University of Otago

Transitioning from Analysing Methylation Profiles in Bulk Populations of Colorectal Cancer Cells to Methylation Profiles of Single Cells

Hannah O’Neill

Supervised by Associate Professor Aniruddha Chatterjee and Dr. Euan Rodger

A thesis submitted in the partial fulfilment of the Degree of Bachelor of Biomedical Science with Honours.
University of Otago, Dunedin, New Zealand.

October 2021

Abstract

Colorectal cancer (CRC) is one of the leading causes of death by cancer in New Zealand, largely due to metastasis of the primary tumour to secondary sites around the body. The initiation, progression and eventual metastasis has been shown to be largely influenced by epimutations. DNA methylation is a stable, heritable epigenetic mark which has been heavily implicated in disease. In cancer contexts, global hypomethylation and focal hypermethylation act to grossly dysregulate the genome, while simultaneously acting as potential biomarkers for initial detection and therapeutic response.

CRC is a molecularly heterogeneous disease, including between patients, tumours and even within the tumours. As well as this, tumours are comprised of immune cells, healthy tissue cells and blood cells in tandem with the neoplastic cells. Traditionally, these populations are sequenced as a whole. Consequently, the methylomes of each cell are coalesced, giving rise to averaged methylation profiles. The emergence of single cell technologies allows us deconvolute heterogeneous populations of cells and identify different cell states and types.

Obtaining high quality data from a single cell is difficult due to the minimal amount of starting DNA, particularly in single-cell bisulfite sequencing as the bisulfite conversion is harsh on the DNA. Hence, the main aim of this project was to optimise and implement single-cell bisulfite sequencing on a sorted human colorectal cancer cell line. Following this, my second aim was to compare the methylation profiles of the single-cell methylation libraries to uncover heterogeneity in the population. In this project, I use a post-bisulfite adaptor tagging (PBAT) method to perform single cell bisulfite sequencing on the CRC human cell line HT29.

Following this, I use a publicly available data set to investigate intra- and inter- tumour heterogeneity.

With a few optimization steps the PBAT method was able to successfully amplify very small numbers of cells, including 100, 10, 5 and two single cell samples. Following this, publicly available data showed even in a small population of single cells, there was evident heterogeneity regarding global, chromosomal and focal promoter methylation.

These results highlighted the heterogeneity which can be unmasked using single cell technologies, even on a small scale. While also confirming renowned biological models such as the global hypomethylation undertaken by cancer cells.

Acknowledgements

Firstly, I would like to thank my supervisors Aniruddha and Euan for their constant support throughout the year. You have both made my first year in research a very welcoming environment. I am so grateful I had intelligent supervisors with the personalities to match. Euan, I knew very little coming into research and I greatly appreciate you explaining both complex and very simple concepts to me. Aniruddha, your passion for research is inspiring. I appreciate that you’re a busy man yet you always make time for your students.

I would also like to thank the Chatterjee lab group, in particular Nick, Sai, Rakesh, Reema and Morgan. Thank you for taking the time out of your own busy lives to help me with my project when I had no idea what to do. The game nights, endless laughs and even lab zoom drinks over lockdown have been highlights of my year. I am so lucky to have joined such an amazing lab group.

To my friends and flatmates, thank you for always believing in me and making this one of my favourite years at university. To Finn and Cassie, thank you for all the good times over the past four years.

To Dad, thank you for your endless support right through all of my schooling. You are my biggest cheerleader. To Lynne, my second mum, thank you for being such a huge role model in my life.

To Mum, my best friend, thank you for calming me down during my many breakdowns. To Steve, my second dad, I will never forget the support you have given me throughout my university experience.

Thank you to Michelle in the Microbiology Department, thank you for taking the time to help with flow cytometry multiple times throughout my project.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 1

Chapter 2

Chapter 3

List of Abbreviations

^{o}C – degrees Celsius

μL – microlitre

ATCC – America Type Culture Collection

bp – base pair

C – cytosine

CAR-T – chimeric antigen receptor T-cell

CD276 – Cluster of Differentiation 276

CDH1 – cadherin 1

CpG – cytosine-guanine dinucleotide

CRC – colorectal cancer

CSC – cancer stem cell

CTC – circulating tumour cell

DMEM – Dulbecco's Modified Eagle Medium

DMR – differentially methylated region

DNA – deoxyribonucleic acid

DNAme – DNA methylation

EBF3 – early B-cell factor 3

EMT – epithelia – mesenchymal transition

FACS – fluorescence activated cell sorting

FBS – foetal bovine serum

FVS450 – fixable viability stain 450

IGF2 – insulin-like growth factor 2

ITH – intra-tumour heterogeneity

kb – kilobase

LCM – laser-capture microdissection

LINE1 – long-interspersed nuclear element-1

LN – lymph node metastases

LOI – loss of imprinting

MELISA – Methylation inference for single cell analysis

ML – liver metastases

mL – millilitre

MLH1 – MutL Homolog 1

MMR – mismatch repair

MP – liver metastases post-treatment

MSI – microsatellite instability

NC – normal adjacent colon tissue

ng – nanogram

nm – nanometre

PBAT – post-bisulfite adaptor tagging

PBS – phosphate buffered saline

PCR – polymerase chain reaction

PS – Penicillin-Streptomycin

PT – primary tumour

RNA – ribonucleic acid

scBS-seq – single cell bisulfite sequencing

scMSRE - single cell methylation sensitive restriction enzyme sequencing

scRNA-seq – single cell RNA sequencing

scRRBS – single cell reduced representation bisulfite sequencing

SCS – single cell sequencing

Chapter 1: Introduction

Many of our current biological models are based on the bulk sequencing of heterogeneous cell populations, yet they may not hold true when scrutinised at a single cell level. This project aimed to carry out a single cell bisulfite sequencing (scBS-seq) method on colorectal cancer (CRC) cells to elucidate the heterogeneity surrounding the disease. The following sections of this chapter introduce key concepts in understanding single cell bisulfite sequencing, including techniques and limitations associated with it. Moreover, it details the implications of this new technology on studying the molecularly heterogeneous disease, colorectal cancer.

1.1 DNA Methylation

Epigenetics; once described as “the most obvious source of dark matter” [4] in the cancer genome – now an exceedingly researched area providing answers to long-asked questions. Epigenetics is described as “the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence.” [5]. One of the main epigenetic alterations is DNA methylation (DNAme), a stable yet reversible mark entailing the addition of a methyl group typically at the 5th position of a cytosine residue adjacent to a guanine residue (CpG site), or sometimes at non-CpG sites (CpH, H = A, C or T). DNAme is involved in many fundamental biological processes throughout the entire life span, including cell-cycle control, cell-fate decisions, X chromosome inactivation [6], genomic imprinting [7], embryonic development, chromosomal stability and transposable element silencing [8]. Many - if not all - of these processes when aberrantly regulated have been implicated in a number of diseases, including heart disease [9], autoimmune diseases [10] and diabetes [11]. Of particular interest to my project, is its role in CRC [12]. Epigenetics is an area of extensive research particularly in development and disease, as a result of its influence on gene expression as well as susceptibility to external factors.

1.1.1 DNA Methylation Dogma

The most widely accepted dogma for DNAme in gene regulation is that it acts to silence genes. It is thought that its presence either acts to recruit repressive transcription factors or physically block any activating transcription factors or RNA polymerase from binding to the promoter [13, 14]. However, when present in a gene body context rather than a gene promoter, it has been linked to transcriptional elongation and alternative splicing [15]. A number of studies have now emerged which display hypermethylation-induced transcriptional activation, not only in disease but in normal developmental processes [16]. This has been shown in melanoma, where the methylation of the EBF3 promoter is associated with an increase of gene expression [17, 18]. Another recent study also found hypermethylation in several genes in prostate cancer patients were associated with upregulation, suggesting its presence in some contexts may act to increase gene expression [19]. This highlights the diverse and not yet fully understood means in which DNAme acts to regulate gene expression. Not to mention, the diverse role DNAme also plays in maintaining structural integrity of the genome [20].

1.2 Single Cell DNA Methylation

Historically, due to technical limitations bulk populations of tissue samples have been analysed for methylome data. Bulk analysis refers to a large population of cells being coalesced followed by sequencing, resulting in an averaged methylome profile. The past few years have seen the emergence of single-cell technologies allowing for the analysis of epigenomes of single cells. A popular analogy when comparing bulk sequencing to single cell sequencing is the “fruit salad/smoothie” analogy. Bulk sequencing is synonymous to a fruit smoothie – in that all the different fruits are blended, and an averaged taste profile of all fruits is produced. Single cell sequencing is synonymous to a fruit salad – you eat each bit of fruit individually and only get that fruits taste profile, without being tainted by any other flavour. To refer this back to methylomes, bulk sequencing is combining a myriad of cells that contain differing methylomes, resulting in averaged data. Single cell sequencing will only provide the methylome of that single cell, allowing for the characterization of rare sub-populations of cells and provides insight into the genetic regulation of heterogeneous populations of cells.

The ability to deconvolute heterogeneous populations of cells and identify different cell states and types is engaging particularly in cancer contexts.

1.2.1 Single Cell Sequencing Techniques

The recent ability to study cells at the singular level has come from new technology enabling the isolation of single cells from a population. This, along with improved amplification steps and growing bioinformatic tools.

The first step to these analyses is the isolation of the single cells from their population. In the earlier stages, fluorescence-activated cell sorting (FACS), laser-capture microdissection (LCM) and micromanipulation were used to physically isolate single cells, usually only tens to hundreds at a time [21]. More recently, separating the cells into droplets via a unique barcode which has been tagged to specific sequence reads of a cell has allowed for the isolation of thousands of cells rather than hundreds [22]. Following this, they can be pooled back together and reactions such as bisulfite conversion can be done on a bulk population, dramatically reducing costs and increasing sensitivity [23]. Of interest to this project, FACS live-dead staining was used to sort cells. This entails using a FVS450 dye which stains cells with permeable membranes. In essence, non-viable/dead cells will have a high fluorescence as a result of higher dye uptake, whereas viable/live cells will take up less dye resulting in a lower fluorescence for sorting. This method was used for my project as I did not need to select for any certain cell type, as all cells were HT29 my selection was for live cells [21].

Once the cells have been isolated, several DNAme analyses can be employed dependant on the research (Table 1). The method used in my project; single cell bisulfite sequencing (scBS-seq) is a single base resolution technique. This is a gold-standard method for whole genome methylation analyses and studies wishing to get high-resolution data [24]. Despite this, it tends to be biased towards CpG rich genomic regions, have a relatively high cost and there is substantial DNA degradation during the bisulfite conversion process. This method was adapted to single cells to minimize adaptor-tagged fragment degradation through post-bisulfite adaptor tagging, in which the bisulfite conversion is performed before the adaptor tagging in library amplification rather than after as is done in bulk sequencing. Single cell methylation sensitive restriction enzyme sequencing (scMSRE) is another technique, in which an enzyme will cut at specific unmethylated cytosine residues, unable to cut 5mC residues. scMSRE is of greater time and cost efficiency than bisulfite methods yet do not have whole genome coverage. However, scMSRE results in less DNA degradation, making it suitable for site-specific or targeted studies. The last common method of note is single cell reduced representation bisulfite sequencing (scRRBS), a combination of enzyme restriction techniques and bisulfite techniques. scRRBS is more cost effective yet has less coverage than scBS-seq. scRRBS is also biased in regions with high CpG density and has low coverage in low CpG dense regions. These techniques have been summarized in table 1. Due to wanting to investigate the methylome of my cells, the most appropriate technique for my project was the scBS-seq technique described above.

Table 1. Single cell methylome sequencing techniques. Adapted from Kashima et al. 2020.

Method	Feature	Key Features	Year of first study	References
scRRBS	Reduced representation bisulfite sequencing	• Bias in regions with high CpG density, limited coverage in regions with low CpG density. • Cost effective.	2013	[25, 26]
scBS-seq	Bisulfite sequencing	• Single base resolution, sensitive to regions with low methylation. • High cost, DNA is degraded by bisulfite treatment.	2014	[24]
scAba-seq	5hmC sequencing	• Low false-positive rate, no chemical degradation.	2016	[27]
scMSRE	Methylation-sensitive restriction enzyme sequencing	• Analysis limited to methylation at restrictions sites. • Doesn’t degrade DNA with chemical treatment.	2021	[28]

The above methods are common for single cell DNAme analyses; however, an increasingly popular single cell technique is the multi-omics method. This method integrates proteome, genomes, epigenomes and transcriptomes from the same cell to give a comprehensive insight into how the omics are all interrelated [29, 30]. This allows for the parallel profiling of multilayers in single cells to identify causal relationships between epigenome regulation and gene expression. This is a favourable technique as lack of methylation does not always infer gene expression and vice versa.

1.3 Epigenetics of Colorectal Cancer

Colorectal cancer (CRC) is the second leading cause of cancer-related deaths in New Zealand, with New Zealand having one of the highest CRC incidences in the world [31]. This incidence has significantly increased in the past few decades, which can in part be attributed to an ageing population, unhealthy modern diets and increased risk factors such as smoking, alcohol consumption and obesity [32]. As described earlier, these environmental influences have a large effect on our epigenomes. Currently, the most effective strategy against CRC mortality is early screening. When CRC is detected in early stages while it is still confined to the bowel wall, surgical cure rates are above 75% [33]. However, this rate drops significantly once the primary tumour has metastasised.

CRC is a highly molecularly heterogeneous disease [34]. Heterogeneity refers to the observation that individual tumour cells will present with distinct morphological and phenotypic profiles. Heterogeneity can be seen within a tumour (intra-tumour), between the primary tumour and its metastases (inter-tumour) and between individuals (inter-individual) [34]. The heterogeneous nature of CRC means patients will often have varied responses to treatment alongside varied prognoses [35].

Collectively, the genome and epigenome of cancer cells are grossly dysregulated. Exploitation of pathways required in normal biological processes results in the initiation, progression, dissemination and metastases of cancer [14, 36]. A common feature of the cancer epigenome is global DNA hypomethylation [37-39]. This is an early event often observed in early neoplastic lesions of the colon [40]. DNA global hypomethylation refers to an overall drop in the level of methylation observed across the cancer genome of neoplastic cells compared to adjacent healthy tissues. This DNA global hypomethylation contributes to the dysregulated cancer genome in a number of ways. Often, this state reflects a decrease in methylation of DNA repeat elements, which recent studies suspect makes up as high as two-thirds of the human genome [41]. The loss of methylation at these typically silenced repeat elements can lead to transposition, chromatin rearrangement and genomic instability contributing to the cancer genome [42].

Moreover, focal hypermethylation in promoters of tumour suppressor genes is another frequented epigenetic event in CRC [43-45]. Through silencing of promoter/enhancer regions or long-regulatory distal elements, genes often involved in cell cycle regulation, DNA repair, apoptosis, angiogenesis, invasion and adhesion can become silenced [46]. On account of this silencing, further genomic instability can also be imposed. Famously, the MLH1 gene involved in DNA mismatch repair (MMR) is frequently silenced in CRC patients with a high microsatellite instability (MSI) phenotype [44]. Another deeply explored gene is CDH1, the gene for the E-cadherin protein [47, 48]. CDH1 is often hypermethylated in CRC, in order to facilitate the dissemination of cells from the primary tumour and subsequent metastasis to new locations in the patient's body [47]. On account of this silencing, further genomic instability can also be imposed.

On the contrary, instances of focal hypomethylation have also been observed in CRC contexts [49]. Focal promoter hypomethylation has been reported in what is known as oncogenes. Oncogenes, when overexpressed, often contribute to cellular proliferation, metastasis and invasion [49]. For example, CD276 has been shown to be frequently hypomethylated and subsequently contribute to tumour progression and chemosensitivity through regulating oncogenic signalling pathways [50]. CD276 also contributes to tumour cell evasion of the immune system [51]. Loss of methylation at imprinted genes in the genome has also been reported, most commonly explored in CRC being the insulin-like growth factor 2 (IGF2) [52]. The loss of imprinting (LOI) means both parental alleles are now expressed, not just one. In the case of IGF2, it's overexpression leads to activation of downstream genes which are powerful cellular proliferators [52].

DNAme, although commonly associated with the silencing of genes, plays an expansive role across the genome. The presence of DNAme does not always infer silencing of gene expression, it also controls telomere length and recombination, chromatin structure and silencing repeat elements [53]. Understanding the complex interactions that act to regulate the cancer genome is essential when analysing not only SCS data, but bulk also.

The distinct methylation phenotypes described above can yield clinically useful biomarkers for the detection of CRC as well as prognostic/therapeutic predictions [54]. In the following section, I will detail how the introduction of SCS in exploring the outlined mechanisms can contribute clinically and theoretically to our understanding of CRC.

1.3.1 Single Cell Methylome Sequencing in Colorectal Cancer

The emergence of single cell bisulfite sequencing in the CRC research field is valuable, when considering the heterogeneity of the disease and the influence of DNAme on the cancer genome as described above. The following section will detail areas of research in CRC which would benefit from the use of scBS-seq, with a depiction of the areas included in figure 1.

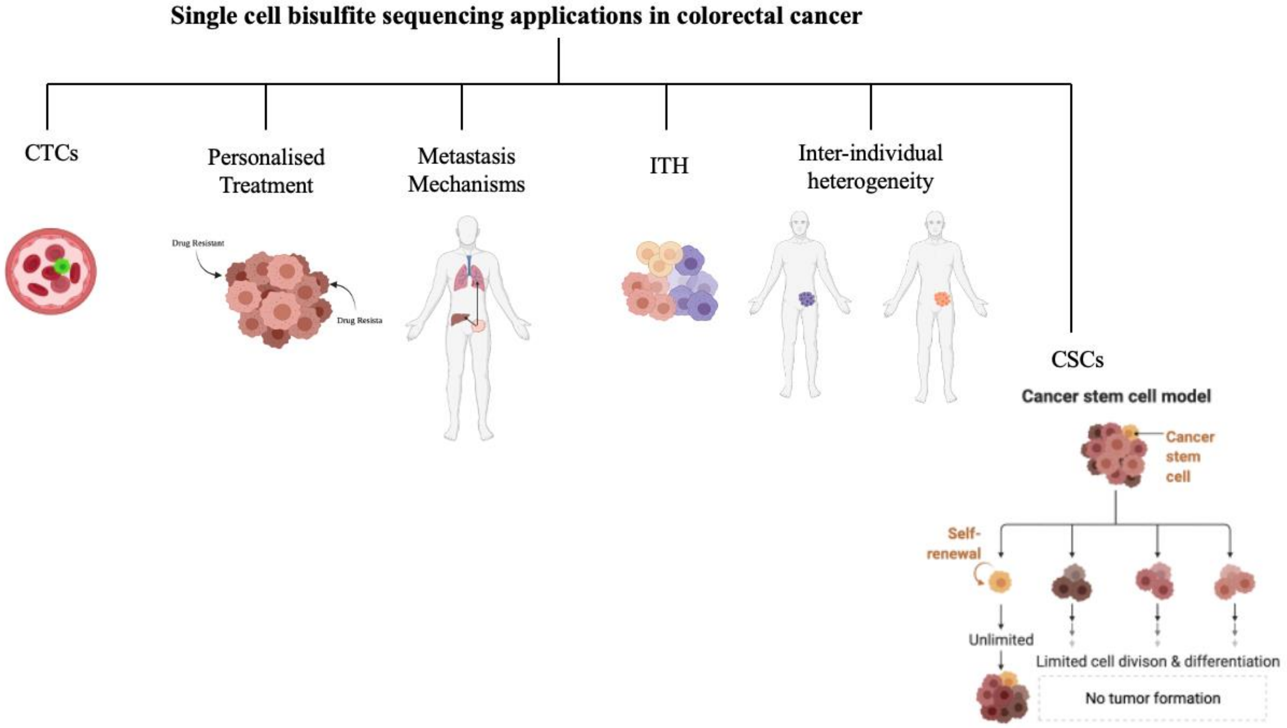

Figure 1 | Schematics of possible applications of single cell bisulfite sequencing in colorectal cancer contexts. Applications include sequencing circulating tumour cells, personalising patient treatment, investigating mechanisms of metastasis, elucidating intra-tumour heterogeneity, investigating inter-individual heterogeneity, and sequencing rare cancer stem cells. CTCs = circulating tumour cells, ITH = intra-tumour heterogeneity, CSCs = cancer stem cells. Figure made using biorender.com.

