)

#Histogram of GpG Methylation of sample
getMethylationStats(myobj[[1]],plot=FALSE,both.strands=FALSE
    )
```

```
#Histogram of CpG coverage
getCoverageStats(myobj[[1]], plot=FALSE, both.strands=FALSE)

#Sample filtered based on coverage, remove less than 10x and
    more than 99.9th percentile of coverage
filtered.myobj.c=filterByCoverage(myobj, lo.count=10, lo.perc=
   NULL, hi.count=NULL, hi.perc=99.9)

normalized.myobj <- normalizeCoverage(filtered.myobj.c)

getMethylationStats(normalized.myobj[[1]], plot=FALSE, both.
   strands=FALSE)

#merge the data
meth <- unite(normalized.myobj, destrand=FALSE) #merges all
   samples on common CpG sites

#--- Clusterings

hc <- clusterSamples(meth, dist="correlation", method="
   complete", plot=FALSE)

dend <- as.dendrogram(hc)
sample_colors <- c("red", "#990033", "#3399FF", "#003399", "
   green", "#336600", "orange","#CC6600")

dend <- dend %>% set("labels_colors", sample_colors)

dist.method <- "correlation"

clust_method <- "complete"

png("C:/Place/name.png", width = 1200, height = 800, res =
   150)
plot(dend, main=paste("CpG methylation Clustering, method:",
    clust_method, ", distance:", dist.method), ylab = "
   Height")
dev.off()

PCASamples(meth)

#------ Genes of interest

# Create GRanges for a few genes of interest
genes_of_interest <- GRanges(seqnames = c("chr1", "chr1", "
   chr1", "chr1", "chr1", "chr1", "chr1", "chr1", "chr1", "
```

```
    chr1", "chr1", "chr2", "chr2", "chr2", "chr3", "chr3", "
    chr3", "chr4", "chr5", "chr5", "chr5", "chr7", "chr7", "
    chr7", "chr7", "chr7", "chr8", "chr8", "chr8", "chr9", "
    chr9", "chr9", "chr9", "chr10", "chr11", "chr11", "chr13"
    , "chr13", "chr13", "chr13", "chr15", "chr16", "chr17", "
    chr17", "chr17", "chr17", "chr17", "chr17", "chr19", "
    chr19", "chr19", "chr20", "chr20", "chr21", "chr22"),
ranges = IRanges(start = c(11066618, 110480752, 115229326,
    153606886, 153614255, 153627926, 223701593, 227728200,
    22710839, 205828025, 214281102, 231453531, 138669157,
    231056845, 13549125, 48403854, 184314495, 1712858,
    114055926, 76403285, 8450701, 98214624, 131110096,
    143390289, 151085831, 726699, 17643795, 41529218,
    2100000, 34957608, 97699625, 136546162, 135932969,
    44959407, 236966, 5596109, 27977717, 112894378,
    114107569, 112894378, 67063763, 56989485, 50508425,
    51165435, 8390702, 75587800,  6655449, 80260852,
    37735833, 2427638, 49690898, 25451594, 64255695,
    39377698, 43999211),
     end = c(11099869, 110491277, 115368072, 153613145,
        153616986, 153632039, 223776018, 227781826, 22921500,
         205850132, 214337131, 231483641, 138780390,
        231173116, 13638422, 48430086, 184335358, 1745171,
        114496500, 76708132, 8486930, 98252232, 131496632,
        143408856, 151144436, 786475, 17801094, 41545030,
        2200000, 34982544, 97853116, 136552541, 135961373,
        44995891, 252984, 5612958, 27988693, 113099742,
        114108820, 113099742, 67195173, 57083531, 50531501,
        51171744, 8397827, 75626849, 6655502, 80398794,
        37817300, 2456959, 49713731, 25585531, 64304820,
        39428528, 44172939)),
strand = c("-", "+", "+","-", "-", "+", "+", "+", "+", "-",
    "+", "-", "-", "+", "+", "-", "+", "+", "+", "+", "+", "-
    ", "+", "-", "+", "+", "-", "+", "+", "+", "-", "+", "-",
     "+", "+", "+", "-", "+", "-", "+", "+", "+", "+", "+", "
    -", "+", "+", "+", "-", "-", "+", "-", "+", "+", "+"),
gene = c("exosc10", "CYMP", "NGF-AS1", "S100A16", "S100A14",
     "S100A1", "CAPN2", "SNAP47", "EPHB2", "PM20D1", "SMYD2",
     "NCL", "NXPH2", "PSMD1", "FBLN2", "PLXNB1", "EIF4G1", "
    TACC3", "KCNN2", "IQGAP2", "MIR4458HG", "TECPR1", "MKLN1"
    , "EPHA1", "AGAP3", "DNAAF5", "MTUS1", "GINS4", "MYOM2",
    "PHF24", "PTCSC2", "NALT1", "UBAC1", "RASSF4", "PSMD13",
    "TRIM6", "URAD", "MCF2L", "RASA3-IT1", "MCF2L", "SMAD3",
    "NLRC5", "MYCBPAP", "NME2", "RNF222", "MYO15B", "MIR4520
    -2", "RNF213", "ZNF573", "LMNB2", "CPT1C", "NINL", "
    PCMTD2", "GET1", "PARVB")
```

```
)

# Convert methylBase to GRanges, done to fix error in
    following step as meth and Granges doesn't work together
meth.gr <- as(meth, "GRanges")