The first mechanism that would benefit from the use of scBS-seq is analysing circulating tumour cells (CTCs). CTCs are capable of detaching from the primary tumour, surviving in the blood stream and eventually colonising a new environment in the body [55]. Studies have shown DNAme heavily influences the ability of a CTC to metastasise through influencing EMT genes as well as stem cell-like genes [56]. On account of this, scrutinising the methylome of CTCs is an essential component to understanding the epigenetic factors which favour these cells over others in the tumour to metastasise. Analysing CTCs has been difficult previously, due to the low DNA content. Using scBS-seq to analyse CTCs has the potential to provide insight into the epigenetic processes that may drive the metastatic cascade of CRC cells and subsequent invasion at new locations, as well as introduce new biomarkers.

Additionally, due to CTCs being found in the blood, being able to extract these cells from blood drawing is far less invasive than other methods such as biopsies. Also of note, one study found that there was a substantial amount of methylome heterogeneity amongst distinct CTCs from the same patients [57]. This could lead to personalised signatures in these CTCs for patient stratification and treatment selection. In essence, scBS-seq allows for investigation of new potential biomarkers and identification of these in patients CTCs.

Other applications of scBS-seq in a CRC context is investigating intra-tumour heterogeneity. Tumours in CRC patients are comprised of a number of different cells within the tumour, including the malignant and healthy cells from the bowel, fibroblasts, immune cells and nerves [58]. Prior to SCS, these cells were all sequenced as a whole, disregarding the distinct methylomes that each of these cells would have. Not to mention, the variation of methylomes observed between tumour cells alone is substantial, let alone including methylomes of healthy cells to further mask any differential areas. Single-cell sequencing provides a tool for understanding the complex heterogeneity that encompasses each individual tumour. Such information will not only augment our current understanding on the CRC methylome but provide clinical insights such as responsiveness to therapy, patient prognoses and disease relapse.

Another branch of cancer research that SCS allows further investigation into is cellular lineage. Understanding lineage can inform on the cancer cell-of-origin, working to characterize and identify when the tumorigenic transformation occurred [59]. It can also trace lineages of cancer stem cells (CSC), a small number of cells thought to be capable to self-renewal, differentiation and tumorigenicity [60]. It is thought CSC’s are resistant to chemotherapy and radiation, a potential cause of metastasis and even initiation of the primary tumour [61]. As mentioned earlier, metastasis of CRC significantly decreases a patient's chance of recovery, and therefore exploring small subpopulations such as CSCs is a vital task which can now be done at higher precision with SCS. Without scBS-seq, CSCs which are likely to have methylomes which diverge from normal neoplastic cells would be hidden as majority of the tumour is not composed of CSCs.

The concept of heterogeneity is of notable importance in the context of therapy, as it only requires one cell to be non-responsive to the therapy to result in a relapse, or in some cases no response at all. For example, it was found glioblastoma patients who had methylated MGMT promoters were more responsive to alkylating agents for treatment as opposed to those without this methylation [62]. One study also showed that different epigenomic subpopulations varied in their response to targeted therapy, with one subpopulation showing greater resistance to imatinib [63]. Using scBS-seq to identify any cells with methylation biomarkers which may be indicative of resistance to certain therapies could inform oncologists on which treatments should be combined for the most effective treatment; a form of personalised treatment. Furthermore, using scBS-seq to monitor the progress of treatment and any potential arising drug-resistant subpopulations through epigenetic biomarkers could be used to direct future steps for therapy. Prior to SCS, these small drug-resistant populations would have had their biomarkers masked by the averaged profile of all the other tumours cells. Of course, methylome sequencing is not yet a frequently used tool in clinical practice, hence the above arguments are merely suggestions of what could be achieved in future if it were to become common practice.

The use of single cell techniques has been popularly used in single cell RNA analyses of cancer rather than methylome analyses. The number of papers available on RNA single cell analyses yields upwards of 8,000 PubMed search results in the past 5 years, comparatively single cell DNAme analysis papers yield less than 300 PubMed search results. This is likely in part due to the difficulty of DNAme analyses. Single cell DNAme only utilises the double stranded DNA in the single cell, whereas there is a lot more RNA present in a single cell. As well as this, the process of amplification is much easier on the RNA than on the DNA during methylation. mRNA is converted to cDNA via reverse transcriptase and subsequently amplified, whereas DNAme analyses require manipulation of the native DNA [64]. As a result of this, there is still a limited amount of research done on single cell methylomes, even less so in a cancer context. The currently available papers which used single cell methylome sequencing in the context of cancer have been summarised in table 2. Nonetheless, the epigenetic influence on CRC and its inherent heterogeneity makes single cell DNAme sequencing an enticing direction of analysis.

Table 2. Examples of single cell methylome analyses in cancer. Adapted from Karemaker and Vermeulen [2].

Cell Line or Tissue	Genome-wide or Gene specific.	Comments	Year Published	References
Hepatocellular carcinoma.	Genome-wide.	Identified subpopulations within tumour.	2016	Hou, Guo [65]
Metastatic breast cancer (mBC) and metastatic castration-resistant prostate cancer (mCRPC).	CDH1 and miR200 promoter s.	CTCs from same patient displayed heterogeneous methylation patterns. Found different methylation patterns at these promoters in mCRPC vs. mBC CTCs suggesting differentially regulated miR200 loops in these two tumour entities.	2017	Pixberg, Raba [57]
Colorectal cancer.	Genome-wide.	Sub-lineages identified in patients found metastases at multiple sites had a common origin.	2018	Bian, Hou [1]
Chronic Lymphocytic Leukaemia.	Genome-wide.	Identified subpopulations which were preferentially expelled from lymph node after treatment.	2019	Gaiti, Chaligne [66]
Glioma.	Genome-wide.	Identified intra-tumoural epigenetic variation which linked subclonal competition and phenotypic state changes.	2021	Johnson, Anderson [67]
CTCs from multiple cancers.	Genome-wide.	Demonstrated tumour origin classification of CTCs based on methylomes.	2021	Chen, Su [68]

1.4 Challenges in Single Cell DNA Methylation

The ability to sequence individual cells to uncover cellular heterogeneity at a mono- or multi-modal level is a big jump in the field, yet not without its challenges. There tends to be coverage non-uniformity, sparse data, false-positives, amplification biases and allelic drop out events [69].

As mentioned earlier, the popularity of scRNA-seq has resulted in a number of bioinformatic tools to aide in making inferences from single cell data. Clustering methods for cell population characterization [70-72] as well as network inference tools [73, 74] have been developed for scRNA-seq techniques. Comparatively, the lack of studies done thus far in scBS-seq is reflected in the lack of bioinformatic tools. In addition, there are no clear DNAme cell type markers as there are transcriptomic cell type markers, making clustering more difficult.

High-throughput single cell DNAme studies have CpG coverage of around 5% [75], while low throughput studies have around 20% [76] genome-wide CpG coverage. This makes it relatively difficult to distinguish cells from one another with large gaps, or to infer the epigenetic control mechanisms of that cell with very sparse coverage. The analysis tool MELISA (MEthyLation Inference for Single cell Analysis) has been created to alleviate these issues. MELISA predicts methylation status of missed CpGs based on information from neighbouring CpGs and from other cells with similar methylation patterns [77]. Although, it must be considered that through using predictive statistical models such as MELISA, the methylation status may be incorrectly predicted and therefore lose a differently methylated site. Reword. better

As expected, when sequencing a minute amount of DNA, the technical noise is very high. The scarce coverage of processed single cell epigenome data requires appropriate normalization of data and these high levels of noise need to be accounted for. Single cell transcriptomics have spike-in standards to control for technical noise, yet strong normalisation strategies have not yet been established for epigenome sequencing. One study combined approximately 100 single cells to identify peaks via algorithms already in place for bulk sequencing, then looked at each cell to see if these peaks were present [78]. This aggregation method however cannot account for cells with specific loci exhibiting low levels of DNAme. A number of methods have also been formed through comparing regions which have similar methylation levels in cis-regulatory elements through ENCODE [79]. Another challenge with this small amount of DNA is the bisulfite conversion process. It is a harsh reaction on a very small amount of DNA, leading to substantial degradation. Although methods have been adapted to account for this, it still remains a challenge.

The potential for contamination throughout the process of single cell DNA methylation is also very high and may skew results. Any DNA that contaminates samples early on will be amplified with the cellular DNA and may provide false results. Also, the multiple rounds of amplifications required means addition of reagents may lead to further contamination. Negative controls at multiple time points such as before treatment, after and throughout amplification may help to alleviate these issues.

Another consideration when analysing SCS data is the target resolution level. This refers to grouping the single cells into different organs, different cell types or up to each single cell at intermediate cell states. Appropriate analysis tools and reference systems such as cell atlases need to be used to reach different levels of resolution. Dependant on the research question, higher resolution may be more insightful, yet increased resolution tends to coincide with a decrease in stability of supporting signals. For example, when researching the EMT process in metastatic tumour cells, a high resolution including intermediate cell states would be most appropriate as the intermediate states influence the phenotype and invasive/migratory properties of the cell as described earlier.

Another context which provides new challenges is clinical implementation. Theoretically, the use of SCS in a clinical context particularly in targeted therapy for cancer patients seems promising. However, more cost and time efficient methods need to be developed which can be easily utilised in clinics. This includes cell preparation and streamlined data analysis pipelines, where only DNAme levels at loci relevant to the clinical phenotype are shown. However, all emerging technologies initially have a high cost, but eventually become very affordable. The cost of single-cell transcriptomic assays has already fallen considerably [80], and it is likely single cell epigenomics will follow.

The level of technical variability associated with single cell methylation should always be considered when drawing conclusions from sc-BS sequencing data.

1.5 Research Aims

The aim of this project was to implement and optimise a single cell bisulfite sequencing protocol on a human sorted colorectal cancer cell line, HT29. The second aim of this project upon completion of the first aim, was to bioinformatically analyse the scBS-seq libraries of the HT29 cells and investigate the DNAme heterogeneity in the population. Due to COVID-19 restrictions, the second aim was altered. My second aim was then to analyse a publicly available CRC DNAme dataset to examine it for DNAme heterogeneity between cells. I hypothesised that DNAme heterogeneity amongst the population would be apparent considering the heterogeneous nature of CRC.

Chapter 2: Materials and Methods

2.1 Cell Culture

HT29 cells were initiated from liquid nitrogen storage and thawed at 37^{o}c. Cells were kept in DMEM + FBS + Penicillin/streptomycin (10% FBS and 1% PS). HT29 is a tumorigenic colorectal cancer cell lines with epithelial morphology. HT29 cells were obtained from the ATCC.

Cell media was changed every 1-2 days. If cells were >80% confluent, cells were split by washing with PBS, trypsinisation for 3 minutes at 37^{o}c, then spun down in falcon tubes containing media. Following this, supernatant was replaced with fresh media and transferred to a new flask. There were 2 passages after thawing of cells before cells were sorted, meaning minimal heterogeneity would have accumulated over this time and cell culture adaptation was minimised.

2.2 Preparation of HT29 Cells for FACS

FVS450 stain was used for live-dead single cell sorting. 3 tubes were prepared for sorting. Tube 1 was an unstained control. Tube 2 contained heat killed cells as a FVS450 control. Tube 3 contained FVS450 stained cells for sorting. Cells were harvested via trypsination and spun down at 350g for 5 minutes. Cells were resuspended in PBS and diluted to approximately 300,000 cells per tube in a 500µl volume. Tube 2 cells were heat killed at 100^{o}c for 5 minutes on a stirrer plate. 1.5µl of FVS450 viability dye was added to 1.5ml of PBS to make the viability mastermix. 500µl of this viability mastermix was added to tube 2 (heat killed cells) and tube 3 (cells for sorting). 500µl of PBS-BSA was added to tube 1, as an unstained control. All tubes were incubated at room temperature for 15 minutes in the dark.

All tubes were washed with MACS buffer and spun down, then repeated. All tubes had MACS buffer added, followed by straining through a 70µm filter.

Cells were sorted using the BD FACSAria into 0.2ml lo-bind DNA strip tubes, each tube containing 2.5µl of RLT Plus Buffer for immediate cell lysis and protein denaturation. Cells were sorted into tubes containing this RLT Plus Buffer as this is lyses the cell and exposes the genomic DNA for future steps in the protocol. Each row of strips contained the following amounts of cells from left to right: 10000, 1000, 100, 10, 5, 1, 1, 1, as depicted in figure 2A.

All following steps were performed in a UV-irradiated laminar-flow hood, with all equipment and reagents which could withstand UV-treatment were UV treated for a minimum of an hour before use.

2.3 Single Cell Bisulfite Sequencing

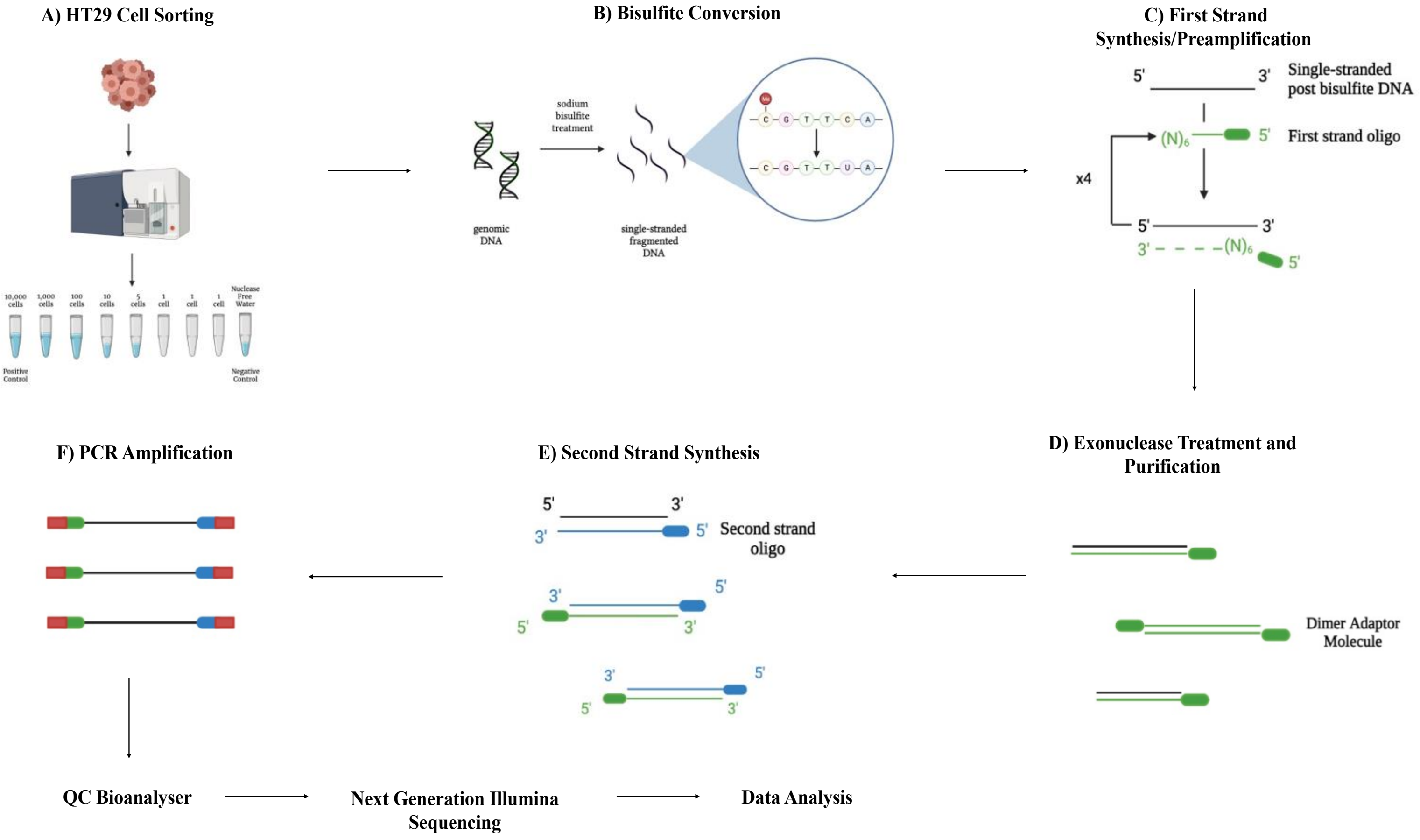

Figure 2 | Overview of protocol followed for single cell bisulfite sequencing process. A) HT29 cells were sorted into varying amounts per tubes via BD FACSAria. B) Genomic DNA was bisulfite treated, followed by C) preamplification. D) Exonuclease and purification treatment is then performed to remove dimers and residual reagents. E) Second strand synthesis is followed by F) PCR amplification and subsequent sequencing and analysis of the library. Adapted from [3].

DNA Bisulfite Conversion

Bisulfite conversion was done directly onto lysed cells, as quantification of genomic DNA is not possible in single cell sequencing. The purpose of this step is to be able to distinguish methylated cytosines from unmethylated cytosines. Genomic DNA from the cells is treated with sodium bisulfite, resulting in the DNA becoming single-stranded and fragmented. Unmethylated cytosines are deaminated, converting them to uracil's, which post-PCR amplification will appear as thymine's. Methylated cytosine's will not be affected by the sodium bisulfite treatment. Therefore, upon sequencing, the cytosines present will only be methylated cytosines. This process can be seen in figure 2B.

The Zymo EZ-methylation kit was used. 790µl of M-solubilization Buffer and 300µl of M-Dilution Buffer was added to the CT conversion powder vial. Vial was wrapped in tin foil and shaken for 10 minutes until all particles completely dissolved. 160µl of M-Reaction Buffer was added to vial post-shaking.

7.5µl of water was added to each lysate sample, followed by 65µl of the CT conversion reagent. Samples were incubated on a thermocycler as follows:

98^{o}C for 00:08:00

65^{o}C for 03:00:00

4^{o}C hold.

50µl of Zymo Magbinding beads was added to 3ml of M-Binding buffer. 305µl of this mixture was added to each sample, then incubated for 5 minutes at room temperature. Beads were then palleted, followed by supernatant removal and an ethanol wash. 100µl of M-Desulphonation buffer was added to each sample, then incubated for 15 minutes. Following this, beads were pelleted and washed twice with ethanol. Beads were then dried at 60^{o}c for 10 minutes, then resuspended in 20µl of the following first strand synthesis mix:

Table 3. Reagents used for first strand synthesis.

Reagent	Amount per sample	Amount for 10 (9+1 for error)
Nuclease Free Water	32.8µl	328µl
10x Blue Buffer	4µl	40µl
dNTP mix (10mM each)	1.6µl	16µl
First strand Oligo (10uM)	1.6µl	16µl
Total	40µl	400µl

Samples were incubated with first strand synthesis mix for 5 minutes; beads were pelleted and 20µl supernatant was transferred to fresh lobind PCR tubes. 20µl of first strand synthesis mix was added to beads again and the process was repeated once more – leaving a total of 40µl of supernatant in each sample.

First strand oligonucleotide sequence: /5SpC3/CT ACA CGA CGC TCT TCC GAT CTN NNN NN (HPLC purification).

First Strand Synthesis / Preamplification

The purpose of this step is preamplification of the genomic DNA post-bisulfite treatment. This limits the loss of informative sequences which are usually lost when bisulfite conversion is completed after adaptor tagging, conserving the complexity of the libraries. A complementary strand is synthesised via random priming and extension to the fragmented BS-converted DNA. These strands are primed by oligonucleotides which contain Illumina adaptor sequences at the 5' end and a 3' stretch of six random nucleotides. There are five rounds of amplification, illustrated in figure 2C, to maximize the number of tagged strands and to create multiple copies of each fragment.

Samples were placed on thermocycler at 65°c for 3 minutes then immediately cooled on an ice block. 1µl of Klenow exo- (50 U/µl, Enzymatics) was added to each sample, then incubated on a thermocycler as follows:

4°C 00:05:00

Slow ramp from 4°C to 37°C at 15s per 1°C

37°C 00:30:00

4°C hold.

Following incubation, samples were heated to 95°c for 45 seconds then immediately cooled on an ice block. 2.5µl of the following first-strand extra cycles mix was added to each sample:

Table 4. Reagents used for preamplification of DNA.

Reagent	Amount per Sample	Amount for 10	Amount for 5 rounds
Nuclease free water	0.65µl	6.5µl	32.5µl
10x Blue Buffer	0.25µl	2.5µl	12.5µl
dNTP mix (10mM each)	0.1µl	1µl	5µl
First strand oligo (10uM)	1µl	10µl	50µl
Klenow exo- (50U/ul)	0.5µl	5µl	25µl
Total	2.5µl	25µl	125µl

Samples were incubated on a thermocycler as follows:

4^{o}C 00:05:00

Slow ramp from 4^{o}C to 37^{o}C at 15s per 1^{o}C

37^{o}C 00:30:00

The process of samples being once again heated to 95^{o}c for 45 seconds and cooled, followed by the addition of 2.5µl mastermix and the thermocycler incubation is repeated an additional three times, totalling four rounds. For the fifth and final round, the final round of PCR was incubated for an additional hour, as follows:

4^{o}C 00:05:00

Slow ramp from 4^{o}C to 37^{o}C at 15s per 1^{o}C

37^{o}C 01:30:00

4^{o}C hold.

Exonuclease Treatment

The purpose of this step is for the Exonuclease I enzyme to degrade any remaining primers or oligonucleotides which did not get incorporated during the first strand synthesis step. This is important otherwise dimer adaptor molecules such as those displayed in figure 3D would be synthesised.

50µl of the following Exonuclease mix was added to each sample:

Table 5. Reagents used for exonuclease treatment.

Reagent	Amount per sample	Amount for 10
Nuclease free water	48µl	480µl
Exonuclease I (NEB)	2µl	20µl

Total	50µl	500µl

Samples were incubated at 37^{o}C for one hour.

First Strand Purification

This step allows for the likes of buffers, nucleotides, degraded primers and enzymes to be washed away. This ensures only tagged strands are present for the second strand purification.

AMPureXP beads were equilibrated to room temperature for 10 minutes. 70µl of AMPureXP beads were added to samples, then incubated at room temperature for 10 minutes. Beads were palleted and washed twice with ethanol. Supernatant was removed and beads were air dried at 50^{o}C for 5 minutes.

Second Strand Synthesis

The second strand oligo is the second adaptor which is integrated similarly to the first strand adaptor for the same purpose, shown in figure 2E.

49µl of the following second strand master mix was added to each sample:

Table 6. Reagents used for second strand master mix.

Reagent	Amount per sample	Amount for 10 samples
Nuclease Free Water	40µl	400µl
10x Blue Buffer	5µl	50µl

dNTP mix (10mM each)	2µl	20µl
Second Strand Oligo (10uM)	2µl	20µl
Total	49µl	490µl

Samples were incubated for 10 minutes at room temperature to elute DNA from beads.

Second strand oligo sequence: TGC TGA ACC GCT CTT CCG ATC TNN NNN N (HPLC Purification).

Samples were then incubated on a thermocycler at 98^{o}C for two minutes, followed by immediate cooling on an ice block. 0.5µl of Klenow exo- (50U/µl, Enzymatics) was added then incubated as following:

4^{o}C 00:05:00

Slow ramp from 4^{o}C to 37^{o}C at 15s per 1^{o}C

37^{o}C 01:30:00

4^{o}C hold.

Second Strand Purification

700µl of AMPure Buffer (AMPure XP beads supernatant) was added to 500µl of nuclease free water. 120µl of this mixture was added to each sample, mixed thoroughly and incubated at room temperature for 10 minutes. Beads were then palleted, washed twice with ethanol and air dried for 50^{o}C at 5 minutes.

Library Amplification

In this final step, the pre-amplified bisulfite treated DNA is amplified to create a sufficient amount of DNA for sequencing. Different index primers are incorporated into each individual sample, allowing samples to be combined during sequencing yet still distinguishable from one another at later analyses. This step is depicted in figure 2F.

Beads were then resuspended in 48µl of the following PCR mastermix:

Table 7. Reagents used for PCR mastermix.

Reagent	Amount per sample	Amount for 10 samples
Nuclease free water	22µl	220µl
KAPA Hifi Ready Mix	25µl	250µl
PE1.0 (10uM)	1µl	10µl
Total	48µl	480µl

PE1.0 Sequence: AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC*T (HPLC Purification). Primer contains the full Illumina P5 and PE read 1 sequence.

2µl of iTag indexing primer (5µM) was added to each sample. The following indexes were added corresponding to the sample they were put in i.e., iTag index primer 1 was added to sample 1 and iTag index primer 2 to sample 2.

Table 8. iTag indexing primers used for each sample.

Primer	Primer sequence	Sequence obtained
iPCR Tag 1	CAAGCAGAAGACGGCATACGAGATAACGTGATGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T	ATCACGTT
iPCR Tag 2	CAAGCAGAAGACGGCATACGAGATAAACATCGGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T	CGATGTTT
iPCR Tag 3	CAAGCAGAAGACGGCATACGAGATATGCCTAAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T	TTAGGCAT
iPCR Tag 4	CAAGCAGAAGACGGCATACGAGATAGTGGTCAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T	TGACCACT
iPCR Tag 5	CAAGCAGAAGACGGCATACGAGATACCACTGTGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T	ACAGTGGT
iPCR Tag 6	CAAGCAGAAGACGGCATACGAGATACATTGGCGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T	GCCAATGT
iPCR Tag 7	CAAGCAGAAGACGGCATACGAGATCAGATCTGGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T	CAGATCTG
iPCR Tag 8	CAAGCAGAAGACGGCATACGAGATCATCAAGTGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T	ACTTGATG

Samples were incubated with indexing primers for 10 minutes at room temperature to elute DNA from beads.

Samples were then incubated on the thermocycler as follows:

95^{o}C 00:02:00

16 Cycles of:

94^{o}C 00:01:20

65^{o}C 00:00:30

72^{o}C 00:00:30

Followed by:

72^{o}C 00:03:00

4^{o}C hold.

33.6µl of AMPureXP beads were added to each sample, then incubated at room temperature for 10 minutes. Beads were palleted, then washed twice with ethanol. 20µl of nuclease free water was added to each sample for 10 minutes to elute DNA from beads. Beads were palleted and DNA containing nuclease free water supernatant was removed and transferred into a fresh PCR tube. Purifying process was repeated once more, so a total of 40µl of the final library was in each tube.

2.4 2100 Agilent High Sensitivity DNA Bioanalyser

Samples were analysed using a qubit fluorometer to quantify the amount of DNA present in each sample. The NanoPhotometer was then used to assess the purity of each sample. Following this, the Agilent High Sensitivity DNA Kit was used to check the library size distribution on a High Sensitivity DNA Chip on an Agilent 2100 Bioanalyzer.

Samples with Qubit concentrations greater than 0.3 ng/μl, with appropriately sized fragments (300 – 500 bp) were selected for library preparation for MiSeq. The final library pool was diluted to 40μl (1.5nM) in preparation for iSeq sequencing.

2.5 Illumina iSeq Sequencing

The multiplexed library was sequenced via Illumina iSeq Sequencer, performed by Rob Weeks (Department of Pathology, University of Otago).

2.6 Publicly Available Data Analyses

Publicly available single cell methylation data was obtained from Bian, Hou [1]. The scTrio-Seq data was obtained from the NCBI Gene Expression Omnibus (GEO) using accession number GSE97693. A table of the selected cells from each subregion is displayed in table 9.

The average methylation status of genes and their promoters was determined by finding the average level of methylated versus unmethylated CpG sites within the promoter/gene region in question based on ENSEMBL gene and promoter regions. A methylation of 0 meant no CpG sites were methylated within the region. A methylation of 1 meant all CpG sites within the region of interest were methylated.

Table 9. CRC patient cells chosen for analysis from public dataset. Data obtained from Bian, Hou [1].

Cell ID	Sampling Region	Methyl GSM_ID	SRR_ID	RNA Expression GSM_ID
507	NC	GSM2697695	SRR5825131	
509	NC	GSM2697696	SRR5825132	
517	NC	GSM2697697	SRR5825133	
521	NC	GSM2697698	SRR5825134	
525	NC	GSM2697699	SRR5825135	
527	NC	GSM2697700	SRR5825136	
529	NC	GSM2697701	SRR5825137	
213	PT1	GSM2697711	SRR5825147	GSM2697231
229	PT1	GSM2697719	SRR5825155	GSM2697237
230	PT1	GSM2697720	SRR5825156	GSM2697238
324	PT2	GSM2697745	SRR5825181	
484	PT2	GSM2697752	SRR5825188	GSM2697285
559	PT2	GSM2697767	SRR5825203	
336	PT3	GSM2697772	SRR5825208	
347	PT3	GSM2697780	SRR5825216	GSM2697308
376	PT4	GSM2697809	SRR5825245	
420	PT4	GSM2697826	SRR5825262	
166	LN1	GSM2697488	SRR5824924	GSM2696959
167	LN1	GSM2697489	SRR5824925	GSM2696960
175	LN1	GSM2697493	SRR5824929	
176	LN1	GSM2697494	SRR5824930	
183	LN2	GSM2697496	SRR5824932	
457	LN2	GSM2697509	SRR5824945	
268	LN3	GSM2697511	SRR5824947	
441	LN3	GSM2697513	SRR5824949	
270	LN3	GSM2697512	SRR5824948	
442	LN3	GSM2697514	SRR5824950	
33	ML1	GSM2697528	SRR5824964	GSM2697017
36	ML1	GSM2697531	SRR5824967	GSM2697020
73	ML2	GSM2697541	SRR5824977	GSM2697051
76	ML2	GSM2697542	SRR5824978	GSM2697053
116	ML3	GSM2697549	SRR5824985	GSM2697078
117	ML3	GSM2697550	SRR5824986	GSM2697079
148	ML4	GSM2697571	SRR5825007	GSM2697104
141	ML4	GSM2697565	SRR5825001	
149	ML4	GSM2697572	SRR5825008	GSM2697105
150	ML4	GSM2697573	SRR5825009	GSM2697106
223	MP1	GSM2697615	SRR5825051	GSM2697152

217	MP1	GSM2697611	SRR5825047	GSM2697148
36	MP2	GSM2697650	SRR5825086	
52	MP2	GSM2697658	SRR5825094	
103	MP3	GSM2697661	SRR5825097	
105	MP3	GSM2697663	SRR5825099	GSM2697199
80	MP4	GSM2697669	SRR5825105	GSM2697204
204	MP4	GSM2697683	SRR5825119	GSM2697214
112	MP5	GSM2697689	SRR5825125	GSM2697220
116	MP5	GSM2697692	SRR5825128	GSM2697223

2.7 Statistical Analysis

One-way ANOVA followed by a Tukey's multiple comparison test was used to compare the methylation differences between normal adjacent tissue, primary tumour and distant metastases. P-values of <0.05, <0.01, <0.001 and <0.0001 were considered statistically significant.

Pearson's correlation co-efficient tests were performed between methylation and expression datasets to investigate any statistical relationships.

Chapter 3: Results

3.1 Implementation and Optimisation of Single Cell Bisulfite Sequencing Protocol

The following chapter addresses aim one of my thesis; to implement and optimise the single cell bisulfite sequencing protocol. The succeeding subsections describe the steps taken to prepare for implementation of the protocol, as well as the process of optimisation.

3.1.1 Fluorescence Activated Cell Sorting

This section explores the sorting of the human colorectal cancer cell line described in aim 1 of my thesis. A range of cell amounts were sorted based on a live-dead stain to ensure viable cancer cells were used for the protocol, as this would have a large influence on the library quality.

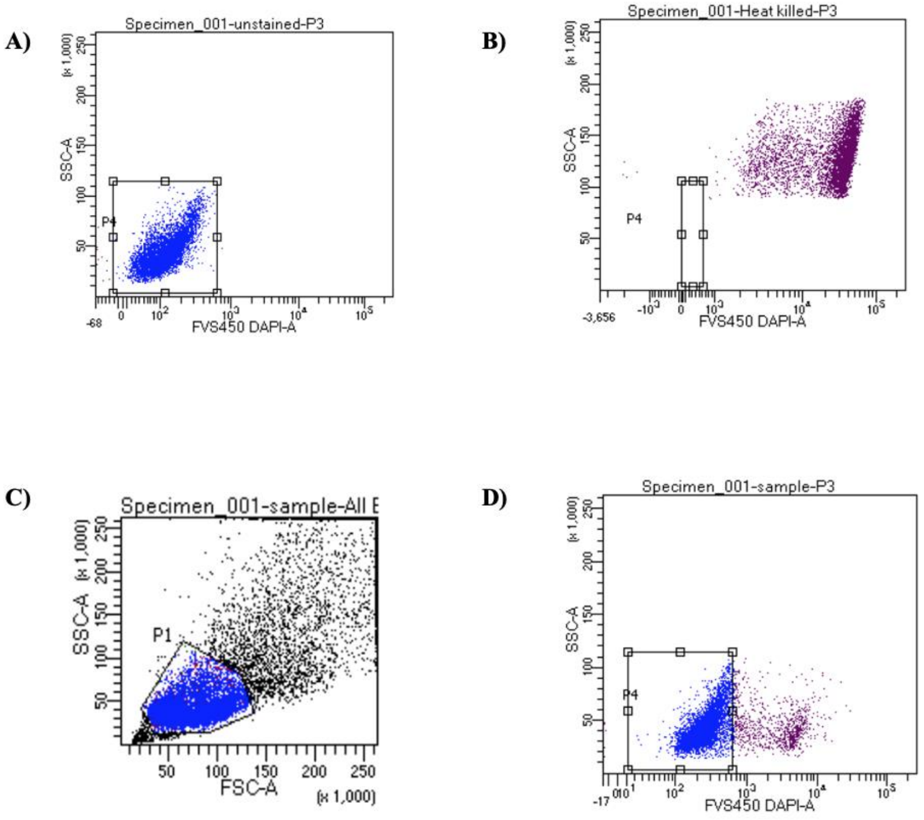

Figure 3 | FACS Sorting of HT29 cells using FVS450 dye to select for viable live cells. A) Positive control. Unstained for viable HT29 cells. Events = 6,132. **B)** Negative control. Heat killed HT29 cells with high fluorescence. Events = 5,411. **C)** Contains entire sample population of HT29 cells for sorting – both live and dead. **D)** Uptake of FVS450 allowed for live and dead population segregation. Viable cells in blue, non-viable in purple. Events = 5,251.

Figure 3 shows live-dead staining for FACS of HT29 cells. A live-dead FVS450 stain was used to discriminate between live and dead cells. FVS450 discriminates between live and dead cells as non-viable/dead cells will have a higher uptake of the dye due to permeable cell membranes, whereas live cells will have less permeable membranes. Figure 3A shows positive control cells, which did not receive any FVS450 dye. As live cells are less likely to take up the dye due to their intact membranes, the unstained cells give reference for what fluorescence live cells will take as they will appear unstained. Figure 3B contained heat killed cells, which were exposed to the FVS450 dye. The heat-killed cells, like other non-viable cells, had permeable membranes resulting in a high uptake of FVS450 dye and consequently higher fluorescence. This acted as a reference point for at which fluorescence a cell was considered non-viable. Figure 3C displays the entire population of HT29 cells exposed to FVS450 dye. Figure 3D depicts the segregated live and dead cell population based on their fluorescence. Cells which were deemed viable were then sorted into tubes containing either 1, 5, 10, 100, 1000 or 10,000 cells.

3.1.2 Optimisation of Single Cell Bisulfite Sequencing Protocol

With respect to the optimisation component of my first aim, the following section explores the steps taken to improve and eventually successfully implement the single cell bisulfite sequencing protocol. It explores each attempts successes and failures, guiding future direction for refinements of the protocol in the following attempt. This section focuses largely on the use of the 2100 Agilent Bioanalyzer High-Sensitivity DNA chip profiles to indicate if the protocol has been successful in creating high-quality libraries.

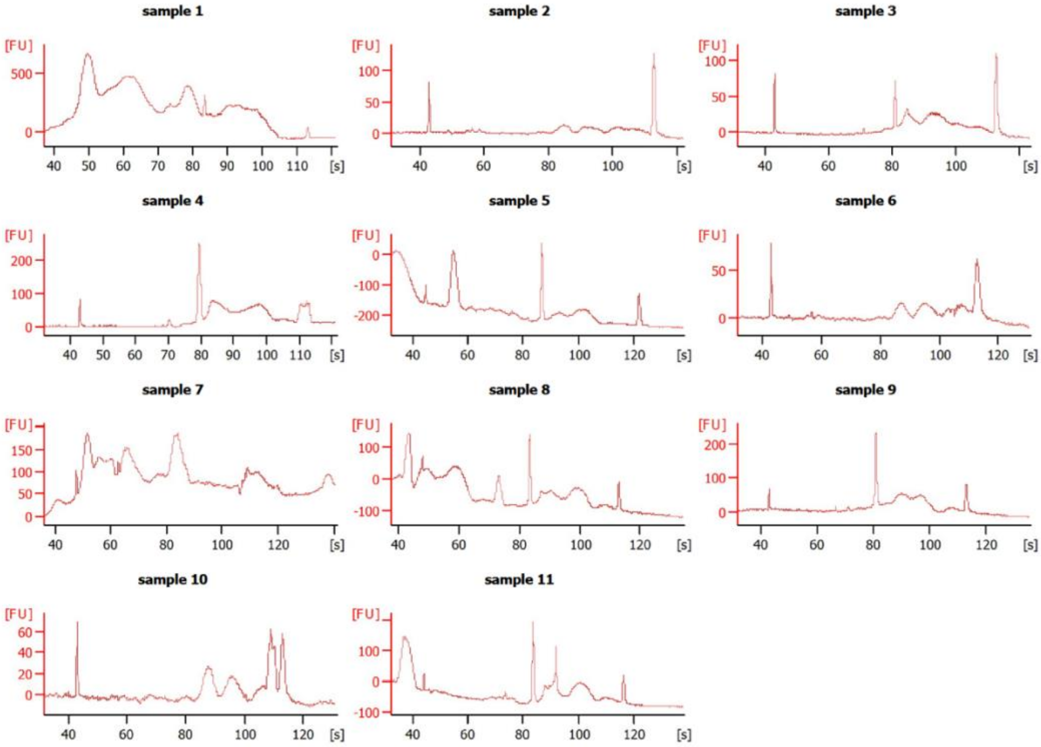

Figure 4 | Agilent 2100 Bioanalyzer profiles for single cell bisulfite libraries. Library profiles of HT29 single cells post-whole genome bisulfite amplification. Attempt 1. Graphics show library size distribution. No peak library length as libraries were of poor quality.