# Get indices of overlapping CpGs
hits <- findOverlaps(meth.gr, genes_of_interest, ignore.
    strand = FALSE)

# Get indices and corresponding gene names
cpg.idx <- queryHits(hits)
gene.names <- genes_of_interest$gene[subjectHits(hits)]

# Subset original methylBase to matching CpGs
overlap.meth <- meth[cpg.idx, ]
meth.mat <- percMethylation(overlap.meth)

# Add gene names to methylation matrix
meth.df <- data.frame(Gene = gene.names, meth.mat)

# Remove NAs if any
#meth.df <- meth.df[!is.na(meth.df$Gene), ]

# Aggregate: average methylation across CpGs for each gene
meth.by.gene <- aggregate(. ~ Gene, data = meth.df, FUN =
    mean)

# Set gene names as rownames and clean up
rownames(meth.by.gene) <- meth.by.gene$Gene
meth.by.gene$Gene <- NULL
colnames(meth.by.gene) <- c("LCG017", "LCG024", "LCG068", "
    LCG071", "LCG077", "LCS110", "LCS112", "LCS116")

#Heatmap
pheatmap(meth.by.gene,
          cluster_rows = TRUE,
          cluster_cols = TRUE,
          show_rownames = TRUE,
          show_colnames = TRUE,
          color = colorRampPalette(c("blue", "white", "red"))
              (50),
          main = "Methylation␣Heatmap␣of␣Genes␣of␣Interest",
          fontsize_row = 7,
          #cellwidth = 5,
          #cellheight = 8
```

```
        )

# ———— Decide DM and then heatmap

# Logistic regression test for DMPs
dmps <- calculateDiffMeth(meth, overdispersion = "MN", test
    = "Chisq")

# Get significant DMP
sigDMPs <- getMethylDiff(dmps, difference = 25, qvalue =
    0.01)

# Top 100 (or other number) significant positions
top <- sigDMPs[order(sigDMPs$qvalue), ]
top <- top[1:min(500, nrow(top)), ]  # avoid errors if <100
    DMPs


# Extract percent methylation for top CpGs
overlap.meth <- meth[cpg.idx, ]
methMatrix <- percMethylation(overlap.meth)
rownames(methMatrix) <- paste(overlap.meth$chr, overlap.meth
    $start, sep = ".")
sigDMPs$chrBase <- paste(sigDMPs$chr, sigDMPs$start, sep = "
    .")
topMethMatrix <- methMatrix[rownames(methMatrix) %in%
    sigDMPs$chrBase, ]

#Gene annotation
library(GenomicFeatures)
library(TxDb.Hsapiens.UCSC.hg38.knownGene)
library(genomation)
library(ChIPseeker)
library(AnnotationDbi)
library(org.Hs.eg.db)

# convert sigDMPs to a GRanges object
sigGR <- GRanges(seqnames = sigDMPs$chr,
                 ranges = IRanges(start = sigDMPs$start, end
                     = sigDMPs$end))

# Load gene annotation (TxDb from Ensembl or UCSC)
#txdb <- TxDb.Hsapiens.UCSC.hg38.knownGene

peakAnno <- annotatePeak(sigGR, TxDb = TxDb.Hsapiens.UCSC.
    hg38.knownGene)
```

```
# Convert peakAnno to a data.frame
peakAnno_df <- as.data.frame(peakAnno)

#Used to view headers
head(peakAnno_df)

# Fix for heatmap

# Create a dataframe with the annotations we want (e.g.,
    Gene, Feature)
annotation_row <- data.frame(
  Gene = peakAnno_df$geneId,
  Feature = peakAnno_df$Promoter
)

rownames(annotation_row) <- peakAnno_df$chrBase  # Ensure
    rownames match

gene_symbols <- mapIds(org.Hs.eg.db,
                       keys = peakAnno_df$geneId,    # Or
                           peakAnno_df$geneID
                       column = "SYMBOL",
                       keytype = "ENTREZID",
                       multiVals = "first")


# Create chr.start identifiers for rownames
peakAnno_df$chrBase <- paste0(peakAnno_df$seqnames, ".",
    peakAnno_df$start)

annotation_row <- data.frame(
  Gene = peakAnno_df$geneSymbol,
  Region = peakAnno_df$annotation
)

rownames(annotation_row) <- peakAnno_df$chrBase

# Subset to just the CpGs in the heatmap
annotation_row <- annotation_row[rownames(topMethMatrix), ]

#Heatmap 1, no annotation
pheatmap(topMethMatrix,
         scale = "row",
         clustering_distance_rows = "euclidean",
         clustering_distance_cols = "euclidean",
```

```
            clustering_method = "complete",
            show_rownames = FALSE,
            main = "Top 13 Differentially Methylated CpGs")

#Heatmap 2, annotation

main_title <- "Differentially Methylated CpGs\nwith Gene
    Annotations"
print(
pheatmap(topMethMatrix,  # Assuming this is your methylation
     matrix
            scale = "row",  # Scale rows (CpGs)
            show_rownames = FALSE,  # Hide CpG names (optional)
            annotation_row = annotation_row,  # Add row
                annotations
            color = colorRampPalette(c("blue", "white", "red"))
                (50),
            main = main_title,
            fontsize_main = 4)
)

# Troubleshooting --

while (!is.null(dev.list())) dev.off()
testMatrix <- matrix(rnorm(100), nrow = 10)
rownames(testMatrix) <- paste0("gene", 1:10)
colnames(testMatrix) <- paste0("sample", 1:10)

pheatmap(testMatrix)

pheatmap(topMethMatrix)
```

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

CHALMERS
UNIVERSITY OF TECHNOLOGY