Results from the first attempt were clamorous, however this was expected for the first attempt at a complex protocol. Following these results, extra precautions were taken to irradicate contamination as much as is feasible. This included increasing periods of UV irradiation of all reagents and equipment capable of withstanding UV-irradiation to two hours. Those not able to withstand UV were cleaned with ethanol frequently. This was done to reduce the large amount of contamination resulting in the sharp messy peaks in the bioanalyser profiles of figure 4.

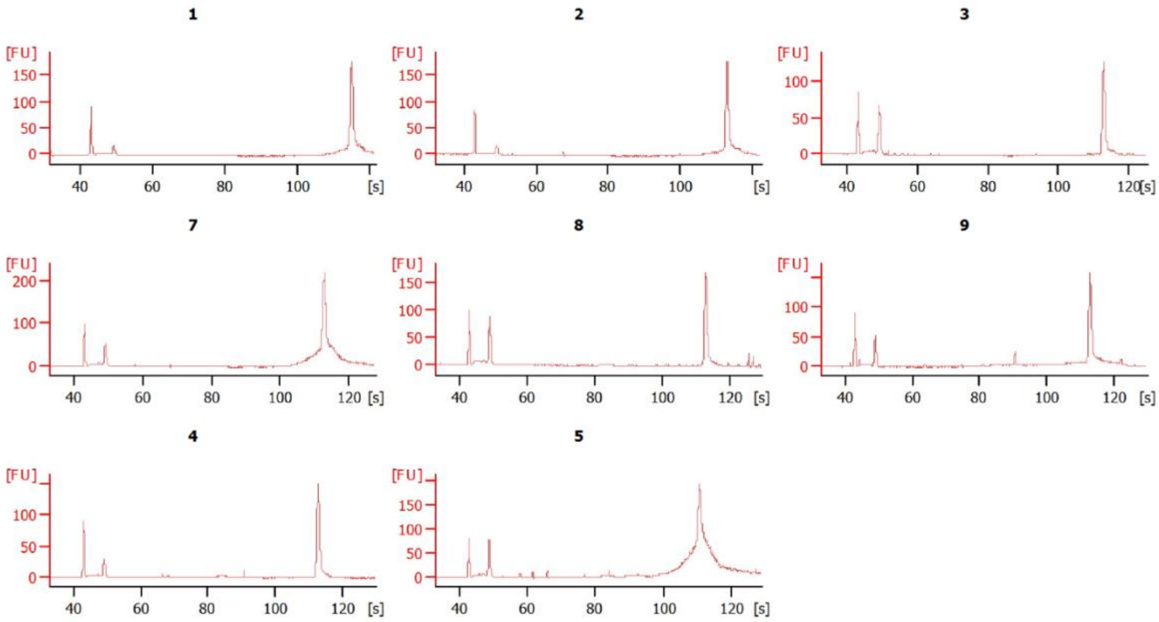

Figure 5 | Agilent 2100 Bioanalyzer profiles for single cell bisulfite libraries. The library profiles of HT29 single cells post-whole genome bisulfite amplification. Graphics show library size distribution. Attempt 2. No peak library length as cells were not successfully amplified. Lower marker = sharp peak at 40 [s], upper marker = sharp peak at 120 [s].

While attempt two (figure 5) did not show signs of a successfully amplified single cell library, that being a smooth peak around 300 – 500 bp mark, in contrast to the earlier Bioanalyzer profile, contamination appeared to be significantly reduced. However, almost all the bioanalyzer profiles in this attempt had sharp peaks near to the lower marker. This is most likely residual adapters which were not correctly washed away. According to the protocol, the sharp peaks are often a result of having the incorrect ratio of AMPureXP beads during purification steps, resulting in left over fragments and oligo concatemers not being removed. Also, likely reflecting the less noisy profiles, the lower and upper markers have worked to a much higher standard than the previous attempt.

For the next attempt (figure 6), the number of transfer steps in the protocol was reduced. The purpose of this was to minimise the opportunities for sample loss to occur as a result of transferring between tubes. Also, extra care was taken when pipetting the AMPureXP beads to ensure the correct ratio was being added to each sample.

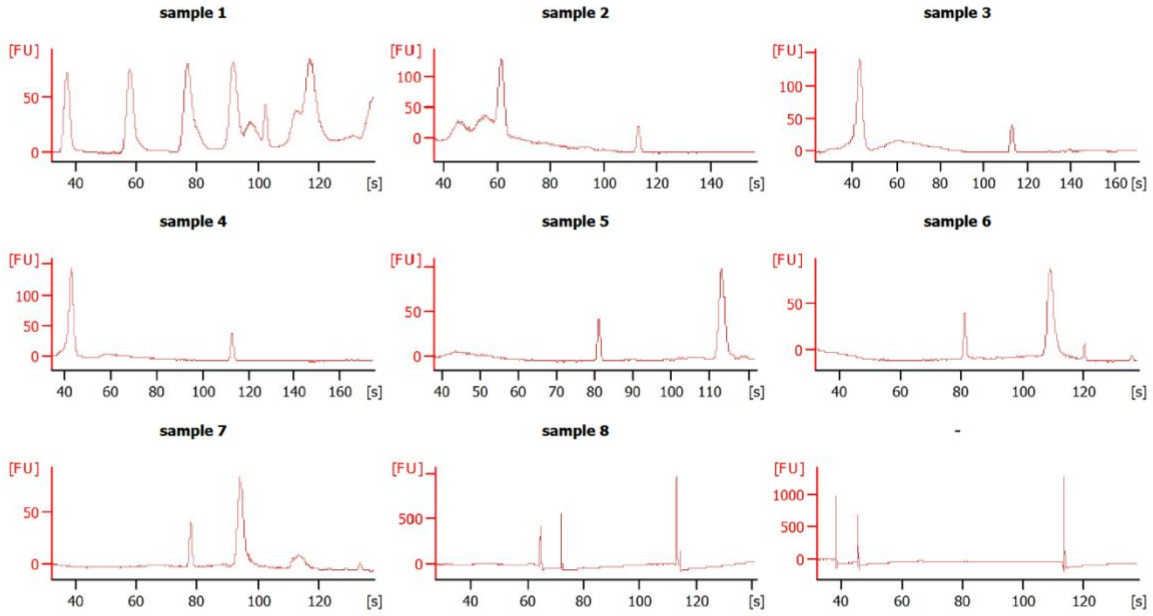

Figure 6 | Agilent 2100 Bioanalyzer profiles for single cell bisulfite libraries. Library profiles of HT29 cells post-whole genome bisulfite amplification. Graphics show libraries size distribution. Attempt 3. Libraries each contain a single cell. '-' denotes an empty well. No peak distribution as libraries were of poor quality.

Sample 3 in figure 6 appears to have a smooth peak just after the lower marker. While the smooth and gradual curve is what you would expect to see, it is peaking at less than 200 bp. I would expect to see a curve that is greater than 200 base pairs with a mean length of 300-600 base pairs and a smooth profile. Additionally, many of the samples, bar 3 and 4, had faulty lower and upper markers.

Following this, cells were resorted via FACS once again. However, they now started with a high number of cells and gradually declined; as 10,000, 1000, 100, 10, 5, 1, 1, 1. This was done to see at what number of cells the protocol ceased to amplify the DNA. It also would provide a positive control as if the 10,000 did not work, there was likely an issue with the protocol or reagents.

Table 10. NanoPhotometer Purity Readings and Qubit Concentrations for each sample of attempt 4.

Sample	DNA Concentration (ng/µL)	A260/A280	A260/A230
Negative control	Concentration too low for qubit to detect.	2.0	1.6
Single cell 1A	Concentration too low for qubit to detect.	2.9	1.4
Single cell 1B	0.642	1.6	1.4
Single cell 1C	0.294	1.8	1.6
5 cells	0.116	2.18	1.4
10 cells	0.252	1.8	1.4
100 cells	0.286	1.7	1.8
1000 cells	Concentration too low for qubit to detect.	1.78	1.6
10,000 cells	5.52	1.98	1.6

The results from Table 10 showed the concentration of the negative control was too low to detect, as expected. However, single cell 1A also had too little DNA to detect, suggesting the DNA was lost at some point or degraded during the bisulfite conversion. Single cell 1B and 1C had expected concentrations - low but still detectable – however, the 5 cell, 10 cell, 100 cell and 1000 cell samples unexpectedly had less DNA present than single cells 1B and 1C.

Samples were once again run on the Agilent 2100 High-Sensitivity DNA Bioanalyzer to determine the fragment length and quality of the libraries, depicted in figure 7.

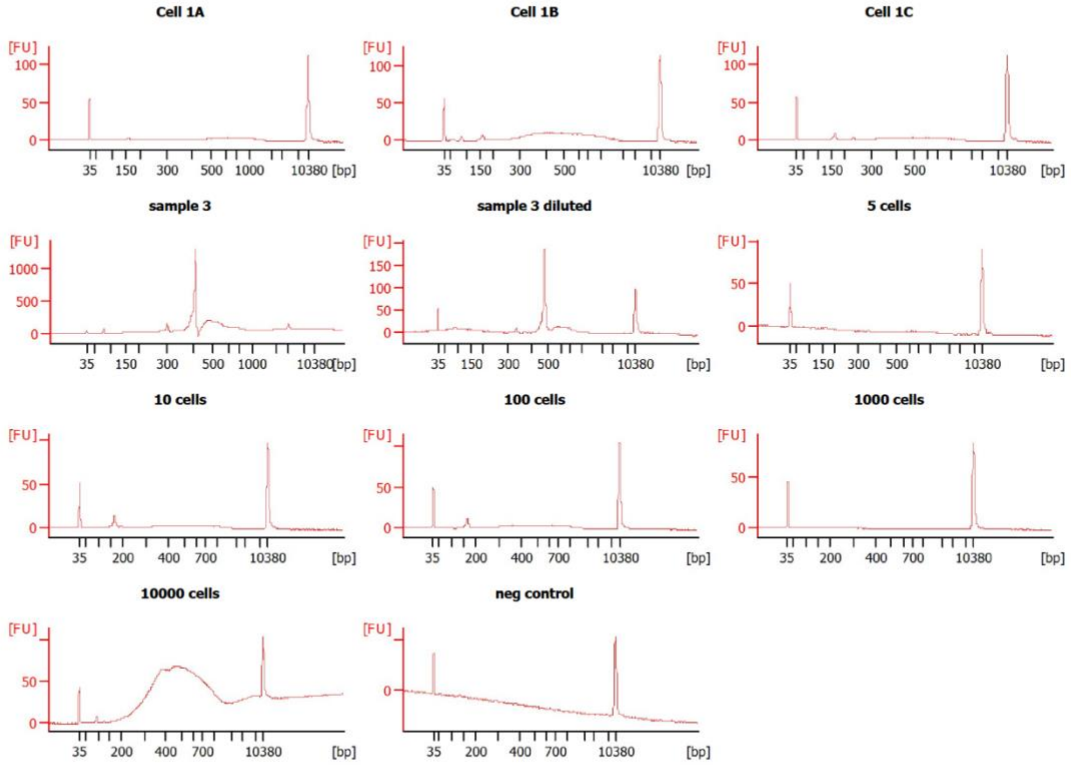

Figure 7 | Agilent 2100 Bioanalyzer profiles for single cell bisulfite libraries. The library profiles of HT29 cells post-whole genome bisulfite amplification. Graphics display libraries size distribution. Attempt 4. Samples 1A, 1B, 1C, "sample 3" and "sample 3 diluted" all contained single cells. Other graphs have number of cells present denoted above them. Lower marker = sharp peak at 35 bp, upper marker = sharp peak at 10380 bp. Negative control = nuclease free water. Positive control = 10,000 cell sample.

As a result of the smooth curve of sample 3 from the previous attempt (figure 6), despite its lower molecular fragment size, it was rerun. As shown in figure 7, upon rerunning it the library presented with a much sharper peak around the 400 – 500 base pair mark, rather than the smooth gradual peak observed in the last run. This was true for both the diluted sample 3 and non-diluted sample 3. The originally seen smooth curve may have been a result of impaired lower and upper markers in the Bioanalyzer run from attempt 3. Despite this, single cell 1B showed a promising potential result, with its library average fragment size ranging from 300 – 600 bp, with a smooth gradual peak. Although it is not a high peak, it is synonymous to the embryonic stem cell examples of a high-quality library on the 2100 Bioanalyzer provided in the protocol. Some small peaks were observed around the 150 bp mark as in previous attempts, indicative of adapters, yet these were very short peaks in comparison.

The 10,000-cell library of figure 7 had a relatively smooth curve between the 300 bp and 700 bp region, indicative of a high-quality library. While single cell 1B also held potential, it was unexpected that the protocol would drop off after the 10,000-cell sample, considering some bulk sequencing methods can capture 1000 cells. I had expected it to drop off potentially around the 10 – 100 cell amounts.

Following the figure 7 bioanalyser results which showed the protocol dropped out at a much higher cell count than anticipated, the amount of transfer steps was once again reduced in an effort to minimise the opportunities for loss of DNA. As well as this, troubleshooting for the protocol suggested increasing PCR cycles if cells have been brought up from frozen, as my HT29 cells had. Therefore, an additional 2 cycles of PCR was performed.

Table 11. NanoPhotometer Purity Readings and Qubit Concentrations for each sample of attempt 5.

Sample	DNA Concentration (ng/μL)	A260/A280	A260/A230
Negative control	Concentration too low for qubit to detect.	1.8	1.6
Single cell 1A	1.59	1.8	1.8
Single cell 1B	1.07	1.9	1.6
Single cell 1C	0.52	1.8	1.4
5 cells	3.34	1.8	1.4
10 cells	1.08	1.9	1.4
100 cells	0.676	1.9	1.4
1000 cells	4.28	1.8	1.6
10,000 cells	9.08	1.8	2.0

Table 11 showed libraries with DNA concentrations more consistent with their original cell count when compared to the previous attempt, in table 10. As well as this, the purity readings were more consistent. The A260/A280 suggested the libraries were relatively pure, while the A260/A230 readings were slightly low in some samples. This could be a result of using an incorrect blank sample which did not match the pH and ionic strength of the samples.

These libraries were then run on the Agilent 2100 Bioanalyzer, as well as single cell 1B from attempt 4's libraries which showed potential. Single cell sample 3 from attempt 3 was also once again ran to confirm if it was a potential high-quality library or not.

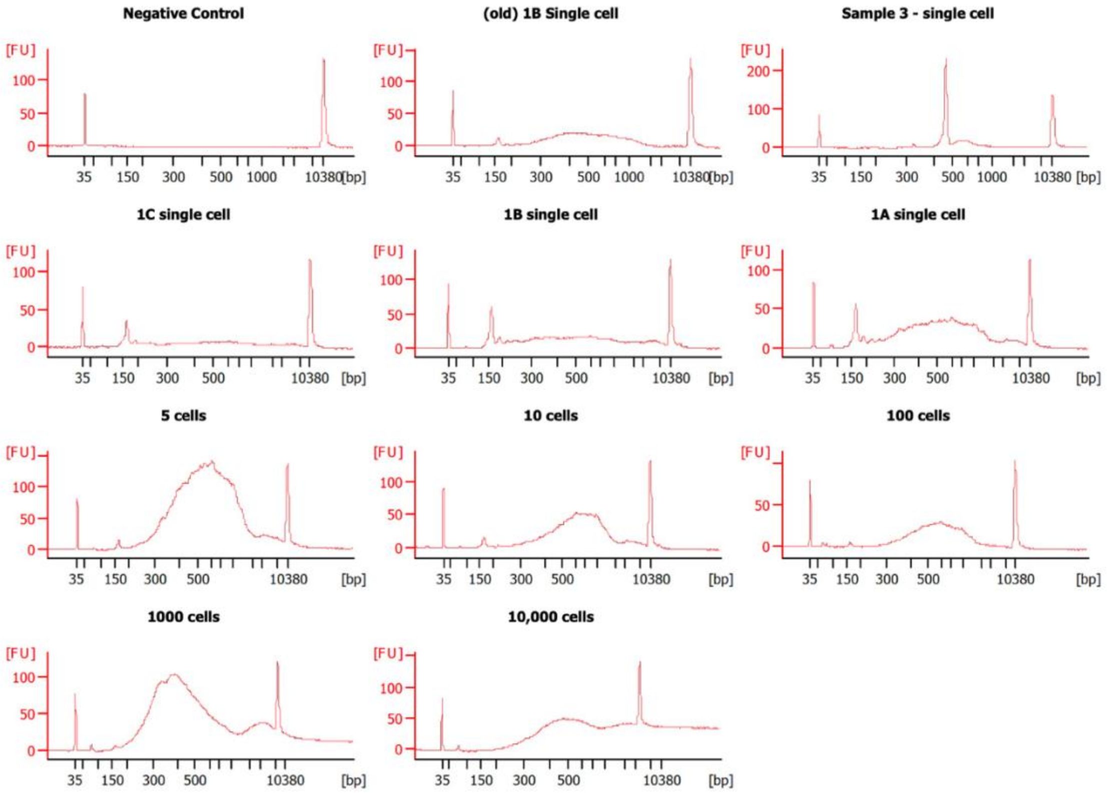

Figure 8 | Agilent 2100 Bioanalyzer profiles for single cell bisulfite libraries. The library profiles of HT29 cells post-whole genome bisulfite amplification. Graphics show libraries size distribution. Attempt 5. Lower marker = sharp peak at 35 bp, upper marker = sharp peak at 10380 bp. Peak library length typically 300 – 600 bp. Negative control = nuclease free water. Positive control = 10,000 cell sample.

Following the previously stated adjustments to the protocol, figure 8 showed the 1000 cells, 100 cells, 10 cells, 5 cells, single cell 1A, 1B and single cell 1B (old) had promising bioanalyzer results. That being, smooth profiles between the 300-600 bp mark. Most profiles had sharp peaks at the 150 bp mark, indicative of residual adapters. These can be trimmed later in the pre-processing step after sequencing. The negative control only showed peaks at the lower and upper base pair markers. As such, it is unlikely any peaks in other samples would be a result of contamination. This excludes sample 3 and (old) 1B as they were prepared in earlier rounds of the protocol.

Following these results, the eligible samples: 1B (old), 1B, 1A, 5 cells, 10 cells, 100 cells and 1000 cells, went on to be sequenced using iSeq.

This chapter has described the methods used to achieve the first aim of my thesis; to optimise and implement the single cell bisulfite sequencing protocol on a human sorted CRC cell line. To summarise, a combination of adjustments were introduced to the protocol after each attempt, which allowed for the acquisition of high-quality libraries.

3.1.3 Low coverage next generation sequencing (using Illumina iSeq) of Single Cell Bisulfite Sequencing Cell Libraries.

Due to time constraints as a result of COVID, I was unable to get the sequences from my libraries back in time to pre-process them.

3.2 Bioinformatic Analysis of CRC Single Cell Patient Data

Upon completion of the first aim, the second aim to bioinformatically analyse the single cell data to uncover heterogeneity that would otherwise be missed in bulk population sequencing could not be completed. This was due to COVID-19 delays preventing the sequencing and pre-processing of my samples in time. Therefore, instead of bioinformatically analysing my own data, I analysed a publicly available data set from a group which used single cell multi-omic methods on CRC patients [1]. This group used the method of scTrio seq – a method which captures the genome, DNA methylome and transcriptome of a single cell. Data from 12 CRC patients was obtained. This data was used to demonstrate what would have been done had I had time to fully sequence and process my own data.

All CRC patients were stage III or stage IV. Where possible, they obtained single cells from each of the patients' adjacent normal colon tissue (NC), primary tumour (PT), lymph node metastases (LN), liver metastases (ML) and liver metastases post chemotherapy treatment (MP). Different areas of the sampling regions were sequenced, depicted numerically; for example, PT1 and PT2 are different areas of the primary tumour. Single cells from all regions were only obtained for patient CRC01, therefore I decided to use this data set in my bioinformatic analyses. In the interest of limited time due to COVID lockdown, analysing all cells from the patient let alone the entire dataset was unattainable. Considering this, I selected the top 10 cells from each sampling region based on the cells with the top mapping efficiency/genome coverage to demonstrate DNAme heterogeneity between CRC cells. The cells chosen for this were provided in table 9. I also chose a relatively even spread of different areas of the sampling region, to evenly represent the sampling region and investigate intra-tumour heterogeneity. The dataset was filtered to only contain methylation status in CpG contexts, as these are typically in promoter regions and influencing gene expression. This data set also provided the corresponding transcriptome for cells that they were able to obtain both the methylome and transcriptome for.

3.2.1 Investigating Global Methylation of CRC Single Cell Patient Data

As stated earlier, the regulation of the cancer genome as a whole is grossly dysregulated. In the following section, I explore patterns between single cell DNAme on a global and chromosomal scale from different tissue regions.

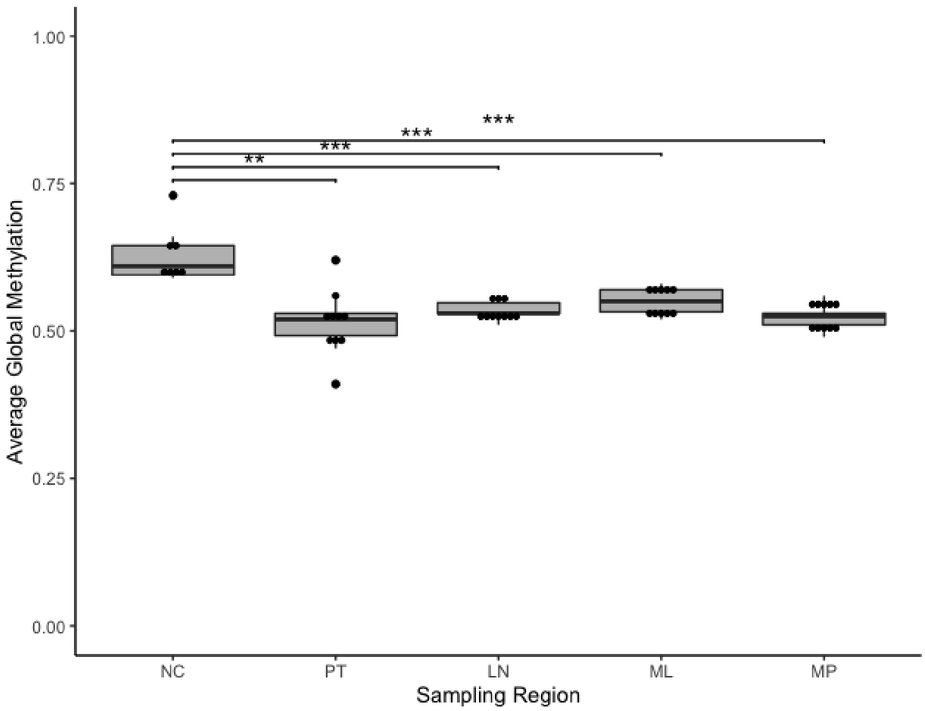

Figure 9 | Average global methylation of each single cell from each sampling region. NC = normal adjacent colon tissue, PT = primary tumour, LN = lymph node metastases, ML = liver metastases, MP = liver metastases post-treatment. Methylation scaled 0-1, 0 being no methylation and 1 being full methylation across CpG sites. Statistical significance was obtained using a one-way ANOVA followed by a Tukey's multiple comparison test. ** = p-value < 0.01. *** = p-value < 0.001.

Both the primary tumours cells as well as the distant metastases displayed significantly less methylation globally compared to the matched normal adjacent colon tissue in figure 9 (p-value <0.01). Interestingly, we see the primary tumour (PT) cells tend to have more variation of global methylation between them compared to distant metastases cells. One cell had as little as 0.4 global methylation while its highest was around 0.6, surpassing the normal adjacent colon (NC) samples median global methylation. The distant metastases: lymph node (LN), liver metastases (ML) and liver metastases post-treatment (MP), showed very little variance and had very similar median global methylation rates, around 0.5.

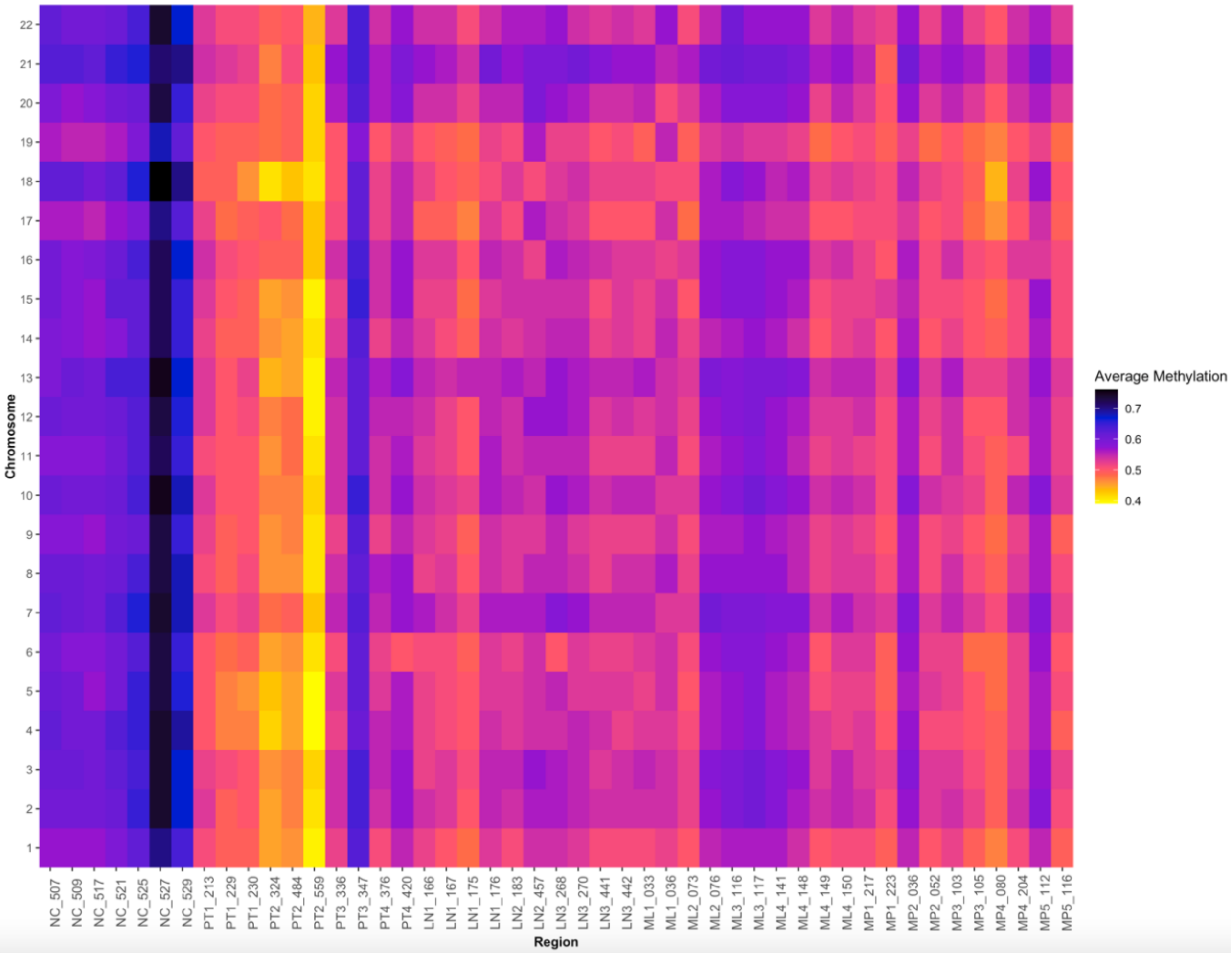

Figure 10 | Heatmap of average chromosomal methylation for each cell from each region. X-axis contains each cell. NC = normal adjacent colon tissue, PT = primary tumour, LN = lymph node metastases, ML = liver metastases, MP = liver metastases post-treatment. Number following region depicts different areas of that region being sampled. Number following underscore is number of that single cell. Colour depicts average methylation of single cell at each chromosome.

NC single cells clearly stand out in figure 10 compared to the carcinogenic regions as having higher methylation of their chromosomes. Curiously, some cells from different regions of the ML had methylation levels almost matching that of NC tissue (figure 10). Also, the majority of single cells from the PT tissue appear to have relatively more hypomethylated chromosomes compared to distant metastases tissues (LN, ML and MP). Inverse to this, another single cell from the PT tissue had high methylation levels across chromosomes, matching that of NC tissue (figure 10).

In terms of chromosomal trends of hypomethylation, there appears to be no obvious single chromosome which is more susceptible to hypomethylation in this group of single cells (figure 1-).

These results were clearly consistent with current literature that observes a global hypomethylation shift in neoplastic cells from normal adjacent colon tissue, even from early stages. This section additionally highlighted that while this phenomenon is observed, there is different global methylation patterns between cells both intra- and inter- tumourally.

3.2.2 Investigating Differentially Methylated Genes in CRC Patient Single Cells.

Having explored the single cell variation of methylation states on a larger scale, let us now consider the methylation states of commonly aberrantly methylated genes in CRC between single cells. As alluded to earlier in this thesis, focal hypermethylation of genes is a common occurrence in neoplastic cells, as is loss of methylation on imprinted genes. This section explores both models to investigate heterogeneity both inter- and intra-tumourally.

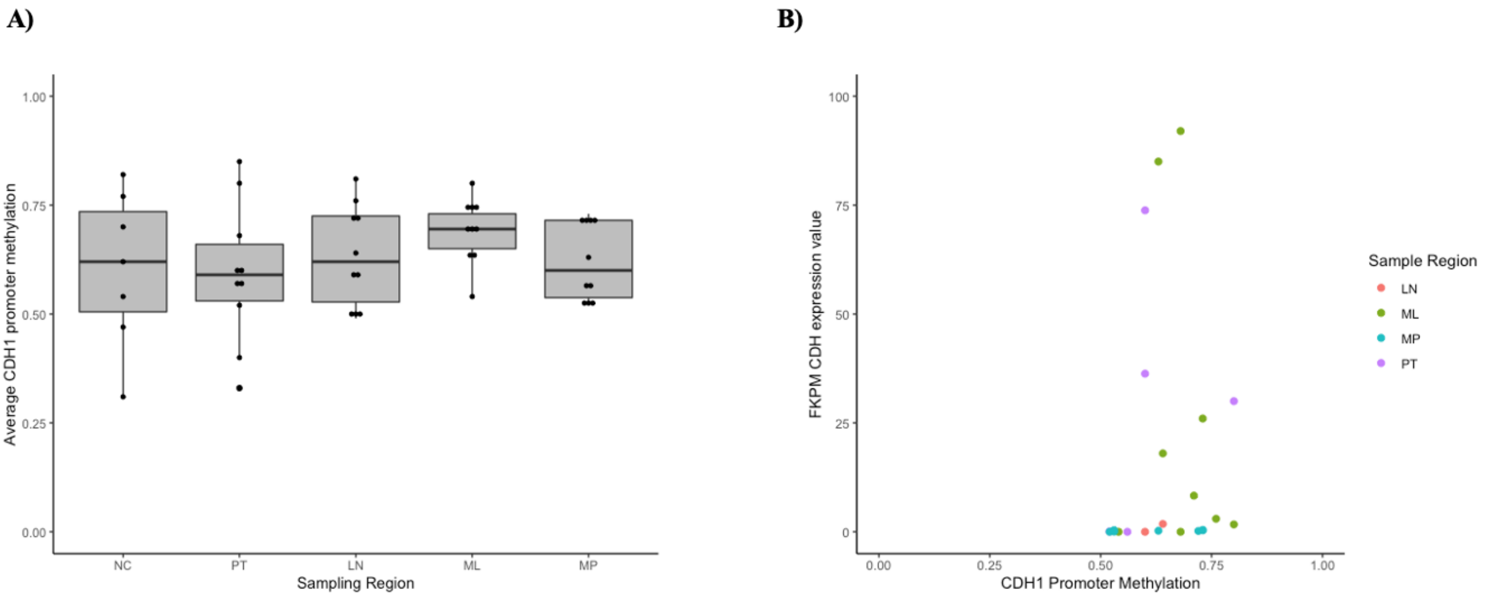

Figure 11 | CDH1 Promoter Methylation and Expression Data from Single Cells of each Sampling Region. A) Average CDH1 promoter methylation of each cell from each sampling region. **B)** Expression of CDH1 transcript vs. promoter methylation of cells with available expression data. NC = normal adjacent colon tissue, PT = primary tumour, LN = lymph node metastases, ML = liver metastases, MP = liver metastases post-treatment. FPKM = Fragments per kilobase of exon per million mapped fragments. Methylation ranges from 0-1, 0 being no methylation and 1 being full methylation of CpG sites across CDH1 promoter. Statistical significance was obtained using a one-way ANOVA followed by a Tukey's multiple comparison test, found no statistically significant differences between the sampling regions.

The E-cadherin gene, CDH1, as described earlier as typically being a point of focal hypermethylation was explored (Figure 11A). The CDH1 promoter region was defined on ENSEMBL (ENSG00000039068) using Hg19. There was no significant difference in promoter hypermethylation between normal tissue, the primary tumour, or its distant metastases. Normal tissue and primary tumour samples appeared to have the greatest variation of CDH1 promoter methylation per cell. The median CDH1 promoter methylation of the single cells from ML was the highest of all the regions (0.0.7), while the median of single cells from the primary tumour was the lowest, at 0.55. Although, these were not statistically significant. The PT region had one cell with very low methylation of 0.3 at the CDH1 promoter, yet another PT single cell had the highest methylation of all the cells, at 0.8.

The cells which had corresponding expression data available were analysed for expression of the CDH1 transcript, which included cells from the PT, LN, ML and MP. The only single cells with CDH1 expression were from sample regions PT and ML. All CDH1 promoter methylation levels fall between 0.5 to 0.8, with no positive or negative correlation to the amount of CDH1 transcript being expressed (Pearson's correlation co-efficient = 0.07). In fact, the highest expression of the CDH1 transcript was in ML samples, which had the highest median CDH1 promoter methylation in figure 11B.

Some ML cells had 100-fold less CDH1 transcript present compared to other ML cells (Figure 11B). Similarly, some PT cells had 0 transcript present, while another had 70 FPKM.

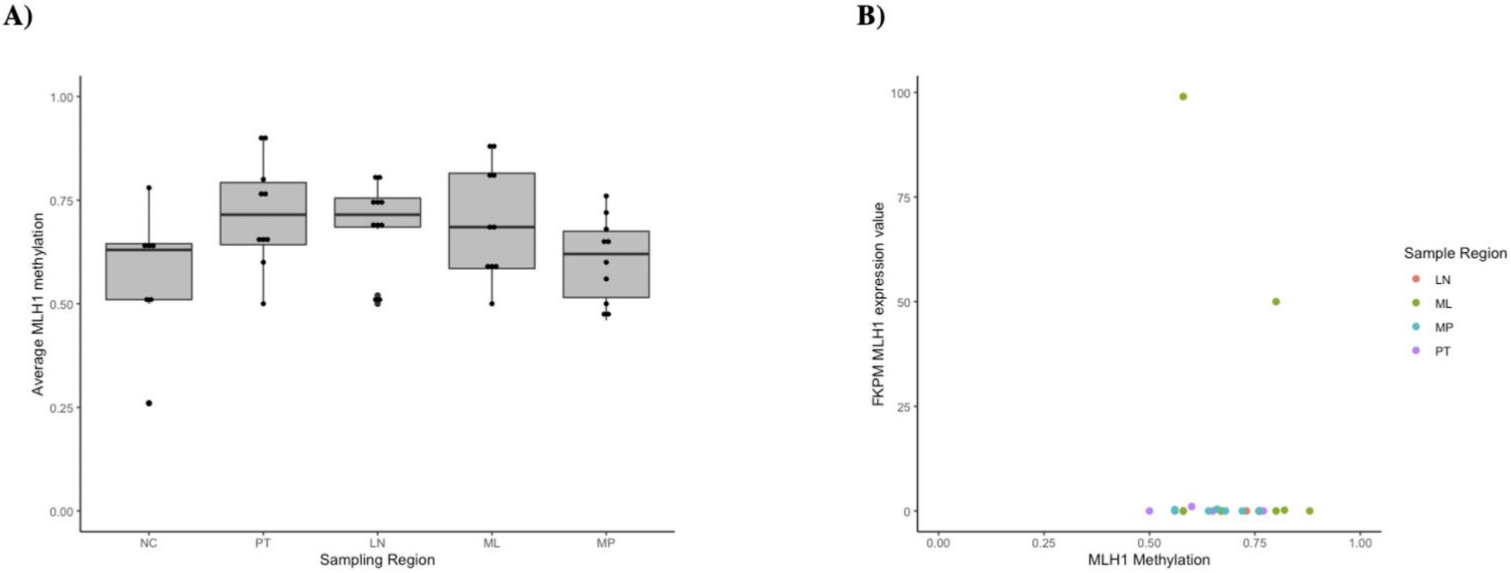

Figure 12 | MLH1 Promoter Methylation and Expression Data from Single Cells of each Sampling Region. A) Average MLH1 promoter methylation of each cell from each sampling region. **B)** Expression of MLH1 transcript vs. promoter methylation of cells with available expression data. NC = normal adjacent colon tissue, PT = primary tumour, LN = lymph node metastases, ML = liver metastases, MP = liver metastases post-treatment. FPKM = Fragments per kilobase of exon per million mapped fragments. Methylation scaled 0-1, 0 being no methylation and 1 being full methylation of CpG sites across the MLH1 promoter. Statistical significance was obtained using a one-way ANOVA followed by a Tukey's multiple comparison test, showing no statistically significant differences between the sampling regions.

All LN and MP cells with expression data available had little to no expression of CDH1 transcript, irrespective of the CDH1 promoter methylation status.

The MLH1 promoter region was defined using ENSEMBL (ENSG00000076242) aligned to Hg19. The MLH1 promoter returned no statistically significant differences in the level of methylation (figure 12A). Although not statistically significant, the PT and distant metastases tended to have higher median promoter methylation compared to NC cells. NC and MP cells had a similar median MLH1 promoter methylation of 0.6, while PT, LN and ML had similar median MLH1 methylation of 0.7. Two single cells in the PT region had the highest MLH1 promoter methylation, at 0.8. However, some cells in the PT region also had low methylation of this promoter at 0.5, synonymous with NC tissue cells.

One cell from the ML tissue had 50 FPKM of MLH1 transcript present, despite having a relatively high level of promoter methylation, at 0.75 (figure 12B). Another ML cell had relatively high MLH1 transcript present of 100 FPKM, with a methylation level similar to that seen in NC cells of 0.55. Majority of cells with MLH1 expression data available had little to no expression of MLH1 transcript, irrespective of the MLH1 promoter methylation. This lack of correlation was confirmed as the Pearson’s correlation co-efficient was -0.09.

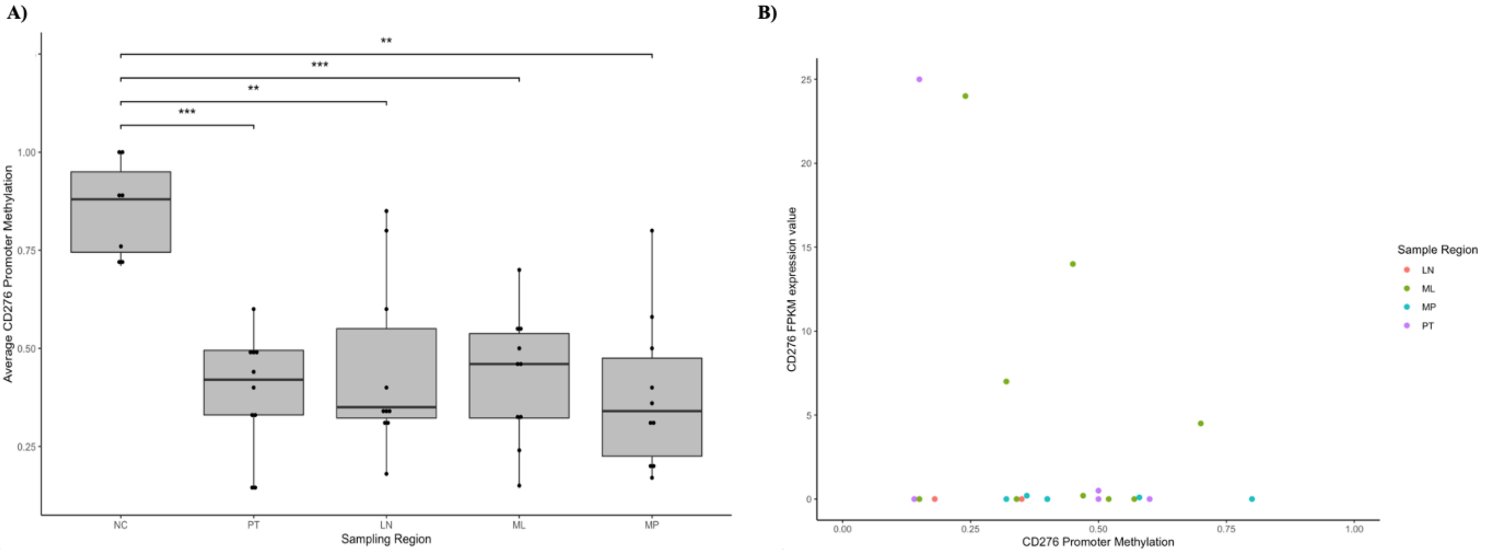

Figure 13 | CD276 promoter methylation and expression data per single cell in each region. A) Average CD276 promoter methylation of each cell from each sampling region. **B)** Expression of CD276 transcript vs. promoter methylation of cells with available expression data. NC = normal adjacent colon tissue, PT = primary tumour, LN = lymph node metastases, ML = liver metastases, MP = liver metastases post-treatment. FPKM = Fragments per kilobase of exon per million mapped fragments. Methylation scaled 0-1, 0 being no methylation and 1 being full methylation of CpG sites across CD276 promoter. Statistical significance was obtained using a one-way ANOVA followed by a Tukey's multiple comparison test, ** = p-value <0.01,

The CD276 promoter region was defined using ENSEMBL (ENSG00000103855) aligned to Hg19. The primary tumour and distant metastases single cells were significantly hypomethylated in comparison to the normal adjacent colon tissue with a decrease in CD276 promoter median methylation of 40% (figure 13A) (p-value <0.01, p-value<0.001). The median CD276 promoter methylation of single cells from tumour samples ranged from 0.3 to 0.4. Meanwhile, the median CD276 promoter methylation of NC cells was up at 0.85. There were still observed outliers, particularly in the LN and MP groups. Two LN single cells had CD276 promoter methylation around 0.85, while cells from this same group had methylation levels as low as 0.2. The MP group also had a single cell with methylation as high as 0.75, as well as cells with a methylation of 0.2.

A slight correlation between CD276 methylation and CD276 transcript expression was observed with a Pearson's correlation co-efficient of 0.33.One primary tumour cell had CD276 promoter methylation of 0.15, with a corresponding CD276 transcript expression of 25 FPKM. Four cells from the ML tissue had CD276 promoter methylation of 0.2, 0.3, 0.45 and 0.7 and corresponding CD276 transcript expression of 24, 7, 14 5 FPKM, respectively (figure 13B).

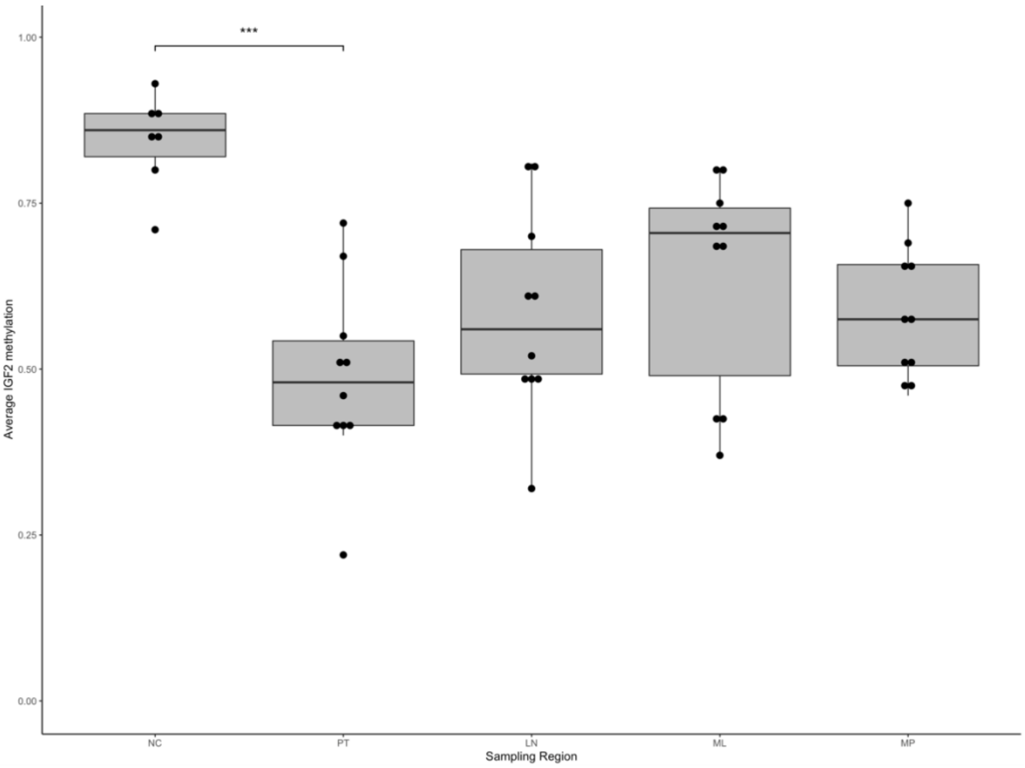

Figure 14 | IGF2 methylation status per cell in each region. NC = normal adjacent colon tissue, PT = primary tumour, LN = lymph node metastases, ML = liver metastases, MP = liver metastases post-treatment. Methylation scaled 0-1, 0 being no methylation and 1 being full methylation of CpG sites across IGF2. Statistical significance was obtained using a one-way ANOVA followed by a Tukey's multiple comparison test. *** = p-value < 0.001.

The methylation of the imprinted gene IGF2 was analysed in these cells (figure 14). This gene was defined using ENSEMBL (ENSG00000167244) aligned against Hg19. IGF2 expression data was not included in figure 14 as all single cells with expression data available displayed little to no IGF2 expression, despite the difference in methylation between normal adjacent tissue and primary tumour tissue. Single cells from the PT region showed a significant decrease in median methylation of IGF2 compared to the NC single cell median methylation, with the single cells of PT methylation median at 0.5 and NC single cells median at 0.8. The NC tissue had relatively little variation of median IGF2 methylation between cells, ranging from 0.7 to 0.85. Whereas the PT had single cells median methylation ranging from as low as 0.22 to as high as 0.8, just below the NC median methylation of IGF2. Interestingly, there was so statistically significant difference between NC cells and the distant metastases. Single cells from the PT sample region had the lowest median of IGF2 methylation amongst all the tissues at 0.5. One PT cell had substantially low IGF2 methylation (methylation value = 0.25), which was the lowest methylation of IGF2 amongst the single cells.

This chapter highlighted the vast amount of heterogeneity in both global and focal methylation contexts between single cells. Three key cancer phenomena were explored in the context of single cells. While the overall results are generally consistent with that seen in bulk sequencing, the variation among tissue samples between single cells is extensive.

Chapter 4: Discussion

In the chapter that follows, I will present the principal findings from the implementation of the scBS-seq protocol, as well as the findings from the public scTrio-seq dataset and their relation to current literature. As well as this, limitations encountered throughout the project are discussed, followed by future directions.

4.1 Generating high-quality single cell libraries

Several aspects of a scBS-seq protocol were fine-tuned to improve the quality of the single cell libraries. This included: reducing contamination of the libraries, improving handling and enhancement of selection and amplification. The following section details the steps taken to optimise the scBS-seq protocol.

The scBS-seq protocol is designed to amplify as little as 6pg of DNA, therefore even a small amount of contamination will easily become amplified. This makes the potential for contamination far higher than usual when using bisulfite sequencing protocols. I encountered this contamination amplification in my first attempt. The sharp random peaks and lack of lower and upper marker peaks in figure 4 suggested contamination of my samples. To address this problem, in the following attempt I increased the UV-irradiation time of equipment prior to the experiment from one to two hours, frequently changed gloves and cleaned anything that entered the cell-culture hood. Following attempts also had a negative control added to ensure there was no contamination. As a result of these adjustments, figure 5, 7 and 8 showed improvements in the level of contamination, with much less noisy spikey bioanalyser profiles and much clearer lower and upper marker peaks. While there were still occasional peaks in figures 5, 7 and 8 following these changes, they were around the 150 bp mark and short which is indicative of residual adaptors rather than contamination.

Following the reduction of contamination, the next aim was to optimise the handling of the initial single cell DNA considering it involves treating 6 pg of DNA with harsh sodium bisulfite followed by amplification. This means, there are 5 steps before preamplification where the 6 pg of DNA could be lost. Figures 5 and 7 presented with libraries that only had lower and upper marker peaks, therefore I had lost the DNA at some point throughout the protocol. To address this, I removed transfer steps where possible. There were extra transfer steps in the protocol typically during bead purification. Therefore, I omitted two of the transfer steps between different tubes to minimise the chance of DNA loss due to remaining in the previous tubes. While figure 7 returned one single cell library (1B) which had a smooth curve between the lower and upper markers between 300 - 700 bp, no other samples displayed high-quality libraries (only had lower and upper marker peaks) apart from the 10,000-cell sample. This was surprising, as I had expected to see at least samples with 100 cells and upwards produce libraries. Although, this was an improvement to all previous attempts as I now had two libraries with smooth curves between 300 – 700 bp. As a result of this, I explored other areas of the protocol which may increase the yield from samples.

As a result of the protocol not successfully amplifying below 1,000 cells in figure 7, I investigated troubleshooting advice for the protocol from the author. The author suggested increasing the number of final PCR cycles if tissue had been kept frozen for a considerable amount of time. The author also suggested to increase the cycles if using human tissue, as the protocol was implemented on a mouse genome which is smaller than the human genome. Therefore, I increased the number of final PCR cycles from 15 to 17. The author also suggested ensuring the AMPureXP ratio was a 0.8x ratio, when I had been using a 0.7x ratio. Therefore, I also increased from 0.7x to 0.8x ratio in the next attempt. This would allow for better size selection of fragments between 300 – 700 bp. Following the adjustment AMPureXP ratio, increasing the number of final PCR amplifications, reducing transfer steps and gaining familiarity with the protocol, I was able to successfully amplify the samples containing single cells (1B old, 1B and 1A) 5 cells, 10 cells, 100 cells, 1000 cells and 10,000 cells. This was shown in figure 8 as the above samples had smooth curves between the lower and upper base pair markers which were between 300 – 700 bp. To confirm the amplification observed was in fact my HT29 cell line, the libraries were sequenced using iSeq. A PhiX spike in was included due to samples being of low-diversity as a result of the BS-treated DNA. Unfortunately, the sequences were not retrieved back in time for pre-processing and adding to this thesis.

In conclusion, taking steps to minimise opportunities for DNA loss, appropriate size selection, and increasing PCR amplification steps lead to the implementation of a scBS-seq protocol [81] on the human-sorted CRC cell line, HT29.

4.1.1 Limitations and future directions for generating high-quality single cell libraries

One major limitation that is currently unavoidable with bisulfite sequencing, is the amount of DNA and therefore information lost in the bisulfite conversion step. This limitation is amplified in single cell bisulfite sequencing, as there are only the two strands of DNA so once that information is lost, it cannot be supplemented by other DNA. While there are other methods for obtaining methylation data which are less harsh on the DNA, such as scMSRE described earlier which uses restriction enzyme digestion, this limits the number of CpG sites and presents biases.

In the future, to increase the final concentration of the libraries, I could increase the number of final PCR cycles again. This however may introduce PCR amplification bias; therefore, I would decrease the final elution amount to obtain libraries with higher concentrations. In future I would use 15μl of nuclease free water to elute the DNA, which is still a sufficient amount for the analyses using qubit, nanodrop, Agilent HS DNA bioanalyser and MiSeq/HiSeq to be performed.

In order to see at which number of cells the protocol would successfully create high quality libraries, I included varying amounts of cells and a negative control. The acquisition of high-quality single cell libraries was achieved finally towards the end of the year, if I was able to attempt the protocol again I would have replaced the higher cell number sample with other controls such as lambda phage as an unmethylated control, allowing for assessment of bisulfite conversion rates. Using an unmethylated lambda phage control checks that the C/T conversion rate is greater than 97%. I would have also added in negative controls at different points throughout the protocol, such as after each amplification to ensure there was no new contamination introduced after adding reagents at different steps.

Additionally, the bioanalyzer profiles of the 1000 cell and 10,000 cell samples produced abnormal bioanalyser profiles towards the upper markers. This is expected, as when performing this protocol on samples with more than 1000 cells, the protocol suggested omitting the extra cycles of first strand synthesis, so it was only performed once [3]. The protocol also suggested reducing the number of final PCR amplifications. In future, this could be done to get tidier overall HS DNA bioanalyser profiles of higher numbers of cells, however the protocol was kept consistently between samples to see at what number of cells it dropped off when I was not receiving any libraries.

Another limitation of my implementation of the scBS protocol was the small number of cells that it can be performed on currently. Despite the success of the protocol, implementing it on as few as 9 samples took three days, with the first two days involving 10+ hours of lab work. While this low sample number was appropriate for my aim, to optimise and implement the protocol, future implementation on patient samples would require this to be on a much larger scale to investigate any clinically relevant results. However, this protocol is sufficient for the analysis of cells such as circulating tumour cells (CTCs), where very few are obtained from the blood at any given time.

As a result of COVID delays elsewhere in the world even before the lockdown, getting certain reagents which were critical for the protocol took extended periods of time, especially considering some were specialised sequences with alterations such as HPLC purifying. As well as this, everything including the PCR machine needed to be kept in the cell-culture hood throughout each day. The protocol took a total of 3 days to get bioanalyzer profiles stage, with the first two days taking 10+. In a lab which shares limited cell culture hoods, it was difficult to book the hood out three full days of the week too often.

4.2 Trends in methylation of single cells in CRC

The scTrio-seq dataset [1] from CRC patients was used to explore DNAme patterns of commonly aberrantly methylated genes involved in pathways that contribute to the initiation and progression of CRC. Global as well as focal methylation patterns were explored, substantiating widely accepted cancer phenomena while also diverging from others.

There are numerous exploited pathways which contribute to the development of CRC, commonly promoting genomic instability, activation of oncogenes and silencing of tumour suppressor genes. This analysis aimed to explore a range of these pathways; globally, imprinted genes, highly conserved genes, genes suspected to be involved in metastatic pathways, as well as DNA mismatch repair genes. This chapter aimed to investigate a wide scope of pathways known to contribute to carcinogenesis in CRC.

4.2.1 Global methylation trends in single cells in CRC

The universally accepted phenomenon in carcinogenic cells of global hypomethylation was highlighted clearly in my results. The global DNA methylation was stable between primary tumour cells and the metastasised cells with 50% global methylation, suggesting the global methylation status did not greatly influence the metastasis of cells. This is not unexpected as many studies have found global hypomethylation to be an early event in carcinogenesis, as early as in precancerous lesions [82]. This is reflected in my results where the major difference was observed between normal adjacent colon cells and tumour cells, with NC cells typically having upwards of 70% global methylation compared to 50% methylation in neoplastic cells. The emergence of single cell sequencing in the presence of this early onset phenomenon is instrumental in a clinical context. Being able to screen normal colon tissue for hypomethylation at a single cell level would allow for the identification of outlying globally hypomethylated cells, which may be beginning neoplastic transformation. The global hypomethylation of this small number of cells at early stages would otherwise be masked in bulk population sequencing.

Past studies have shown epigenetic alterations in tumours post-treatment which allows for drug-resistance of certain cells in the population [83, 84]. Despite this, there was no difference in global methylation between liver metastases cells and liver metastases cells post-treatment, although literature typically reports more focal epigenetic influences in drug tolerance rather than global [85-87].

The primary tumour displayed the greatest amount of intra-tumour heterogeneity in respect to global and chromosomal methylation per cell, with the lowest cell exhibiting 40% global methylation yet another with almost 70% global methylation. Cells from the same sampling region of the primary tumour (i.e., PT1) had, expectedly, the most similar chromosomal methylation. This was expected as cells from within the same region will be closer in lineage and are therefore more likely to be homogeneous. Interestingly, the PT2 region had cells with chromosomal methylation around 40%, compared to the PT3 region cells, which had a higher chromosomal methylation around 60-70%. This was able to show the contrast between cells from the same primary tumour tissue, highlighting intra-tumour heterogeneity which would have otherwise been masked and presented as an average methylation rate.

Interestingly, 5 single cells from the liver metastases displayed higher chromosomal methylation upwards of 60%, compared to 40-50% global methylation seen at the primary tumour, lymph node and even liver metastases post-treatment cells. This may be a result of the vastly different environments in which the different cancer cells are required to grow in. However, this finding contradicts current literature which found liver metastases tended to have a late increase of hypomethylation accompanied by an increase of proto-oncogenes suspected to be regulated by hypomethylated LINE1 elements [88, 89].

Previous literature has often described chromosomes more susceptible to chromosomal hypomethylation as well as chromosomal hypermethylation in CRC. Chromosomes 18 and 5 have been described to display greater hypermethylation due to carrying frequently hypermethylated genes[90]. Whereas chromosomes 22, 17 and 15 were described as typically hypomethylated due to carrying frequently hypomethylated genes [90]. My results were not consistent with this literature, as no chromosome displayed consistent hypomethylation or hypermethylation across tumour cells compared to normal adjacent colon cells.

In future, looking at the structural genomic alterations of the cancer genome as a result of this global hypomethylation could further explore the impact of demethylation beyond the scope of expression. Often, this global hypomethylation is more representative of repeat elements such as LINES and pericentromeric regions becoming demethylated and contributing to the structurally aberrant cancer genome, than hypomethylation of specific genes. Therefore, it would be interesting to compare the structural abnormalities of the cancer genome in the PT cell with the 0.25 methylation, compared to the other PT single cells with higher average global methylation and see if the chromosomal abnormalities reflected the lower global methylation.

The concept of global hypomethylation being a trait of cancerous cells was gained through the sequencing of bulk populations of tumour and normal tissue. My results were consistent with this literature but were able to add a layer of information which displayed the heterogeneity of this phenomenon both within and between tumours.

4.2.2 Focal hypermethylation of single cells in CRC

This section follows the findings from exploring genes which are commonly focally hypermethylated at their promoters in CRC. The presence of focal hypermethylation at certain gene promoters has been shown to act as biomarkers, alter gene expression and influence therapeutic response. This section discusses two commonly hypermethylated genes in CRC from my analysis, CDH1 and MLH1.

I expected to see an increase in CDH1 promoter methylation of single cells from neoplastic tissues compared to normal colon, considering hypermethylation of CDH1 and loss of E-cadherin has been reported frequently in CRC studies [47, 91, 92], with reports of as many as 93% of patients displaying CDH1 promoter methylation and decreased E-cadherin expression [48]. The methylation of the CDH1 promoter region in my results, however, were sporadic. There was no obvious hypermethylation of the famous tumour suppressor gene. The lack of significant differences between groups in the CDH1 promoter methylation could be a result of different definitions of the promoter region. Additionally, I had a small sample size of cells per group (10 per tumour group, 7 for normal colon tissue), meaning a larger scale analysis of cells may have displayed promoter hypermethylation overall. The difference between my data and literature may also reflect the very aim of single cell sequencing; not every cell is going to have CDH1 promoter methylation, but identifying cells which do through the use of single cell technologies could have predictive power of future metastasis [91].

The expression of the CDH1 transcript (E-cadherin) was still prevalent in >70% of primary tumour and liver metastases cells, therefore despite my results inconsistencies with the literature, the normal methylation of the CDH1 promoter correlates with E-cadherin still being expressed. On the contrary, some cells with 50% CDH1 promoter methylation had no expression of E-cadherin, but it has been shown that CDH1 is also controlled by other elements such as TWIST1 and SNAI1 [93]. Overall, there was no discernible correlation between CDH1 promoter methylation and CDH1 expression (Pearson's correlation co-efficient = 0.07), highlighting the importance of not inferring gene expression on account of promoter hypomethylation.

In an effort to explore another pathway which contributes to CRC initiation and progression, MLH1, a critical component of the DNA mismatch repair (MMR) system was analysed [94]. Previous literature surrounding MLH1 promoter methylation in CRC has been relatively conflicting, with some studies observing as little as 0.0% MLH1 promoter methylation [95] to as high as 66.9% [96]. Akin to the CDH1 promoter methylation, there was no significant differences between normal colon tissue and tumour tissue. The MLH1 promoter methylation of single cells fluctuated irrespective of the tissue they were from. While the data did not show significant differences in MLH1 promoter methylation between the neoplastic regions and normal tissue, the purpose of single cell sequencing is predominantly to recognise small populations or single cells which are outliers from the group. Acknowledging this, there were multiple cells from both the primary tumour tissue and liver metastases tissue which presented with hypermethylated MLH1 promoter regions (methylation = 0.85), which may render these cells more susceptible to microsatellite instability (MSI) [97], which could potentially give rise to a sub-population with that phenotype.

It should also be acknowledged that MLH1 promoter methylation is largely influenced by location of the tumour, gender, and age of the patient. Typically, older, female CRC patients with right sided primary tumour location have hypermethylated MLH1 promoters [98]. I would have expected to see greater MLH1 promoter hypermethylation, as the patient data I analysed was from a 50-year-old female, however with a left-sided primary site. One possible reason I did not see this hypermethylation, could reflect the patients BRAF/KRAS mutation status. MLH1 promoter methylation has shown evidence of being common in CRC patients with BRAF mutations [99], but less common in those with KRAS mutations [100]. The BRAF/KRAS status of this patient was not explored and therefore, the lack of MLH1 promoter methylation may have reflected this. This could be explored in the future however as the genome sequence was also recovered in the scTrio-seq method.

Albeit there was no observed difference in the MLH1 promoter methylation between normal colon and neoplastic tissues, the expression data is consistent with previous studies; with >90% of neoplastic cells showing no expression of MLH1 transcript. This means while the MLH1 expression is absent, it may not be a consequence of MLH1 promoter hypermethylation. Rather, as commonly reported, a genetic mutation inactivating MLH1 in CRC cells [101], which would explain the lack of correlation between promoter methylation and expression in my analysis (Pearson’s correlation co-efficient = -0.09).

While there were no statistically significant differences between normal adjacent colon tissue and carcinogenic tissue methylation of the above genes, intra-tumour heterogeneity was observed across all genes of interest between all tissue regions.

4.2.3 Focal Hypomethylation of Single Cells in CRC

This section explores the inverse of the previous section, discussing the results of single cell data of commonly hypomethylated genes in CRC.

The event of loss of imprinting (LOI) of imprinted genes has been thoroughly explored in cancer, with IGF2 being commonly studied [102]. I chose to explore this gene as it is a selective marker of CRC progression and staging [52], hence using single cell sequencing to identify small sub-populations of cells which may harbour such selective markers has relevant clinical implications. My analysis showed a significant drop in IGF2 methylation from a 85% median of normal colon tissue cells to a median of 50% in primary tumour cells (p-value < 0.001), consistent with published literature [103, 104]. Importantly, one cell displayed very low IGF2 methylation at 0.25, which has been shown to be associated with metastasis and correlated with high mortality [105]. In a clinical context, identifying the cells such as those with low methylation in the PT, LN and ML tissues may have prognostic significance.

Despite the fact there was no expression of IGF2 despite a 35% decrease in methylation from normal tissue cells to primary tumour cells, it has been shown that LOI of IGF2 acts to increase expression of other genes involved in cellular proliferation, such as MCM5, MCM3, CDC6, LIG1 and CCNE1 [106]. This may explain why some studies have found a correlation between IGF2 LOI and increased IGF2 expression [107], yet a paradoxical expression of IGF2 has also been observed, where LOI of IGF2 in the tumour was associated with a decrease in IGF2 transcript compared to normal tissues [108]. LOI of IGF2 has also been associated with global chromatin instability, which may suggest the significant (p-value < 0.001) 35% median methylation decrease contributes to the cancer genome in this way, rather than increased IGF2 expression [109]. IGF2 is also regulated by a DMR upstream of another imprinted gene, H19, which acts to repress it [110]. Therefore, IGF2 may have lost imprinting but the regulatory DMR upstream of H19 may not be hypermethylated, therefore still repressing IGF2 expression. In future, the methylation status of the DMR upstream of

H19 could be investigated [111]. Not only this, but previous studies have found IGF2 transcription is controlled by several promoters [112]. Essentially, there are a number of reasons that may explain why no increase in IGF2 expression was observed in my data despite the significant decrease in methylation (p-value < 0.001). Using single cell sequencing to investigate the LOI trend in cancer is beneficial as contamination of tumour tissues with normal cells is likely to result in an underestimation of LOI.

The cells from normal colon tissue had relatively homogeneous IGF2 methylation, with all cells bar one having median IGF2 methylation between 0.8 and 0.9. However, the primary tumour and its distant metastases cells showed greater heterogeneity between the single cells, having a 55% difference in median IGF2 methylation between their lowest methylated cells and highest methylation. This calls attention back to the highly heterogeneous nature of CRC, proving this heterogeneity is not observed in normal colon tissue and therefore it is not reflective of the tumours tissue of origin.

Another commonly exploited normal biological pathway in CRC are immune checkpoints. The significant decrease in the methylation of CD276 promoter region between normal colon tissue and neoplastic tissues, is synonymous with recently published studies [113, 114]. CD276 is of interest as it has immunosuppressant functions [115] as well as aiding in metabolic reprogramming of cancer cells [116]. While there was a statistically significant drop in CD276 methylation of neoplastic tissues compared to normal colon (p-value < 0.01), most cells did not have high expression of CD276. There was evidence of a slight correlation between methylation of CD276 and CD276 transcript expression, with a Pearson’s correlation coefficient of 0.33. Moreover, considering CD276’s immunosuppressant functions, I would have expected to see a higher expression of CD276 in the lymph nodes considering these cells were under more stress from the immune system. Of note however, one cell had very low methylation level of 0.18 with a corresponding high CD276 mRNA expression, of 25 FPKM. Cells with these properties have been shown to be associated with poor prognosis [113], and this would have otherwise been masked had it been sequenced with the bulk population. This is also clinically relevant as CAR-T therapy targeting CD276 is currently being tested as a potential therapeutic target [117], however this single cell data contained single cells which had high methylation of CD276 and little corresponding expression, meaning these cells would be resistant to therapy. As a repercussion, therapy-resistant cells would persist and consequently result in relapse. While this data focused on CD276, it highlights the principle of small sub-populations/single cells being non-responsive to certain therapies due to heterogeneity amongst the population. Additionally, the drop in methylation of the promoter region could potentially be used as a biomarker, as it appears to occur even in the primary tumour.

Both CD276 and IGF2 showed significant decreases of methylation from normal adjacent tissue cells to carcinogenic cells. Similar to observations of focal hypermethylation and global trends, the cells which were from neoplastic tissues of the patients exhibited far more heterogeneity compared to cells from the matched colon healthy tissue.

4.2.4 Limitations and Future Directions for Analysing Single Cell CRC Patient Data

Multiple technical limitations were encountered while analysing this dataset, owing mostly to the nature of the data being obtained from single cell omics methods. The below section traverses these limitations, followed by suggestions for potential future directions that could be taken with this dataset to augment our current understanding of the cancer genome and its regulation.

One major limitation to my single cell data analysis was the inconsistent number of CpG sites in promoter regions of my target gene. CpG sites were filtered out which had a sequencing read depth of less than three, as I would have had little confidence in calling CpG sites with a lower sequencing read depth than this. Even a sequencing read depth of three is low compared to traditional bisulfite sequencing, however due to the nature of single cell sequencing it is not possible to get read depths matching those generated in bulk population sequencing, currently recommended to have a 30X coverage [118]. Allele specific and strand specific differences in methylation would also be difficult to detect, as the probability of covering both alleles and strands is unlikely for every single CpG genome wide. This was highlighted in my analyses as numerous CpG sites were absent in promoter regions of certain cells, but present in others. This resulted in a significantly smaller number of CpG sites within the promoter regions and consequently unequal ratios leading to skewed results. This created bias, as promoter regions with less CpG site data available was more likely to skew towards the extremes of no methylation, or complete methylation. One solution offered to this problem in single cell sequencing is to merge up to 10 homogeneous cells bioinformatically to reconstruct the methylome, however due to the heterogeneous nature of CRC, this solution cannot be applied.

Another limitation of the data analysis was the small sample size per region. This may explain why commonly observed cancer DNAme phenomena were not reflected in my analysis. While an increase in sample size would potentially change this, in the special circumstances of single cell sequencing, the purpose is to identify rare sub-populations of cells, in which case the argument of power in numbers is moot.

One constraint which was encountered when comparing expression and methylation relationships between tissue regions, was the absence of normal adjacent colon tissue single cell expression data. This meant there was no control to compare the tumour region single cell expression to, to see if there was in fact a significant difference between them. The relationship between methylation and expression in some neoplastic cells was still explored where possible and informative, it just meant comparisons to the normal colon tissue was not viable. There were also more liver metastases cells with expression data available, explaining why often ML cells had expression of interest genes and other regions appeared not to. This over- and under- representation of tissue regions meant making any inferences about tissues over or under expressing certain genes was limited. The reason some cells were not able to have their transcriptomes captured is a result of the nature of scTrio sequencing. Capturing the genome, methylome and transcriptome from one single cell is an intricate task and capturing all three from each cell is not always possible. Despite this, the methylation data was still valuable, as it could still uncover potential biomarkers at these regions or influence the cancer genome in ways other than regulation of gene expression.

The single cell data was also filtered to include only CpG sites for my analysis. This was done as my focus in my limited time frame was promoter regions of genes of interest, which are typically rich in CpG islands for regulation [119]. The potential role of non-CpG DNAme in cancers is becoming more appreciated, with numerous studies finding non-CpG DNAme implicated in the initiation and progression of cancer [120, 121]. On account of this, analysing single cell data with all methylated cytosines irrespective of their surrounding bases, could give more insight into these non-CpG functions in cancer.

Due to time constraints as previously mentioned, only select cells from one patient was analysed in this thesis. However, there were 12 patients with hundreds of single cells available for analysis. My analysis also focused primarily on methylation data from these cells, however there was also genomic and transcriptomic data available. In future, analysing the entire dataset would provide a much greater picture of the interactions contributing to the disorderly cancer genome. As highlighted often throughout this thesis, DNAme has versatile effects on the genome, expanding beyond the scope of gene expression. Integrating data from the genome, transcriptome and methylome would exhibit interrelation of the three, enhancing the ability to identify cellular populations, cellular trajectories, and lineage tracing. Not only this, exploring the entire dataset would also unveil inter-individual heterogeneity between the 12 CRC patients. All in all, the dataset used for my analysis provided an abundance of information with triple-omics data from a diverse range of patient samples at different stages, genders, and ages. It is a powerful dataset with copious domains that could be explored in the future, such as comparisons between individual's primary tumours of different CRC stages, or liver metastases post-treatment to observe any different responses to chemotherapy.

No cells from liver metastases post-treatment showed significant differences in methylation of the above genes compared to liver metastases cell pre-treatment. Research has shown the methylation status of certain genes render tumour cells drug resistant [122].Therefore, in future studies, considering methylomes of pre- and post- treatment liver metastases cells are available, examining the difference in methylation between ML and MP cells of these genes may be used clinically to identify any sub-populations present in the tumours which may be chemotherapy resistant.

Single cell sequencing can be used to build upon principals already established from bulk sequencing. Bulk sequencing found the differences in methylation between normal tissues and tumours which led to my choice of interest genes to investigate for this thesis. Single cell sequencing allows for greater elucidation of these mechanisms and has relevant clinical applications.

Chapter 5: Conclusion

Overall, I was able to successfully optimise and implement the generation of single cell bisulfite sequencing libraries on the HT29 CRC cell line. Three single cell libraries were of high enough quality for sequencing, as well as other small cell amounts of 5 and 10 which are not attainable by traditional bulk population bisulfite sequencing methods. Furthermore, I was able to analyse publicly available scTrio-seq data to compare methylomes of single cells from a tumour and its matched normal colon tissue. This confirmed my hypothesis that a vast amount of heterogeneity would be observed between cells of CRC tumours, compared to normal tissue. This was seen at both a global and focal level, providing awareness into the amount of information lost when these cells are coalesced and sequenced as one. This section also highlighted the complex layers of gene expression, and that DNAme does not always infer expression.

The implementation of this protocol has provided the base work for future experiments to be done possibly with patient samples. While the bioinformatic analyses covered a broad range of pathways exploited in CRC and explored currently accepted biological models of cancer genome dysregulation in a single cell context.

The initial plan of this project to address my hypothesis was to be achieved by optimising and implementing the scBS protocol, followed by analysis of my own data to see how greatly the cells methylomes varied from one another. Unfortunately, due to COVID-induced time restrictions, public single cell methylome data had to be used to address the second aim. Nonetheless, I was able to show that consistent with my hypothesis, heterogeneity both intra- and inter-tumourally in CRC cells is very prevalent and as such, single cell sequencing is a powerful tool in the field.

Chapter 6: References

1. Bian, S., et al., *Single-cell multiomics sequencing and analyses of human colorectal cancer.* Science, 2018. **362**(6418): p. 1060-1063.
2. Karemaker, I.D. and M. Vermeulen, *Single-Cell DNA Methylation Profiling: Technologies and Biological Applications.* Trends in Biotechnology, 2018. **36**(9): p. 952-965.
3. Clark, S.J., et al., *Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq).* Nat Protoc, 2017. **12**(3): p. 534-547.
4. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
5. Eccleston, A., et al., *Epigenetics.* Nature, 2007. **447**(7143): p. 395-395.
6. Mohandas, T., R.S. Sparkes, and L.J. Shapiro, *Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation.* Science, 1981. **211**(4480): p. 393-6.
7. Li, E., C. Beard, and R. Jaenisch, *Role for DNA methylation in genomic imprinting.* Nature, 1993. **366**(6453): p. 362-5.
8. Walsh, C.P., J.R. Chaillet, and T.H. Bestor, *Transcription of IAP endogenous retroviruses is constrained by cytosine methylation.* Nature Genetics, 1998. **20**(2): p. 116-117.
9. Pepin, M.E., et al., *Genome-wide DNA methylation encodes cardiac transcriptional reprogramming in human ischemic heart failure.* Laboratory Investigation, 2019. **99**(3): p. 371-386.
10. Vento-Tormo, R., et al., *DNA demethylation of inflammasome-associated genes is enhanced in patients with cryopyrin-associated periodic syndromes.* Journal of Allergy and Clinical Immunology, 2017. **139**(1): p. 202-211.e6.
11. Ahmed, S.A.H., et al., *The role of DNA methylation in the pathogenesis of type 2 diabetes mellitus.* Clinical Epigenetics, 2020. **12**(1): p. 104.
12. Kerachian, M.A., et al., *Crosstalk between DNA methylation and gene expression in colorectal cancer, a potential plasma biomarker for tracing this tumor.* Scientific Reports, 2020. **10**(1): p. 2813.

13. Moore, L.D., T. Le, and G. Fan, *DNA methylation and its basic function.* Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology, 2013. **38**(1): p. 23-38.

14. Chatterjee, A., E.J. Rodger, and M.R. Eccles, *Epigenetic drivers of tumourigenesis and cancer metastasis.* Semin Cancer Biol, 2018. **51**: p. 149-159.

15. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nature Reviews Genetics, 2012. **13**(7): p. 484-492.

16. Smith, J., et al., *Promoter DNA Hypermethylation and Paradoxical Gene Activation.* Trends Cancer, 2020. **6**(5): p. 392-406.

17. Kim, J., et al., *Aberrant DNA methylation and tumor suppressive activity of the EBF3 gene in gastric carcinoma.* International Journal of Cancer, 2012. **130**(4): p. 817-826.

18. Chatterjee, A., et al., *Genome-wide methylation sequencing of paired primary and metastatic cell lines identifies common DNA methylation changes and a role for EBF3 as a candidate epigenetic driver of melanoma metastasis.* Oncotarget, 2017. **8**(4): p. 6085-6101.

19. Rauluseviciute, I., F. Drabløs, and M.B. Rye, *DNA hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation.* BMC Medical Genomics, 2020. **13**(1): p. 6.

20. Kulis, M. and M. Esteller, *DNA methylation and cancer.* Adv Genet, 2010. **70**: p. 27-56.

21. Yaron, J.R., et al., *A convenient, optimized pipeline for isolation, fluorescence microscopy and molecular analysis of live single cells.* Biological Procedures Online, 2014. **16**(1): p. 9.

22. Tambe, A. and L. Pachter, *Barcode identification for single cell genomics.* BMC Bioinformatics, 2019. **20**(1): p. 32.

23. Guo, C., et al., *CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics.* Genome Biology, 2019. **20**(1): p. 90.

24. Smallwood, S.A., et al., *Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity.* Nature Methods, 2014. **11**(8): p. 817-820.

25. Guo, H., et al., *Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing.* Genome research, 2013. **23**(12): p. 2126-2135.

26. Guo, H., et al., *Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing.* Nat Protoc, 2015. **10**(5): p. 645-59.

27. Mooijman, D., et al., *Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction.* Nat Biotechnol, 2016. **34**(8): p. 852-6.

28. Niemöller, C., et al., *Bisulfite-free epigenomics and genomics of single cells through methylation-sensitive restriction.* Communications Biology, 2021. **4**(1): p. 153.

29. Hu, Y., et al., *Single Cell Multi-Omics Technology: Methodology and Application.* Frontiers in Cell and Developmental Biology, 2018. **6**(28).

30. Stephenson, E., et al., *Single-cell multi-omics analysis of the immune response in COVID-19.* Nature Medicine, 2021. **27**(5): p. 904-916.

31. Blackmore, T., et al., *The characteristics and outcomes of patients with colorectal cancer in New Zealand, analysed by Cancer Network.* N Z Med J, 2020. **133**(1513): p. 42-52.

32. Guraya, S.Y., *Association of type 2 diabetes mellitus and the risk of colorectal cancer: A meta-analysis and systematic review.* World J Gastroenterol, 2015. **21**(19): p. 6026-31.
33. Markowitz, S.D., et al., *Focus on colon cancer.* Cancer Cell, 2002. **1**(3): p. 233-6.
34. Molinari, C., et al., *Heterogeneity in Colorectal Cancer: A Challenge for Personalized Medicine?* Int J Mol Sci, 2018. **19**(12).
35. Boisen, M.K., et al., *Primary tumor location and bevacizumab effectiveness in patients with metastatic colorectal cancer.* Ann Oncol, 2013. **24**(10): p. 2554-2559.
36. Kastan, M.B., C.E. Canman, and C.J. Leonard, *P53, cell cycle control and apoptosis: implications for cancer.* Cancer Metastasis Rev, 1995. **14**(1): p. 3-15.
37. Zheng, C., et al., *Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing.* Cell, 2017. **169**(7): p. 1342-1356.e16.
38. Wahlfors, J., et al., *Genomic hypomethylation in human chronic lymphocytic leukemia.* Blood, 1992. **80**(8): p. 2074-80.
39. Lin, C.H., et al., *Genome-wide hypomethylation in hepatocellular carcinogenesis.* Cancer Res, 2001. **61**(10): p. 4238-43.
40. Grady, W.M. and J.M. Carethers, *Genomic and epigenetic instability in colorectal cancer pathogenesis.* Gastroenterology, 2008. **135**(4): p. 1079-99.
41. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.
42. Rodriguez, J., et al., *Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers.* Cancer Res, 2006. **66**(17): p. 8462-9468.
43. Ashktorab, H., et al., *DNA methylome profiling identifies novel methylated genes in African American patients with colorectal neoplasia.* Epigenetics, 2014. **9**(4): p. 503-12.
44. Cunningham, J.M., et al., *Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability.* Cancer Res, 1998. **58**(15): p. 3455-60.
45. Melotte, V., et al., *N-Myc downstream-regulated gene 4 (NDRG4): a candidate tumor suppressor gene and potential biomarker for colorectal cancer.* J Natl Cancer Inst, 2009. **101**(13): p. 916-27.
46. Esteller, M., et al., *DNA methylation patterns in hereditary human cancers mimic sporadic tumorigenesis.* Hum Mol Genet, 2001. **10**(26): p. 3001-7.
47. Wheeler, J.M., et al., *Hypermethylation of the promoter region of the E-cadherin gene (CDH1) in sporadic and ulcerative colitis associated colorectal cancer.* Gut, 2001. **48**(3): p. 367-71.
48. Azarschab, P., et al., *Epigenetic control of the E-cadherin gene (CDH1) by CpG methylation in colectomy samples of patients with ulcerative colitis.* Genes Chromosomes Cancer, 2002. **35**(2): p. 121-6.
49. Ashktorab, H. and H. Brim, *DNA Methylation and Colorectal Cancer.* Curr Colorectal Cancer Rep, 2014. **10**(4): p. 425-430.
50. Tekle, C., et al., *B7-H3 contributes to the metastatic capacity of melanoma cells by modulation of known metastasis-associated genes.* Int J Cancer, 2012. **130**(10): p. 2282-90.
51. Nygren, M.K., et al., *B7-H3 and its relevance in cancer; immunological and non-immunological perspectives.* Front Biosci (Elite Ed), 2011. **3**: p. 989-93.

52. Belharazem, D., et al., *Carcinoma of the colon and rectum with deregulation of insulin-like growth factor 2 signaling: clinical and molecular implications.* J Gastroenterol, 2016. **51**(10): p. 971-84.
53. Sallustio, F., L. Gesualdo, and A. Gallone, *New findings showing how DNA methylation influences diseases.* World J Biol Chem, 2019. **10**(1): p. 1-6.
54. Chen, W.D., et al., *Detection in fecal DNA of colon cancer-specific methylation of the nonexpressed vimentin gene.* J Natl Cancer Inst, 2005. **97**(15): p. 1124-32.
55. Pixberg, C.F., et al., *Characterization of DNA Methylation in Circulating Tumor Cells.* Genes (Basel), 2015. **6**(4): p. 1053-75.
56. Lyberopoulou, A., et al., *Identification of Methylation Profiles of Cancer-related Genes in Circulating Tumor Cells Population.* Anticancer Res, 2017. **37**(3): p. 1105-1112.
57. Pixberg, C.F., et al., *Analysis of DNA methylation in single circulating tumor cells.* Oncogene, 2017. **36**(23): p. 3223-3231.
58. Giraldo, N.A., et al., *The clinical role of the TME in solid cancer.* British Journal of Cancer, 2019. **120**(1): p. 45-53.
59. Rycaj, K. and D.G. Tang, *Cell-of-Origin of Cancer versus Cancer Stem Cells: Assays and Interpretations.* Cancer research, 2015. **75**(19): p. 4003-4011.
60. Singh, S.K., et al., *Identification of a cancer stem cell in human brain tumors.* Cancer Res, 2003. **63**(18): p. 5821-8.
61. Diehn, M. and M.F. Clarke, *Cancer stem cells and radiotherapy: new insights into tumor radioresistance.* J Natl Cancer Inst, 2006. **98**(24): p. 1755-7.
62. Wick, W., et al., *MGMT testing—the challenges for biomarker-based glioma treatment.* Nature Reviews Neurology, 2014. **10**(7): p. 372-385.
63. Litzenburger, U.M., et al., *Single-cell epigenomic variability reveals functional cancer heterogeneity.* Genome Biology, 2017. **18**(1): p. 15.
64. Haque, A., et al., *A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications.* Genome Medicine, 2017. **9**(1): p. 75.
65. Hou, Y., et al., *Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas.* Cell Res, 2016. **26**(3): p. 304-19.
66. Gaiti, F., et al., *Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia.* Nature, 2019. **569**(7757): p. 576-580.
67. Johnson, K.C., et al., *Single-cell multimodal glioma analyses identify epigenetic regulators of cellular plasticity and environmental stress response.* Nat Genet, 2021. **53**(10): p. 1456-1468.
68. Chen, H., et al., *Single-cell DNA methylome analysis of circulating tumor cells.* Chin J Cancer Res, 2021. **33**(3): p. 391-404.
69. Yuan, G.-C., et al., *Challenges and emerging directions in single-cell analysis.* Genome Biology, 2017. **18**(1): p. 84.
70. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902.e21.
71. žurauskienė, J. and C. Yau, *pcaReduce: hierarchical clustering of single cell transcriptional profiles.* BMC Bioinformatics, 2016. **17**(1): p. 140.
72. Kiselev, V.Y., et al., *SC3: consensus clustering of single-cell RNA-seq data.* Nat Methods, 2017. **14**(5): p. 483-486.

73. Moignard, V., et al., *Decoding the regulatory network of early blood development from single-cell gene expression measurements.* Nat Biotechnol, 2015. **33**(3): p. 269-276.

74. Ocone, A., et al., *Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data.* Bioinformatics, 2015. **31**(12): p. i89-96.

75. Luo, C., et al., *Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex.* Science, 2017. **357**(6351): p. 600-604.

76. Angermueller, C., et al., *Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity.* Nature Methods, 2016. **13**(3): p. 229-232.

77. Kapourani, C.-A. and G. Sanguinetti, *Melissa: Bayesian clustering and imputation of single-cell methylomes.* Genome biology, 2019. **20**(1): p. 61-61.

78. Zamanighomi, M., et al., *Unsupervised clustering and epigenetic classification of single cells.* Nature Communications, 2018. **9**(1): p. 2410.

79. Bravo González-Blas, C., et al., *cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data.* Nat Methods, 2019. **16**(5): p. 397-400.

80. Salomon, R., et al., *Droplet-based single cell RNAseq tools: a practical guide.* Lab Chip, 2019. **19**(10): p. 1706-1727.

81. Clark, S.J., et al., *Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity.* Genome Biol, 2016. **17**: p. 72.

82. Molnár, B., et al., *Gene promoter and exon DNA methylation changes in colon cancer development - mRNA expression and tumor mutation alterations.* BMC Cancer, 2018. **18**(1): p. 695.

83. Turner, J.G., et al., *ABCG2 expression, function, and promoter methylation in human multiple myeloma.* Blood, 2006. **108**(12): p. 3881-9.

84. Gagnon, J.F., et al., *Irinotecan inactivation is modulated by epigenetic silencing of UGT1A1 in colon cancer.* Clin Cancer Res, 2006. **12**(6): p. 1850-8.

85. Roesch, A., et al., *Overcoming intrinsic multidrug resistance in melanoma by blocking the mitochondrial respiratory chain of slow-cycling JARID1B(high) cells.* Cancer Cell, 2013. **23**(6): p. 811-25.

86. Bélanger, A.S., et al., *Regulation of UGT1A1 and HNF1 transcription factor gene expression by DNA methylation in colon cancer cells.* BMC Mol Biol, 2010. **11**: p. 9.

87. Kosuri, K.V., et al., *An epigenetic mechanism for capecitabine resistance in mesothelioma.* Biochem Biophys Res Commun, 2010. **391**(3): p. 1465-70.

88. Hur, K., et al., *Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis.* Gut, 2014. **63**(4): p. 635-46.

89. Orjuela, S., et al., *The DNA hypermethylation phenotype of colorectal cancer liver metastases resembles that of the primary colorectal cancers.* BMC Cancer, 2020. **20**(1): p. 290.

90. Kim, Y.H., et al., *Epigenomic analysis of aberrantly methylated genes in colorectal cancer identifies genes commonly affected by epigenetic alterations.* Ann Surg Oncol, 2011. **18**(8): p. 2338-47.

91. Kim, S.A., et al., *Loss of CDH1 (E-cadherin) expression is associated with infiltrative tumour growth and lymph node metastasis.* Br J Cancer, 2016. **114**(2): p. 199-206.

92. Caldeira, J.R., et al., *CDH1 promoter hypermethylation and E-cadherin protein expression in infiltrating breast cancer.* BMC Cancer, 2006. **6**: p. 48.

93. Kroepil, F., et al., *Down-regulation of CDH1 is associated with expression of SNAI1 in colorectal adenomas.* PLoS One, 2012. **7**(9): p. e46665.

94. Li, X., et al., *MLH1 promoter methylation frequency in colorectal cancer patients and related clinicopathological and molecular features.* PLoS One, 2013. **8**(3): p. e59064.

95. Belshaw, N.J., et al., *Use of DNA from human stools to detect aberrant CpG island methylation of genes implicated in colorectal cancer.* Cancer Epidemiol Biomarkers Prev, 2004. **13**(9): p. 1495-501.

96. Kumar, K., et al., *Distinct BRAF (V600E) and KRAS mutations in high microsatellite instability sporadic colorectal cancer in African Americans.* Clin Cancer Res, 2009. **15**(4): p. 1155-61.

97. Truninger, K., et al., *Immunohistochemical analysis reveals high frequency of PMS2 defects in colorectal cancer.* Gastroenterology, 2005. **128**(5): p. 1160-71.

98. Haraldsdottir, S., et al., *Patients with colorectal cancer associated with Lynch syndrome and MLH1 promoter hypermethylation have similar prognoses.* Genet Med, 2016. **18**(9): p. 863-8.

99. Bettstetter, M., et al., *Distinction of hereditary nonpolyposis colorectal cancer and sporadic microsatellite-unstable colorectal cancer through quantification of MLH1 methylation by real-time PCR.* Clin Cancer Res, 2007. **13**(11): p. 3221-8.

100. de Vogel, S., et al., *MGMT and MLH1 promoter methylation versus APC, KRAS and BRAF gene mutations in colorectal cancer: indications for distinct pathways and sequence of events.* Ann Oncol, 2009. **20**(7): p. 1216-22.

101. Ellison, A.R., J. Lofing, and G.A. Bitter, *Human MutL homolog (MLH1) function in DNA mismatch repair: a prospective screen for missense mutations in the ATPase domain.* Nucleic Acids Res, 2004. **32**(18): p. 5321-38.

102. Kasprzak, A. and A. Adamek, *Insulin-Like Growth Factor 2 (IGF2) Signaling in Colorectal Cancer-From Basic Research to Potential Clinical Applications.* Int J Mol Sci, 2019. **20**(19).

103. Cui, H., et al., *Loss of <i>IGF2</i> Imprinting: A Potential Marker of Colorectal Cancer Risk.* Science, 2003. **299**(5613): p. 1753-1755.

104. Vu, T.H. and A.R. Hoffman, *Promoter-specific imprinting of the human insulin-like growth factor-II gene.* Nature, 1994. **371**(6499): p. 714-7.

105. Baba, Y., et al., *Hypomethylation of the IGF2 DMR in colorectal tumors, detected by bisulfite pyrosequencing, is associated with poor prognosis.* Gastroenterology, 2010. **139**(6): p. 1855-64.

106. Kaneda, A., et al., *Enhanced sensitivity to IGF-II signaling links loss of imprinting of IGF2 to increased cell proliferation and tumor risk.* Proc Natl Acad Sci U S A, 2007. **104**(52): p. 20926-31.

107. Woodson, K., et al., *Loss of insulin-like growth factor-II imprinting and the presence of screen-detected colorectal adenomas in women.* J Natl Cancer Inst, 2004. **96**(5): p. 407-10.

108. Zhao, R., et al., *Loss of imprinting of the insulin-like growth factor II (IGF2) gene in esophageal normal and adenocarcinoma tissues.* Carcinogenesis, 2009. **30**(12): p. 2117-2122.

109. Leick, M.B., et al., *Loss of imprinting of IGF2 and the epigenetic progenitor model of cancer.* Am J Stem Cells, 2012. **1**(1): p. 59-74.

110. Srivastava, M., et al., *H19 and Igf2 monoallelic expression is regulated in two distinct ways by a shared cis acting regulatory region upstream of H19.* Genes Dev, 2000. **14**(10): p. 1186-95.

111. Cui, H., et al., *Loss of imprinting in colorectal cancer linked to hypomethylation of H19 and IGF2.* Cancer Res, 2002. **62**(22): p. 6442-6.

112. Issa, J.P., et al., *Switch from monoallelic to biallelic human IGF2 promoter methylation during aging and carcinogenesis.* Proc Natl Acad Sci U S A, 1996. **93**(21): p. 11757-62.

113. Wang, C., et al., *Potential Therapeutic Targets of B7 Family in Colorectal Cancer.* Front Immunol, 2020. **11**: p. 681.

114. Kraan, J., et al., *Endothelial CD276 (B7-H3) expression is increased in human malignancies and distinguishes between normal and tumour-derived circulating endothelial cells.* Br J Cancer, 2014. **111**(1): p. 149-56.

115. Lee, Y.H., et al., *Inhibition of the B7-H3 immune checkpoint limits tumor growth by enhancing cytotoxic lymphocyte function.* Cell Res, 2017. **27**(8): p. 1034-1045.

116. Lim, S., et al., *Immunoregulatory Protein B7-H3 Reprograms Glucose Metabolism in Cancer Cells by ROS-Mediated Stabilization of HIF1α.* Cancer Res, 2016. **76**(8): p. 2231-42.

117. Picarda, E., K.C. Ohaegbulam, and X. Zang, *Molecular Pathways: Targeting B7-H3 (CD276) for Human Cancer Immunotherapy.* Clin Cancer Res, 2016. **22**(14): p. 3425-3431.

118. Ziller, M.J., et al., *Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing.* Nat Methods, 2015. **12**(3): p. 230-2, 1 p following 232.

119. Vavouri, T. and B. Lehner, *Human genes with CpG island promoters have a distinct transcription-associated chromatin organization.* Genome Biology, 2012. **13**(11): p. R110.

120. Kinoshita, H., et al., *Methylation of the androgen receptor minimal promoter silences transcription in human prostate cancer.* Cancer Res, 2000. **60**(13): p. 3623-30.

121. Kouidou, S., et al., *Non-CpG cytosine methylation of p53 exon 5 in non-small cell lung carcinoma.* Lung Cancer, 2005. **50**(3): p. 299-307.

122. Chekhun, V.F., et al., *Role of DNA hypomethylation in the development of the resistance to doxorubicin in human MCF-7 breast adenocarcinoma cells.* Cancer Lett, 2006. **231**(1): p. 87-93.